Hickey, D. R., & Turner, D. H. (1985) *Biochemistry 24*, 2086–2094.

Johnston, P. D., & Redfield, A. G. (1981) *Biochemistry 20*, 1147–1156.

Kierzek, R., Caruthers, M. H., Longfellow, C. E., Swinton, D., Turner, D. H., & Freier, S. M. (1987) *Biochemistry* (in press).

Kim, S. H., Suddath, F. L., Quigley, G. J., McPherson, A., Sussman, J. L., Wang, A. H. J., Seeman, N. C., & Rich, A. (1974) *Science (Washington, D.C.) 185*, 435–440.

Ladner, J. E., Jack, A., Robertus, J. D., Brown, J. S., Rhodes, D., Clarke, B. F. C., & Klug, A. (1975) *Proc. Natl. Acad. Sci. U.S.A. 72*, 4414–4418.

Markiewicz, W. T., Biala, E., & Kierzek, R. (1984) *Bull. Pol. Acad. Sci., Chem. 32*, 433–451.

Matteucci, M. D., & Caruthers, M. H. (1980) *Tetrahedron Lett. 21*, 719–722.

Noller, H. F. (1984) *Annu. Rev. Biochem. 53*, 119–162.

Patel, D. J., Kozlowski, S. A., Marky, L. A., Rice, J. A., Broka, C., Dallas, J., Itakura, K., & Breslauer, K. J. (1982) *Biochemistry 21*, 437–444.

Petersheim, M., & Turner, D. H. (1983) *Biochemistry 22*, 256–263.

Richards, E. G. (1975) in *Handbook of Biochemistry and Molecular Biology*: *Nucleic Acids* (Fasman, C. D., Ed.) 3rd ed., Vol. I, p 197, CRC, Cleveland, OH.

Romaniuk, P. J., Hughes, D. W., Gregoire, R. J., Neilson, T., & Bell, R. A. (1979a) *J. Chem. Soc., Chem. Commun.*, 559–560.

Romaniuk, P. J., Hughes, D. W., Gregoire, R. J., Bell, R. A., & Neilson, T. (1979b) *Biochemistry 18*, 5109–5116.

Salser, W. (1977) *Cold Spring Harbor Symp. Quant. Biol. 42*, 985–1002.

Sprinzl, M., Möll, J., Meissner, F., & Hartman, T. (1985) *Nucleic Acids Res. 13*, r1–r49.

Steger, G., Hofmann, H., Förtsch, J., Gross, H. J., Randles, J. W., Sänger, H. L., & Riesner, D. (1984) *J. Biomol. Struct. Dyn. 2*, 542–571.

Sussman, J. L., & Kim, S.-H. (1976) *Biochem. Biophys. Res. Commun. 68*, 89–96.

Tinoco, I., Jr., Borer, P. N., Dengler, B., Levine, M. D., Uhlenbeck, O. C., Crothers, D. M., & Gralla, J. (1973) *Nature (London), New Biol. 246*, 40–41.

Tsong, T. Y., Hearn, R. P., Wrathall, D. P., & Sturtevant, J. M. (1970) *Biochemistry 9*, 2666–2677.

Uhlenbeck, O. C., & Gumport, R. I. (1982) *Enzymes (3rd Ed.) 15*, 31–58.

Uhlenbeck, O. C., Martin, F. H., & Doty, P. (1971) *J. Mol. Biol. 57*, 217–229.

Woese, C. R., Gutell, R., Gupta, R., & Noller, H. F. (1983) *Microbiol. Rev. 47*, 621–669.

Compact Units in Proteins[†]

Micheal H. Zehfus and George D. Rose*

Department of Biological Chemistry, The Milton S. Hershey Medical Center, The Pennsylvania State University, Hershey, Pennsylvania 17033

Received November 5, 1985; Revised Manuscript Received April 23, 1986

ABSTRACT: An explicit measure of geometric compactness called the coefficient of compactness is introduced. This single value figure of merit identifies those continuous segments of the polypeptide chain having the smallest solvent-accessible surface area for their volume. These segments are the most compact units of the protein, and the larger ones correspond to conventional protein domains. To demonstrate the plausibility of this approach as a method of identifying protein domains, the measure is applied to lysozyme and ribonuclease to discover their constituent compact units. These units are then compared with domains, subdomains, and modules found by other methods. To show the sensitivity of the method, the measure is used to successfully differentiate between native and deliberately misfolded proteins [Novotný, J., Bruccoleri, R., & Karplus, M. (1984) *J. Mol. Biol. 177*, 787–818]. Methods that utilize only backbone atoms to define domains cannot distinguish between authentic and misfolded molecules because their backbone conformations are virtually superimposable. Compact units identified by this method exhibit a hierarchic organization. Such an organization suggests possible folding pathways that can be tested experimentally.

