Characterization of Fold Diversity among Proteins with the Same Number of Amino Acid Residues

Gustavo A. Arteca[†,‡] and O. Tapia*[,‡]

Département de chimie et biochimie, Laurentian University, Sudbury, Ontario, Canada P3E 2C6, and Department of Physical Chemistry, Uppsala University, Box 532, S-751 21 Uppsala, Sweden

Received February 19, 1998

Chain entanglements provide a simple and global measure of folding in a macromolecule. The complexity of these entanglements can be expressed by the pattern of projected bond−bond crossings, or "overcrossings", associated with the molecular backbone. In this work, we use this approach to characterize quantitatively the range of tertiary folds observed in proteins with a given chain length. To discriminate among folding features, we use two shape descriptors derived from the probability distribution of overcrossings: the *mean overcrossing number,* $\bar{N}$, and the *most probable overcrossing number, N**. The values of $\bar{N}$ and *N** relate to the content of secondary structure in a protein as well as its global three-dimensional organization. We propose a measure of folding diversity based on the properties of these descriptors. In addition, we discuss the application of our method to study how tertiary folds evolve during protein dynamics.

INTRODUCTION

Native protein conformations are characterized by a specific three-dimensional (3D) organization of secondary structural elements within one or several folding units.[1] To some extent, this tertiary structure (or "tertiary fold") is preserved with priority over chemical composition, since the number of structural motifs appears to be limited despite variations in primary sequence.[2] It is indeed possible that protein sequences are naturally selected on the basis of how rapidly they achieve specific structural features.[3] Accordingly, characterizing the 3D architecture of a protein[4] and establishing its spatial homology with other known structures[2e,5] remains a central issue in molecular biology. In this work, we present a new viewpoint for classifying and comparing the *folding patterns accessible* to different proteins with the same polymer length. Our main objective is to characterize the large-scale molecular shape in native protein conformations. To this end, we propose a global measure of *absolute* folding features. We shall not be concerned here with relative differences between akin proteins or with comparisons based on sequence homology or evolutionary relationships.

Our approach to assessing the range of accessible folding patterns is also motivated by progress in computer-simulated protein folding and unfolding transitions. Recent molecular dynamics studies of folding/unfolding in vacuo[6] and solution[7] provide insights into the properties of intermediate conformers inferred experimentally.[8] Establishing whether some structural features prevail along all possible pathways is an essential piece of information toward a general theory for the mechanism leading to the native structure.[9] In this context, our goal is to develop a tool to survey the space of molecular shapes traversed by transient configurations during folding dynamics. These transients will have the same chain length and composition but not necessarily a quasi-native structure. In order to understand the nature of these "favored" molecular shapes, we must estimate the range of *stable folding features,* i.e., the shapes for the native states of other proteins with the same chain length. To this end, it is important to have a measure of folding diversity for proteins of variable composition but constant number of amino acids. This is the task we address below.

There are several approaches to characterize the shape of a protein.[10] Standard descriptors of size and anisometry are limited in the sense that they use only the molecular geometry. A more convenient approach, which we adopt here, is to monitor the polymer *entanglements*.[11] This technique makes explicit use of the primary sequence connectivity, in addition to the molecular geometry, leading to more discriminating shape descriptors.

Topological (or "permanent") entanglements can take place when macromolecular chains form knots or links. However, since bonds cannot move through each other, an open chain can also behave as self-entangled during a brief period of time. The complexity of these *transient* self-entanglements can be measured in terms of the distribution of projected bond−bond crossings.[11a] This distribution has been used to assess structural homologies between proteins[11a] as well as the extent of some domain motions.[11c] Presently, we employ two descriptors of overcrossings, the *mean overcrossing number* and the *most probable overcrossing number*, to characterize protein folds and assess their diversity. Even though these two descriptors behave in a similar fashion in random chains,[11b] they exhibit some interesting differences in highly organized, nonrandom structures such as protein native states. In the following sections, we discuss the interrelation between these two descriptors for various families of equal-length proteins. We close by sketching how our procedure can be used to monitor the evolution of shape features during computer-simulated unfolding transitions.

OVERCROSSING SPECTRA AND PROTEIN TERTIARY STRUCTURE

In a 3D molecular chain, bonds do not cross each other. However, if a bond lies over another along a "viewing"

* Corresponding author. Phone: +46-18-471-3659. Fax: +46-18-508-542. E-mail: orlando.tapia@fki.uu.se).
† Laurentian University.
‡ Uppsala University.

10.1021/ci990323i CCC: $18.00
Published on Web 05/21/1999

direction, it will produce a *projected* crossing. The term *overcrossing* refers to these projected crossings. (They are also known as "double points" in knot theory, where they are used in computing topological invariants.[12] Different projections will yield different numbers of overcrossings, and the pattern of overcrossings in space can be used as a shape descriptor. In practice, one can use the distribution $A_N(n)$, which measures the probability of observing N overcrossings in a 2D projection of an n-monomer bond network. Algorithms for the computation of $A_N(n)$ are discussed in the literature.[11a] Below, we restrict ourselves to protein backbones, where a residue is represented only by its α-carbon.[1c]

The entire $\{A_N(n)\}$ distribution can be useful for some structural analyses.[11,13] Its main features are contained in two descriptors associated with moments of the distribution: the *mean overcrossing number*, $\bar{N}$, and the *most probable overcrossing number*, N^*. The $\bar{N}$ value can be computed directly using line integrals:[14]

$$\bar{N} = \frac{1}{2\pi}\sum_{i=1}^{n-3}\sum_{j=i+2}^{n-1}\int_0^1\int_0^1 \frac{|(\dot{\gamma}_i(s)\times\dot{\gamma}_j(t))\cdot(\gamma_i(s)-\gamma_j(t))|}{||\gamma_i(s)-\gamma_j(t)||^3}\,ds\,dt \quad (1)$$

where γ_i is the line segment connecting the i and $i+1$ α-carbons, and $\dot{\gamma}_i$ the corresponding parametric derivative along the oriented backbone. The evaluation of N^* must be done numerically.[11a] Both $\bar{N}$ and N^* give information on the extent of the polymer self-entanglements. As a chain stretches toward a rod, $N^*, \bar{N} \to 0$. In contrast, $\bar{N}$ and N^* increase as the chain entangles with an increase in secondary structure.[11b] (Helical motifs produce higher overcrossing numbers than β-sheets.[15]) However, $\bar{N}$ and N^* depend *differently* on how secondary structural elements are organized relative to each other in space. As we show below, the plane (N^*, $\bar{N}$) yields a discriminating map of tertiary structures based on the concept of overcrossings. In principle, the values of (N^*, $\bar{N}$) are not strongly correlated with molecular size or anisometry. However, large molecular sizes in a fixed-length polymer can only be achieved by increasing its anisometry and decreasing its entanglements.[16]

Figure 1 shows the overcrossing spectra for two distinct proteins with $n = 340$. The results represent behaviors commonly found among native states. One example is ribonucleotide reductase (Brookhaven PDB code lav8), with a highly skewed distribution characterized by $N^* \ll \bar{N}$. The second case is one chain of porin (PDB code lgfp), with a symmetrical distribution, characterized by $N^* \approx \bar{N}$. For clarity, we have added some representative projections associated with low and high overcrossing numbers (denoted by "I" and "II", respectively). These snapshots are only indicative; there can be other projections that yield the same number of overcrossings.

