A Measure of Folding Complexity for *D*-Dimensional Polymers

Gustavo A. Arteca*

Département de Chimie et Biochimie, Laurentian University, Ramsey Lake Road, Sudbury,
Ontario, Canada P3E 2C6

Received May 31, 2002

A measure of folding characterizes aspects of the instantaneous organization of a polymer chain in space. For three-dimensional polymers ($D = 3$), one such measure is the *mean overcrossing number*. An intuitively similar property, the *radial intersection number*, has been proposed as a tool to characterize "folding features" in two-dimensional polymers ($D = 2$). In this work, we show rigorously that these measures are indeed related and that they can be derived as particular cases within a single, unified formulation. The present approach provides an analytical expression for a measure of folding complexity that can be applied to generic *D*-dimensional polymers. In the case $D = 2$, we show results for models derived from experimental structures by using optimized multidimensional scaling transformations for data compression.

INTRODUCTION

Basic aspects of polymer shape can be conveyed by conformer size or anisometry.[1,2] A more detailed analysis is required, however, whenever one wants to distinguish among polymer *folding features*. This is particularly important in homology analyses of protein native states[3,4] or in studying conformer population shifts during protein folding-unfolding transitions.[5,6]

Folding complexity takes into account both chain geometry and connectivity. A common approach is to use measures of chain self-entanglements, inspired in the study of three-dimensional (3D) knotted chains and links.[7] One such parameter is the mean overcrossing (or "average crossing") number, $\overline{N}$.[8] (For clarity, we use here the notation $\overline{N}^{(3D)}$, to emphasize the dimensionality of the object.) This quantity is an *absolute descriptor of folding*, in the sense that it assigns a numerical value that is intrinsic to a given backbone and not relative to a reference fold. Conceptually simple, $\overline{N}^{(3D)}$ is the number of projected bond-bond crossings, averaged over all directions in space. This descriptor satisfies a number of known or conjectured *mean* properties when averaged over *accessible random configurations.*[8c,8d,9] In a rodlike conformation (i.e., those with no entanglements), $\overline{N}^{(3D)}$ reaches the minimum value of zero. If we consider a chain with a more convoluted fold, e.g. a protein, $\overline{N}^{(3D)}$ increases depending on secondary and tertiary structure. Figure 1 illustrates the typical differences that can be expected in $\overline{N}^{(3D)}$ for conformers with distinct folding features. The figure shows two configurations of the same "polymer," whose backbone is a 21-bead polygon with a contour length of 21.157 Å. These conformers have similar mean sizes; the radii of gyration, R_g, for A and B are 1.480 Å and 1.401 Å, respectively. However, their folding features are clearly distinct: a long "strand" is *outside* the helix in conformer A and *inside* the helix in conformer B. (We follow the usual drawing convention, whereby one "hides" a curve section whenever a bond is *behind* another one.) Even though entanglements are *dynamically transient* in linear polymers,[1] it is evident that the instantaneous configuration B in Figure 1 is *strongly self-entangled*. These differences in entanglement complexity are readily captured by the overcrossing numbers. In the depicted projection, conformer A has $N = 3$ overcrossings. When all projections are taken into account, we find a *min* $N = 0$ and a mean overcrossing number $\overline{N}^{(3D)} = 6.05 \pm 0.03$ for the same conformer. (The error bar corresponds to one standard deviation derived with five different random distributions up to 25 000 projections.) In contrast, the projection depicted in Figure 1 for conformer B has $N = 9$ overcrossings; when all projections are accounted, we find *min* $N = 7$ and a mean overcrossing number $\overline{N}^{(3D)} = 10.52 \pm 0.02$. In other words, polymer self-entanglements translate typically in much larger $\overline{N}^{(3D)}$ values.

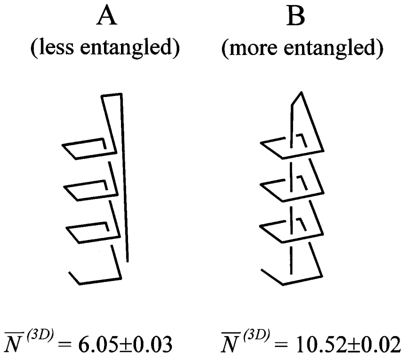
A
(less entangled)
B
(more entangled)
$\overline{N}^{(3D)} = 6.05 \pm 0.03$
$\overline{N}^{(3D)} = 10.52 \pm 0.02$

Figure 1. Entanglement complexity for two conformations of the same three-dimensional (3D) curve, as measured in terms of the 3D mean overcrossing number, $\overline{N}^{(3D)}$. Intuitively, it is clear that the B conformer is more entangled than conformer A since its long "strand" is folded *inside* the helix. The large difference in $\overline{N}^{(3D)}$ values (shown below the snapshots) reflects these distinct folding features. Note that the $\overline{N}^{(3D)}$ descriptor contains information of *all* (infinitely many) planar linear projections of the 3D curve. The goal of the alternative $\overline{N}^{(2D)}$ descriptor is to convey a similar information by using a *single* nonlinear projection of the original 3D backbone.

