Unified perspective on proteins: A physics approach

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

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

Amos Maritan, Flavio Seno, and Antonio Trovato
INFM and Dipartimento di Fisica "G. Galilei," Università di Padova, Via Marzolo 8, 35131 Padova, Italy
(Received 9 April 2004; published 27 October 2004)

We study a physical system which, while devoid of the complexity one usually associates with proteins, nevertheless displays a remarkable array of proteinlike properties. The constructive hypothesis that this striking resemblance is not accidental not only leads to a unified framework for understanding protein folding, amyloid formation, and protein interactions but also has implications for natural selection.

DOI: 10.1103/PhysRevE.70.041905 PACS number(s): 87.14.Ee, 82.35.Lr, 87.15.Aa, 61.41.+e

I. INTRODUCTION

The revolution in molecular biology [1] sparked by the discovery [2] of the structure of the DNA molecule 50 years ago has led to a breathtakingly beautiful description of life. Life employs well-tailored chain molecules to store and replicate information, to carry out a dizzying array of functionalities, and to provide a molecular basis for natural selection. The complementary base pairing mechanism in DNA combined with its double-helix structure serves as a repository of information and provides a pretty mechanism for replication [2]. The replication is prone to errors or mutations and these errors, which are the basis of evolution, are in turn copied in future generations [3]. Using the RNA molecule as an intermediary, the information contained in the DNA genes is translated into proteins, which are linear chains of amino acids. Unlike the DNA molecule, which adopts a limited number of related structures, protein molecules [4–6] fold into thousands of native state structures under physiological conditions. For proteins, form determines functionality and the rich variety of observed forms underscores the versatility of proteins. There then follows a complex orchestrated dance in which proteins catalyze reactions, interact with each other, and finally feed back into the gene to regulate the synthesis of other proteins [1].

A protein molecule is large and has many atoms. In addition, the water molecules surrounding the protein play a crucial role in its behavior. At the microscopic level, the laws of quantum mechanics can be used to deduce the interactions but the number of degrees of freedom is far too many for the system to be studied in all its detail. When one attempts to look at the problem in a coarse-grained manner [7] with what one hopes are the essential degrees of freedom, it is very hard to determine what the effective potential energies of interaction are. This situation makes the protein problem particularly daunting and no solution has yet been found.

Over many decades, much experimental data has been accumulated yet theoretical progress has been somewhat limited. The problem is highly interdisciplinary and touches on biology, chemistry, and physics and it is often hard to distill the essential features of each of the multiple aspects of the problem. The great successes of quantum chemistry in the determination of the structure of the DNA molecule [2] and in the spectacular prediction that helices and sheets [8–10] are the building blocks of protein structures have spurred much work using detailed chemistry on understanding the protein problem. Such work has been very insightful in providing useful hints on how proteins behave at the atomic scale in performing their tasks. The missing feature, of course, in such a theoretical approach is that it treats each protein as a special entity with all the attendant details of the sequence of amino acids, their intricate side chain atoms, and the water molecules. Such an approach, while quite valuable, neither has as a goal nor can lend itself to a unified way of understanding seemingly disparate phenomena pertaining to proteins. Reinforcing this, experiments, which are very challenging, are carried out on one protein at a time and cry out for an understanding of the behavior of an individual class of protein.

The lessons we have learned from physics are of a different nature. The history of physics is replete with examples of the elucidation of connections between what seem to be distinct phenomena and the development of a unifying framework, which, in turn, leads to new observable consequences [11,12]. There have been many attempts at using physics-based approaches for understanding proteins. These have provided valuable insights on how one might think about the problem and have served as a means of understanding experimental data. Yet no simple unification has been achieved in a deeper understanding of the key principles at work in proteins.

We restrict ourselves to globular proteins which display the rich variety of native state structures. There are other interesting and important classes of proteins [13] such as membrane proteins and fibrous proteins which we do not consider here. Our goal here is to present a different approach to understanding proteins—our focus is on understanding the origin of protein structures and how they form the basis for both functionality and natural selection. Our work points to a unification of the various aspects of all

proteins: symmetry and geometry determine the limited menu of folded conformations that a protein can choose from for its native state structure; these structures are in a marginally compact phase in the vicinity of a phase transition and are therefore eminently suited for biological function; these structures are the molecular target for the powerful forces of evolution; proteins are well-designed sequences of amino acids which fit well into one of these predetermined folds; and proteins are prone to misfolding and aggregation leading to the formation of amyloids, which are implicated in debilitating human diseases [14,15] such as Alzheimer's, light-chain amyloidosis, and spongiform encephalopathies.

We present a discussion of the nature of the denatured state (which can loosely be thought of as the collection of unfolded conformations) and its possible key role in the protein folding problem. We also show how disordered proteins could fit into our unified framework.

The problem of how life was created is a fascinating one. Our focus is on looking at life on earth and asking how it works. The lessons we learn provide hints to the answers of deep and fundamental questions that have been pondered by our ancients: Was life on earth inevitable? Then there is the question posed by Henderson [16] about whether the nature of our physical world is biocentric. Is there a need for fine-tuning in biochemistry to provide for the fitness of life in the cosmos or even less ambitiously for life here on earth? Surprisingly, as we will show, a physics approach turns out to be valuable for thinking about these questions.

The main text of the paper contains the principal ideas and details of the calculations are relegated to the Appendixes. In Sec. II, we introduce the description of a protein as a thick polymer chain and highlight the differences in its phase diagram with respect to the usual string and bead model. In Sec. III, we make a comparison of the predictions obtained from the simple tube model against experimental data available on protein native state structures. In Sec. IV, we introduce a more refined model in which the tube picture is reinforced with the geometrical constraints that arise in the formation of hydrogen bonds and discuss the resulting phase diagram for an isolated peptide chain. In Sec. V, we discuss several consequences of our model including the nature of the free energy landscape, the innate propensity of proteins to aggregate into amyloidlike forms, and the role played by proteins as the targets of natural selection in molecular evolution. In Sec. VI, we discuss the nature of the denatured state of proteins and its possible role in protein folding. In the final Sec. VII, we conclude with a summary.

II. PHASES OF MATTER: FROM SPHERES TO TUBES

The fluid and crystalline phases of matter can be readily understood [17] in terms of the behavior of a simple system of hard spheres. The standard way of ensuring the self-avoidance of a system of uniform hard spheres is to consider all pairs of spheres and require that their centers are no closer than their diameter. Studies of hard spheres have a venerable history [18] including early work by Kepler on the packing of cannonballs in a ship's hold. Each hard sphere can be thought of as a point particle or a zero-dimensional object

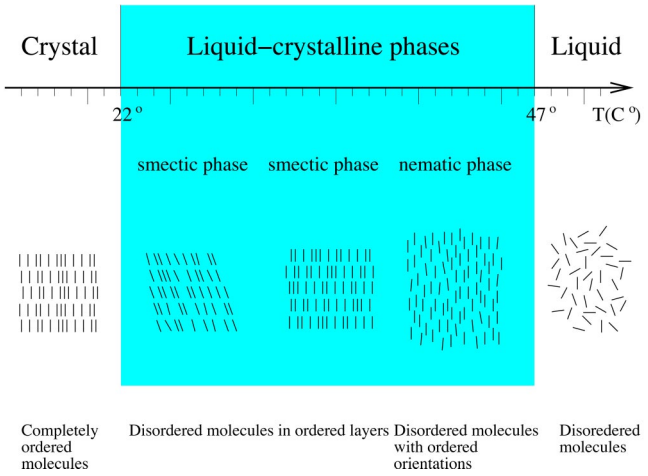

FIG. 1. (Color online) Schematic phase diagram for hard rods highlighting the rich behavior and the new (with respect to hard spheres) liquid crystal phases exhibited at intermediate temperatures.

with its own private space of spatial extent equal to its radius. Generalizing to a one-dimensional object, one must consider a line or a string, with private space associated with each point along the line, leading to a uniform tube of radius of cross section or thickness Δ, with its axis defined by the line. (Likewise, one could consider a collection of interacting tubes.) The generalization of the hard sphere constraint to the description of the self-avoidance of a tube of nonzero thickness is as follows [19] (see Appendix A). Consider all triplets of points along the axis of the tube. Draw circles through each of the triplets and ensure that none of the radii is less than the tube thickness [20]. This prescription surprisingly entails discarding pairwise interactions and working with effective three-body interactions [19,21,22].

One may visualize a tube as the continuum limit of a *discrete* chain of tethered disks or coins [21] of fixed radius separated from each other by a distance a in the limit of $a \rightarrow 0$. The inherent anisotropy associated with a coin (the heads to tails direction being different from the other two perpendicular to it) reflects the fact that there is a special local direction at each position defined by the locations of the adjacent objects along the chain. An alternative description of a discrete chain molecule is a string and bead model in which the tethered objects are spheres. The key difference between these two descriptions is the different symmetry of the tethered objects. Upon compaction of a chain of spheres, each individual sphere tends to surround itself isotropically with other spheres, unlike the tube situation in which nearby tube segments need to be placed parallel to each other. Even for unconstrained particles, deviations from spherical symmetry (replacing a system of hard spheres with one of hard rods, for example) lead to rich new liquid crystal phases [23,24] (see Fig. 1). Likewise, we find that the tube and a chain of tethered spheres exhibit quite distinct behaviors with one exception—in the presence of an attractive self-interaction favoring compaction, the chain of coins and the string and bead model behave similarly in the limit of vanishing ratios of the radii of the coin and sphere to the range

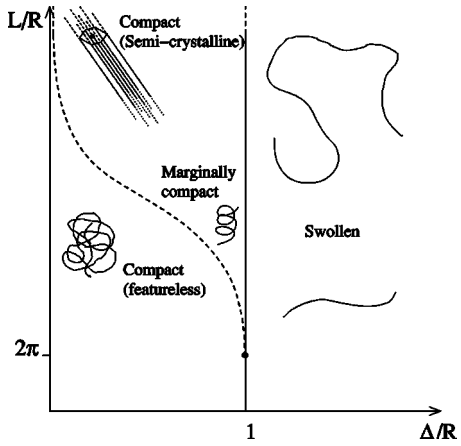

FIG. 2. Sketch of the zero-temperature phase diagram of a tube in the continuum, subject to a self-attraction promoting compaction. There are two phases when the tube length L is long compared to the range of attractive interaction R. One obtains the semicrystalline phase (with parallel/antiparallel alignment between different stretches of the tube which then fill the space with hexagonal symmetry, as depicted in the figure) when the tube thickness Δ is small compared to R and a swollen phase when Δ is large compared to R. There are interesting finite size effects in the semicrystalline phase. In the thin tube limit, on decreasing the length there is a crossover from the semicrystalline phase with overall cylindrical symmetry to a featureless compact phase with spherical symmetry when $L/R \sim (\Delta/R)^{-2}$. There is an unusual finite size effect when $\Delta \sim R$ near the confluence of three phases (at $L=2\pi R$, $\Delta=R$ for a chain in the continuum): the semicrystalline phase, the featureless compact phase, and the swollen phase. A marginally compact phase is obtained in this regime and displays a dramatic entropy reduction, with the choice structure being a helix with a well-defined pitch to radius ratio (see Fig. 4). Other structures such as hairpins and sheets are present in the marginally compact phase for *discrete chains* (see Fig. 3) [22].

of attraction. A detailed comparison between the chain of coins (tube) and the string and bead model with a bending rigidity energy term is carried out in Appendix B.

Figure 2 is a sketch of the phase diagram, at zero temperature, of a homopolymer of length L and thickness Δ with the range of attractive interaction R. This phase diagram has been obtained using detailed computer simulations accompanied by an approximate mean field theory [22] and can be understood on the basis of physical arguments. For large values of L/R, there are two distinct phases. When Δ/R is large, the tube is very thick compared to the range of attractive interactions and one obtains a swollen phase with equal weight for all self-avoiding conformations. One finds a very large degeneracy with no tendency toward compaction. On the other hand, for small Δ/R, one has a semicrystalline phase [25] in which the tube is stretched out locally with nearby sections parallel to each other. A similar structure is also obtained for many long tubes—the arrangement is akin to piling up logs parallel to each other with each log surrounded by six other logs in a hexagonal array, the optimal packing in two dimension of coins of radius Δ. Such structures are similar to those found in the Abrikosov flux lattice [26] and bear a resemblance to liquid crystal order.

Liquid crystals are a delicate state of matter of rodlike molecules which adopt many distinct arrangements sensitive to external electric and magnetic fields [23,24]. A liquid crystal phase that is analogous to the semicrystalline phase is the nematic phase in which the molecules move as in a regular liquid but with an alignment of their axes. Unlike a spin system in which an up spin is different from a down spin, in the nematic phase, all that matters is the direction of the axis of the particle—there is no up-down distinction—and this change in symmetry leads to a first order phase transition between the disordered isotropic and the ordered nematic phases. Likewise, the phase transition between the semicrystalline phase at low temperatures and a high temperature disordered phase in which there is no compaction of the tube is a first order transition as in the melting of ice into water. At the transition temperature, there is a coexistence of the two phases (e.g., pieces of ice floating in a glass of water) and an abrupt transition between the two states. One might call such a system a two-state system—one has water and/or ice but nothing in between.

When the tube is short, one would expect finite size effects [27] to come into play. In most physical systems, such finite size effects are intuitively obvious corrections to the bulk scenario and arise from the effects of the finite boundaries. For our tube, the simplest situation occurs in the swollen phase where the finite size effects are not important—short fat tubes continue to adopt open conformations. At the other extreme of small Δ/R, as one reduces the length of the tube, the overall symmetry of the folded object crosses over from that of a cylinder (corresponding to the Abrikosov flux lattice–like phase akin to the hexagonal arrangement of parallel, straight logs) to a sphere when $L \sim R^3/\Delta^2$ and one obtains one out of many degenerate featureless compact conformations. Physically, for a short tube, there are many more conformations that can be accommodated in the spherical topology than in the cylindrical topology without any accompanying sacrifice in the attractive interaction energy.

There is a confluence of three distinct types of structures: the swollen conformations, the semicrystalline phase, and the featureless compact conformations, when $\Delta \sim R \sim L$ (Fig. 2). This interplay leads to quite remarkable finite size effects: one obtains a *marginally compact* phase with a huge reduction in the degeneracy compared to the featureless compact phase and the swollen phase. On raising the temperature, one again finds a two-state behavior and the finite size analog of a first order transition between the marginally compact phase and the disordered phase. The first order transition occurs because it is necessary for different nearby tube segments to snap into position right alongside each other and parallel to each other in order to avail themselves of the attraction. The inherent anisotropy of a tube along with the fact that Δ is of order R leads to this requirement. Such two-state behavior can, in the simplest scenario, be associated with a transition state [28] along suitably chosen reaction coordinates. The structures of choice [21,29] in the marginally compact phase, for a *discrete chain*, are helices, kissing hairpins, regular hairpins, and sheets (Figs. 3 and 4). Helices, hairpins, and sheets are indeed characterized by a parallel placement of nearby tube segments. The marginally compact phase is poised in the vicinity of a phase transition to the swollen

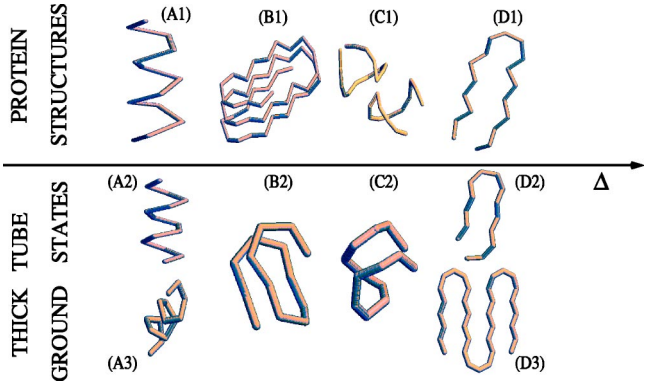

FIG. 3. (Color online) Building blocks of biomolecules and ground state structures associated with the marginally compact phase of a short tube corresponding to a discrete chain of tethered disks of radius Δ. The axis in the middle indicates the direction along which the tube thickness Δ increases. The top row shows some of the building blocks of biomolecules, while the bottom row depicts the corresponding structures obtained as the ground state conformations of a short tube. (A1) is an α helix of a naturally occurring protein, while (A2) and (A3) are the helices obtained in our calculations—(A2) has a regular contact map (i.e., a matrix whose elements, corresponding to residue pairs, are either 0 or 1 depending on whether the two given residues are in contact or not) whereas (A3) is a distorted helix in which the distance between successive atoms along the helical axis is not constant but has period 2. (B1) is a helix of strands in the alkaline protease of *Pseudomonas aeruginosa*, whereas (B2) shows the corresponding structure obtained in our computer simulations. (C1) shows the "kissing" hairpins of RNA and (C2) the corresponding conformation obtained in our simulations. Finally (D1) and (D2) are two instances of quasiplanar hairpins. The first structure is from the same protein as before (the alkaline protease of *Pseudomonas aeruginosa*) while the second is a typical conformation found in our simulations. The sheetlike structure (D3) is obtained for a longer tube (see [21] for more details). The biomolecular structures in the top row are shown in the C^{α} representation for proteins, and in the *P* representation for RNA kissing hairpins.

phase and the structures are therefore flexible [30] and sensitive to the right types of perturbations.

III. TUBES AND PROTEINS

There is a truly remarkable coincidence between the structures one obtains in the marginally compact physical state of matter of short tubes and the building blocks of protein native state structures (Fig. 3). Proteins [13] are linear chains of amino acids, of which there are 20 naturally occurring types with distinct side chains. The backbone and several of the side chains are hydrophobic and, under physiological conditions, globular proteins fold rapidly and reproducibly to somewhat compact conformations called their native state structures. In their native states, a hydrophobic core is created which is space filling and water is expelled from the interior. Even though there are hundreds of thousands of proteins in human cells, the total number of distinct folds that they adopt in their native states is only of the order of a few thousand [31–33]. Furthermore, these structures seem to be evolutionarily conserved [34–36]. Proteins are relatively short chain molecules and indeed longer globular proteins form domains which fold autonomously [37]. The building blocks of protein structures are helices, hairpins, and almost planar sheets (Fig. 3). Strikingly, short tubes, with no *heterogeneity*, in the marginally compact phase form helices with the same pitch to radius ratio as in real proteins [38] (Fig. 4) and almost planar sheets made up of zigzag strands. It is interesting to note that the helix is a very natural conformation for a tube and occurs without any explicit introduction of hydrogen bonding. Recent work on the denatured state of short amino acid sequences has suggested that the polyproline II helix might be the preferred structure in that phase, even though it does not entail the formation of any hydrogen bonds [39]. As in the tube case, small globular proteins show a two-state behavior [40–44] and recent experiments [4,45,46] have been successful in mapping out the nature of the transition state in several cases.