Using the snapshots in Figure 1, we can interpret the different A_N distributions. In the first case, protein lav8 resembles an α-helical bundle, comprising elongated helices packed in quasi- parallel fashion. Here, a large fraction of the projections (e.g., lav8(I)) are nearly perpendicular to some helical axis, producing a relatively small number of over-

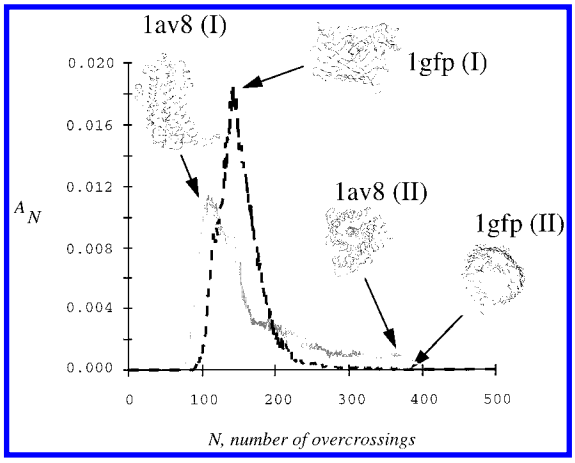

Figure 1. Overcrossing spectra for two representative (single chain) proteins with $n = 340$. Ribonucleotide reductase (lav8) resembles an α-helical bundle and is characterized by $N^* \ll \bar{N}$. Porin (lgfp) is a compact, fairly symmetrical structure characterized by $N^* \approx \bar{N}$. Representative projections associated with low and high overcrossing numbers are indicated by "I" and "II", respectively.

crossings.[9] Very high N values can be obtained when projecting *along* the helical axes (lav8(II)), but their probability is minimal. In other words, the condition $N^* \ll \bar{N}$ is mainly associated with a nearly parallel packing of secondary structural elements. (If the packed elements were β-strands, the A_N distribution would be similar, but shifted toward smaller N values.)

The second protein in Figure 1 provides an interesting contrast. The porin lgfp chain is a compact, fairly symmetrical arrangement of nonparallel β-strands. Most projections lead to similar overcrossing numbers (e.g., lgfp(I)). High N values are still possible (e.g., lgfp(II)), but they span a small fraction of all projections. The resulting $N^* \approx \bar{N}$ condition is therefore associated with a uniform spatial distribution of chain segments. When the chain is compact (or has a large content of secondary structure), we expect $N^* \approx \bar{N}$, with large $\bar{N}$ values. If the chain is relatively noncompact, then we will still have $N^* \approx \bar{N}$, but with smaller $\bar{N}$ values.

Whereas the overcrossing spectra satisfying $N^* \approx \bar{N}$ and $N^* \ll \bar{N}$ are the most frequent ones among proteins, other remarkable overcrossing patterns are also possible. Figure 2 shows the case of transducin (PDB code ltbg), also with $n = 340$ amino acids. In this case, we find a distribution of overcrossings with *two* principal peaks. The distribution is skewed toward the smaller N values, leading to $N^* > \bar{N}$. This situation is found in a small subset of structures: compact proteins with high content of secondary structure, where the helices and/or strands are not all packed parallel to each other.

Figures 3 and 4 provide further examples of these representative behaviors for proteins with $n = 316$ residues. In Figure 3, we find the case of annexin V, characterized by two peaks in the overcrossing spectrum and the condition $N^* > \bar{N}$. As in Figure 2, this protein exhibits two dominant types of projections. Projections leading to the low-N peak correspond to directions along a molecular cavity. The high-N peak is related to projections along directions perpendicular to several helical axes. Note that the helices are packed in

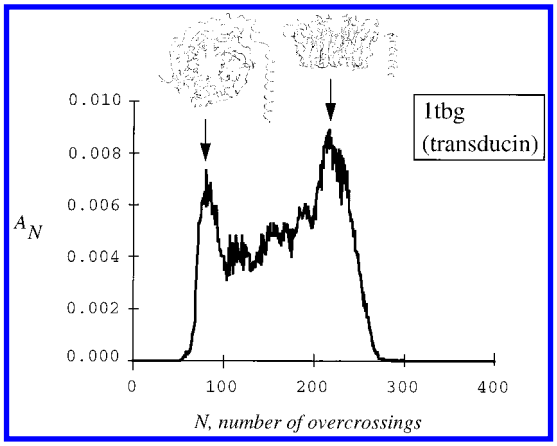

Figure 2. Overcrossing spectrum for transducin (PDB code ltbg, $n = 340$). This protein belongs to the rare class where the mean overcrossing number $\overline{N}$ is *smaller* than the most probable overcrossing number N^*. The figure includes representative projections associated with each of the two main peaks in the overcrossing distribution.

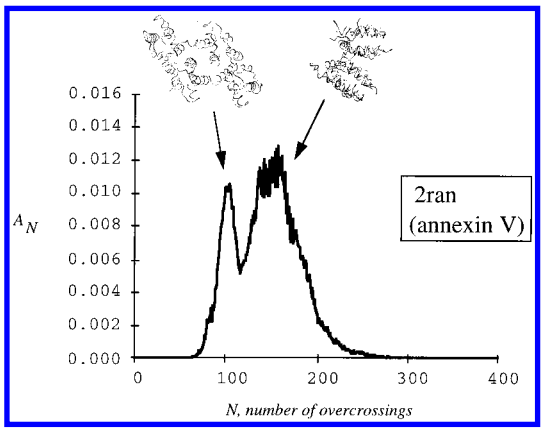

Figure 3. Overcrossing spectrum for annexin V (PDB code 2ran, $n = 316$). This protein contains α-helices packed in nonparallel fashion and is characterized by $N^* > \overline{N}$. Representative projections associated with each of the two main peaks in the overcrossing distribution are included.

distinct groups: some are parallel to each other, some are at significant angles from each other. This pattern translates into the $N^* > \overline{N}$ relation between shape descriptors. Finally, Figure 4 shows two examples of more common overcrossing distributions. The protein ltde has two well-separated distinct domains. Consequently, it produces mostly low N values along the directions where the two domains are clearly "visible." The resulting distribution is characterized by $N^* \ll \overline{N}$. In contrast, the compact protein lonr exhibits no significant separation of domains, and it leads to a symmetrical distribution with $N^* \approx \overline{N}$.