In this work, we discuss the rigorous computation of similar descriptors for objects of different dimensionality,

* Corresponding author phone: + 1 - (705) 675-1151, ext. 2117; fax: + 1 - (705) 675-4844; e-mail: Gustavo@laurentienne.ca.

10.1021/ci020289z CCC: $25.00 © 2003 American Chemical Society
Published on Web 11/16/2002

particularly, two-dimensional (2D) polymer chains. Let us state briefly the main problem addressed here. (The detailed analysis is given in the next section.) Consider a 3D-polymer with n identical monomers. The chain backbone, which contains the main information on folding features, can be modeled as 3D-polygon of $n-1$ bonds with identical bond length, b. The mean overcrossing number of a 3D-chain configuration can be written in terms of bond pairs.[10] Let $\overline{N}_{ij}^{(3D)}$ be the contribution to the total mean overcrossing number $\overline{N}^{(3D)}$ corresponding to the i-th and j-th bonds, then[10b]

$$\overline{N}^{(3D)} = \sum_{i}\sum_{j} \overline{N}_{ij}^{(3D)} = 2\sum_{i<j}\sum_{j} \overline{N}_{ij}^{(3D)} \tag{1}$$

Let us now write the ith-bond as a parametrized vector $\gamma_i(s)$ over the unit interval

$$\gamma_i :\ I \rightarrow \mathscr{R}^3,\ \gamma_i(s) = \mathbf{R}_i + s\,(\mathbf{R}_{i+1} - \mathbf{R}_i),\ s \in I = [0,1] \tag{2}$$

where $\{\mathbf{R}_i\}$ the node coordinates measured from the backbone centroid. The bond vector has continuous parametric derivative $\dot{\gamma}_i(s) = \partial\gamma_i(s)/\partial s$ and components: $\gamma_i(s) = \sum_{m=1}^{3} \gamma_{im}\,(s)\,\mathbf{e}_m$, in terms of Cartesian unit vectors. In terms of these quantities, the pair contribution $\overline{N}_{ij}^{(3D)}$ can be computed analytically by using a path integral formalism based on Gauss mappings[9−11] (briefly discussed in the next section). The result is[10b]

$$\overline{N}_{ij}^{(3D)} = \frac{1}{4\pi}\int_0^1\int_0^1 \frac{|(\dot{\gamma}_i\,(s) \times \dot{\gamma}_j\,(t))\cdot(\gamma_i\,(s) - \gamma_j\,(t))|}{||\gamma_i\,(s) - \gamma_j(t)||^3}\, ds\, dt \tag{3}$$

which adopts a simple analytical form in the case of a polymer network in a cubic lattice. In particular, it emerges from eq 3 that the large-distance contribution to the overcrossing number of two perpendicular bonds is $\overline{N}_{ij}^{(3D)} \sim 1/r_{ij}^2$, where r_{ij} is the mean distance between the i-th and j-th bonds.[9e] This fast decrease of $\overline{N}_{ij}^{(3D)}$ with the bond-bond distance ensures that well separated bonds do not contribute much to the total $\overline{N}^{(3D)}$. The value of $\overline{N}^{(3D)}$ is dominated by contributions from any large regions where densely packed bonds appear in entangled loops.

For the analysis of folding features in 2D-polymers, a proposed descriptor is the "*mean radial intersection number*", here denoted by $\mathscr{F}$.[12] The parameter $\mathscr{F}$ is an intuitive extension of the notion of "crossings along a line of sight" to the case of a 2D-figure. The concept is very simple, as explained in Figure 2. This figure shows a generic open 2D-curve. Within the plane, "lines of sight" are visualized as the dashed straight lines containing the centroid of the curve (indicated by the central black dot). If we now consider the number of intersections between these lines and the curve, we can compute a number $\mathscr{F}$ associated with each possible "viewing" direction. (The figure shows examples leading to one, two, and four of such intersections.) The final value $\mathscr{F}_{chain}$ for the curve, averaged over all possible directions, is the descriptor used for 2D-polymers. The $\mathscr{F}_{chain}$ index can be computed analytically as a sum of bond contributions.[12] For a generic bond between the pair of nodes (i,j), the local intersection number $\mathscr{F}_{ij}$ is the fraction of the circumference of the unit circle subtended by the angle between the (i,j)-nodes, as shown in Figure 3: $\mathscr{F}_{ij} = |\theta_{ij}\,|/2\pi$.[12b] In terms of coordinates measured from the curve's centroid (cf. Figure 3), we have $\theta_{ij} = arccos\ \{\mathbf{R}_i\cdot\mathbf{R}_j/R_iR_j\}$, with $R_i = ||\mathbf{R}_i||$. Finally, the mean intersection number for an open 2D-chain becomes[12b]

