Proteins and polymers

Jayanth R. Banavar[a)]
Department of Physics, 104 Davey Laboratory, The Pennsylvania State University, University Park, Pennsylvania 16802

Trinh Xuan Hoang
Institute of Physics and Electronics, Vietnamese Academy of Science and Technology, 10 Dao Tan, Hanoi, Vietnam

Amos Maritan
Dipartimento di Fisica 'G. Galilei' and Istituto Nazionale di Física Nuclear (INFN), Università di Padova, Via Marzolo 8, 35131 Padova, Italy

(Received 7 March 2005; accepted 2 May 2005; published online 21 June 2005)

Proteins, chain molecules of amino acids, behave in ways which are similar to each other yet quite distinct from standard compact polymers. We demonstrate that the Flory theorem, derived for polymer melts, holds for compact protein native state structures and is not incompatible with the existence of structured building blocks such as α helices and β strands. We present a discussion on how the notion of the thickness of a polymer chain, besides being useful in describing a chain molecule in the continuum limit, plays a vital role in interpolating between conventional polymer physics and the phase of matter associated with protein structures. © *2005 American Institute of Physics*. [DOI: 10.1063/1.1940059]

Proteins are chain molecules made up of small chemical entities called amino acids. In spite of their small size, the diverse physical and chemical attributes of the twenty types of naturally occurring amino acids and the history-dependent role played by evolution, globular proteins exhibit a range of striking common characteristics.[1] Traditional attempts at creating a framework for understanding proteins using ideas from polymer physics have been largely unsuccessful as stated by Flory.[2] "Synthetic analogs of globular proteins are unknown. The capability of adopting a dense globular configuration stabilized by self-interactions and of transforming reversibly to the random coil are peculiar to the chain molecules of globular proteins alone." The standard models of polymer physics do not provide an explanation for why there are a relatively small number (of order thousand) native state folds,[3] why they are inevitably made up of helices and sheets[4] and how these folds are adapted for biological function especially enzymatic activity.

In this paper, we seek to bridge this apparent gap between polymer physics and the physics of compact biomolecules. We do this in two complementary ways: first, we study the average behavior of compact protein native state structures and show that, in spite of being made up mainly of α helices and β strands, the Flory theorem derived for polymer melts[5,6] holds reasonably well for native state protein structures as well; second, we demonstrate that the notion of an anisotropic chain of nonzero thickness is valuable for extrapolating from conventional polymer physics to the phase used by nature to house protein structures.

Let us begin with an analysis of protein native state structures from the protein data bank[7] to assess the validity of the Flory theorem. We consider a coarse-grained description in which each amino acid is represented by its C^{α} atom, the hinges of the protein backbone. It is well known from Flory's work in polymer physics that polymer melts or even a long compact polymer has very interesting substructure.[5,8,9] The basic idea is that a short labeled piece of a polymer chain from within such a dense melt exhibits statistics (distributions and an end-to-end distance) which are characteristic of random-walk behavior. Physically, the effective absence of any interaction is believed to arise from the inability of the chain to discern whether it is making contacts with itself or with other chains. Does the presumed validity of the Flory theorem and the existence of Gaussian random-walk statistics for short chain segments preclude structures built up from helices and sheets? Interestingly, it has been suggested recently[10] that the model denatured proteins can exhibit random coil statistics in spite of having significant secondary structure.

Our principal results are summarized in Figs. 1–5 and demonstrate that for compact proteins, characterized by an end-to-end distance scaling approximately as the cube root of the protein size (see Fig. 1):

(1) The Flory theorem is found to hold (Fig. 2) for protein segments made up of more than 48 amino acids. The existence of secondary motifs results in an effective persistence length of this order beyond which one obtains Gaussian statistics (Fig. 3) accompanied by random-walk behavior.
(2) The validity of the Flory theorem is *not* incompatible with the existence of secondary motifs.[9]
(3) One can understand the crossover in Fig. 2 by studying correlation functions of the tangent and the binormal vectors along the chain (Figs. 4 and 5).

