Lattice Tube Model of Proteins

Jayanth R. Banavar,[1] Marek Cieplak,[2] and Amos Maritan[3]

[1]104 Davey Laboratory, Department of Physics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA
[2]Institute of Physics, Polish Academy of Sciences, Aleja Lotników 32/46, 02-668 Warsaw, Poland
[3]Dipartimento di Fisica and INFN, Universitá degli Studi di Padova, via Mazzolo 8, 35100, Padova, Italy

(Received 8 May 2004; published 29 November 2004)

We present a new lattice model for proteins that incorporates a tubelike anisotropy by introducing a preference for mutually parallel alignments in the conformations. The model is demonstrated to capture many aspects of real proteins.

DOI: 10.1103/PhysRevLett.93.238101 PACS numbers: 87.15.He, 87.15.Aa, 87.15.Cc

There have been several physics-based attempts to distill the essential features of the protein problem, and notable success in capturing many of the key ingredients has been achieved using lattice models [1]. Such coarse-grained descriptions allow a virtually exact analysis of many properties and provide a useful framework for understanding experimental results. Indeed, valuable progress has been made within the simplified description of a lattice model with just two types of amino acids denoted by H and P representing hydrophobic and polar behaviors. The principal theme of this Letter is to present a new lattice model of proteins, which takes into account a key attribute of chain molecules—the context of amino acids within a chain. We benchmark the behavior of this model with the well-studied HP lattice model and show that the new model captures several attributes of real proteins.

There are clear hints, manifested by the many common characteristics of proteins [2], that proteins may be simpler than one might expect. Protein structures are constructed in a modular manner from common building blocks—helices, hairpins, and sheets connected together by tight turns. Further, the total number of distinct protein folds seems to be of the order of just a few thousand [3].

The simplest model of an unconstrained object is a hard sphere. A collection of hard spheres exhibits both fluid and crystalline phases on changing the volume fraction. When objects are tethered together in the form of a chain, it is no longer appropriate to consider them as spheres. There is a special direction that one may associate with each object which is tangent to the chain and is defined by the adjoining particles along the chain. It is therefore more appropriate to model the objects making up a chain by means of discs or coins, for which the heads-to-tails direction is distinct from the two other directions. This picture of tethered coins leads to a tubelike description of a chain molecule [2]. Just as symmetry plays a key role in determining the nature of ordering of unconstrained particles (the phases associated with a collection of spheres are vastly simpler than the liquid crystal phases of anisotropic objects), the anisotropy inherent in a tube leads to new behavior. Recent work [2] has shown that the tube picture can be used to understand the conventional polymer phases and the novel phase of matter used by nature to house protein native state structures in a unified way and for the development of a framework for understanding the common character of proteins.

There are three key features of a tube description that one ought to incorporate in a lattice model: self-intersections of a tube are not allowed, the local radius of curvature of a tube can be no smaller than the tube radius, and, in a compact state, there is a tendency for nearby tube segments to be parallel (indeed, both helices and sheets have tube segments alongside and parallel to each other leading to a cooperative placement of hydrogen bonds). The first two features are built into a model of a self-avoiding chain on a lattice. Our focus here is on considering the effects of introducing the third.

In order to illustrate the key idea, we consider a 16 amino acid (aa) self-avoiding chain on a square lattice. In the standard HP model, one ascribes a favorable energy -1 for a HH contact (two H aa which are *not* next to each other in sequence but sit next to each other in the lattice) and zero energy otherwise. Here, in addition, we pay attention to the context that the contact occurs in. Figure 1 illustrates three distinct types of contacts (denoted by an index m) depending on the degree to which the segments

