Geometry of proteins: Hydrogen bonding, sterics, and marginally compact tubes

Jayanth R. Banavar
Department of Physics, 104 Davey Laboratory, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

Marek Cieplak
Institute of Physics, Polish Academy of Sciences, 02-668 Warsaw, Poland

Alessandro Flammini
School of Informatics, Indiana University, Bloomington, Indiana 47408, USA

Trinh X. Hoang
Institute of Physics and Electronics, VAST, 10 Dao Tan, Hanoi, Vietnam

Randall D. Kamien
Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

Timothy Lezon
Department of Physics, 104 Davey Laboratory, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

Davide Marenduzzo
Mathematics Institute, University of Warwick, Coventry CV4 7AL, England

Amos Maritan and Flavio Seno
Dipartimento di Fisica 'G. Galilei', Università di Padova, Via Marzolo 8, 35131 Padova, Italy
and Sezione INFN, Università di Padova, I-35131 Padova, Italy

Yehuda Snir
Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

Antonio Trovato
Dipartimento di Fisica "G. Galilei", Sezione INFN, Università di Padova, I-35131 Padova, Italy
(Received 11 July 2005; revised manuscript received 3 February 2006; published 22 March 2006)

The functionality of proteins is governed by their structure in the native state. Protein structures are made up of emergent building blocks of helices and almost planar sheets. A simple coarse-grained geometrical model of a flexible tube barely subject to compaction provides a unified framework for understanding the common character of globular proteins. We argue that a recent critique of the tube idea is not well founded.

DOI: 10.1103/PhysRevE.73.031921 PACS number(s): 87.14.Ee, 82.35.Lr, 87.15.Aa, 87.15.Cc

The protein problem [1–3] is one of formidable complexity. The number of degrees of freedom of the protein atoms as well as the surrounding water molecules, which play an essential role in the folding process, is enormous. In addition, a protein chain is relatively short compared to macromolecular polymer chains and one might therefore expect significant nonuniversal behavior with the details mattering a great deal. Furthermore, the sequences of proteins have been subject to evolution and natural selection, a history dependent process. Yet there are striking patterns that one observes in protein behavior.

All proteins fold rapidly and reproducibly [4] and their native state structures are made of common building blocks: helices and zig-zag strands assembled into almost planar sheets. For globular proteins to serve vital enzymatic roles, their folded structures need to be flexible. The total number of distinct folds adopted by globular proteins is only of the order of a few thousand [5], a remarkably small number compared to the profusion of structures one might have expected for compact chains comprising a few hundred monomers. Furthermore, it is believed that the folds are evolutionarily conserved [6,7]. Many protein sequences adopt the same native state conformation [8]. Once a sequence has selected its native state structure, it is able to tolerate a significant degree of mutability except at certain key locations [9].

It has been suggested that these common attributes of globular proteins [10–13] reflect a deeper underlying unity in their behavior. Yet, a protein molecule along with the surrounding water molecules constitutes a system of great complexity. Such a system can be described at many levels. At the finest level, one would simply treat the entire system with all the degrees of freedom with the laws of quantum mechanics. The difficulties associated with a first-principles quantum mechanical approach include the large number of degrees of freedom; the necessity of calculating the interactions during

the dynamical process of folding, with the solvent taken into account in an accurate manner; and, even if the interactions were known exactly, the limitations of present-day computers in being able to accurately follow the dynamics through the folding process. Understanding such a system at this level of description is a daunting task and has not yet been achieved.

Any alternative coarse-graining procedure implies the determination of effective interactions that are postulated to arise on integrating out the degrees of freedom of the water. For example, Pitard *et al.* [14] have studied the folding and anisotropic collapse of a microscopic continuous model of a homopolymer chain where each monomer carries a dipole moment. In an equilibrium description of any such coarse-grained model, the effective potential not only depends on the protein conformation as represented by the values of the coordinates of the atoms of the protein but is also a function of the temperature. The averaging is envisioned to be carried out under the assumption of an instantaneous equilibration of the fine details represented by the coordinates of the water molecules. However, the folding of a protein is not an equilibrium situation but entails dynamical processes that cannot be captured within an equilibrium description.

