Random Walks in the Space of Conformations of Toy Proteins

Rose Du,[1] Alexander Yu. Grosberg,[2] and Toyoichi Tanaka[1]

[1]*Department of Physics and Center for Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139*

[2]*Department of Physics, University of Minnesota, 116 Church Street SE, Minneapolis, Minnesota 55455*

(Received 22 April 1999)

Monte Carlo dynamics of the lattice toy protein of 48 monomers is interpreted as a random walk in an abstract (discrete) space of conformations. To test the geometry of this space, we examine the return probability $P(T)$, which is the probability to find the polymer in the native state after T Monte Carlo steps, provided that it starts from the native state at the initial moment. Comparing computational data with the theoretical expressions for $P(T)$ for random walks in a variety of different spaces, we show that conformation spaces of polymer loops may have nontrivial dimensions and exhibit negative curvature characteristics of Lobachevskii (hyperbolic) geometry.

PACS numbers: 87.14.Ee, 61.25.Hq, 87.15.Aa

Levinthal's paradox [1] is universally considered the essence of the protein folding problem. In its most direct form, the paradox revolves around the exponentially large number of possible conformations being immeasurably larger than what a protein can conceivably test within the observable time scale. Levinthal's paradox arises from thinking of protein folding as a search in a conformation space resembling a golf course with just one hole representing the native state. To resolve the problem, it has been conjectured in the literature [2–6] that volume interactions between monomers should provide an energetic bias towards the native state. However undoubtedly correct, this should not overshadow the necessity to understand the search in conformation space. Indeed, if we start from an open coil conformation, then volume interactions cannot provide any significant bias for a while, at least until after some minimal number of contacts has been formed. In macroscopic terms, this initial stage is an uphill climb over an entropic barrier. In microscopic terms, it is a random walk in conformation space. In order to initiate the energy-driven downhill slide towards the native state, or to enter the funnel-shaped [3,4,6] area of the free energy landscape [7], a fluctuation has to provide a sufficient decrease in entropy. In other words, a random walk has to bring the system into the specific region in conformation space, which corresponds to the transition (macro)state, most likely to a critical nucleus of some kind [2,8–10]. The system searches for this critical (macro)state through a random walk in conformation space, largely unaffected by heteropolymeric interaction energies, such that in some cases it may be similar to the kinetics of a homopolymer collapse [11]. Unfortunately, this problem remains out of reach of current simulation techniques.

In order to pave the way to it, we will consider in this work another related problem of random walks in conformation space, namely, that of fluctuations around the native state. This is itself a pressing issue in protein folding theory. Indeed, understanding these fluctuations is necessary in order to address the corrections to mean field theory which is formulated in terms of the random energy model (see review [12] and references therein).

We will restrict ourselves with the standard lattice protein model of 48 monomers. We choose to work with the particular native state conformation addressed in [13]. In order to remain in the vicinity of the native state, we can permanently fix the contacts which form the nucleus. The nucleus conformation, as conjectured in [13], is shown in Fig. 1. As a matter of fact, determination of the nucleus remains a subject of scrutiny and heated debate [14–18]. The various models of a single nucleus [8], of multiple nuclei [10], and of nucleation classes [18] are being debated but the choice among the different models is not the subject of this work. The nucleus from the work [13] which is used here has been chosen arbitrarily among a large number of possible conformations with nuclei surrounded by loops.

In order to address purely entropic factors, we examine polymer with no interactions—except for the constraints of polymer connectivity, excluded volume, and fixed contacts. The resulting structure is essentially that of many loops with fixed ends in the nucleus. For the discrete lattice model, we consider the conformation space as a graph in which the conformations are represented by nodes on the graph. If two conformations can be interconverted via a single Monte Carlo move (end flip, corner flip, or crankshaft), their corresponding nodes are connected by edges on the graph [3,4,19]. A Monte Carlo run is thus equivalent to a random walk on that graph [20].

