Minimum Energy Compact Structures of Random Sequences of Heteropolymers

Carlos J. Camacho and D. Thirumalai

Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742

(Received 29 March 1993)

We enumerate the number of minimum energy compact structures (MECS) for a two dimensional heteropolymer model consisting of hydrophobic (H) and hydrophilic beads averaged over all possible sequences as a function of the total number of beads N and the number of H beads N_H. The analysis of the average number of compact structures, which grows exponentially with N, gives no indication of logarithmic corrections to the free energy. For intermediate values of the ratio N_H/N, we estimate that the average number of MECS—which has a minimum at $N_H^{\mathrm{min}}/N = 0.6 \pm 0.1$—does not grow with N. Our results suggest that the late stages of protein folding should be restricted to the sampling of MECS only.

PACS numbers: 87.10.+e, 61.41.+e, 64.60.Cn

The argument that the random search of all available conformations of even a moderately sized protein takes too much time has been used to suggest that kinetic processes must dictate the folding of globular proteins. The estimate of folding times using this argument is incompatible with biological folding time scales (approximately seconds). The apparent conflict between biological folding times and the times estimated by random sampling of conformation space is known as the Levinthal paradox [1]. Levinthal's arguments ignore correlations between the various residues and further assume that all conformations have equal weight. Thus the number of allowed conformations of a protein with N residues is estimated as follows:

$$C_N(\mathrm{ideal}) = z^N , \tag{1}$$

where each residue can be found in any of the z states. If the polypeptide chain is treated as a walk embedded in a regular space lattice, z would be the lattice coordination number.

If one accounts for the simplest of these correlations, namely, excluded volume interactions between residues, the number of conformations is drastically reduced. The importance of self-avoidance in the context of protein folding seems to have been first emphasized by Dill [2]. However, it is easy to ascertain that taking this simple correlation alone does not rationalize the biological folding time scale for proteins. One can further assume that proteins are tightly packed structures and that the initial events in the folding process lead to the collapse of the random coil into a set of compact structures. These considerations lead to a further decrease in the number of conformations [2], and therefore result in a much smaller estimate of the folding time. Nevertheless, it still remains astronomically large. To illustrate the above arguments let us consider the folding of a protein of N residues. For our purposes we will ignore side chains and view the protein simply as a polymer of N sites on a regular lattice. The number of self-avoiding walks (SAWs) (mostly unfolded structures) scales with N as [3]

$$C_N(\mathrm{unfolded}) \simeq a N^{\gamma - 1} z_{\mathrm{eff}}^N , \tag{2}$$

where γ is a universal exponent, z_{eff} is an effective nonuniversal coordination number, and $a \sim O(1)$ is a nonuniversal amplitude. In $d=3$ and $d=2$, $\gamma \simeq 1.16$ [4] and exactly 55/32 [5], respectively. If the rate of conformational sampling is assumed to be 10^{14} sec^{-1} then a random search of all SAWs in a cubic lattice, where $z_{\mathrm{eff}} \simeq 4.684$ [4], would take approximately 10^{53} sec for an $N=100$ chain. The number of compact structures, on the other hand, can be written in its most general form as [6]

$$C_N(\mathrm{compact}) \approx b \bar{z}^N z_1^{N^{(d-1)/d}} N^{\gamma_c - 1} , \tag{3}$$

where $\ln \bar{z}$ is proportional to the free energy that depends on the temperature and the lattice, z_1 is a surface fugacity, d is the space dimension, γ_c measures possible logarithmic corrections to the free energy, and $b \sim O(1)$ is a nonuniversal amplitude. For the purpose of estimation, we will use the mean field expression $\bar{z} = z/e$ [7], $z_1 = 1$ and $\gamma_c = 1$. Hence, if one restricts the random search to the set of compact structures only, then the time needed to find the native state becomes about 10^{20} sec if the protein is again confined to a cubic lattice (i.e., $z=6$).