The helix is a natural, compact conformation of a short, flexible tube. This motivated us [12,13,15–19] to investigate the phase behavior of a flexible tube subject to compaction in order to investigate whether it is related to and can explain protein behavior. The tube is anisotropic and may be thought of as the continuum limit of a discrete chain of disks or coins. Unlike a chain of spheres, a chain of coins accurately captures the symmetry of a chain molecule because associated with each object along the chain is a special local axis defined by the tangent to the chain and represented by the axis perpendicular to the face of the disk. The amino acids have sidechains which stick out in a direction lying approximately in the plane of the disk. Unlike an ordinary garden hose, the tube is one in which each disk orients itself in such a way that the sidechain sticks out at an angle of around 143° from the normal vector [20] joining the disk center to the center of the circle passing through the center of the disk and the centers of its two adjacent neighbors. The tube model does not arise from an integration of some of the degrees of freedom of a microscopic model.

For a short discrete tube, with less than 20 residues (with the same bond length and typical thickness of a polypeptide chain), helices and planar hairpins and sheets are found to be the preferred structures in a marginally compact phase in which the attractive forces promoting compaction barely set in. This is due to the self tuning of two key length scales, the thickness of the tube and the interaction range between the centers of the disks, to be comparable to each other. When the tube thickness is much larger than the interaction range, one cannot avail of the attractive interaction and one obtains a highly degenerate swollen phase. In the other extreme in which the tube thickness is much smaller than the interaction range, one obtains a highly degenerate compact phase—there is a great deal of flexibility in the relative placement of nearby tube segments. The marginally compact phase opens up in the vicinity of the phase transition between these two phases, when the two length scales become comparable to each other. In the marginally compact phase, there is a great reduction in the degeneracy of the ground state structures with a requirement that nearby tube segments be right alongside and parallel to each other.

Two basic requirements must be met by neighboring tube segments in the marginally compact phase in order for them to maximally avail of the attraction that has barely set in. First, the anisotropy of a tube requires that neighboring tube segments be parallel to each other rather than be perpendicular and consequently progressively separating from each other. Second, because the range is such that the attraction has just set in, it is crucial that neighboring segments not only be approximately parallel to each other but right alongside each other. A simple way of understanding how a protein is automatically poised to be in the marginally compact phase is by noting that hydrophobicity, which drives the self attraction of a tube, requires that the buried area associated with the tube be as large as possible. This drive ensures that neighboring tube segments are placed right next to each other to facilitate effective screening of the water.

The α helix is tightly packed with the main chain atoms fitting snugly within the helix. Likewise, in a sheet, the space between neighboring strands is occupied by the main chain atoms. In both cases, the scaffolding is provided by hydrogen bonds between the *N-H* group of one amino acid and the $C=0$ group of another. Both the tube size and the range of the interaction are governed by the geometry of the protein determined by quantum chemistry and more specifically the locations of the main chain atoms. The amazingly perfect fit of the quantum chemistry, e.g., the planarity of the peptide bond and the lengths of the covalent and hydrogen bonds, to the structures in the marginally compact phase is especially noteworthy.

This simple tube model is closely related to the seminal contributions of Pauling [21–23] and Ramachandran [24]. Both of them considered the protein backbone which is the common part of all proteins. Pauling and his co-workers explored the types of structures that are consistent with both the backbone geometry and the formation of hydrogen bonds, which would then provide the scaffolding for such structures. They predicted that helices and sheets are the structures of choice in this regard. Ramachandran and his co-workers considered the role of excluded volume or steric interactions between nearby amino acids along the sequence in reducing the available conformational phase space (see Ref. [25] for a recent assessment of such effects on longer sequence stretches and Ref. [26] for a discussion of steric restrictions in protein folding). Astonishingly, the two significantly populated regions of the Ramachandran plot correspond to the α helix and the β strand. Even though backbone hydrogen bonds and steric constraints are not related to each other, they are both promoters of helices and sheets. One might ask whether this concurrence of events is a mere accident. The results from the simple tube model provide a clue that the answer might be negative suggesting that proteins, which obey physical law, may have been selected to conform to the tube geometry through steric interactions between nearby amino acids along the sequence and hydrogen bonds between backbone atoms. Hydrogen bonds serve to enforce the parallelism of nearby tube segments [27], a feature of

both helices and sheets while steric constraints emphasize the non-zero thickness of the tube.

