Scaling regimes of molecular size and self-entanglements in very compact proteins

Gustavo A. Arteca
Département de Chimie et Biochimie, Laurentian University, Ramsey Lake Road, Sudbury, Ontario, Canada P3E 2C6
(Received 18 May 1994; revised manuscript received 4 November 1994)

Establishing interrelations between size, compactness, and three-dimensional shape in biomolecules is important for a better understanding of the factors governing their folding features and their biological function. Whereas size and compactness can be characterized by parameters such as the radius of gyration, the description of folding features is less well defined. Recently, we have introduced the probability of overcrossings in two-dimensional projections of a rigid backbone as a descriptor of self-entanglements. This function provides a simple and intuitive characterization: the more complex the entanglements, the larger the mean number of overcrossings. In this work, we study relationships between size and entanglements on a special subclass of biomolecules with a global structural constraint: the family of native protein conformations which are the most compact within a range of amino acid residue numbers. Initially, we use a set of 373 experimental protein backbones exhibiting very diverse lengths, composition, and structural features. Within this set, we have located the proteins with the smallest radii of gyration for fixed ranges of monomer numbers. For this class of proteins, we observe power-law scaling behavior in size and entanglement complexity in terms of the residue number. The results suggest that there are two distinct regimes of scaling characterizing short compact proteins and long compact proteins, respectively. The change of regime appears to be localized roughly around 300 amino acid residues. We propose that this difference correlates with a change in the content of secondary structure for compact proteins: the content of β strands for short chains is almost twice as large as that of longer chains, the latter in turn being much richer in α helices. In summary, the work establishes that in forming a very compact polypeptide there are some constraints among the number of residues, the radius of gyration, the entanglement complexity, and the content of secondary structure of the molecular chain. Thus the degree of compactness in heteropolymers appears to exhibit more complex features than those found in homopolymers.

PACS number(s): 87.15.He, 82.20.Wt, 05.90.+m

I. INTRODUCTION

The characterization of a polymer's *shape,* as opposed to its size and *compactness,* is a task that cannot fully be solved with a single technique or addressed by a single descriptor. Whereas global geometric descriptors, such as the radius of gyration R_G, provide well-known measures of size (and, to some degree, compactness),[1] there is no unique feature that can be associated with "shape" as a property of a polymer chain. This lack of precise characterization hampers the analysis of rigid or flexible protein backbones. A common strategy is to analyze the occurrence of highly structured sections of a backbone (e.g., α helices and β strands). However, this may not be very informative in monitoring unfolding or melting, when the backbone is in a rather unstructured state, midway between the native structure and a random coil [1–3]. Visually inspecting the polymers in such a state often fails to reveal any valuable information.

The primary and tertiary structure of a protein contain the two key elements needed for a detailed description of the α-carbon backbone: the *connectivity* of the main chain atoms and their spatial positions (the *geometry*), respectively. (We use the term "connectivity" in the sense of a matrix establishing which α-carbon atoms are "linked" because of being associated to sequential residues in the primary structure.) Any useful description of the degree of folding in a backbone has to take into account these two pieces of information. Although there is no natural measure for the degree of folding, a recent approach provides a satisfactory representation of the complexity and type of entanglements found in a chain [4–7]. The methodology uses the probability of finding a rigid three-dimensional (3D) backbone in a two-dimensional (2D) projection with N overcrossings or double points. For an n-residue backbone, these probabilities are indicated as $\{A_N(n)\}$, and they constitute a global descriptor of self-entanglements in a chain [6]. (We use the concept of "self-entanglement" as a broad notion conveying the occurrence of turns, loops, and folds in a *single* polymer molecule.) For a given n value, a chain may exhibit radically different types of folds. A chain adopting a β-strand form will exhibit essentially no self-entanglements. In contrast, the chain will appear mildly entangled if folded as an α helix, and even further entangled in some random coil configurations. Even though some of the latter structures may exhibit comparable geometries, their shapes can readily be distinguished by using the probability distribution $\{A_N(n)\}$. A key observation is as follows: the more entangled structures will have a higher mean number of overcrossings [6]. This mean number $\bar{N}$ is given as

$$\bar{N}=\sum_{N=0}^{\max N} N A_N(n) , \tag{1}$$

1063-651X/95/51(3)/2600(11)/$06.00

where max N is the maximum number of overcrossings compatible with a given architecture. For linear (nonbranched) chains, max $N=(n-2)(n-3)/2$, $n\geq 3$. Depending on the actual folding, the distribution $\{A_N(n)\}$ will vary and so will the $\bar{N}$ value. A related shape descriptor is the maximum probability of overcrossings, A^*. These two parameters, derived from the overcrossing probabilities $\{A_N(n)\}$, have been used to monitor structural stability of short chains along molecular dynamics trajectories [7] and the configurational state of random polymers and some proteins [8]. Recently, we have shown that configurational averages of these shape descriptors also follow approximate power-law scaling with the number of residues, n [8]. This type of behavior is well known for the configurationally averaged radius of gyration [9] $\langle R_G \rangle$,

$$\langle R_G \rangle \sim k n^{\nu} , \tag{2}$$

where the critical exponent ν takes the value $\frac{1}{2}$ for ideal chains in θ conditions (poor solvent) [10], approximately 0.588 in good solvents (the fully developed excluded-volume interaction above θ conditions) [11], and $\frac{1}{3}$ in the case of "collapsed" polymers below the θ conditions [12,13]. An additional scaling regime has recently been proposed [14]. From now on, we shall refer to ν as the "size exponent."

For the shape descriptors of entanglement complexity in model polymers, we have found that similar relations hold [8]:

$$\langle A^* \rangle \sim a n^{b} , \tag{3}$$

$$\langle \bar{N} \rangle \sim \alpha n^{\beta} . \tag{4}$$

For self-avoiding walks with variable excluded-volume interaction and $50\leq n\leq 500$, we have estimated that $b\approx -1.00\pm 0.03$ and $\beta\approx 1.4\pm 0.1$ [8]. The exponents appeared to change little with the excluded volume. That is, they do not depend strongly on whether the dominant configurations are ideal, compact, or swollen [8]. The dependence on the configurational state is mostly conveyed by the preexponential factors a and α in Eqs. (3) and (4). These values for b and β must be considered as "effective" critical exponents for medium-size polymers, and may not represent the correct behavior for $n\rightarrow\infty$. Orlandini *et al.* [15] have recently studied the behavior of the mean number of overcrossings in self-avoiding walks with very large n values and estimated $\beta\approx 1.13$ as a lower bound to the asymptotic limit in ideal and good solvents (above the θ point). From now on, we shall refer to β as the "entanglement exponent."

The number of monomers needed to reach the truly asymptotic limit of $\bar{N}$ appears to be beyond the typical number of amino acid residues found in proteins. As we showed in Ref. [8], the effective exponent $\beta\approx 1.4$ describes accurately not only the configurationally averaged shapes of model polymers but also the average shape of most proteins. In Ref. [8], a set of 197 proteins was analyzed. The set was diverse and contained no structural bias. The proteins included were selected among all available structures in the Brookhaven Protein Data Bank (PDB) [16]. We found that, in a set with all possible configurational features (i.e., compact as well as swollen proteins) and all possible numbers of amino acid residues, Eq. (2) was not satisfied. In contrast, Eqs. (3) and (4) for the shape descriptors held more accurately. This result indicated an apparent universality in the scaling of shape descriptors for proteins with very diverse structures.