The N dependence of C_N given by Eqs. (2) and (3) results in estimates for folding times that are incompatible with biological time scales. Nevertheless, the arguments set forth above suggest that under folding conditions interactions in proteins must rapidly lead to a state of higher compactness, and in all likelihood the process of folding involves only a search among a subclass of compact structures. Good candidates for this subclass of structures are *minimum energy compact structures* (MECS). It is well known that the attractive interactions between certain residues in proteins result in well defined folded structures. In this paper, we have estimated the effect of the attractive energy between hydrophobic residues on C_N by undertaking a primarily numerical study of the MECS in a two dimensional heteropolymer model introduced by Lau and Dill [8]. It is shown that the number of MECS is considerably less than the number of dense compact structures. The relatively small number of MECS indicates that the effective attractive interaction between certain (hydrophobic) residues may direct the folding into a relatively small set of minimum energy compact structures. The transition to the native state (believed to be unique) takes place from these MECS. In fact, the existence of these pockets in free energy surface has already been demonstrated in both off and on lattice simulations [9,10]. More recently, a direct observation of such basins of attraction has been made in some lattice models of proteins [11], thus lending credence to this notion.

The heteropolymer model consists of self-avoiding walks of N sites on the square lattice. Sites in the SAW can be either hydrophobic (H) or polar (P). The energy of a given configuration is

$$E=-\varepsilon\sum_{i>j\in H}\delta_{|\mathbf{r}_i-\mathbf{r}_j|,a}\,, \tag{4}$$

where $\mathbf{r}_i$, $i=1,2,\ldots,N$, are the coordinates of the SAW sites, $\varepsilon>0$, and a is the lattice spacing. The interactions involving P sites are taken to be zero. Hence, minimum energy structures correspond to those configurations with the largest number of nearest neighbors involving H sites only.

For this model, we have defined *compact structures* (CS) as those conformations with the largest number of nearest neighbors. The minimum energy condition, however, imposes an additional constraint in the space of CS. We have quantified the effect of this condition by enumerating and classifying the MECS according to the number of sites N, and the number of H sites N_H. We have computed the set of self-avoiding CS [12] for $N\leq 30$ using the Martin algorithm [13]. We have fixed the direction of the first step of all structtures in order to eliminate the fourfold symmetry of the square lattice. Our series of compact structures is in complete agreement with that of Chan and Dill [14]. For fixed values of N and N_H, we have generated all possible $\binom{N}{N_H}$ sequences. Finally, for every sequence, we have computed the corresponding energy for *all* CS, thus obtaining the MECS for $N\leq 22$ [15]. We have also performed identical computations for larger chains, with $N=23$-26, using a finite number (about 10^4) of randomly generated sequences.

We are interested in computing the number of MECS averaged over all possible sequences, $\overline{C_{N,N_H}(\mathrm{MECS})}$, as a function of N and N_H. However, it is not clear whether the direct (or so-called annealed) average coincides with the most probable value of the number of MECS. In fact, there are few sequences for which the number of MECS equals the total number of CS. One way around this problem is to compute the so-called quenched average of the number of MECS, namely, $\mathcal{C}_{N,N_H}=\exp[\overline{\ln\{C_{N,N_H}(\mathrm{MECS})\}}]$. From the viewpoint of evolution one can argue that since useful mutations take place on a long time scale random sequences can be treated as being quenched. A three-dimensional plot of $\mathcal{C}_{N,N_H}$ as a function of N and the number of P sites $N_P=N-N_H$ is shown in Fig. 1. Clearly, $\mathcal{C}_{N,N_H}$ and $\overline{C_{N,N_H}(\mathrm{MECS})}$ are equal to C_N(compact) for $N_H=0$, 1, and N. The most striking aspect of this figure is the dramatic decrease in the average number of MECS observed for intermediate values of N_H, the actual minimum being at N_H^{min}. (Data are available upon request.)

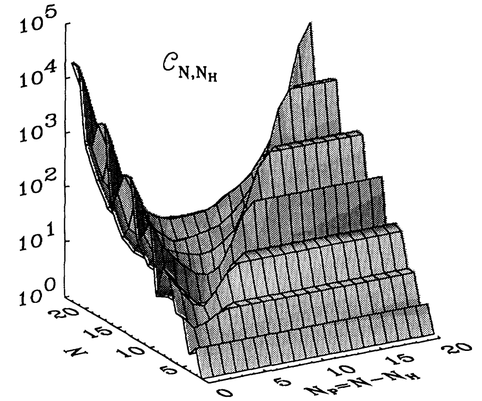

FIG. 1. The average number of minimum energy compact structures $\mathcal{C}_{N,N_H}$ on the square lattice is plotted as a function of the number of monomers N and the number of polar monomers $N_P=N-N_H$ for $N\leq 22$ (for $N_P>N$, N_H have been set equal to 0).