The exposure of nonpolar residues to solvent is thermodynamically disfavored in both model compounds and proteins (Kauzmann, 1959; Nozaki & Tanford, 1971). As a result, the nonpolar residues of proteins tend to aggregate, minimizing the total hydrophobic surface area they expose to solvent. This phenomenon, known as the hydrophobic effect, is thought to be a major force driving the protein-folding process (Kauzmann, 1959; Tanford, 1980). Consistent with this idea, the average surface area that residues lose upon folding is found to scale linearly with hydrophobicity (Rose et al., 1985). For protein molecules then, a strong correlation exists between hydrophobicity and surface area. This correlation suggests, in turn, an underlying relationship between conformational free energy and molecular surface area.

In this paper, a normalized measure of molecular surface area is introduced. The parameter, called the coefficient of compactness, is applied to X-ray-elucidated proteins to identify those continuous-chain regions having the smallest accessible surface area for their volume. Geometrically, these units are compact, while, energetically, they have efficiently minimized their conformational free energy through the reduction of exposed surface area.

[†] This research was supported by grants from the National Institutes of Health [GM10824 (M.H.Z.) and GM29458 (G.D.R.)].

[*] Author to whom correspondence should be addressed.

0006-2960/86/0425-5759$01.50/0

These compact, low-energy substructures are likely candidates to be folding intermediates and to exhibit autonomous structural stability. As such, they satisfy the usual definition for protein domains. However, the word "domain" has often been used casually in recent literature, and the term has been adapted to describe a wide variety of ideas. To avoid confusion, we refer to the "domains" found here as "compact units".

The method of classifying units by their coefficient of compactness offers two major advantages over many of the existing methods used to find domains. First, units may be of any size, with segment termini determined independently for each unit. Thus, unit definitions are specified by the molecular structure itself, without implicit constraints imposed by the algorithm. Second, the compactness of the units is explicitly quantitated. Quantification allows the discovered units to be evaluated and ranked numerically.

Compactness should not be confused with the familiar idea of packing density (Richards, 1977). The relationship between compactness and packing density is discussed in the following section of this paper, after which the coefficient of compactness is defined in detail.

Next, the utility of the approach is demonstrated by identifying all compact units in two familiar proteins, lysozyme and ribonuclease. The discovered units are seen to be physically meaningful, and the larger ones are in good agreement with domains found by several other methods.

The sensitivity of this approach is then illustrated by using it to differentiate between correctly and deliberately misfolded protein structures. Misfolded molecules were generated by using pairs of proteins of known structure and identical sequence length (Novotný et al., 1984). To create misfolded molecules, each protein's native backbone conformation was used, in turn, as a structural template for the other protein. The conformations of both the native template and its misfolded analogue were then adjusted to minimize their internal energies. After this regime, the backbone conformations of the misfolded analogues bear no resemblance to their own native structures but, instead, retain conformations similar to their templates. These structures cannot be differentiated energetically, since both have plausible internal energies (op. cit.). Nevertheless, the method used here can successfully distinguish authentic molecules from misfolded ones by evaluating the compactness of the constituent units.

Finally, the existence of hierarchic relationships among units is explored by displaying a protein's complete set of compact units in a hierarchic organizational plot. Analysis of these plots suggests possible folding pathways. Some experiments designed to detect these folding pathways are also suggested.

Compactness and Packing Density

Packing density, defined as the ratio of summed atom volumes to total protein volume, has often been emphasized in the geometric analysis of X-ray-elucidated protein structures. The observed packing density of globular proteins is remarkably efficient, rivaling that of small organic crystals (Richards, 1977). However, packing density is not the only factor involved in the evaluation of compactness. When packing efficiency is uniform, the overall shape of an object becomes the facor that determines compactness. This point is illustrated in Figure 1. Here, the two arrangements of cubes have identical volumes and packing densities, but intuitively, arrangement B is the more compact.

In a set of equal-volume objects, the member with the smallest surface area is the most compact. In Figure 1, B is more compact than A because it has less surface area. Since it can be proven that a sphere has the smallest surface area for its enclosed volume, an arbitrary object can be ranked against this standard. A simple way to effect this ranking is to divide the surface area of an arbitrary object by the surface area of the sphere of equal volume. The resultant dimensionless ratio is the measure of compactness used here; it is termed the coefficient of compactness.

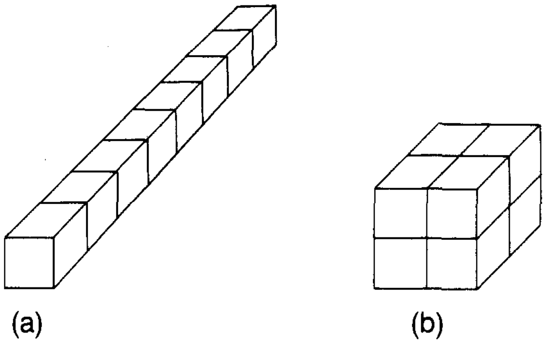

FIGURE 1: Two arrangements of identical cubes with equal volume and packing density but differing compactness. Arrangement B, with smaller surface area, is the more compact.

Coefficient of Compactness

Geometrically, the coefficient of compactness is the surface area of a segment normalized by the minimum possible surface area. Numerically, the coefficient of compactness is a proportionality constant relating a segment's actual surface area to its minimum value. This measure of compactness is universally applicable and is used here at the level of molecular structure.