As stated above, we consider here a problem of fluctuations around the native state; in this case, random walk starts in the origin, which is the native state, and we ask how frequently it returns back to the origin. Thus, our plan is as follows: we will perform a long Monte Carlo run of the above described loop model, starting from the folded native state, and we will record all the time moments (or Monte Carlo steps) of spontaneous folding, or random arrival to the origin, which is the native state. This will give us the return probability, $P(T)$ [21], as a function of Monte

 0031-9007/00/84(8)/1828(4)$15.00

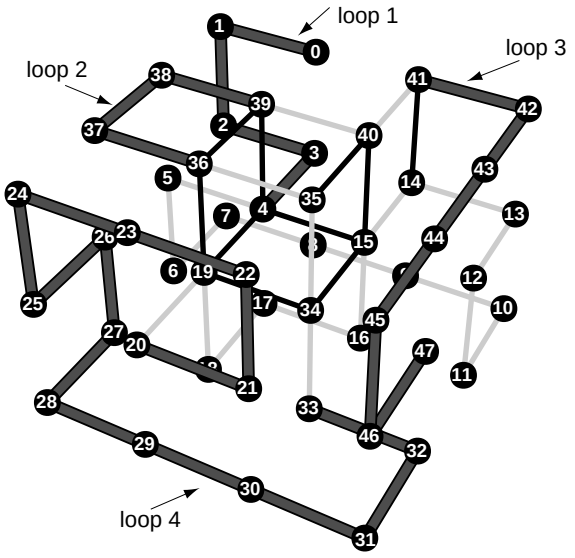

FIG. 1. The 48-mer conformation. The dark contacts indicate critical folding nucleus for this particular conformation, according to the data of the work by Mirny [13]. The loops considered in the present work are shown as thick lines.

Carlo time, T. In order to interpret these data, we will compare them with the following summary of the known results for $P(T)$ in a variety of different (discrete) spaces:

(i) For a random walk of T steps in an unbounded Euclidean space (cubic lattice) of dimension d, the probability of return is

$$P(T) = (2\pi T/d)^{-d/2} \sim T^{-d/2}. \tag{1}$$

(ii) Equation (1) holds for a fractal space with noninteger d; in this case d is the spectral dimension [22].

(iii) It is important for us to consider a random walk in a bounded region, because the set of conformations is always finite for lattice models, and thus, the region of interest in the conformation space is also finite. For a random walk in a bounded "cavity" in Euclidean space, or on a bounded fractal,

$$P(T) \simeq \begin{cases} (\frac{2\pi}{d}T)^{-d/2}, & \text{when } T < T^* \sim R^2, \\ \frac{1}{\mathcal{M}}\frac{z_0}{\bar{z}}, & \text{when } T > T^* \sim R^2. \end{cases} \tag{2}$$

Here R is the "size" of the allowed conformation space (graph), $\mathcal{M}$ is the total number of allowed conformations (or graph nodes), z_0 and $\bar{z}$ are the numbers of possible Monte Carlo moves (or incident graph edges), respectively, for the native state and averaged over all states. Equation (2) means that a random walk does not feel the bounds at "small" times until it arrives at the boundary. At later times, it covers all of the available region in a uniform manner, then the probability for visiting each point (conformation), α, is simply proportional to the number of ways, z_α, incident to that point. To better understand the meaning of R, it is useful to define the "distance" $D_{\alpha\beta}$ between two conformations, α and β, as the minimal number of elementary moves necessary to convert α into β. Then, according to graph theory, the diameter of the graph representing our conformation space should be defined as the maximum of $D_{\alpha\beta}$ over all pairs of conformations. Our R is then typically on the order of one-half of this diameter. Note that in the limit of very long loops, $N \gg 1$, we expect the space diameter and R to scale as $R \sim N$.

(iv) For a random walk in a d-dimensional Lobachevskii space [23,24]

$$P(T) \sim \begin{cases} T^{-d/2}\exp(-T\lambda/2d), & \text{when } T < T^*, \\ \frac{1}{\mathcal{M}}\frac{z_0}{\bar{z}}, & \text{when } T > T^*, \end{cases} \tag{3}$$

where λ is the Gaussian curvature (inverse squared curvature radius) of the space. This formula can be explained in the following simple way. At the scale $T \ll 1/\lambda$, when typical distance from the origin remains smaller than the curvature radius, the space appears effectively flat and Eq. (3) reduces to (1). On the other hand, for the intermediate T, between $1/\lambda$ and R^2, $P(T)$ is dominated by the exponential term, which can be understood if one remembers that a Cayley tree graph is the discrete counterpart of Lobachevskii space.