In conclusion, variations in folding pattern can give rise to a diverse range of N^* and $\overline{N}$ values, even in proteins with the same number of residues n. In the next section, we discuss the general properties of the (N^*, $\overline{N}$)-maps.

DIVERSITY OF FOLDING FEATURES IN FAMILIES OF PROTEINS WITH THE SAME NUMBER OF RESIDUES

We have searched the Brookhaven PDB for n values that provide large sets of distinct proteins. Very low and very high n values are not possible since they produce few native folds as single chains. The optimum range for statistical significance was found to be $75 \leq n \leq 340$. Within this range, we have selected a number of protein families yielding at least six different n-residue chains. Only cases where *all* residues are clearly identified in the electron density have been considered. This condition eliminates a large fraction of the PDB entries, where residues with unassigned coordinates are common. We have also excluded repeated structures and entries for multiple refinements. The conditions to include a structure in our data set are stringent: typically, three quarters of all structures listed in the PDB as having an n-residue chain must be rejected. As a result of our selection, we have retained for analysis the sets of proteins corresponding to $n = 75$, 129, 175, 212, 238, 269, 312, and 340. For these structures, we have computed overcrossing numbers and generated the (N^*, $\overline{N}$)-maps of folding features. In order to increase precision, the N^* are $\overline{N}$ values which are computed by averaging over several different sets of randomized rigid projections.[11a] As a result, N^* appears as a real number instead of an integer.

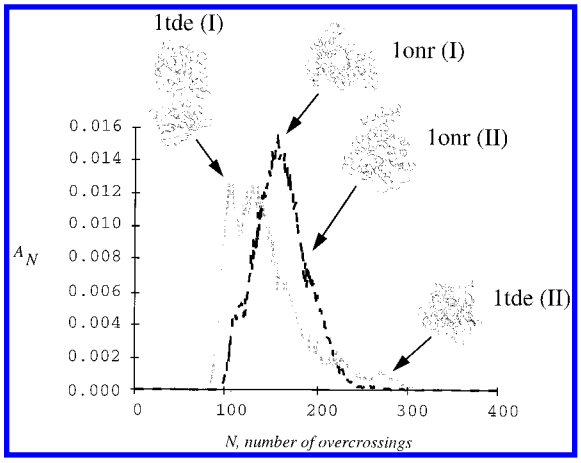

Figure 4. Overcrossing spectra for two representative (single chain) proteins with $n = 316$. Thioredoxin reductase (ltde) has well-separated domains and is characterized by $N^* \ll \overline{N}$. Transaldose B (lonr) is a compact protein, where we find $N^* \approx \overline{N}$. Representative projections associated with low and high overcrossing numbers are indicated by "I" and "II", respectively.

The maps corresponding to the two lowest n values appear in Figures 5 and 6. In Figure 5, we include a snapshot of the tertiary fold for every structure in the data set. In Figure 6, the number of available structures is much larger, and we have only indicated the position of some representative folds within the (N^*, $\overline{N}$)-map. As Figures 5 and 6 illustrate, different folds can span a significant range of (N^*, $\overline{N}$) values. An analysis of these figures reveals some simple trends for the location of folding features in an (N^*, $\overline{N}$)-map:

(i) High α-helical content and compact packing are associated with large $\overline{N}$ values. This is clearly seen in Figure 6, where CheY protein, lysozymes, interleukin 4, and cytochrome C′ have comparable $\overline{N}$ values. However, the different spatial organization of the helices in these chains is reflected in the N^* values. Note that parallel packing of helices produces low N^* (e.g., cytochrome C′), whereas nonparallel packing leads to larger N^* (e.g., CheY protein).

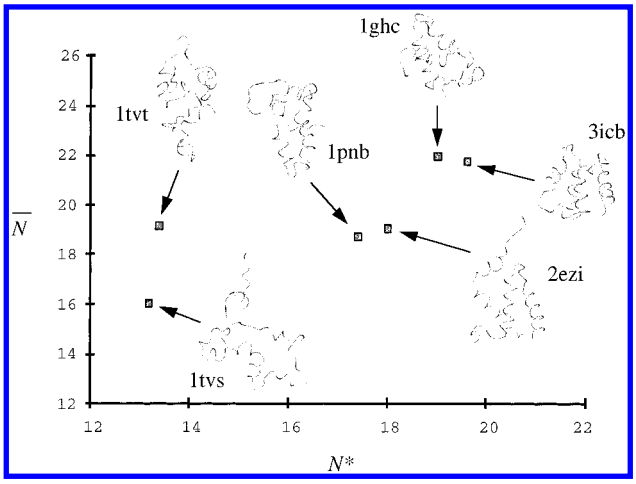

Figure 5. Diversity of tertiary folds for all selected proteins with $n = 75$. A representative view is given for each of the displayed proteins.

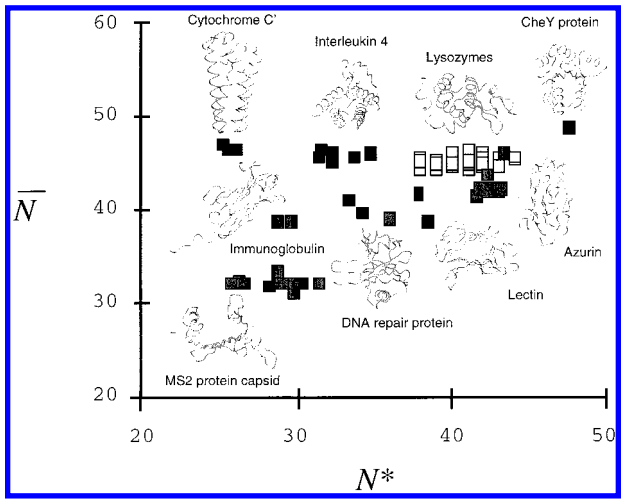

Figure 6. Diversity of tertiary folds for all, selected proteins with $n = 129$. There are 145 proteins in this set, 105 of which are mutations of hen egg-white lysozyme. The approximate location of representive tertiary folds is given.