Given an arbitrary chain segment of known structure, we define its *coefficient of compactness*, *Z*, as

$$Z = \frac{\text{accessible surface area of segment}}{\text{surface area of sphere of equal volume}} \quad (1)$$

This parameter is not new, having been described previously as a roughness index (Richards, 1977), a globularity index (Wodak & Janin, 1981), and a possible means of finding protein domains (Wetlaufer, 1973).

When evaluating surface roughness, Richards (1977) observed that all proteins, regardless of size, have a roughness index of approximately 2. While closely related, this index is not identical with the coefficient of compactness because it is calculated in a slightly different manner. Both the roughness index and the coefficient of compactness are determined by dividing an accessible surface area by the surface area of a normalizing sphere. However, the two methods employ differing spheres. The roughness index uses the sphere with volume equal to the molecular volume of the unit, while the coefficient of compactness uses a sphere with the volume that is enclosed within the accessible surface. The latter volume includes a region that extends 1.4 Å beyond the molecular surface and is therefore larger than the molecular volume. The normalizing term (i.e., the denominator) in the coefficient of compactness is thus larger than the one used in the roughness index, and consequently, the coefficient of compactness for a unit will be smaller than its roughness index. The average *Z* value for a set of 20 globular proteins is 1.64 ± 0.08 (data not shown).

Methods

The ideal way to find compact units using the coefficient of compactness is to evaluate *Z* for every continuous-chain segment of every size and then to select units with minimum *Z* values as compact units. These calculations, however, would require considerable computer time, making this approach unwieldy. Alternatively, all continuous segments can be screened to eliminate the bulk of noncompact chains, leaving a limited number of relatively compact segments for exact evaluation. The latter strategy is the one used here. The continuous segments are first screened to locate all local compactness minima. These local minima are then grouped into structurally compact regions, and Z is evaluated exactly for a limited number of units within each compact region to determine the overall compactness of the region and to bracket the location of the most compact unit within the region. After the noncompact regions are eliminated, Z is evaluated exactly for a larger number of units near each region's compactness minima to precisely locate the most compact unit within the region. When this most compact unit has a Z value equal to or less than that of native globular proteins, it is considered an actual compact unit (domain).

A desirable feature of this procedure is that each compact unit is discovered and defined independently. No attempt is made to partition the protein into contiguous nonoverlapping segments. Thus, the units may overlap, or there may be regions entirely devoid of units. For example, if an extended strand bridges two units but does not contribute to the compactness of either, then the strand is not included in either unit. Conversely, were two units to overlap at a helix that contributes to the compactness of each, then this helical element would be included in both units.

Another desirable feature of the approach is that each unit's compactness is quantified in its Z value, allowing direct comparisons of compactness to be made between different units. This feature is used later in the section on misfolded proteins.

The procedure is now described in greater detail. First, to find local minima in compactness, Z values are *estimated* for all continuous units containing 4, 8, 12, ... residues by using computationally fast approximations of surface area and volume (Zehfus et al., 1985). While the absolute Z values derived from these estimates are only approximate, relative Z values of closely related units can be used to accurately identify local minima of compactness.

Next, the local minima are grouped to define the limits of structurally compact regions. Z is then evaluated *exactly* for units of 4, 8, 12, ... residues to bracket the location of the true minmum within each region and to eliminate regions that are noncompact.

Finally, Z is evaluated exactly for units of *all* sizes within each bracketed interval. The unit with the lowest Z value is then chosen to represent that compact region.

In this procedure a Z value of 1.72 was used as the cutoff between compact and noncompact units. This value is 1 standard deviation larger than the average Z value found for a set of 20 globular proteins. Units with Z values greater than 1.72 are therefore significantly less compact than most native globular proteins.

In exact computations of Z, the accessible surface area of each unit was determined by the method of Lee and Richards (1971), using a step size of 0.25 Å. To determine the volume enclosed within this surface, first the Van der Waals radius of each atom was augmented by 1.4 Å, the radius of a water-sized sphere. Then the volume of the unit was determined by using the method of Pavlov and Fedorov (1983), in which the molecule is embedded in a cubic lattice (0.25 Å edge length), and cubes contained within the augmented atomic radii were summed. Both the areas and volumes determined by these methods are precise to at least 1%, so the overall precision of the Z values presented here is 2% or better.

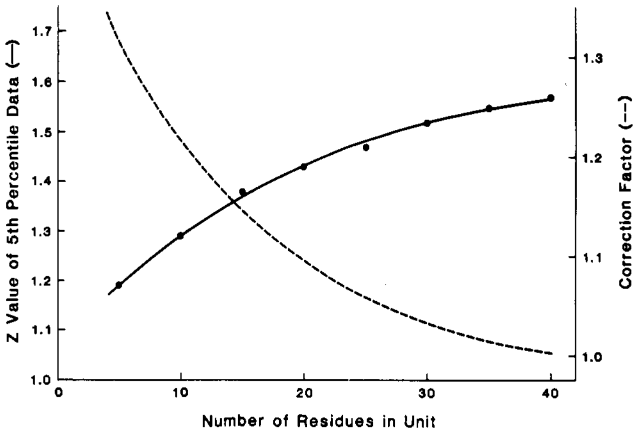

FIGURE 2: Apparent size dependency of Z for units of 40 residues or less together with a compensating correction curve. Points along the solid line correspond to data from pooled distributions of Z values for the proteins staphylococcal nuclease (2SNS), superoxide dismutase (2SOD), cytochrome *c* (3CYT), subtilisin (1SBT), parvalbumin (3CPV), and theromolysin (3TLN) of the Brookhaven protein data base (Bernstein et al., 1977). The dashed line is the compensating correction curve, which when multiplied by the solid line, yields a constant standard value of 1.57 for units of all sizes. The equation for the correction curve is given in eq 2.