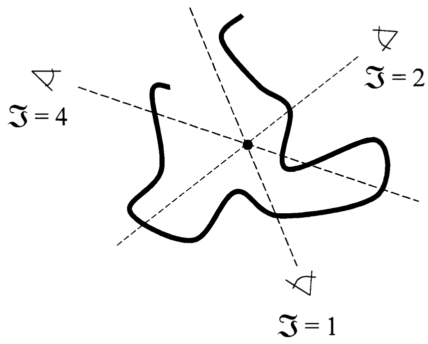

Figure 2. The radial intersection number for a planar curve. An intersection number $\mathscr{F}$ is associated with each "viewing direction", represented by dashed lines. Each line contains the centroid of the curve (black dot). Some examples of $\mathscr{F}$ values are highlighted. The mean value for the curve, $\mathscr{F}_{chain}$, is the value of $\mathscr{F}$ per line averaged over all possible "viewing directions".

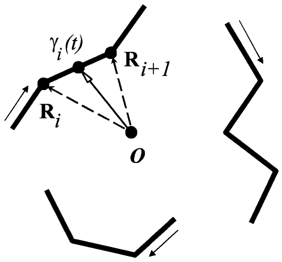

Figure 3. Geometrical information required for the computation of the mean radial intersection number, $\mathscr{F}$. The vectors $\{\mathbf{R}_i\}$ represent the position of the nodes of the network measured from the centroid O. The $(i,i+1)$-bond is represented as a parametrized segment $\gamma_i\ (t)$, where $\gamma_i\ (0) = \mathbf{R}_i$ and $\gamma_i\ (1) = \mathbf{R}_{i+1}$.

$$\mathscr{F}_{chain} = 2\sum_{i<j}\sum_{j} \epsilon_{ij}\,\mathscr{F}_{ij} = \frac{1}{\pi}\sum_{i=1}^{n-1} |\theta_{i,i+1}| \tag{4}$$

where $\{\epsilon_{ij}\}$ is the connectivity matrix, i.e., $\epsilon_{ij} = 1$ if nodes (i,j) are connected, and zero otherwise. As discussed above, eq 4 provides a 2D-descriptor of folding features conceived as an imitation of the self-entanglement descriptor for 3D-curves (eq 3). Yet, any rigorous relation between these two properties is lacking. In this work, we show that, indeed, $\mathscr{F}_{chain}$ and $\overline{N}^{(3D)}$ are the 2D- and 3D-versions of the same type of descriptor. In other words, $\mathscr{F}_{chain}$ can be seen as equivalent to a $\overline{N}^{(2D)}$ descriptor. The formulation discussed in the next section provides a single, unified approach to characterizing the shape of generic D-dimensional polymer models and a strategy to design new shape descriptors.

D-DIMENSIONAL FOLDING DESCRIPTORS FROM GAUSS MAPPINGS

Below, we show that the mean radial intersection number $\mathscr{F}$ of a 2D-curve is the proper equivalent in 2-space of the mean overcrossing number of a 3D-chain. The connection between these two descriptors stems from representing them as path integrals of Gauss mappings.

A Gauss (or "spherical") map assigns a vector to the unit *D*-sphere. Let us review briefly how $\overline{N}^{(3D)}$ is computed using these maps.[10] A projected bond-bond crossing corresponds to a vector joining two points on different bonds; this vector is perpendicular to a plane tangent to a sphere enclosing the 3D-chain. (We consider here the smallest sphere centered at the backbone's centroid.) Since each bond is parametrized over the unit interval, we can write the mapping as

$$f^{(3D)}\!: \; I \times I \rightarrow S^2, I = [0,1] \tag{5a}$$

in terms of two overcrossing bond vectors, *i* and *j*, defined as in eq 2[10b]

$$f_{ij}^{(3D)}(s,t) = \frac{\gamma_i(s) - \gamma_j(t)}{||\gamma_i(s) - \gamma_j(t)||} \in S^2 \subset \mathscr{R}^3 \tag{5b}$$

with the following components: $f_{ij}^{(3D)}(s,t) = \sum_{k=1}^{3} f_{ij,k}^{(3D)}\, \mathbf{e}_k$, $f_{ij,k}^{(3D)} = (\gamma_{ik}(s) - \gamma_{jk}(t))/||\gamma_i(s) - \gamma_j(t)||$. The $\overline{N}_{ij}^{(3D)}$ contribution is a fraction of the 2-sphere's surface weighted by the cardinality of the inverse mapping[10b]

$$\overline{N}_{ij}^{(3D)} = \int_{S^2} card\{[f_{ij}^{(3D)}]^{-1}\,(\mathbf{p})\}\, d\sigma / \int_{S^2} d\sigma \tag{6}$$

with $d\sigma$ the surface element on S^2 at a point **p**. The numerator in eq 6 can be computed as an integral over the image of *f* on S^2 or as the reciprocal image of *f* on its $I \times I$ domain: [10b,11]

$$\overline{N}_{ij}^{(3D)} = \frac{1}{4\pi} \int_{f_{ij}^{(3D)}(I \times I)} |d\sigma| = \frac{1}{4\pi} \int_{I \times I} |f_{ij}^{(3D)} * d\sigma| \tag{7}$$

