Appendix A

Explicit Evaluation of U_{di}'

$U_{di}' =$

$$\begin{bmatrix} 0 & \tau_2' & \tau_1' & w(\tau_1'\tau_2+\tau_2'\tau_1) \\ \mathcal{S}_2' & m(SM)_2 & \tau_1'\mathcal{S}_2+\tau_1\mathcal{S}_2' & w(SM)_2(m\tau_1+\tau_2') \\ \mathcal{S}_1' & \mathcal{S}_1'\tau_2+\mathcal{S}_1\tau_2' & m(SM)_1 & w(SM)_1(m\tau_2+\tau_2') \\ \mathcal{S}_1'\mathcal{S}_2+\mathcal{S}_1\mathcal{S}_2' & (SM)_2(\mathcal{S}_1'+m\mathcal{S}_1) & (SM)_1(\mathcal{S}_2'+m\mathcal{S}_2) & 2m(SM)_1(SM)_2 \end{bmatrix} \quad \text{(A-1)}$$

The index i has been dropped for convenience. The subscripts 1 and 2 label chains 1 and 2, respectively. $\mathcal{S}_j'$ and τ_j' are defined in eq 31c and 31a, respectively.

If both chains 1 and 2 contain blocks composed at m identical residues (the residues need not be the same in both chains), then

$$\mathcal{S}' = s\frac{d\mathcal{S}}{ds} = G(m,s) \quad \text{(A-2)}$$

$$\tau' = \sigma s\mathcal{S} + \sigma sG(m,s) \quad \text{(A-3)}$$

$$SM = s^m \quad \text{(A-4)}$$

where

$$G(m,s) = \frac{(m-1)s^{m+1} - ms^m + s}{(s-1)^2} \quad \text{(A-5)}$$

Note that

$$G(1,s) = 0 \quad \text{(A-6)}$$

$$G(m,1) = m(m-1)/2 \quad \text{(A-7)}$$

References and Notes

(1) Szent-Györgyi, A. G.; Cohen, C.; Kendrick-Jones, J. *J. Mol. Biol.* **1971**, *56*, 239–58.
(2) Cohen, C.; Holmes, K. *J. Mol. Biol.* **1963**, *6*, 423–32.
(3) Lowey, S.; Kucera, J.; Holtzer, A. *J. Mol. Biol.* **1963**, *7*, 234–44.
(4) Cowgill, R. *Biochemistry* **1974**, *13*, 2467–74.
(5) Cohen, C.; Szent-Györgyi, A. G. *J. Am. Chem. Soc.* **1957**, *79*, 248.
(6) Holtzer, A.; Clark, R.; Lowey, S. *Biochemistry* **1965**, *4*, 2401–11.
(7) Woods, E. *Biochemistry* **1969**, *8*, 4336–44.
(8) Caspar, D.; Cohen, C.; Longley, W. *J. Mol. Biol.* **1969**, *41*, 87–107.
(9) Hodges, R.; Sodek, J.; Smillie, L.; Jurasek, L. *Cold Spring Harbor Symp. Quant. Biol.* **1972**, *37*, 299–310.
(10) Johnson, P.; Smillie, L. *Biochem. Biophys. Res. Commun.* **1975**, *64*, 1316–22.
(11) Lehrer, S. *Proc. Natl. Acad. Sci. U.S.A.* **1975**, *72*, 3377–81.
(12) Stewart, M. *FEBS Lett.* **1975**, *53*, 5–7.
(13) McLachlan, A.; Stewart, M. *J. Mol. Biol.* **1975**, *98*, 293–304.
(14) Stone, D.; Smillie, L. *J. Biol. Chem.* **1978**, *253*, 1137–48.
(15) Mak, A.; Lewis, W.; Smillie, L. *FEBS Lett.* **1979**, *105*, 232–4.
(16) (a) Woods, E. *Aust. J. Biol. Sci.* **1976**, *29*, 405–18. (b) Crimmins, D.; Isom, L.; Holtzer, A. *Comp. Biochem. Physiol.* **1981**, *69B*, 35–46.
(17) Wu, C.-S.; Ikeda, K.; Yang, J.-T. *Biochemistry* **1981**, *20*, 566–70.
(18) Mattice, W.; Srinivasan, G.; Santiago, G. *Macromolecules* **1980**, *13*, 1254–60.
(19) (a) Poland, D. C.; Scheraga, H. A. *Biopolymers* **1965**, *3*, 305. (b) Poland, D. C.; Scheraga, H. A. *Ibid.* **1965**, *3*, 335.
(20) Poland, D.; Scheraga, H. "Theory of Helix-Coil Transitions in Biopolymers"; Academic Press: New York, 1970.
(21) Zimm, B.; Bragg, J. *J. Chem. Phys.* **1959**, *31*, 526–35.
(22) Nagai, K. *J. Phys. Soc. Jpn.* **1960**, *15*, 407.
(23) Lifson, S.; Roig, A. *J. Chem. Phys.* **1961**, *34*, 1963–74.
(24) Gibbs, J.; DiMarzio, E. *J. Chem. Phys.* **1958**, *28*, 1247–8. *Ibid.* **1959**, *30*, 271–82.
(25) Lehman, G.; McTague, J. *J. Chem. Phys.* **1968**, *49*, 3170–9.
(26) Lifson, S.; Zimm, B. *Biopolymers* **1963**, *1*, 15–23.
(27) Lifson, S. *Biopolymers* **1963**, *1*, 25–32.
(28) Lifson, S.; Allegra, G. *Biopolymers* **1964**, *2*, 65–8.
(29) Zimm, B. *J. Chem. Phys.* **1960**, *33*, 1349–56.
(30) Ananthanarayanan, V.; Andreatta, R.; Poland, D.; Scheraga, H. *Macromolecules* **1971**, *4*, 417–24.
(31) Alter, J.; Taylor, G.; Scheraga, H. *Macromolecules* **1972**, *5*, 739–46.
(32) Matheson, R.; Nemenoff, R.; Cadinaux, F.; Scheraga, H. *Biopolymers* **1977**, *16*, 1567–85.
(33) Hecht, M.; Zweiful, B.; Scheraga, H. *Macromolecules* **1978**, *11*, 545–51.
(34) Lehrer, S. *J. Mol. Biol.* **1978**, *118*, 209–26.
(35) Edwards, B.; Sykes, B. *Biochemistry* **1980**, *19*, 2577–83.
(36) Byron, F.; Fuller, R. "Mathematics of Classical and Quantum Physics"; Addison-Wesley: Reading, MA, 1969; Vol. I.
(37) Flory, P. "Statistical Mechanics of Chain Molecules"; Wiley: New York, 1969.
(38) Crothers, D.; Kallenbach, N. *J. Chem. Phys.* **1966**, *45*, 917–27.
(39) Poland, D.; Scheraga, H. *Physiol. Chem. Phys.* **1969**, *1*, 389.
(40) Nagasawa, M.; Holtzer, A. *J. Am. Chem. Soc.* **1964**, *86*, 538–43.
(41) Olander, D.; Holtzer, A. *J. Am. Chem. Soc.* **1968**, *90*, 4549–60.
(42) Mattice, W.; Robinson, R. *Biopolymers* **1981**, *20*, 1421–34.
(43) Mayer, J.; Mayer, M. "Statistical Mechanics"; Wiley: New York, 1940; Chapter 9.
(44) Schultz, G.; Schirmer, R. "Principles of Protein Structure"; Springer-Verlag: New York, 1979.
(45) Chou, P.; Fasman, G. *Biochemistry* **1974**, *13*, 222–45.
(46) Levitt, M. *Biochemistry* **1978**, *17*, 4277–85.
(47) Parry, D. *J. Mol. Biol.* **1975**, *98*, 519–35.
(48) Smillie, L.; Pato, M.; Pearlstone, J.; Mak, A. *J. Mol. Biol.* **1980**, *136*, 199–202.
(49) Crimmins, D.; Holtzer, A. *Biopolymers* **1981**, *20*, 925–50.
(50) Pato, M.; Mak, A.; Smillie, L. *J. Biol. Chem.* **1981**, *256*, 593–601.
(51) Hodges, R.; Saund, A.; Chong, P.; St.-Pierre, S.; Reid, R. *J. Biol. Chem.* **1981**, *256*, 1214–24.

