Restrictions of Sequence on the Thickness of Globular Protein Molecules

(model/theoretical/polarity of amino acids/random distribution/surface residues)

RONALD E. GATES* AND HARVEY F. FISHER

Veterans Administration Hospital, 4801 Linwood Blvd., Kansas City, Missouri 64128; and the University of Kansas School of Medicine, Lawrence, Kans. 66044

Communicated by David E. Green, March 26, 1971

ABSTRACT A model for globular protein molecules based on a linear and random sequence of polar and nonpolar amino acids was developed. Polar amino acids were assumed to be always in contact with the aqueous solvent, while nonpolar amino acids were assumed to have a tendency to be buried. Considerations of amino-acid dimensions indicated that only runs of four or more nonpolar amino acids in a row could allow some amino acids to be more than 1-nm (10-Å) removed from the surface of the protein. The expected volume fraction of a protein molecule that could be more than 1-nm removed from the surface was obtained with the assumption of a random sequence. Calculation of this volume fraction for a number of simple geometric shapes indicated that some nonpolar amino acids must be exposed to solvent and that the maximum average thickness of globular proteins should be 3-4 nm. Good agreement with published protein dimensions was obtained.

As x-ray diffraction studies continue to reveal the intricate and diverse folded patterns of increasing numbers of protein molecules, the riddle of the rules under which these similar-looking peptide chains assume their precise and amazingly varied conformations has become more puzzling. One of us (1) has developed an approach to the problem that considers a protein molecule as being composed of an outer, exposed layer occupied by polar amino acids and an inner core occupied by nonpolar residues. From the amino-acid composition and molecular volume alone, information concerning the state of association or the shape of the protein molecule was obtained. We propose to extend this approach to the rules governing the folding of a protein molecule by examining the geometrical consequences that arise when the notion of a linear sequence of polar and nonpolar amino acids is introduced.

THE MODEL

In many globular proteins the average residue volume is about 12.5 nm³, which suggests an average residue dimension of 0.5 nm. If we envision a protein molecule as a rigid, three-dimensional geometric shape immersed in an aqueous solution, those amino acids that are in contact with the solvent will lie in a surface layer of average thickness "d" (2), which is about 0.5 nm. The remaining amino acids will be in contact only with other amino-acid residues. The first assumption is that in a protein molecule, polar amino acids are almost always exposed (3, 4), while nonpolar amino acids have a tendency to be buried, as suggested by Kauzmann (5).

Table 1 presents a polarity analysis for amino acids by x-ray diffraction results from several proteins (6-9). The table was compiled by calculation of the ratio of the number buried to the total number of a given amino acid in all of the

proteins considered. Since the α and β chains of hemoglobin and sperm-whale myoglobin form a virtually identical class of structures, and since elastase, chymotrypsin, and trypsin form another such class, the respective numbers for lysozyme and subtilisin were multiplied by three to crudely normalize the data. The assignment of buried and exposed amino acids from x-ray crystallographic data is difficult, and sometimes the residues are listed as exposed, surface, or buried (8). We have treated the surface residues as exposed, since they are certainly in the outer 0.5-nm shell of the protein molecule. While Table 1 is arranged in order from the most polar to the least polar amino acid, there is no clear-cut quantitative jump in the fraction buried that would allow an unambiguous assignment of polarity. Therefore, the precise value at which smaller values imply polar amino acids and larger values imply nonpolar amino acids must be arbitrarily chosen. We have assigned proline, glycine, tryptophan, alanine, phenylalanine, leucine, cysteine, valine, isoleucine, and methionine as nonpolar. Finally, it should be emphasized that polar and nonpolar are defined here to mean residues that have a tendency to be exposed or to be buried, respectively. Since these tendencies are inferred from experimental data, theoretical justification is neither offered nor necessary.