The function $f{*}d\sigma$ in eq 7 is the pullback of $d\sigma$ under map *f*.[13] In terms of the exterior derivatives along the (i,j)-bonds, denoted by $f_{ij} * \dot{\gamma}_i$ and $f_{ij} * \dot{\gamma}_j$, the reciprocal image of the surface element becomes $|f * d\sigma| = |f{\cdot}(f{*}\, \dot{\gamma}_i \times f{*}\, \dot{\gamma}_j)|\; ds\, dt$, with derivatives given as[13]

$$f_{ij}^{(3D)} * \dot{\gamma}_i = \sum_{k=1}^{3} \sum_{m=1}^{3} \dot{\gamma}_{im} \frac{\partial f_{ij,k}^{(3D)}}{\partial \gamma_{im}} \mathbf{e}_k,$$

$$\text{where: } \dot{\gamma}_{im}(s) = \frac{\partial \gamma_{im}(s)}{\partial s} \tag{8}$$

Using (8) on (5b), we get $f_{ij}^{(3D)} * \dot{\gamma}_i = \dot{\gamma}_i/||\gamma_i - \gamma_j|| - \{(\gamma_i - \gamma_j)/||\gamma_i - \gamma_j||^3\}(\dot{\gamma}_i{\cdot}(\gamma_i - \gamma_j))$; therefore

$$|f_{ij}^{(3D)} * d\sigma| = \frac{|(\gamma_i - \gamma_j){\cdot}(\dot{\gamma}_i \times \dot{\gamma}_j)|}{||\gamma_i - \gamma_j||^3}\, ds\, dt \tag{9}$$

The pair contribution $\overline{N}_{ij}^{(3D)}$ in (3) emerges from (7) and (9). For a 3D-chain the result is[10b]

$$\overline{N}_{chain}^{(3D)} = \frac{1}{2\pi} \sum_{i=1}^{n-2} \sum_{j=i+2}^{n} \int_0^1 \int_0^1 \frac{|(\dot{\gamma}_i(s) \times \dot{\gamma}_j(t)){\cdot}(\gamma_i(s) - \gamma_j(t))|}{||\gamma_i(s) - \gamma_j(t)||^3}\, ds\, dt \tag{10}$$

We can now derive the 2D-descriptor of folding features that matches the previous formulation. The procedure would then be as follows. First, we define a $\overline{N}^{(2D)}$ descriptor for a 2D-chain, written as a sum over pair contributions $\overline{N}_{ij}^{(2D)}$. Second, we propose a Gauss map associated with 2D "crossings". Then, we compute $\overline{N}_{ij}^{(2D)}$ as a line integral of the reciprocal image of this mapping. Finally, we compare the result with the intuitive approach in eq 4.

Let us then consider a 2D-chain as a sequence of bond vectors parametrized as in eq 2

$$\gamma_i\!: \; I \rightarrow \mathscr{R}^2, \gamma_i(s) = \mathbf{R}_i + s\,(\mathbf{R}_{i+1} - \mathbf{R}_i) = \sum_{m=1}^{2} \gamma_{im}(s)\, \mathbf{e}_m,\; s \in I \tag{11}$$

For the sake of a simpler notation, yet with no loss of generality, we shall write $\overline{N}^{(2D)}$ as a sum over *pairs of nodes* given that a 2D-intersection is defined for a *single* bond

$$\overline{N}^{(2D)} = \sum_i \sum_j \epsilon_{ij}\, \overline{N}_{ij}^{(2D)} = 2 \sum_{i<j} \sum_j \epsilon_{ij}\, \overline{N}_{ij}^{(2D)} \tag{12a}$$

with $\{\epsilon_{ij}\}$ the connectivity matrix. Since the bonds in a simple chain are $(i,i{+}1)$-pairs, we have

$$\overline{N}_{chain}^{(2D)} = 2 \sum_{i=1}^{n-1} \overline{N}_{i,i+1}^{(2D)} \tag{12b}$$