Shape and Surface Features of Globular Proteins

M. Prabhakaran and P. K. Ponnuswamy*

Department of Physics, Autonomous Postgraduate Centre, University of Madras, Tiruchirapalli 620 020, Tamil Nadu, India.

Received May 20, 1981

ABSTRACT: By constructing best-fitting ellipsoidal shapes and computing surface areas from volume packing considerations, we investigate shape anisotropy in globular proteins. The spatial positions from the centroids of the fitted ellipsoids and the water contact areas of the amino acid residues in a set of protein molecules are used to obtain information about the demarcation between buried and exposed spatial levels of the residues, the local surface roughness of the protein, and the intrinsic residue contribution to the molecular shape. Crystal data form the basis for this study.

Introduction

Size and shape are two exceedingly important characteristics of globular proteins in assessing the specificity of their interactions with other molecular entities and consequently, their biological functions. Studies have been made to investigate these subtle features from various points of view. The size and shape characters were studied by analyzing amino acid composition,[1] surface roughness,[2,3] solvent exclusion,[4] surface symmetry,[5] and topological considerations.[6] Spherical representation was found to be

0024-9297/82/2215-0314$01.25/0

Table I
Parameters Pertaining to the Shape and Surface Features in Globular Proteins

protein[a]	molecular surface area A_s	solvent contact area A_c	best-fitting ellipsoidal area A_e	surface roughness index $R_c = A_c/A_s$	shape anisotropy index $R_e = A_e/A_s$
pancreatic trypsin inhibitor	1980.68	1181.98	2257.00	0.597	1.139
high-potential iron protein	2443.08	1433.64	3157.00	0.589	1.292
carp myogen	2889.29	1679.70	3681.23	0.588	1.274
cytochrome c_2	3047.88	1812.22	4215.00	0.595	1.383
ribonuclease S	3245.57	2168.21	5417.00	0.668	1.669
lysozyme	3352.60	2108.75	5228.00	0.629	1.560
flavodoxin	3508.89	2050.26	4263.00	0.584	1.215
staphylococcal nuclease	3632.26	2460.96	6857.00	0.678	1.888
myoglobin	3622.99	2407.12	5203.00	0.664	1.436
concanavalin A	4815.93	2968.67	6431.00	0.616	1.335
α-chymotrypsin	4832.26	3156.64	7473.00	0.653	1.547
elastase	4973.66	3300.35	8558.00	0.664	1.721
subtilisin BPN'	5183.70	3067.43	7808.00	0.592	1.506
thermolysin	5721.77	3554.29	10824.00	0.616	1.875

[a] See Levitt and Greer[11] for the original references of the crystal studies.

an inadequate model for protein shape.[7] In the ellipsoidal model[2] the estimated surface areas were found to be lower than the solvent-accessible areas. These preliminary studies stress the need for further investigations. In a previous report, we reported the spatial behavior of amino acid residues in protein globules in terms of their observed positions and amounts of exposure to solvent by treating the protein as an ellipsoidal matrix. In this paper we try to relate the spatial positions of the residues with their degrees of exposure by way of a linear regression analysis. The amount of anisotropy in the shape is then brought out by the calculated deviations from the assumed linear connection between the spatial positions and the degrees of exposure to solvent by the constituent residues. The contributions of the 20 kinds of amino acid residues to the anisotropic character of the protein are quantitatively evaluated and discussed. The surface roughness in the proteins is studied by computing the exposed contact areas of the residues and the geometrical surface area of the protein.