The second assumption in the model is that the polar and nonpolar residues are randomly arranged in the linear sequence. Many investigators have examined known protein sequences for deviations from a random order of residues (10-15). No significant deviations have thus far been found. Avoiding speculations concerning the rationale of this viewpoint, we shall only attempt to justify the assumption as it applies to our model, which uses experimental data. For reasons that will become apparent in the next section, the

TABLE 1. Polarity analysis for the amino acids in human hemoglobin α and β chains and sperm-whale myoglobin (6), in bovine chymotrypsin A, bovine trypsin, and porcine elastase (7), in chicken lysozyme (8), and in subtilisin BPN' (9)

Amino acid	Fraction buried	Amino acid	Fraction buried	Amino acid	Fraction buried
Arg	0	Asp	0.126	Ala	0.468
Glu	0	Thr	0.195	Leu	0.570
Asn	0.019	His	0.222	Cys	0.625
Lys	0.024	Pro	0.273	Val	0.651
Gln	0.044	Gly*	0.275	Ile	0.660
Ser	0.118	Trp	0.333	Met	0.813
Tyr	0.123	Phe	0.353		

* Present Address: St. Jude Children's Research Hospital, Memphis, Tenn. 38011

* Data from subtilisin BPN' was not available and is not included.

number of times a run of four nonpolar amino acids occurs in a real sequence, n_4 , will be compared to the expected value of n_4 for a random sequence. Runs of four refer to overlapping runs, so, for a sequence of five nonpolar amino acids in a row, n_4 is equal to two. If we define the number of polar amino-acid residues in a sequence as " a ," and the number of nonpolar residues as " b ," the total number of residues in a sequence is $a + b$. If we assume a random arrangement of polar and nonpolar amino acids, the expected value of n_4 can be determined from standard probability-theory methods as

$$E(n_4) = [b(b-1)(b-2)(b-3)] / [(a+b)(a+b-1)(a+b-2)]. \quad [1]$$

In Table 2, the actual and expected values of n_4 are compared for several different amino-acid sequences by the use of the previously deduced polarity assignment. The number of standard deviations of the actual value from the expected value is also shown. Since the absolute value of this number of deviations is less than one in about half of the cases, and never exceeds three, this distribution of values seems to follow a normal distribution curve that has 68.3% of the values within one standard deviation and 99.7% of the values within three standard deviations. Because the distribution of values seems to approximate a normal distribution, we conclude that the number of runs of four nonpolar amino acids is adequately determined by assuming a random arrangement of polar and nonpolar amino acids in the sequence. Small sample size, nonindependent samples due to consideration of related protein molecules, and an unknown probability distribution render this conclusion tentative.

THE GEOMETRIC CONSEQUENCES OF THE MODEL

One way of discussing the geometric restrictions imposed on a protein molecule is in terms of thickness. The diameter of a sphere or a cylindrical rod and the height of a cylindrical disk are examples of the thickness of geometric shapes. If we consider the restrictions of sequence, the maximum thickness of a protein molecule according to our model must be related to the number and length of runs of nonpolar amino acids. The observation that all runs of nonpolar amino acids are very short would imply a severe restriction on thickness; if all runs were very long would imply no restriction by sequence. In quantitative terms, we define Δ_2 as the fraction of amino acids in the geometric shape that are not in the surface layer or in a second layer, also of dimension " d ," immediately under the surface layer. The maximum value of Δ_2 is determined by the sequence and can be calculated as follows. Consider the section of sequence ABBBBBA, where A and B refer to polar and nonpolar residues, respectively. If we assume that no nonpolar residue is in the surface layer, and we approximate the dimension of an amino acid in the direction of the chain as $\frac{2}{3}d$, at least three nonpolar residues must be in the second layer. The number of residues that are not required to be in the first or second layer is two. The latter number corresponds exactly to the number of overlapping runs of four nonpolar amino acids in the sequence. Therefore, Eq. 1 for a random sequence can be used to give the expected, maximum fraction of residues not in the first or second layer, $E(\Delta_{2,\max}) = E(n_4)/(a+b)$. If both " a " and " b " are large, and the fraction of nonpolar amino acids is identified as $\Delta = b/(a+b)$, then a very good approximation is