(ii) A noncompact arrangement of α-helices yields smaller $\overline{N}$ and N^* values than case i. This situation can be seen when comparing the calcium binding protein (3icb) and the DNA binding domain of bacteriophage transposase (2ezi) in Figure 5.

(iii) Since a β-strand overcrosses itself much less than any helix,[15] compact structures with high β-strand content lead to smaller $\overline{N}$ and N^* values than compact structures with high helical content. This is clearly seen in Figure 6 if one compares azurin, lectin, and immunoglobulin to the proteins in case i. The former proteins have comparable $\overline{N}$ values, yet they can be distinguished by their position along the N*-axis. Once again, N^* diminishes as the secondary structural elements are packed less compactly (cf. azurin and immunoglobulin).

(iv) Noncompact proteins with little secondary structure lead to the smallest values for both $\overline{N}$ and N^*. Commonly, this situation is found in viral proteins. Examples are found in the horse anemia virus TAT (1tvs in Figure 5) and the MS2 viral capsid (in Figure 6).

(v) Variations in sequence and small configurational changes are reflected more strongly in N^* than in $\overline{N}$, as

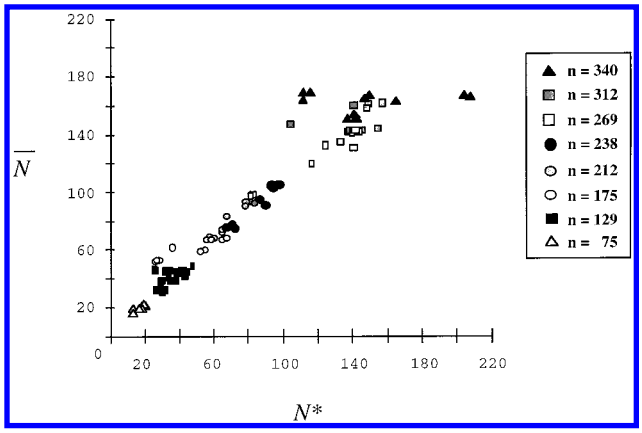

Figure 7. Relation between mean overcrossing number $\overline{N}$ and most probable overcrossing number N^* for a eight families of proteins with the same number of amino acid residues. This diagram contains those in Figures 5 and 6. The unsual variation of N^* values associated with $n = 340$ may be due to the presence of both multidomain and compact single domain proteins.

Table 1. Estimation of the Diversity of Protein Folding Features Encountered in the Protein Sets with a Constant Number of Residues

no. of residues	no of proteins	ΔN^*	$\Delta\overline{N}$	$(\Delta\overline{N})_{expected}$	diagnosis of diversity
75	6	6.4	5.93	8 ± 2	medium
129	145	22.4	17.78	17 ± 4	large
175	11	40.4	19.41	25 ± 7	large
212	16	27.6	27.11	33 ± 9	medium
238	10	30.6	29.49	39 ± 11	medium
269	11	61.6	45.74	46 ± 13	*large (fitted)*
316	12	53.2	19.06	58 ± 15	small
340	11	95.1	17.29	64 ± 17	small

illustrated by the group of lysozymes in Figure 6. This result is simply a consequence of the fact that N^* (an integer number) is always computed with a larger intrinsic error than the mean overcrossing number $\overline{N}$.

The analysis above indicates how variations in tertiary fold can be recognized in the shape diagram. The cases i−iv represent the widest range of situations that could be found among common proteins. We would expect a similar diversity to be accessible to proteins of all possible chain lengths. One can gauge the actual degree of such diversity by assessing how significant is the range of $\overline{N}$ and N^* values observed in a given family of proteins with n residues. To this end, one can estimate the maximum *expected* variations in $\overline{N}$ and N^* as a function of n (indicated as $\Delta\overline{N}$ and ΔN^*, respectively). As commented in point v above, $\Delta\overline{N}$ should provide a better analysis tool. This descriptor has smaller intrinsic numerical error and thus will reflect more accurately the variations in folding features.

Figure 7 collects the results for all selected sets of proteins. In addition, Table 1 lists the *observed* $\Delta\overline{N}$ and ΔN^* variations within these protein families. From the (N^*, $\overline{N}$)-map in Figure 7, we make the following observations: (a) On average, N^* and $\overline{N}$ are linearly correlated as the number of residues increase. (b) The fluctuations $\Delta\overline{N}$ and ΔN^* about the mean can be large depending on n. Note that Figure 7 indicates only that N^* and $\overline{N}$ increase equally fast with n, not that they increase linearly with n. As discussed

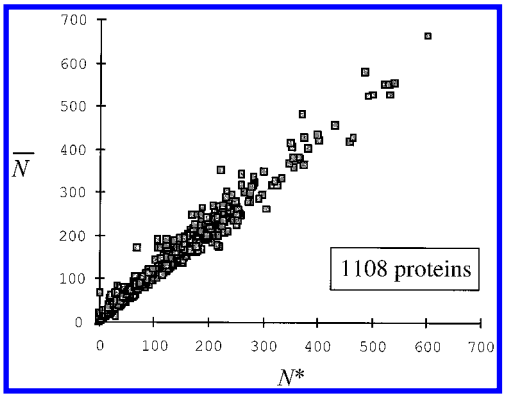

Figure 8. Relation between mean overcrossing number $\overline{N}$ and most probable overcrossing number N^* for a large set of proteins with variable length. The mean behavior is represented by $\overline{N} \approx 1.05N^*$ (see text).

elsewhere, we expect both N^* and $\overline{N}$ to increase *faster* than linearly with the number of residues.[11b]

In order to assess the magnitude of the $\Delta\overline{N}$ and ΔN^* fluctuations, we must establish the *mean* behavior of $\overline{N}$ and N^* in generic native states. To this end, we have computed these descriptors within a very large set of protein chains. Our ensemble is an expanded version of the one used in refs 11d and 11e, which follows the "natural" population of chain lengths for native folds.[17] This set excludes repeated structures, does not oversample any particular n value, and is not biased toward any special structural feature. The results for 1108 distinct proteins are shown in Figure 8. A least-square fitting over the points in this figure gives

$$\overline{N} \approx (1.05 \pm 0.01)N^* + (4.7 \pm 1.2) \qquad (2)$$

with correlation coefficient $C = 0.989$. (Error bars are 95% confidence intervals.) The result is similar if the intercept is set to zero: $\overline{N} \approx (1.08 \pm 0.01)N^*$, with $C = 0.988$. (Interestingly, a similar scaling is obtained in very simple polymer models, as discussed in Appendix A.)