Our previous work was not concerned with any special structural feature, either local or global [8]. However, for many applications of relevance to biochemistry and biophysics, it is important to establish relations between size and shape restricted to macromolecules with well-defined structural profiles. In this work, we carry out a detailed analysis of the scaling behavior of molecular size and molecular shape descriptors for a class of proteins with a specified structural (global) constraint: those whose actual native conformation is the most compact among all other proteins of comparable length.

A number of measures of compactness have been proposed in the literature, based on the number of chain contacts in a lattice model [17,18] and on the deviation from a reference structure [19]. Here, we are interested in a simple procedure that characterizes the compactness in proteins at their actual experimental (native) structures, such as the ones available from the Brookhaven Protein Data Bank. The radius of gyration defined by the main chain (α-carbon) atoms is a possible alternative, although it neglects the distinct contribution to the excluded volume of the variable monomers in heteropolymers. Chan and Dill have defined a "minimum radius of gyration" which takes into account the protein composition, by considering an idealized close packing of the sequence of amino acids within a sphere [20]. In our case, we resort to a simpler procedure which provides a set of very compact proteins with little composition bias. This set of proteins is determined by using a large and diverse set of experimental structures and then selecting those with the smallest R_G value within a given range in the number of amino acid residues. In this manner, by selecting various ranges, the extracted radii of gyration become somewhat "averaged" over variable compositions. For completeness, we include in this work an analysis of compactness for this set of proteins by employing also the criterion in Ref. [20].

For the above group of compact proteins, we analyze the interrelation between size and molecular shape. From these results, we derive conclusions on the conditions prevalent to form very compact proteins with a given chain length. Size and compactness are described with the radius of gyration. Shape features are conveyed by the complexity of self-entanglements, measured in terms of $\bar{N}$ and A^*.

The article is organized as follows. Section II discusses the working set of proteins, the technical details on the evaluation of molecular size and shape descriptors, and their precision. We discuss also the criteria followed for the location of the subset of most compact proteins and compare it with other procedures in the literature. Section III presents the scaling behavior of various descriptors for compact proteins. A brief analysis of the estima-

tion of asymptotic exponents in the limit of very large number of amino acid residues is also given. Section IV discusses the change in the content of secondary structure for compact proteins of variable length. We propose a correlation between the secondary structural content and the scaling behavior of shape and size descriptors. Section V closes with further comments and a summary of conclusions.

II. WORKING SET OF PROTEINS AND ACCURACY OF SHAPE DESCRIPTORS

Compactness and entanglement complexity are analyzed on a set of 373 proteins (from now on referred to as the "working set of proteins"). The set includes all *distinct* proteins deposited in the Brookhaven PDB by 1992 (i.e., discounting structural updates and equivalent proteins obtained from different biological sources). Moreover, we have added to the working set several mutant proteins and variations on the same protein whenever some significant difference could have been expected from the biological source. Finally, the set also includes several proteins deposited as "prerelease" in the PDB. Briefly, this large working set [21] represents accurately the structural variety of all proteins known to date.

In all cases, we have considered the actual number of residues in the experimental 3D structures. This number may differ sometimes from the number of amino acids in the primary sequence. In the cases where a protein forms a quaternary structure, only one protein monomer was considered at a time. Figure 1 gives a histogram of the proteins in the working set as a function of the number of amino acid residues, n. Note that the most frequently found backbone lengths correspond to values of n between 100 and 350. Proteins with more than 600 residues are uncommon.

For all structures in the above set, we have computed size and entanglement complexity descriptors. Following the common approach, the analysis is confined to the α-carbon backbone (i.e., only one main chain atom is retained per residue). Figure 2 displays the radius of gyration R_G as a function of n. The lack of a single scaling behavior is evident. Figure 2 reveals that the set includes proteins in open as well as compact structures [22]. Note that the dispersion is not symmetrical: proteins with smaller radius of gyration appear to follow a better defined scaling [8,23]. We focus our study on this latter subset of proteins. These correspond to the "most compact proteins," since they exhibit the smallest radius of gyration compatible with a given number of amino acid residues.

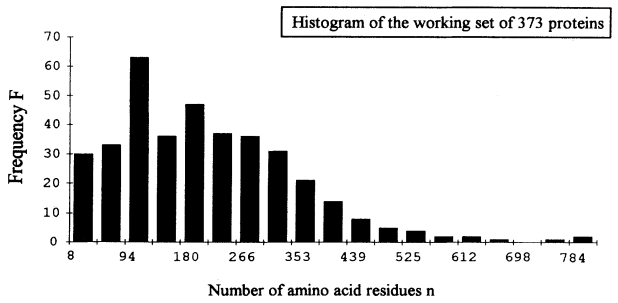

FIG. 1. Histogram of the working set of 373 proteins classified in terms of the number of amino acid residues. (The smallest protein in the set has 8 residues and the largest 823. Frequencies are given for regular intervals of length 43. Note the lack of any proteins between $n = 698$ and 741. We did not find any proteins with these numbers of residues in the data bank. As the figure indicates, most proteins are found in the range $100 < n < 350$.)

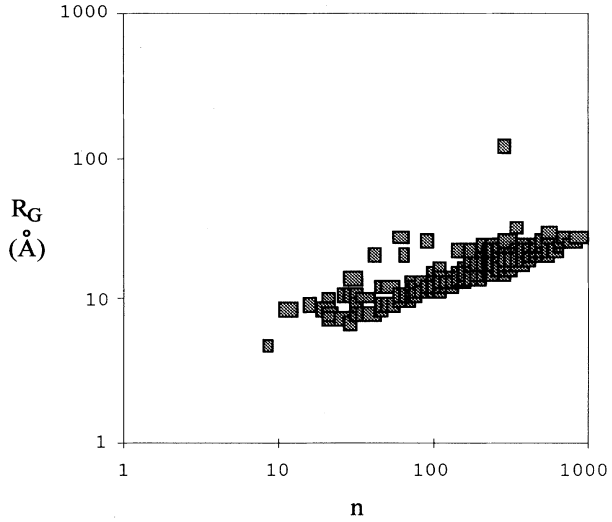

FIG. 2. Log_{10}-log_{10} plot of the radius of gyration as a function of the number of amino acid residues for the proteins in the entire working set.