Methods

Construction of Best-Fitting Shape for a Protein. The best-fitting shape for a protein was constructed by using atomic coordinates taken from crystal structure studies: To begin, the largest and the smallest values in each of the coordinates x, y, and z of the centroids of the residues in the protein molecule and their respective mean values $\bar{x}$, $\bar{y}$, and $\bar{z}$ were determined. Keeping $\bar{x}$, $\bar{y}$, and $\bar{z}$ as the center, we first constructed an approximate ellipsoid with semiaxes a, b, and c, respectively, parallel to x, y, and z, each having an initial length equal to the distance of the farthest residue from the center. This ellipsoid was then iteratively reduced to a size just sufficient to enclose all the residues in the protein molecule. To achieve this result the following steps were followed: (1) The length of each semiaxis was decreased in steps of 0.1 Å until the size of the ellipsoid and the product abc were reduced to a minimum. (2) At each step of the above translational process, the ellipsoid was independently rotated about two of its axes (b and c) up to 90° in steps of 10°, each time checking the abc product and the state of residue accommodation; by this process, the smallest size possible for the ellipsoid just sufficient to accommodate all the residues of the protein corresponding to the possible minimum of the product abc was determined. (3) A further refinement of the size was carried out by repeating the rotational procedure at 1° intervals (up to 10°) and the best ellipsoidal shape for the protein corresponding to the global minimum value of abc was determined.

The spatial position d_i of a residue i with respect to the centroid of the protein is taken to be

$$d_i = \frac{x_i^2}{a^2} + \frac{y_i^2}{b^2} + \frac{z_i^2}{c^2} \quad (1)$$

where x_i, y_i, and z_i are the Cartesian coordinates of the centroid of residue i, the origin coinciding with the protein centroid: d_i varies from 0 to 1 and these extreme values correspond to the most interior and the most exterior residues, respectively. The value $1 - d_i$ then indicates the depth of residue i in the protein matrix from the surface.

Calculation of Molecular Areas. The contact surface area of atoms is a set of disconnected patches representing those portions of the atomic van der Waals surface which are in contact with the surface of a probing agent.[2] These contact areas were calculated with the algorithm of Lee and Richards.[9] The molecular surface area was calculated by using the following equation given by Richards:[2]

$$A_s = 5.67M^{2/3} \quad (2)$$

The surface area of an ellipsoid with semiaxes a, b, and c was calculated from the formula

$$S = 2\pi c^2 + \frac{2\pi ab}{\sin \varphi}[(\sin^2 \varphi)E(\varphi,k) + (\cos^2 \varphi)F(\varphi,k)] \quad (3)$$

where $k = a^2(b^2 - c^2)/b^2(a^2 - c^2)$, $\cos \varphi = c/a$, $F(\varphi,k) = \int_0^{\varphi} d\upsilon/(1 - k^2 \sin^2 \upsilon)^{1/2}$, and $E(\varphi,k) = \int_0^{\varphi} d\upsilon\,(1 - k^2 \sin^2 \upsilon)^{1/2}$. Elliptic integral values E and F were taken from Byrd and Friedman.[10]

Results and Discussion

Shape Anisotropy in Proteins. The surface areas computed by the use of eq 2, the ellipsoidal areas computed by the use of eq 3, and the solvent contact areas computed with a probe radius of 1.4 Å are compiled in Table I for 14 proteins. The computed ellipsoidal volume for a protein inevitably contains voids arising from the anisotropic character of the protein shape. The surface area of a protein, on the other hand, is derived from volume packing considerations and therefore it could expand through unoccupied space to assume the ellipsoidal shape