It has long been realized that due to the rather strong corrections to the leading asymptotic behavior, including some probable surface terms [6], series of compact structures are very difficult to analyze beyond leading order. These corrections are evident in Fig. 2 where C_N(compact) and $\overline{C_{N,N_H^{\mathrm{min}}}(\mathrm{MECS})}$ are plotted as a function of N. The steplike shape of C_N(compact) shows a clear $\sqrt{N}$ ($\propto$ perimeter) periodicity in the size of each dip. To get the overall trend of C_N(compact), we have coarse grained the series over $\sqrt{N}$ intervals, and the resulting averages $\langle C_N(\mathrm{compact})\rangle_{\sqrt{N}}$ are shown as full cir-

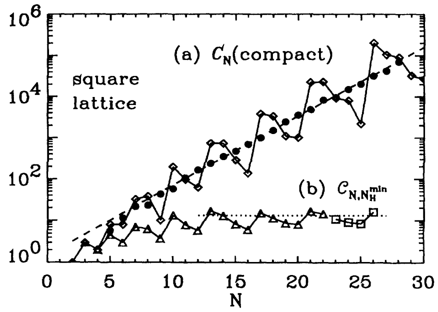

FIG. 2. The number of compact structures (a) C_N(compact) and of (b) the average number of minimum energy compact structures for N_H^{min}, $\mathcal{C}_{N,N_H^{\mathrm{min}}}$ are plotted as a function of N, and are denoted by open symbols. The series (a) is coarse grained over $\sqrt{N}$ intervals, and the averages $\langle C_N(\mathrm{compact})\rangle_{\sqrt{N}}$ are indicated as full circles. The dashed line represents the theoretical estimate $(z/e)^N$ with $z=4$. The dotted line corresponds to the estimate (b) $\mathcal{C}_{N,N_H^{\mathrm{min}}}=13$.

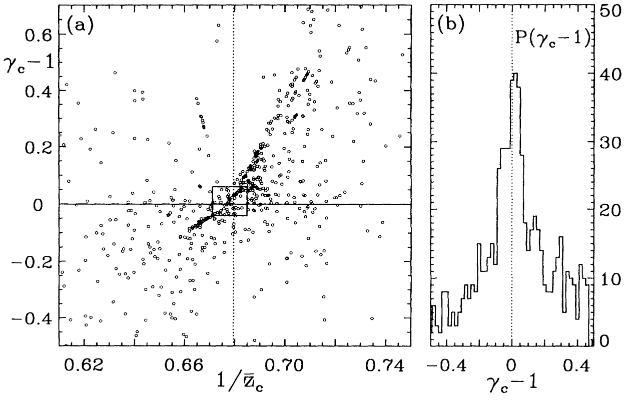

FIG. 3. (a) Differential approximant analysis of the average number of compact structures $\langle C_N(\text{compact})\rangle_{\sqrt{N}}$ on the square lattice. Each symbol represents one candidate for $1/\bar{z}$ and γ_c-1. The vertical dotted line indicates the mean field value $1/\bar{z}_c=e/4=0.679\,57$. The accumulation near $1/\bar{z}_c$ suggests $\gamma_c-1\simeq0$; the rectangle encloses our best estimates. (b) Histogram $P(\gamma_c-1)$ of all possible candidates for γ_c-1. The dotted line indicates $\gamma_c-1=0$.

cles in Fig. 2. For values of N for which the periodicity is not well defined, direct interpolation between nearest or next to nearest neighbors is performed. The reliability of the mean field estimate $\bar{z}=z/e$ [7], with $z=4$ on the square lattice, is confirmed by the dashed line in Fig. 2. It is noteworthy that the appropriately averaged series of $C_N(\text{compact})$ is surprisingly smooth, suggesting that corrections to the leading exponential term in (3) may vanish. We have investigated this possibility by applying well established series-analysis techniques including differential approximants [16] to $\langle C_N(\text{compact})\rangle_{\sqrt{N}}$. This has led to the conclusion that

$$\gamma_c=1.01\pm0.05 \quad\text{and}\quad \bar{z}=1.47_5\pm0.01_5\,. \tag{5}$$

The differential approximant analysis of the series is shown in Fig. 3. No surface term is apparent from our analysis of the *averaged* series. Indeed, our results do not show any significant deviation from the mean field values $\bar{z}=4/e$ [17] and $\gamma=1$ [see Fig. 3(b)], suggesting that the compact phase of polymers may in fact be described by a mean field theory.