[a)]Electronic mail: banavar@psu.edu

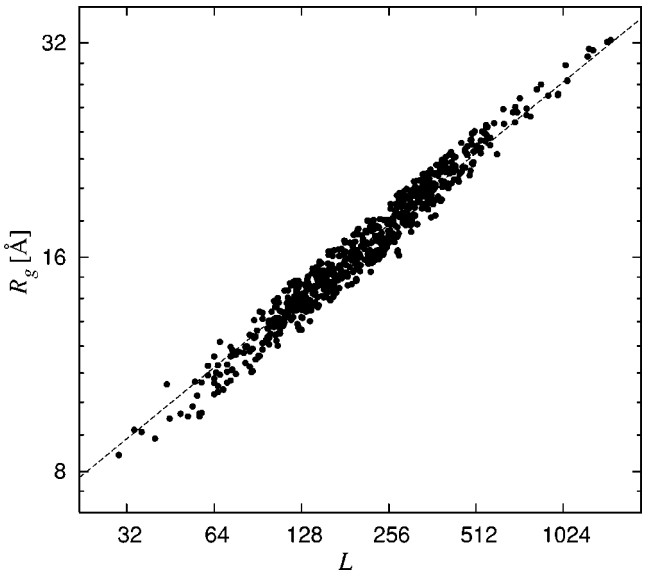

FIG. 1. Log-log plot of the radius of gyration R_g of a set of 700 proteins obtained from the Protein Data Bank (PDB) of (Ref. 11) versus their length L or the number of constituent amino acids.

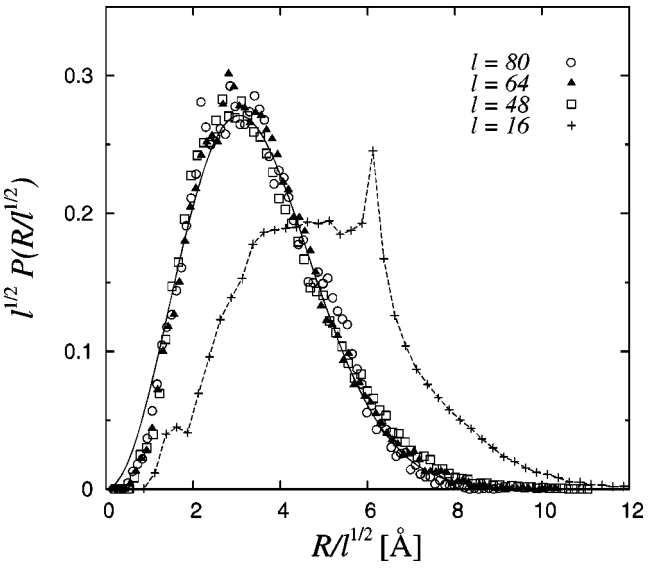

FIG. 3. Statistics of the end-to-end distance of segments of proteins of length l. For $l=48$, 64, and 80, the distributions show a nice collapse to the form expected for Gaussian statistics: the solid line denotes the function $P(x)=1/\sigma^3\sqrt{2/\pi l}x^2\exp(-x^2/2\sigma^2)$, where $\sigma=2.164$ Å. For $l=16$, where the presence of secondary motifs play a major role, the distribution is qualitatively different from the other sizes and exhibits a peak arising from the presence of α helices.

Our results vividly demonstrate that proteins exhibit properties that are not incompatible with those of generic compact polymers. However, as stated before, the standard models of polymer physics do not account for the rich phase of matter associated with protein native state structures. In order to proceed, let us recall that a dominant structural motif used in biomolecular structures is the helix.[12,13] An everyday object which, on compaction, can be coiled naturally and efficiently into a helical shape is a garden hose or a tube.[14] A tube can be thought of as a thick polymer, a polymer chain endowed with a natural thickness. We will proceed to study the attributes of a tube and its relationship with conventional descriptions of polymers.