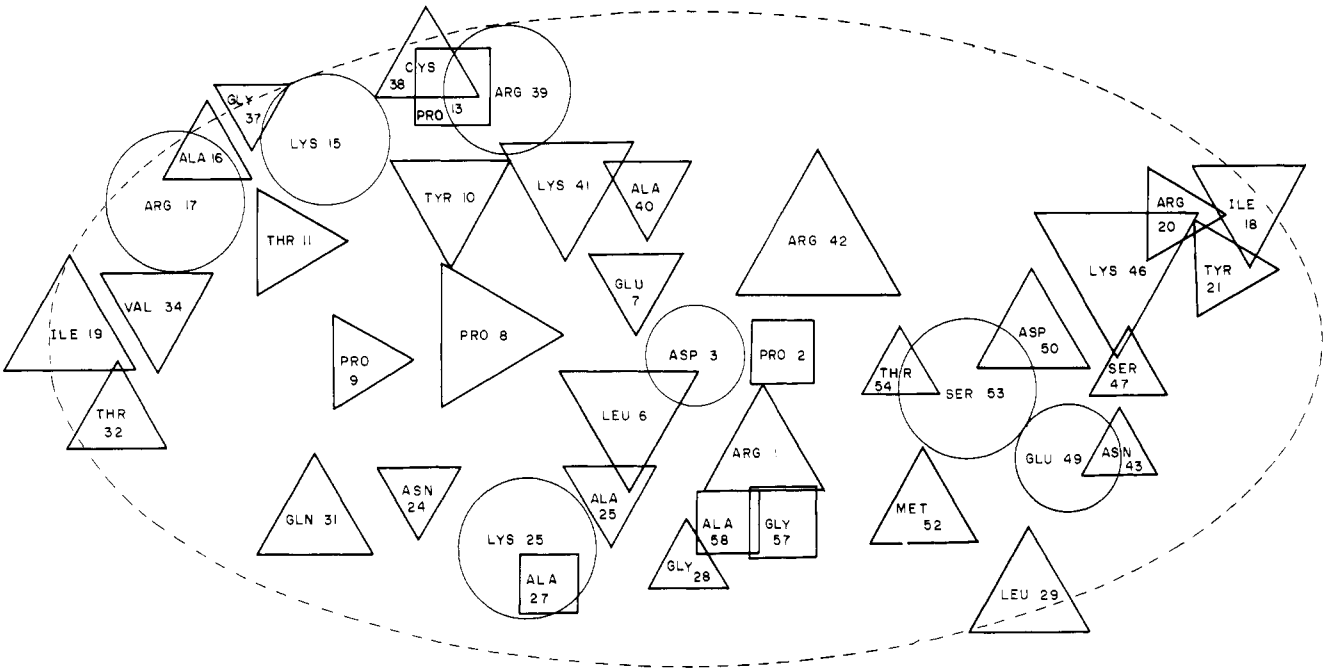


Figure 1. Mollweide projection[12] of amino acid residues in the protein PTI. The amino acid residues are represented by six symbols depending on their spatial positions (circles for residues with d = 0.83–1.0, squares for d = 0.66–0.83, upward pointing triangles for d = 0.5–0.66, downward pointing triangles for d = 0.33–0.5, right-pointing triangles for d = 0.16–0.33, and left-pointing triangles for d = 0.0–0.16); the area of the symbol is taken to be proportional to the extent of the solvent contact area of the residue. Residues with solvent contact areas less than 10 Å^2 have been omitted. The minor axis of the projection represents the z direction of the protein and the major axis represents the angular orientation of the amino acids in the xy plane. The spatial positions of amino acids in the protein are taken as the mean of their atomic coordinates weighted by their accessible contact areas.

with minimum volume enclosing all the residues. The ratio of the ellipsoidal area to the surface area (R_e) thus measures the shape anisotropy in the protein. The computed values of this ratio for the 14 proteins are given in Table I. This anisotropic index varies between 1.14 (PTI) and 1.89 (staphylococcal nuclease and thermolysin), the average value being 1.53. It should be noted that this index increases with molecular weight, although not uniformly. A comparison of the indices for the three proteins elastase, chymotrypsin, and concanavalin A, which are of nearly equal molecular weight, establishes the characteristic shapes of these proteins. Concanavalin A, roughly spherical, has the lowest R_e value among them. The two homologous family proteins elastase and chymotrypsin have different R_e values. Subtilisin BPN′, having a higher molecular weight than elastase and chymotrypsin, has a lower R_e index due to its nearly spherical shape. The largest unoccupied volume, found in the globule of staphylococcal nuclease, a protein of moderate size, is reflected by its higher R_e index. These specific points reinforce the fact that the shape of a globular protein is its characteristic feature, and a general relation between an ellipsoidal area and the real packing volume of the protein could not be realized.[2]

Surface Roughness in Proteins. As already stated, the contact areas are the discontinuous patches which are in contact with the probe molecule. With a plain and smooth surface, the contact area of a protein would be equivalent to its surface area calculated from packing volume, and the contact area would decrease as the surface roughness increases. Therefore the ratio of the contact area to the surface area (R_c) will be a measure of the molecular surface roughness in the respective protein. The values of R_c given in Table I for the 14 proteins (for a probe radius of 1.4 Å) are uniformly smaller than the corresponding anisotropic R_e indices, and the value does not vary significantly from protein to protein. The invariance of this surface roughness index indicates that at the atomic level the surface roughness prevalent is more or less similar in all the proteins and hence it could be taken to be governed mostly by the geometries of the residues and the local interactions. A comparison with the molecular area computed by Finney[3] for the protein ribonuclease S shows that there is a decrease of about 70% in the molecular area for contact with the probe molecule due to its surface roughness, and this result could be taken to be approximately true for other proteins also.

Solvent Contact Areas and Spatial Positions of Residues. In this section, we study the anisotropy in the shapes of globular proteins by computing the relationship between the solvent contact surface areas of residues and their spatial positions in the protein ellipsoids As an illustrative example, we first describe the solvent contact areas of the amino acid residues in the protein PTI and bring out the shape and surface features of this protein. The spatial positions of the residues were computed by the use of eq 1, and the solvent contact areas for the residues were computed by use of a program kindly supplied by Richards. The residues were classified into six groups depending upon their spatial positions and were placed in the ellipsoid following the principles of Mollweide projection:[12] the residues in each group were represented by a specific symbol, the area of the symbol being proportional to the solvent contact area of the respective residue in the group. Such a projection for PTI is shown in Figure 1. For the sake of clarity, the residues with solvent contact areas less than 10 Å^2 were omitted in this figure. We note from this figure a few interesting points as to the shape and surface features of PTI: The residues in the outer zones (the sixth and fifth) invariably possess larger contact areas, except the two smaller members Ala and Gly. Absence of residues in the innermost zone is due to the residues' inaccessibility to solvent. Thr-10, Arg-42, and Lys-46 of interior zones possess relatively larger contact areas, contributing to the irregular shape of the protein. Residues with different exposure and spatial positions (Pro-13,