For the working set of proteins, we have computed the shape descriptors $\bar{N}$ and A^*. The strategy and algorithms for the calculation of $\{A_N(n)\}$ for protein backbones have been discussed in detail elsewhere and will not be repeated here [4]. For completeness, we make only a few remarks on the numerical evaluation of descriptors $\bar{N}$ and A^*. The overcrossing probabilities in the chain are computed in practice from a large number of random projections of the protein backbone to planes tangent to the smallest sphere (centered at the center of mass) enclosing the backbone completely. The overcrossing descriptors are evaluated from an average of six overcrossing spectra $\{A_N(n)\}$ for each protein. The six spectra correspond to computations involving 4000, 6000, 8000, 10 000, 15 000, and 20 000 random projections, respectively. The accuracy for a given number of projections decreases with the length of the chain. Figure 3 illustrates the accuracy achieved in the most disadvantageous cases, namely, those of aconitase (top) and glycogen phosphorylase (bottom), which are the largest proteins in our working set. Despite the large error, the results are accurate enough to estimate A^* and $\bar{N}$ reliably. Note that the different structural features in these proteins are reflected by different overcrossing spectra. In the case of aconitase, we observe the presence of a "shoulder" about $N \sim 580$.

Figure 4 gives the shape descriptor $\bar{N}$ for all proteins in

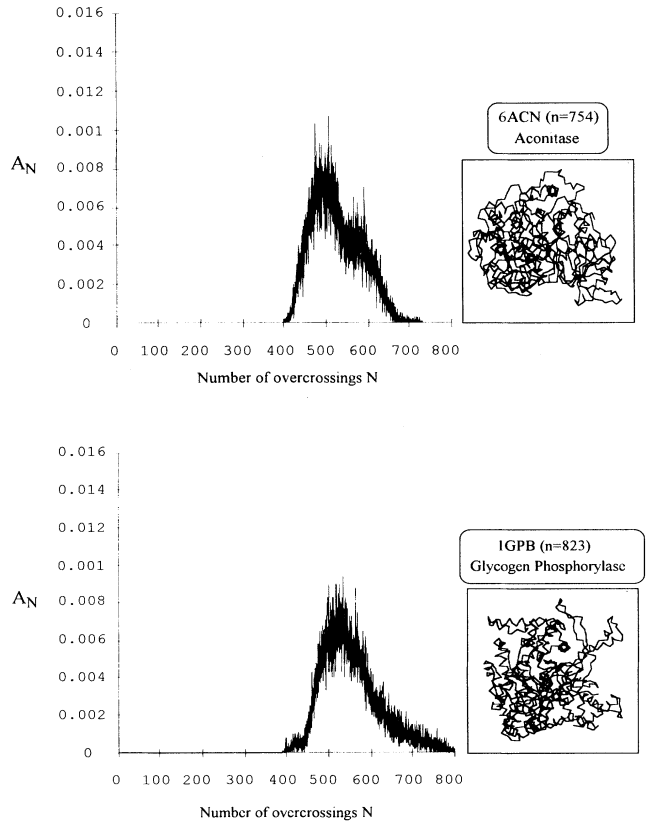

FIG. 3. Overcrossing spectra of the two largest proteins in the working set. [Since there are very few long proteins, aconitase ($n=754$) and glycogen phosphorylase ($n=823$) end up also included in the set of most compact proteins known to date. The figure on the right illustrates one possible placement of the protein backbone. These two proteins illustrate the accuracy limit with which we can evaluate numerically the overcrossing probabilities. However, despite the large relative uncertainties, the differences in folding features between these two structures can readily be recognized in the spectra.]

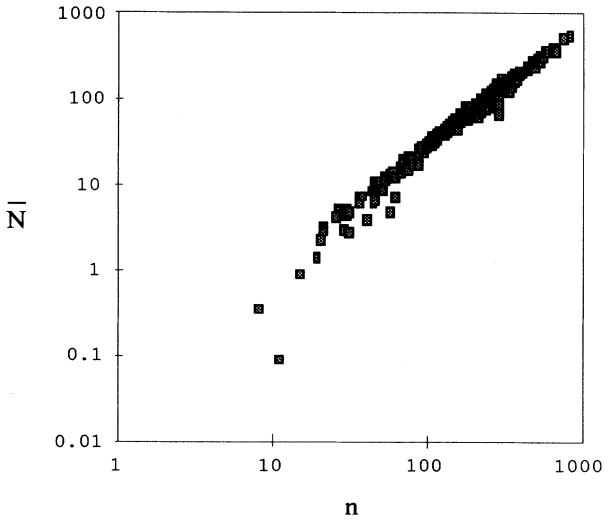

FIG. 4. Log_{10}-log_{10} plot of the mean number of overcrossings, $\bar{N}$, as a function of the number of amino acid residues for the proteins in the entire set.

the working set. The scaling behavior for $n > 100$ is evidently better defined than that for R_G in Fig. 2. Due to the seemingly small dependence of $\bar{N}$ on protein configurational state, it is difficult to extract information from Fig. 4 on any differential scaling associated with specific structural features. For this reason we resort to the size descriptor R_G to select a subset of compact proteins.

The definition of the subset of most compact proteins depends on the choice of an "amino acid residue window" Δn. Because sets of proteins with exactly the same number of residues n are very limited, one is forced to select structures with the smallest R_G within a "window" of n values of length Δn. We have explored several possible choices of Δn values. If Δn is too small, proteins that are not very compact are included in the set. In contrast, if Δn is too large, the final subset is very small. In this work, we discuss the results obtained with two reasonable choices, $\Delta n = 50$ and 10. Short polypeptides have been excluded since they trivially lead to small R_G values. For this reason, the selections have started from $n = 20$. In the case of $\Delta n = 50$, this corresponds to finding the smallest protein in each of the ranges [20,49], [50,99], [100,149], etc., up to [800,849]. Similarly, in the case $\Delta n = 10$, the ranges considered are [20,29], [30,39], etc., up to the last one, [820,829].

A $\Delta n = 50$ window is an optimum choice since it renders a fairly regular and complete set of compact proteins, with only the range $650 < n < 749$ failing to produce a sample. The set of 15 proteins is listed in Table I together with their size and shape descriptors, as well as their biochemical function. Most proteins in the set are involved with electron transport, oxygen transport, or redox and hydrolysis reactions [20]. The set certainly contains the most compact proteins known for $n < 600$. For longer chains, this statement must be qualified by the fact that very few proteins are found in the PDB. The choice of a $\Delta n = 10$ window is less satisfactory: a set of 59 compact proteins with $n \geq 20$ is obtained, with 22 other ranges failing to produce a sample. The interrelations between size, shape, and content of secondary structures for the compact proteins obtained with the $\Delta n = 50$ and 10 windows are discussed in Secs. III and IV.

The above use of the radius of gyration R_G (defined by the α-carbon coordinates) to estimate compactness neglects the steric contribution of the various side chains. However, the use of large windows, such as $\Delta n = 50$ and $\Delta n = 10$, within a set of proteins with a large variation in composition should produce a sample of structures which are very compact over all possible compositions. Note that we do not select the protein with the smallest R_G for a given n, but rather select the value of n which provides the smallest possible R_G among all proteins within a range of amino acid residues. This is a very stringent criterion. Nevertheless, we have further tested this choice by analyzing the compactness of the proteins in Table I with a different criterion. A composition-

TABLE I. Geometric and shape descriptors for the most compact proteins found in the working set of 373, classed within ranges of $\Delta n = 50$. (The proteins listed correspond to those with the smallest radius of gyration found in the ranges of n: 0–49, 50–99, 100–149, . . . , and 800–849. Note that there are no proteins in the list for windows of 650–699 and 700–749 amino acid residues. In order to describe properly features in medium-size to large-size proteins, molecules with $n < 20$ were excluded from the analysis.)