$$E(\Delta_{2,\max}) \cong \Delta^4. \quad [2]$$

TABLE 2. Random sequence analysis of proteins

Protein	n_4	$E(n_4)$	No. of deviations
Myoglobin (sperm whale)	3	9.25	-1.79
Hemoglobin α (human)	2	11.84	-2.58
Hemoglobin β (human)	11	15.19	-1.00
Lysozyme (chicken)	6	7.75	-0.55
Lactalbumin α^* (bovine)	1	4.51	-1.39
Ribonuclease (bovine)	2	2.51	-0.27
Cytochrome c (human)	7	4.65	+0.93
Cytochrome c (<i>Candida krusei</i>)	5	5.58	-0.21
Subtilisin (<i>Bacillus subtilis</i> BPN')	31	25.61	+0.97
Subtilisin (<i>Bacillus subtilis</i> Carlsburg)	34	25.89	+1.46
Glyceraldehyde 3-phosphate dehydrogenase (lobster)	14	28.37	-2.54
Glyceraldehyde 3-phosphate dehydrogenase (porcine)	17	27.37	-1.78
Papain* (papaya)	12	14.21	-0.52
Carboxypeptidase A α^* (bovine)	16	13.60	+0.55
Tryptophan synthetase α (<i>E. coli</i>)	33	34.38	-0.23
Chymotrypsin A (bovine)	18	16.98	+0.22
Trypsin (bovine)	9	14.64	-1.29
Elastase* (porcine)	11	16.49	-1.19
γ G1 immunoglobulin* (Eu)-Light (human)	7	6.64	+0.12
γ G1 immunoglobulin* (Eu)-Heavy (human)	37	24.47	+2.19
Growth hormone (human)	0	7.08	-2.24

* All sequence data is from Dayhoff (16), except for lactalbumin α (8), papain (17), carboxypeptidase A α (18), elastase (7), and γ G1 immunoglobulin (19).

Eq. 2 can be used to relate the maximum thickness of globular protein molecules to the fraction of nonpolar amino acids. If we identify " d^3 " as the average volume of an amino acid and $2yd$ as the thickness of a particular shape,

$$\Delta_2 = [(y-2)/y]^m, \quad [3]$$

where $m = 3$ for one sphere or for several, identical spheres, $m = 2$ for an infinite, uniform rod, and $m = 1$ for an infinite, uniform disk. It must be emphasized that the thickness obtained in each case refers to an average thickness. A collection of nonidentical spheres will have diameters both larger and smaller than the diameter predicted for a collection of identical spheres. Finally, if we combine Eqs. 2 and 3, the expected, maximum, average thickness is given by

$$E(2yd)_{\max} = 4d/(1 - \Delta^{4/m}). \quad [4]$$

This relationship between the fraction of nonpolar amino acids and the thickness is shown in Fig. 1 for the three reference shapes. If the previously deduced polarity assignment is used, most globular proteins have a fraction of nonpolar amino acids between 0.4 and 0.6. With these values, and considering the shape of maximum thickness, i.e., spheres, our model then predicts that the expected, maximum, averaged thickness of globular proteins should be between 3 and 4 nm.

An expected value refers to the expected fraction of amino acids not required to be in the first two layers, as determined from a random sequence. A maximum value indicates that all

FIG. 1. Relationship between the fraction of nonpolar amino acids, Δ , and the thickness, $2yd$, for a collection of identical spheres, for an infinite uniform rod, and for an infinite uniform disk. This relationship is based on the assumption that the fraction of amino acids that are not within $2d = 1$ nm ($10 \text{ \AA}$) of the surface is Δ^4 . The vertical dashed line indicates the thickness at which some nonpolar amino acids must be exposed and the dotted curve is the result for spheres, based on the assumption that all of the nonpolar amino acids are buried.

of this expected fraction has, in fact, been treated as not being in the first two outside layers. Since other constraints may prevent this from being realized, only a maximum thickness can be predicted. Finally, thickness is determined as an average value, since only uniform shapes have been used.