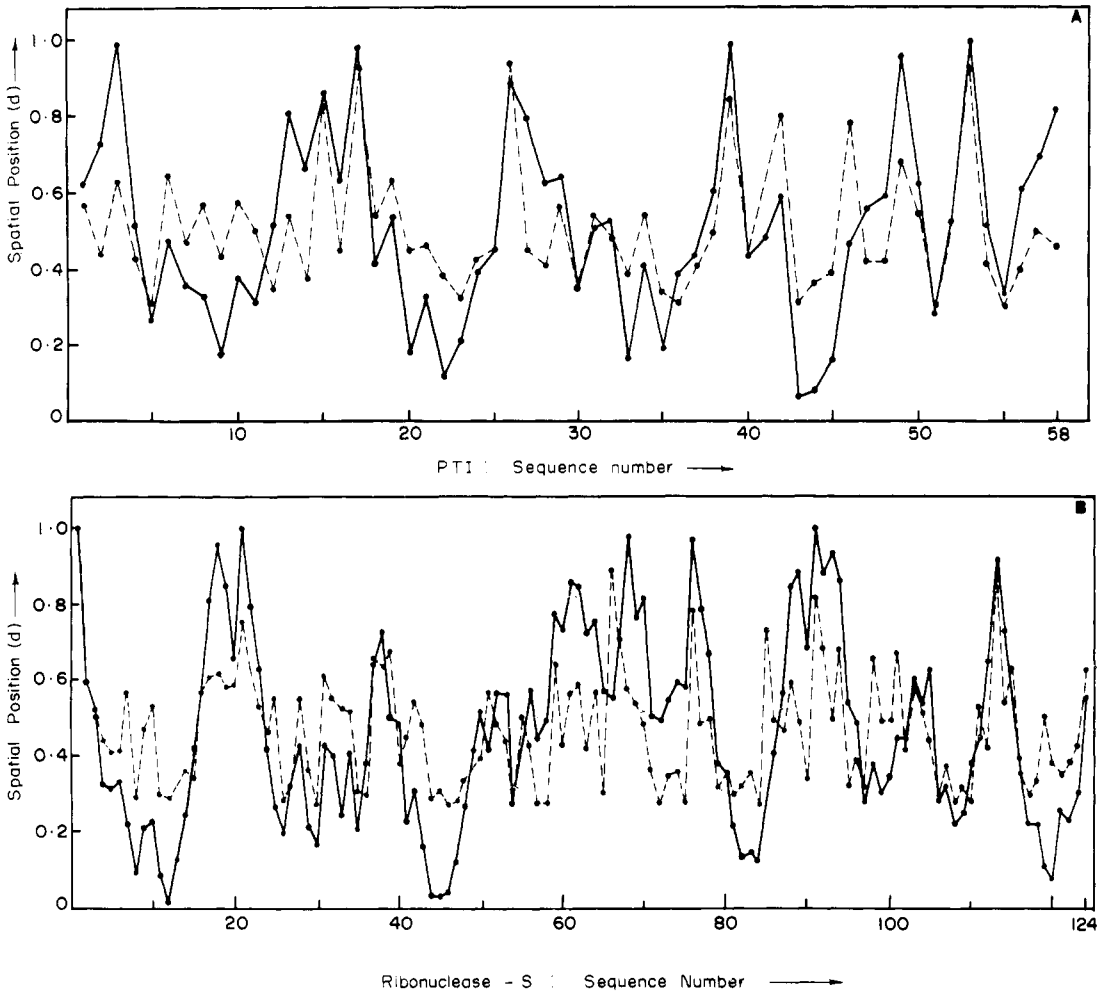


Figure 2. Estimated spatial positions (---) from regression analysis and the observed spatial positions (——) for the residues in the proteins (A) PTI and (B) ribonuclease S.

Arg-17, Ile-19, Thr-32, Val-34, Cys-38, and Arg-39) are also noted; the two interlinked hydrophobic domains[13] around the amino acid residues 20 and 43 are absent: the unoccupied portions in a regular spherical shape could also be noted in this kind of projection.

A calculation of the deviations in the positions of outer residues from the ellipsoidal surface throws light on the role of these residues in the development of irregular surface features in a protein. From the computation of correlations between the various powers of contact areas and spatial positions of residues, i.e., $(A_c)^n$ vs. d_i, in the constructed regular-shaped ellipsoid, it is found that the linear relationship is the best fit between the calculated contact area of a residue and its spatial position. Hence using the solvent contact areas and observed spatial positions for the residues, we carried out a linear regression analysis for each of the 14 proteins. In Figure 2, the estimated spatial positions derived from this regression analysis and the observed spatial positions in the constructed ellipsoid are plotted for the two proteins PTI and ribonuclease S. The similarity between the two profiles for each protein demonstrates the linear relationship between the two parameters. Deviations are also noted at a few residue sites where the estimated exposure is greater in comparison with the noted spatial position. Residue segments 7–11, 20–23, and 43–46 in the plot of PTI and residue segments 7–13, 41–48, 81–84, and 117–122 in the plot of ribonuclease S indicate a continuous difference, pointing out the high deviations from the regular shape in these regions.

Table II
Results of Linear Regression Analysis on the Data of Solvent Contact Areas and Spatial Positions of Residues in Various Globular Proteins

protein	inter-cept	slope	standard dev	corr coeff
pancreatic trypsin inhibitor	0.306	0.0099	0.0246	0.671
high-potential iron protein	0.274	0.0161	0.0216	0.712
carp myogen	0.358	0.012	0.0196	0.612
cytochrome c_2	0.336	0.0112	0.0176	0.645
ribonuclease S	0.273	0.0112	0.0182	0.632
lysozyme	0.307	0.0097	0.0169	0.6200
flavodoxin	0.352	0.013	0.0151	0.705
staphylococcal nuclease	0.246	0.0091	0.0162	0.608
myoglobin	0.341	0.0114	0.0145	0.683
concanavalin A	0.310	0.0088	0.0138	0.501
α-chymotrypsin	0.320	0.0097	0.0122	0.595
elastase	0.333	0.0093	0.0152	0.320
subtilisin BPN'	0.298	0.0136	0.0105	0.682
thermolysin	0.321	0.0113	0.0111	0.574

The regression lines for the 14 proteins are displayed in Figure 3 and the slopes and intercepts of these lines together with the corresponding standard deviations and the correlation coefficients are given in Table II. An analysis of the data of Table II and a look at Figure 3 provide valuable information as to the shapes of the various proteins. The intercept of the regression line for a protein in effect indicates that the residues having depth