The scaling of MECS appears to be dramatically different from that of $C_N(\text{compact})$. If we assume that the same periodicities observed in $C_N(\text{compact})$ are also present in $\mathcal{C}_{N,N_H^{\min}}$, we can estimate $\mathcal{C}_{N,N_H^{\min}}\sim12\text{–}14$ (see dotted line in Fig. 2); thus, $\mathcal{C}_{N,N_H^{\min}}$ approaches a constant value *independent* of N [18]. The annealed average $\overline{C_{N,N_H}(\text{MECS})}$ behaves similarly, the only difference seems to be that the asymptotic behavior is obtained for larger values of N. In any case, our results clearly demonstrate that the average number of MECS in model proteins is far fewer than the corresponding number of compact structures [19].

At this point, we summarize the total number of relevant configurations found for a chain of $N=100$ residues under all the aforementioned conditions [20]:

Dimension	Lattice	Ideal	Unfolded	Compact	MECS
$d=2$	Square	10^{60}	10^{43}	10^{17}	13
$d=3$	Simple cubic	10^{78}	10^{67}	10^{34}	$\cdots$

We expect the number of CS and MECS to depend only on the lattice structure, and therefore, for $d=3$ the number of MECS should also be a constant.

These numbers confirm Levinthal's conclusion that no random search, *not even among compact structures,* will be able to fold a protein on the time scale of seconds. However, a folding kinetic scheme that includes a very fast collapse of the protein structure, plus a relatively fast location of the low-lying states that would correspond to one of the MECS [21], may well be enough to allow folding of a protein to occur on a biological time scale. Indeed, in an earlier kinetic study [11] we have argued that the time scale for reaching a favorable MECS should scale as τ_0N^ζ, where $\zeta\sim2\text{–}3$ corresponds to a dynamical folding exponent, and τ_0 is a microscopic time constant. Based on the rough estimates of Ref. [11], this time extrapolates to milliseconds for a protein with $N=100$ residues. Finally, we are led to conclude that the rate limiting step for folding should correspond to the transition between the minimum energy compact states and the native state. It is interesting to note that, based solely on our estimates for the size of the space of conformations, the aforementioned *three stage kinetic* scheme (unfolded → compact → MECS → native) arises as a natural scenario for folding on a biologically accessible time scale. This kinetic scheme is further supported by recent simulations of lattice models of proteins, where explicit studies of the kinetics of approach to the native state from a denatured state show that folding occurs in three distinct stages [11].

From a physical point of view, there should be an optimum value, or range of values, of N_H/N for proteins to adopt well defined three dimensional structures [2]. In fact, if $N_H/N\ll1$, i.e., most of the residues are polar, proteins would dissolve in water. On the other hand, if $N_H/N\to1$ proteins would collapse and precipitate out of solution. On an average, proteins have a ratio of hydrophobic residues (as defined in Ref. [22]) of $N_H/N\simeq0.54$.

On very general grounds, we expect that in the limit

$N \rightarrow \infty$, $\mathcal{C}_{N,N_H^{\min}}$ should occur at some well defined ratio $N_H^{\min}/N$. However, does $N_H^{\min}/N$ approach unity as hydrophobic sites form the core of the protein and polar sites distribute on the surface? Or is $N_H^{\min}/N$ close to the minimum possible ratio for which a compact ground state is obtained, apparently about 0.5? From our very limited data, it seems that the latter possibility is more likely as $N \rightarrow \infty$, in agreement with the ratio of hydrophobic residues found in proteins [22]. Our data suggest a rough estimate for $N_H^{\min}/N$ is 0.6 ± 0.1. The fewest number of MECS are found when hydrophobic monomers are somewhat more prevalent than polar monomers.