We can now use eq 2 together with the known dependence of $\overline{N}$ on n[11e]

$$\overline{N} \approx (0.055 \pm 0.006)(n-3)^{1.37\pm0.02} \qquad (3)$$

to estimate the *mean* values $\overline{N}$ and N^* for any chain length. Equation 3 takes into account that a three-residue chain does not overcross. The effective scaling exponent $\beta \approx 1.37 \pm 0.02$ represents well the empirical behavior for medium-size proteins ($n < 700$). The asymptotic result for very large n appears to be $\beta = 1.2 \pm 0.1$.[11d]

We can now estimate the "expected" fluctuations in overcrossings for a family of n-residue proteins with maximum diversity of tertiary folds. Since the fluctuation $\Delta\overline{N}$ has less intrinsic numerical error, we shall use only this property for the rest of our analysis. We can now invoke the following key notion: *the* $\Delta\overline{N}$ *values arising from random conformational fluctuations satisfy also the scaling relation eq 3*.[11b] As a result, we can "calibrate" the estimated maximum folding diversity by fitting a relation $\Delta\overline{N} \sim a(n-3)^{\beta}$. In our case, we determine the constant "a" by using a reference family of proteins. From the results in Table 1, we will define that the family of $n = 269$ proteins has "large diversity". (Similar results are obtained by adopting another reasonable choice, e.g., the family of $n = 129$ proteins.) By fitting the result in Table 1 for $n = 269$, we obtain a general law for the expected range of fluctuations in the mean overcrossing number for a family of proteins with n residues:

$$(\Delta\overline{N})_{\text{expected}} = (0.022 + 0.006)(n-3)^{1.37\pm0.02} \qquad (4)$$

Using eq 4, one would expect $\Delta\overline{N} \approx 17 \pm 4$ for $n = 129$, in good agreement with the observed fluctuation ($\Delta\overline{N} \approx 18$). Thus, our assessment is that the available experimental structures for proteins with $n = 129$ residues contain already *the largest possible diversity of folding features*. In contrast, eq 4 yields $\Delta\overline{N} \approx 58$ for $n = 316$ proteins, a much larger value than the observed one (cf. Table 1). In this case, one would diagnose that the available PDB structures for this family of proteins represent only a small fraction of all possible folding features. All other families in Figure 7 can be analyzed in this fashion. The results and our assessment are summarized in the last two columns of Table 1. As discussed in the next section, this approach can now be used as a diagnostic tool for assessing the shapes of transient protein conformations.

A remark is in order regarding the interpretation of a given diagnosis. There are several reasons, besides small sample size, for diagnosing a family of proteins as having "small diversity" in folding features. On the one hand, there are intrinsically fewer sequences foldable to a single chain as n increases.[17] On the other hand, the diversity can also be reduced because longer proteins tend to fold by domains.[1,2,18] The available molecular shapes may then be restricted to those defined by the relative orientation of smaller (independently-folded) subunits. In this case, the features associated with single-domain global tertiary folds would not be accessible. Establishing which of these two effects is more dominant may be useful for understanding the folding mechanism of large proteins. Insights into this issue require the analysis of a larger set of experimental 3D structures.

ALTERNATIVE MOLECULAR SHAPE DESCRIPTORS

A. Molecular Size and Anisometry. The molecular shape descriptors discussed in the previous sections combine information on the 3D geometry and primary sequence (or "topology") of a protein. We believe that there are two important characteristics that make the present approach useful: (a) Proteins with the same chain length but different content of secondary structure and 3D organization appear in distinct regions of an (N^*, $\overline{N}$)-map. (b) Proteins with different chain lengths but similar structural features have similar $\overline{N}/N^*$ ratios, but they are clearly separated in an (N^*, $\overline{N}$)-map by the absolute values of the shape descriptors. These two factors together yield an effective tool to discriminate among folding features. This analysis may be difficult when using *other* descriptors of macromolecular shape. For example, let us suppose that we monitor backbone size and anisometry for the proteins in Figure 7. The former property is described by the standard radius of gyration, R_g, and the latter by the asphericity, Ω,[19]

$$\Omega = \frac{(I_2 - I_1)^2 + (I_3 - I_2)^2 + (I_3 - I_1)^2}{2(I_3 + I_2 + I_1)^2} \qquad (5)$$

given in terms of the principal moments of inertia $\{I_i\}$. The same structures in Figure 7 are replotted in Figure 9 in the form of an (R_g, Ω)-map. As Figure 9 shows, this map shows no clear scaling with polymer length. Moreover, different folds can have the same R_g and Ω values. As a result, tertiary folds are not properly discriminated when using molecular size and anisometry.

B. Difference Overcrossing Spectra. The example above illustrates the advantage of using entanglement descriptors. Further refinements for comparing folding features or assessing their diversity can still be implemented. For example, we can use the difference of overcrossing spectra as a *relative* measure of entanglement between two tertiary folds. A simple descriptor would be a *root-mean-square deviation* of overcrossing probabilities, $\sigma(K, K')$, defined as

$$\sigma(K, K') = [\sum_{N=0}^{N_{max}} (A_N^{(K)} - A_N^{(K')})^2]^{1/2} \qquad (6)$$

between two conformers K and K'. (These can correspond to two instantaneous configurations of the same protein, or to the backbones of different equal-length proteins.) Whereas it is possible that different tertiary folds can yield similar $\overline{N}$ values, they will always be distinguished by the shape of their *complete* overcrossing spectra. Low $\sigma(K, K')$ values can only correspond to nearly identical folds. Note that $\sigma(K, K')$ is computed in a much simpler fashion than the standard root-mean-square deviation coordinates. In the latter case, one must perform an optimum alignment and minimize over all possible deviations. By contrast, overcrossing numbers are an intrinsic single-molecule property, independent of molecular orientation. As a result, no alignment or minimization would be needed to compute $\sigma(K, K')$.