Let us make the *constructive hypothesis* that the extraordinary similarity between the structures adopted by short tubes in the marginally compact phase and the building blocks of protein native state structures is not a mere coincidence. We *postulate* instead that the tube picture presented above is a paradigm for understanding protein structures. Quite generally, such postulates are of limited utility unless one is able to unify seemingly unrelated aspects of the problem and make predictions amenable to experimental verification. In our case, while the tube idea is theoretical, there is a wealth of experimental data already available on proteins. Before we proceed to explore the consequences of our hypothesis, we will first link the tube picture with the protein problem using experiments as the guide.

Let us begin by asking whether the backbone of a protein can be described as a tube. Figure 5 indeed shows that, in its native state, the protein backbone can be thought of as the axis of a tube of approximate radius of cross section (Δ) equal to 2.7 Å. Interestingly, there are small variations in the tube radius especially in the vicinity of backward bends [48]. The tuning of the two length scales Δ and *R* to be comparable to each other happens automatically for proteins: the sizes of the amino acid side chains determine both the tube thickness and the range of interactions. Steric interactions lead to a vast thinning of the phase space that protein structures can explore [49,50]. Physically, the notion of a thick chain or a tube follows directly from steric interactions in a protein—one needs room around the backbone to house the amino acid side chains without any overlap. The same side chains that determine the tube thickness also control the range of attraction—the outer atoms of the side chain interact through a short range interaction screened by the water. This self-tuning is a quite remarkable feature of proteins.

The rapid folding of small proteins can be understood in terms of the inherent anisotropy of a tube and the self-tuning of the two key length scales, the tube thickness and the range of the attractive interactions. In the marginally compact phase, in order to take advantage of the attractive interactions, nearby segments of the tube have to snap into place parallel to each other and right up against each other. As stated before, both in the tube picture and in proteins, the helix and the sheet are characterized by such parallel space

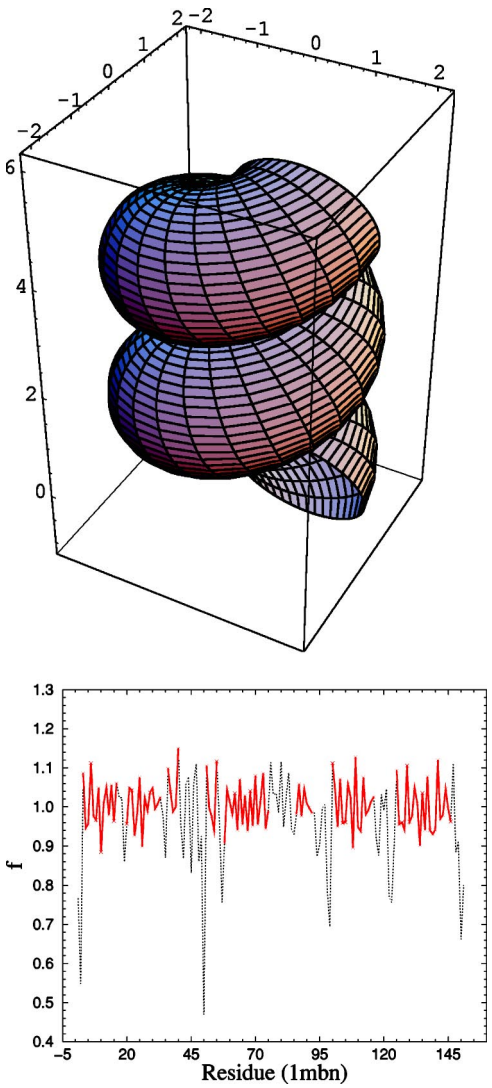

FIG. 4. (Color online) (a) Space filling *optimal* helix, with a pitch to radius ratio $c^* \approx 2.512$ (drawn using MATHEMATICA). As explained in Appendix C, this optimal value is determined by requiring that the radius of curvature of the helical curve is equal to half the minimum distance of closest approach between different turns of the helix. The corresponding tube (that can be thought of as being inflated uniformly around the curve) is optimally space filling since it stops growing when reaching its maximum thickness both *locally* (the radius of curvature) and *nonlocally* (half the minimum distance of closest approach between different turns) at the same time (see Appendixes A and C). Such an optimality criterion is shared by some of the conformations selected as ground states in our simulations in the marginally compact phase such as helices or planar hairpin and sheets shown in Fig. 3, when it is properly translated for the case of a discrete chain [see below and Eq. (A5) in Appendix A]. It can be shown [21] that the planarity of hairpins and sheets is a consequence of this optimal space filling criterion. The same geometrical feature is strikingly found to hold, within 3%, for α helices occurring in the native state of natural proteins [38]. (b) Plot of the ratio $f_i = \rho_{NL}(i)/\rho_L(i)$ of the nonlocal radius of curvature $\rho_{NL}(i) \equiv \min_{j<k} r(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k)$ (with $\{j,k\} \neq \{i-2, i-1\}, \{i-1, i+1\}, \{i+1, i+2\}$) over the radius of curvature $\rho_L(i) \equiv r(\mathbf{r}_{i-1}, \mathbf{r}_i, \mathbf{r}_{i+1})$ as a function of the residue index i for the native state structure of sperm whale myoglobin (Protein Data Bank code 1mbn), where $\mathbf{r}_i$ refers to the spatial coordinates of the C^α atom of the ith residue, $1 \leqslant i \leqslant 153$ [see Appendix A for the definition of the triplet radius $r(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k)$]. In correspondence with the eight α helices present in the myoglobin fold, shown as the solid (red) parts in the plot, the values of f_i oscillate around unity, demonstrating that helices in natural proteins are *optimally* space filling in the sense described above.

filling alignment of nearby tube segments. In proteins, such an arrangement serves to expel the water from the protein core. As shown by Pauling and co-workers [8,9], hydrogen bonds provide the scaffolding for both helices and sheets and place strong geometrical constraints stemming from quantum chemistry.

IV. BEYOND THE TUBE ARCHETYPE: A REFINED TUBE MODEL INFORMED BY PROTEIN DATA

We turn now to a marriage of the tube idea and the wealth of information available from a variety of experimental probes [4,51] in preparation for the task of exploring the consequences of our hypothesis. Recall that three-body local and nonlocal radii constraints describe the self-avoidance of a tube [19] (see Appendix A). For a discrete chain, the local three-body radius is defined as the radius of a circle drawn through three consecutive nodes of the chain (in the limit of a continuous chain the local three-body radius is equal to the radius of curvature). The nonlocal radius at a given node is defined to be the smallest among all the radii of circles drawn through that node and all pairs of other nodes except for its adjacent nodes [see also the caption of Fig. 4(b)]. Unlike unconstrained matter for which pairwise interactions suffice, for a chain molecule, it is necessary to define the context of the object that is part of the chain. This is most easily carried out by defining a local Cartesian coordinate system (see Fig. 6) whose three axes are defined by the tangent to the chain at that point, the normal, and the binormal which is perpendicular to both the other two vectors. A study [52] of the experimentally determined native state structures of proteins from the Protein Data Bank [53] reveals that there are clear amino acid aspecific geometrical constraints on the relative orientation of the local coordinate systems due to sterics and also associated with amino acids which form hydrogen bonds with each other (see Fig. 15 in Appendix D).

Recently [52], we have carried out Monte Carlo simulations of short *homopolymers*, chains made up of just one type of amino acid, subject to these geometrical constraints and physically motivated interaction energies, a local bending energy penalty e_R, an overall hydrophobicity e_W, and effective hydrogen bond energies (see Appendix E for details about the refined tube model and the simulations). The resulting phase diagram and the associated structures for short homopolymers of length 24 are depicted in Fig. 7. In keeping

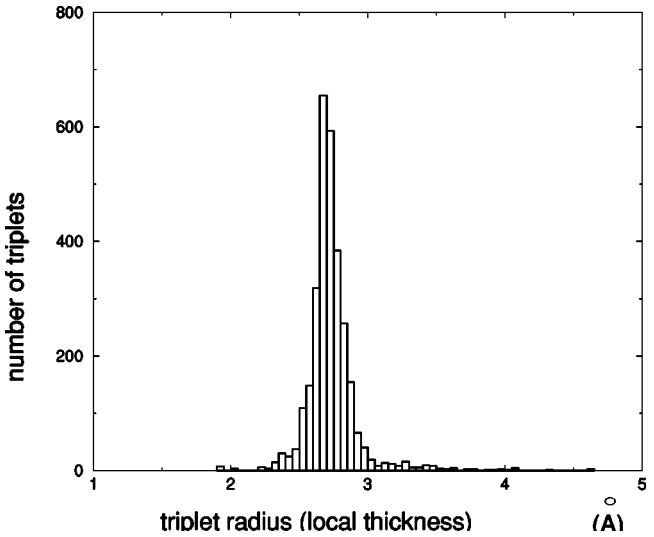

FIG. 5. Histogram of local thicknesses computed for all residues of different protein native structures, when the virtual chain formed by the backbone C^{α} atoms is viewed as a discretized thick tube. At a given residue the local thickness is simply the minimum triplet radius over all triplets containing that residue (see Appendix A for the definition of triplet radius and for an explanation of how such a quantity arises within a tube description).

with the behavior of the archetype tube discussed earlier, in the vicinity of the swollen phase, one obtains distinct assembled tertiary structures, quite akin to real protein structures, on making small changes in the interaction parameters. The striking similarity between the observed structures and real protein structures suggests that our model captures the essential ingredients responsible for the limited menu of protein native structures.

The marginally compact phase has distinct structures including a single helix, a bundle of two helices, a helix formed by β strands, a β hairpin, three-stranded β sheets with two distinct topologies and a β-barrel-like conformation. These structures are the stable ground states in different parts of the phase diagram. Furthermore, conformations such as the β-α-β motif are found to be competitive local minima.

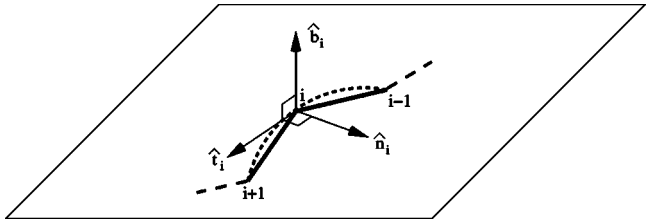

FIG. 6. Sketch of the local coordinate system. For each C^{α} atom i (except the first and the last ones), the axes of a right-handed local coordinate system are defined as follows. The tangent vector $\hat{t}_i$ is parallel to the segment joining $i-1$ with $i+1$. The normal vector $\hat{n}_i$ joins i to the center of the circle passing through $i-1$, i, and $i+1$ and it is perpendicular to $\hat{t}_i$. $\hat{t}_i$ and $\hat{n}_i$ along with the three contiguous C^{α} atoms lie in a plane shown in the figure. The binormal vector $\hat{b}_i$ is perpendicular to this plane. The vectors $\hat{t}_i$, $\hat{n}_i$, $\hat{b}_i$ are normalized to unit length.

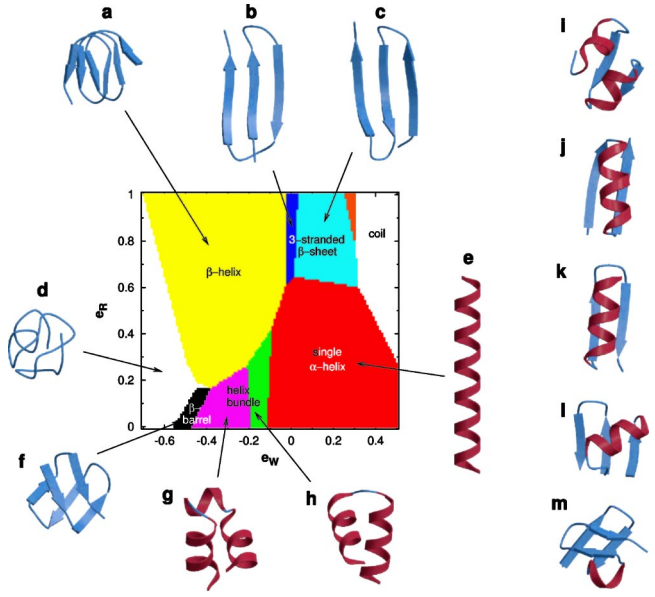

FIG. 7. (Color online) Phase diagram of ground state conformations. The ground state conformations were obtained by means of Monte Carlo simulations of chains of 24 C^{α} atoms. e_R and e_W denote the local radius of curvature energy penalty and the solvent mediated interaction energy, respectively (see Appendix E). Over 600 distinct local minima were obtained in our simulations in different parts of parameter space starting from a randomly generated initial conformation. The temperature is set initially at a high value and then decreased gradually to zero. (a), (b), (c), (e), (f), (g), and (h) are the Molscript representations of the ground state conformations which are found in different parts of the parameter space as indicated by the arrows. The helices and strands are assigned when local or nonlocal hydrogen bonds are formed according to the rules described in Appendix E. Conformations (i), (j), (k), (l), and (m) are competitive local minima. In the shaded phase (orange line) on the top right that does not correspond to any conformation, the ground state is a two-stranded β hairpin (not shown). Two distinct topologies of a three-stranded β sheet are found corresponding to conformations shown in (b) and (c), respectively. The region (white online) in the left of the phase diagram has large attractive values of e_W and the ground state conformations are compact globular structures with a crystalline order induced by hard sphere packing considerations [54] and not by hydrogen bonding [conformation (d)].

The specific structure depends on the precise values of the local radius of curvature penalty (a large penalty forbids tight turns associated with helices resulting in an advantage for sheet formation) and the strength of the hydrophobic interactions (a stronger overall attraction leads to somewhat more compact well-assembled tertiary structures). The topology of the phase diagram allows for the possibility of conformational switching leading to the conversion of an α helix to a β topology on changing the hydrophobicity parameter analogous to the influence of denaturants or alcohol in experiments [55].

V. CONSEQUENCES OF THE PROTEIN-TUBE HYPOTHESIS

We now turn to a study of some of the consequences of our postulate that the tube is a useful paradigm for under-

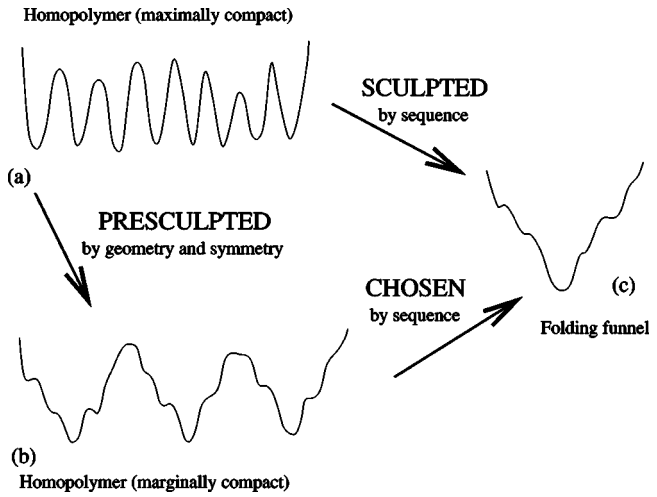

FIG. 8. Simplified one-dimensional sketches of energy landscape. The quantity plotted on the horizontal axis schematically represents a distance between different conformations in the phase space and the barriers in the plots indicate the energy needed by the chain in order to travel between two neighboring local minima. (a) Rugged energy landscape for a homopolymer chain with an attractive potential promoting compaction as, e.g., in a string and bead model. There are many distinct maximally compact ground state conformations with roughly the same energy, separated by high energy barriers (the degeneracy of ground state energies would be exact in the case of both lattice models and off-lattice models with discontinuous square-well potentials). (b) Presculpted energy landscape for a homopolymer chain in the marginally compact phase. The number of minima is greatly reduced and the width of their basin increased by the introduction of geometrical constraints. (c) Funnel energy landscape for a protein sequence. As folding proceeds from the top to the bottom of the funnel, its width, a measure of the entropy of the chain, decreases cooperatively with the energy gain. Such a distinctive feature, crucial for fast and reproducible folding, arises from careful sequence design in models whose homopolymer energy landscape is similar to (a). In contrast, funnel-like properties already result from considerations of geometry and symmetry in the marginally compact phase (b), thereby making the goals of the design procedure the relatively easy task of stabilization of one of the presculpted funnels followed by the more refined task of fine-tuning the putative interactions of the protein with other proteins and ligands.

standing protein structures and behavior. We will benchmark these against experimental evidence to assess their validity.

A. Energy landscape of proteins

There have been many previous studies of proteins from a physics point of view [56]. The standard approach is to assume an overall attractive short range potential which serves to lead to a compact conformation of the chain in its ground state. In the absence of amino acid specificity or when one deals with a homopolymer, there is a huge number of highly degenerate ground states comprising all maximally compact conformations with high barriers between them [see Fig. 8(a)]. The ground state degeneracy and the height of the barriers grow exponentially with the length of the homopolymer. The role played by sequence heterogeneity is to break the degeneracy of maximally compact conformations, leading to a unique ground state conformation which, of course, depends on the amino acid sequence. Yet, for a typical random sequence, the energy landscape is still very rugged and is virtually the same as in Fig. 8(a). A model protein moving in such a rugged landscape can be subject to trapping in local minima and may not be able to fold rapidly, so that glassy behavior may ensue due to such trapping. Bryngelson and Wolynes [57] suggested that there is a principle of minimal frustration at work for well-designed sequences in which there is a nice fit between a given sequence and its native state structure carving out a funnel-like landscape [58] which promotes rapid folding and avoids the glassy behavior [Fig. 8(c)].