A more refined tube model [12,13] was subsequently introduced by incorporating the geometrical constraints of backbone hydrogen bonds and a local bending energy penalty term. In its simplest form, the model describes the hompolymer character of the main backbone chain. At odds with conventional belief, it was suggested that the gross features of the energy landscape of proteins result from the amino acid aspecific common features of all proteins and that protein structures lie in a marginally compact phase, analogous to the simple tube model. This landscape is *(pre)sculpted* by general considerations of geometry and symmetry and has around a thousand broad minima corresponding to putative native state structures. For each of these minima, the desirable funnel-like behavior [28] is already achieved at the *homopolymer* level. The interplay of the three energy scales, hydrophobic, hydrogen bond, and bending energy, stabilizes marginally compact structures, and also provides the close cooperation between energy gain and entropy loss needed for the sculpting of a funneled energy landscape. Further, the marginally compact phase is poised in the vicinity of a phase transition to the swollen phase and confers exquisite sensitivity to the structures within the phase [13].

In a recent manuscript, Hubner and Shakhnovich [29] (HS) have presented a critique of the tube model. They state: "The tube model predicts that geometrical and topological factors alone, without inclusion of more chemically detailed hydrogen bonding interactions, determine global features of protein folds such as protein-like secondary structure." They then make the premise: "Therefore, if tube models have implications for real proteins, one would expect similar formation, upon collapse, of helices and secondary structure motifs in a model that accurately represented the geometric and topological properties of amino acid chain in terms of excluded volume and torsional degrees of freedom (as opposed to a featureless tube), but is devoid of explicit hydrogen bonding." This expectation is unfounded, since the simple tube model does predict the emergence of secondary structure (helices and sheets) in the absence of explicit hydrogen bonding for very short chains. While the "compaction of a realistic protein chain model without consideration of hydrogen bonding does not necessarily result in helical geometries" [29], excluded volume and packing of a short tube are sufficient to understand the emergence of proteinlike secondary structure. Furthermore, in Ref. [30] there was no attempt made to explain the existence of β sheets by invoking "a change in the relative sizes of the solvent and tube," but rather the results of the numerics were described in terms of common folding motifs.

Let us consider the coarse-graining description of HS, in which protein coordinates representing all atoms are represented as impenetrable hard spheres of physical radii and the degrees of freedom associated with the water molecules are subsumed in a knowledge-based atomic interaction potential consisting of weak nondirectional van der Waals interactions and stronger hydrogen bonds which are highly dependent upon geometry. This representation of treating atoms as hard spheres and replacing the quantum mechanics with effective classical potentials is a coarse-graining which only works as long as the essential ingredients underlying the system are captured adequately. What HS demonstrate is that, in their model system, classical potentials mimicking directional hydrogen bond formation and van der Waals effects promoting overall compaction lead to parts of the sequence folding into helices. It is then not surprising that throwing away the hydrogen bonds and retaining just the van der Waals interactions leads to no helix formation in the HS model [29]. This result merely suggests that at this scale of description, and for chain lengths considered by HS, the directional hydrogen bonds play a key role.

A short self avoiding tube subject to a self attraction promoting compaction, in its marginally compact phase, curls up into a helix with a specific pitch to radius ratio [15,30] close to that observed in real protein helices and also forms zig-zag strands which assemble into almost planar sheets [16–18]. Interestingly, this model, which is sufficient for understanding individual secondary motifs of a protein, does not require the incorporation of any classical potential mimicking hydrogen bond formation as in the HS model. The directionality of the hydrogen bonds is crudely captured by the inherent anisotropy of a tube. Because the simplest description of any chain molecule is effectively that of a tube, this result applies to *any generic* polymer chain, provided it is poised in the marginally compact part of the phase diagram. It is interesting to note that synthetic oligomers have been shown to fold into helices without the presence of hydrogen bonds [31].