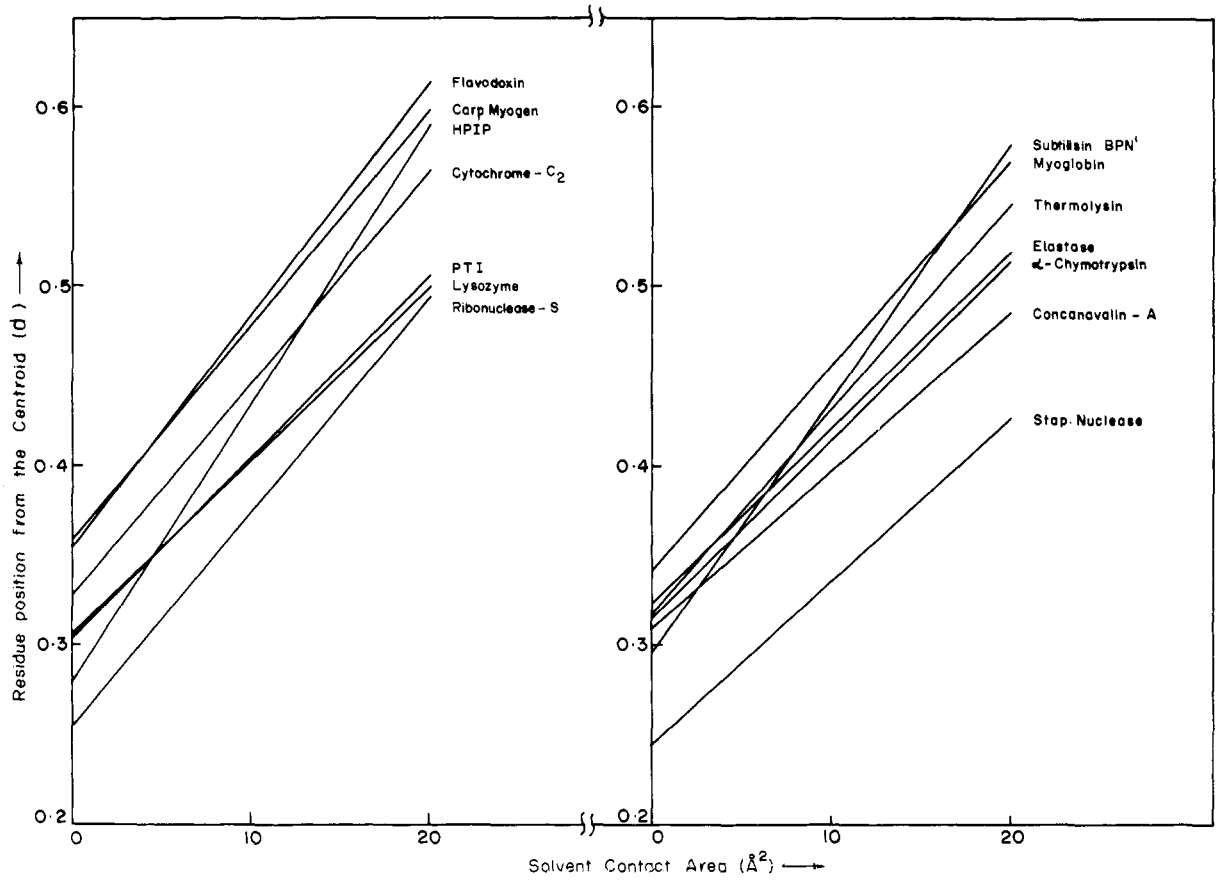


Figure 3. Linear regression lines for the 14 proteins (displayed in two sets for the sake of clarity) obtained from the data of contact areas and spatial positions of amino acid residues.

up to this value are completely buried, and exposure to solvent starts from this demarcation level. The slope indicates the rate at which the exposure increases the positional coordinate. A comparison of the regression lines of the two smallest proteins, PTI and high-potential iron protein, interestingly points out that the inhibitor protein is tightly packed compared to the iron protein, whereas the latter exposes its residues to a larger extent than the former for the same depth. Lysozyme and ribonuclease S, two proteins of nearly equal number of residues, again exhibit characteristic differences in their shapes. The proteins concanavalin A, chymotrypsin, and elastase have similar features, but elastase has a low correlation coefficient and a high standard deviation due to its greater anisotropy. Subtilisin BPN′, a high molecular weight protein, has a low intercept and a high slope. The protein staphylococcal nuclease, with appreciable unoccupied volume, has comparatively lower values for both its intercept and slope. Using the contact area and depth parameters for all the residues in the 14 proteins, we computed a general regression line. This line has an intercept of 0.33 and a slope of 0.0095. These two parametrical values correspond, respectively, to the starting spatial position for solvent exposure, i.e., the demarcation point between the interior and exterior parts of the protein globule and the factor to relate solvent contact area to spatial position. This demarcation point agrees fairly well with the demarcation level predicted for a spherical protein by Janin.[14] The correlation coefficient between the two parameters (contact area and depth) for the residues in the 14 proteins was found to be 0.58. The average parametrical values of the slope and intercept of the regression line could be used to obtain information as to the completely buried zone and the exposure pattern of residues in any theoretical protein model geometry, provided the probable axial values of the ellipsoid are known from preliminary experimental studies (such as hydrodynamic and light scattering) on the protein.

Residue Contributions to Protein Anisotropy. In order to bring out the role of individual amino acid residues in the development of the shapes of the proteins, a linear regression analysis was performed on the spatial positions and contact areas of individual residues in the 14 proteins put together. The computed regression lines for the 20 amino acid residues are displayed in Figure 4, and the corresponding values of intercepts, slopes, standard deviations, and correlation coefficients are given in Table III. A study of Figure 4 and Table III provides valuable information: Among the polar residues, His, Arg, Thr, and Gln have lower intercept values, indicating their ability to be accessible to solvent even at highly interior positions, while Asp, Asn, and Lys have relatively higher intercept values, showing their inability to be interior-cum-accessible members; Lys has the lowest slope and highest intercept, indicating that this residue is always exposed to solvent to the maximum extent. Of the nonpolar residues, Cys and Met have small slope values, indicating their buriedness at all times; Ile, Leu, and Val are more or less similar in their spatial to exposure behavior; Pro has an intercept value to demonstrate its surface-preferring character; Trp and Tyr are more or less comparable with the behavior of Glu in terms of exposure demarcation in spatial position, but the values of their slopes show that their exposure is high at lower spatial position in comparison with Glu. The values of the standard deviations and correlation coefficients further confirm that Lys and Asn are always loosely packed with greater exposure, and Cys and Met are always tightly packed with lower exposure. The smaller sized residues Gly, Ala, and Ser show the best correlation between their spatial positions and solvent contact areas,