Indeed, given a sequence of amino acids, with all the attendant details of the side chains and the surrounding water, one obtains a funnel-like landscape with the minimum corresponding to its native state structure. Each protein is characterized by its own landscape. In this scenario, the protein sequence is all important and the protein folding problem, besides becoming tremendously complex, needs to be attacked on a protein-by-protein basis.

In contrast, our model calculations show that the large number of common attributes of globular proteins [29,59] reflects a deeper underlying unity in their behavior. At odds with conventional belief, a consequence of our hypothesis is that the gross features of the energy landscape of proteins result from the amino acid specific common features of all proteins. This landscape is *(pre)sculpted* by general considerations of geometry and symmetry [Fig. 8(b)]. Our unified framework suggests that the protein energy landscape ought to have thousands of broad minima corresponding to putative native state structures. The key point is that for each of these minima the desirable funnel-like behavior is already achieved at the homopolymer level *in the marginally compact part of the phase diagram* (see Fig. 7). The self-tuning of two key length scales, the thickness of the tube and the interaction range, to be comparable to each other and the interplay of the three energy scales, hydrophobic, hydrogen bond, and bending energy, in such a way as to stabilize marginally compact structures also provide the close cooperation between energy gain and entropy loss needed for the sculpting of a funneled energy landscape.

Recent work has shown that the rate of protein folding is not too sensitive [44,60] to large changes in the amino acid sequence [60,61], as long as the overall topology of the folded structure is the same. Furthermore, mutational studies [44–47] have shown that, in the simplest cases, the structures of the transition states are also similar in proteins with similar native state structures.

Sequence design [62] would favor the appropriate native state structure over the other putative ground states leading to an energy landscape conducive for rapid and reproducible folding of that particular protein. Nature has a choice of 20 amino acids for the design of protein sequences. A presculpted landscape greatly facilitates the design process. Indeed, within our model, we find that a crude design scheme, which takes into account the hydrophobic (propensity to be buried) and polar (desire to be exposed to the water) character of the amino acids, is sufficient to carry out a successful

design of sequences with one or the other of the structures shown in Fig. 7. The matching of the hydrophobic profile of the designed sequence to the burial profile [63] (as measured by the number of neighbors within the range of the hydrophobic interaction) leads to the correct fold in a Monte Carlo simulation. As examples, the sequence HPPHHPHHPPPPPPHHPHHPPPPP, with $e_R=0.3$ uniformly for all residues, $e_W=-0.4$ for contacts between H and H, and $e_W=0$ for other contacts, has as its ground state the two-helix bundle structure [Fig. 7(h)] whereas HPHHHPPPPHHPPHHPPPPHHHPP prefers the $\beta\alpha\beta$ motif [Fig. 7(j)]. It is interesting to note that the $\beta\alpha\beta$ motif is only a local minimum in the phase diagram of a homopolymer but is stabilized by the designed sequence. Also, as is seen experimentally, many protein sequences adopt the same native state conformation [64]. Once a sequence has selected its native state structure, it is able to tolerate a significant degree of mutability except at certain key locations [45,46,62,65]. Furthermore, multiple protein functionalities can arise within the context of a single fold [66].

One of the successful methods of protein structure prediction is based on threading [67]. The basic idea is entirely consistent with our findings—one uses pieces of native state structures of longer proteins as possible candidate structures of a shorter protein—but the technique is simpler because instead of determining the structure from *ab initio* calculations, one merely has to select from among the putative native state structures. The documented success of the threading method confirms that each protein does not fashion its own native state structure but merely selects from the menu of predetermined folds.

B. Amyloid phase of proteins

A range of human diseases such as Alzheimer's, spongiform encephalopathies, and light-chain amyloidosis lead to degenerative conditions and involve the deposition of plaquelike material in tissue arising from the aggregation of proteins [14,15,68,69]. In the case of prions [69], one observes a transition from α to β rich structures which favors aggregation and causes bovine spongiform encephalopathy disease. It has been argued [70] that the formation of amyloid fibrils occurs in a hierarchical way starting from a chiral β strand. The resulting structures arise from a competition between the free energy gain from the aggregation and the elastic energy cost of the distortion. A variety of proteins not involved in these diseases also form aggregates very similar to those implicated in the diseased state [15,68]. This suggests [15] that the tendency for proteins to aggregate is a generic property of polypeptide chains with the specific sequence of amino acids playing at best a secondary role. Can one understand this general tendency of proteins to form amyloids within our framework?

Let us recall the semicrystalline polymer phase which one obtains when the tube is sufficiently long (or when there are many interacting tubes) and is subject to attractive interactions leading to compaction. In this phase, the tube is stretched out locally with nearby sections parallel to each other (or has the tubes stacked parallel to each other in a periodic arrangement) and does not have the richness we associate with protein native state structures. Returning to the protein, one may ask whether there are structures which are the analogs of those found in the semicrystalline phase.

In order to assess the role played by the interaction between multiple short proteins, let us first consider our model homopolymer chain made up of 36 identical amino acids in the marginally compact phase of the refined tube model (see Sec. IV and Appendix E) with the hydrophobic parameter and the local bending energy penalty chosen so that the ground state is a single long helix.

On making two incisions in the chain to create three distinct chains each containing 12 amino acids, the ground state of the system appears to be a bundle of three helices [see Fig. 9(d)]. This helix structure, however, is stable only at very low temperatures. At intermediate temperatures, close to but lower than the temperature of the specific heat peak, it is destabilized in favor of aggregated β helices [Figs. 9(a) and 9(b)] or sandwiches of β sheets [Fig. 9(c)], due to entropic effects. Cutting a single chain into parts increases the entropy of the system. Unbonded chains are more flexible and this promotes the formation of interchain hydrogen bonds. The β-sheet structures also show an increased flexibility comparing to the helix bundle, and they have better kinetic accessibility from a disordered globule. While the appearance of β-sheet conformations in the case of three chains seems to have an entropic origin, it seems likely that the ground state of a system of multiple chains does in fact consist of aggregated β sheets. Indeed simulations of five or ten chains have shown that β structures are the most likely choice (see Fig. 10).

The formation of β-sheet structured protein aggregates is favored with respect to other possible aggregates such as helix bundles [which we actually detect in our simulations; see Fig. 9(d)]. In the latter case hydrogen bonds are saturated within a single helix so that aggregation is driven exclusively by the effective hydrophobic attraction between different helices. On the other hand, for structures such as those shown in Figs. 9(a)–9(c) hydrogen bonds are formed between different chains and several β strands are left unsaturated at both "ends" of the aggregate, which can then readily grow by hydrogen bonding to other chains.

The refined tube model can be used to explore the free energy landscape of a homopolymer chain in the vicinity of its folding transition temperature, operationally defined as the specific heat peak temperature. (Of course, there is no real phase transition for finite size systems such as proteins.) Figure 11(a) is a contour plot of the free energy at a temperature higher than the folding transition temperature for the parameter values $e_W=-0.08$ and $e_R=0.3$ for which the ground state is an α helix. The free energy landscape has just one minimum corresponding to the denatured phase whose typical conformations are still somewhat compact. The contour plot at the folding transition temperature [Fig. 11(b)] has three local minima corresponding to an α helix, a three-stranded β sheet, and the denatured state. At lower temperatures, the α helix is increasingly favored and the β sheet is never the global free energy minimum.

That a β-sheet structure is a significant competitor with a large basin of attraction in a region where the stable phase is

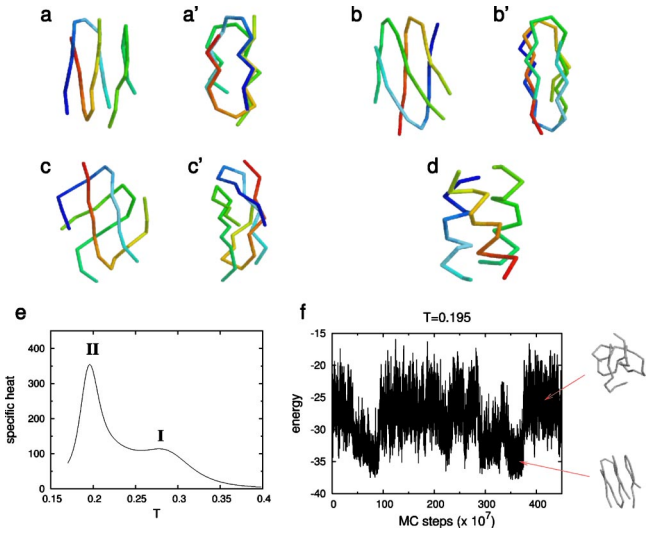

FIG. 9. (Color online) Aggregated structures formed by three chains of length 12. We show the lowest energy conformations obtained in long simulations for three 12-residue chains confined within a cubic box of side $L=80$ Å at $T=0.19$ (a), $T=0.18$ (b), and $T=0.16$ (c). The conformations shown in (a'), (b'), and (c') are the same as in (a), (b), and (c), respectively, but viewed from a different angle. The parameters used in the model are $e_W=-0.08$ and $e_R=0.2$ which correspond to having a single helix ground state in the case of a single chain. The simulations start with random extended conformations for all chains and are carried out with pivot and crank-shaft moves that are accepted or rejected based on the METROPOLIS criterion. Moves that bring the residues out of the box are not allowed. The bundle of three helices (d) is a putative ground state of the system and was obtained in a simulation at a very low temperature ($T=0.05$) starting with isolated single helices. This conformation has the lowest energy among those shown but is not the equilibrium conformation at intermediate temperatures. Indeed, a simulation run at $T=0.18$ starting with conformation (d) leads to the helix bundle being converted into the β-helix-like conformation shown in (b) which is the dominant equilibrium conformation at this temperature. (e) The aspecific heat as function of temperature for the system of three 12-residue peptides. The data shown were obtained using the weighted histogram technique [71] based on long equilibrium simulations at various temperatures between 0.16 and 4. The small shoulder (I) corresponds to a condensation of separated peptides into a disordered globule. The large peak (II) corresponds to a transition from disordered globule to the β-helix-like phase. (f) The energy as a function of time (in Monte Carlo steps) during a long simulation at a temperature corresponding to the maximum of the specific heat, $T=0.195$. The simulation shows several transitions between the disordered globule phase and the β-helix-like phase.

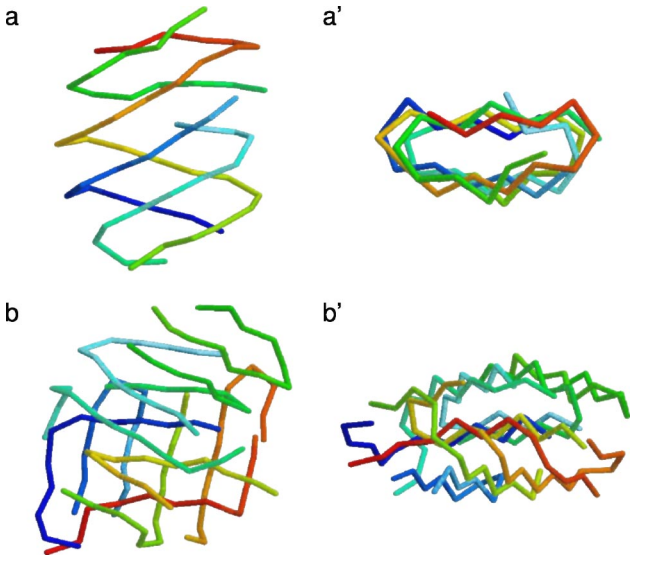

FIG. 10. (Color online) Aggregated structures formed by five and ten chains of length 12 with $e_W=-0.08$, $e_R=0.2$. We show the lowest energy conformations obtained in long simulations for five chains at $T=0.18$ (a), and for ten chains at $T=0.2$ (b). The five-chain system is confined within a cubic box of side $L=80$ Å whereas the ten-chain system is confined within a cubic box of side $L=100$ Å. The conformations shown in (a') and (b') are the same as those in (a) and (b) but viewed from a different angle.

a helix (see Fig. 11) reinforces the possibility that the interaction between several proteins could stabilize the formation of extended hydrogen bonded β sheets via the aggregation of individual chains (see [72] for experimental evidence that the increased propensity for extended single chain β conformations as the temperature is increased could indeed drive the formation of β aggregates). These kinds of structures, which resemble the basic structures associated with amyloid fibrils, thus seem to belong to the general class of predetermined folds, but this time for multiple proteins, and ought to be seen ubiquitously in generic proteins [15,68]. This suggests that the key to the prevention of such aggregates is the stabilization of helices in such proteins and evolutionary mechanisms such as proteasomes, molecular chaperones [73], and ubiquitination enzymes [15,68].

Our results show the generic tendency for multiple chains of amino acids to form aggregated amyloids rather than maintain their proteinlike shape. Interestingly, nature has, on suitable occasions, thwarted the tendency of a single long chain to form amyloid by dividing the protein into substantially independent domains which fold autonomously and are then assembled together. This suggests that the variety of protein folds increases with length up to a certain point at which they are supplanted by the formation of domains or amyloids.

In a recent paper, Fandrich and Dobson [74] suggested that "amyloid formation and protein folding represent two fundamentally different ways of organizing polypeptides into ordered conformations. Protein folding depends critically on the presence of distinctive side chain sequences and produces a unique globular fold. By contrast, … amyloid formation arises primarily from main chain interactions that are, in some environments, overruled by specific side chain contacts." Our results are in complete accord with the suggestion that amyloid structures may arise from the generic properties of the proteins with the details of the amino acid side chains playing a secondary role. However, our work suggests that instead of an "inverse side chain effect in amyloid structure formation" [74], there is a unifying theme in the behavior of proteins. Just as the class of cross-linked β structures are determined from geometrical considerations, the menu of protein native state structures is also determined by the com-

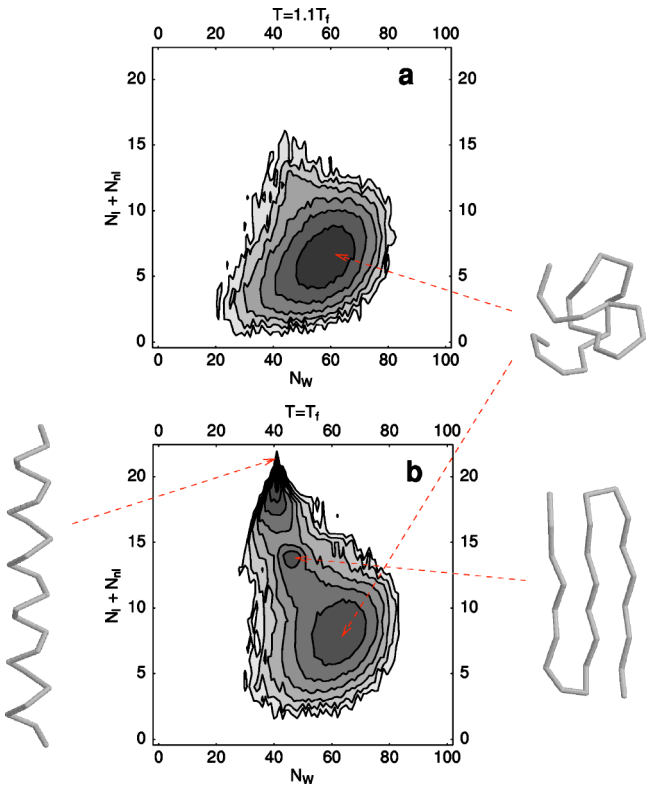

FIG. 11. (Color online) Contour plots of the effective free energy (a) at high temperature ($T=0.22$) and (b) at the folding transition temperature $T_f=0.2$ for a single 24-residue homopolymeric chain, with $e_W=-0.08$, $e_R=0.3$. The effective free energy, defined as $F(N_l+N_{nl},N_W)=-\ln P(N_l+N_{nl},N_W)$, is obtained as a function of the total number of hydrogen bonds N_l+N_{nl} and the total number of hydrophobic contacts N_W from the histogram $P(N_l+N_{nl},N_W)$ collected in equilibrium Monte Carlo simulations at constant temperature. The spacing between consecutive levels in each contour plot is 1 and corresponds to a free energy difference of $k_B\tilde{T}$, where $\tilde{T}$ is the temperature in physical units. The darker the color, the lower the free energy value. There is just one free energy minimum corresponding to the denatured state at a temperature higher than the folding transition temperature (a) whereas one can discern the existence of three distinct minima at the folding transition temperature (b). Typical conformations from each of the minima are shown in the figure.

mon attributes of globular proteins: the inherent anisotropy associated with a tube and the geometrical constraints imposed by hydrogen bonds and steric considerations.

C. Natural selection and protein interactions

Traditionally, the framework of evolution in life works through two aspects of organization called the genotype and the phenotype. The genotype is the heritable information encoded in the DNA, which is translated through the RNA molecules into proteins. The phenotype is valuable for adaptation and at the molecular level plays a key role in natural selection. One conventionally assumes that there is a selection of phenotypes which leads to an enhancement in the numbers of the genotype. Furthermore, mutations of the genotype lead to the possibility of new phenotypes.

Let us consider the situation at two levels: the sequence level (which is the genotype because it is a direct translation from the evolving DNA molecules) and the structure level, which we can think of as the phenotype. As pointed out by Maynard Smith [75], as the sequence undergoes mutation, there must be a continuous network that the mutated sequences can traverse without passing through any intermediaries that are nonfunctioning. Thus, one seeks a connected network in sequence space for evolution by natural selection to occur. There is considerable evidence, accumulated since the pioneering suggestion of Kimura [76] and King and Jukes [77], that much of evolution is neutral. The experimental data strongly support the view that the "random fixation of selectively neutral or very slightly deleterious mutants occur far more frequently in evolution than selective substitution of definitely advantageous mutants" [78]. Also "those mutant substitutions that disrupt less the existing structure and function of a molecule (conservative substitutions) occur more frequently in evolution than more disruptive ones" [78]. Thus while one has a "random walk" in sequence space that forms a connected network, there is no similar continuous variation in structure space [36,79].

These facts are in accord with our result of a presculpted energy landscape that is shared by all proteins and has thousands of local minima corresponding to putative native state structures—not too few because that would not lead to sufficient diversity and not too many because that would lead to too rugged a landscape with little hope that a protein could fold reproducibly and rapidly into its native state structure. Indeed, many proteins share the same native state fold and often the mutation of one amino acid into another does not lead to radical changes in the native state structure, underscoring the fact that it is not the details of the amino acid side chains that sculpt the energy landscape but rather some overarching features of symmetry and geometry that are common to all proteins. In this respect, the phase of matter that comprises the native state structures is one that is possibly determined by physical law rather than by the plethora of microscopic details in analogy with the limited menu of possible crystal structures.

