Conformations of Proteins in Equilibrium

Cristian Micheletti,[1] Jayanth R. Banavar,[2] and Amos Maritan[1]
[1]International School for Advanced Studies (S.I.S.S.A.) and INFM, Via Beirut 2-4, 34014 Trieste, Italy
[2]Department of Physics and Center for Materials Physics, 104 Davey Laboratory,
The Pennsylvania State University, University Park, Pennsylvania 16802

(Received 8 March 2001; published 3 August 2001)

We introduce a simple theoretical approach for an equilibrium study of proteins with known native-state structures. We test our approach with results on well-studied globular proteins, chymotrypsin inhibitor (2ci2), barnase, and the alpha spectrin SH3 domain, and present evidence for a hierarchical onset of order on lowering the temperature with significant organization at the local level even at high temperatures. A further application to the folding process of HIV-1 protease shows that the model can be reliably used to identify key folding sites that are responsible for the development of drug resistance.

DOI: 10.1103/PhysRevLett.87.088102 PACS numbers: 87.15.By, 64.60.Cn, 87.10.+e

Recent experimental and theoretical advances [1] have shown that the topology of the native structure of a protein plays an important role in determining many of its attributes. The number of distinct native-state conformations of proteins is limited [2]—often several distinct sequences fold into the same native-state structure. The native-state structures of proteins contain secondary motifs (helices and sheets) in lower dimensional manifolds which are curled into neat patterns (somewhat analogous to the packing of clothes in a suitcase) and play a central role in the folding process [3–5].

The problem of protein folding entails the study of the nonequilibrium dynamics in a rugged free-energy landscape [6]. A valuable starting point for attacking such a problem is through a thorough equilibrium analysis of proteins with known native-state structures. This would be useful for the determination of the folding transition temperature by, for example, monitoring the temperature at which thermodynamic quantities such as the specific heat show a peak. Such a study would lead to clear indications of the equilibrium conformations of proteins as a function of the temperature and provide a detailed picture of the onset of native-state-like order on lowering the temperature through the folding transition temperature. Ideally, one would like information on the non-native contacts and their role in facilitating the onset of native-state ordering. Furthermore, it would be useful to incorporate amino-acid specific interactions whenever such information is available.

At the present time, there is no simple theoretical framework for accomplishing all these objectives. Go-like models [7] have proved to be useful for incorporating the role of the topology of the native-state structure in the folding process. In such models, one ascribes a favorable attractive energy to the native contacts. There have been numerous studies of Go-like models, which capture the notion of minimal frustration [6], but even here, for realistic off-lattice calculations, it is hard, if not impossible, to carry out a full exploration of phase space in order to deduce equilibrium averages. As mentioned previously, the ruggedness of the free-energy landscape carves out only a small part of phase space that the system tends to be in leading to nonergodic behavior even for modest size proteins. Commonly used dynamics such as Monte Carlo or molecular dynamics tend to result in the system being compartmentalized in phase space because of barriers that are difficult to surmount as the temperature is lowered.

Recent progress has been made in the development of physically motivated topology-based models [3,8–12]. The models vary greatly in complexity and analytic tractability. The only energy contributions are postulated to arise from the establishment of native interactions, and therefore it is impossible to recover information on non-native contacts. In addition, a huge reduction in phase space is achieved by introducing suitable constraints on contiguous spin variables along the chain. As a consequence, the "true" hierarchical formation of the native-state structure may lead to incompatibilities with the phase-space constraints.

In this paper, we present a simple model for calculating the equilibrium properties of heteropolymers or proteins with known native states. The model builds on the importance of the native-state topology by assigning an attractive interaction between nearby amino acids that are known to make native-state contacts. It is also possible to incorporate amino-acid-specific interactions into the model. While the model, in its present simple form, does not accurately represent the effects of self-avoidance [13], it nevertheless ensures that the native state is the true ground state and satisfies all the steric constraints. The connectedness of the chain and its entropy are captured in a simple, but nontrivial, manner. The most significant advantage of the model is that it can be used to explore the equilibrium thermodynamics without being hampered by inaccurate or sluggish dynamics. A self-consistent approximation is used to reduce the model to a Gaussian [14] form. The latter lends itself to a straightforward determination of equilibrium quantities that identify the key folding sites [3], such as those targeted by drugs against viral enzymes [15,16].