The dotted curve in Fig. 1 shows the case for one sphere or for several identical spheres when sequence is ignored. The maximum, averaged thickness when sequence is not considered is calculated by the assumption that all nonpolar amino acids are buried. If Δ_1 is identified as the fraction of buried amino acids, i.e., the fraction of amino acids not in the outer layer of thickness " d ," it follows that the basic equation in this case is

$$\Delta = \Delta_1 = [(y - 1)/y]^m, \quad [5]$$

where m assumed the same values as in Eq. 3. It is apparent that a random sequence imposes severe restraints on the acceptable dimensions for protein molecules. Furthermore, since the allowed thickness of the three shapes is less when sequence is considered than when all the nonpolar residues are buried, it follows that all of the nonpolar residues cannot be buried. The discontinuity in all three curves in Fig. 1 at 2.2 nm reflects the transition from no nonpolar residues exposed to the necessary exposure of some nonpolar residues. In general, for a given fraction of nonpolar amino acids, an infinite disk can bury more than an infinite rod, which can bury more than a collection of identical spheres. For these three shapes, this implies that the infinite geometric shape of greatest thickness buries the least number of nonpolar residues. Of course, this conclusion does not necessarily apply to finite geometric shapes.

Considering sources of error, we must emphasize that while the assumption of a random sequence is an oversimplification and introduces some degree of error, the ambiguities of the polarity assignment dominate the error analysis. Even the limited amount of data in Table 1 suggest a continuous

range of polarity rather than just two classes. However, modification of the polarity assignment would only introduce quantitative changes, while the qualitative conclusions that a random sequence limits the thickness and forces some nonpolar amino acids to be exposed would be unchanged. Although we have considered only protein molecules formed from one polypeptide chain, virtually no change in the model is required for several chains of reasonable length that form one molecule, if all the chains have random sequences and similar fractions of nonpolar amino acids.

REAL MOLECULES

In this section, the predictions of the model will be compared with data from real protein molecules. A fundamental test of our model involves the comparison of the predicted and actual values of the number of overlapping runs of four buried residues. For myoglobin (6) and lysozyme (8) the predicted values of 9 and 8, respectively, are considerably larger than the actual values of 0 and 2, respectively, determined from the sequence and x-ray diffraction data. This lack of agreement should not be interpreted as a deficiency of the model since the predicted value is claimed to represent only an upper limit for the real value. However, for elastase (7) the predicted value of 16 is less than the actual value of 22. This discrepancy, coupled with the fact that the sequential arrangement of polar and nonpolar residues in elastase has only 11 overlapping runs of four nonpolar amino acids (see Table 2), suggests that a modification of the polarity assignment may be needed, as discussed previously. In any event, small protein molecules do not provide an adequate test of the model. The maximum thickness of a single elastase molecule calculated from a perfect shape can only be 3.6 nm, as compared with a predicted thickness from our model of 3.4 nm. Comparison of actual and predicted dimensions for large protein molecules whose dimensions, if they were perfect spheres, would greatly exceed our predictions is necessary to test the model.

While the sequence of buried and exposed residues is not available for larger proteins, x-ray diffraction results on the dimensions are available. The best known case is the hemoglobin tetramer, which on the basis of its total number of amino acids could approximate a sphere of 5.2 nm diameter. Actually, the tetramer has an internal cavity, of width 0.8–1 nm, that is filled with solvent and lined with polar residues (20). Since the maximum dimension of the tetramer is 6.4 nm (21), the thickness as defined here does not exceed 2.8 nm. If the tetramer is treated as four independent spheres with the fraction of nonpolar amino acids as 0.56, the expected, maximum, average thickness is 3.7 nm. If the tetramer is treated as a circular torus, the predicted thickness is 2.9 nm. It seems that the internal cavity is one solution to the constraint placed on the hemoglobin tetramer by a random sequence of polar and nonpolar amino acids. Another tetrameric molecule that solves the restrictions of sequence by the formation of internal cavities is lactate dehydrogenase (EC 1.1.1.27, ref. 22). Each subunit, of molecular weight 35,000, has a deep groove or cleft on the molecular surface. Furthermore, the four subunits are arranged in the molecule in such a way that three mutually-perpendicular holes of at least 1-nm width pass through the molecule and intersect in the center to form a large cavity. The two proteins, bovine carboxypeptidase A α (EC 3.4.2.1) and human carbonic anhydrase C (EC 4.2.1.1) have 307 and about 260 residues, respectively, from which spheres of diameter 4.2 nm and 4 nm,