Anfinsen [41] wrote in 1973, "Biological function appears to be more a correlate of macromolecular geometry than of chemical detail." There has been much recent progress in extracting information on biological function and protein interactions [80] from the structure of proteins and the complexes they form [81]. A protein structure chosen from the predetermined menu of folds contains information on the topology of the folded state. Additionally, one can glean information on the nature of the exposed surface, crystal packing, and the existence of clefts or other geometrical features (which are often the active sites of enzymes). The picture is completed by knowledge of the sequence of amino acids that folds into the structure using which one can infer the amino acid composition of the exposed surfaces, the location of mutants and conserved residues, and evolutionary relationships. For some structural families, function is highly conserved, whereas for others, one can use the types of information described above to guess the function [82].

Biological reactions are accelerated by factors of more than a billion by enzymatic proteins. Enzymes not only pro-

vide for great catalytic efficiency but are also extremely specific in their function. The principal mechanism [83] underlying the tremendous enhancement of the reaction rate by the enzymes is the lowering of the free energy of the transition state of the reaction through their specific binding to the substrate or the reactant(s). In its native state, an enzyme adopts a structure chosen from the menu of predetermined folds. Strikingly, only a small part of this structure is important for the enzymatic action. Generally, there are a few amino acids, associated with the active site, which are responsible for the catalytic activity. In close proximity, one also finds the substrate binding site which provides the specificity, often through the classic lock and key mechanism.

An illustration of enzymatic action and the role of molecular evolution is provided by the protease family of proteins. In a living cell, there is turnover of proteins with new proteins being continually synthesized along with the degradation of existing proteins. Proteins responsible for degradation through the hydrolysis of peptide bonds are called proteases. Under physiological conditions, peptide bonds are stable for a period of around a hundred years. The proteases are able to enhance the degradation rate selectively by factors of around a billion. There are several classes of proteases including serine proteases (such as chymotrypsin, a digestive enzyme) with a very reactive serine residue, cysteine proteases (such as papain, which is a digestive enzyme derived from papaya) with cysteine playing the role of serine, aspartyl proteases (such as renin which controls blood pressure) which employs a pair of aspartate groups, and metalloproteases (such as collagenase responsible for collagen degradation in osteoarthritic cartilage), which use a bound metal ion such as zinc to accelerate the hydrolysis.

In serine proteases, the catalytic triad comprises three amino acids, serine, histidine, and aspartate, bound to each other through hydrogen bonds, whose presence leads to the proton being moved away from the serine and the creation of a reactive alkoxide ion. The same triad is implicated in all serine proteases. Indeed, an example of convergent evolution is provided by subtilisin (an enzyme that resembles chymotrypsin in its action and is made by certain soil bacteria) and its family members, which possess the catalytic triad but have a quite different structure from chymotrypsin. Here Nature uses different folds from the presculpted energy landscape which, on appropriate sequence design, have the same catalytic triad and perform similar tasks.

The limited menu of possible protein folds provides a marvellous opportunity for divergent evolution. This corresponds to proteins whose native state structure and the catalytic triad are the same but with distinct differences in the nature of the binding site. The binding site in chymotrypsin is adjacent to the active site and is a hydrophobic cavity which facilitates hydrolysis of the peptide bonds on the carboxyl side of aromatic or large hydrophobic amino acids such as Trp, Tyr, Phe, Met, and Leu. Relatively small changes in the amino acid sequence, which maintain both the native state structure and the active triad lead to other proteins such as trypsin (a digestive protein made in the pancreas which cleaves after positively charged amino acids lysine and arginine due to a change of one of the hydrophobic amino acids in the binding cavity to a negatively charged aspartic acid), elastase (a protein made both in the pancreas and by white blood cells in which two glycines in the binding cavity are replaced by much larger amino acids valine and threonine allowing the enzyme to specifically target elastin, which is an important building block of blood vessel walls and ligaments—elastase is able to cleave proteins after a glycine and alanine because of the small size of the binding cavity), thrombin (a larger enzyme, the tail end of which bears a significant similarity to the sequence of amino acids of chymotrypsin and trypsin and cleaves proteins only at arginine-glycine linkages; thrombin is a complex regulatory protease which converts a usually soluble blood protein fibrinogen into the insoluble fibrin causing a blood clot and the cessation of bleeding), plasmin (an enzyme which cleaves proteins after lysine and arginine and dissolves blood clots), cocoonase (which also cleaves after lysine and arginine in the silk strands of the cocoon after the transformation of a caterpillar into a silk moth), and acrosin (an enzyme which plays a pivotal role in fertilization by creating a hole in the protective sheath around the egg and allowing sperm-egg contact).

As we have seen, evolution along with natural selection allow Nature to use variations on the same theme facilitated by the rich repertory of amino acids to create enzymes that are able to catalyze a remarkable array of diverse and complex tasks in the living cell. The key point, of course, is that in order for molecular evolution to work in this manner, one needs the constant backdrop of folds not shaped by the sequence but determined by physical law. Were the folds not immutable and themselves subject to Darwinian evolution, the possibility of creating so many subtle and wonderful variations on the same theme would not exist. The presculpted landscape is the crucial feature that leads to a predetermined menu of immutable folds.

It is known that key functional sites exhibit a high degree of conservation [84]. Interestingly, coevolutionary analysis has been useful in identifying protein-protein interactions [85]. Structural similarity, independent of evolutionary homology, can be the key reason why proteins with different folds share some commonality in enzymatic activity or ligand binding [86]. Conversely, there are protein structures such as the TIM (Triose phosphate Iso Merase) barrel [87] which are very versatile and are able to house proteins that are able to carry out multiple functionalities. Even though the proteins are able to perform diverse catalytic tasks, Nagano *et al.* [87] find that the active site is generally found at the C terminal end of the barrel sheets and that there are "striking structural superpositions" of the metal-ligating and catalytic residues.

Nooren and Thornton [88] have pointed out that "The structure and affinity of a PPI [protein-protein interaction] is tuned to its biological function and the physiological environment and control mechanism. PPIs presumably evolve to optimize 'functional' efficacy. This does not necessarily involve strong interactions. Clearly, weak transient interactions that are efficiently controlled are also very important in cellular processes."

There are several attractive features of the picture we have developed based on the tube-protein hypothesis. First, protein structures lie in the vicinity of a phase transition to

the swollen phase which confers on them exquisite sensitivity, especially in the exposed parts of the structure, to the effects of other proteins and ligands. The flexibility of different parts of the protein depends on the amount of constraints placed on them from the rest of the protein [30]. From this point of view, it is easy to understand how loops, which are not often stabilized by backbone hydrogen bonds, can play a key role in protein functionality.

It is useful to reconsider how nature uses the variety of amino acids for sequence design. The existence of a presculpted energy landscape with broad minima corresponding to the putative native state structures and the existence of neutral evolution demonstrate that the design of sequences that fit a given structure is relatively easy, leading to many sequences that can fold into a given structure. This freedom facilitates the accomplishment of the next level task of evolution through natural selection: the design of optimal sequences, which not only fold into the desired native state structure, but also fit in the environment of other proteins. A useful protein is one that can interact with other proteins in a synergistic manner and at the same time is not subject to the tendency to aggregate into the harmful amyloid form. This suggests that protein engineering studies aimed at improving enzymatic function ought to be carried out in a two-step manner: first, the family of sequences that fold into a desired target structure needs to be selected and a finer design needs to be carried out in the context of the substrates and the other proteins that the target protein interacts with. Unlike the generality of geometry and symmetry that leads to the menu of native state folds, what we have here is a problem of chemistry acting within the fixed background of the physically determined structures. These considerations suggest that, when the information becomes available, protein-protein interaction networks [89] can be fruitfully viewed not only as the interactions between proteins but also as the interactions between the structures that house them.

The characteristics required for protein native state structures to be targets of an evolutionary process are stability and diversity. Stability is needed because one would not want to mutate away a DNA molecule able to code for a useful protein, and diversity, in order to allow evolution to build complex and versatile forms. The mechanism for natural selection arises naturally in this context—DNA molecules that code for amino acid sequences that fit well into one of these predetermined folds and have useful functionality thrive at the expense of molecules that create sequences that are not useful. Indeed, in this picture, sequences and functionality evolve in order to fit within the constraints of these folds, which, in turn, are immutable and determined by physical law.

VI. THE DENATURED STATE OF PROTEINS

Progress occurs in science through the use of constructive hypotheses with a careful assessment of their consequences. Experiments not only provide valuable hints for selecting between competing hypotheses but are also the ultimate tests of a given hypothesis. There are strong hints from protein experiments that the protein-tube hypothesis is valid. It provides a unification of the various aspects of all proteins: one obtains a presculpted energy landscape with relatively few folds, one can rationalize how a protein might fold in a cooperative manner into its native state conformation, there is the possibility of straightforward design of optimal sequences that fit into a desired structure, the structures are in a marginally compact phase in the vicinity of a phase transition and have the flexibility needed for biological function, and one can understand the formation of amyloids and the role played by the protein structures as a molecular basis for natural selection.

Protein sequence design provides an optimal fit of the sequence with one among the menu of presculpted conformations. The question arises of course as to how a given sequence is able to reach its native state conformation or its home starting from its denatured conformation. The answer to this question entails the understanding of its denatured state [39,90–96]. Unlike the native state which is a somewhat tightly bound set of marginally compact conformations, one envisions the denatured state as an ensemble of somewhat open conformations that the protein adopts when it is not under physiological conditions.

While one may naively think that the denatured state is devoid of any interesting features, recent work has underscored the possibility that the number of accessible conformations is severely reduced compared to a random chain [39,90,91,93–96] leading to biases in the chain direction that persist over the entire length of the protein [94]. Indeed, Shortle [94] has argued that "long-range structure, which cannot be removed by strongly denaturing conditions, could arise predominantly from local steric hindrance." He goes on to state that "not only does the ribosome determine the primary structure of each protein it makes, it also establishes the topological space in which that protein chain will be confined for the rest of its existence."

We build on these insights and the presumed validity of our protein-tube hypothesis by making a second hypothesis that just as there is a one-way correspondence between a sequence and its native state structure, there could exist a similar correspondence between the sequence and its denatured state. In this view, the denatured state can thought of as an address of the native state conformation and lies within its basin of attraction.

Unlike the native state, the denatured state has a larger entropy and comprises somewhat open conformations. Because of this, water plays a quite crucial role in the denatured state. Both the above factors lead to local interactions [97,98] playing a more important role than nonlocal interactions in the denatured state. As can be seen from Fig. 15(a), the local bending energy term is amino acid specific. In addition, in the spirit of the tube model, one might ask whether there are extra geometrical constraints between the local frames of reference (see Fig. 6) of neighboring amino acids along the chain. (As discussed earlier, at the nonlocal level, hydrogen bonds linking different parts of the chain do place geometrical constraints on the reference frames associated with these locations.) Physically, such correlations arise from the fact that in addition to the C^{α} atom that we have considered as a

surrogate for the amino acid, all amino acids but glycine have a C^{β} atom to which the side chain is attached. In a chain of coins (see Sec. II), this corresponds to breaking the symmetry in the plane of the coin. Thus one would quite generally expect that side chain interactions would lead to correlations between the local coordinate frames of nearby amino acids along the sequence [99]. Remarkably, the local steric constraints [49] and the hydrogen bonds [8,9] act in concert and both promote helices and sheets in the native state.

One can ask what the effects of such a local interaction are in the absence of any nonlocal interaction promoting the compaction of the chain. Let us first consider a homopolymer made of just one kind of amino acid. A simple chain molecule with a local bending constraint leads to a tangent-tangent correlation $\langle \hat{t}_i \cdot \hat{t}_{i+n} \rangle$ (see Fig. 6 for the definition of the tangent vector) that decays exponentially in n. Adding a local binormal-binormal interaction term leads quite generally (see the example in Appendix F) to the tangent-tangent correlation decaying exponentially with sequence separation but being modulated with oscillatory behavior. This generic behavior underscores the fact that the class of denatured conformations are not merely featureless but rather already have short-range structure built into them. Indeed, there is a clear reduction in the entropy, due to the short-range binormal-binormal interaction, which is reflected in the oscillations.

The situation is vastly more interesting when one considers a specific sequence of amino acids in its denatured state. It is clear that one ought to have amino acid specific correlations between neighboring coordinate frames which reflect the nature and size of the side chains. An amino acid like proline with its cumbersome side chain configuration can lead to strong constraints in its vicinity whereas glycine which lacks the C_{β} atom and a side chain can provide great flexibility [100] and act as a joker in a card game. Even if such correlations reflect small but systematic deviations from the average behavior [101], these can build up in a very specific way along the sequence leading to a clear imprinting of the native state conformation even in the denatured state. In this context, it is interesting to note that Shortle [94] has shown that "denaturation by at least three different agents—truncation, urea, and acid—gives rise to essentially the same persistent native-state like topology." Furthermore, the alteration of the denatured state by even a single mutation [102] provides further evidence for the structure inherent in the denatured state.

We have shown that the menu of native state structures is determined from generic considerations. Sequence specificity is key in determining whether a given sequence fits particularly well into one of these conformations. Because the menu is large (thousands of conformations), one has diversity. However, because the menu is not too large, a well-designed sequence is able to fold rapidly into its native state conformation. Our hypothesis is that local sequence-specific interactions alone lead to a denatured state which is a reflection of the native state. The denatured state lies in the basin of attraction of the native state and the folding process simply entails the action of the appropriate nonlocal interactions in leading to the protein adopting the native state conformation.

The situation is somewhat reminiscent of a content-addressable memory [103] in which partial information is converted by the brain to recover the complete information. Such content-addressable memories [103] as well as the energy landscape [104] suitable for prebiotic evolution [105] have been modeled through spin glasses [106]. The energy landscape of spin glasses is also characterized by diversity and stability arising from randomness and frustration, which is quite distinct from the the physical mechanisms of short tubes in the marginally compact phase. In conventional spin glasses, randomness, which plays a role somewhat similar to amino acid specific interactions in proteins [107], through frustration sculpts an energy landscape with many local minima. Indeed, a nonrandom exchange interaction between spins would lead to periodic order with much simpler behavior. In spin glasses, starting from a random spin configuration, it is hard to reach a specific local minimum unless the exchange constants are tuned in a clever way as in a content-addressable memory. The landscape is not invariant on changing the exchange interactions and can be fashioned at will. For proteins, on the other hand, our analysis shows that a rich landscape is obtained even in the absence of any sequence heterogeneity and the nature of the ground states is determined by geometry and symmetry and is therefore immutable [35].

An interesting consequence of the type of denatured state described above along with the existence of the presculpted landscape is the possibility of disordered proteins [108]—sequences that are in temporally fluctuating denatured form but which fold in the presence of distinct substrates to carry out vital multiple functionalities. In our picture, these sequences need appropriate stabilizing influences to fold. In the absence of these influences (substrates), the protein is denatured and is located, colloquially, on the fence between different native state structures. Given that finite size effects are severe for proteins, the presence of different substrates (leading to different boundary conditions) would not only favor one competing structure over the others but also result in folding to that structure. The simultaneous existence of the distinct folds in the energy landscape allows the protein to choose from among them depending on the precise nature of the stabilizing influence.

VII. SUMMARY AND PERSPECTIVE

Symmetry and geometry place strong constraints on the types of infinite sized crystal structures and there are exactly 230 distinct space groups in three dimensions [109]. Proteins are finite sized objects. Our analysis demonstrates that the same kind of symmetry and geometrical considerations lead to a finite number of protein folds. This number grows with the size of the protein but is limited by the fact that proteins beyond a characteristic length from either autonomous domains or amyloids. Unlike the crystalline state of matter, proteins are characterized by an inherent anisotropy due to their tubelike character. A given crystalline structure transcends the material that is housed in it—common salt adopts the face-centered-cubic lattice structure as also the well-packed cannonballs of Kepler [18]. Likewise, different se-

quences of proteins can be housed in the same protein fold and yet be able to perform different functionalities [66]. Protein structures are modular in form being simple assemblages of helices and strands connected by tight turns.

The unified picture leads to a single free energy landscape with two distinct classes of structures. The amyloid phase is dominated by β strands linked to each other in a variety of forms whereas the native state structure menu is an assembly of α helices and β structures. Nature has exploited these native state structures in the context of the work horse molecules of life. The selection mechanism for genetic evolution at the molecular level lies in the ability of the protein encoded by the gene to fold well into one of the predetermined folds and have useful function. Unfortunately, however, the proximity of this beautiful phase to the generic amyloid phase underscores how life can easily malfunction as soon as aggregational tendencies of proteins come to the fore. One cannot but marvel at the robustness of life.

An imperfect analogy to the protein problem is a township consisting of around a thousand houses (protein structures), each with its own distinctive style (topology), determined by geometry and symmetry. The form of a house (structure) is the basis of useful functionality. A person (protein) whose tastes (sequence) are (is) especially matched to a given style of house (native state structure) would choose to live in it. Of course, many people (proteins) with similar though not identical tastes (sequences) might choose the same style of house (native state structure). If a person were to arrive in this town, how would she/he know which house to move into? One way would be to explore all the house styles until the dream house is identified. A vastly more efficient situation would occur if the person arrives at the township in the vicinity of the house that she/he will eventually occupy. This would require that the location of the starting point (the denatured conformation) is encoded by the tastes of the person (the sequence) and is within the basin of attraction of her/his dream home (native state structure). This, as yet unproven, scenario would greatly facilitate the folding of a protein into its native state structure accounting for its "surprising simplicity" [44].

The protein problem, which lies at the intersection of many disciplines, is highly complex. Evolution complicates the situation even further. Human design allows for an engineer to devise entirely new ways of accomplishing certain tasks—a classic example is the replacement of vacuum tubes with semiconductor transistors. Nature does not have this luxury in evolutionary design. Nature takes what she has, tinkers with it, and builds on it. Thus the notion of optimal design is not particularly relevant and the future is very strongly correlated with the present and the past. A slightly different turn of events could have led to conspicuously different life forms. This picture of Nature muddling along through evolution combined with the inherent complexity of proteins makes the problem very daunting. Yet within this complexity there is a stunning simplicity provided by the fixed backdrop of the protein folds determined by physical law in the context of which sequences and functionalities are shaped by evolution.