In the continuum, a nonzero chain thickness serves a valuable purpose. Consider first a polymer chain of vanish-

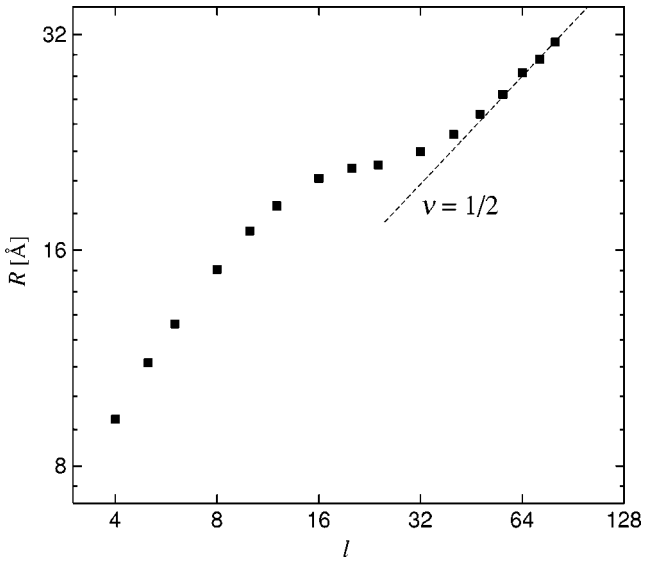

FIG. 2. Log-log plot of the end-to-end distance R versus l for protein segments. The plot was obtained by averaging over all segments of length l selected from the data set depicted in Fig. 1. For a given l, R was determined as an average over all segments of that length in proteins whose lengths are greater than $l^{3/2}$, in order to avoid finite-size effects (Ref. 9). The error bars are of the order of the size of the symbols. Note the plateau which indicates that R is only slowly increasing with l around 24. For values of l larger than 48, we find that $R\sim l^{1/2}$.

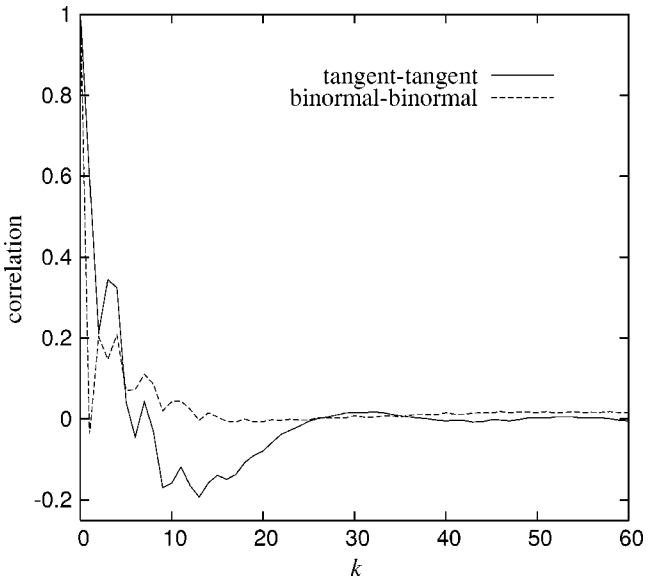

FIG. 4. Plot of the tangent-tangent and binormal-binormal correlation functions along the protein sequence derived from our data set. The tangent vector at location i is defined as an unit vector pointing along the line joining the positions of the $i-1$th and the $i+1$th amino acids. The normal vector is defined by joining the ith location to the center of the circle drawn through three amino acid $(i-1,i,i+1)$ locations. The binormal is perpendicular to the plane defined by the tangent and the normal. Note that: (a) the negative tangent-tangent correlation at sequence separation k around 13 corresponds to a turning back, on average, of the chain direction and is related to the crossover shown in Fig. 2; (b) the binormal-binormal correlation remains nonzero for large separations.