Note that, in contrast, $\overline{N}^{(3D)}$ is written as a sum of *bond pairs'* contributions. Here, $\overline{N}_{ij}^{(2D)}$ stands for the contribution of a *single bond*, the one formed by the $(i,i{+}1)$-pair. Accordingly, the Gauss map for an intersection (or "crossing") associated with a single bond takes the form

$$f^{(2D)}\!: \; I \rightarrow S^1 \tag{13}$$

as it projects a single point to the unit circle S^1. Using (11), the map for the $(i,i{+}1)$-bond is

$$f_i^{(2D)}(s) = \frac{\gamma_i(s)}{||\gamma_i(s)||} \in S^1 \subset \mathscr{R}^2 \tag{14}$$

with the following components: $f_i^{(2D)}(s) = \sum_{k=1}^{2} f_{i,k}^{(2D)}\, \mathbf{e}_k$, $f_{i,k}^{(2D)} = \gamma_{ik}(s)/||\gamma_i(s)||$. As in eq 6, we then write

$$\overline{N}_{i,i+1}^{(2D)} = \int_{S^1} card\{[f_i^{(2D)}]^{-1}\,(\mathbf{p})\}\, \boldsymbol{dl} / \int_{S^1} \boldsymbol{dl} \tag{15}$$

with ***dl*** the differential of arc length at point **p**. It follows from the equalities in (7) that

$$\overline{N}_{i,i+1}^{(2D)} = \frac{1}{2\pi} \int_{f_i^{(2D)}(I)} |\boldsymbol{dl}| = \frac{1}{2\pi} \int_I |f_i^{(2D)} * \boldsymbol{dl}| \tag{16}$$

where the reciprocal image of the "volume differential" on S^1 becomes $|f * dl| = ||f{*}\dot{\gamma}_i||\; dt$, in terms of the exterior derivative along the *i*-th bond. When applying eq 8 to the $f^{(2D)}$ mapping, we obtain $f_i^{(2D)}{*}\dot{\gamma}_i = \dot{\gamma}_i/||\gamma_i|| - \{\gamma_i/||\gamma_i||^3\}(\dot{\gamma}_i{\cdot}\gamma_i)$, with which the node pair contribution becomes

$$\overline{N}^{(2D)}_{i,i+1} = \frac{1}{2\pi}\int_0^1 \frac{\{||\dot{\gamma}_i||^2\,||\gamma_i||^2 - (\dot{\gamma}_i\cdot\gamma_i)^2\}^{1/2}}{||\gamma_i(t)||^2}\,dt = \frac{1}{2\pi}\int_0^1 \frac{||\dot{\gamma}_i(t)\times\gamma_i(t)||}{||\gamma_i(t)||^2}\,dt \quad (17)$$

It only remains to perform the computations in (17) for a generic 2D-chain. If, for simplicity, all bonds have the same length b, then we have $||\dot{\gamma}_i||^2 = ||\mathbf{R}_{i+1} - \mathbf{R}_i||^2 = R_i^2 + R_{i+1}^2 - 2\mathbf{R}_i\cdot\mathbf{R}_{i+1} = b^2$. Simple algebra shows that the numerator in (17) is *independent* of the parameter t

$$||\dot{\gamma}_i(t)\times\gamma_i(t)||^2 = ||\dot{\gamma}_i||^2\,||\gamma_i||^2 - (\dot{\gamma}_i\cdot\gamma_i)^2 = R_i^2\,R_{i+1}^2\,sin^2\,\theta_{i,i+1} \quad (18a)$$

whereas the denominator takes the form

$$||\gamma_i(t)||^2 = R_i^2 + t^2\,b^2 + 2t\{\mathbf{R}_i\cdot\mathbf{R}_{i+1} - R_i^2\} \quad (18b)$$

Accordingly, the main contribution to eq 17 becomes the integral:[14]

$$\int_0^1 \frac{dt}{||\gamma_i(t)||^2} = \frac{1}{R_{i+1}\,R_i\,sin\,\theta_{i,i+1}} \times arctan\left(\frac{t\,b^2 + [\mathbf{R}_i\cdot\mathbf{R}_{i+1} - R_i^2]}{R_{i+1}\,R_i\,sin\,\theta_{i,i+1}}\right)\Bigg|_0^1 \quad (19)$$

Replacing into $\overline{N}^{(2D)}_{i,i+1}$ and using $arctan\,(x) = -\,arctan\,(-x)$), we obtain

$$\overline{N}^{(2D)}_{i,i+1} = \frac{1}{2\pi}\left|arctan\left(\frac{R_{i+1}^2 - \mathbf{R}_i\cdot\mathbf{R}_{i+1}}{R_{i+1}\,R_i\,sin\,\theta_{i,i+1}}\right) + arctan\left(\frac{R_i^2 - \mathbf{R}_i\cdot\mathbf{R}_{i+1}}{R_{i+1}\,R_i\,sin\,\theta_{i,i+1}}\right)\right| \quad (20)$$