The conformation of a protein is specified by the location of the C^α atom of the ith amino acid in sequence, $\vec{r}_i$. In the native state, $\vec{r}_i \equiv \vec{r}_i^{\,0}$. A simple effective Hamiltonian (energy function) that captures the essential features of proteins is

$$\tilde{\mathcal{H}} = \frac{T}{2} K \sum_{i=1}^{N-1} (\vec{r}_{i,i+1} - \vec{r}^{\,0}_{i,i+1})^2 + \sum_{i,j} \frac{\Delta_{i,j}}{2} X_{i,j} \theta[-X_{i,j}], \tag{1}$$

where $\vec{r}_{i,j} \equiv \vec{r}_i - \vec{r}_j$, T is the temperature ($K_B = 1$), θ is the step function,

$$\theta(x) = \begin{cases} 1 & \text{if } x > 0\,, \\ 0 & \text{otherwise}\,, \end{cases} \tag{2}$$

Δ is the contact matrix, whose element Δ_{ij} is 1 if residues i and j are in contact in the native state (i.e., their C_α separation is below the cutoff $c = 6.5$ Å) [17] and 0 otherwise, and

$$X_{i,j} = (\vec{r}_{i,j} - \vec{r}^{\,0}_{i,j})^2 - R^2\,. \tag{3}$$

The first term in the Hamiltonian involves harmonic interactions between successive "beads" in the chain [18]. The temperature factor ensures that the free-energy contributions of the peptide chain are constant over the range of temperatures relevant to the folding process. The second term in (1) provides an attractive interaction when amino acids which are in contact in the native-state conformation are in the proximity of each other [7]. Furthermore (1) guarantees that a specific native target structure is the ground state among all possible three dimensional structures. It is straightforward to generalize the Hamiltonian so that all the pairwise interactions do not have the same strength but are different and reflect the amino-acid-specific interactions. In standard off-lattice approaches, the interaction between nonbonded amino acids at a distance d is taken to be a square well potential, or some type of Lennard-Jones interaction. Our choice in Eq. (1) is a sort of "harmonic well" which, while being physically sound and viable, is suitable for a self-consistent treatment, as explained below. The location of the outer rim of the well is controlled by R, which can be set to a few angstroms ($R = 3$ Å the present study) to reflect the fact that, when the separation of two residues is appreciably different from the native one, their interaction is negligible. In its present form, the model is complex and not amenable to a simple attack. While the first term has a simple quadratic form, the second term is difficult to deal with because of the step function.

The key observation is that a dramatic simplification is accomplished on making a self-consistent Gaussian approximation within which the partition function is $\mathcal{Z} = \int \Pi_i d^3 r_i \, e^{-\beta \mathcal{H}}$, with

$$\mathcal{H} = \frac{T}{2} K \sum_{i=1}^{N-1} (\vec{r}_{i,i+1} - \vec{r}^{\,0}_{i,i+1})^2 + \sum_{i,j} \frac{\Delta_{i,j}}{2} X_{i,j} p_{i,j}\,, \tag{4}$$

where the $p_{i,j}$'s are determined self-consistently, in a spirit similar to a local mean field approximation:

$$p_{i,j} \equiv \langle \theta[R^2 - (\vec{r}_{i,j} - \vec{r}^{\,0}_{i,j})^2] \rangle_{\mathcal{H}}\,. \tag{5}$$

