# Self-similarity in entanglement complexity along the backbones of compact proteins

Gustavo A. Arteca
*Département de Chimie et Biochimie, Laurentian University, Ramsey Lake Road, Sudbury, Ontario, Canada P3E 2C6*

(Received 25 February 1997; revised manuscript received 5 May 1997)

The mean overcrossing number is a useful descriptor of the nature and complexity of self-entanglements in polymer conformations. We show that this descriptor exhibits a degree of self-similarity along the backbones of protein native states. We have estimated the scaling exponent for the power-law behavior of the mean number of overcrossings as a function of the contour length *within* a fixed (compact) backbone. The reported scaling behavior is found in self-entanglements and not in descriptors of molecular size. The result provides a useful criterion for the elucidation of protein conformations. [S1063-651X(97)11909-2]

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

Despite the specific relation between protein composition and its biochemical function, most protein native states are known to share several basic three-dimensional (3D) structural features [1]. The interpretation of these similarities is subject of much debate [2,3]. If we look for only essential *global shape features*, qualitative similarities could be used to define an ''average'' native state, containing all rough 3D features common to most proteins. It would be desirable that such an average state should be characterized by a *scaling regime*, as defined by the dependence of some shape descriptor with the number of amino acid residues, $n$. However, standard descriptors of molecular size used in polymer statistics (e.g., the radius of gyration, $R_G$) do not exhibit a clear regime for proteins and thus do not provide a characterization for the ''average'' native state [4–6]. A study of global molecular shape similarities within a large protein database requires alternative descriptors. These should meet some criteria: (i) to be not explicitly dependent (*a priori*) on molecular size; (ii) to exhibit scaling with the number of residues; (iii) to take into account the 3D ''trace'' of the backbone. Measures of *chain self-entanglements* provide such a descriptor for polymers [7–9]. Here, we expand the analysis of these descriptors and report on their scaling behavior for experimental protein backbones.

Self-entanglements convey the ''twists, turns, and folds'' found along a polymer chain in a rigid configuration. These features depend on the connectivity of the polymer backbone, and not only on the spatial position of the monomers [7,8]. As a result, self-entanglements provide a better tool than molecular size or anisometry descriptors for comparing ''folding topologies'' [9–11]. Previous work in the literature has used measures of backbone entanglement to assess global homologies between protein folds [9] and to monitor changes in three-dimensional shape during conformational rearrangements triggered by ligand binding [11].

Self-entanglements can be characterized by geometrical or topological descriptors. A simple geometrical descriptor uses the notion of ''*backbone overcrossings*'' [7,8]. Overcrossings (or ''double points'' in knot theory) are the points where two bonds appear to cross in a regular 2D projection of a rigid 3D conformation. A simple descriptor of entanglement is the *mean number of overcrossings* $\overline{N}$ (or ''average crossing number''), computed as an average over all possible rigid projections [7–9]. Its configurational average over an ensemble of accessible conformers will be indicated by $\langle \overline{N} \rangle$. As a shape descriptor, $\langle \overline{N} \rangle$ conveys briefly the folding nature of accessible conformers. In an unentangled chain (e.g., a rodlike rigid polymer or a polymer in a ''good'' solvent), most 3D projections will not produce overcrossings, and thus we expect small $\langle \overline{N} \rangle$ values. The more entangled the chain, the larger the $\langle \overline{N} \rangle$ value. Note that very compact chains will necessarily produce large overcrossing numbers. However, in polymers with intermediate compactness (e.g., globular proteins) the degree of entanglement is not directly related to molecular size. Two polymers with similar molecular size or anisometry may yet be distinguished by their folding features [10,11]. Therefore, mean overcrossing numbers provide a powerful additional tool to discriminate between polymers according to their folding patterns.

The physical relevance of these descriptors is becoming apparent. In the case of knotted DNA, $\langle \overline{N} \rangle$ correlates with the electrophoretic diffusion velocity [12,13]. Also, the value of $\langle \overline{N} \rangle$ for a tubelike polymer knot with maximum cross sectional diameter appears to be a topological invariant [14]. These recent developments show that a descriptor of entanglement can provide more insights into the nature of chain configurations than simple molecular size descriptors. In addition, they also indicate that geometrical descriptors of entanglement can yield information on polymer topology. In this work, we apply these notions to study linear polymers.