If we consider that $arctan\,(x) = arccos\,([1{+}x^2]^{-1/2})$, the choice $x = (R_{i+1}^2 - \mathbf{R}_i\cdot\mathbf{R}_{i+1})/R_i\,R_{i+1}\,sin\,\theta_{i,i+1}$, i.e., $[1{+}x^2]^{-1/2} = (R_i/b)\,sin\,\theta_{i,i+1}$, simplifies the expression (20) into

$$\overline{N}^{(2D)}_{i,i+1} = \frac{1}{2\pi}\left|arccos\left(\frac{R_i}{b}sin\,\theta_{i,i+1}\right) + arccos\left(\frac{R_{i+1}}{b}sin\,\theta_{i,i+1}\right)\right| \quad (21)$$

Finally, we apply the equality $arccos\,(y) + arccos\,(z) = arccos\,\{yz - [(1-y^2)(1-z^2)]^{-1/2}\}$, valid for $y + z \geq 0$.[14] By choosing $y = (R_i/b)\,sin\,\theta_{i,i+1}$ and $z = (R_{i+1}/b)\,sin\,\theta_{i,i+1}$ and noting that

$$yz - [(1-y^2)(1-z^2)]^{-1/2} = \mathbf{R}_i\cdot\mathbf{R}_{i+1}/R_i\,R_{i+1} = cos\,\theta_{i,i+1} \quad (22)$$

we obtain from (21) the main result of this work:

$$\overline{N}^{(2D)}_{i,i+1} = |\theta_{i,i+1}|/2\pi \quad (23)$$

Finally substituting (23) into the result (12b) for the entire chain, we arrive at an equivalence between the 2D-descriptor based on the Gauss map and the intuitive radial intersection number:

$$\overline{N}^{(2D)}_{chain} = 2\sum_{i=1}^{n-1}\overline{N}^{(2D)}_{i,i+1} = \frac{1}{\pi}\sum_{i=1}^{n-1}|\theta_{i,i+1}| \equiv \mathscr{F}_{chain} \quad (24)$$

In other words, the $\mathscr{F}_{chain}$ index is the equivalent of the 3D-overcrossing number $\overline{N}^{(3D)}$ for a 2D-curve, provided that the Gauss mapping is chosen as in (13) and (14). The only difference between $\overline{N}^{(2D)}_{chain}$ and the $\overline{N}^{(2D)}$ descriptor for 2D-curves in ref 10b is a reference point. Since a chain comprising a single bond (or an even-n chain of collinear bonds) has $\mathscr{F}_{chain} = 1$, the choice

$$\overline{N}^{(2D)} = \overline{N}^{(2D)}_{chain} - 1 \quad (25)$$

ensures that a rodlike conformer has folding complexity “zero” in both 2- and 3-space.

FURTHER COMMENTS AND CONCLUSIONS

We have shown a general strategy to build descriptors of folding complexity based on spherical mappings. The approach contains the indexes discussed in the literature for 3D- and 2D-polymer chains as particular examples. In the standard parametric notation for a straight-line bond γ_i, the *3D-descriptor based on crossings between bond pairs* is found to be

$$\overline{N}^{(3D)}_{chain} = \frac{1}{2\pi}\sum_{i=1}^{n-2}\sum_{j=i+2}^{n}\int_0^1\int_0^1 \frac{|(\dot{\gamma}_i(s)\times\dot{\gamma}_j(t))\cdot(\gamma_i(s) - \gamma_j(t))|}{||\gamma_i(s) - \gamma_j(t)||^3}\,ds\,dt \quad (26a)$$

whereas the *2D-descriptor based on radial intersections with single bonds* is

$$\overline{N}^{(2D)}_{chain} = \frac{1}{\pi}\sum_{i=1}^{n-1}\int_0^1 \frac{||\dot{\gamma}_i(t)\times\gamma_i(t)||}{||\gamma_i(t)||^2}\,dt \quad (26b)$$

These relations are not restricted to chains with constant bond length. For 2D-polymer chains, such models can be derived by multidimensional scaling. These transformations, commonly used for data compression,[15,16] are based on nonlinear mappings that mimic the shape features of a higher-dimensional data set. An example of these is the Sammon transform,[15] which allows one to generate a 2D-set of points that optimally preserves the pattern of distances among points in an initial 3D-set. This mapping has been proposed as a tool to generate planar projections of protein backbones[17] and compare their properties with those of simpler 2D-models.[12b]

For illustration, Figure 4 shows examples of Sammon projections for selected protein backbones, together with their folding characterization. The 2D-models have been generated using the algorithm by Rauber et al.[18] (see ref 12b for a discussion). The figure gives the proteins’ codes according to the Protein Data Bank, together with the number of amino acid residues n. The folding descriptor values correspond to $\overline{N}^{(2D)}$, defined as in eq 25. In the case of small compact structures such as those of 1ina and 1ccd proteins, most radial lines produce 5 or 6 intersections and thus lead to $\overline{N}^{(2D)} \approx$