FOLDING FEATURES DRAWN FROM MOLECULAR DYNAMICS SIMULATIONS

We discuss here an application of our technique in monitoring the evolution of tertiary folds along molecular dynamics (MD) trajectories. As an illustrative example, we have chosen the results of a recent computer simulation of in vacuo unfolding of hen egg-white lysozyme.[6a] In this simulation, unfolding into elongated conformers is caused by weak (but systematic) centrifugal forces. The conformational transition appears to take place in a series of sharp steps, a finding compatible with experimental results for this protein.[8a] The MD trajectory was generated with GROMOS87,[20] using the neutral protein model with polar hydrogens and electrostatic hydrogen bonding.[21] The starting point for the simulation was the crystal structure of disulfide-intact lysozyme, PDB code lhel. When using in vacuo boundary conditions, the system can rotate freely after equilibration. In practice, unfolding was kept via a strong coupling to a Berendsen thermostat at $T = 293$ K.[22] Under a weak coupling, lysozyme does not unfold but resembles the crystal structure. We refer to the latter as the "in vacuo native structure". For further details on the simulation, see ref 6a.

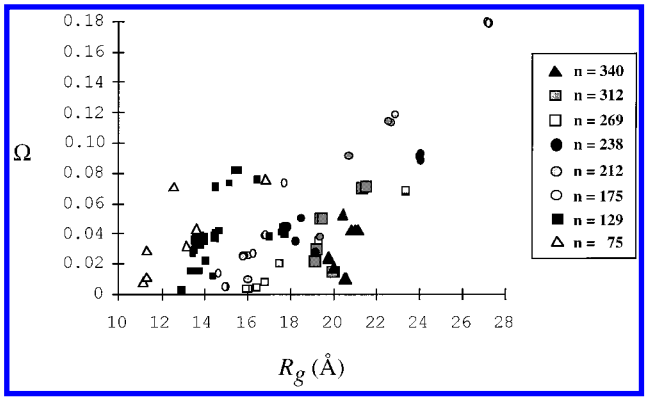

Figure 9. Molecular size and anisometry map for the families of proteins in Figure 7. The lack of correlation between the radius of gyration R_g, the asphericity Ω, and the number of monomers n shows that these descriptors alone cannot classify the accessible folding features.

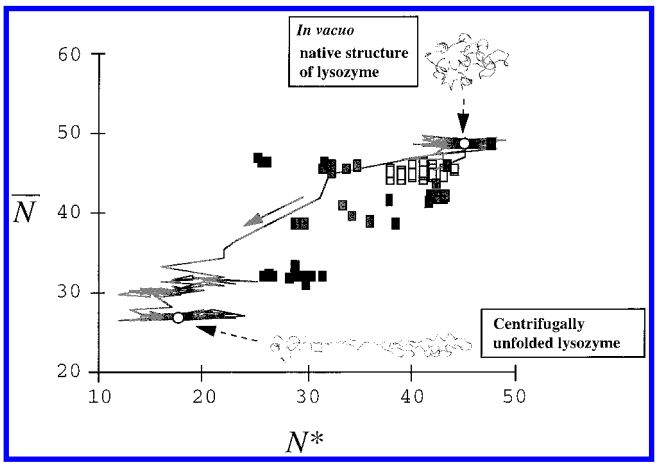

Figure 10. Evolution of entanglement features during an unfolding simulation of in vacuo lysozyme. The accessible native folds for $n = 129$ proteins (cf. Figure 6) are superimposed as a reference. The snapshots indicated correspond to the equilibrium in vacuo lysozyme structure prior to unfolding (top right) and a fully unfolded conformation (bottom left). The arrow indicates the direction of time evolution along the trajectory. Note the occurrence of shape transitions where $\overline{N}$ decreases at constant N^* values. None of the structures found during these transitions resembles the native folds of proteins with the same chain length (see text).

Using our present approach, we can contrast the folding features observed along the MD trajectory with those of all proteins having $n = 129$ residues. To this end, we have computed the entanglement descriptors for conformational transients as 5 ps snapshots along a l.1 ns trajectory. The results appear in Figure 10, contrasted with those for the proteins in Figure 7. The direction of time evolution is indicated by an arrow. We have also included the representative tertiary folds for the in vacuo native structure and a fully unfolded transient. From Figure 10, we can make a number of key observations on the molecular shape features accessible to unfolding transients of lysozyme:

(i) The in vacuo native structure is slightly shifted in the (N^*, $\overline{N}$)-map with respect to the lysozyme crystal structures (gray squares). Nevertheless, the folding features of the crystal structure are preserved, since the shift conserves the $\overline{N}/N^*$ ratio. This effect is simply due to compactification caused by the in vacuo boundary conditions.[6a] This stage lasts ca. 550 ps along the MD trajectory.

(ii) During the initial unfolding (between 550 and 570 ps), there is a rapid decrease in N^* and a much smaller change in $\overline{N}$. Qualitatively, the transients move along a line connecting the shape of the lysozymes with those of interleukin 4. This is consistent with conformational changes where the secondary structure is mostly unaffected. Note that the transient molecular shapes never resemble those of the organized packing in cytochrome C′ (cf. Figure 7).

(iii) Once full unfolding take place (between 600 and 900 ps), the conformational transients lose compactness and secondary structure, as indicated by a decrease in both $\overline{N}$ and N^*. The final structures reach $\overline{N}$ values comparable to those of noncompact proteins of the same length (e.g., viral capsids), yet they have even smaller N^* values. Therefore, these transients correspond to structures *without equivalent* among the $n = 129$ proteins, as illustrated by the snapshot in Figure 10 for the fully unfolded lysozyme.

(iv) Configurational transitions within the set of unfolded structures can still be recognized in the (N^*, $\overline{N}$)-map. At long times (over 900 ps of simulation), we observe a sharp transition where the N^* value is conserved but $\overline{N}$ decreases. This transformation can be associated with either (a) a global α-helix to β-strand conversion that conserves the molecular size or (b) a simple loss of compactness, i.e., the conformation opens further without changing much of its residual secondary structure. Case a would be unlikely as a fast unfolding process; comparison with the evolution of molecular size along the MD trajectory[6a] shows that the last transition observed in Figure 10 corresponds indeed to case b.

CONCLUSION

All the applications discussed in this work are based on the use of molecular shape descriptors that measure backbone entanglements. On the one hand, the approach allows one to classify and compare folding features in proteins with variable size. In addition, our analysis provides a tool to monitor the evolution of folding features during conformational rearrangements. The molecular shapes visited by the transient configurations can be interpreted in the context of those accessible to proteins of the same length. In this sense, our method can yield new insights into the mechanism for protein unfolding. Relaxation and refolding dynamics could be analyzed in a similar fashion. In this case, the present approach would be a useful tool for determining the molecular shape features "favored" by folding pathways and folding intermediates.