The emergence of a proteinlike secondary structure without the need of explicit hydrogen bonds, for short chains within the context of the simple tube model, does *not* imply, however, that we "challenge the view that hydrogen bonding plays an important role in protein structure," as stated by HS. The simple tube model, which describes a generic polymer chain, needs to be refined in order to capture the properties of a polypeptide chain. A more realistic yet still simple geometrical model considers amino acid aspecific geometrical constraints arising from the chemistry of hydrogen bonds and steric effects and leads to assembled tertiary structures even for a chain consisting of just one type of amino acid [12,13]. It has been shown that this refined model provides behavior in remarkable accord with that of proteins. The marginally compact phase within this model also provides a simple explanation for the generic formation of amyloid [32], and elucidates the role of sequence design in promoting the fitness of proteins in the environment of cell products and it shows how the limited menu of geometrically determined folds act as targets of natural selection [13].

Let us discuss some familiar phases of matter—the fluid phase, the crystal phase, and the liquid crystal phase. The simplest way to understand the fluid and crystal phases is by means of a system of hard spheres [33]. Note that the hard sphere description in this context or, for that matter, in the HS model is itself an emergent property [34]. At low densities one obtains a fluid phase, whereas at higher packing fractions one obtains crystalline order. Liquid crystal phases [35] arise when the objects making up the material are no longer isotropic. Consider the formation of smectic liquid crystals. Though Onsager showed that long enough rods will, in general, form nematic phases independent of their precise

geometry, the same is not true for smectics. Indeed, spherocylinders undergo a nematic-to-smectic phase transition at high enough density [36] whereas ellipsoids do not seem to form smectics at any density [37]. Again, the fact that the latter does not form the smectic phase is not indicative of the failure of excluded volume to predict and control liquid crystalline phases; rather, it highlights the sensitivity to the details of the specific model, just as the HS model shows that removal of the hydrogen bonds destroys the tendency to form helices.

Consider the sodium chloride structure adopted by ionic crystals such as NaCl, LiCl, KBr, and AgCl. The NaCl structure is a fcc arrangement for the Cl ions with the sodium ions occupying the octahedral holes. Let us consider how the structure of the Cl ions may be determined. One can do a very careful quantum mechanical calculation and show that this fcc structure arises from considerations of electrovalent bonding. Alternatively, following the pioneering work of Kepler [38] or the everyday experience of grocers, one realizes that a collection of spherical cannonballs or apples are best packed in a fcc lattice. One may then be emboldened to suggest that considerations of packing, periodicity and the correct symmetry (note that a packing of cubes instead of spheres would not lead to a fcc lattice but rather a simple cubic lattice) are the essential ingredients that determine the menu of possible crystal structures. In other words, the essential elements underlying the fcc structure are not the details of the interatomic interactions or even the quantum mechanics which describes the interactions of all matter but rather the considerations of geometry and symmetry. It is of course remarkable that nature has found such a perfect fit between the quantum interactions in NaCl and the fcc structure.

The HS exercise has a simple analogy. Let us say that a claim was made that close packing of spheres leads to a fcc structure without invoking charges and electrovalent bonding. Consider now doing a calculation with effective potential energies of interaction incorporating the electrovalent interactions on a microscopic model of the Cl ions and finding that one recovers the fcc lattice structure correctly. This would suggest that the model studied has enough features to produce the right answer. Let us then imagine that on leaving out the electrostatic interactions, one finds in this model that the structure is no longer fcc. Would one conclude from this observation that the original claim that close packing of spheres leads to a fcc structure is wrong? Of course not. Such a result would merely serve to show that, in the model being studied, the electrovalent interactions were important to get the right result. Indeed, it is well known that the structure of NaCl at the atomic level is in fact described by electrovalent interactions. Back to the protein context, the importance of hydrogen bonds in determining protein structure has been recognized for more than five decades. The HS finding was contained in a statement in Hoang *et al.* [12], "Our work here underscores the importance of hydrogen bonds in stabilizing both helices and sheets simultaneously (without any need for adjustment of the tube thickness) allowing the formation of tertiary arrangements of secondary motifs. Indeed, the fine-tuning of the hydrogen bond and the hydrophobic interaction is of paramount importance in the selection of the marginally compact region of the phase diagram in which protein native folds are found." The utility of the tube paradigm arises from its ability, in the marginally compact phase, to capture the essential ingredients underlying helix and sheet formation.