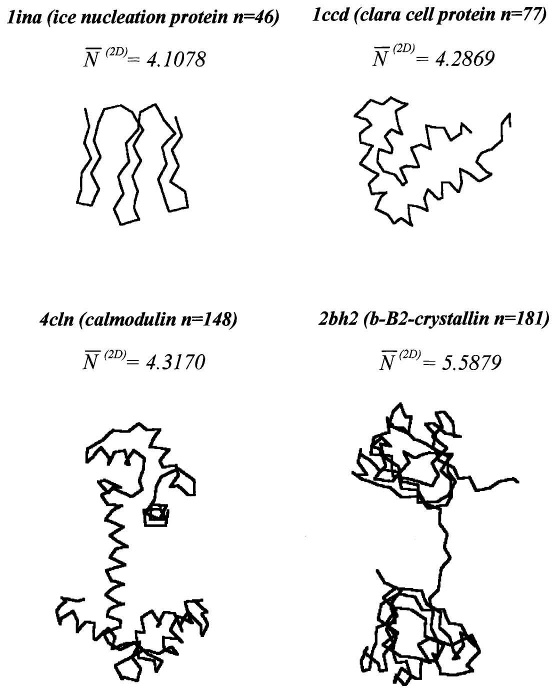
1ina (ice nucleation protein n=46)
$\overline{N}^{(2D)}$= 4.1078
1ccd (clara cell protein n=77)
$\overline{N}^{(2D)}$= 4.2869
4cln (calmodulin n=148)
$\overline{N}^{(2D)}$= 4.3170
2bh2 (b-B2-crystallin n=181)
$\overline{N}^{(2D)}$= 5.5879

Figure 4. Folding complexity index, $\overline{N}^{(2D)} = \mathscr{F}_{chain} - 1$, for 2D-models of selected protein native states. The 2D-models are Sammon projections of the backbones of the corresponding 3D-proteins. This nonlinear projection is designed to maximize the conservation of the original shape features of the 3D-backbone.

4. In contrast, the proteins 4cln and 2bh2 have radial lines producing higher intersection numbers, yet their final $\overline{N}^{(2D)}$ values are not very large. This latter behavior translates the fact that, although these proteins have longer backbones, they adopt noncompact folds with well separated domains. Note, for example, that 4cln has two domains joined by a long α−helix; each domain has a comparable entanglement complexity to that seen in the smaller 1ccd protein. Since the 4cln domains are *not* entangled with each other, the resulting $\overline{N}^{(2D)}$ value is not much different from that in 1ccd. In contrast, the two domains in 2bh2 are also well separated, yet individually more self-entangled; as a result, the $\overline{N}^{(2D)}$ value is larger than that in 4cln. Still, these two $\overline{N}^{(3D)}$ values are smaller than those corresponding to compact proteins with the same number of amino acid residues. For example, 2uce is a compact protein with the same chain length as 4cln($n = 148$), yet its more entangled native state produces $\overline{N}^{(2D)} = 9.53$. Similarly, 5sga has the same chain length as 2bh2 ($n = 181$), but it exhibits a more entangled and compact native fold with $\overline{N}^{(2D)} = 13.11$. These results serve to remind the reader that both $\overline{N}^{(2D)}$ and $\overline{N}^{(3D)}$ are *global* descriptors of entanglement complexity; they are computed *from* local contributions but they convey *only large-scale folding features*.

Other proteins can be studied in a similar fashion. The combined approach of multidimensional scaling and a 2D-descriptor of folding is convenient because the accurate computation of (26a) is much more time-consuming than performing a Sammon projection. On the other hand, solving for (26b) is very fast.

Following the approach in this work, it is possible to design alternative descriptors of folding complexity by varying the nature of the Gauss map. Here, we have considered a 2D-descriptor based on *single* bond contributions (cf. eq 14) in order to compare with the indices proposed in ref 12. One could of course adopt a different approach whereby the $\overline{N}^{(2D)}$ index is computed by contributions of *pairs* of bonds, as in the Gauss map (5b). We believe, however, that this approach would produce little advantage since a single "two-bond crossing" in 2-space can be detected by the occurrence of two single-bond radial intersections, i.e. $\mathscr{F} = 2$. Thus, two-bond contributions can simply be taken into account by adding a constant to the single-bond $\overline{N}^{(2D)}$ index, as done in eq 25. In contrast, a single-bond approach to computing the $\overline{N}^{(3D)}$ index would be futile because the projection of a single bond does not contribute to the area of S^2 (i.e., the numerator in eq 6 will vanish).