We conclude by revisiting the classic theoretical work of Pauling *et al.* [8,9] and Ramachandran and Sasisekharan [49]. 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. They predicted that helices and sheets are the structures of choice in this regard [Figs. 12(a) and 12(b)]. Ramachandran and his co-workers carried out their pioneering work more than a decade after Pauling. They considered the role of excluded volume or steric interactions between the adjacent amino acids in reducing the available conformational phase space [Fig. 12(c)]. Astonishingly, the two significantly populated regions of the Ramachandran plot correspond to the α helix and the β strand. Even though hydrogen bonds and sterics are not related to each other, they are both promoters of helices and sheets. Is this concurrence of events a mere accident? The marginally compact phase of short tubes has helices and sheets as its preferred structures. In order for Nature to take advantage of this phase of matter, proteins, which obey physical laws, may have been selected to conform to the tube geometry. Hydrogen bonds serve to enforce the parallelism of nearby tube segments, a feature of both helices and sheets, while sterics emphasizes the nonzero thickness of the tube and serves to position it in the marginally compact phase. Because the marginally compact phase is a finite size effect, proteins tend to be relatively short compared to conventional macromolecules including DNA. Indeed, proteins seem to be a vivid example of the adaptation of Nature to her own laws.

In his insightful book *The Fitness of the Environment*, Henderson extended the notion of Darwinian fitness to argue that "the fitness of environment is quite as essential a component as the fitness which arises in the process of organic evolution." Strikingly, the chemistry of proteins ensures that they are self-tuned to occupy the marginally compact phase of short tubes. One cannot but marvel at how several factors, the steric interactions, hydrogen bonds which provide the scaffolding for protein structures, the constraints placed by quantum chemistry on the relative lengths of the hydrogen and covalent bonds and the near planarity of the peptide bonds, and the key role played by water, all reinforce and conspire with each other to place proteins in this phase of matter.

Proteins have proved to be difficult to understand because of their inherent complexity with 20 types of amino acids and the role played by water, because they are relatively short molecules compared to generic manmade polymers and are therefore likely to be characterized by "nonuniversal" behavior, and because of the complexities associated with the random process of evolution. Nevertheless, our work suggests that there is an underlying stunning simplicity. While sequences and functionalities of proteins evolve, the folds that they adopted, which in turn determine function, seem to be determined by physical laws and are not subject to Darwinian evolution. In that regard, these folds may be thought of as immutable or Platonic. Protein folds do not evolve—rather, the menu of possible folds is determined by physical law. In that sense, it is as if evolution acts in the theater of life and shapes sequences and functionalities but does so within the fixed backdrop of the Platonic folds.

Henderson [16] wrote "The properties of matter and the course of cosmic evolution are now seen to be intimately

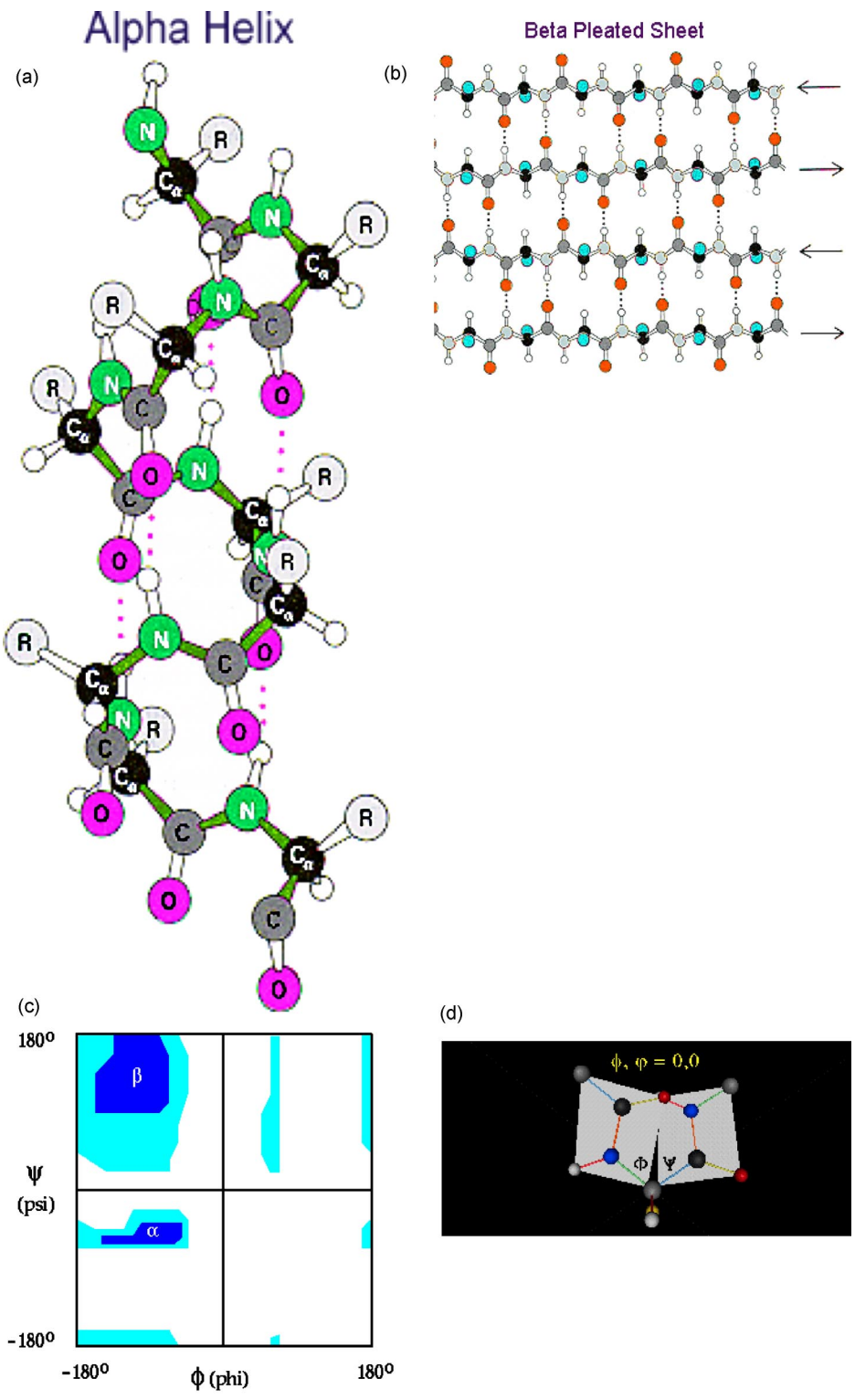

FIG. 12. (Color online) Pauling and Ramachandran revisited: The top row depicts the "classic" structures of an α helix (a) and a pleated β sheet (b). The main-chain backbone atoms and the C^{β} atoms of the side chain groups are shown [color codes are different for (a) and (b)]. Hydrogen bonds, which stabilize the structures, are shown as dashed lines. In the bottom row we show the Ramachandran plot (c) describing how the torsional degrees of freedom (ψ, ϕ), the backbone dihedral angles within an all atom representation, are constrained by steric effects. The shaded (colored online) areas in the plot correspond to allowed regions in conformational space. The structures (a) and (b) stabilized by hydrogen bonding indeed lie squarely within the sterically accessible regions. An example of a dipeptide conformation disallowed because of steric hindrance is shown in (d).

related to the structure of the living being and to its activities; they become, therefore, far more important in biology than has been previously suspected. For the whole evolutionary process, both cosmic and organic, is one, and the biologist may now rightly regard the universe in its very essence as biocentric." His intriguing ideas continue to provoke thought even as we strive to understand the connections between life and the laws of nature.

ACKNOWLEDGMENTS

We are indebted to Phil Anderson, Buzz Baldwin, Paul Chaikin, Hue-Sun Chan, Marek Cieplak, Morrel Cohen, Michael Denton, Ken Dill, Chris Dobson, Russell Doolittle, Sam Edwards, Alessandro Flammini, Leo Kadanoff, Arthur Lesk, Tim Lezon, Davide Marenduzzo, Cristian Micheletti, Harold Scheraga, Michele Vendruscolo, and Saraswathi

Vishveshwara for useful discussions and especially to George Rose for teaching us much about proteins. This work was supported by PRIN 2003, FISR 2001, INFM, NASA, NSF IGERT Grant No. DGE-9987589, the Templeton Foundation, and NSF MRSEC at Penn State.

APPENDIX A: THREE-BODY DESCRIPTION OF A TUBE

In this appendix we will describe how a suitable three-body potential [19] characterizes the self-avoidance of a tube of thickness Δ whose axis $\mathcal{C}$ is a *smooth* curve $\mathbf{r}(s)$, parametrized by its arclength s with $0 \leqslant s \leqslant L$, L being the total length of the tube. The tube is a one-dimensional generalization of the zero-dimensional hard sphere case as described in the text. The self-avoidance of an ensemble of hard spheres, each of radius Δ, can be ensured by requiring that none of the distances between all pairs of sphere centers is less than 2Δ.

Let us consider, first, a closed curve, i.e., $\mathbf{r}(0)=\mathbf{r}(L)$. At each position s along the curve $\mathcal{C}$ we consider an infinitesimally thin circular disk of radius Δ, $\Sigma(s,\Delta)$, centered at the point $\mathbf{r}(s)$ and perpendicular to the tangent vector $d\mathbf{r}(s)/ds$ at s. The tube is simply the union of all the disks. The self-avoidance is imposed by requiring that pairs of disks at different points do not intersect, $\Sigma(s,\Delta)\cap\Sigma(s',\Delta)=\varnothing\ \forall s,s'$.

There is an easier way to implement the self-avoidance (steric constraints), which underscores the key difference between the hard sphere and the tube problem. Indeed, in the latter case, there are two classes of lengths which are relevant to the steric interaction: the radius of curvature $|\ddot{\mathbf{r}}(s)|^{-1}$ at each position s and the closest approach distances [note that $|\dot{\mathbf{r}}(s)|=1$ within the arclength parametrization]. A closest approach occurs at, say, points $\mathbf{r}(s_1)$ and $\mathbf{r}(s_2)(s_1 \neq s_2)$ when $\mathbf{r}(s_1)-\mathbf{r}(s_2)$ is perpendicular to both tangent vectors at s_1 and s_2. For a *smooth* closed curve there is at least one such closest approach. It is rather intuitive [110,111] that a necessary and sufficient condition for the self-avoidance is that Δ be less than the minimum among $|\ddot{\mathbf{r}}(s)|^{-1}\ \forall s$ and $(1/2)|\mathbf{r}(s_1)-\mathbf{r}(s_2)|\ \forall s_1,s_2$ where $\dot{\mathbf{r}}(s_i), i=1,2$, are both perpendicular to $\mathbf{r}(s_1)-\mathbf{r}(s_2)$. This minimum is called the thickness $\Delta(\mathcal{C})$, of the curve $\mathcal{C}$ [110,111]. The minimum among the closest approach distances is analogous to the minimum among all distances between pairs of centers in the hard sphere problem. The fact that the tube is a linear object introduces another length in the problem which is the minimum among all radii of curvature and it is local in nature in the sense that it involves nearby points of the curve $\mathcal{C}$: the radius of curvature at position s represents the radius of the circle that best approximates the curve $\mathcal{C}$ at s.

We now turn to a reformulation of the tube self-avoidance constraint in a much more appealing way that makes it more similar to the self-avoidance recipe for hard spheres. Following Ref. [20] let us consider a triplet of positions along the curve $\mathcal{C}$ (instead of a pair of centers as in the hard sphere problem), $\mathbf{r}_i \equiv \mathbf{r}(s_i), i=1,2,3$. These positions define a plane and hence a unique circle through them whose radius is

$$r(\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3)=\frac{|\mathbf{r}_2-\mathbf{r}_1||\mathbf{r}_3-\mathbf{r}_1||\mathbf{r}_3-\mathbf{r}_2|}{4A(\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3)}, \tag{A1}$$

where $A(\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3)$ is the area of the triangle whose vertices are $\mathbf{r}_1,\mathbf{r}_2$, and $\mathbf{r}_3$. The theorem proved in Ref. [20] states that

$$\Delta'(\mathcal{C}) \equiv \min_{s_1,s_2,s_3} r(\mathbf{r}(s_1),\mathbf{r}(s_2),\mathbf{r}(s_3)) = \Delta(\mathcal{C}), \tag{A2}$$

where the s's do not need to be distinct. Indeed it is easy to show that when $s_1,s_2,s_3 \to s$ then $r(\mathbf{r}(s_1),\mathbf{r}(s_2),\mathbf{r}(s_3)) \to |\mathbf{r}(s)|^{-1}$, the radius of curvature at s. Furthermore it is not difficult to show that the search for the minima in Eq. (A2) can be restricted to

$$\lim_{s_2\to s_1} r(\mathbf{r}(s_1),\mathbf{r}(s_2),\mathbf{r}(s_3)) \equiv r(\mathbf{r}_1,\mathbf{r}_1,\mathbf{r}_3), \tag{A3}$$

which is the radius of the circle through the point $\mathbf{r}(s_3)$ and $\mathbf{r}(s_1)$ and tangent to the curve at the latter point.

Let us assume that the minimum in Eq. (A2) is reached at three distinct points $\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3$ and let us consider the sphere of radius $\Delta'(\mathcal{C})$, Eq. (A2), through them. If it is not tangent to the curve in at least two of the points $\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3$ we have a contradiction. Indeed if the sphere is tangent to the curve at one or none of the three points we can shrink the sphere slightly still keeping three intersections with the curve. However this is a contradiction since due to the definition of thickness, Eq. (A2), any sphere of radius less than $\Delta'(\mathcal{C})$ cannot intersect the curve in more than two points. Thus say that one of the points where the tangency occurs is $\mathbf{r}_1$. Since the circle through $\mathbf{r}_1$ and $\mathbf{r}_2$ and tangent to the former lies on the sphere it implies that $r(\mathbf{r}_1,\mathbf{r}_1,\mathbf{r}_2) \leqslant \Delta'(\mathcal{C})$ which is a contradiction unless the equality holds. This demonstrates that the minimum in Eq. (A2) is never exclusively reached at three distinct points along the curve. The above argument leads also to the proof of the theorem. In fact if the other tangency point is, say, $\mathbf{r}_2$, then in addition to $r(\mathbf{r}_1,\mathbf{r}_1,\mathbf{r}_2)=\Delta'(\mathcal{C})$, one also has $r(\mathbf{r}_2,\mathbf{r}_2,\mathbf{r}_1)=\Delta'(\mathcal{C})$. One may immediately prove that this can occur only if the tangent vectors $\dot{\mathbf{r}}(s_i), i=1,2$, are perpendicular to $\mathbf{r}(s_1)-\mathbf{r}(s_2)$. Thus $\min_{s_1,s_2,s_3} r(\mathbf{r}(s_1),\mathbf{r}(s_2),\mathbf{r}(s_3))$ captures simultaneously both the radius of curvature and the distances of closest approaches, consequently proving the equality (A2).

The local thickness of the tube (global radius of curvature in Ref. [20]), at each $\mathbf{r}(s) \in \mathcal{C}$, may be defined as

$$\Delta_{\mathbf{r}(s_1)}(\mathcal{C}) = \min_{s_2,s_3} r(\mathbf{r}(s_1),\mathbf{r}(s_2),\mathbf{r}(s_3)). \tag{A4}$$

Of course the thickness $\Delta(\mathcal{C})$ is the minimum of $\Delta_{\mathbf{r}(s)}(\mathcal{C})$ as $\mathbf{r}(s)$ varies on $\mathcal{C}$. Another theorem proved in Ref. [20] states that if $\mathcal{C}$ can be deformed smoothly in order to maximize the thickness without changing the knot type, the resulting curve $\mathcal{C}^*$, called the "ideal shape" of the given knot type, has $\Delta_{\mathbf{r}(s)}(\mathcal{C}^*)=\Delta(\mathcal{C})$ for all points where $|\ddot{\mathbf{r}}(s)| \neq 0$. Figure 5 is a histogram of local thicknesses for a sample of native protein structures. The variations of the local thickness around the average value 2.7 Å is about 7%.

What we learn from the above mathematical framework is that a mere pairwise interaction does not suffice to describe the steric constraint of a tube whose axis is a string $\mathcal{C}$ [19]. This is because, in addition to the distance between two points on a string, one also needs to know the context, i.e., the local direction of the string in the proximity of the points themselves. Let us consider a three-body potential $V(r(\mathbf{r}_1,\mathbf{r}_1,\mathbf{r}_3))$ characterizing the interaction between three

particles on the axis of the string in terms of the radius of the circle through them (notice that this potential is invariant under translation, rotation, and permutation of the three points). $V(r)$ could be the same as commonly used in the hard sphere problem, i.e., $V(r)=\infty$ when $r<\Delta$ and $V(r)=0$ otherwise (in the hard sphere problem r is half of the distance between a pair of sphere centers). This length scale neatly solves the contextual problem mentioned above. When two parts of a chain come together, the radius of a circle passing through two of the particles on one side of the chain and one particle from the other side of the chain turns out to be a measure of the distance of approach of the two sides of the chain. On the other hand, when one considers three particles consecutively along the chain, the radius of the circle passing through them is simply the local radius of curvature. Indeed when three such particles form a straight line, the radius goes to infinity and the three particles essentially become noninteracting. The straight line configuration is the best that the particles can do in terms of staying away from each other given that they are constrained to be neighbors along the chain. In the case of a polymer chain, such as a protein, a tube whose axis is a *smooth* string is clearly an approximation. One ought to introduce a *discrete* curve $\{\mathbf{r}_1,\mathbf{r}_1,\ldots,\mathbf{r}_N\}$, and the continuous variable s of the curve now becomes discrete. In correspondence with the considerations above, one may again define the thickness of a discrete curve $\mathcal{C}$ as [20]

$$\Delta(\mathcal{C})=\min_{i,j,k} r(\mathbf{r}_i,\mathbf{r}_j,\mathbf{r}_k), \tag{A5}$$

where now i,j, and k are all distinct. For a discrete curve there is no guarantee that the minimum is obtained from among $r(\mathbf{r}_i,\mathbf{r}_j,\mathbf{r}_k)$ with at least two of the three indices separated by one unit (e.g., $j=i\pm1$), but one can still distinguish between a local and a nonlocal contribution to the thickness, according to whether i,j,k are consecutive along the chain or not. In the latter case, the minimum obtained from among $r(\mathbf{r}_i,\mathbf{r}_j,\mathbf{r}_k)$ gives half the minimum distance of closest approach computed for the discrete chain. Similarly, there is no simple restriction of the triplets when one deals with open continuous curves with free ends.