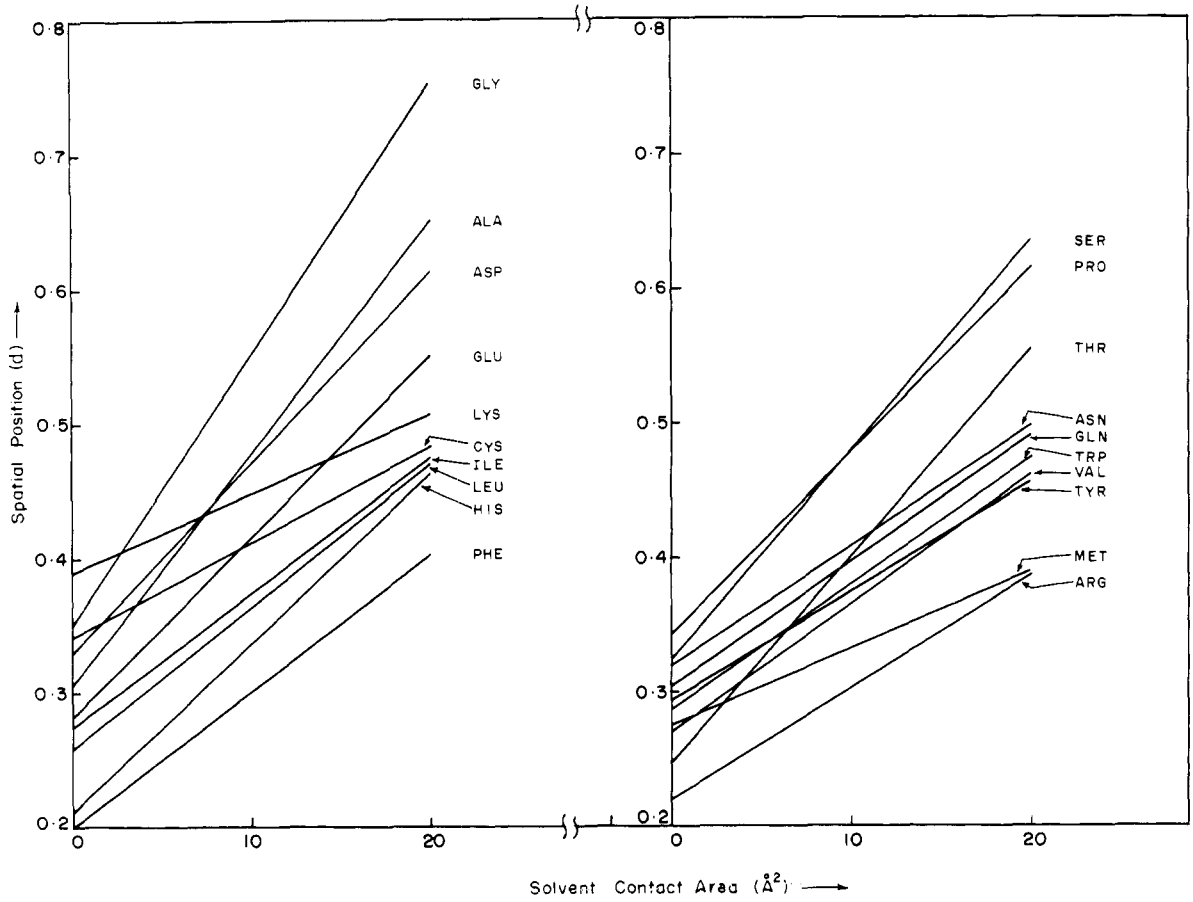


Figure 4. Linear regression lines for the 20 amino acid residues (displayed in two sets for the sake of clarity) obtained from the data of contact areas and spatial positions of the residues in 14 proteins put together.

Table III
Average Parameters from the Regression Analysis on Solvent Contact Area and Spatial Positions of Amino Acid Residues in Globular Proteins

residue	intercept	slope	stand dev	corr coeff	av of sum of square of dev	av of sum of dev
Ala	0.305	0.0175	0.0123	0.687	0.0439	−0.0525
Asp	0.335	0.0140	0.0173	0.632	0.043	−0.0627
Cys	0.339	0.0074	0.0276	0.263	0.0517	0.008
Glu	0.282	0.0135	0.0177	0.669	0.0355	−0.0047
Phe	0.195	0.0104	0.0182	0.577	0.0514	0.143
Gly	0.352	0.0201	0.0144	0.670	0.0638	−0.106
His	0.215	0.0125	0.0285	0.594	0.0642	0.094
Ile	0.278	0.0100	0.0166	0.564	0.0449	0.0581
Lys	0.391	0.0058	0.0142	0.407	0.0474	0.0944
Leu	0.262	0.0104	0.0137	0.541	0.0349	0.0541
Met	0.280	0.0054	0.0260	0.328	0.0676	0.108
Asn	0.322	0.0090	0.0189	0.489	0.0588	0.023
Pro	0.346	0.0136	0.0227	0.600	0.056	−0.063
Gln	0.306	0.0093	0.0207	0.527	0.042	0.04
Arg	0.227	0.0083	0.0239	0.590	0.039	0.112
Ser	0.326	0.0155	0.0127	0.692	0.038	−0.077
Thr	0.251	0.0152	0.0143	0.713	0.031	−0.014
Val	0.291	0.0096	0.0218	0.529	0.028	0.0466
Trp	0.291	0.0092	0.0236	0.632	0.021	0.053
Tyr	0.293	0.0081	0.0206	0.495	0.037	0.053

meaning that they could appear anywhere within the protein matrix with corresponding exposure. The slope and intercept values of the residues are found to be unique in their own respects and unrelated to any individual physicochemical property of the residues.