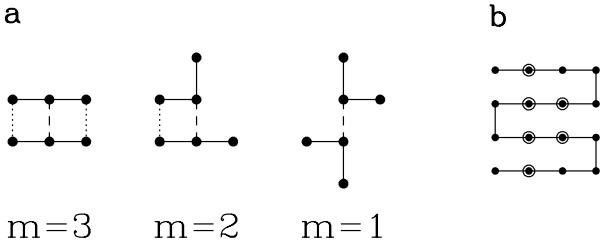

FIG. 1. (a) Sketch of three contact environments in the THP model. The dashed line denotes a contact. (b) The optimal structure for the THP model. The circles represent the hydrophobic core and have H aa in them more than 87% of the time for the sequences that fold into this conformation when $\epsilon = 1$.

containing the aa in contact are parallel to each other. The energy assigned to a HH contact of type m in the tube HP (THP) model is denoted by e_m. In what follows, let us choose e_m to be -3, -2, and -1 for $m = 3$, 2, and 1, respectively, thereby favoring the parallelism of nearby segments. In the standard HP model $e_m = -1$ independent of m.

In order to understand the role of sequence heterogeneity, we generalize the model so that the energies are given by

$$E = \sum_{i<j} e_m \Delta(i-j,m)[\delta_{i,\mathrm{H}}\delta_{j,\mathrm{H}} + (1-\epsilon)D_{ij}], \quad (1)$$

where

$$D_{ij} = (\delta_{i,\mathrm{H}}\delta_{j,\mathrm{P}} + \delta_{i,\mathrm{P}}\delta_{j,\mathrm{H}} + \delta_{i,\mathrm{P}}\delta_{j,\mathrm{P}}) = 1 - \delta_{i,\mathrm{H}}\delta_{j,\mathrm{H}}. \quad (2)$$

Here, $\Delta(i-j,m)$ is equal to 1 if the amino acids i and j form a contact and 0 otherwise. When such a contact exists, the energy of attraction associated with it depends on the index m. $\delta_{i,\mathrm{H}}$ is defined to be equal to 1 if amino acid i is hydrophobic and 0 if it is polar. Similarly, $\delta_{i,\mathrm{P}} = 1 - \delta_{i,\mathrm{H}}$ is equal to 1 if the amino acid i is polar. Depending on the choice of the e_m parameters, one obtains the HP or THP models. The limiting cases correspond to $\epsilon = 1$, i.e., the "standard" THP or HP models, and $\epsilon = 0$—the case of a homopolymer made of H amino acids.

For the 16-aa chain, all sequences and all possible conformations can be enumerated exactly. There are interesting differences in the energy landscape of the HP and the THP models. One may determine the sequences which have a unique ground state and the number of distinct designable conformations, which house these sequences, as a function of ϵ (see inset of Fig. 2). For a homopolymer ($\epsilon = 0$), the HP model has no designable structure—all compact conformations are degenerate and have the same energy. Thus in the absence of sequence specificity, there is no preselection of proteinlike structures among compact conformations. Thus in the absence of sequence specificity, there is no proteinlike behavior. When a weak heterogeneity is introduced by turning on a small ϵ, the HP energy landscape becomes rugged and each of the 69 maximally compact conformations becomes designable but with a weak thermodynamic stability. Thus the funnel-like energy landscape [4] arises only on turning on the full degree of heterogeneity.

This is in sharp contrast to the behavior of the THP model—here, even for a homopolymer, one obtains a unique ground state, akin to either a helix or a sheet in two dimensions (see Fig. 1), selected not by considerations of the chemistry of the aa but rather by the overarching principles of geometry and symmetry. Interestingly, in the limit of small ϵ, all $2^{16} = 65\,536$ sequences have a unique ground state in the THP model and none in the HP model. When $\epsilon = 1$, one obtains 10 579 and 1539 designable sequences in the THP and HP models, respectively (see Fig. 2), folding into 684 and 456 distinct folded structures. Furthermore, the number of sequences folding into the most designable structure are 637 and 26 for the two models.