Protein (PDB code)	n	R_G (Å)	$\bar{N}$	A^*	Protein function
2ETI	28	7.010	4.77	0.2882	Protein (trypsin) inhibitor
1RDG	52	9.557	11.85	0.1297	Electron transport
4FD1	106	12.071	33.46	0.0548	Electron transport
2SOD	151	13.798	56.24	0.0550	Oxidoreductase
1SGT	223	15.790	101.02	0.0277	Hydrolase (serine proteinase)
3TEC	279	16.210	151.36	0.0259	Hydrolase (serine proteinase)
3CPA	306	17.866	153.80	0.0186	Hydrolase
1ALD	363	19.192	197.72	0.0144	Aldehyde lyase
7ENL	436	21.029	250.13	0.0128	Carbon-oxygen lyase
1GLY	470	21.004	279.34	0.0142	Hydrolase
1COX	502	21.814	301.76	0.0131	Oxidoreductase
1GAL	581	22.831	373.99	0.0084	Oxidoreductase
1LLA	600	23.955	367.40	0.0136	Oxygen transport
6ACN	754	25.290	524.48	0.0088	Carbon-oxygen lyase
1GPB	823	27.547	552.58	0.0087	Glycogen phosphorylase

dependent virtual "minimum" radius of gyration, $(R_G)_m$, has been proposed in Ref. [20]. This radius $(R_G)_m$ is calculated as $(\frac{3}{5})^{1/2}R^*$, where R^* is the radius of a sphere formed by the idealized close packing of the particular amino acids in a protein within a uniform spherical distribution leading to the same actual molecular volume [24]. Therefore $(R_G)_m$ becomes dependent on composition through the total volume enclosed by the protein molecular surface. Results for $(R_G)_m$, further assuming that the maximum packing density is equal to that at the core of a globular protein, are given in Ref. [20] for ten proteins with $42 \leq n \leq 316$. The comparison between $(R_G)_m$ and the actual R_G value derived from the α-carbon coordinates can be used to assess the compactness of the protein. Chan and Dill show results for ten proteins, exhibiting deviations from 3% to 15% over the ideal $(R_G)_m$ values. Following this criterion, we find that the majority of the proteins in Table I are indeed the most compact found within their Δn windows. For the proteins in Table I with $52 \leq n \leq 306$, we find $[R_G - (R_G)_m]/(R_G)_m \approx (3.8 \pm 0.9)\%$. In contrast, longer proteins are also found to be maximally compact according to this criterion but their R_G values are at least 10% above the corresponding $(R_G)_m$. The significance of this change in the extent of the compactness is addressed in the next sections.

Summarizing, the results indicate that the selected proteins (Table I) are among the most compact derived when invoking various criteria. Consequently, we believe that the relation between their molecular shapes and the number of amino acid residues n should reflect the constraints required to adopt very compact folding over all possible variations in composition. In other words, the features observed should relate to compactness in absolute terms and not to the frequency of occurrence of a particular amino acid among the proteins considered in the working set.

III. MOLECULAR SIZE AND ENTANGLEMENT COMPLEXITY IN COMPACT PROTEINS

We have studied the behavior of all molecular shape descriptors for the set of most compact proteins in Table I. From our analyses, the first protein in the set (2ETI, $n = 28$) appeared to be too short to be included in the analysis of scaling behavior. The results below are confined to the remaining 14 proteins ($n \geq 52$). For a proper comparison, the proteins selected with the $\Delta n = 10$ window start also from $n = 52$ (1RDG).

Figure 5 shows the results for the radius of gyration using the $\Delta n = 50$ window. A linear regression produces the following estimation of the critical exponent ν [cf. Eq. (2)]:

$$\ln R_G \sim (0.380 \pm 0.023) \ln n + (0.72 \pm 0.13) , \quad \mathcal{C}_\nu = 0.9955 , \quad 52 < n < 823 \ [14 \text{ pts.}] , \tag{5}$$

where slope and intercept are always given with 95% confidence intervals, $\mathcal{C}_\nu$ is the correlation coefficient associated with the estimation of ν, and R_G is expressed in Å. The value in square brackets indicates the number of compact proteins in the fitting. The exponent estimated in (5) for the most compact proteins, $\nu \approx 0.38 \pm 0.02$, does not agree with the behavior expected from the collapsed-polymer model [23]. As mentioned in Sec. I, such a model would require an exponent $\nu \approx 0.333$. This latter behavior appears in Fig. 5 as a dashed line. The agreement with the collapsed-polymer model seems to be restricted to the case of short compact proteins. Proteins longer than *ca.* 300 residues deviate systematically from

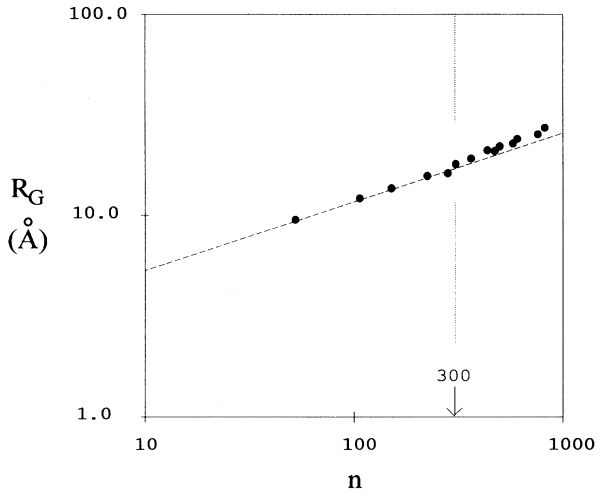

FIG. 5. Log_{10}-log_{10} plot of the radius of gyration as a function of the number of amino acid residues, restricted to the family of most compact proteins in the working set, derived with a $\Delta n = 50$ window and $n > 50$. (Notice the systematic deviation with respect to the collapsed-polymer model for chains longer than 300 residues. The dashed line represents the slope $\frac{1}{3}$, expected for collapsed homopolymers.)

this model behavior (cf. Fig. 5). Regression analyses confined to short and long proteins, respectively, indicate a difference in scaling behavior:

$$\ln R_G \sim (0.339 \pm 0.046) \ln n + (0.92 \pm 0.23) ,$$
$$\mathcal{C}_\nu = 0.9952 , \quad 52 < n < 306 \quad [6 \text{ pts.}] , \tag{6}$$

$$\ln R_G \sim (0.413 \pm 0.047) \ln n + (0.52 \pm 0.30) ,$$
$$\mathcal{C}_\nu = 0.9918 , \quad 306 < n < 823 \quad [9 \text{ pts.}] . \tag{7}$$