From an evolutionary point of view, our results add support to the hypothesis that proteins could have originated from random sequences. Since roughly half of the naturally occurring amino acids are hydrophobic [22], it follows that on an average a random heteropeptide sequence should have a ratio of $N_H/N \simeq 0.5$. Our numerical results strongly suggest that this ratio is not only very convenient to stabilize proteins in solution [2], but also it is close to the apparent optimum ratio needed to achieve the fewest number of MECS. Notice that here we do not appeal to any stability argument and that our conclusions follow from explicit computations of the MECS.

In summary, using very simple considerations, in particular, the chemical heterogeneity of the amino acid sequence and hydrophobiclike forces, we have shown that the number of minimum energy compact structures is far less than the number of compact structures. It appears that in the intermediate stages of protein folding, minimum energy compact structures should act as basins of attraction for most pathways. Hence, as long as the protein is able to reach the MECS on a millisecond time scale the folding process can occur on a biological time scale. Our calculations along with the kinetic scheme reported earlier [11] provide a rationale for resolving the Levinthal paradox.

We are grateful to Michael E. Fisher for a critical reading of the manuscript and to Alexander M. Gutin for suggesting that we perform a quenched average of the number of MECS. Financial support from NSF is acknowledged.

[1] C. Levinthal, in *Mossbauer Spectroscopy in Biological Systems,* edited by P. Debrunner, J. C. M. Tsibris, and E. Münck (University of Illinois Press, Urbana, 1968).

[2] K. A. Dill, Biochemistry **24**, 1501 (1985).

[3] M. E. Fisher and B. J. Hiley, J. Chem. Phys. **44**, 616 (1961).

[4] See, e.g., A. J. Guttmann, J. Phys. A **20**, 1839 (1986).

[5] B. Nienhuis, Phys. Rev. Lett. **49**, 1062 (1982).

[6] (a) M. E. Fisher, Physics (Long Island City, N.Y.) **3**, 255 (1967); M. E. Fisher and B. U. Felderhof, Ann. Phys. (N.Y.) **58**, 176 (1970); (b) T. G. Schmalz, G. H. Hite, and D. J. Klein, J. Phys. A **17**, 445 (1984); (c) A. L. Owczarek, T. Prellberg, and R. Brak, Phys. Rev. Lett. **70**, 951 (1993).

[7] H. Orland, C. Itzykson, and C. de Dominicis, J. Phys. (Paris), Lett. **46**, L353 (1985).

[8] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).

[9] J. D. Honeycutt and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **87**, 3526 (1990); Biopolymers **32**, 695 (1992).

[10] P. E. Leopold, M. Montal, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **89**, 9721 (1992).

[11] C. J. Camacho and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **90**, 6369 (1993).

[12] Series of compact structures date back to W. J. C. Orr, Trans. Faraday Soc. **43**, 12 (1947); see also Ref. [3]. More recently, H. S. Chan and K. A. Dill have studied these series in the context of protein folding; see, e.g., Proc. Natl. Acad. Sci. U.S.A. **87**, 6388 (1990), and references therein.

[13] J. L. Martin, in *Phase Transitions and Critical Phenomena,* edited by C. Domb and M. S. Green (Academic, New York, 1974), Vol. 3.

[14] H. S. Chan and K. A. Dill, Macromolecules **22**, 4559 (1989).

[15] Notice that MECS are not necessarily the same as minimum energy structures. In fact, there are some sequences for which the ground state is not compact (as defined here).

[16] M. E. Fisher and H. Au-Yang, J. Phys. A **12**, 1677 (1979); **13**, 1517 (1980); D. L. Hunter and G. A. Baker, Jr., Phys. Rev. B **19**, 3808 (1979).

[17] Other analyses of CS have yielded even better agreement with the mean field expression of Orland *et al.*; see, e.g., references in Ref. [7] and Ref. [6(b)].

[18] This conclusion was conjectured by A. M. Gutin (private communication), based on results of the discrete random energy model; see A. M. Gutin and E. I. Shakhnovich, J. Chem. Phys. (to be published).

[19] This possibility was already hinted at by Dill in Ref. [2].

[20] Values of $z_{\rm eff}$ were taken from Ref. [4]. The estimate quoted for MECS corresponds to $N_H^{\min}$.

[21] As far as we know, Ref. [11] shows the first direct evidence for the close connection between the number of low-lying states and folding time scales.

[22] S. H. White and R. E. Jacobs, J. Mol. Evol. **36**, 79 (1993).