Addendum to ''Quantitative measure of folding in two-dimensional polymers''

Gustavo A. Arteca*
Département de Chimie et Biochimie, Laurentian University, Ramsey Lake Road, Sudbury, Ontario, Canada P3E 2C6
(Received 19 July 1999)

Recently, we introduced a measure of folding complexity for two-dimensional polymers, $\overline{N}$, the *mean radial intersection number* [Phys. Rev. E **59**, 4209 (1999)]. In this addendum, we expand on three aspects of the previous work. First, we provide an analytical expression for $\overline{N}$ that is valid for two-dimensional networks. Second, we show that the power-law scaling $\overline{N} \approx n^{\beta}$, with n the number of monomers, has a different critical exponent for random and self-avoiding walks. Finally, we find that the folding features in optimized projections of experimental three-dimensional (native) protein backbones fall between the latter limit models. [S1063-651X(99)08511-6]

PACS number(s): 87.15.By, 05.50.+q, 02.70.Lq

In addition to being a natural representation for adsorbed polymers, two-dimensional (2D) molecular models are also useful to test the validity of current ideas on structure and dynamics for more realistic (three-dimensional) macromolecules. (See Ref. [1] for a review on the literature.) In this context, 2D *self-avoiding* walks have been used when modeling protein folding dynamics [2], whereas 2D *self-intersecting* walks occur in projected polymer knots [3] or optimal (nonlinear) projections of proteins backbones [4]. A great deal of work has been devoted to understanding the statistical properties of these models; in particular, the mean molecular size for the accessible conformers. Recently, we introduced a complementary tool for analysis that discriminates between conformers with similar molecular size but different folding features [1]. Our descriptor of 2D folding, denoted by $\overline{N}$, is an extension of the *mean overcrossing number* used for the characterization of entanglements in 3D linear [5,6] and knotted polymers [7]. The shape descriptor $\overline{N}$ takes into account geometrical and topological features of a chain, i.e., the atomic positions and the chain connectivity. As a result, $\overline{N}$ characterizes features not conveyed by standard descriptors of molecular size; e.g., the mean radius of gyration $\langle R_g^2 \rangle^{1/2}$. For this reason, we refer to $\overline{N}$ as a *descriptor of folding complexity*. The configurationally averaged molecular size, given by $\langle R_g^2 \rangle^{1/2}$, and the configurationally averaged folding complexity, given by $\langle \overline{N} \rangle$, are correlated *only* in limit cases. For example, conformers with the largest size have the simplest folds (e.g., a rodlike chain with $\overline{N} \approx 0$), whereas the smallest conformers exhibit the maximum compatible folding complexity. Otherwise, R_g^2 and $\overline{N}$ are not correlated and together provide a discriminating pair of shape descriptors. In this work, we discuss a number of properties of $\langle \overline{N} \rangle$ that extend our discussion in Ref. [1].

The descriptor $\overline{N}$ is computed from the intersections between the molecular chain and a selected set of ''reference'' lines. By choosing these reference lines as those containing the centroid of the polymer, the approach becomes the 2D extension of the one used to calculate mean overcrossing numbers in 3D chains [5]. In Ref. [1], the practical algorithm to compute $\overline{N}$ was as follows: (i) Enclose the polymer inside the smallest circle centered at the chain's centroid; (ii) choose a point $\mathbf{p}_1$ on the circle, and compute the number of intersections I_1 between the diameter line associated with $\mathbf{p}_1$ and backbone bonds; (iii) the descriptor $\overline{N}$ is computed from the average number of intersections,

$$\overline{I} = (1/m) \sum_{j=1}^{m \gg 1} I_j .$$

Our choice of $\overline{N} = \overline{I} - 1$ ensures that a rodlike chain is assigned $\overline{N} = 0$ both in two and three dimensions (i.e., no ''entanglements''). As shown below, the descriptor $\overline{N}$ can also be computed analytically in terms of chain geometry and connectivity.