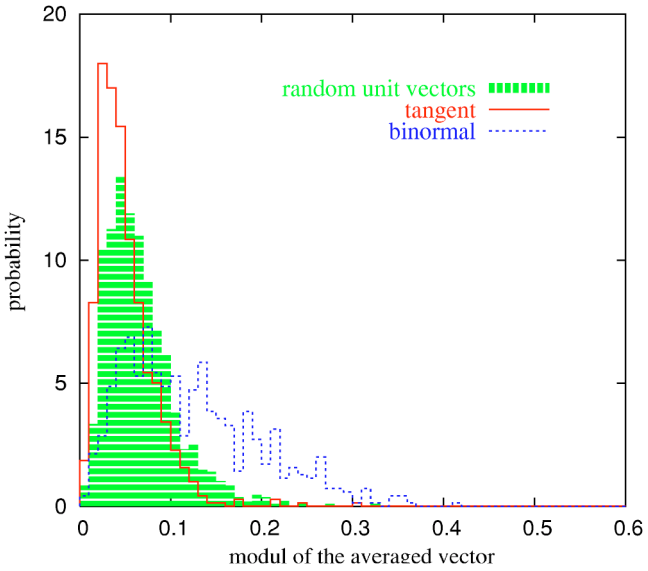

FIG. 5. Histogram of the magnitudes of the average tangent and binormal vectors for each protein in our data set. For each protein, we measured the magnitude as $1/N|\Sigma_{i=1}^{N}\mathbf{v}_i|$, where $\mathbf{v}_i$ is either the unit tangent or the unit binormal vector at location i and N is the number of such vectors for a given protein. For comparison, a histogram of the magnitudes of the average of randomly oriented vectors is shown as the shaded histogram. (Here $\mathbf{v}_i$ was selected to be a randomly oriented unit vector.) Note that several proteins have a significant nonzero mean binormal vector due to the presence of α helices.

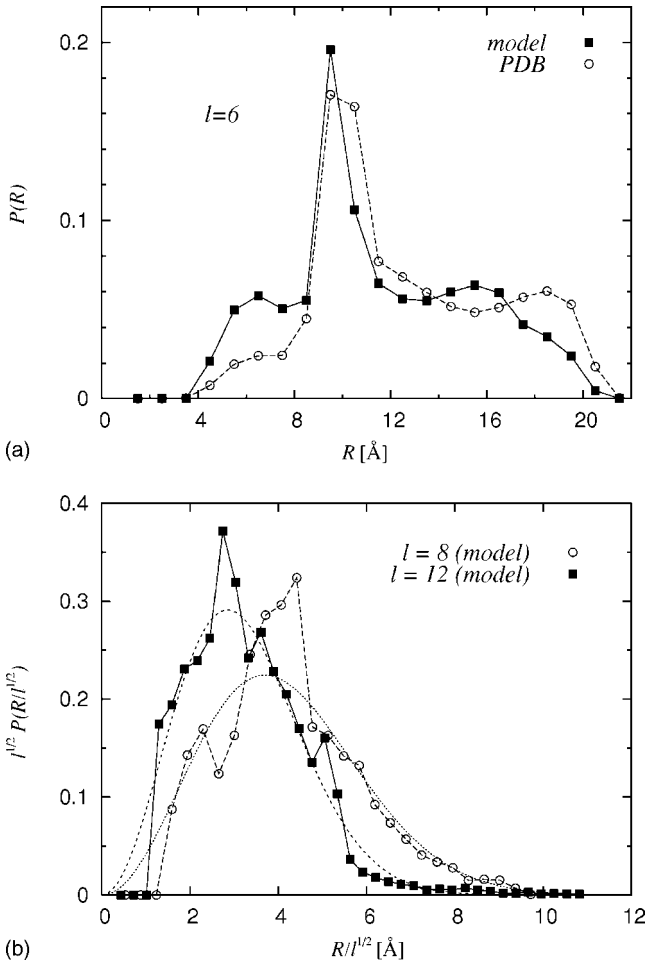

FIG. 6. (a) Statistics of the end-to-end distance of segments of length $l=6$ taken from model protein structures (Ref. 18) and from PDB structures. The peak in the distributions arises from the presence of α helices. (b) Same as Fig. 3 but for segments of the model structures of lengths $l=8$ and 12. The fits to the Gaussian form given in the caption of Fig. 3 yield $\sigma=2.61$ Å for $l=8$ and $\sigma=2.08$ Å for $l=12$.