ACKNOWLEDGMENT

We thank I. Velázquez and C. Reimann for valuable discussions and for making available the unfolding MD trajectory and N. Grant for comments on the manuscript. G.A.A. acknowledges support from the Natural Sciences and Engineering Research Council (NSERC) of Canada, and O.T. is grateful to the Swedish Natural Science Council (NFR) for financial support.

APPENDIX A: OVERCROSSING DESCRIPTORS FOR RANDOM AND SELF-AVOIDING WALKS

We can put the average behavior obtained for proteins in the context of the results for random polymers. To this end, we consider two elementary models: the *random walks* (i.e., walks with no excluded volume) and the *self-avoiding walks* (i.e., walks "swollen" by excluded volume interaction between monomers). The simplest form of these models in the continuum is a necklace of n-beads, with constant bond length l and a variable radius of excluded volume, r_{ex}, around a bead. The random walks (or "freely jointed chains" in the polymer literature) correspond to the limit $r_{ex} \rightarrow 0$. The evaluation of molecular shape properties for these models is discussed in the literature.[11b,23] The descriptor $\overline{N}$ (or N^*) depends only on n and r_{ex}, and not on l. We can now reinterpret these results in terms of an (N^*, $\overline{N}$)-map. For random walks, the configurationally averaged descriptors follow the law

$$\langle\overline{N}\rangle \approx (1.10 \pm 0.01)\langle N^*\rangle + (3.3 \pm 0.1) \qquad \text{(A.1)}$$

with $C = 0.99996$ and 95% confidence intervals. Equation A.1 has been derived by fitting the results for a series of chain lengths comparable to those of our protein families: $n = 31$, 54, 70, 94, 129, 150, 198, 246, 296, and 390. (For each n value, the descriptors $\langle\overline{N}\rangle$ and $\langle N^*\rangle$ correspond to averages over 10^3 random configurations.) Despite the dramatic difference between the physical models, it is remarkable that the scaling (eq A.1) resembles that of the average native states for proteins (cf. eq 2). (Random walks are more compact and have overcrossing distributions with extended large-N tails, thus a slightly larger $\overline{N}/N^*$ ratio.) This result is a strong indication that the underlying mathematical structure of the overcrossing probability distribution may be determined by general properties common to all curves in space. Of course, only the $\overline{N}/N^*$ ratios are similar between random walks and protein native states. The actual values for the descriptors are very different. As a rule of thumb, we find that n-bead random walks produce roughly mean overcrossing numbers similar to those for the native states of longer proteins, with ca. $1.6n$ amino acid residues.

Consistent with the above observations, we have also found that the $\overline{N}/N^*$ ratio depends little on excluded volume. These analyses are difficult because long chains with large excluded volume are not computationally feasible. In the case of self-avoiding walks with $n = 129$ beads and five excluded volume radii, $r_{ex} = 0$, $0.13l$, $0.26l$, $0.39l$, and $0.53l$, we get

$$\langle\overline{N}\rangle \approx (1.08 \pm 0.04)\langle N^*\rangle + (4.5 \pm 2.1) \qquad \text{(A.2)}$$

with $C = 0.99981$ and 95% confidence intervals. Again, the coincidence of results suggests a general overcrossing probability distribution, common to the "averaged state" of both random and nonrandom polymers. In Appendix B, we discuss briefly some possible candidates for a model of the general $\{A_N\}$ distribution.

APPENDIX B: LINEAR RELATIONS BETWEEN THE $\overline{N}$ AND N^* DESCRIPTORS IN MODEL OVERCROSSING PROBABILITY DISTRIBUTIONS FOR RANDOM POLYMERS

When averaged over all conformations accessible to an n-monomer random chain, the $\{A_N\}$ distribution takes the form of a smooth single-peak function, skewed toward the

large N values. It resembles a χ^2 distribution, roughly similar to those of lgfp and lonr in Figures 1 and 4, respectively. The qualitative shape is not changed whether the chains have excluded volume or not. Using the results in Appendix A, we test here if the average distribution for general walks in the continuum, $(A_N)_{\text{walks}}$, is representable by model distributions of the form:

$$(A_N)_{\text{walks}} \approx N^x \, e^{-a\, f(N)}, \quad x > 0 \quad \text{and} \quad a > 0 \quad \text{(B.1)}$$

with some positive defined function $f(N)$. Normalization can be done analytically only in the cases of exponential and Gaussian functions, i.e., $f(N) = N$ and $f(N) = N^2$, respectively. The computation of the shape descriptors $\bar{N}$ and N^* in these two cases is straightforward. We comment here on the significance of the results.

The overcrossing distribution in random polymers does not seem to fall exponentially with N. Using eq B.1 with $f(N) = N$, we obtain $N^* = x/a$ for the most probable overcrossing number and $\bar{N} = (x + 1)/a$, for the mean overcrossing number. The resulting linear correlation

$$\bar{N} = N^* + a^{-1} \quad \text{(B.2)}$$

disagrees with the results from computer simulations which indicate a ratio $\bar{N}/N^* > 1$. When using a Gaussian function, the scaling between these descriptors is independent of a:

$$\bar{N} = \left(\frac{2}{x}\right)^{1/2} \frac{\Gamma\left(\frac{x+2}{2}\right)}{\Gamma\left(\frac{x+1}{2}\right)} N^* \quad \text{(B.3)}$$

where $\Gamma(y) = (y - 1)!$ is the Gamma function. Equation B.3 is consistent with the results in Appendix A, since it gives $\bar{N}/N^* > 1$ for $x > 0$. Equations A.1 and A.2 suggest $x \approx 4 \pm 1$. However, eq B.3 has non-zero intercept; therefore, a Gaussian-like model of an overcrossing probability distribution is applicable, at best, to very long chains. In addition, we note that *both* the exponential and the Gaussian models produce the relation $A^* = (N^*)^{-1}$, between the most probable number of overcrossings and its associated probability A^*. Results in the literature suggest instead $A^* = (N^*)^{-1/\beta}$, with a scaling exponent $\beta = 1.2 \pm 0.1$.[11b,d] This leads us to believe that the actual distribution for overcrossings in random polymers with arbitrary chain length cannot be represented by eithe $f(N) = N$ or $f(N) = N^2$ in eq B.1.[24]

REFERENCES AND NOTES

(1) (a) Levitt, M.; Chothia, C. *Nature* **1976**, *261*, *552*. (b) Janin, J.; Wodak, S. J. *Prog. Biophys. Mol. Biol.* **1983**, *42*, 21. (c) Brändén, C.; Tooze, J. *Introduction to Protein Structure;* Garland: New York, 1991.