APPENDIX B: TUBE VERSUS STRING AND BEADS MODEL

In this appendix we summarize the main differences between the thick polymer (TP) model that we deal with in this work and the Edwards model [112] (EM) in the presence of a bending rigidity term (the analog of the Edwards model in the discrete case is the usual string and bead model). In both cases, one may add a twist rigidity term, which we neglect here, for simplicity. Let us consider the case of continuous chains. The Hamiltonian for the generalized EM is

$$\mathcal{H}_{\mathrm{EM}}(\{\mathbf{r}\})=\frac{1}{2}\int_0^L \dot{\mathbf{r}}(s)^2 ds+\frac{\kappa_b}{2}\int_0^L \ddot{\mathbf{r}}(s)^2 ds+\frac{v_2}{6}\int_0^L\int_0^L \delta(\mathbf{r}(s)-\mathbf{r}(s'))ds\,ds'+\frac{v_3}{90}\int_0^L\int_0^L\int_0^L \delta(\mathbf{r}(s)-\mathbf{r}(s'))\times\delta(\mathbf{r}(s)-\mathbf{r}(s''))ds\,ds'\,ds''. \tag{B1}$$

The self-avoidance in the TP model [113] is given by

$$\mathcal{H}_{\mathrm{TP}}(\{\mathbf{r}\})=\int_0^L\int_0^L\int_0^L V(R_c(\mathbf{r}(s),\mathbf{r}(s'),\mathbf{r}(s'')))ds\,ds'\,ds'', \tag{B2}$$

where $R_c(\mathbf{r}(s),\mathbf{r}(s'),\mathbf{r}(s''))$ is the radius of the circle through the three points $\mathbf{r}(s)$, $\mathbf{r}(s')$, $\mathbf{r}(s'')$ and

$$V(r)=\begin{cases}\infty & \text{if } r<R_0,\\ -1 & \text{if } R_0<r<R_1,\\ 0 & \text{if } R_1<r.\end{cases} \tag{B3}$$

Note that in the limit of a continuous chain the EM needs the introduction of singular potentials, in order to deal with the fact that a two-body potential is unable to distinguish whether two nearby beads are far apart or not along the chain. Within the context of the Edwards model such singularities can then be treated successfully within a perturbative renormalization group approach [115]. On the other hand, the need of singular potentials is deftly avoided when using the three-body prescription implied by the thickness constraint in the TP model [113].

The details of the discretization scheme matter for a discrete chain. First, the discretization introduces a natural cutoff length scale. Second, the three-body potential V of Eq. (B2) cannot be used by itself for a discrete chain. Indeed, in the absence of a two-body repulsion, the chain would collapse onto a circle of radius between R_0 and R_1 and would wind repeatedly along it.

1. High temperature phase

It is well known that in the high temperature regime the critical behavior of the EM in the limit of very long chains is governed by the exponent $\nu\simeq0.58$, so that a typical length ξ measuring the spatial extension of the chain scales as $\xi\sim L^{\nu}$, where L is the chain length. The chain is *swollen* with respect to the Gaussian random walk behavior for which $\nu=1/2$. The same feature holds for the TP model; in the high temperature regime the different symmetry properties induced by the inherent anisotropy of a thick tube are averaged out, and a chain of coins shares the same critical behavior as a chain of spheres.

Interestingly, other features such as the form of the two-point tangent-tangent correlation function along the chain differentiate the TP model from the EM. In the absence of twisting rigidity [the *intrinsic* twist of the chain, as defined by the *torsion* of the corresponding curve for the EM or the axis of the tube for the TP model, is described by the energy term $(\kappa_t/2)\int_0^L \dot{\hat{b}}(s)^2 ds$, where $\hat{b}(s)$ is the *binormal* vector which is part of the Frenet triad], one gets a simple exponential decay in both cases. However, when the twisting rigidity κ_t is introduced, the EM exhibits an oscillatory decaying correlation function for any value of κ_t, whereas the TP model crosses over from simple to oscillatory decay on increasing κ_t (see Appendix F) [116]. The existence of a transition line in the parameter space (κ_b,κ_t) separating simple from oscillatory decay is an interesting feature of the TP model.

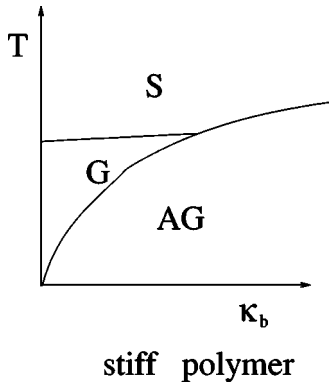

FIG. 13. Schematic phase diagram of an EM in the temperature (T) bending rigidity (κ_b) plane [117]. The different phases are S= swollen, G= globule, and AG= asymmetric globule.

2. Persistence length

Another similarity between the TP model and the EM is the following: at any *fixed* value of the length L of the chain and of the temperature T, the persistence length l_p (which is a measure of the distance along the chain after which the tangent vectors become uncorrelated) diverges both for the TP model, in the limit $\Delta \to \infty$ of infinite thickness, and for the EM, in the limit $\kappa_b \to \infty$ of infinite bending rigidity. The thickness constraint indeed stiffens the chain locally. Yet a closer look reveals an important difference between the TP model and the EM. In an ideal case in which nonlocal interactions are disregarded we get a different scaling behavior. For the EM, $l_p \sim \kappa_b / k_B T$, whereas in the TP model the persistence length does *not* increase at low temperatures, a first hint that the low temperature behavior of the TP model may be radically different from that of the EM.

3. Low temperature phase

The anisotropy inherent in the thick tube description strongly affects the behavior of the TP model at low temperatures, as can be seen by comparing the bending rigidity/temperature phase diagram (Fig. 13) for the EM and the corresponding thickness/temperature phase diagram (Fig. 14) for the TP model *in the thermodynamic limit*.

Let us first note that whereas the swollen (S) and the (disordered compact) globule (G) phase share similar features in the two cases (but see the above discussion concerning the correlation function properties in the swollen phase), the asymmetric globule (AG) (semicrystalline) phase is different. The persistence length diverges (strictly at $T=0$) with the chain length for both EM and TP model (the chain is locally straight). For the latter, this arises from the interplay of the thickness constraint and the interaction promoting compaction, so that the resulting ground state conformation will have tube segments aligned with respect to one another similar to the Abrikosov flux lattice, filling the space with hexagonal symmetry. For the former, this is a mere consequence of the local bending rigidity, so that ground state conformations will likely consist of planes stacked onto each other with parallel (or antiparallel) alignment within the same plane, but not necessarily between different planes.

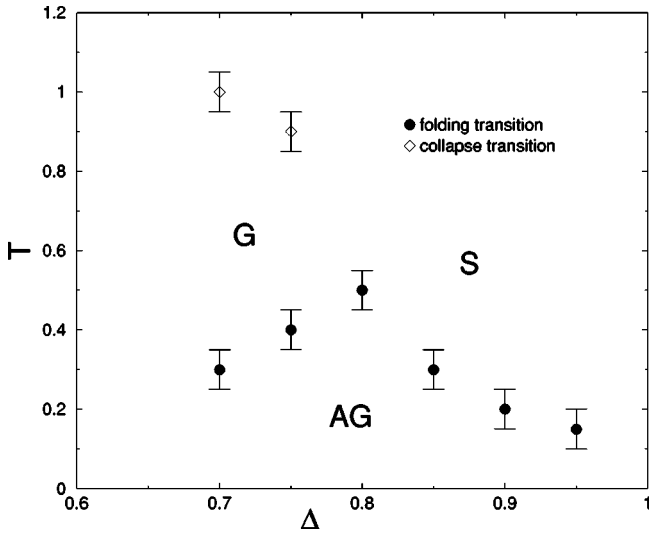

FIG. 14. Phase diagram for a thick polymer chain in the temperature-thickness (T, Δ) plane obtained with Monte Carlo simulations (see [22] for details). The different phases are S= swollen, G= globule, and AG= asymmetric globule.

A second crucial difference is that *in the limit of zero temperature* the EM is in the AG phase for all finite values of the bending rigidity, whereas the TP model exhibits a transition from the AG phase to the swollen phase with increasing thickness. This has profound consequences, especially when finite size effects are taken into account. It is instructive to revisit the phase diagram at $T=0$ for a TP in the plane $(L/R, \Delta/R)$ (see Fig. 2). If $\Delta > R$, the chain cannot take advantage of the attraction, the length of the chain does not play any role and in the $L \to \infty$ limit one gets the critical behavior of the swollen phase. When $\Delta < R$, in the thermodynamic limit $L \to \infty$, the chain is in the asymmetric globule phase resembling the Abrikosov flux lattice with hexagonal symmetry, but different phenomena occur for finite chain length. If $L < 2\pi R$ *all* parts of the chain are able to interact with each other. When $L > 2\pi R$, this is still true for small enough thickness. However, as the thickness increases, this is not possible anymore and the chain adopts a conformation which optimizes the attractive interaction. In the long chain limit the boundary line between the two regimes scales as $L/R \sim (R/\Delta)^2$. This result is obtained by equating the volume occupied by the tube $L\Delta^2$ to the volume of the sphere of attraction R^3. For shorter chains the compact regime at intermediate thickness in which the chain seeks to compact itself within the constraint of the thickness is indeed *marginal*, being sandwiched between the featureless compact [118] and the swollen regimes described above. It is precisely in this window of parameter space that we find marginally compact ground state structures such as space filling helices. This finite size feature of the TP model is quite robust independent of the details introduced, for instance in the discrete case. None of these features are present for the EM case for which there is no dependence whatsoever on the bending rigidity at $T=0$.

APPENDIX C: OPTIMAL HELIX

In this appendix we derive the value c^* of the pitch/radius ratio c of an optimal space filling helix. The radius of curva-

ture of such a helix equals the tube radius and is equal to half the minimum distance of closest approach between different turns of the helix.

The parametric equation of a helix is

$$\mathbf{x}(t) = (r\cos t, r\sin t, vt), \tag{C1}$$

where the pitch/radius ratio is $c=2\pi v/r$. The tangent and the acceleration vectors are

$$\dot{\mathbf{x}}(t) = (-r\sin t, r\cos t, v), \tag{C2}$$

$$\ddot{\mathbf{x}}(t) = (-r\cos t, -r\sin t, 0). \tag{C3}$$

Since $\dot{\mathbf{x}}(t)\cdot\ddot{\mathbf{x}}(t)=0$, the radius of curvature is simply given by

$$\rho_L = \frac{|\dot{\mathbf{x}}(t)|^2}{|\ddot{\mathbf{x}}(t)|} = r\left(1+\frac{v^2}{r^2}\right) \tag{C4}$$

independently of t.

We define the nonlocal radius of curvature as half the distance of closest approach between successive turns of the helix. Fix a point $\mathbf{A}=\mathbf{x}(t)$ on the curve, and compute the distance $d(s,t)=|\mathbf{B}-\mathbf{A}|$ from a second point $\mathbf{B}=\mathbf{x}(s)$ moving along the curve as a function of s. The nonlocal radius is then

$$\rho_{NL}(t) = \frac{1}{2}\min_{s\neq t}\{d(s,t)\}, \tag{C5}$$

with the requirement that $\partial d(s,t)/\partial s=0$ at some $s^*\neq t$, implying that $\mathbf{B}-\mathbf{A}$ is perpendicular to the tangent vector $\dot{\mathbf{x}}(s^*)$. Note that the nonlocal radius need not exist (for open curves) and is in principle a varying function of t, when the curve is not invariant under translation along it.

Because the helix is invariant under translation along the curve, implying that $\mathbf{B}-\mathbf{A}$ is perpendicular also to the tangent vector $\dot{\mathbf{x}}(t)$, we can choose $t=0$ so that

$$d^2(s) \equiv d^2(s,0) = r^2\left(2(1-\cos s)+\frac{v^2}{r^2}s^2\right). \tag{C6}$$

The condition allowing one to get extremal points for $d^2(s)$ is

$$\sin s + s\frac{v^2}{r^2} = 0. \tag{C7}$$

One trivial solution of this equation is $s=0$ and there is no other solution for sufficiently high pitch to radius ratio $c=2\pi v/r$. If c is decreased, new solutions appear, two at a time, the smaller a maximum and the greater a minimum, corresponding to the increasing packing of helix turns. We are interested in the minimum s^* corresponding to $\mathbf{A}$ and $\mathbf{B}$ staying on two consecutive turns, that is $\pi<s^*<2\pi$. For sufficiently low c, the above equation then defines the implicit function $s^*(c)$, and one has

$$\rho_{NL} = \frac{1}{2}d[s^*(c)] \tag{C8}$$

for all points of the helix. In the limit $c\ll 1$, one has $s^*\simeq 2\pi$ and thus $\rho_{NL}\simeq \pi v = p/2$, where $p=2\pi v$ is the pitch of the helix, as expected. The particular value $c=c^*$, for which the local and the nonlocal radii of curvature are equal, is then defined by

$$\frac{d[s^*(c^*)]}{2r} = 1+(c^*)^2. \tag{C9}$$

Thus, according to the definition of thickness given in Appendix A, $\Delta_{\text{helix}}=\rho_L$ if $c>c^*$, since the radius of curvature is smaller than the nonlocal radius. A tube swelling around the helix would stop increasing due to local singularities, leaving space between the successive turns of the helix. On the other hand, the nonlocal radius is smaller than the radius of curvature if $c<c^*$, implying $\Delta_{\text{helix}}=d[s^*(c)]/2$. In such a case, the tube would stop swelling due to self-intersection between different turns, leaving a hole in the middle of the helix. At $c=c^*$, one obtains an optimal space filling helix with a special pitch to radius ratio of $c^*\approx 2.512$ (shown in Fig. 4).

APPENDIX D: GEOMETRICAL CONSTRAINTS DETERMINED FROM EXPERIMENTAL DATA

In this appendix we describe the data analysis used to elucidate the geometrical constraints imposed by sterics and hydrogen bonds (see Fig. 15). We have used a database of 600 different protein native structures [119] consisting of sequences varying in length from 44 to 1017, with low sequence homology and covering many different three-dimensional folds according to the Structural Classification of Proteins scheme [120]. Panel (a) depicts the histogram of the local radius of curvature associated with two classes of triplets, the first [shown in light gray (red online)] featuring strong α-helix-forming amino acids (LEU, ALA, GLU) and the second [shown in dark gray (blue online)] featuring β-strand formers (VAL, ILE, TYR) [13] and underscores the vital role of chemistry in choosing from among the menu of native state folds. The vertical dashed line indicates the threshold length scale chosen in the model for the curvature energy penalty. The remaining panels show histograms for several quantities involved in the definition of hydrogen bonds: the C^α-C^α distance between $i,i+3$ atoms given that i, $i+1$, $i+2$, $i+3$ all belong to a helix (b) and between i,j (with $j>i+4$) atoms given that i, j, belong to a β strand (c); the scalar products $\hat{b}_i\cdot\hat{b}_j$ (d) and $(\hat{b}_i+\hat{b}_j)\cdot\hat{r}_{ij}/2$ (e) for i,j contacts [with $|j-i|=3$light gray (red online) and with $|j-i|>4$ dark gray (blue online) provided that no closer interstrand contact is present among $i\pm1, j\pm1$] ($\hat{b}_i$ is the binormal vector at atom i and $\hat{r}_{ij}$ is the vector joining atoms i and j normalized to unit length); and the scalar product $\hat{b}_i\cdot\hat{b}_{i+1}$ for consecutive residues along a β strand (f). In each case, the dashed lines and arrows depict the approximate constraints used in our model. All histograms are normalized in such a way that a flat distribution would have a constant unit height.