Let $\{\mathbf{R}_i\}$ be the node coordinates of the 2D polymer backbone, and $\{\mathbf{R}_i'\}$ the center-of-mass coordinates. Let $\{\varepsilon_{ij}\}$ be the connectivity matrix, where $\varepsilon_{ij} = 1$ if the ith and jth nodes are connected, and zero otherwise. The mean radial intersection descriptor $\overline{N}$ can be computed as a sum of *individual bond contributions*. Consider radial lines stemming from the chain's centroid, O', as in Fig. 1. All radial lines within the two shaded areas indicated as $A(i,j)$ will intersect the bond segment $\mathbf{R}_j' - \mathbf{R}_i'$. The number of intersections with a bond can be computed as the fraction of the circle's area corresponding to A. In turn, this fractional area is determined by the angle between the position vectors for the bond in question. Thus, for the bond $\mathbf{R}_j' - \mathbf{R}_i'$, the fractional number of intersections $I(i,j)$ becomes $I(i,j) = \pi^{-1} \arccos\{\mathbf{R}_j' \cdot \mathbf{R}_i' / \|\mathbf{R}_i'\| \|\mathbf{R}_j'\|\}$. When adding these contributions over all chain nodes ($\varepsilon_{ij} \neq 0$), we obtain

$$\overline{N} = \frac{1}{\pi} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \varepsilon_{ij} \arccos\left\{ \frac{\mathbf{R}_i^j \cdot \mathbf{R}_j^i}{\|\mathbf{R}_i'\| \|\mathbf{R}_j'\|} \right\} - 1, \qquad (1)$$

which gives the mean radial intersection number $\overline{N}$ for configurations in arbitrary topologies.

In Ref. [1] we considered adsorbed polymer configurations modeled by 2D self-avoiding walks (SAWs). For these systems, we showed that the configurational average $\langle \overline{N} \rangle$ ex-

*Electronic address: Gustavo@nickel.laurentian.ca

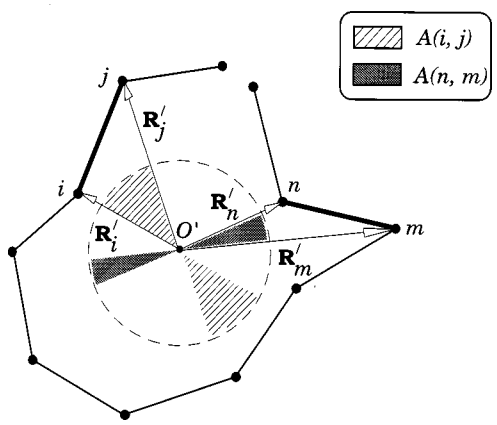

FIG. 1. Geometrical interpretation of the mean radial intersection number. The areas $A(i,j)$ and $A(n,m)$ contain the radial lines intersecting the bonds $\mathbf{R}'_j-\mathbf{R}'_i$ and $\mathbf{R}'_m-\mathbf{R}'_n$, respectively. The radial intersection descriptor $\overline{N}$ uses these areas to convey the folding features of a two-dimensional polymer chain (see text).

hibits power-law scaling with the number of monomers n. Recent work using nonlinear projections of protein backbones [4] shows that *self-intersecting* walks can also occur in 2D biopolymers. Here, we use Eq. (1) to analyze some properties of $\overline{N}$ for walks with bond-bond intersections.

First, we have computed $\langle\overline{N}\rangle$ as an average over 10^3 uncorrelated configurations for 2D random walks (RWs) with n monomers. Figure 2 compares these results with those for SAWs from Ref. [1]. (Results for RWs and SAWs appear as black and white squares, respectively.) Using polymer chains with $10\leqslant n\leqslant 1000$, the scaling behavior for random walks is found to be