It is important to consider Lobachevskii geometry because, as we mentioned, the conformation space diameter scales as N for very long loops in the $N \gg 1$ limit, while the number of conformations scales exponentially with N. This means that there is an exponential growth of the number of conformations as a function of the distance from any given conformation (e.g., native). Such an exponential growth is the signature of Cayley tree or Lobachevskii geometry [25]. Thus, we expect the window of exponential behavior at large N or for longer loops.

(v) The analysis of loop conformations which arise from a fixed nucleus of contacts becomes more complicated if we have multiple loops. Still, if loops are independent of one another, then a simple estimate for the conformation space of all the loops combined can be obtained. Consider k loops each having a fraction f_i of the total number of movable monomers. When a monomer is chosen for a Monte Carlo move, the probability that it will be from loop i is f_i. Assume that each loop lives in an unbounded Euclidean space so that the probability for loop i to fold after t_i Monte Carlo steps is $P(t_i) = t_i^{-d_i/2}$, neglecting constant factors. The probability for all loops to return after time T, $P(T)$, is thus

$$P(T) = \sum_{t_1,\ldots,t_k=0}^{T} \delta\left(T - \sum_i t_i\right) \prod_{i=1}^{k} f_i^{t_i} P_i(t_i)\, \frac{T!}{\prod t_i!}. \tag{4}$$

Using Stirling's approximation, and noticing that due to the combinatorial factor, the sum is dominated at large T by the term in which $t_i = f_i T$, we obtain for independent loops

$$P(T) \sim T^{-d_{\text{eff}}/2}, \quad \text{where } d_{\text{eff}} = \sum_{i=1}^{k} d_i\,. \tag{5}$$

(vi) In reality, different loops are not independent. To some extent they obstruct each other's folding. In general, for obstructing loops, we expect

$$d_{\text{eff}} \leq \sum_{i=1}^{k} d_i \,. \tag{6}$$

With the summary of mathematical results for the return probability in different geometries, we can now proceed to the computer experiments on the loops shown in Fig. 1.

The return probability for loops 1, 2, and 3 is shown in Fig. 2 as a function of return time. The log-log graphs indicate clearly the power law dependence characteristic of Euclidean geometry (1). The dimensions in loop space for loops 1, 2, and 3, according to Fig. 2, are $d_1 = 1.74$, $d_2 = 0$, and $d_3 = 2.64$, respectively.

To see why the dimension for loop 2 is zero, note that there are only two possible positions for loop 2. At any given time the probability for loop 2 to be in one position or the other is $1/2$, which means that the probability of return must be $P(T) = 1/2$ or $\ln P(T) = -0.69$, consistent with the result shown in Fig. 2. The leveling off expected according to Eq. (2) in a bounded space is also seen for loop 1. The saturation level [which corresponds to $1/P \approx \exp(3.3) \approx 27$] and saturation time ($T^* \approx 67$) are roughly consistent with both the first and the second line of Eq. (2) given that the number of conformations for loop 1 is $\mathcal{M} = 51$, and the number of allowed moves for the native state is $z_0 = 4$. Loop 3 will undoubtedly exhibit leveling off as well, but at longer times, since loop 3 is much longer. We did not reach this in our Monte Carlo experiment.

To see the effect of having multiple loops on the loop space dimension, we examine conformations in which both loops 1 and 3 are allowed to move and those in which loops 2 and 3 are allowed to move. The reason for this choice of loops is that loops 1 and 3 (and loops 2 and 3) are sufficiently far apart on the conformation that their interactions are negligible, consistent with our simple analytical estimate. The effective combined dimension for loops 1 and 3 is $d_{13} = 4.40 \approx d_1 + d_3$ and for loops 2 and 3 is $d_{23} = 2.52 \approx d_2 + d_3$ (see Fig. 2) in agreement with the approximations given in Eq. (5).