APPENDIX E: DETAILS OF MODEL AND MONTE-CARLO SIMULATIONS

The protein backbone is modeled as a chain of C^α atoms with a fixed distance of 3.8 Å between successive atoms along the chain, an excellent assumption for all but non-*cis*-proline amino acids [13]. The geometry imposed by chemis-

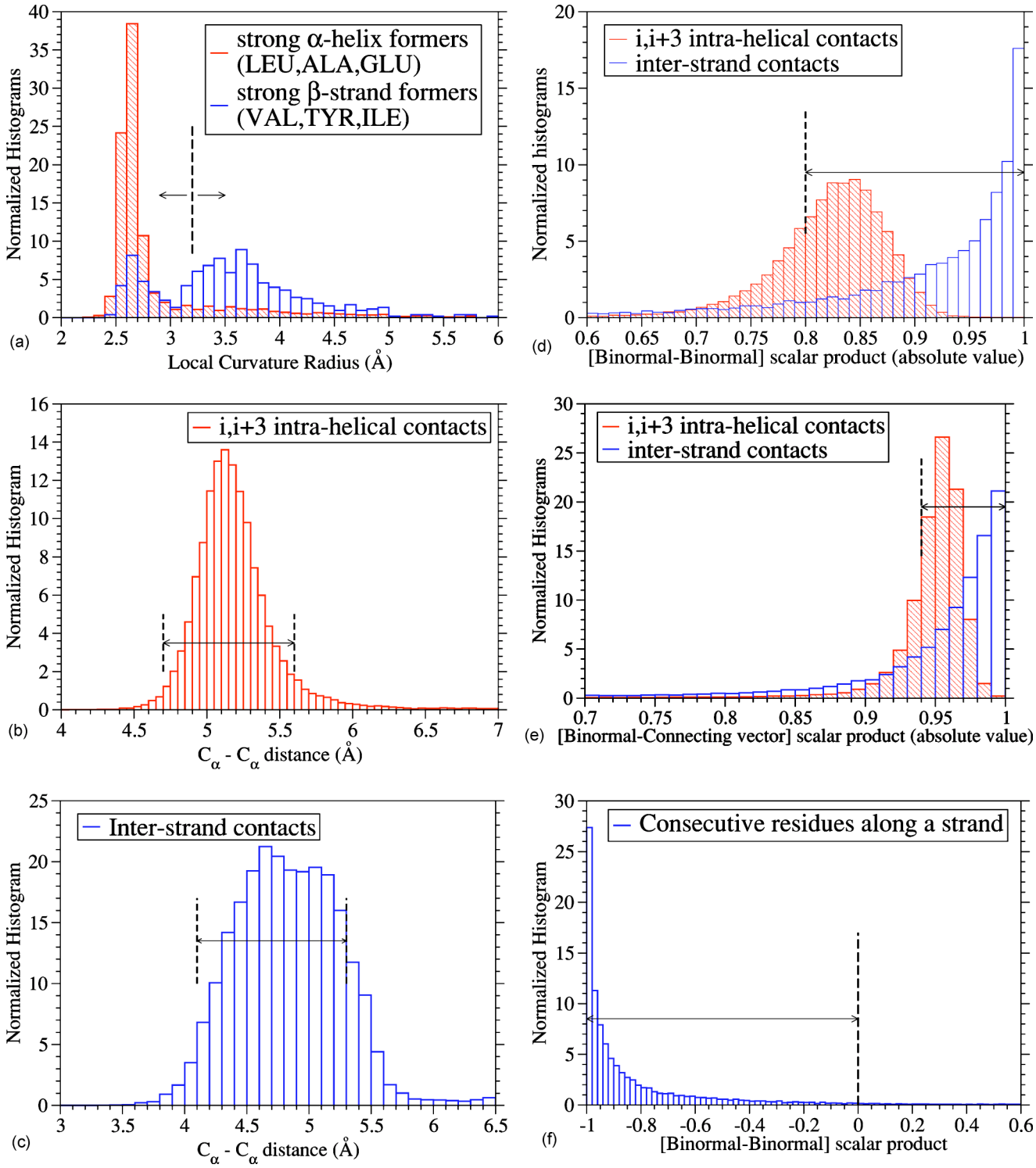

FIG. 15. (Color online) Statistical analysis of several quantities computed for residues classified as participating in secondary structures in protein native state structures from the Protein Data Bank. Light (dark) gray [red (blue) online] histograms refer to residues participating in α helices (β strand).

try dictates that the bond angle associated with three consecutive C^{α} atoms is between 82° and 148°.

1. Tube geometry

Self-avoiding conformations of the tube whose axis is the protein backbone are identified by considering all triplets of C^{α} atoms and drawing circles through them and ensuring that none of their radii is smaller than the tube radius [19]. At the local level, the three-body constraint ensures that a flexible tube cannot have a radius of curvature any smaller than the tube thickness in order to prevent sharp corners whereas, at the nonlocal level, it does not permit any self-intersections.

The backbone of C^α atoms is treated as a flexible tube of radius 2.5 Å, a constraint imposed on all (local and nonlocal) three-body radii, an assumption validated for protein native structures [48].

2. Sterics

Steric constraints require that no two nonadjacent C^α atoms are allowed to be at a distance closer than 4 Å. Ramachandran and Sasisekharan [49] showed that steric considerations based on a hard sphere model lead to clustering of the backbone dihedral angles in two distinct α and β regions for nonglycyl and nonprolyl residues. The two backbone geometries that allow for systematic and extensive hydrogen bonding [8–10] are the α helix and the β sheet obtained by a repetition of the backbone dihedral angles from the two regions, respectively [43]. Short chains rich in alanine residues, which are a good approximation to a stretch of the backbone, can adopt a helical conformation in water (see [121] for a detailed discussion of experimental conditions necessary to achieve this). However, when one has more heterogeneous side chains, the helix backbone could sterically clash with some side-chain conformers resulting in a loss of conformational entropy [122]. When the price in side-chain entropy is too large, an extended backbone conformation results pushing the segment towards a β-strand structure [43]. These steric constraints are approximately imposed through an energy penalty (denoted by e_R) when the local radius of curvature is between 2.5 Å and 3.2 Å. (The magnitude of the penalty does not depend on the specific value of the radius of curvature provided it is between these values.) There is no cost when the local radius exceeds 3.2 Å. Note that the tube constraint does not permit any local radius of curvature to take on a value less than the tube radius, 2.5 Å.

3. Hydrogen bonds

We do not allow more than two hydrogen bonds to form at a given C^α location. In our representation of the protein backbone, local hydrogen bonds form between C^α atoms separated by three along the sequence with an energy defined to be -1 unit, whereas nonlocal hydrogen bonds are those that form between C^α atoms separated by more than four along the sequence with an energy of -0.7. This energy difference is based on experimental findings that the local bonds provide more stability to a protein than do the nonlocal hydrogen bonds [123]. Cooperativity effects [124] are taken into account by adding an energy of -0.3 units when consecutive hydrogen bonds along the sequence are formed. There is some latitude in the choice of the values of these energy parameters. The results that we present are robust to changes (at least of the order of 20%) in these parameters.

4. Geometrical constraints due to hydrogen bonding

For hydrogen bond formation between atom i and j, the distance between these atoms ought to be between 4.7 Å and 5.6 Å (4.1 Å and 5.3 Å) for the local (nonlocal) case [see Fig. 15(b) for the local case]. A study of protein native state structures reveals an overall nearly parallel alignment of the axes defined by three vectors: the binormal vectors at i and j and the vector $\mathbf{r}_{ij}$ joining the i and j atoms. A hydrogen bond is allowed to form only when the binormal axes are constrained to be within 37° of each other, whereas the angle between the binormal axes and that defined by $\mathbf{r}_{ij}$ ought to be less than 20° [see Fig. 15(e)]. Additionally, for the cooperative formation of nonlocal hydrogen bonds, one requires that the corresponding binormal vectors of successive C^α atoms make an angle greater than 90° [see Fig. 15(f)]. The first and the last residues of the chain are special cases since their binormal vectors are not defined. In order for such residues to form a hydrogen bond (with each other or with other internal residues in the chain), it is required that the angle between the associated ending peptide link and the connecting vector to the other residue participating in the hydrogen bond is between 70° and 110°. As in real protein structures, when helices are formed, they are constrained to be right handed. This is enforced by requiring that the backbone chirality associated with each local hydrogen bond is positive. The chirality is defined as the sign of the scalar product $(\mathbf{r}_{i,i+1}\times\mathbf{r}_{i+1,i+2})\cdot\mathbf{r}_{i+2,i+3}$ [125].

5. Hydrophobic interactions

The hydrophobic (hydrophilic) effects mediated by the water are captured through a relatively weak interaction e_W (either attractive or repulsive) between C^α atoms which are within 7.5 Å of each other. Note that hydrogen bonds can easily be formed between the amino acid residues in an extended conformation and the water molecules. Within our model, the intrachain hydrogen bond interaction introduces an effective attraction, because water molecules are not explicitly present. The hydrophobicity scale is thus renormalized (e.g., even when e_W is weakly positive, there could be an effective attraction resulting in structured conformations such as a single helix or a planar sheet). A negative e_W is, in any case, crucial for promoting the assembly of secondary motifs in native tertiary arrangements.

Monte Carlo simulations are carried out with pivot and crankshaft moves commonly used in stochastic chain dynamics [126]. A METROPOLIS procedure is employed with a thermal weight $\exp(-E/T)$, where E is the energy of the conformation and T is the effective temperature.

APPENDIX F: CORRELATION FUNCTIONS IN THE DENATURED STATE

We will consider a polypeptide chain in a *phase* where the local interactions dominate the behavior of the correlation functions, to be studied below, at least at short and intermediate distances along the chain. We thus neglect the steric interactions apart from the effect that they have on neighboring nodes of the chain. The correlation functions we will consider involve the unit vectors $\hat{t}_i$ parallel to $\mathbf{r}_{i+1}-\mathbf{r}_i$ and the binormal $\hat{b}_i \equiv (\hat{t}_i\times\hat{t}_{i-1})/|\hat{t}_i\times\hat{t}_{i-1}|$. Note that, in order to facilitate the calculations, our definition for the tangent vector is different from the one used in Fig. 6 in Sec. IV. The geometrical constraints of hydrogen bond formation are associated with the binormal vector, whose definition is

unchanged—the binormal vector is perpendicular to the plane defined by $\mathbf{r}_{i-1},\mathbf{r},\mathbf{r}_{i+1}$. Let $\theta_i \in (0,\pi)$ be the angle between $\hat{t}_i$ and $\hat{t}_{i-1}$ and $\phi_i \in (-\pi,\pi)$ the angle by which the plane defined by $\mathbf{r}_{i-1},\mathbf{r}_i$, and $\mathbf{r}_{i+1}$ is rotated along the axis $\hat{t}_i$ with respect to the plane defined by $\mathbf{r}_{i-2},\mathbf{r}_{i-1}$, and $\mathbf{r}_i$. Quite generally the joint probability distribution of angles, $\mathcal{P}(\theta_2,\phi_3,\theta_3,\phi_4,\ldots)$ will depend on the entire ensemble of interactions including the steric interactions. However in the *phase* we wish to study we will assume that this probability distribution can be factorized, i.e., we will consider the case where we have probability distributions $\rho(\theta_i,\phi_i)$ for each pair of angles θ_i,ϕ_i with $i=3,4,\ldots$ and

$$\mathcal{P}(\theta_2,\phi_3,\theta_3,\phi_4,\theta_4,\ldots)=\rho_2(\theta_2)\prod_{i\geqslant 3}\rho_i(\theta_i,\phi_i), \tag{F1}$$

where the contribution for the angle θ_2 between the first two vectors of the chain, $\hat{t}_1$ and $\hat{t}_2$, has been selected out. The average with respect to $\mathcal{P}$ will be written as $\langle\cdot\rangle_{\mathcal{P}}$ whereas the average with respect to $\rho_i(\theta,\phi)$ will be denoted simply as $\langle\cdot\rangle_i$. In the case of a protein sequence the $\rho_i(\theta_i,\phi_i)$ depends explicitly on the type of amino-acids in the neighborhood of the *i*th position. It is this dependence that ultimately will determine the propensity of a given segment of the protein sequence to be in a given secondary structure. One can straightforwardly derive the following recursion relations:

$$\hat{t}_i=-\hat{t}_{i-2}\frac{\sin\theta_i\cos\phi_i}{\sin\theta_{i-1}}+\hat{t}_{i-1}(\cos\theta_i\sin\theta_i\cos\phi_i\cot\theta_{i-1})+\hat{b}_{i-1}\sin\theta_i\sin\phi_i, \tag{F2}$$

$$\hat{b}_i=+\hat{t}_{i-2}\frac{\sin\phi_i}{\sin\theta_{i-1}}-\hat{t}_{i-1}\cot\theta_{i-1}\sin\phi_i+\hat{b}_{i-1}\cos\phi_i. \tag{F3}$$

If one wishes to calculate the correlation function $\langle\mathbf{x}\cdot\hat{t}_i\rangle_{\mathcal{P}}$, where $\mathbf{x}$ is $\hat{t}_2$, $\cot\theta_2\hat{t}_2$, $\hat{b}_2$, or any other combination of them, then one needs to introduce other correlation functions in order to have a closed ensemble of recursion equations. By defining the vector

$$\mathbf{V}_i=\begin{pmatrix}\langle\mathbf{x}\cdot\hat{t}_i\rangle_{\mathcal{P}}\\ \langle\mathbf{x}\cdot\hat{t}_{i-1}\rangle_{\mathcal{P}}\\ \langle\mathbf{x}\cdot\hat{t}_i\cot\theta_i\rangle_{\mathcal{P}}\\ \langle\mathbf{x}\cdot\hat{b}_i\rangle_{\mathcal{P}})\end{pmatrix}, \tag{F4}$$

the recursion equations can be written in a compact form in terms of the vectors $\mathbf{V}$'s and the *transfer* matrix $\mathcal{T}_i$,

$$\mathbf{V}_i=\mathcal{T}_i\mathbf{V}_{i-1}, \tag{F5}$$

where the nonzero matrix elements of $\mathcal{T}_i$ are

$$t_{i1,1}=\langle\cos\theta\rangle_i, t_{i1,2}=-\left\langle\frac{1}{\sin\theta}\right\rangle_{i-1}\langle\sin\theta\cos\phi\rangle_i,$$

$$t_{i1,3}=\langle\sin\theta\cos\phi\rangle_i, t_{i1,4}=\langle\sin\theta\sin\phi\rangle_i,$$

$$t_{i2,1}=1,$$

$$t_{i3,1}=\langle\cot\theta\cos\theta\rangle_i, t_{i3,2}=-\left\langle\frac{1}{\sin\theta}\right\rangle_{i-1}\langle\cos\theta\cos\phi\rangle_i,$$

$$t_{i3,3}=\langle\cos\theta\cos\phi\rangle_i, t_{i3,4}=\langle\cos\theta\sin\phi\rangle_i,$$

$$t_{i4,2}=\left\langle\frac{1}{\sin\theta}\right\rangle_{i-1}\langle\sin\phi\rangle_i,$$

$$t_{i4,3}=-\langle\sin\phi\rangle_i, t_{i4,4}=\langle\cos\phi\rangle_i. \tag{F6}$$

Thus given the initial condition $\mathbf{V}_2$, which depends only on $\rho_2(\theta)$, all successive $\mathbf{V}$'s can be calculated recursively using Eq. (F5). Let us discuss the case of a uniform stretch where $\rho_i(\theta,\phi)$, and thence $\mathcal{T}_i$, does not depend on i (the subindices i will be omitted in this case). If the left and right eigenvectors of $\mathcal{T}$, $\mathbf{W}_\mu$, and $\mathbf{W}^\mu$, respectively, corresponding to the eigenvalue λ_μ, form a complete basis set the general solution of Eq. (F5) can be written as

$$\mathbf{V}_n=\sum_{\mu=1}^{4}\lambda_\mu^{n-2}\mathbf{W}_\mu\cdot\mathbf{V}_2\mathbf{W}^\mu. \tag{F7}$$

If all the eigenvalues are real and positive and if $\lambda\equiv\max_{\mu=1,\ldots,4}\{\lambda_\mu\}$ then for large n

$$\langle\hat{t}_2\cdot\hat{t}_{n+2}\rangle\sim\lambda^n, \tag{F8}$$

and likewise for $\langle\hat{b}_2\cdot\hat{b}_n\rangle$. On physical grounds, we expect that $\lambda<1$, so that the correlation functions decay exponentially with the distance measured along the chain. However, it is quite common that some eigenvalues are complex. Since the matrix $\mathcal{T}$ is real, complex eigenvalues occur in pairs of complex conjugate values. If the pair $\lambda_\pm=\exp(\pm i\chi-1/\xi)$ (χ and ξ are both real and positive) corresponds to the maximum modulus eigenvalue, then at large n we get, for example, for the tangent-tangent correlation,

$$\langle\hat{t}_2\cdot\hat{t}_{n+2}\rangle\sim\cos(\chi_0+n\chi)e^{-n/\xi}, \tag{F9}$$

where χ_0 depends on the initial conditions. Thus there is still an exponential decay with a correlation length ξ (in units of chain bond length), but there is also an oscillatory modulation with another length scale $1/\chi$, which corresponds to short range order along the chain (notice that, in one-dimensional systems such as our chain, long range order cannot occur if the interactions are short range as in the present case). This type of behavior, with $1/\chi\sim 3.6$, would be expected on a stretch of chain that adopts a helical conformation with 3.6 amino acids per turn.

We end this appendix with an example of such behavior which can be worked out in full detail. For the case $\rho_i(\theta,\phi)=\rho_i(\theta,-\phi)$ (which corresponds to invariance under chirality flipping), and ρ_i is independent of i for $i>2$, then $t_{k,4}=t_{4,k}=0$ with $k\neq 4$. This implies that the matrix $\mathcal{T}$ becomes block diagonal with an eigenvalue equal to $\langle\cos\phi\rangle$ and $\langle\mathbf{x}\cdot\hat{b}_n\rangle$ decays exponentially with a correlation length $-1/\ln\langle\cos\phi\rangle$. Furthermore, since $t_{1,2}t_{3,3}=t_{1,3}t_{3,2}$, one eigenvalue is zero and the remaining two are the solutions of the second order equation $\lambda^2+b\lambda+c=0$ with

$$b = -t_{1,1} - t_{3,3}, \quad c = t_{1,1}t_{3,3} - t_{1,3}t_{3,1} - t_{1,2}. \tag{F10}$$

Thus, if $b^2-4c>0$, the two solutions are real and the tangent-tangent correlation decays exponentially to zero. On the other hand, if $b^2-4c<0$, the two solutions are complex conjugate of each other, as described above, and in the particular case we are considering, i.e., $\rho_i(\theta,\phi)=\rho_i(\theta,-\phi)$, one finds for all n, that

$$\langle \hat{t}_1 \cdot \hat{t}_{n+2} \rangle = \frac{\cos(\chi_0 + n\chi)}{\cos \chi_0} e^{-n/\xi}, \tag{F11}$$

where $\xi=-2/\ln c$, $\chi=\arccos(-b/2\sqrt{c})$, and χ_0 depends on the initial conditions.

[1] B. Alberts *et al.*, *Molecular Biology of the Cell* (Garland, New York, 2002); D. Voet and J. G. Voet, *Biochemistry* (Wiley, New York, 2003).

[2] J. D. Watson and F. H. C. Crick, Nature (London) **171**, 737 (1953).

[3] M. Kimura, *Population Genetics, Molecular Evolution and the Neutral Theory* (University of Chicago Press, Chicago, 1994).

[4] A. Fersht, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (Freeman, New York, 1999).

[5] A. M. Lesk, *Introduction to Protein Architecture: The Structural Biology of Proteins* (Oxford University Press, Oxford, 2000).

[6] A. V. Finkelstein and O. Ptistyn, *Protein Physics: A Course of Lectures* (Academic Press, New York, 2002).

[7] J. R. Banavar and A. Maritan, Proteins **42**, 433 (2001).

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

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

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

[11] A. P. Lightman, *Great Ideas in Physics* (McGraw-Hill, New York, 2000).

[12] F. J. Dyson, *Origins of Life* (Cambridge University Press, Cambridge, U.K., 1999).

[13] T. E. Creighton, *Proteins: Structures and Molecular Properties* (Freeman, New York, 1993).