Z is a dimensionless ratio and should show no dependency on unit size. However, the configurational freedom of very small segments is restricted in comparison to larger units, so smaller segments are biased toward more compact arrangements, with correspondingly lower Z values. This bias results in an apparent size dependency. When distributions of Z values are examined as a function of size, the upper and lower extremes of the distributions are seen to increase as the segment size increases, up to approximately 40 residues. Above this level the effect becomes negligible, and the lower limit observed for Z values remains relatively constant.

Figure 2 illustrates the relationship between unit size and Z value. Here distributions of Z from several proteins have been pooled, and the datum corresponding to the lowest fifth percentile rank plotted as a function of unit size (solid line). To adjust for the apparent size dependency, a compensating correction curve, shown in Figure 2 as a dashed line, was calculated. The equation for this line is given by eq 2, and

$$\text{correction factor} = 0.488e^{(-0.068 \times \text{number of residues})} + 0.970 \qquad (2)$$

the sum of squared deviations of points from this curve is 0.00016. This correction curve, when multiplied by the size curve, yields a standard value of $Z = 1.57$ for units of all sizes. This standard value corresponds to the fifth percentile datum in the set of all units of 40 residues in length from these six proteins. While the correction curve was constructed by using the fifth percentile rank data, the choice is not critical. The curve would not change significantly if 10th, 20th, or even 30th percentile data had been used instead. All data reported here have been corrected in this manner.

Coordinates for lysozyme (6LYZ) and ribonuclease A (4RSA) were obtained from the the Brookhaven data base (Bernstein et al., 1977). Energy minimized coordinates for correctly and incorrectly folded sea worm hemerythrin and mouse VL-domain immunoglobulin k-chain (Novotný et al., 1984) were provided by Jiři Novotný.

Results for Lysozyme and Ribonuclease

The compact units of lysozyme and ribonuclease are listed in Table I. These units are structurally familiar and visually reasonable. For example, several of the units consist of full

Table I: Compact Units of Lysozyme and Ribonuclease A

	unit[a]	number of residues	Z[b]
lysozyme		129	1.62
	1–124	124	1.58
	1–116	116	1.58
	8–111	104	1.57
	28–111	84	1.59
	39–98	60	1.48
	48–95!	48	1.47
	31–76	46	1.55
	42–84!	43	1.43
	30–60	31	1.53
	58–83	26	1.49
	8–29	22	1.52
	42–60	19	1.52
	78–94	17	1.53
	16–32	17	1.58
	98–113	16	1.51
	104–113	10	1.52
	29–38	10	1.54
	24–32	9	1.58
	3–11	9	1.56
	89–96	8	1.54
ribonuclease A		124	1.68
	4–124	121	1.64
	7–110	104	1.70
	24–124	101	1.70
	40–124	85	1.69
	40–110?	71	1.74
	10–48	39	1.53
	47–80	34	1.62
	5–37	33	1.62
	10–41	32	1.63
	24–54	31	1.67
	80–103	24	1.63
	56–76	21	1.54
	29–46	18	1.54
	20–37	18	1.58
	107–122	16	1.51
	62–73	12	1.54
	86–97	12	1.57
	2–12	11	1.53
	20–29	10	1.58
	29–37	9	1.54

[a] Units denoted with ? have marginal coefficients of compactness while those denoted with ! have exceptionally low Z values. [b] The Z values of units with less than 40 residues have been normalized to a 40-residue value (see text and Figure 2).

or partial helices and their adjacent turns. In lysozyme these include units 89–96, 3–11, 24–32, 29–38, and 98–113, and in ribonuclease, units 20–29, 20–37, 29–37, and 2–12. This last unit is independently stable in solution (Brown & Klee, 1971; Bierzywski et al., 1982). Loop structures, which can be both simple and compound, are also in evidence, including 29–46, 62–73, and 86–97 of ribonuclease and 58–83 of lysozyme. On a slightly larger scale are the helix–loop–helix and β–loop–β structures: 8–29 of lysozyme and 107–122 of ribonuclease, respectively. Larger yet are a 3-stranded β sheet in lysozyme (42–60) and a distorted β–α–β unit in ribonuclease (10–48).