ACKNOWLEDGMENT

This work has been supported by a discovery grant from NSERC (Canada) and the Canada Research Chairs' Program.

REFERENCES AND NOTES

(1) de Gennes, P.-G. *Scaling Concepts in Polymer Physics*; Cornell University Press: Ithaca, 1985.

(2) Haber, C.; Ruiz, S. A.; Wirtz, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 10792.

(3) Chothia, C.; Hubard, T.; Brenner, S.; Barns, H.; Murzin, A. *Annu. Rev. Biophys. Biomol. Struct.* **1997**, *26*, 597.

(4) Arteca, G. A.; Tapia, O. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 642.

(5) (a) Hoaglund-Hyzer, C. S.; Counterman, A. E.; Clemmer, D. E. *Chem. Rev.* **1999**, *99*, 3037. (b) Jarrold, M. F. *Annu. Rev. Phys. Chem.* **2000**, *51*, 179.

(6) Arteca, G. A.; Reimann, C. T.; Tapia, O. *Mass Spectrom. Rev.* **2001**, *20*, 402.

(7) (a) Stasiak, A.; Katritch, V.; Bednar, J.; Michoud, D.; Dubochet J. *Nature* **1996**, *384*, 122. (b) Katritch, V.; Bednar, J.; Michoud, D.; Scharein, R. G.; Dubochet, J.; Stasiak, A. *Nature* **1996**, *384*, 142. (c) Vologodskii, A. V.; Crisona, N. J.; Laurie, B.; Pieranski, P.; Katritch, V.; Dubochet, J.; Stasiak, A. *J. Mol. Biol.* **1998**, *278*, 1. (d) Diao, Y.; Ernst, C. *Topology Appl.* **1998**, *90*, 1. (e) Buck, G.; Simon, J. *Topology Appl.* **1999**, *91*, 245. (f) Dai, X.; Diao, Y. *J. Knot Theory Ramif.* **2000**, *9*, 713.

(8) (a) Arteca, G. A.; Mezey, P. G. *Biopolymers* **1992**, *32*, 1609. (b) Arteca, G. A. *Biopolymers* **1993**, *33*, 1829. (c) Orlandini, E.; Tesi, M. C.; Whittington, S. G.; Sumners, D. W.; Janse van Rensburg, E. J. *J. Phys. A* **1994**, *27*, L333. (d) Arteca, G. A. *Phys. Rev. E* **1995**, *51*, 2600. (e) Orlandini, E.; Tesi, M. C.; Whittington, S. G. *J. Phys. A* **2000**, *33*, L181. (f) Huang, J.-Y.; Lai, P.-Y. *Phys. Rev. E* **2001**, *63*, 021506.

(9) (a) Cantarella, J.; Kusher, R. B.; Sullivan, J. M. *Nature* **1998**, *392*, 237. (b) Buck, G. *Nature* **1998**, *392*, 238. (c) Kholodenko, A. L.; Vilgis, T. A. *Phys. Rep.* **1998**, *298*, 251. (d) Grassberger, P. *J. Phys. A* **2001**, *34*, 9959. (e) Arteca, G. A. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 329.

(10) (a) Arteca, G. A.; Caughill, D. I. *Can. J. Chem.* **1998**, *76*, 1402. (b) Arteca, G. A. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 550.

(11) (a) Calugareanu, G. *Rev. Math. Pure Appl.* **1959**, *4*, 5. (b) Arnol'd, V. I. *Sel. Math. Sov.* **1986**, *5*, 327. (c) Freedman, M. E.; He, Z.-X. *Ann. Math.* **1991**, *134*, 189. (d) Brynson, S.; Freedman, M. E.; He, Z.-X.; Wang, Z. *Bull. Am. Math. Soc. (NS)* **1993**, *28*, 99.

(12) (a) Arteca, G. A.; Zhang, S. *Phys. Rev. E* **1999**, *59*, 4209. (b) Arteca, G. A. *Phys. Rev. E* **1999**, *60*, 6206.

(13) O'Neill, B. *Elementary Differential Geometry*; Academic Press: Boston, 1966.

(14) Gradshtein, I. S.; Ryzhik, I. M. *Tablitsy Integralov, Summ, Riadov i Proizvedenii [Tables of integrals, sums, series, and products]*; FM Publishers: Moscow, 1962.

(15) Sammon, J. W., Jr. *IEEE Trans. Comput. C* **1969**, *18*, 401.

(16) Kohonen, T. *Self-organizing Maps*; Springer: Berlin, 2001.

(17) Barlow, T. W.; Richards, W. G. *J. Mol. Struct. (THEOCHEM)* **1997**, *398*, 483.

(18) Rauber, T. W.; Barata, M. M.; Steiger-Garção, A. S. *Proc. Int. Conf. Fault Diagnosis (Tooldiag'93)* **1993**, *3*, 906.

CI020289Z