ing thickness in the continuum. It is well known[5] that the end-to-end distance R of a swollen, self-avoiding chain scales approximately as the third/fifth power of its length L. In the absence of any other length scale in the problem (recall that we are dealing with a chain of zero thickness in the continuum), one is led to a fundamental problem in simple dimensional analysis in expressing the relationship $R\sim L^{0.6}$—both R and L have units of length and there is no other length scale in the problem which can be used to fix the correct dimension in the scaling relation. In order to study a chain molecule in the continuum, the traditional approach has been to use the powerful machinery of renormalization group theory.[15] A tube of nonzero thickness circumvents this problem by providing the required additional length scale naturally, even in the continuum. Indeed, one may write a scaling form $R(L,b,\Delta)=LF(L/\Delta,b/\Delta)$, where Δ is the tube thickness. The continuum limit can be safely taken by letting b go to 0 leading to $R=LF(L/\Delta,0)\sim\Delta^{1-\nu}L^{\nu}$.

An interesting issue in polymer physics is the description, in the continuum, of a closed chain with certain knot topologies. One, of course, requires physically that the knot number be preserved in any dynamics. A string described by in standard continuum approach is necessarily characterized by an infinitesimal thickness and allows changes in the knot topology with a finite-energy cost rendering the model somewhat unphysical in this regard. This problem is cured by the tube description. Hard spheres have been studied for centuries and their self-avoidance is ensured by considering all pairs of spheres and requiring that their centers are no closer than the sphere diameter. Strikingly, the generalization of this result to a tube entails a simple modification of the standard pairwise interactions.[16] For each pair of points along the tube axis, one draws two circles both passing through the two points and each one tangential to the axis at one or the other location. One then simply requires that none of the radii is smaller than the tube radius.[16,17] The use of many-body potentials is an essential ingredient for describing a tube in the continuum.[16] The many-body potential replaces the pairwise self-interaction potential and ought not to be thought of as a higher-order correction.

The coarse-grained flexible tube model captures two essential ingredients of proteins—the space within a tube roughly allows for the packing of the protein atoms and local steric effects are encapsulated by constraints on the local radius of curvature; the effects of the geometrical constraints imposed by the chemistry of backbone hydrogen bonds are represented by the inherent anisotropy of a tube (a tube, when discretized, may be imagined to be a chain of disks). The generic compact polymer phase arises for long tubes with a thickness much smaller than the range of attractive interactions promoting compaction.

Recent work[18] has shown that the low-energy conformations adopted by tubelike polymers with certain constraints on symmetry and geometry are made up of helices and

sheets akin to marginally compact protein secondary structures. For classes of short homopolymers characterized by generic geometrical constraints arising from backbone hydrogen bonds and sterics and with mild variations in their overall hydrophobicity and local curvature energy penalty parameters, one obtains a free-energy landscape,[18] determined by geometry and symmetry, with multiple minima corresponding to the menu of folds. We have generated a thousand structures with low energies of a homopolymer of length $N=48$. The structures are local energy minima in simulated annealing simulations. A refined set of about 320 proteinlike structures is obtained by choosing only those that are marginally compact ($7.6\ \text{Å} < R_g < 12\ \text{Å}$) and have a sufficient amount of secondary structure content (the fraction of residues participating in either a helix or a sheet is larger than 60% of the total number of residues). Strikingly, Fig. 6(a) shows that the behavior of short segments of real proteins and the model structures are qualitatively similar to each other. The deviation from Gaussian behavior in both cases is due to the presence of secondary structures, whose characteristic length scale is smaller for the model structures than for real proteins. Interestingly, even for relatively short segment lengths ($l=8,12$) in the model structures, one observes statistical behavior somewhat similar to that of Gaussian chains [Fig. 6(b)] along with significant deviations, most notably a peak due to the presence of the secondary structures. Due to the limited chain length that one can reliably study in the model we are not able to observe the crossover to the regime predicted by Flory.