A diversity of methods used to identify domains can be found in the literature, including visual inspection (Wetlaufer, 1973; Phillips, 1966; Drenth et al., 1968), distance maps (Liljas & Rossman, 1974; Gō, 1983), clustering (Crippen, 1978), cutting planes (Rose, 1979), minimization of interface area (Wodak & Janin, 1981), minimization of specific volume (Lesk & Rose, 1981), maximization of solvent exclusion (Rashin, 1981), and isolation of coherent regions in a normal mode analysis (Levitt et al., 1985). A compilation of results from these methods is shown in Table II.

Meaningful comparison of our results (Table I) with those of other methods (Table II) is necessarily limited to continuous-chain segments. Moreover, in our method a given stretch of chain can belong to more than one unit or can be excluded from any unit. These features complicate the comparison with other methods.

Table II: Domain Assignments for Ribonuclease and Lysozyme

ref	Lysozyme domains[a]
Wetlaufer (1973)	(1–38, 88–100), 39–87 or (1–38, 39–87), 88–100
Liljas and Rossman (1974)	40–85
Crippen (1978)	(1–36, 111–129), 37–110
Rose (1979)	1–40, 41–129
Rashin (1981)	8–29, 39–98, 105–116
Janin and Wodak (1983)	(1–38, 88–129), 39–87
Levitt et al. (1985)	(1–39, 87–129), 40–86
	subdomains
Crippen (1978)	1–17, 18–21, 22–36, 37–48, 49–55, 56–66, 67–72, 73–86, 87–101, 102–111, 112–129
Rose (1979)	1–40, 41–85, 41–60, 61–85, 86–129
Gō (1983)	1–30, 31–55, 56–84, 85–108, 109–129
	Ribonuclease
ref	domains[a]
Wetlaufer (1973)[b]	(11–49, 80–105), (1–10, 52–76, 105–124)
Crippen (1978)[b]	1–50
Rose (1979)	1–49, 50–124
Rashin (1981)[b]	23–124
Janin and Wodak (1983)[b]	1–56, 57–110
Levitt et al. (1985)	(20–46, 80–102), (47–79, 103–124)
	subdomains
Crippen (1978)[b]	1–27, 28–36, 37–50, 51–67, 68–92, 93–114, 115–124
Rose (1979)	1–49, 81–101, 81–124, 50–80, 102–124

[a] Units in parentheses are discontinuous domains. [b] Analysis performed on ribonuclease S.

Subject to these qualifications, the compact units listed in Table I can be compared with units ("domains", "modules", "intermediates",...) found by other methods. In general, large, continuous-chain units are similar in most methods. Typically, lysozyme is divided into three major regions: (i) 1··8–29··38, (ii) 37··40–85··98, and (iii) 88··111–100··129. The corresponding compact units from Table I are (i) 8–29, (ii) 39–98, and (iii) 98–113. Similarly, ribonuclease is generally divided into two regions: (i) 1··11–49··56 and (ii) 52··57–76··100, and the corresponding compact units from Table I are (i) 10–48 and (ii) 47–80.

Larger differences are apparent for the smaller units. In Table I, lysozyme has a number of smaller units similar to those of Rose (1979), a few similar to those of Gō (1983), but almost none similar to those of Crippen (1978). For ribonuclease, there is again a good correlation with work of Rose (1979) and a poor correlation with work of Crippen (1978).

Misfolded Proteins

To assess sensitivity, the coefficient of compactness was used to analyze misfolded proteins. Novotný et al. (1984) generated misfolded molecules from two X-ray-elucidated proteins with the same number of residues but entirely different chain folds. These proteins were (1) hemerythrin, an all-α protein, and (2) the VL domain of mouse immunoglobulin, an all-β protein. To design misfolded chains, Novotný et al. (1984) interchanged the backbone dihedral angles of these two proteins and then energy minimized the resultant structures. Both misfolded chains converged to a reasonable internal energy while retaining the incorrectly assigned backbone structure (op. cit.).

Hemerythrin, the VL domain of immunoglobulin, and their misfolded analogues were analyzed as described in the preceding sections. Results are listed in Table III. In general,