The presence of common global features among polymer configurations produces a well-defined scaling law for $\langle \overline{N} \rangle$ in terms of the monomer number $n$. Lattice [15] and off-lattice [16] polymer simulations indicate $\langle \overline{N} \rangle \sim B n^{\beta}$, where the scaling exponent $\beta$ changes little with the monomer-monomer interaction. It can be easily proved that $1 \leqslant \beta \leqslant 2$ in three dimensions [15]. Numerical results for polymer models suggest a lower bound of $\beta > 1.12$ [15], an upper bound of $\beta < 1.40$ [16] using medium size chains, and an extrapolated value of $\beta \approx 1.2$ in very long chains. Conjectural arguments appear to support a value below 1.40 [17]. Recently, we have shown that a similar exponent is found within an ensemble of ca. 350 experimental protein native states [8]. It should be noticed that this scaling law is only found in a geometric descriptor of entanglements. Descriptors of molecular size (e.g., $R_G$) do not exhibit a clear scaling within the same set of structures.

A scaling law in the shape of actual backbones is a useful tool for modeling protein structure [5]. For instance, it can be used to discard unlikely features when testing possible folding patterns. Here, we extend further the previous work on proteins by analyzing the scaling behavior of $\overline{N}$ in much greater detail. We are interested in the following two questions: (a) Is there a differential scaling behavior (i.e., a variation in $\beta$ exponent) associated with whether protein native states are compact or not? (b) Consider a ''series'' of chains defined by increasingly longer partial sections of a given backbone. Is there a scaling behavior in the mean overcrossing numbers associated with this sequence of backbones, regardless of primary sequence? Such a behavior would imply a measure of self-similarity in entanglements along protein backbones.

In order to address these two questions, we have considered a working set with 904 single-monomer protein backbones in their experimental native states (as deposited in the Brookhaven Protein Data Bank, PDB [18]). Each structure is analyzed in terms of the $\alpha$-carbon backbone (i.e., one atom per amino acid residue). Our ensemble mimics the known distribution of backbone lengths and it has no bias with respect to a particular secondary structural feature. The set spans a large variety of 3D structures, molecular sizes, and primary sequences.

The $\overline{N}$ values have been computed to an accuracy of three significant figures, using an algorithm explained elsewhere [9]. We will distinguish between the mean number of overcrossings of the entire backbone and of a section of it. Consider the section of an $n$-residue backbone (in its native state) containing its first $n_1 \geqslant 3$ residues. The mean overcrossing number associated with this section is denoted by $\overline{N}(n_1,n)$. [Since three consecutive atoms define a plane and do not overcross, then $\overline{N}(3,n)=0$ for all $n \geqslant 3$.]

We study the molecular shape features of (a) the full chain, characterized by $\overline{N}(n,n)$, (b) the sequence of backbone segments with lengths $n_k = kn_1 \leqslant n$, $k = 1,2,3,\ldots,\max(k) = \mathrm{int}[n/n_1]$. For the full backbone, we test a scaling law:

$$\overline{N}(n,n) \sim Bn^{\beta}, \tag{1}$$

whereas for sections of a given backbone we test an alternative scaling:

$$[\overline{N}(n_k,n)]_n \sim Gn_k^{\gamma}, \tag{2}$$

where $[\overline{N}(n_k,n)]_n$ indicates that $n_k$ varies and $n$ is kept constant (i.e., a scaling law within a single native state). We shall use the terms ''global'' and ''local'' scaling exponents when referring to $\beta$ and $\gamma$, respectively. In general, the exponents $\beta$ and $\gamma$ can be different. Similar scaling laws can also be tested in other descriptors, e.g., by defining $[R_G(n_k,n)]_n$ for the molecular size.

First, we address the question of the possible dependence of exponent $\beta$ on backbone compactness. Figure 1 indicates a well-defined scaling behavior in the working set of 904 proteins. (Such a behavior is not found in $R_G$ or other molecular size descriptors.) A mean scaling law for the entire set can be extracted from a $\ln(\overline{N})$-$\ln(n)$ regression. Restricting the correlation to the 684 ''long'' proteins ($n \geqslant 100$) in our set, we obtain