$$\ln\langle\overline{N}\rangle=(0.56\pm0.01)\ln(n-2)-(0.5\pm0.1), \qquad (2)$$

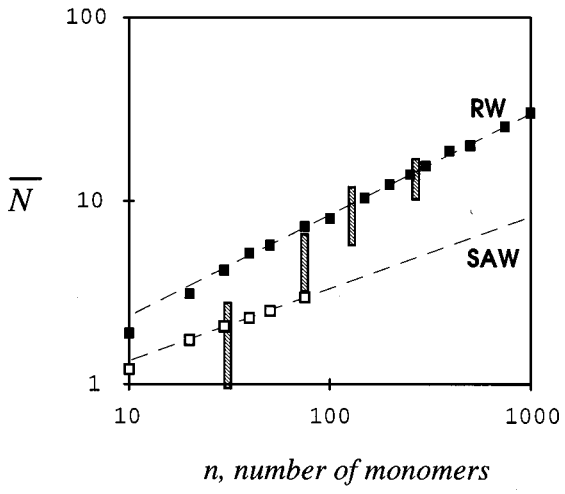

FIG. 2. Scaling behavior of the folding complexity descriptor $\overline{N}$ for a number of model 2D polymers. [Black squares correspond to random walks, RWs, whereas white squares correspond to self-avoiding walks, SAWs. Gray bars correspond to the span of $\overline{N}$ values for the Sammon projections of protein backbones having the following numbers of amino acid residues $n=31$ (10 proteins), $n=75$ (6 proteins), $n=129$ (9 proteins), and $n=269$ (11 proteins). See text for details.]

with 95% confidence intervals and a correlation coefficient $\mathcal{C}=0.9993$. [A fitting to a simpler n^{β} power law gives similar results: $\ln\langle\overline{N}\rangle=(0.58\pm0.02)\ln n-(0.6\pm0.1)$, with $\mathcal{C}=0.9984$. A scaling law $(n-2)^{\beta}$, as in Eq. (2), ensures that the exact result $\overline{N}=0$ holds for a two-node chain.] To assess the quality of our sampling, we have also monitored the mean polymer size. Using the same configurations, we find the following asymptotic behavior for the mean radius of gyration: $\ln\langle R_g^2\rangle^{1/2}=(0.498\pm0.006)\ln n+(0.44\pm0.03)$, $\mathcal{C}=0.9998$. This value is in good agreement with the well-known result for random walks, $\langle R_g^2\rangle^{1/2}=n^{1/2}$. (Note that 2D random walks correspond to polymers in the poor-solvent regime. In contrast, 3D random walks, also characterized by a size exponent $\nu=1/2$, resemble polymers in an ideal solvent.) Our results indicate that the folding complexity for SAWs and RWs is characterized by different critical exponents. Our conservative estimates for the exponent β in 2D walks are (a) for random walks, $\beta_{(\mathrm{RW})}=0.57\pm0.03$, (b) for self-avoiding walks with no excluded volume, $\beta_{(\mathrm{SAW})}=0.40\pm0.05$ [1]. It is worth noting that the difference in β values is significant for 2D walks, whereas the exponents for 3D RWs and SAWs *are virtually coincident* [5]. Present results should motivate further work towards understanding the dimensional dependence (and exact values) of β.

Random and self-avoiding walks provide a reference to compare the behavior of other 2D models of biopolymers. In particular, we are interested in *optimized* 2D projections resulting from multidimensional scaling. These techniques employ a nonlinear mapping to produce a single projection to a lower-dimensional space under the condition of optimally preserving the shape pattern of the initial data set. One such technique is the Sammon algorithm, commonly exploited for data compression and pattern recognition [8,9]. In our case, the Sammon mapping projects a set of n nodes in 3-space, with coordinates $\{\mathbf{X}_i\}$, to another set of n nodes in 2-space with coordinates $\{\mathbf{R}_i\}$. The mapping proceeds by minimizing a ''stress-function'' E that takes into account the pattern of distances in each space. A common choice is [8,9]

$$E=\left\{\sum_{i,j<i} w_{ij}\right\}^{-1}\sum_{i,j<i} w_{ij}^{-1}(X_{ij}-R_{ij})^2, \qquad (3)$$

where $X_{ij}=\|\mathbf{X}_i-\mathbf{X}_j\|$ and $R_{ij}=\|\mathbf{R}_i-\mathbf{R}_j\|$ are the distances in 3- and 2-space, respectively, and $\{w_{ij}\}$ are statistical weights that can be chosen according to various criteria (e.g., $w_{ij}=X_{ij}$). The optimum set of 2D coordinates is generated from an initial set of random positions $\{\mathbf{R}_i^{(0)}\}$. New positions can be generated iteratively by using minimization techniques [8–10]. Here, we use a steepest-descent Sammon algorithm, as implemented by Rauber *et al.* [11], where