Results are comparable, yet less accurate, when using the smaller $\Delta n = 10$ window. For the complete range of proteins we find

$$\ln R_G \sim (0.402 \pm 0.022) \ln n + (0.63 \pm 0.12) ,$$
$$\mathcal{C}_\nu = 0.9810 , \quad 52 < n < 823 \quad [56 \text{ pts.}] , \tag{8}$$

whereas for the subsets of short and long chains, divided as before at about 300 residues, the results are

$$\ln R_G \sim (0.341 \pm 0.025) \ln n + (0.92 \pm 0.13) ,$$
$$\mathcal{C}_\nu = 0.9850 , \quad 52 < n < 306 \quad [26 \text{ pts.}] , \tag{9}$$

$$\ln R_G \sim (0.403 \pm 0.084) \ln n + (0.63 \pm 0.51) ,$$
$$\mathcal{C}_\nu = 0.8773 , \quad 306 < n < 823 \quad [31 \text{ pts.}] . \tag{10}$$

Comparing the slopes in Eqs. (5)–(7) and (8)–(10), we note an agreement between the apparent scaling exponents for both regimes. However, the correlation is very poor for long chains if the window for locating compact proteins is made as small as $\Delta n = 10$. This is an indication that some noncompact structures are included in the set with the smaller Δn window. For this reason, we shall restrict the analysis of differential scaling in other shape descriptors to the more controlled set defined with the larger $\Delta n = 50$ window.

The mean numbers of overcrossings, $\bar{N}$, are displayed in Fig. 6 as a function of the number of residues for the most compact proteins. A comparison with Fig. 4 indicates a smaller dispersion when only compact proteins are considered, especially in the case of $\Delta n = 50$. For this shape descriptor, we also observe different scaling behavior for short and long compact proteins. A linear regression for all compact proteins gives

$$\ln \bar{N} \sim (1.40 \pm 0.04) \ln n + (-3.0 \pm 0.2) ,$$
$$\mathcal{C}_\beta = 0.9989 , \quad 52 < n < 823 \quad [14 \text{ pts.}] , \tag{11}$$

whereas the restrictions to short and long compact proteins provide

$$\ln \bar{N} \sim (1.48 \pm 0.08) \ln n + (-3.4 \pm 0.4) ,$$
$$\mathcal{C}_\beta = 0.9992 , \quad 52 < n < 306 \quad [6\text{pts.}] , \tag{12}$$

$$\ln \bar{N} \sim (1.31 \pm 0.07) \ln n + (-2.4 \pm 0.4) ,$$
$$\mathcal{C}_\beta = 0.9984 , \quad 306 < n < 823 \quad [9 \text{ pts.}] , \tag{13}$$

respectively. Note that the quality of the correlations for R_G and $\bar{N}$ is comparable in the case of the compact proteins, whereas it differs largely when viewed over the

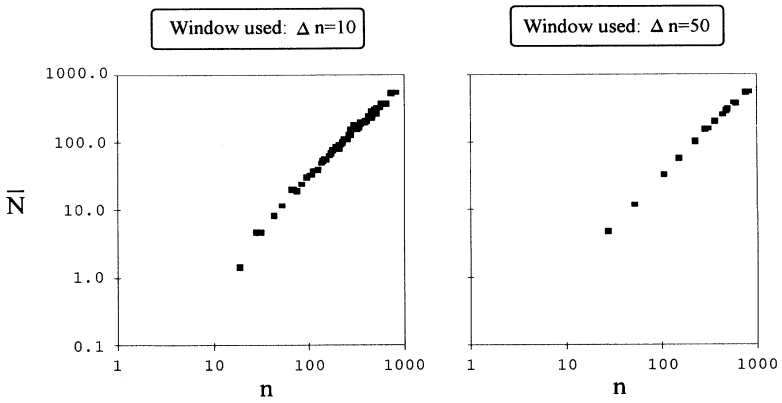

FIG. 6. Log_{10}-log_{10} plot of the mean number of overcrossings, $\bar{N}$, as a function of the number of amino acid residues. [Results are given for the most compact proteins in the working set, derived with two windows, $\Delta n = 10$ (left) and $\Delta n = 50$ (right).]

whole working set of proteins. As was found in the case of the radius of gyration, the difference in scaling exponents β between the two series of compact proteins is significant [cf. Eqs. (12) and (13)]. The results above suggest the following.

(a) The most compact proteins with fewer than 300 amino acid residues are characterized by a "size exponent" $\nu \sim 0.34 \pm 0.05$ and an "entanglement exponent" $\beta \sim 1.5 \pm 0.1$.

(b) For compact proteins with more than 300 residues the corresponding exponents are $\nu \sim 0.41 \pm 0.05$ and $\beta \sim 1.3 \pm 0.1$. Therefore only the shorter compact proteins seem likely to appear as truly "collapsed" configurations.

The reliability of the approximate point for the change in scaling regimes (ca. 300 residues) has been checked in two different ways. Table II shows the estimated exponents ν and β obtained by least-squares fittings of series of five consecutive proteins from Table I, as well as their correlation coefficients $\mathcal{C}_\nu$ and $\mathcal{C}_\beta$. The five proteins considered define an interval $[n_i, n_f]$ of approximately 200 residues in length, for proteins with less than 582 residues (i.e., $n_f < 582$). The quality of the correlations for both exponents depends on the location of the interval $[n_i, n_f]$. From Table II, the following cases can be distinguished.

(i) The correlation is high in the range of [52,279] residues, which provides the exponents $\nu \sim 0.33$ and $\beta \sim 1.50$.

(ii) The correlation decreases when the interval $[n_i, n_f]$ is displaced towards longer compact proteins. The poorest correlation in both exponents appears in the range [223,436] centered about $n \approx 329$.

(iii) Both correlations improve again over the ranges [306,502] and [363,581], giving exponents $\nu \sim 0.40$ and $\beta \sim 1.36$.

(iv) Correlations are poor again over the last three series of five proteins, although the derived exponents ν and β are consistent with the values found in (iii). A likely reason for these poorer correlations is that the ranges of amino acid residues become erratic over the last three ranges, due to the reduced sample of available experimental compact proteins. Whereas the range of residues in the cases (i)–(iii) is $n_f - n_i \approx 208 \pm 13$, the last three ranges are 164, 284, and 321, respectively.

In conclusion, the results in Table II are consistent with the existence of two distinct scaling regimes of size and shape for compact proteins, approximately separated at $n \approx 300$. Note that the change in correlations is exactly the same for both exponents, even though they are associated with two independent properties ("size" and "shape"). The more clearly defined behaviors are found, approximately, for proteins with numbers of residues $50 < n < 300$ ("short" proteins) and $300 < n < 550$ ("long" proteins). We believe that the correlations involving *only* proteins with more than 450 residues are not reliable, due to the irregularity of the sampling in the region.

The correlations in Table II are derived by sequences of linear fittings with a constant number of data points. This approach does not produce fittings with a constant interval length in a logarithmic plot. For this reason, we have also tested linear correlations of the form $\ln R_G$ vs $\ln n$ and $\ln \bar{N}$ vs $\ln n$, where the length of the interval is kept constant [i.e., $\ln(n_f/n_i) \approx \text{const}$]. Linear regressions were performed on groups of compact proteins satisfying approximately the condition $\ln(n_f/n_i) \approx 1.0$, and the results appear in Table III. (Note that the number of proteins in each correlation is no longer constant.) The findings in Table III are consistent with our previous observations: short compact proteins and long compact proteins can be differentiated by their pairs of apparent scaling exponents ν and β. The values given for these pairs in Tables II and III agree. Moreover, the change between the two regimes appears to take place somewhere midway in the interval [151,436] (cf. Table III). This interval is centered at $n = 293$, which provides an estimation of the transition point not far from the previous one $n \approx 300$ (cf. Fig. 2).