In summary, we have shown that there is a natural bridge, provided by the chain thickness, between polymer physics and the physics of biomolecular structures. The thickness provides a physically motivated cut-off length scale which allows for a well-defined continuum limit. The Flory theorem is found to hold for proteins in spite of the structured building blocks of protein native state structures. Our results suggest that the powerful arsenal of techniques of polymer physics can be brought to bear on the protein problem and conversely, the notion that the chain molecules are inherently anisotropic and have a nonzero thickness provide a new perspective in the field of polymer physics.

This work was supported by PRIN 2003, INFN, NASA, NSF IGERT Grant No. DGE-9987589, NSF MRSEC at Pennsylvania State, and the NSC of Vietnam (Grant No. 410704).

[1] J. R. Banavar and A. Maritan, Rev. Mod. Phys. **75**, 23 (2003).
[2] P. J. Flory, *Statistical Mechanics of Chain Molecules* (Wiley, New York, 1969).
[3] C. Chothia, Nature (London) **357**, 543 (1992).
[4] T. E. Creighton, *Proteins, Structure, and Molecular Properties*, 2nd ed. (Freeman, New York, 1993).
[5] P. J. Flory, *Principles of Polymer Chemistry* (Cornell University Press, Ithaca, 1953); A. Y. Grosberg and A. R. Khokhlov, *Statistical Physics of Macromolecules* (AIP, New York, 1994); M. Rubinstein and R. Colby, *Polymer Physics* (Oxford University Press, New York, 2003); C. Vanderzande, *Lattice Models of Polymers* (Cambridge Univeristy Press, Cambridge, 1998).
[6] H. Orland, J. Phys. I **4**, 101 (1994).
[7] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, Nucleic Acids Res. **28**, 235 (2000).
[8] E. N. Govorun, V. A. Ivanov, A. R. Khokhlov, P. G. Khalatur, A. L. Borovinsky, and A. Y. Grosberg, Phys. Rev. E **64**, 040903 (2001).
[9] R. Lua, A. L. Borovinsky, and A. Y. Grosberg, Polymer **45**, 717 (2004).
[10] N. C. Fitzkee and G. D. Rose, Proc. Natl. Acad. Sci. U.S.A. **101**, 12497(2000).
[11] H. M. Berman, J. Westerbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, Nucleic Acids Res. **28**, 235 (2000).
[12] J. D. Watson and F. H. C. Crick, Nature (London) **171**, 737 (1953).
[13] L. Pauling, R. B. Corey, and H. R. Branson, Proc. Natl. Acad. Sci. U.S.A. **37**, 205 (1951).
[14] A. Maritan, C. Micheletti, A. Trovato, and J. R. Banavar, Nature (London) **406**, 287 (2000); A. Stasiak and J. H. Maddocks, *ibid.* **406**, 251 (2000).
[15] K. G. Wilson, Rev. Mod. Phys. **55**, 583 (1983).
[16] J. R. Banavar, O. Gonzalez, J. H. Maddocks, and A. Maritan, J. Stat. Phys. **110**, 35 (2003).
[17] O. Gonzalez and J. H. Maddocks, Proc. Natl. Acad. Sci. U.S.A. **96**, 4769 (1999).
[18] T. X. Hoang, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan, Proc. Natl. Acad. Sci. U.S.A. **101**, 7960 (2004).