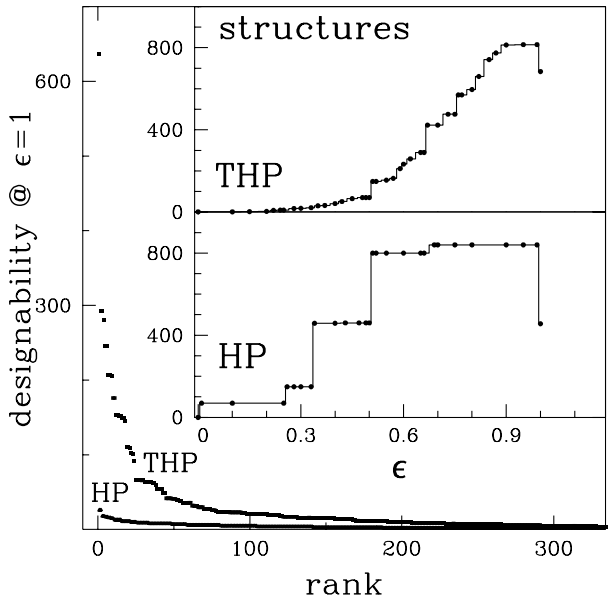

FIG. 2. Rank ordered values of the number of sequences that fold into the given structure for all of the designable structures at $\epsilon = 1$. The inset shows the number of designable structures as a function of ϵ (see text).

The thermodynamic stability of a sequence is characterized by the folding transition temperature, T_f, at which the equilibrium probability of being in the native state conformation is equal to $\frac{1}{2}$. The spread in the values of T_f is nearly 3 times larger in the THP model than in the HP case. The most stable THP sequence folds into the structure shown in Fig. 1(b), whereas the most stable HP sequence folds to a structure which is not maximally compact. In order to describe the folding kinetics, we take a sequence at a temperature equal to its T_f value and consider ten batches of 101 trajectories and determine the first passage time to the native state starting from an unfolded conformation. The time evolution [5] is a Monte Carlo process which satisfies detailed balance. The kinetic moves consist of single bead moves (the kink flips and rotations of the terminal segments) with probability 0.2 and of two bead "crankshaft" moves with probability 0.8. A median folding time is determined for each batch and averaged over all batches to yield a measure of the folding time, t_{fold}. Our calculations were carried out for 12 sequences in each model (the top ten sequences in T_f values and the sequences ranked 20 and 30). In all cases, the THP model exhibits more rapid two-state folding than the HP model with the ratio of the folding times for the 12 sequences ranging between 0.10 and 0.47.

The framework of evolution in life works through both the DNA molecule and the functionally useful protein molecule. Mutations of the DNA molecule lead to the possibility of new proteins, whose selection, in turn, leads to an enhancement of the number of such DNA molecules. As pointed out by Maynard Smith [6], 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.

We have investigated the topology of connections [7] between the designable structures resulting from point mutations in the sequence (the change of one aa from H to P or vice versa). Indeed, while one has a "random walk" in sequence space that forms a connected network, there is no similar continuous variation in structure space. When $\epsilon = 1$, 39.3% or 605 of the HP sequences do not belong to the connected network envisioned by Maynard Smith in contrast to the THP model for which only 13 of the 10 579 sequences, i.e., 0.12%, do not belong to the network. The THP model is vastly better connected than the HP model, as illustrated in Fig. 3. The former exhibits approximate scale-free behavior [8] while the latter is more akin to a random network with low mean coordination number (Fig. 4).

As an illustration, we conclude with an analysis of a few hundred proteins [9] to determine the propensity of amino acid pairs in contact [10] to be in specific environments characterized by the m index introduced above. Specifically, we look at the type of contact between aa k and aa l along the sequence and categorize it in the following manner: the specific aa pair involved in the contact, their sequence separation $s = | k - l |$ equal to 2, 3, 4, or greater than 4 and the number of contacts m between the two groups of aa $(k-1, k, k+1)$ and $(l-1, l, l+1)$ which can range between 1 and 9. (The geometry of the lattice model in two dimensions allows for only three values, 1, 2, or 3, of the contact environment index m.) We have determined

$$\chi_2(k, l, s, m) = \frac{[n(k, l, s, m) - p(k, l, s, m)]^2}{p(k, l, s, m)}. \quad (3)$$

Here n is the number of contacts and p the expected number of contacts based on chance: $p(k, l, s, m) = aq(k, l, s)$, where q is the number of the specific aa pairs at distance s and $a = \sum_{kl} n(k, l, s, m) / \sum_{kl} q(k, l, s)$. A large value of χ_2 corresponds to a strong signal that aa k and aa l prefer to make or avoid a contact in the environment defined by the s and m indices (Table I) and would be useful in the development of scoring functions for protein structure recognition [9].