Table III: Compact Units of Authentic and Misfolded Proteins

			VL Domain like Conformations					
VL domain			misfolded hemerythrin			misfolded–adjusted hemerythrin		
unit[a]	no. of residues	Z	unit	no. of residues	Z	unit	no. of residues	Z
1–113	113	1.55	1–113	113	1.79	1–113	113	1.68
2–99	98	1.62	2–98	97	1.90	2–99	98	1.78
19–110	92	1.61	19–112	94	1.86	18–110	93	1.75
19–98	80	1.58	18–98	81	1.84	18–98	81	1.72
28–98	71	1.57	23–98	76	1.85	28–98	71	1.75
28–90	63	1.61	28–82	55	1.83	28–82	55	1.72
19–81	63	1.63	19–82	64	1.85	19–82	64	1.73
39–94	56	1.51	37–95	59	1.81	37–95	59	1.71
39–81	43	1.54	37–81	45	1.76	37–82	46	1.66
29–64	36	1.62	29–61	33	1.78	29–61	33	1.69
52–83	32	1.51	52–82	31	1.68	52–83	32	1.60
39–70	32	1.54	37–73	37	1.72	37–72	36	1.61
89–110	23	1.66	91–108	18	1.68	89–110	22	1.66
39–59	21	1.58	38–61	24	1.71	38–61	24	1.63
65–85	21	1.61	67–85	19	1.73	64–83	20	1.67
52–71!	20	1.47	52–72	21	1.63	52–72	21	1.55
8–20	13	1.68	5–24	20	1.81	5–24	20	1.74
42–53	12	1.61	41–54	14	1.72	41–53	13	1.59
28–38	11	1.52	31–36	6	1.61	31–36	6	1.55
95–104	10	1.56	96–102	7	1.67	96–102	7	1.66
72–77	6	1.55	71–78	8	1.70	71–78	8	1.62
			Hemerythrin-like Conformations					
hemerythrin			misfolded VL domain			misfolded–adjusted VL domain		
unit	no. of residues	Z	unit	no. of residues	Z	unit	no. of residues	Z
1–113	113	1.55	1–113	113	1.81	1–113	113	1.79
2–113	112	1.54	2–113	112	1.79	2–113	112	1.77
8–113	106	1.53	18–113	96	1.74	20–109	90	1.70
20–102	83	1.59	21–102	82	1.73	21–98	78	1.68
19–87	69	1.64	24–87	64	1.72	24–91	68	1.66
51–113	63	1.63	47–109	63	1.75	47–109	63	1.79
8–54	47	1.52	8–58	51	1.69	10–58	49	1.60
47–86	40	1.55	47–85	39	1.56	47–85	39	1.55
20–55	36	1.48	24–57	34	1.54	24–55	32	1.46
51–81	31	1.53	55–82	28	1.57	55–81	27	1.55
75–102	28	1.56	72–101	30	1.63	72–109	38	1.71
54–74	21	1.59	56–67	12	1.56	56–67	12	1.51
10–29	20	1.49	9–31	23	1.63	10–29	20	1.58
30–47!	18	1.43	31–52	22	1.53	31–51	21	1.43
80–94	15	1.51	80–98	19	1.63	80–97	18	1.67
40–52	13	1.59	40–50	11	1.62	40–50	11	1.55
97–107	11	1.53	99–107	9	1.55	99–108	10	1.57
51–60	10	1.55	47–58	12	1.55	45–58	14	1.55
19–26	8	1.52	19–28	10	1.58	20–26	7	1.55
80–86	7	1.52	80–85	6	1.55	78–86	9	1.57
73–79	7	1.56	72–79	8	1.50	72–79	8	1.49

[a] Exceptionally compact native units are indicated with !.

unit boundaries are similar between authentic molecules and their misfolded counterparts, reflecting their almost identical backbone structures. Slightly more variation in unit boundaries is seen in the α-helical hemerythrin-like structures, since the inclusion or deletion of one turn of helix will change a unit's definition by plus or minus four residues.

In general, compact units from misfolded proteins (Table III, columns 4–6) are significantly less compact than those from authentic molecules. In fact nearly half of the misfolded structures have Z values of 1.72 or greater and therefore are not considered compact by the criterion used here. The inherent noncompactness of the misfolded structures cannot be an artifact of energy minimization, since both native and misfolded structures were subjected to the same minimization protocol.

Recognizing that their structures had excessive surface area, Novotný et al. (1984) attempted to alleviate this problem by individually adjusting solvent-exposed side chains into less exposed conformations. It can be seen (Table III, columns 7–9) that side-chain adjustment yields some improvement. One-fourth of the adjusted units have compactness comparable to that of their native counterparts, although another one-fourth of the units remain noncompact.

Three units in the misfolded–adjusted proteins are significantly more compact than their native counterparts: residues 72–79, 40–50, and 56–67 in hemerythrin-like immunoglobulin. Examination of these units reveals that the misfolded sequences contain residues with significantly shorter side chains than corresponding residues in the native sequence. Enhanced compactness, in these cases, is probably due to the presence of these abridged side chains.

Hierarchic Organization

Hierarchic organizational plots of the four native proteins used in this paper are shown in Figure 3. Each plot consists of a set of triangles, one for every compact unit. The largest triangle in each plot represents the entire protein, while the smaller subsumed triangles represent the constituent compact units. The height of a triangle is proportional to the size of the unit, and the base defines the linear extent of the unit.

The structural hierarchy depicted by these plots suggests possible folding pathways. Lysozyme, Figure 3A, serves as