TABLE II. Estimation of the scaling exponents ν and β for the radius of gyration R_G and the mean number of overcrossings, $\bar{N}$, for the most compact proteins found in the working set of 373, classed within ranges of $\Delta n = 50$. The table gives ν and β estimated with 95% confidence and the respective correlation coefficients $\mathcal{C}_\nu$ and $\mathcal{C}_\beta$. [The exponents are computed by analyzing series of five consecutive proteins. The changes in the correlation coefficients suggest two distinct scaling regimes. The two regimes appear to be characterized by (1) $\nu \approx 0.33$, $\beta \approx 1.5$ for $50 < n < 300$, and (2) $\nu \approx 0.40$, $\beta \approx 1.3$ for $300 < n < 550$. Results involving proteins with $n > 600$ may not be reliable since it is difficult to assess if the few structures in this range correspond to truly compact proteins.]

Range of five proteins considered ($n_i \rightarrow n_f$)	ν (±95% error)	$\mathcal{C}_\nu$	β (±95% error)	$\mathcal{C}_\beta$
52→279	0.33±0.05	0.996	1.50±0.08	0.9994
106→306	0.34±0.10	0.987	1.49±0.18	0.9978
151→363	0.36±0.16	0.973	1.45±0.26	0.9951
223→436	0.45±0.20	0.971	1.31±0.32	0.9914
279→470	0.49±0.20	0.977	1.24±0.30	0.9916
306→502	0.40±0.10	0.990	1.36±0.07	0.9996
363→581	0.36±0.12	0.984	1.35±0.08	0.9995
436→600	0.40±0.22	0.955	1.27±0.29	0.9924
470→754	0.39±0.16	0.978	1.33±0.24	0.9952
502→823	0.44±0.19	0.974	1.27±0.27	0.9933

TABLE III. Estimation of the scaling exponents ν and β for compact proteins in the limit of a large number of amino acid residues. The computations are performed by linear regressions in ranges of n values of approximately the same length in a logarithmic plot. The results below correspond to ranges of approximately length 1, that is, including all proteins with numbers of monomers between n_i and the closest integer to $n_f \sim e \times n_i$. (The exponents ν and β are given with their statistical errors at a 95% confidence level. The corresponding correlation coefficients are $\mathcal{C}_\nu$ and $\mathcal{C}_\beta$, respectively.)

Range of proteins considered ($n_i \rightarrow n_f$)	$\ln(n_f/n_i)$	ν ($\pm$95%)	$\mathcal{C}_\nu$	β ($\pm$95%)	$\mathcal{C}_\beta$
52→151	1.066	0.34±0.16	0.999	1.46±0.03	1.0000
106→279	0.968	0.31±0.14	0.989	1.55±0.19	0.9992
151→436	1.060	0.39±0.11	0.979	1.41±0.18	0.9959
223→600	0.990	0.43±0.06	0.988	1.30±0.09	0.9969
279→754	0.994	0.43±0.06	0.987	1.28±0.09	0.9971
306→823	0.989	0.41±0.05	0.992	1.31±0.07	0.9984

Results in Table III can be used to estimate the scaling behavior for virtual, very long compact proteins ($n \rightarrow \infty$). Figures 7 and 8 display the approximate scaling exponents ν and β as a function of $1/n_f$. The results include linear regressions on the groups of proteins satisfying the constraints $\ln(n_f/n_i) \approx 1.0$ and $\ln(n_f/n_i) \approx 1.5$ as closely as possible. (None of the linear regressions includes exclusively proteins with $n > 450$.) Despite the large dispersions, we can make rough assessments of the scaling exponents ν_∞ and β_∞ in the large n limit:

$$\nu_\infty \sim 0.44 \pm 0.06 \ , \quad \beta_\infty \sim 1.20 \pm 0.14 \ . \tag{14}$$

The result for ν_∞ would imply a conformational state below the θ conditions for long compact proteins, yet far from the collapsed-polymer situation. The value for β_∞ agrees with the result in Ref. [15] for self-avoiding walks with attractive interaction ($\beta_\infty \sim 1.18$). These values for $n \gg 1$ are conjectural and derived from the available sample of compact proteins. It remains to be seen if they hold after very long compact proteins are discovered or experimentally synthesized.

The discussion of shape descriptors for compact proteins can be completed by considering the maximum probability of overcrossings, A^*. The dispersion for this shape descriptor is larger than the dispersions in R_G and $\bar{N}$ for the same sets of proteins. In the case of $\Delta n = 50$, a linear regression gives

$$\begin{aligned} &\ln A^* \sim (-1.05 \pm 0.09) \ln n + (2.1 \pm 0.5) \ , \\ &\mathcal{C}_b = 0.9889 \ , \quad 52 < n < 823 \quad [14 \text{ pts.}] \ , \end{aligned} \tag{15}$$

which indicates a scaling exponent $b \sim -1.0$, as found for model polymers [8]. However, no difference in the behavior for short and long compact chains can be assessed from A^*. For the analysis of scaling in the entanglement complexity of proteins with specific structural features one must use the mean number of overcrossings.

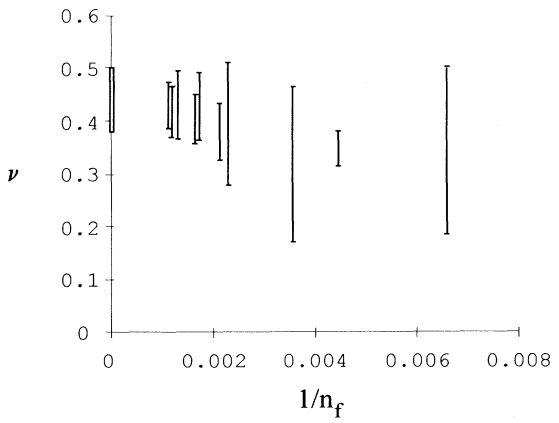

FIG. 7. Estimation of the scaling exponent ν for the radius of gyration in the limit of a very large number of residues. The effective exponents ν have been computed by linear regressions of $\ln R_G$ vs $\ln n$, with intervals of $[n_i, n_f]$ with an approximately constant $\ln(n_f/n_i)$ ratio, for the compact proteins in the $\Delta n = 50$ window. [The figure shows the results obtained with approximate $\ln(n_f/n_i)$ ratios of 1.0 and 1.5. The estimated extrapolated value is shown as an open white sector in the limit $n_f \rightarrow \infty$.]