Statistical Significance of the Results. To evaluate the possible deviations from the assumed linear relationship between spatial positions and contact surface areas of residues, two statistical test quantities, the square of the deviation and the average of the sum of algebraic deviations of each amino acid in the individual protein study, were computed and are given in Table III. These parameters provide additional information with regard to the contributions of residues to the protein shapes. The performances of Val, Ser, and Leu are uniformly good, both as individuals (as revealed from the individual standard deviations) and in their behavior with other residues (as shown by the average of the sum of the squares of the deviations) in the proteins. But high deviations are found for His and Met in both these aspects. While the performance of Gly by itself is good, it shows high deviations in connection with other residues, whereas a reverse performance is found with Trp. The average of the algebraic sum of the deviations of estimated values from their spatial positions in ellipsoids consists of both negative and positive values. The negative values of Ala, Asp, Glu, Gly, Pro, Ser, and Thr show that their spatial positions are outside their estimated positions from the exposure measurement. The size could be attributed as the reason for the residues Ala, Gly, Ser, and Thr, whereas the acidic residues, due to their charged atoms, place themselves well in the exterior, burying the other atoms inside. The bend-forming residue Pro prefers exterior positions in spite of its relatively low accessibility. It is found that these residues invariably possess high values for the slope in comparison to the reported average slope value for the proteins. The high positive values for Phe and Met indicate that these two residues could themselves prefer to be more buried than the extent revealed by their solvent contact areas. The positive value for His is due to the difference in the intercept values of this residue and that of the general regression line and the positive value for Lys is due to the

difference in the corresponding slope values.

Conclusion

The present investigation confirms the fact that it is not possible to describe the shape of the globular protein by means of a well-defined geometry. Construction of best-fitting ellipsoids and the calculation of solvent contact areas for the residues in a set of globular proteins demonstrate a definite relationship between the contact areas of residues and their spatial positions from the centroids of the ellipsoids. The presence of clefts and crevices in the protein globule is due to the peculiar placement of certain residues against their intrinsic spatial preferences. The varying exposure behavior of residues determines the prevailing surface shape. We note, more or less, a uniform atomic level surface roughness and a differential macroscopic shape in the proteins, which means that the local short-range interactions influence the surface roughness whereas the long-range interactions influence the shape.

Acknowledgment. This work was partly supported by a grant to P.K.P. from the Department of Science and Technology, Government of India.

References and Notes

(1) Gibbs, A. J.; Knowles, G. R. *J. Gen. Virol.* **1977**, *38*, 167–168.
(2) Richards, F. M. *Annu. Rev. Biophys. Bioeng.* **1977**, *6*, 151–176.
(3) Finney, J. L. *J. Mol. Biol.* **1978**, *119*, 415–441.
(4) Greer, J.; Bush, B. L. *Proc. Natl. Acad. Sci. U.S.A.* **1978**, *75*, 303–307.
(5) Morgan, R. S.; MCadon, J. M. *Int. J. Pept. Protein Res.* **1980**, *15*, 177–180.
(6) White, E. J. M. *Biochem. J.* **1980**, *187*, 297–302.
(7) Gates, R. K. *J. Mol. Biol.* **1979**, *127*, 345–351.
(8) Prabhakaran, M.; Ponnuswamy, P. K. *J. Theor. Biol.* **1980**, *87*, 623–637.
(9) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379–400.
(10) Byrd, P. F.; Friedman, M. D. "Handbook of Elliptic Integrals for Engineers and Scientists"; Springer-Verlag: New York, 1971.
(11) Levitt, M.; Greer, J. *J. Mol. Biol.* **1977**, *114*, 181–239.
(12) Kellaway, G. P. "Map Projections"; Methuen and Co. Ltd.: London, 1970.
(13) Ponnuswamy, P. K.; Prabhakaran, M. *Biochem. Biophys. Res. Commun.* **1980**, *57*, 1582–1590.
(14) Janin, J. *Nature (London)* **1971**, *277*, 491–492.

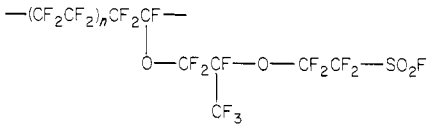
Crystallinity in Perfluorosulfonic Acid Ionomers and Related Polymers

Howard W. Starkweather, Jr.

E. I. du Pont de Nemours and Company, Central Research and Development Department, Experimental Station, Wilmington, Delaware 19898. *Received October 14, 1981*

ABSTRACT: In spite of a large number of bulky side groups, perfluorosulfonic acid ionomers and their precursors have significant levels of crystallinity. The crystal structure appears to be similar to that of poly(tetrafluoroethylene) at elevated temperatures. The size of the crystallites is greater than the average separation between side groups. A model is suggested which is consistent with these observations. The polymer is thought to be arranged in bilayers with ionizable side groups extending on either side. An optical diffraction pattern based on this model is similar to the X-ray diffraction patterns. In earlier work it was concluded that ionic domains are separated by walls 10 Å thick, the correct magnitude for two layers of fluorocarbon. The model is also consistent with the results of recently published small-angle X-ray experiments.

Perfluorosulfonic acid ionomers are derived from copolymers of tetrafluoroethylene (TFE) and a perfluorovinyl ether terminating in a sulfonyl fluoride group and having the structure



Since the molecular weights of TFE and the comonomer are 100 and 446, respectively, the equivalent weight (EW) per sulfonyl fluoride group is

$$EW = 100n + 446$$

The copolymers of greatest interest have about 6–13 TFE units per comonomer unit.

The sulfonyl fluoride groups can be hydrolyzed to give sulfonic acid groups or various metal sulfonates. The acid and salt forms are hydrophilic. Gierke[1] concluded that the ionic domains are about 40 Å in diameter and are separated by walls of fluorocarbon about 10 Å thick.

The samples discussed here were prepared in the same manner as those in earlier work by Gierke and co-workers.[1,2]

Table I
DSC Data

EW	T_m, °C peak/end	ΔH_f, cal/g
	SO_2F Form	
1050	236/254	1.16
1100	231/257	1.48
1350	251/278	3.34
1500	256/283	4.13
1790	266/286	6.92
	SO_3H Form	
1100	207/235	1.06
1350	231/253	1.68
1500	249/265	4.98

Degree of Crystallinity

Data from DSC scans at 20 °C/min on a series of copolymers in the sulfonyl fluoride form are summarized in Table I. As shown in Figure 1, the heat of fusion (ΔH_f) is a linear function of equivalent weight and extrapolates to zero at EW = 910 or a TFE/comonomer mole ratio of about 5. Thus, some crystallinity remains at a very high concentration of bulky side groups. The melting point also increases with increasing equivalent weight. The data are

0024-9297/82/2215-0320$01.25/0