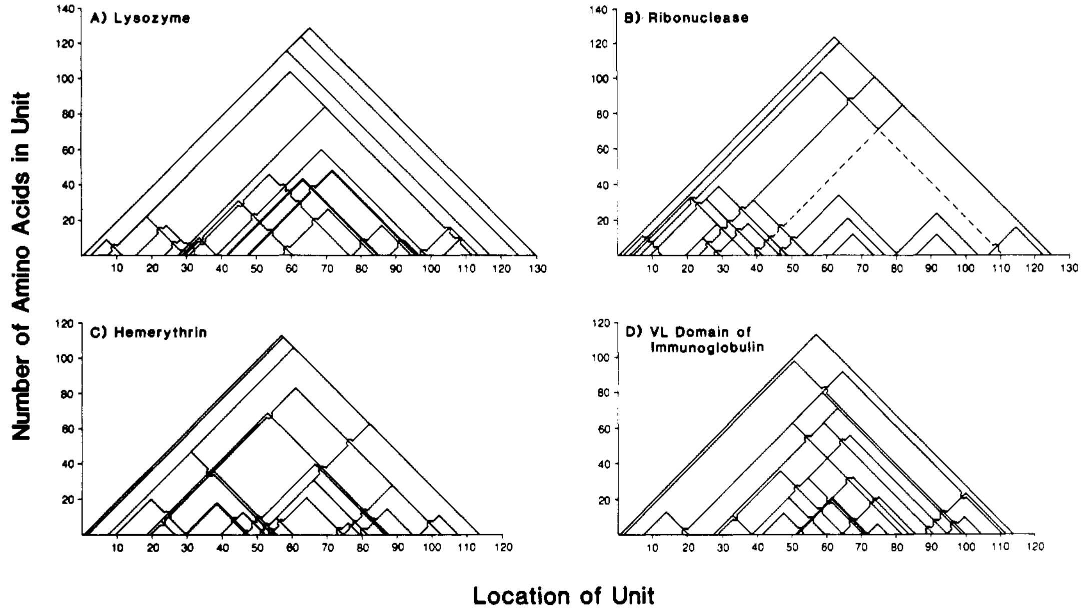

FIGURE 3: Hierarchic organizational plots for (A) lysozyme, (B) ribonuclease, (C) hemerythrin, and (D) the VL domain of immunoglobulin. Each compact unit is represented as a triangle. The apex of the triangle marks the center of the unit on the abscissa and the size of the unit on the ordinate. The endpoints of the base of the triangle mark the N- and C-termini of the units on the abscissa. The outermost triangle in each plot represents the entire protein; interior triangles are the constituent compact units. Exceptionally compact units ($Z \leq 1.47$) are represented by heavier lines, while marginal units ($1.62 \leq Z \leq 1.75$) are represented by dashed lines.

a first example. This protein contains six compact primitives, i.e., small units that cannot be further subdivided: 3–11, 16–32 or 29–38, 42–60, 58–83, 89–96, and 104–113. Such regions are possible sites of nucleation in protein folding. In this protein the primitives consist of helices, loops, and a β–turn–β structure. The largest and most compact of these primitives is the complex loop occurring between residues 58 and 83, which provides a core structure for the rest of the protein.

This core can be enlarged by growth along either of two alternative paths. In one case, the primitive 42–60, a β–turn–β structure, is added to the N-terminal end of the unit to form the highly compact 42–84 unit. In the other case, a single β-strand is added to the N-terminal end, while a helix is added to the other end to form the 48–95 unit. The next larger unit, containing residues 39–98, may be formed from either of these two intermediates. Compact units 29–38 and 98–113 are then added to both ends of this growing structure to form the 28–111 unit, which then incorporates the 8–29 unit to become the 8–111 unit. Finally, the remaining residues are added to complete the protein.

It should be emphasized that this folding pathway is not unique. At least one branch point occurs where different segments can coalesce with a given precursor unit to yield differing higher order structures. This feature is seen in virtually all proteins. Although differences in Z values may indicate that one pathway is favored over another, it is likely that these choices represent genuine alternative folding pathways (Harrison & Durbin, 1985).

Ribonuclease, Figure 3B, is interesting because it contains several small units, no intermediate sized units, and a few relatively noncompact large units. This protein is composed of four distinct modules: 10–48, 47–80, 80–103, and 107–122. Three of these modules exhibit simple folding pathways. The 107–122 module is a primitive β–turn–β structure. The 80–103 module is nucleated at a β–loop–β primitive (86–97), which then extends its two adjacent β-strands to form the module. The 47–80 module is formed in three steps: nucleation in a β–loop–β primitive (62–73); extension of the β-strands to form the 56–76 unit; then addition of a helical segment to the N-terminal end. The 10–48 module at first appears complicated. However, if only the more compact units are used, the folding pathway resolves in a simple manner. Here, the nucleation is in segment 29–37, which then adds a β-strand to its C-terminal end to become the 29–46 unit, and this, in turn, adds a β–loop–α to its N-terminal to become the completed module. Once these individual modules are formed there appears to be no satisfactory way to assemble them into compact higher order structures. This "dead end" occurs because the plot displays only continuous protein domains, and the actual domains of ribonuclease are discontinuous. Examination of the protein indicates that the first module folds against the third, and the second against the fourth. While present procedure precludes the automatic analysis of discontinuous domains, it is nevertheless possible to calculate Z values for discontinuous units. The Z value of the 10–48 + 80–103 unit is 1.50, while that of the 47–80 + 107–122 unit is 1.51. These discontinuous units are quite compact, more so than any continuous higher order structure found here. Thus, they appear to represent examples of discontinuous compact domains.