The conformation space becomes even more interesting with longer loops such as loop 4 (see Fig. 1), which goes from monomer 20 to 33. As Fig. 3 indicates, the behavior of $P(T)$ for this loop is consistent with Lobachevskii geometry [first line of Eq. (3)]. A least-squares fit of $P(T)$ gives $d_4 = 1.5$, and $\lambda = 4.9 \times 10^{-8}$. This value of $\lambda = 1/4\pi r^2$ corresponds to the curvature radius about $r \sim 1000$. Thus, at rather small scales below r the conformation space of the loop 4 is a usual fractal graph, while at larger scales it branches exponentially like a Cayley tree. Physical nature of branching in the conformation space is very simple: when two different pieces of polymer are close together, each piece can move either on one or on the other side of the second piece, and to switch from one side to the other it has to go back, which is precisely the description of the bifurcation point on the Cayley tree. The rather large numerical value of r can be speculated to be of the same origin as that of the large values of the entanglement length N_e known in polymer dynamics [26]. It can be also compared to the characteristic length associated with formation of nontrivial knots upon the loop closure [27]. Undoubtedly, $P(T)$ for loop 4 will reach a plateau [second line of Eq. (3)] at the level about 10^{-8}; we did not pull our simulation quite that far, since we do not expect to see anything interesting there.

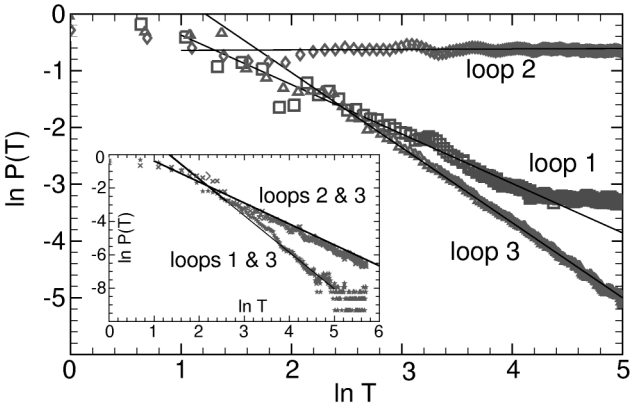

FIG. 2. Log-log plot of the return probability as a function of return time for loop 1, loop 2, and loop 3. The dimensions in conformation space for loops 1, 2, and 3 are $d_1 = 1.74$, $d_2 = 0$, and $d_3 = 2.64$. The lines shown have slopes of $-d_i/2$. Inset: Probability of return for loops 1 and 3, and for loops 2 and 3 combined. The corresponding dimensions are $d_{13} = 4.40 \approx d_1 + d_3$ and $d_{23} = 2.52 \approx d_2 + d_3$.

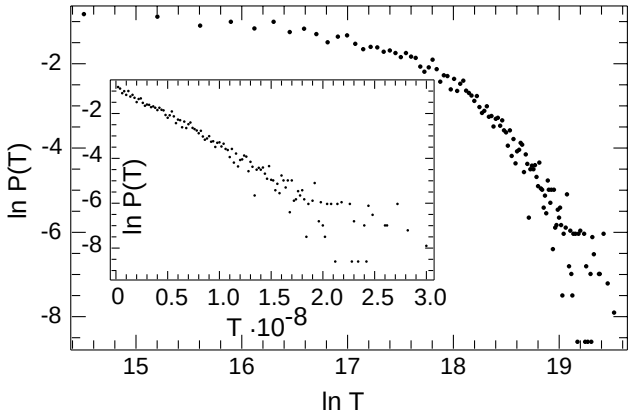

FIG. 3. Return probability as a function of return time for loop 4. Since the return time is typically very long, it is difficult to gather statistics. To circumvent this problem, the data were binned over the time intervals of 2×10^6. Thus, the apparent crossover of the log-log plot at small times reflects this binning rather than any physical peculiarity of the loop 4 dynamics. The inset shows the same data in semilog scale and indicates exponential behavior at long times, which is consistent with Lobachevskii geometry. The least-square fit yields $d = 1.5$ and $\lambda = 4.9 \times 10^{-8}$.

To conclude, we have shown that there exists a nontrivial geometry of the conformation space, with noninteger dimensions and negative (Lobachevskii) curvature. While the very fact of negative curvature, or exponential branching, of the phase space is due to the general property that volume is exponential in radius in this space, the actual curvature may be different for loops including chain ends, such as loops 1 and 3. Returning to the introduction, we repeat that the numerical part of the present work tested mainly short excursions, or local geometry of the conformation space in the vicinity of native state, which is directly relevant for the problem of fluctuations in the folded phase. In regards to another more frequently discussed problem of folding time, which is defined as mean first passage time starting from an arbitrary open conformation, the question is how long does it take for a random walk to bring the system into the critical region ω, where energetic bias takes over? To answer this, in addition to the overall conformation space geometry, one has to know more about ω: what is the shape and fractal dimension of ω, what is its boundary, etc. Although we do not know answers to these pressing questions, we do know that the Levinthal-style estimate of folding time as $t \sim |\Omega|/|\omega|$, where Ω is the entire conformation space, and $|\cdots|$ means the number of conformations in the domain "$\cdots$," is valid only for very special ways of embedding ω inside Ω [lower line in (3)]. Thus, in addition to the question of energetic bias towards the native state, an understanding of random walks in conformation space is crucial to the understanding of protein folding.