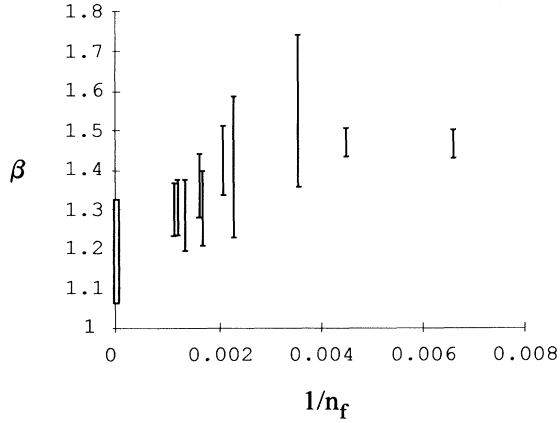

FIG. 8. Estimation of the scaling exponent β for the mean number of overcrossings in the limit of a very large number of residues. The effective exponents β have been computed by linear regressions of $\ln \bar{N}$ vs $\ln n$, within intervals of $[n_i, n_f]$ with an approximately constant $\ln(n_f/n_i)$ ratio, for the compact proteins in the $\Delta n = 50$ window. [The figure shows the results obtained with approximate $\ln(n_f/n_i)$ ratios of 1.0 and 1.5. The estimated extrapolated value is shown as an open white sector in the limit $n_f \rightarrow \infty$.]

IV. VARIATIONS IN THE CONTENT OF SECONDARY STRUCTURE FOR COMPACT PROTEINS

The distinct behavior of molecular size and entanglement complexity for short and long compact native protein structures may reflect a fundamental difference in their structures. As discussed above, the "size exponent" ν for the short chains is close to that of the collapsed-polymer model ($\nu=\frac{1}{3}$), whereas a larger value is found in the longer chains. The result implies that the short compact proteins are actually "more compact" than the compact proteins with $n > 300$ residues.

The occurrence of a more complex scaling in proteins than in homopolymers is in principle not surprising. A heteropolymer can have a ground-state (native) structure which is not maximally compact due to the occurrence of specific interactions which will vary with each primary sequence [25,26]. Early studies of hydrophobicity indicate that the stability of the globular state of a protein depends on its composition and chain length, and that not all chain lengths can be stabilized in spheroidal globules [25]. Lattice heteropolymer models predict that long chains will be stabilized in nonspherical conformations comprising several globular domains, and therefore will be less compact than short chains [25]. This difference in compactness and shape for long and short heteropolymer chains agrees qualitatively with our finding of differential scaling in compact proteins. The transitional critical length of $n \approx 300$ residues appears to be shorter than the values derived from polymer models with excluded volume and hydrophobic interactions.

In order to interpret the distinct shape and size scalings from another viewpoint, we have analyzed the secondary structural content in our family of compact proteins. Here, the content of secondary structure is simply conveyed by the two dominant features: the percentage of amino acid residues forming part of α helices or β strands. Other less dominant features, such as β turns, have been omitted. The number of residues belonging to either secondary structural motif has been extracted from the standard PDB files. (This information is derived using the standard dihedral angles for helices or sheets [27].) This information is not provided to the user at every entry in the data bank. Whenever the classification was not available, we omitted the protein from the analysis.

Our results are summarized in Figs. 9 and 10, using the $\Delta n = 50$ and 10 windows, respectively. The content of helices and strands for the most compact proteins is shown as a function of the number of residues, n.

Even though secondary structural content is not available for all proteins in Table I, Fig. 9 gives a strong indication of a systematic change in the percentage of β strands and α helices for short and long compact proteins. The content of β strands is almost twice as large for the short proteins (ca. 35% vs ca. 17%, respectively). Moreover, the change in content clearly takes place between 250 and 300 amino acids, which is close to the number of residues where a change in scaling behavior was found in the size and entanglements. This change in β-strand content seems to be accompanied by an opposite

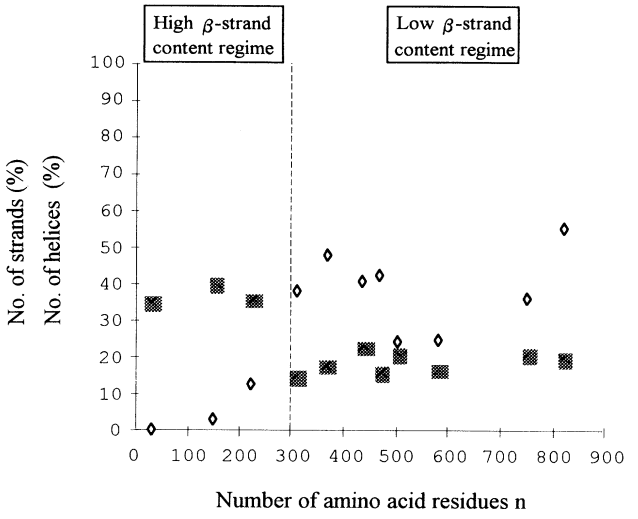

FIG. 9. Content of secondary structure in the most compact proteins derived from the working set with a $\Delta n = 50$ window. [Note the approximate inversion in the content of α helices (diamonds) and β strands (squares) for the compact proteins above and below $n \approx 300$. The actual ranges of amino acid residues used are 20–49, 50–99, 100–149, 150–199, . . . , 800–849.]

change in α-helical content, although the latter is more erratic.

Figure 10 is consistent with these observations. In this case, we note that the majority of the structures with $n > 300$ have a β-strand content of ca. 15–20 %, whereas proteins with $n < 300$ exhibit a larger content. An opposite behavior is found for the percentage of helices. Once again, the clearest separation point appears to be close to

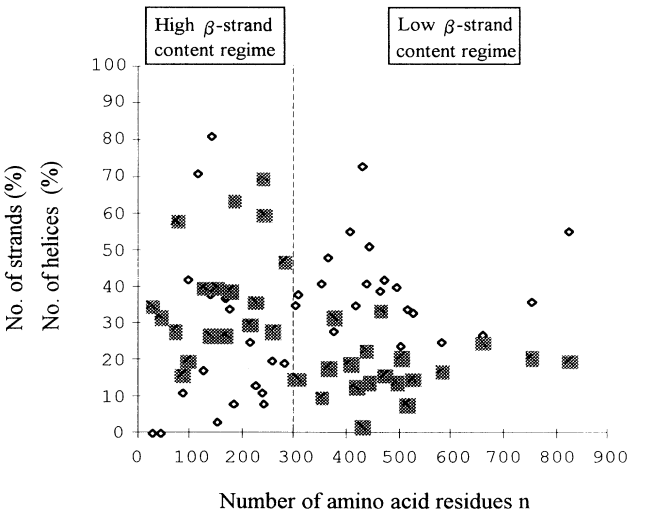

FIG. 10. Content of secondary structure in the most compact proteins derived from the working set with a $\Delta n = 10$ window. [Note the approximate inversion in the content of α helices (diamonds) and β strands (squares) for the compact proteins above and below $n \approx 300$. The result is consistent with the one observed in Fig. 9 for a larger Δn window. The actual ranges of amino acid residues used are 20–29, 30–39, 40–49, . . . , 820–829.]

300 residues. The larger dispersion in Fig. 10 compared with Fig. 9 is consistent with the inclusion of more noncompact structures with a smaller Δn window.