The hierarchic organization of hemerythrin is shown in Figure 3C. The primitive units in this protein are exclusively helices and helix–turn–helix structures. The dominant progression leads from the 30–47 primitive to the 8–54 unit along two different pathways. In one, the C-terminal helix of the 30–47 primitive is extended while the 19–26 primitive coalesces with the other end, forming the 19–55 unit. In the other, the 19–26 primitive grows to become the 10–29 unit, which then combines with the 30–47 primitive. Independently, the 51–113 unit is formed from three primitives, 54–75, 80–94, and 97–107. These two major domains then merge, resulting in the final protein. Other pathways, involving either the 20–102 or the 18–87 units, can also be traversed.

The VL-domain of immunoglobulin (Figure 3D), serves as the final example. This protein contains six compact units, 8–20, 28–38, 42–53, 52–71, 72–77, and 95–104. In this case, the primitives are loops and β–turn–β structures, the largest and most compact of which is a double loop containing residues 52–71. This loop appears to serve as a core for the two β-sheets of this protein. The next events in the hierarchy are the addition of the two primitives, 42–53 or 72–77, to either the N- or C-terminal end of the core, resulting in the 39–70 or 52–83 units. The net effect of these additions is to add β-strands to the growing β-sheet structure. When either of these pieces has been added to the core, the other may attach, yielding the 52–83 unit. This latter unit then adds a β-strand in its C-terminal region to become the highly compact 39–94 unit, which can, in turn, recruit strands at either end in a succession of steps, leading to increasingly larger structures, and ultimately to the entire protein.

These folding pathways are, of course, speculative and are intended only as guides for the design of experiments. Such experiments might include (1) monitoring the protein under partially denaturing conditions to see whether hypothesized compact cores can be detected (Creighton, 1978), (2) isolating some of the larger compact units (Wetlaufer, 1981), and (3) modifying the interfaces between compact units to perturb folding by changing unit–unit interactions (Ackers & Smith, 1985).

In summary, the coefficient of compactness is a highly sensitive parameter that can be used to dissect proteins into physically reasonable units. Compact units found by this method comprise a structural hierarchy that suggests folding pathways. Work is currently in progress to analyze a large data base of proteins with these methods.

ACKNOWLEDGMENTS

We thank Jiři Novotný for useful discussion and protein coordinates, Mark Handschumacher and Fred Richards for the program used to calculate accessible surface area, Jacque Leszczynski for small unit data, and an anonymous reviewer for insightful suggestions.

Registry No. Ribonuclease, 9001-99-4; lysozyme, 9001-63-2.

REFERENCES

Ackers, G. K., & Smith, F. R. (1985) *Annu. Rev. Biochem. 54*, 597–629.

Bernstein, F. C., Koetzle, T. G., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rogers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977) *J. Mol. Biol. 122*, 535–542.

Bierzywski, A., Kim, R. S., & Baldwin, R. L. (1982) *Proc. Natl. Acad. Sci. U.S.A. 79*, 2470–2474.

Brown, J. E., & Klee, W. K. (1971) *Biochemistry 10*, 470–476.

Creighton, T. E. (1978) *Prog. Biophys. Mol. Biol. 33*, 231–297.

Crippen, G. M. (1978) *J. Mol. Biol. 126*, 315–332.

Drenth, J., Jansonius, N., Koekoek, R., Swen, H., & Wolthers, B. (1968) *Nature (London) 218*, 929–932.

Gō, M. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 1964–1968.

Harrison, S. C., & Durbin, R. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 4028–4030.

Kauzmann, W. (1959) *Adv. Protein Chem. 14*, 1–64.

Lee, B. K., & Richards, F. M. (1971) *J. Mol. Biol. 55*, 379–400.

Lesk, A. M. & Rose, G. D. (1981) *Proc. Natl. Acad. Sci. U.S.A. 78*, 4304–4308.

Levitt, M., Sander, C., & Stern, P. S. (1985) *J. Mol. Biol. 181*, 423–447.

Liljas, A., & Rossman, M. G. (1974) *Annu. Rev. Biochem. 43*, 475–507.

Pavlov, M. Y., & Fedorov, B. A. (1983) *Biopolymers 22*, 1507–1522.

Phillips, D. C. (1966) *Sci. Am. 215*, 78–90.

Novotný, J., Bruccoleri, R., & Karplus, M. (1984) *J. Mol. Biol. 177*, 787–818.

Nozaki, Y., & Tanford, C. (1971) *J. Biol. Chem. 246*, 2211–2217.

Rashin, A. A. (1981) *Nature (London) 291*, 85–86.

Richards, F. M. (1977) *Ann. Rev. Biophys. Bioeng. 6*, 151–176.

Rose, G. D. (1979) *J. Mol. Biol. 134*, 447–470.

Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H., & Zehfus, M. H. (1985) *Science (Washington, D.C.) 229*, 834–838.

Tanford, C. (1980) *The Hydrophobic Effect*, 2nd ed., Wiley, New York.

Wetlaufer, D. B. (1973) *Proc. Natl. Acad. Sci. U.S.A. 70*, 697–701.

Wetlaufer, D. B. (1981) *Adv. Protein Chem. 34*, 61–92.

Wodak, S. J., & Janin, J. (1981) *Biochemistry 20*, 6544–6552.

Zehfus, M. H., Seltzer, J. P., & Rose, G. D. (1985) *Biopolymers 24*, 2511–2519.