The work was supported by the NSF Grant No. DMR-9616791. A. G. acknowledges useful discussion with S. Nechaev.

[1] C. Levinthal, in *Mossbauer Spectroscopy in Biological Systems,* edited by P. Debrunner, J. C. M. Tsibris, and E. Munck (University of Illinois Press, Urbana, 1968).
[2] A. R. Fersht, Curr. Opin. Struct. Biol. **5**, 79–84 (1995).
[3] M. R. Betancourt and J. N. Onuchic, J. Chem. Phys. **103**, 773–787 (1995).
[4] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins **21**, 167–195 (1995).
[5] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, J. Mol. Biol. **252**, 460–471 (1995).
[6] K. A. Dill and H.-S. Chan, Nat. Struct. Biol. **4**, 10–19 (1997).
[7] It is worth stressing the difference between energy landscape and free energy landscape. The former is a function of the huge number of microscopic coordinates which define a conformation with all the microscopic details. The latter, on the other hand, involves a certain preaveraging, or coarse-graining, and is a function of relatively few macrovariables. Unfortunately, these two concepts are frequently confused in the literature.
[8] V. Abkevich, A. Gutin, and E. Shakhnovich, Biochemistry **33**, 10 026–10 036 (1994).
[9] A. Finkelstein and A. Badretdinov, Folding & Design **2**, 115 (1997); **3**, 67 (1998).
[10] D. K. Klimov and D. Thirumalai, J. Mol. Biol. **282**, 471–492 (1998).
[11] A. Halperin and P. M. Goldbart, Phys. Rev. E **61**, 565–573 (2000).
[12] V. S. Pande, A. Yu. Grosberg, and T. Tanaka, Rev. Mod. Phys. **72**, 259–314 (2000).
[13] L. A. Mirny, Ph.D. thesis, Harvard University, 1998.
[14] P. G. Wolynes, Folding & Design **3**, R107 (1998).
[15] E. I. Shakhnovich, Folding & Design **3**, R108–R111 (1998).
[16] D. Thirumalai and D. K. Klimov, Folding & Design **3**, R112–R118 (1998).
[17] V. S. Pande, A. Yu. Grosberg, T. Tanaka, and D. S. Rokhsar, Curr. Opin. Struct. Biol. **8**, 68–79 (1998).
[18] V. S. Pande and D. S. Rokhsar, Biochemistry **95**, 1490–1494 (1998).
[19] R. Du, V. S. Pande, A. Yu. Grosberg, T. Tanaka, and E. I. Shakhnovich, J. Chem. Phys. **111**, 10 375–10 380 (1999).
[20] There is a potential source of terminological confusion, since polymer conformations are often described in terms of some walker in 3D space. We stress that we are speaking here about a completely different random walk. In our case, the walker is the entire protein chain, and the walk is being performed in the abstract space of conformations.
[21] $P(T)$ is the return probability $P(T)$, not the first return probability; it is the probability that the system is in the origin at T, no matter how many times it may have visited origin before T.
[22] D. Stauffer and A. Aharony, *Introduction to Percolation Theory* (Taylor & Francis, London, 1992).
[23] S. P. Novikov and A. T. Fomenko, *Basic Elements of Differential Geometry and Topology* (Kluwer Academic, Dordrecht, 1990).
[24] J. C. Gruet, Stoch. Stat. Rep. **56**, 53–61 (1996) (in French).
[25] S. Nechaev, *Statistics of Knots and Entangled Random Walks* (World Scientific, Singapore, 1996).
[26] M. Doi and S. F. Edwards, *The Theory of Polymer Dynamics* (Oxford, Oxford, 1986).
[27] T. Deguchi and K. Tsurusaki Phys. Rev. E **55**, 6245–6248 (1997).