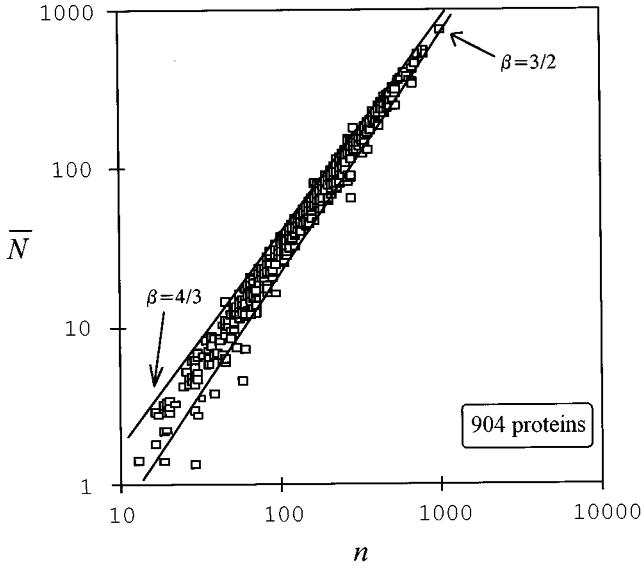

FIG. 1. Scaling behavior of the mean number of overcrossings $(\overline{N})$ $\alpha$-carbon backbones with $n$ amino acid residues. [Each square represents a backbone in its native state. The straight lines provide qualitative bounds to the ''effective'' scaling exponent $\beta$.]

$$\overline{N}(n,n) \sim (0.050 \pm 0.005) n^{(1.39 \pm 0.02)}, \quad C = 0.9892, \tag{3}$$

where $C$ is the correlation coefficient and the errors corresponds to 95% confidence intervals. The scaling exponent $\beta = 1.39 \pm 0.02$ refines previous estimations derived within a much smaller sample of proteins [8]. [For completeness, we have also checked an asymptotic behavior of the form $\overline{N}(n,n) \sim B'(n-3)^{\beta}$, which satisfies the limit $\overline{N}(3,3)=0$. The results are comparable to those in Eq. (3): $B' = 0.055 \pm 0.006$ and $\beta = 1.37 \pm 0.02$.]

The large working set used here allows us to do further refinements and address the first question mentioned above. As suggested by Fig. 1, two slightly different scaling laws could be conjectured. Our results are bound between those corresponding to two limit $\beta$ values. Proteins with large mean overcrossing numbers appear to follow a law with scaling exponent $\beta \approx 4/3$. In contrast, native states with a minimal number of overcrossings follow a law with $\beta \approx 3/2$. The overcrossing numbers for large $n$ would appear to be already in the asymptotic regime, since no systematic curvature is observed after $n > 100$. Thus, the difference in scaling does not seem to be an artifact due to low residue numbers. (The exceptional points found Fig. 1 are mostly $\alpha$-helical proteins. These are not typical native states, and follow a different scaling law.

Proteins whose mean overcrossing numbers grow as $\overline{N} \sim n^{4/3}$ appear to include those with maximally compact backbones, i.e., those with a minimum value of backbone radius of gyration within a range of $n$ values [8]. We have checked the consistency of this observation by correlating $R_G^2$ and $\overline{N}$ values for compact proteins with $n < 300$ [8]. We obtain