[14] J. W. Kelly, Curr. Opin. Struct. Biol. **8**, 101 (1998).

[15] C. M. Dobson, Nat. Rev. Drug Discovery **2**, 154 (2003).

[16] L. J. Henderson, *Fitness of the Environment: An Inquiry into the Biological Significance of the Properties of Matter* (Macmillan, Basingstoke, England, 1913).

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

[18] G. G. Szpiro, *Kepler's Conjecture* (Wiley, New York, 2003).

[19] J. R. Banavar, O. Gonzalez, J. H. Maddocks, and A. Maritan, J. Stat. Phys. **110**, 35 (2003).

[20] A theorem proved by O. Gonzalez and J. Maddocks [Proc. Natl. Acad. Sci. U.S.A. **96**, 4769 (1999)] in the context of knot topologies shows the remarkable connection between the three-point recipe and the tube thickness.

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

[22] D. Marenduzzo, A. Flammini, A. Trovato, J. R. Banavar, and A. Maritan, SISSA Report, 2002 (unpublished).

[23] P. G. de Gennes and J. Prost, *The Physics of Liquid Crystals* (Oxford University Press, Oxford, 1995).

[24] S. Chandrasekhar, *Liquid Crystals* (Cambridge University Press, Cambridge, U.K., 1977).

[25] D. C. Bassett, *Principles of Polymer Morphology* (Cambridge University Press, Cambridge, U.K., 1981); G. Strobel, *The Physics of Polymers* (Springer, New York, 1997).

[26] M. Tinkham, *Introduction to Superconductivity* (McGraw-Hill, New York, 1996).

[27] *Finite Size Scaling and Numerical Simulation of Statistical Systems*, edited by V. Privman (World Scientific, New York, 1990).

[28] H. Eyring and A. E. Stern, Chem. Rev. (Washington, D.C.) **24**, 253 (1939).

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

[30] D. J. Jacobs, A. J. Rader, L. A. Kuhn, and M. F. Thorpe, Proteins **44**, 150 (2001); A. J. Rader, B. M. Hespenheide, L. A. Kuhn, and M. F. Thorpe, Proc. Natl. Acad. Sci. U.S.A. **99**, 3540 (2002).

[31] C. Chothia and A. V. Finkelstein, Annu. Rev. Biochem. **59**, 1007 (1990).

[32] C. Chothia, Nature (London) **357**, 543 (1992).

[33] T. Przytycka, R. Srinivasan, and G. D. Rose, Protein Sci. **2**, 409 (2002).

[34] C. Chothia, J. Gough, C. Vogel, and S. A. Teichmann, Science **300**, 1701 (2003).

[35] M. Denton and C. Marshall, Nature (London) **410**, 417 (2001).

[36] E. V. Koonin, Y. I. Wolf, and G. P. Karev, Nature (London) **420**, 218 (2002).

[37] P. L. Privalov, Adv. Protein Chem. **35**, 1 (1982).

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

[39] R. V. Pappu and G. D. Rose, Protein Sci. **11**, 2437 (2002); Z. S. Shi, C. A. Olson, G. D. Rose, R. L. Baldwin, and N. R. Kollenbach, Proc. Natl. Acad. Sci. U.S.A. **99**, 9190 (2002).

[40] A. Ginsburg and W. R. Carroll, Biochemistry **4**, 2159 (1965).

[41] C. B. Anfinsen, Science **181**, 223 (1973).

[42] S. E. Jackson, Folding Des. **3**, R81 (1998).

[43] R. L. Baldwin and G. D. Rose, Trends Biochem. Sci. **24**, 26 (1999).

[44] D. Baker, Nature (London) **405**, 39 (2002).

[45] V. Villegas, J. C. Martinez, F. X. Aviles, and L. Serrano, J. Mol. Biol. **283**, 1027 (1998); F. Chiti *et al.*, Nat. Struct. Biol. **6**, 1005 (1999).

[46] J. C. Martinez and L. Serrano, Nat. Struct. Biol. **6**, 1010 (1999); D. S. Riddle *et al.*, *ibid.* **6**, 1016 (1999).

[47] See F. Ding, N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley,

and E. I. Shakhnovich, Biophys. J. **83**, 3525 (2002) for a direct molecular dynamics observation of the protein folding transition state ensemble.
[48] J. R. Banavar, A. Maritan, C. Micheletti, and A. Trovato, Proteins **47**, 315 (2002).
[49] G. N. Ramachandran and V. Sasisekharan, Adv. Protein Chem. **23**, 283 (1968).
[50] R. Srinivasan and G. D. Rose, Proteins **22**, 81 (1995).
[51] S. W. Englander, N. W. Downer, and H. Teitelbaum, Annu. Rev. Biochem. **41**, 903 (1972); C. K. Woodward and B. D. Hilton, Annu. Rev. Biophys. Bioeng. **8**, 99 (1979); J. B. Udgaonkar and R. L. Baldwin, Nature (London) **335**, 694 (1988); A. Matouschek, J. T. Kellis, Jr. L. Serrano, M. Bycroft, and A. R. Fersht, *ibid.* **346**, 440 (1990); S. E. Radford and C. M. Dobson, Philos. Trans. R. Soc. London, Ser. B **348**, 17 (1995); A. R. Fersht, Curr. Opin. Struct. Biol. **5**, 79 (1995).
[52] T. X. Hoang, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan, Proc. Natl. Acad. Sci. U.S.A. **101**, 7960 (2004).
[53] H. M. Berman *et al.*, Nucleic Acids Res. **28**, 235 (2000).
[54] Y. Zhou, C. K. Hall, and M. Karplus, Phys. Rev. Lett. **77**, 2822 (1996).
[55] K. D. Wilkinson and A. N. Meyer, Arch. Biochem. Biophys. **250**, 390 (1986); E. Dufour and T. Haertle, Protein Eng. **4**, 185 (1990); M. M. Harding, D. H. Williams, and D. N. Woolfson, Biochemistry **30**, 3120 (1991); P. Fan, C. Bracken, and J. Baum, *ibid.* **32**, 1573 (1993); K. Shiraki, K. Nishikawa, and Y. Goto, J. Mol. Biol. **245**, 180 (1995).
[56] V. S. Pande, A. Y. Grosberg, and T. Tanaka, Rev. Mod. Phys. **72**, 259 (2000).
[57] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524 (1987).
[58] 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).
[59] J. D. Bernal, Nature (London) **143**, 663 (1939).
[60] D. Perl *et al.*, Nat. Struct. Biol. **5**, 229 (1998).
[61] D. S. Riddle *et al.*, Nat. Struct. Biol. **4**, 805 (1997); D. E. Kim, H. Gu, and D. Baker, Proc. Natl. Acad. Sci. U.S.A. **95**, 4982 (1998).
[62] 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 *et al.*, *ibid.* **96**, 11 211 (1999); Y. Wei, S. Kim, D. Fela, J. Baum, and M. H. Hecht, *ibid.* **100**, 13 270 (2003).
[63] G. D. Rose, A. R. Geselowitz, G. J. Lesser, R. H. Lee, and M. H. Zehfus, Science **229**, 834 (1985); S. Miller, J. Janin, A. M. Lesk, and C. Chothia, J. Mol. Biol. **196**, 641 (1987); C. Lawrence, I. Auger, and C. Mannella, Proteins **2**, 153 (1987).
[64] 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).
[65] E. Shakhnovich, V. Abkevich, and O. Ptitsyn, Nature (London) **379**, 96 (1996); L. A. Mirny, V. I. Abkevich, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **95**, 4976 (1998); M. Vendruscolo, C. M. Dobson, E. Paci, and M. Karplus, Nature (London) **409**, 641 (2001).
[66] L. Holm and C. Sander, Proteins **28**, 72 (1997).
[67] D. T. Jones, W. R. Taylor, and J. M. Thornton, Nature (London) **358**, 86 (1992).
[68] S. E. Radford and C. M. Dobson, Cell **97**, 291 (1999); M. Bucciantini *et al.*, Nature (London) **416**, 507 (2002); M. Dumoulin *et al.*, *ibid.* **424**, 783 (2003); F. Chiti *et al.*, *ibid.* **424**, 805 (2003).
[69] S. B. Prusiner, Proc. Natl. Acad. Sci. U.S.A. **95**, 13 363 (1998).
[70] A. Aggeli *et al.*, Proc. Natl. Acad. Sci. U.S.A. **98**, 11 857 (2001).
[71] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **63**, 1195 (1989).
[72] W. Y. Yang, E. Larios, and M. Gruebele, J. Am. Chem. Soc. **125**, 16 220 (2003).
[73] A. L. Horwich, Proc. Natl. Acad. Sci. U.S.A. **96**, 11 033 (1999).
[74] M. Fandrich and C. M. Dobson, EMBO J. **21**, 5682 (2002).
[75] J. Maynard Smith, Nature (London) **225**, 563 (1970).
[76] M. Kimura, Nature (London) **217**, 624 (1968).
[77] J. L. King and T. H. Jukes, Science **164**, 788 (1969).
[78] M. Kimura, Ann. Genet. **19**, 153 (1976).
[79] G. Tiana, B. E. Shakhnovich, N. V. Dokholyan, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **101**, 2846 (2004).
[80] C. von Mering *et al.*, Nature (London) **417**, 399 (2002).
[81] J. M. Thornton, A. E. Todd, D. Milburn, N. Borkakoti, and C. A. Orengo, Nat. Struct. Biol. **7**, 991 Suppl. S (2000).
[82] C. A. Orengo, A. E. Todd, and J. M. Thornton, Curr. Opin. Struct. Biol. **9**, 374 (1999).
[83] M. Garcia-Viloca, J. Gao, M. Karplus, and D. G. Truhlar, Science **303**, 186 (2004).
[84] O. Lichtarge and M. E. Sowa, Curr. Opin. Struct. Biol. **12**, 21 (2002).
[85] C. S. Goh and F. E. Cohen, J. Mol. Biol. **324**, 177 (2002); P. J. Bickel, K. J. Kechris, P. C. Spector, G. J. Wedemayer, and A. N. Glazer, Proc. Natl. Acad. Sci. U.S.A. **99**, 14 764 (2002).
[86] S. J. Campbell, N. D. Gold, R. M. Jackson, and D. R. Westhead, Curr. Opin. Struct. Biol. **13**, 389 (2003).
[87] N. Nagano, C. A. Orengo, and J. M. Thornton, J. Mol. Biol. **321**, 741 (2002).
[88] I. M. A. Nooren and J. M. Thornton, EMBO J. **22**, 3486 (2003); see also L. L. Looger, M. A. Dwyer, J. J. Smith, and H. W. Hellinga, Nature (London) **423**, 185 (2003).
[89] A. C. Gavin *et al.*, Nature (London) **415**, 141 (2002); Y. Ho *et al.*, *ibid.* **415**, 180 (2002).
[90] K. A. Dill and D. Shortle, Annu. Rev. Biochem. **60**, 795 (1991).
[91] R. V. Pappu, R. Srinivasan, and G. D. Rose, Proc. Natl. Acad. Sci. U.S.A. **97**, 12 565 (2000).
[92] K. W. Plaxco and M. Gross, Nat. Struct. Biol. **8**, 659 (2001).
[93] G. D. Rose, Adv. Protein Chem. **62**, xv (2002); R. L. Baldwin, *ibid.* **62**, 361 (2002).
[94] D. Shortle, Adv. Protein Chem. **62**, 1 (2002).

[95] X. Duara, A. Glattli, P. Gee, C. Peter, and W. F. Van Gunsteren, Adv. Protein Chem. **62**, 341 (2002).
[96] B. Zagrovic, C. D. Snow, S. Khaliq, M. R. Shirts, and V. S. Pande, J. Mol. Biol. **323**, 153 (2002).
[97] The importance of local information is underscored by the fact that many schemes for secondary structure prediction do moderately well with just local sequence information. See also the discussion presented by Srinivasan and Rose [50] on the importance of local interactions in protein folding.
[98] K. W. Plaxco, K. T. Simons, and D. Baker, J. Mol. Biol. **277**, 985 (1998); H. S. Chan, Nature (London) **392**, 761 (1998).
[99] S. Rackovsky and H. A. Scheraga, Macromolecules **11**, 1168 (1978); **13**, 1440 (1980); **14**, 1259 (1981); **15**, 1340 (1982); Acc. Chem. Res. **17**, 209 (1984).
[100] W. G. Miller, D. A. Brant, and P. J. Flory, J. Mol. Biol. **23**, 57 (1967).
[101] F. Krieger, B. Fierz, O. Bieri, M. Drewello, and T. Kiefhaber, J. Mol. Biol. **332**, 265 (2003).
[102] D. Shortle and A. K. Meeker, Proteins **1**, 81 (1986); D. Shortle, W. E. Stites, and A. K. Meeker, Biochemistry **29**, 8033 (1990); J. M. Flanagan, M. Kataoka, T. Fujisawa, and D. M. Engleman, *ibid.* **32**, 10 359 (1993).
[103] J. J. Hopfield, Proc. Natl. Acad. Sci. U.S.A. **79**, 2554 (1982).
[104] P. W. Anderson, Proc. Natl. Acad. Sci. U.S.A. **80**, 3386 (1983).
[105] M. Eigen, P. Schuster, W. Gardiner, and R. Winkler-Oswatisch, Sci. Am. **244**, 78 (1981).
[106] K. Binder and A. P. Young, Rev. Mod. Phys. **58**, 801 (1986); M. Mezard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[107] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins **21**, 167 (1995); D. K. Klimov and D. Thirumalai, Phys. Rev. Lett. **76**, 4070 (1996); H. S. Chan and K. A. Dill, Proteins **30**, 2 (1998).
[108] A. K. Dunker, C. J. Brown, and Z. Obradovic, Adv. Protein Chem. **62**, 26 (2002); H. J. Dyson and P. E. Wright, Curr. Opin. Struct. Biol. **12**, 54 (2002).
[109] P. B. Yale, *Geometry and Symmetry* (Dover, Mineola, NY, 1988).
[110] G. Buck and J. Simon, in *Lectures at Knots*, edited by S. Suzuki (World Scientific, Singapore, 1997).
[111] R. A. Litherland, J. Simon, O. Durumeric, and E. Rawdon, Topol. Appl. **91**, 233 (1999).
[112] M. Doi and S. F. Edwards, *The Theory of Polymer Dynamics* (Clarendon, New York, 1993).
[113] An interesting problem for thick polymers in the continuum limit is the measure associated with the curve corresponding to the viable centerline of a thick tube. Indeed, for the Edwards model, the first term in the effective energy in Eq. (B1) is used to define a measure on the ensemble of all continuous curves embedded in three-dimensional space, the Wiener measure. It is well known that this gives a nonzero weight to all curves which are continuous but not differentiable (C^0 curves in the standard notation). In particular, the curves for which the tangent can be defined are a subset of zero measure in this ensemble. It can be proved rigorously (see, e.g., Ref. [114]) that if a curve is a viable centerline for a tube of nonzero thickness, then it must satisfy more stringent smoothness requirements, namely, it should be $C^{1,1}$, i.e., it is differentiable and its derivative is a Lipschitz function (see [114] and references therein). An interesting problem, which we do not consider here, is the definition of a suitable measure in the continuum limit for such thick polymer centerlines.
[114] O. Gonzalez, J. H. Maddocks, F. Schuricht, and H. von der Mosel, Calculus Var. Partial Differ. Equ. **14**, 29 (2002).
[115] H. Yamakawa, *Modern Theory of Polymer Solutions* (Harper and Row, New York, 1971); P. G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, NY, 1979); G. des Cloiseaux and J. F. Jannink, *Polymers in Solution: Their Modeling and Structure* (Clarendon Press, Oxford, 1990).
[116] D. Marenduzzo, C. Micheletti, H. Seyed-allaei, A. Trovato, and A. Maritan (unpublished).
[117] E. Pitard, T. Garel, and H. Orland, J. Phys. I **7**, 1201 (1997); S. Doniach, T. Garel, and H. Orland, J. Chem. Phys. **105**, 1601 (1996); S. Lise, A. Maritan, and A. Pelizzola, Phys. Rev. E **58**, R5241 (1998).
[118] Note that the featureless compact regime in Fig. 2 is *not* the same as the disordered globule regime which is found in the EM case (phase G in Fig. 13). While, in the former case, basically all parts of the chain manage to interact with all other parts, the same does not happen in the latter case. Instead, this compact phase is akin to the packed conformation of independent hard spheres, being dominated by an isotropic two-body repulsion, absent in the Hamiltonian (B2) for which Fig. 2 was derived. We find a similar phase in our Monte Carlo simulations in the low thickness, low temperature regime (see Fig. 14) [22] because, as explained above, such a repulsion is needed in the discrete case.
[119] I. Chang, M. Cieplak, R. I. Dima, A. Maritan, and J. R. Banavar, Proc. Natl. Acad. Sci. U.S.A. **98**, 14350 (2001).
[120] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, J. Mol. Biol. **247**, 536 (1995).
[121] R. T. Ingwall, H. A. Scheraga, N. Lotan, A. Berger, and E. Katchalski, Biopolymers **6**, 331 (1968); S. Marqusee, V. H. Robbins, and R. L. Baldwin, Proc. Natl. Acad. Sci. U.S.A. **86**, 5286 (1989); J. M. Scholtz, E. J. York, J. M. Stewart, and R. L. Baldwin, J. Am. Chem. Soc. **113**, 5102 (1991); E. J. Spek, C. A. Olson, Z. Shi, and N. R. Kallenbach, *ibid.* **121**, 5571 (1999); J. A. Vila, D. R. Ripoll, and H. A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **97**, 13 075 (2000); Biopolymers **58**, 235 (2001).
[122] T. P. Creamer and G. D. Rose, Proc. Natl. Acad. Sci. U.S.A. **89**, 5937 (1992).
[123] Z. Shi, B. A. Krantz, N. Kallenbach, and T. R. Sosnick, Biochemistry **41**, 2120 (2002).
[124] B. Fain and M. Levitt, Proc. Natl. Acad. Sci. U.S.A. **100**, 10 700 (2003).
[125] M. Cieplak and T. X. Hoang, Biophys. J. **84**, 475 (2003).
[126] A. D. Sokal, Nucl. Phys. B (Proc. Suppl.) **47**, 172 (1996).