Consider a theoretical challenge of determining the crystal structure for a material such as NaCl. One route would be to study the quantum chemistry of the material in detail and calculate from first principles that the correct structure is a face-centered-cubic crystal. Alternatively, one might opt to first catalog the list of possible structures based on considerations of space filling and translational symmetry and then select the best fit structure from this list. The key point is that the structure transcends the chemical housed in it and is determined by the overarching constraints of geometry and symmetry. The fact that many protein sequences adopt the same fold and that the menu of possible folds is limited [39] strongly suggest that similar considerations may be at play here as well even though proteins are neither infinite in extent nor periodic. The close packing of a flexible tube *in the marginally compact phase* is then the analog of the grocer's packing of apples for this problem.

In conclusion, we believe that the results of the HS analysis do *not* disprove the tube idea.

ACKNOWLEDGMENTS

We thank P. De Los Rios for enlightening discussion. This work was supported by the 2P03B-032-25 (Ministry of Science and Informatics, Poland), PRIN MURST 2003, NASA, NSF IGERT Grant No. DGE-9987589 and the NSF MRSEC at Penn State.

[1] T. E. Creighton, *Proteins: Structures and Molecular Properties* (W.H. Freeman and Company, New York, 1993).
[2] A. Fersht, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (W.H. Freeman and Company, New York, 1999).
[3] A. V. Finkelstein and O. Ptistyn, *Protein Physics: A Course of Lectures* (Academic Press, New York, 2002).
[4] C. B. Anfinsen, Science **181**, 223 (1973).
[5] C. Chothia, Nature (London) **357**, 543 (1992).
[6] M. Denton and C. Marshall, Nature (London) **410**, 417 (2001).
[7] C. Chothia, J. Gough, C. Vogel, and S. A. Teichmann, Science **300**, 1701 (2003).
[8] J. U. Bowie, J. F. Reidhaar-Olson, W. A. Lim, and R. T. Sauer, Science **247**, 1306 (1990); W. A. Lim and R. T. Sauer, J. Mol. Biol. **219**, 359 (1991); D. W. Heinz, W. A. Baase, and B. W. Matthews, Proc. Natl. Acad. Sci. U.S.A. **89**, 3751 (1992); B. W. Matthews, Annu. Rev. Biochem. **62**, 139 (1993).
[9] J. S. Richardson and D. C. Richardson, Trends Biochem. Sci.

14, 304 (1989); W. F. DeGrado, Z. R. Wasserman, and J. D. Lear, Science **243**, 622 (1989); M. H. Hecht, J. S. Richardson, D. C. Richardson, and R. C. Ogden, *ibid.* **249**, 884 (1990); C. P. Hill, D. H. Anderson, L. Wesson, W. F. DeGrado, and D. Eisenberg, *ibid.* **249**, 343 (1990); C. Sander and R. Schneider, Proteins **9**, 56 (1991); S. Kamtekar, J. M. Schiffer, H. Y. Xiong, J. M. Babik, and M. H. Hecht, Science **262**, 1680 (1993); A. P. Brunet *et al.*, Nature (London) **364**, 355 (1993); A. R. Davidson and R. T. Sauer, Proc. Natl. Acad. Sci. U.S.A. **91**, 2146 (1994); M. W. West, W. X. Wang, J. Patterson, D. J. Mancia, J. R. Beasley, and M. H. Hecht, *ibid.* **96**, 11211 (1999); Y. Wei, S. Kim, D. Fela, J. Baum, and M. H. Hecht, *ibid.* **100**, 13270 (2003).

[10] J. D. Bernal, Nature (London) **143**, 663 (1939).

[11] J. R. Banavar and A. Maritan, Rev. Mod. Phys. **75**, 23 (2003).

[12] T. X. Hoang, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan, Proc. Natl. Acad. Sci. U.S.A. **101**, 7960 (2004).