$$\mathbf{R}_i^{(s+1)}=\mathbf{R}_i^{(s)}-\alpha\left[\frac{\partial E/\partial\mathbf{R}_i}{|\partial^2 E/\partial\mathbf{R}_i^2|}\right]_{\mathbf{R}_i=\mathbf{R}_i^{(s)}}, \quad s=0,1,2,\ldots, \qquad (4)$$

with α a coupling constant (or ''learning parameter'') taken between 0.15 and 0.45. In our case, we test the robustness of the final set of projected coordinates by repeating the procedure with different initial randomizations of the 2D data and α values. Whenever multiple solutions appear, we choose the one with the lowest min E error. This technique yields a set

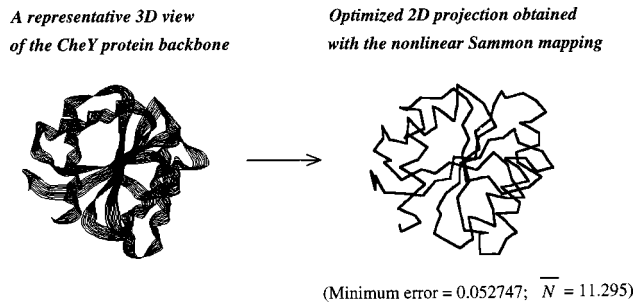

FIG. 3. Example of a 2D nonlinear projection of a protein backbone obtained using the nonlinear Sammon mapping. (This protein has the largest $\overline{N}$ value among all proteins with 129 residues.)

of 2D coordinates $\{\mathbf{R}_i\}$ that can be used, together with the initial chain topology, to compute the shape descriptor in Eq. (1). Until now, this method had only been used as a graphical tool for displaying 3D protein backbones [4]. Here, we have generated the Sammon projections for proteins of various lengths, in order to understand some general properties of the resulting ''optimized'' 2D conformers.

Figure 3 illustrates the results for the backone of CheY protein, containing $n=129$ amino acid residues. [All 3D coordinates correspond to experimental structures deposited in the Protein Data Bank (PDB) [12].] On the right, we show the optimized Sammon projection. On the left, a ribbon trace of the 3D backbone indicates how the projection indeed preserves the overall shape.

Figure 3 gives also the value for the folding descriptor $\overline{N}$ in CheY protein (PDB code 1cye). Among all *distinct* proteins with $n=129$ residues found in the PDB, 1cye has the smallest radius of gyration and the largest $\overline{N}$ value. This result suggests that compactness in 3-space may translate into a maximum folding complexity for the corresponding 2D Sammon projection. In order to analyze the significance of this observation, we have computed the Sammon projections for families of unrelated proteins (with complete backbone coordinates in the PDB) sharing the same chain length. With these data, we have made an estimate of the *range of folding complexity* accessible to small- and medium-size proteins. The following n values provide sets with large diversity in molecular shapes: $n=31$ (10 proteins), $n=75$ (6 proteins), $n=129$ (9 proteins), $n=269$ (11 proteins). For these proteins, we have computed optimized nonlinear 2D projections and shape descriptors. Illustrative cases appear in Fig. 4. In this figure, 3icb and 1tib have the most ''entangled'' folds for proteins with lengths $n=75$ and 269, respectively.

Our main observation appears in Fig. 2, where the shape descriptors for all the proteins considered are contrasted with the results for 2D walks. (The walks in Fig. 2 have a constant bond length of 3.8 Å, in order to allow a proper comparison