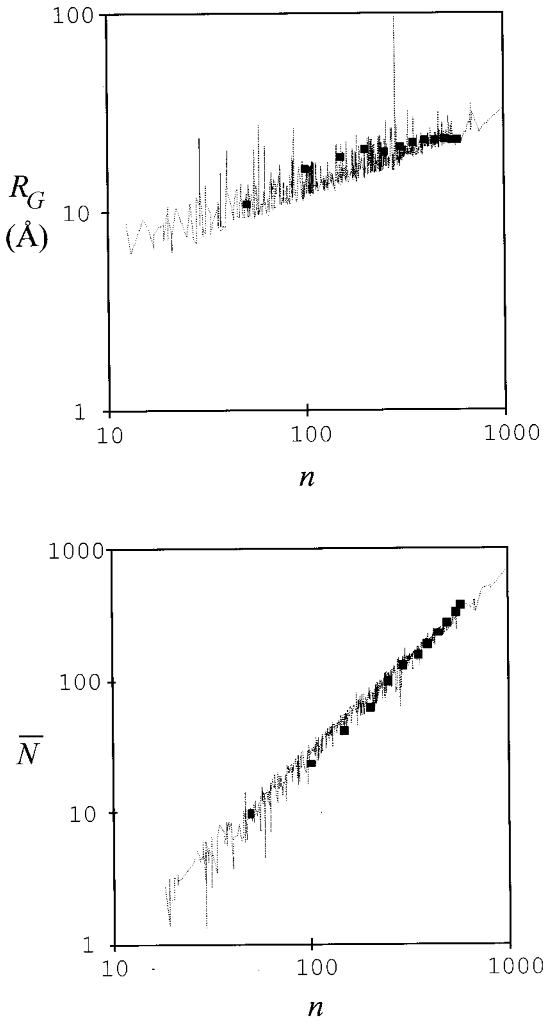

FIG. 2. Compared scaling behavior of the radius of gyration ($R_G$) and the mean number of overcrossings ($\overline{N}$) for sequential sections of the 1GAL protein. [The results for this protein (black squares) are contrasted with the average scaling of the entire set of full backbones (thin lines).]

$R_G^2 \sim \overline{N}^{\delta}$, with $\delta \approx 0.50 \pm 0.03$, and $C = 0.9930$. Since short, very compact proteins are found in the collapsed polymer regime, (i.e., $R_G \sim n^{1/3}$ [5]), the expected scaling coefficient should be $\beta = 2/3\delta \approx 4/3$.

The above mean scaling behavior of an ''average'' native state provides a reference for the analysis of ''local'' self-entanglements, i.e., *within* a given protein backbone. Our main result in this work is that a law such as Eq. (2) does indeed take place, at least for compact proteins.

Figure 2 illustrates the typical results for the mean overcrossing numbers along a compact backbone, $[\overline{N}(n_k, n)]_n$. The example shows glucose oxidase (PDB code 1GAL). This protein has the smallest $R_G$ value among the backbones with residue number $550 \leqslant n \leqslant 600$. Figure 2 compares the behavior of $R_G$ for the partial backbone segments ($[R_G(n_k, n)]_n$) with that of $[\overline{N}(n_k, n)]_n$. [The CPU time required to characterize the ''local scaling'' grows as $t_{\mathrm{CPU}} \approx k^3 t(n_1)/3$, where $t(n_1)$ is the CPU time needed to compute the first section of $n_1$ residues. The analysis of 1GAL, with $n = 581$ and $n_1 = 50$, requires ca. 9 h CPU on a DEC 255/233 AXP workstation.]

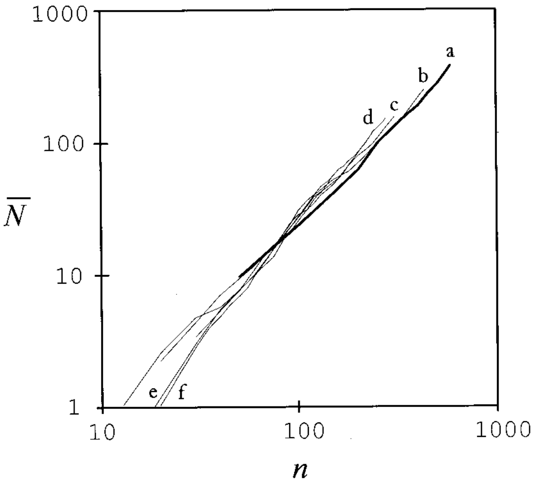

FIG. 3. Self-similarity in overcrossing numbers along the chains of selected proteins with maximally compact backbones. [The letters stand for (a) 1GAL (thick line), (b) 7ENL, (c) 3CPA, (d) 3TEC; (e) 1SGT, (f) 2SOD.]

Figure 2 contrasts the values of $[R_G(n_k, n)]_n$ and $[\overline{N}(n_k, n)]_n$ for 1GAL with those for the *full backbones* of 904 native states (thin line). Two general observations are illustrated in Fig. 2: (a) There is no ''local'' scaling law for molecular size. As the length of the backbone segments increases, the radius of gyration initially increases. However, in segments longer than $n_k > 250$ the center of mass shifts in such a way that the radius of gyration $[R_G(n_k, n)]_n$ remains essentially constant. All segments longer than $n_k > 250$ cannot be distinguished according to molecular size, even though the protein chain becomes more convoluted as $n_k$ increases. (b) There is a clear ''local'' scaling for chain self-entanglements. Whereas the molecular size remains constant in segments longer than 250 residues, the chain entanglements continue to increase according to a well-defined power law. Qualitatively, the local scaling for $[\overline{N}(n_k, n)]_n$ in 1GAL is similar to the behavior found in the ''average'' native state (thin line).