respectively, could be formed. However, each protein has a low fraction of nonpolar amino acids. From the amino-acid composition of carboxypeptidase (18), $\Delta = 0.46$. From the amino-acid composition of carbonic anhydrase (23), $\Delta = 0.47$. If we assume a random sequence, the maximum sphere-diameter allowed for these values of Δ is 3.1–3.2 nm. The results of x-ray diffraction studies on carboxypeptidase (24) and on carbonic anhydrase (25) indicate that both molecules have solved the restriction of a random sequence in a similar manner, having their active site in a deep pocket or large cavity with a zinc atom at the bottom and near the center of the molecule.

More data concerning the structure of globular protein molecules is available from electron microscopy. However, the practical limits of resolution of micrographs of negatively stained protein allow only reasonable estimates of dimensions to be made, and holes or cavities in a molecule could go undetected, which would cause the thickness to be overestimated. Therefore, the data from this technique is used with the understanding that the thickness could be less than measured. *Escherichia coli* glutamine synthetase (EC 6.3.1.2) of molecular weight about 600,000 exists as twelve approximately spherical subunits of diameter 4.5 nm, arranged into two hexagonal rings (26). Membrane adenosine triphosphatase (EC 3.6.1.3) from *Streptococcus faecalis* is seen as six approximately spherical subunits of diameter 4 nm, arranged in a hexagonal ring (27). However, it seems to us that a protein of molecular weight about 400,000 would be too large for the proposed structure. Recently two models have been proposed for the structure of the smallest active oligomer (molecular weight about 300,000) of bovine-liver glutamate dehydrogenase (EC 1.4.1.3). One model has six subunits approximating prolate ellipsoids, with twice the minor axis being 4.3 nm (28), arranged in two trimeric rings. The other model has twelve approximately spherical foldons, of diameter about 3.5 nm (29), with their centers located at the vertices of a regular icosahedron. The subunit structure of mammalian fructose diphosphate aldolase (EC 4.1.2.13), of molecular weight about 160,000, is seen as a tetramer of four approximately spherical subunits of diameter 4 nm (30). These data indicate that even large globular proteins are constructed from smaller folded units of dimensions smaller than 4 nm, in accordance with our predictions.

DISCUSSION

We have developed here a model for globular protein molecules based on a random sequence of polar and nonpolar amino acids. While some authors have attempted to relate protein structure and the nonrandomness of sequence, the unique feature of our model is that the random nature of the sequence is one factor that determines the structure, and, in fact, allows the restrictive requirements of sequence to be considered from composition data without knowledge of the exact sequence. The restrictions of a random sequence require that some nonpolar residues be exposed and that the thickness be severely limited. If we assume the fraction of nonpolar amino acids to be between 0.4 and 0.6, the expected, maximum, average thickness of globular proteins should be between 3 and 4 nm. This thickness applies to the diameter of spheres and should be smaller for other shapes, such as a rod or a disk, or even for slightly deformed spheres in contact with one another over a non-negligible fraction of their surface. The improbability of forming perfect, smooth spheres from a collection of amino acids suggests that real thicknesses

will, in general, be less than the predicted maximum thickness. Finally, of course, it should be emphasized that this treatment has not exhausted the restrictions of sequence. We have not considered the limitations imposed by sequence on the folding into close proximity of residues widely separated on the polypeptide chain. Such additional restrictions may require that even more nonpolar amino acids be exposed and that the thickness be less than we have predicted. In conclusion, it does seem, within the limits of the uncertainties, that large globular proteins do comply with the constraints imposed by a random sequence of polar and nonpolar amino acids.

Supported in part by grants from the National Institutes of Health (GM15188) and the National Science Foundation (GB20923).