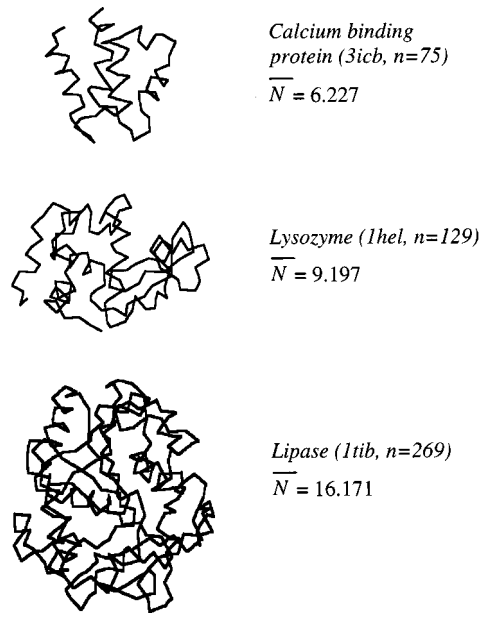

FIG. 4. Illustrative examples of folding complexity in the Sammon projections for proteins with variable length. (Proteins 3icb and 1tib have the largest $\overline{N}$ values among proteins with their same chain length.)

with protein backbones with the same n value.) In Fig. 2, gray bars indicate the span of $\overline{N}$ values over all protein native states with $n=31$, 75, 129, and 269. Our results indicate a clear *crossover behavior* in the folding complexity: (a) short protein chains are projected onto 2D structures that are not maximally compact and exhibit the folding complexity of SAWs with the same length; (b) protein chains with $n>100$ residues are projected onto 2D structures whose folding features are similar to those of compact RWs with the same length.

The above change in folding features may reflect essential differences in the 3D native states; e.g., the fact that longer proteins consist of multiple folding units (''domains''). It is possible that the pattern of inter-residue distances in multidomain proteins can only give rise to 2D nonlinear projections with a large number of self-intersections. Pragmatically, one can exploit this behavior as one more piece of information to be used when designing tentative folds for a given protein. Our results establish constraints to the ranges of $\overline{N}$ values that are consistent with two- and three-dimensional backbones in native states. Accordingly, they provide a criterion to decide whether some folding features are reasonable for the native state of an n-residue protein.

I would like to thank T. W. Rauber (Vitória, Brazil) for making available the program Tooldiag for pattern recognition. This work was supported by grants from NSERC (Canada).

[1] G. A. Arteca and S. Zhang, Phys. Rev. E **59**, 4209 (1999).
[2] H. S. Chan and K. A. Dill, Annu. Rev. Biophys. Biophys. Chem. **20**, 447 (1991); R. E. Bleil, C. F. Wong, and H. Rabitz, J. Phys. Chem. **99**, 3379 (1995); M. S. Li and M. Cieplak, Phys. Rev. E **59**, 970 (1999).
[3] E. Guitter and E. Orlandini, J. Phys. A **32**, 1359 (1999).
[4] T. W. Barlow and W. G. Richards, J. Mol. Struct.: THEOCHEM **398**, 483 (1997).
[5] G. A. Arteca, Phys. Rev. E **49**, 2417 (1994); **51**, 2600 (1995).
[6] A. L. Kholodenko and D. P. Rolfsen, J. Phys. A **29**, 5677

(1996); A. L. Kholodenko and T. A. Vilgis, Phys. Rep. **298**, 251 (1998).

[7] A. Stasiak, V. Katritch, J. Bednar, D. Michoud, and J. Dubochet, Nature (London) **384**, 122 (1996); V. Katritch, J. Bednar, D. Michoud, R. G. Scharein, J. Dubochet, and A. Stasiak, *ibid.* **384**, 142 (1996); A. V. Vologodskii, N. J. Crisona, B. Laurie, P. Pieranski, V. Katritch, J. Dubochet, and A. Stasiak, J. Mol. Biol. **278**, 1 (1998); J. Cantarella, R. B. Kusher, and J. M. Sullivan, Nature (London) **392**, 237 (1998); G. Buck, *ibid.* **392**, 238 (1998).

[8] J. W. Sammon Jr., IEEE Trans. Comput. **C-18**, 401 (1969).

[9] D. de Ridder and R. P. W. Duin, Pattern Recogn. Lett. **18**, 1307 (1997).

[10] T. Kohonen, *Self-organizing Maps* (Springer, Berlin, 1995).

[11] T. W. Rauber, M. M. Barata, and A. S. Steiger-Garção, in *Proceedings of the International Conference on Fault Diagnosis (Tooldiag'93)*, edited by M. Labarrere (Cert-Onera, Toulouse, 1993), Vol. 3, p. 906.

[12] 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**, 2417 (1977).