The same essential features are also found in other compact proteins. (The discussion below is restricted to the case of very compact proteins. These provide the best case for analysis, since they exhibit the clearest scaling behavior in molecular size and self-entanglements [5].) Figure 3 shows the change in $[\overline{N}(n_k, n)]_n$ for a number of proteins, chosen according to the compactness criterion used in Ref. [5]. Roughly the same scaling is observed, except in very short segments. Since these compact chains span the typical lengths $(200 < n < 600)$, the occurrence of local scaling in $\overline{N}$ appears to be a feature independent of protein length and primary sequence.

In summary, the results in Figs. 2 and 3 point toward the following conjecture: from the viewpoint of self-entanglements, a (nontrivially short) section of length $n_k$ of a compact $n$-residue backbone in its native state resembles itself the ''average'' native state of another compact protein with a ''full'' backbone of $n_k$ residues.

TABLE I. Local scaling exponents ($\gamma$) for the mean number of overcrossings $\overline{N}(n_k,n)$, for selected proteins with maximally compact $\alpha$-carbon backbones. [The errors are indicated at 95% confidence level. The correlation coefficient $C$ and the number of $n_k$ values in the correlation are given in the last column.]

| PDB code ($n$) | $\gamma$ | $C$ {max($k$)} |
|---|---|---|
| 1SGT (223) | $1.64\pm0.11$ | 0.9947 {14} |
| 3TEC (279) | $1.54\pm0.05$ | 0.9982 {19} |
| 3CPA (307) | $1.50\pm0.07$ | 0.9979 {12} |
| 1ALD (363) | $1.60\pm0.02$ | 0.9999 {10} |
| 7ENL (436) | $1.44\pm0.12$ | 0.9932 {12} |
| 1GLY (470) | $1.43\pm0.05$ | 0.9995 {8} |
| 1COX (502) | $1.59\pm0.05$ | 0.9993 {9} |
| 1GAL (581) | $1.57\pm0.05$ | 0.9989 {11} |

Nevertheless, the resemblance between an $n_k$-residue segment and a full $n_k$-residue native protein is only qualitative. For a quantitative comparison, Table I shows the estimated local scaling exponents $\gamma$ [Eq. (2)] for a number of compact proteins, including those in Fig. 3. In all cases, the segments have been defined starting from the first available residue, but the length of the segment has been varied according to the protein length. [The results do not change significantly if one takes another amino acid as a starting point. It must be noted that Eq. (2) is the only form of self-similarity we have found. Other sequences of ''partial'' backbones were tested. For instance, we considered the backbones obtained by skipping every $n_1/s$ residues, where $s=1,2,4,8,\ldots$ . This sequence, resembling the standard test for fractality, exhibits no scaling in $\overline{N}$.]

Table I shows that a power-law behavior is valid for all the proteins tested. Within the precision of the calculations, we estimate an average local scaling exponent for compact proteins:

$$\gamma=1.54\pm0.08. \tag{4}$$

This exponent represents a mean over the entire set; our results do not rule out a small dependence of $\gamma$ with protein sequence.

The difference between local and global exponents appears to be significant. The scaling law of $\overline{N}(n,n)$ for the *complete* backbones of the selected group of compact proteins is consistent with the results for all other proteins. (For the proteins in Table I, a regression $\ln[\overline{N}(n,n)]-\ln(n)$ gives $\beta=1.32\pm0.10$, $C=0.9969$, in agreement with Eq. (1).) These results indicate that (a) the global scaling exponent $\beta$ for the subset of compact proteins is close to its ''effective'' lower bound within the set of native states ($\beta\approx4/3$), (b) the local scaling exponent $\gamma$ for compact proteins is closer to the ''effective'' upper bound to the scaling exponent for native states ($\beta\approx3/2$).

These results can be interpreted as follows. In terms of self-entanglements, long segments of a compact backbone appear to behave as proteins in *noncompact* native states. This conclusion is consistent with our observations for the molecular size. As we consider shorter sections of an $n$-residue backbone, we observed that the mean overcrossing number decreases whereas the molecular size is constant. Therefore, these segments should resemble less compact proteins (which are characterized by $\beta\approx3/2$). The present local scaling law [Eq. (2)] provides a precise quantitative expression for the shape of these segments.