[13] J. R. Banavar, T. X. Hoang, A. Maritan, F. Seno, and A. Trovato, Phys. Rev. E **70**, 041905 (2004).

[14] E. Pitard, T. Garel, and H. Orland, J. Phys. I **7**, 1201 (1997).

[15] A. Maritan, C. Micheletti, A. Trovato, and J. R. Banavar, Nature (London) **406**, 287 (2000).

[16] J. R. Banavar, A. Maritan, C. Micheletti, and A. Trovato, Proteins **47**, 315 (2002).

[17] J. R. Banavar, A. Flammini, D. Marenduzzo, A. Maritan, and A. Trovato, ComPlexUs **1**, 4 (2003).

[18] D. Marenduzzo, A. Flammini, A. Trovato, J. R. Banavar, and A. Maritan, J. Polym. Sci., Part B: Polym. Phys. **43**, 650 (2005).

[19] T. Lezon, J. R. Banavar, and A. Maritan, J. Phys.: Condens. Matter **18**, 847 (2006).

[20] B. Park and M. Levitt, J. Mol. Biol. **258**, 367 (1996).

[21] L. Pauling, R. B. Corey, and H. R. Branson, Proc. Natl. Acad. Sci. U.S.A. **37**, 205 (1951).

[22] L. Pauling and R. B. Corey, Proc. Natl. Acad. Sci. U.S.A. **37**, 729 (1951).

[23] D. Eisenberg, Proc. Natl. Acad. Sci. U.S.A. **100**, 11207 (2003).

[24] G. N. Ramachandran and V. Sasisekharan, Adv. Protein Chem. **23**, 283 (1968).

[25] R. V. Pappu, R. Srinivasan, and G. D. Rose, Proc. Natl. Acad. Sci. U.S.A. **97**, 12565 (2000).

[26] N. C. Fitzkee and G. D. Rose, Protein Sci. **13**, 633 (2004).

[27] J. R. Banavar, M. Cieplak, and A. Maritan, Phys. Rev. Lett. **93**, 238101 (2004).

[28] P. E. Leopold, M. Montal, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **89**, 8271 (1992); P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, Science **267**, 1619 (1995); K. A. Dill and H. S. Chan, Nat. Struct. Biol. **4**, 10 (1997).

[29] I. A. Hubner and E. I. Shakhnovich, Phys. Rev. E **72**, 022901 (2005).

[30] Y. Snir and R. D. Kamien, Science **307**, 1067 (2005).

[31] J. C. Nelson, J. G. Saven, J. S. Moore, and P. G. Wolynes, Science **277**, 1793 (1997); D. J. Hill, M. J. Mio, R. B. Prince, T. S. Hughes, and J. S. Moore, Chem. Rev. (Washington, D.C.) **101**, 3893 (2001); T. Nakano and Y. Okamoto, *ibid.* **101**, 4013 (2001).

[32] C. M. Dobson, Nat. Rev. Drug Discovery **2**, 154 (2003); D. M. Fowler *et al.*, PLoS Biol. **4**, 1 (2006).

[33] P. M. Chaikin and T. C. Lubensky, *Principles of Condensed Matter Physics* (Cambridge University Press, Cambridge, 2000).

[34] R. B. Laughlin, *A Different Universe: Reinventing Physics from the Bottom Down* (Basic Books, New York, 2005).

[35] P. G. de Gennes and J. Prost, *The Physics of Liquid Crystals* (Oxford University Press, NewYork, 1995); S. Chandrasekhar, *Liquid Crystals* (Cambridge University Press, Cambridge, 1977).

[36] A. Stroobants, H. N. W. Lekkerkerker, and D. Frenkel, Phys. Rev. Lett. **57**, 1452 (1986).

[37] D. Frenkel, B. M. Mulder, and J. P. McTague, Phys. Rev. Lett. **52**, 287 (1984).

[38] G. G. Szpiro, *Kepler's Conjecture: How Some of the Greatest Minds in History Helped Solve One of the Oldest Math Problems in the World* (Wiley, Hoboken, NJ, 2003).

[39] D. T. Jones, W. R. Taylor, and J. M. Thornton, Nature (London) **358**, 86 (1992).