Based on this distinct behavior, we suggest that the short and long compact proteins available to date indeed differ in structure. A very compact protein with fewer than ca. 300 residues can be formed with a relatively large percentage of β sheet, typically 30–40 %. When the number of residues is over 300, such a large content of β sheet does not appear to be compatible with a structure of the same degree of compactness. This difference is possibly an expression of the known fact that longer proteins have the capability to form stable supersecondary structures, such as α/β barrels [27], which are not accessible to small proteins. For instance, some 250 residues are needed to form α/β barrels [27]. It is therefore conceivable that, by organizing supersecondary motifs, long compact proteins can be stabilized in conformations which are "more open" than the stable conformations accessible to shorter compact proteins. Supersecondary structures for long proteins, though compact, would thus not resemble "collapsed" polymer configurations.

The variation in the content of secondary structure with chain length for compact proteins agrees with the finding from lattice models that long compact chains can deviate from sphericity by forming smaller stable globular domains [25]. From the present work, such domains should exhibit a low content of β-sheet structure. Domains with high β-sheet content appear to provide a less dense packing for long chains than domains with a high α-helical content.

V. CONCLUSIONS

In this work, we have analyzed the scaling behavior of the molecular size and the entanglements for very compact proteins of variable length. The most reliable set requires the choice of a relatively large Δn window of amino acid residues and thus produces a small sample. An acceptable compromise is found with $\Delta n = 50$. This condition leads to a set of 14 proteins, mostly involved with transport, redox, and hydrolytic functions. Despite their different local features, these proteins share a common global structural feature: maximum conformational compactness within a range of chain lengths. Scaling properties of this special subset have been studied in detail. Due to the large size and variety of the initial working set (373 structures), we believe that the proteins selected are indeed among the most compact known to date, at least for $n < 450$.

Our main findings are as follows.

(a) A definite scaling for R_G is observed when one studies those proteins with the smallest R_G values compatible with a given range $[n, n+\Delta n]$ of amino acid residues. Such a scaling behavior does not exist in the initial working set of structures. For the family of compact proteins, definite scaling is also found in the descriptors of self-entanglements in molecular chains (the "molecular shape").

(b) Two scaling regimes appear to take place in terms of protein length. Compact proteins with less than 300 residues are characterized by a scaling exponent $\nu = 0.34$, close to the value for collapsed polymers, and an entanglement exponent $\beta = 1.5$. Longer compact proteins provide a larger ν exponent and a smaller β exponent. This is an indication that the longer proteins are "less compact:" a larger ν value leads to bigger, more open chains. Consistently, a smaller β value leads to smaller mean numbers of overcrossings, i.e., chains with less complex entanglements.

(c) The change in scaling regimes correlates well with a change in the content of secondary structure. Shorter proteins become compact by achieving a large content of β strands. In contrast, the longer proteins are stabilized in less compact ("noncollapsed") structures by forming supersecondary motifs which involve a smaller content of sheets. One such possible motif is an α/β barrel [27].

In conclusion, the present work proposes some relationships among the number of residues, the radius of gyration, the entanglement complexity, and the content of secondary structure needed to form a very compact (possibly "collapsed") polypeptide. We believe that these results should be valuable for a better understanding of protein folding features [28,29], the structure of dynamic intermediates [3,30,31], and for improvements in the *de novo* synthesis of proteins [32].

ACKNOWLEDGMENTS

I thank Professor S. G. Whittington (Toronto) for sending his manuscript of Ref. [15] prior to publication and M. Payette (Sudbury) for his collaboration in computing shape descriptors for protein backbones. This work has been supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

[1] K. A. Dill and D. Shortle, Annu. Rev. Biochem. **60**, 795 (1991).
[2] J. M. Thornton, Curr. Opinion Struct. Biol. **2**, 888 (1992).
[3] V. E. Bychkova and O. B. Ptitsyn, Chemtracts-Biochem. Mol. Biol. **4**, 133 (1993).
[4] G. A. Arteca and P. G. Mezey, Biopolymers **32**, 1609 (1992).
[5] E. J. Janse van Rensburg, D. W. Sumners, E. Wasserman, and S. G. Whittington, J. Phys. A **25**, 6557 (1992).
[6] G. A. Arteca, Biopolymers **33**, 1829 (1993).
[7] G. A. Arteca, J. Phys. Chem. **97**, 13 831 (1993).
[8] G. A. Arteca, Phys. Rev. E **49**, 2417 (1994).
[9] P.-G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, 1985).
[10] P. J. Flory, *Statistical Mechanics of Chain Molecules* (Interscience, New York, 1969).
[11] J.-C. Le Guillou and J. Zinn-Justin, Phys. Rev. B **21**, 3976 (1980).
[12] H. E. Stanley, J. Phys. A **10**, L211 (1977).
[13] I. M. Lifshitz, A. Y. Grosberg, and A. R. Khokhlov, Rev. Mod. Phys. **50**, 683 (1978).
[14] P. Biswas and B. J. Cherayil, J. Chem. Phys. **100**, 4665

(1994).
[15] E. Orlandini, M. C. Tesi, S. G. Whittington, D. W. Sumners, and E. J. Janse van Rensburg, J. Phys. A **27**, L333 (1994).
[16] L. L. Walsh (personal communication).
[17] H. S. Chan and K. A. Dill, Macromolecules **22**, 4559 (1989).
[18] E. Shakhnovich and A. Gutin, J Chem. Phys. **93**, 5967 (1990).
[19] J. D. Honeycutt and D. Thirumalai, Proc. Nat. Acad. Sci. U.S.A. **87**, 3526 (1990).
[20] H. S. Chan and K. A. Dill, J. Chem. Phys. **95**, 3775 (1991).
[21] The whole list of the proteins in the study can be obtained upon request to the author.
[22] The one point completely outside the main cluster corresponds to a very unusual structure: the muscle protein tropomyosin (coded 2TMA), which is a 284-residue 97% α-helical coiled coil. This is the only example we found in the PDB of a very long, purely helical structure.
[23] T. G. Dewey, J. Chem. Phys. **98**, 2250 (1993).
[24] S. Miyazawa and R. L. Jernigan, Macromolecules **18**, 534 (1985).
[25] K. A. Dill, Biochemistry **24**, 1501 (1985).
[26] M. Karplus and E. Shakhnovich, in *Protein Folding,* edited by T. E. Creighton (Freeman, New York, 1992).
[27] C. Brändén and J. Tooze, *Introduction to Protein Structure* (Garland, New York, 1991).
[28] C. Chothia and A. V. Finkelstein, Annu. Rev. Biochem. **50**, 537 (1990).
[29] D. T. Jones, W. R. Taylor, and J. M. Thornton, Nature **358**, 86 (1992).
[30] V. Daggett and M. Levitt, Proc. Nat. Acad. Sci. U.S.A. **89**, 5142 (1992).
[31] V. Daggett and M. Levitt, Curr. Opinion Struct. Biol. **4**, 291 (1994).
[32] J. S. Richardson and D. C. Richardson, in *Proteins: Form and Function,* edited by R. A. Bradshaw and M. Purton (Elsevier, Cambridge, England, 1990).