The tube idea reveals a deep underlying simplicity in the protein problem. In standard approaches, the sequence of amino acids is believed to play a key role in sculpting the free energy landscape and determining its native state structure. Here, instead, the free energy landscape is sculpted predominantly by symmetry and geometry and the sequence plays a vital role in the *selection* of the

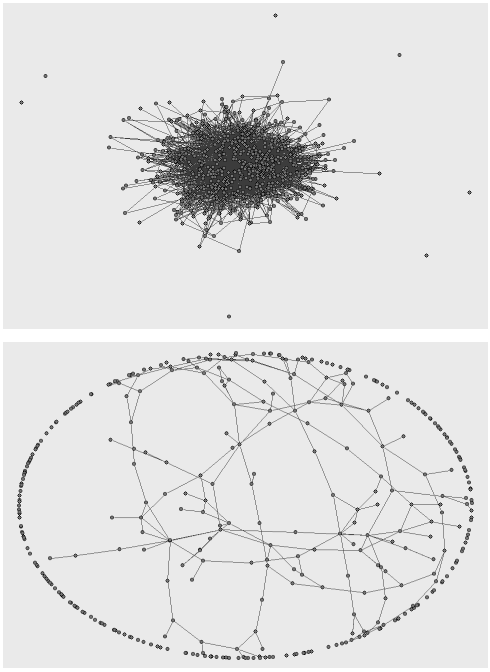

FIG. 3. Network topologies (using Pajek) of designable structures resulting from point mutations in the sequence. The top and bottom panels are for the THP and HP models, respectively.

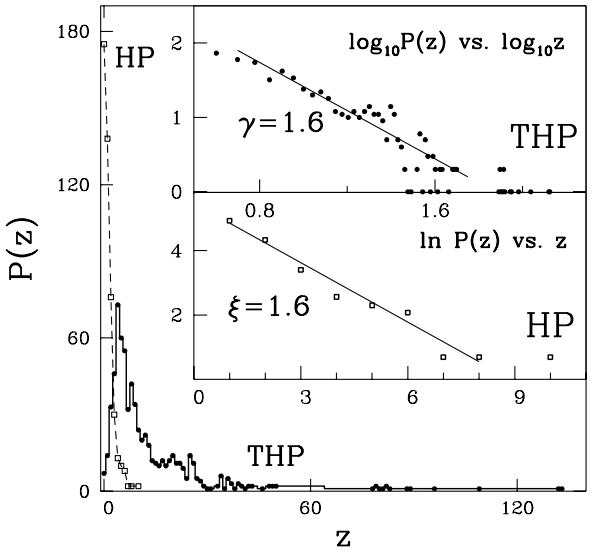

FIG. 4. Probability distribution, $P(z)$, of the effective coordination number for the network of designable structures shown in Fig. 3. The inset is a plot of the same data in a log-log scale (the top panel) for the THP model and in a log-linear scale for the HP model. The results illustrate the approximate validity of $P(z) \sim z^{-\gamma}$ and $P(z) \sim \exp -z/\xi$ for the THP and HP models, respectively.