The scaling laws in mean overcrossing numbers can be valuable towards the elucidation of protein structure from the primary sequence [19–22]. Current procedures use homology modeling to predict structural content. However, for new sequences with little similarity to others in a database, success in predicting secondary structure rarely surpasses 70% [20–22]. Our results provide an additional criterion to test the reliability of a model folding pattern. We propose here an improved algorithm that would work as follows: (a) After an initial guess for the tertiary structure, one computes the $[\overline{N}(n_k,n)]_n$ values for a sequence of $\{n_k\}$ backbone segments. (b) If the results in (a) show no power-law scaling or a local scaling exponent far from Eq. (4), the proposed 3D structure could be rejected in a first approximation. (c) New tentative structures could be tried by changing the packing of secondary structural elements. An acceptable 3D structure should show scaling in the ''local'' mean overcrossing numbers, $[\overline{N}(n_k,n)]_n$, while maintaining the ''local'' radius of gyration, $[R_G(n_k,n)]_n$, virtually constant over a range of segment lengths. (d) If no structure is found to satisfy the criterion (c), then one should revise the secondary structure content and begin the test again from (a).

We believe that this method could also be an additional tool to improve the algorithms for designing sequences with desired structural features (*de novo* or ''inverse'' folding [19,23]).

Finally, the present results show that the simultaneous analysis of *distinct* shape descriptors (e.g., molecular size and chain self-entanglements) can lead to valuable insights into polymer configurations. Previously, we have shown that a comparative study of size and entanglements allows one to assess the structural and dynamic stability of linear polymer with variable composition and at various temperatures [10,11]. In these cases, we showed the conditions for the persistence over time of certain global folding features. The present work could allow one to extend these notions to the conservation of local folding patterns, e.g., secondary structure. To this purpose, we are currently testing the conditions under which the present local scaling in mean overcrossing numbers is also also found in the average configuration of linear polymer models.

I thank Denise Caughill for her collaboration in the computations and Naomi Grant for her comments on the manuscript. This work was supported by FRUL and by Grants from NSERC (Canada).

[1] C. Chothia, Nature (London) **357**, 543 (1993).

[2] C. A. Orengo, D. T. Jones, and J. M. Thornton, Nature (London) **372**, 631 (1994).

[3] K. Mizuguchi and N. Go, Curr. Opin. Struct. Biol. **5**, 377 (1995).

[4] T. G. Dewey, J. Chem. Phys. **98**, 2250 (1993).

[5] G. A. Arteca, Phys. Rev. E **51**, 2600 (1995).

[6] G. A. Arteca, Phys. Rev. E **54**, 3044 (1996).

[7] G. A. Arteca and P. G. Mezey, Biopolymers **32**, 1609 (1992).

[8] E. J. Janse van Rensburg, D. W. Sumners, E. Wasserman, and S. G. Whittington, J. Phys. A **25**, 6557 (1992).

[9] G. A. Arteca, Biopolymers **33**, 1829 (1993).

[10] G. A. Arteca, Macromolecules **29**, 7594 (1996).

[11] G. A. Arteca, Biopolymers **39**, 671 (1996).

[12] A. Stasiak, V. Katritch, J. Bednar, D. Michoud, and J. Dubochet, Nature (London) **384**, 122 (1996).

[13] V. Katritch, J. Bednar, D. Michoud, R. G. Scharein, J. Dubochet, and A. Stasiak, Nature (London) **384**, 142 (1996).

[14] A. Yu. Grosberg, A. Feigel, and Y. Rabin, Phys. Rev. E **54**, 6618 (1996).

[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] G. A. Arteca, Phys. Rev. E **49**, 2417 (1994).

[17] A. L. Kholodenko and D. P. Rolfsen, J. Phys. A **29**, 5677 (1996).

[18] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rogers, O. Kennard, T. Shimanouchi, and M. Tasumi, J. Mol. Biol. **112**, 535 (1977).

[19] J. S. Richardson and D. C. Richardson, in *Proteins: Form and Function*, edited by R. A. Bradshaw and M. Purton (Elsevier, Cambridge, 1990), pp. 173–182.

[20] B. Rost, C. Sander, and R. Schneider, J. Mol. Biol. **235**, 13 (1994).

[21] L. Holm and C. Sander, Proteins **19**, 165 (1994).

[22] F. Eisenhaber, B. Persson, and P. Argos, Crit. Rev. Biochem. Mol. Biol. **30**, 1 (1995).

[23] R. Leplae, A. Lahm, and A. Tramontano, Biopolymers (Peptide Sci.) **37**, 377 (1995).