(2) (a) Richardson, J. S. *Adv. Protein Chem.* **1981**, *34*, 167. (b) Richardson, J. S. *Methods Enzymol.* **1985**, *115*, 349. (c) Chothia, C. *Nature* **1993**, *357*, 543. (d) Blundell, T. L; Johnson, M. S. *Protein Eng.* **1993**, *2*, 877. (e) Murzin, A. G. *Curr. Opin. Struct. Biol.* **1998**, *8*, 380.

(3) Shakhnovich, E. I. *Phys. Rev. Lett.* **1994**, *72*, 3907.

(4) (a) Thornton, J. M. *Curr. Opin. Struct. Biol.* **1992**, *2*, 888. (b) Orengo, C. A. *Curr. Opin. Struct. Biol.* **1994**, *4*, 429.

(5) (a) Orengo, C. A.; Flores, T. P.; Taylor, W. R.; Thronton, J. M. *Protein Eng.* **1993**, *6*, 485. (b) Holm, L.; Sander, C. *J. Mol. Biol.* **1993**, *233*, 123. (c) Feng, Z.-K.; Sippl, M. J. *Fold. Design* **1996**, *1*, 123. (d) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. *J. Mol. Biol.* **1995**, *247*, 536. (e) Holm, L.; Sander, C. *Proteins* **1998**, *31*, 88. (f) Chou, K.-C.; Liu, W.-M.; Maggiora, G. M.; Zhang, C.-T. *Proteins* **1998**, *31*, 97.

(6) (a) Reimann, C. T.; Velázquez, I.; Tapia, O. *J. Phys. Chem. B* **1998**, *102*, 2277. (b) Reimann, C.T.; Velázquez, I.; Tapia, O. *J. Phys. Chem. B* **1998**, *102*, 9344. (c) Marchi, M.; Ballone, P. *J. Chem. Phys.* **1999**, *110*, *1*.

(7) (a) Lazaridis, T.; Karplus, M. *Science* **1997**, *278*, **1928**. (b) Daura, X.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. *J. Mol. Biol.* **1998**, *280*, 925. (c) Duan, Y.; Kollman, P. A. *Science* **1998**, *282*, 740.

(8) (a) Miranker, A.; Robinson, C. V.; Radford, S. E.; Aplin, R.T.; Dobson, C. M. *Science* **1993**, *262*, 896. (b) Shelimov, K. B.; Jarrold, M. F. *J. Am. Chem. Soc.* **1996**, *118*, 10313. (c) Shelimov, K. B.; Jarrold, M. F. *J. Am. Chem. Soc.* **1997**, *119*, 2987. (c) McLafferty, F. W.; Guan, Z.; Haupts, U.; Wood, T. D.; Kelleher, N. L. *J. Am. Chem. Soc.* **1998**, *120*, 4732. (d) Gross, D. S.; Schnier, P. D.; Rodriguez-Cruz, S. E.; Fagerquist, C. K.; Williams, E. K. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 3143. (e) Valentine, J. S.; Anderson, J. S.; Ellington, A. D.; Clemmer, D. E. *J. Phys. Chem. B* **1997**, *101*, 3891. (f) Reimann, C. T.; Sullivan, P. A.; Axelsson, J.; Quist, A. P.; Altman, S.; Roepstorff, P.; Velázquez, I.; Tapia. O. *J. Am. Chem. Soc.* **1998**, *120*, 7608.

(9) Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 2426.

(10) (a) Mezey, P. G. *Shape in Chemistry;* VCH Publishers, New York, 1993. (b) Artymiuk, P. J.; Grindley, H. M.; Poirette, A. R.; Rice, D. W.; Ujah, E. C.; Willett, P. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, *54*. (c) Arteca, G. A. In *Reviews in Computational Chemistry;* Lipkowitz, K., Boyd, D. B., Eds.; VCH Publishers: New York, 1996; Vol. 9.

(11) (a) Arteca, G. A. *Bipolymers* **1993**, *33*, 1829. (b) Arteca, G. A. *Phys. Rev. E* **1994**, *49*, 2417. (c) Arteca, G. A. *Bipolymers* **1996**, *39*, 671. (d) Arteca, G. A. *Phys. Rev. E* **1995**, *51*, 2600. (e) Arteca, G. A. *Phys. Rev. E* **1997**, *56*, 4516.

(12) Murasugi, K. *Knot Theory and Its Applications;* Birkhäuser: Boston, 1996.

(13) (a) Arteca, G. A.; Mezey, P. G. *Biopolymers* **1992**, *32*, 1609. (b) Arteca, G. A.; Nilsson, O.; Tapia, O. *J. Mol. Graph.* **1993**, *11*, 193.

(14) Arteca, G. A.; Caughill, D. I. *Can. J. Chem.* **1998**, *76*, 1402.

(15) Arteca, G. A. *Can. J. Chem.* **1995**, *73*, 241.

(16) (a) Arteca, G. A. *Biopolymers* **1995**, *35*, 393. (b) Arteca, G. A. *Macromolecules* ***1996***, *29*, 7594.

(17) White, S. H. *Annu. Rev. Biophys. Biomol. Struct.* **1994**, *23*, 407.

(18) (a) Garel, J. R. In *Protein Folding*; Creighton, T. E., Ed.; Freeman: New York, 1992. (b) Xu, D.; Nussinov, R. *Fold. Design* **1997**, *3*, 11.

(19) (a) Rudnick, J.; Gaspari, G. *J. Phys. A* 1986, *19*, L 191. (b) Rudnick, J.; Gaspari, G. *Science* **1987**, *237*, 384. (c) Diehl, H. W.; Eisenriegler, E. *J. Phys. A* **1989**, *22*, L 87. (d) Baumgärtner, A. *J. Chem. Phys.* **1993**, *98*, 7496.

(20) Van Gunsteren, W. F.; Berendsen, H. J. C. *Groningen Molecular Simulation (GROMOS) Library Manual;* Biomos, Groningen, 1987.

(21) Åqvist, J.; van Gunsteren, W. F.; Leijonmarck, M.; Tapia, O. *J. Mol. Biol.* **1985**, *83*, *461*.

(22) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.

(23) Arteca, G. A. *J. Phys. Chem. B* **1997**, *101*, 4097.

(24) Ribbon pictures were made with Molmol program (Karadi, R.; Billeter, M.; Wüthrich, K. *J. Mol. Graph.* **1996**, *14*, 51.).

CI990323I