TABLE I. The list of aa pairs with $\chi_2(k, l, s, m)$ larger than 65. Amino acids with long and/or forked side groups (L, K, Q, R, E) are more likely to form local contacts with a large m. On the other hand, the smallest amino acid, G, is much less likely to form such contacts, as evidenced by the aversion in the pairs G-G, G-P, and G-S for $s = 2$ and $m = 8$. The propensity of aa A to participate in short range contacts with a large m (also for $s = 4$) is also due to its size: A is small enough to allow for participation in conformational twists, but it is sufficiently big to facilitate formation of many contacts. The C-C covalent attraction results in nonlocal contacts over a range of m values.

aa Pairs	Attraction/aversion	s	m
V-I	Attraction	2	4
AL-AEQKR	Attraction	2	8
G-PSG	Aversion	2	8
A-AQIR L-ALQ	Attraction	3	8
A-A L-LA E-R	Attraction	4	6
G-V	Aversion	4	6
L-IFVLMWY	Attraction	>4	1
V-IFVMW F-FWY	Attraction	>4	1
I-FIWM C-C M-FY	Attraction	>4	1
A-G G-DST	Aversion	>4	1
L-IFVLMWY W-Y	Attraction	>4	2
V-I-IF F-FW C-C	Attraction	>4	2
L-LF I-V C-C	Attraction	>4	3
C-C	Attraction	>4	4
V-V-I I-I	Attraction	>4	5
V-LVIFT I-I	Attraction	>4	6

native state from a predetermined menu. Unlike sequences and functionalities, which are shaped by the powerful forces of evolution, the menu of putative native state structures is immutable and is determined by physical law. Indeed, this fixed backdrop provides the initial basis for selection in molecular evolution. DNA which make proteins that are able to fold readily into one of the predetermined folds pass the initial screening. An additional level of filtering completes the selection process of proteins that are not only good folders but are also able to interact well with ligands and other proteins and play a useful functional role.

We are grateful to Istvan Albert for helpful advice. This work was supported by KBN (Grant No. 2 P03B 032 25), COFIN MURST 2003, INFM, NASA, NSF IGERT Grant No. DGE-9987589, and the NSF MRSEC at Penn State.

[1] K. A. Dill, S. Bromberg, K. Z. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan, Protein Science **4**, 561 (1995).

[2] J. R. Banavar and A. Maritan, Rev. Mod. Phys. **75**, 23 (2003); T. X. Hoang, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan, Proc. Natl. Acad. Sci. U.S.A. **101**, 7960 (2004); J. R. Banavar, T. X. Hoang, A. Maritan, F. Seno, and A. Trovato, Phys. Rev. E **70**, 041905 (2004).

[3] C. Chothia and A.V. Finkelstein, Annu. Rev. Biochem. **59**, 1007 (1990); C. Chothia, Nature (London) **357**, 543 (1992).

[4] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, Science **267**, 1619 (1995); Z. Luthey-Schulten, B. E. Ramirez, and P. G. Wolynes, J. Phys. Chem. **99**, 2177 (1995); K. A. Dill and H. S. Chan, Nat. Struct. Biol. **4**, 10 (1997).

[5] M. Cieplak, M. Henkel, J. Karbowski, and J. R. Banavar, Phys. Rev. Lett. **80**, 3654 (1998).

[6] J. Maynard Smith, *Evolutionary Genetics* (Oxford University Press, New York, 1998), 2nd ed.

[7] E. Bornberg-Bauer and H. S. Chan, Proc. Natl. Acad. Sci. U.S.A. **96**, 10 689 (1999); Y. Cui, W. H. Wong, E. Bornberg-Bauer, and H. S. Chan, Proc. Natl. Acad. Sci. U.S.A. **99**, 809 (2002); H. S. Chan and E. Bornberg-Bauer, Appl. Bioinformatics **1**, 121 (2002).

[8] R. Albert and A.-L. Barabasi, Rev. Mod. Phys. **74**, 47 (2002).

[9] I. Chang, M. Cieplak, R. I. Dima, A. Maritan, and J. R. Banavar, Proc. Natl. Acad. Sci. U.S.A. **98**, 14351 (2001).

[10] J. Tsai, R. Taylor, C. Chothia, and M. Gerstein, J. Mol. Biol. **290**, 253 (1999); M. Cieplak and T. X. Hoang, Biophys. J. **84**, 475 (2003).