1. Fisher, H. F., *Proc. Nat. Acad. Sci. USA* **51**, 1285 (1964).
2. Waugh, D. F., *Advan. Protein Chem.*, **9**, 325 (1954).
3. Perutz, M. F., *Eur. J. Biochem.*, **8**, 455 (1969).
4. Klotz, I. M., *Arch. Biochem. Biophys.*, **138**, 704 (1970).
5. Kauzmann, W., *Advan. Protein Chem.*, **14**, 37 (1959).
6. Perutz, M. F., J. C. Kendrew, and H. C. Watson, *J. Mol. Biol.*, **13**, 669 (1965).
7. Shotton, D. M., and B. S. Hartley, *Nature*, **225**, 802 (1970).
8. Browne, W. J., A. C. T. North, D. C. Phillips, K. Brew, T. C. Vanaman, and R. L. Hill, *J. Mol. Biol.*, **42**, 65 (1969).
9. Wright, C. S., R. A. Alden, and J. Kraut, *Nature*, **221**, 235 (1969).
10. Yčas, M., in *Symposium on Information Theory in Biology*, ed. H. P. Yockey, R. L. Platzman, and H. Quastler (Pergamon Press, New York, 1956), p. 78.
11. Urbakh, V. Y., *Bull. Math. Biophys.*, **25**, 213 (1963).
12. Williams, J., J. B. Clegg, and M. O. Mutch, *J. Mol. Biol.*, **3**, 532 (1961).
13. Thiebaut, H. J., and H. H. Pattee, *J. Theor. Biol.*, **17**, 121 (1967).
14. Zimmerman, J. M., N. Eliezer, and R. Simha, *J. Theor. Biol.*, **21**, 170 (1968).
15. Miller, W. G., and C. V. Goebel, *Biochemistry*, **11**, 3925 (1968).
16. Dayhoff, M. O., in *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Silver Spring, Md., 1969), Vol. 4.
17. Husain, S. S., and G. Lowe, *Biochem. J.*, **114**, 279 (1969).
18. Bradshaw, R. A., L. H. Ericsson, K. A. Walsh, and H. Neurath, *Proc. Nat. Acad. Sci., USA*, **63**, 1389 (1969).
19. Edelman, G. M., B. A. Cunningham, W. E. Gall, P. D. Gottlieb, U. Rutishauser, and M. J. Waxdal, *Proc. Nat. Acad. Sci., USA*, **63**, 78 (1969).
20. Perutz, M. F., H. Muirhead, J. M. Cox, L. C. G. Goaman, F. S. Mathews, E. L. McGandy, and L. E. Webb, *Nature*, **219**, 29 (1968).
21. Perutz, M. F., H. Muirhead, J. M. Cox, and L. C. G. Goaman, *Nature*, **219**, 131 (1968).
22. Adams, M. J., D. J. Haas, B. A. Jeffry, A. McPherson, H. L. Mermall, M. G. Rossman, R. W. Schevitz, and A. J. Wonacott, *J. Mol. Biol.*, **41**, 159 (1969).
23. Nyman, P., and S. Lindskog, *Biochim. Biophys. Acta*, **85**, 141 (1964).
24. Reeke, G. N., J. A. Hartsuck, M. L. Ludwig, F. A. Quiocho, T. A. Steitz, and W. N. Lipscomb, *Proc. Nat. Acad. Sci. USA*, **58**, 2220 (1967).
25. Fridborg, K., K. K. Kannan, A. Liljas, J. Lundin, B. Strandberg, R. Strandberg, B. Tilander, and G. Wiren, *J. Mol. Biol.*, **25**, 505 (1967).
26. Valentine, R. C., B. M. Shapiro, and E. R. Stadtman, *Biochemistry*, **6**, 2143 (1968).
27. Schnebli, H. P., A. E. Vatter, and A. Abrams, *J. Biol. Chem.*, **245**, 1122 (1970).
28. Eisenberg, H., and E. Reisler, *Biopolymers*, **9**, 113 (1970).
29. Fiskin, A. M., E. F. J. von Bruggen, and H. F. Fisher, *Biochemistry*, **10**, 2396 (1971).
30. Penhoet, E., M. Kochman, R. Valentine, and W. J. Rutter, *Biochemistry*, **6**, 2940 (1967).