Physically, $p_{i,j}$ represents the equilibrium probability of the formation of the i-j contact at temperature T. $p_{i,i+1}$ can be conveniently frozen to 1 to reflect the strength of the peptide chain. With the functional form given in (4), the partition function can be readily deduced because all integrals are Gaussian. The free energy, $F = -T \ln \mathcal{Z}$, is

$$F = -\frac{3N}{2} \ln(2\pi) - \frac{3}{2} \ln(\det M) - \frac{R^2}{2T} \sum_{lm} \Delta_{lm} p_{lm}\,, \tag{6}$$

where the inverse matrix M^{-1} is defined as

$$M^{-1}_{i,j} = \begin{cases} K(2 - \delta_{i,1} - \delta_{i,N}) + 2\sum_l \Delta_{i,l} p_{i,l}/T & \text{for } i = j\,, \\ -2p_{i,j}\Delta_{i,j}/T + K[-\delta_{i,j+1} - \delta_{i,j-1}] & \text{for } i \neq j\,. \end{cases} \tag{7}$$

The self-consistency relations for the probabilities $p_{ij} = \langle \theta_{i,j} \rangle_{\mathcal{H}}$ are satisfied by finding the fixed point of the recursion equation,

$$p'_{ij} = (2\pi G_{i,j})^{-3/2} \int d^3 r \, e^{-r^2/2G_{i,j}} \theta(R^2 - r^2), \tag{8}$$

where the right-hand side depends on the p_{ij} obtained at the previous iteration through the matrix M, which enters the definition of $G_{i,j} = M_{i,i} + M_{j,j} - M_{i,j} - M_{j,i}$. The solution of the recursion equations for the $p_{i,j}$, Eq. (8), entails the evaluation of incomplete Γ functions and converge to the fixed point very fast and typically an accuracy of 10^{-4} is reached in a few dozen iterations. Thus the Gaussian nature of the Hamiltonian allows a straightforward analytic attack on the problem which, when combined with a rapidly convergent iterative procedure on a computer, allows one to determine the equilibrium properties of any protein with ease.

We present here the results for the globular proteins 2ci2, barnase, and the α-spectrin SH3 domain (PDB codes: 2ci2, 1a2p, and 1shg, respectively) for the simple case of uniform attractive interactions between the amino acids which form the native contacts. In all cases, it is straightforward to determine various thermodynamic quantities as a function of temperature and identify [19] the folding transition temperature as one at which the specific heat exhibits a maximum (see, e.g., Fig. 1). The width of the specific heat peak at the folding transition in Fig. 1 is significantly smeared out compared to experiment [20] and theory [21]. The cooperativity of the model can be easily controlled by adjusting the value of K, with smaller values leading to sharper "transitions." In addition, a change of K also impacts on the average amount of native structure that is formed at the native state. Because we are particularly interested in characterizing the progressive formation of the

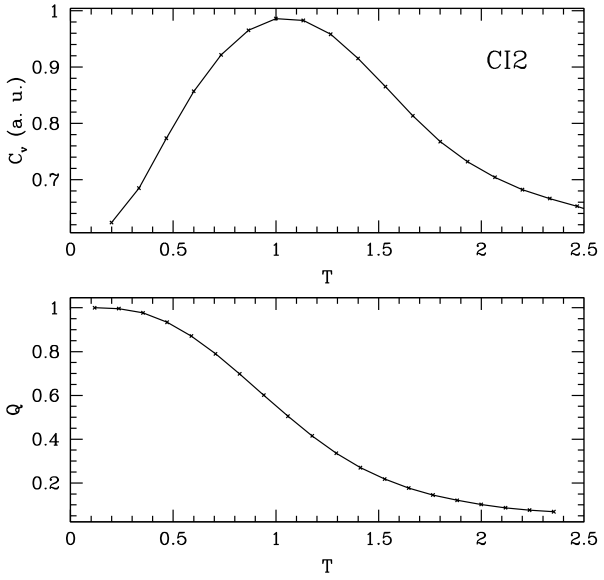

FIG. 1. Plot of the specific heat (in arbitrary units) and the native-state overlap as a function of temperature for the protein, chymotrypsin inhibitor. The temperature is measured in units of the folding transition temperature (identified through the maximum of the specific heat).

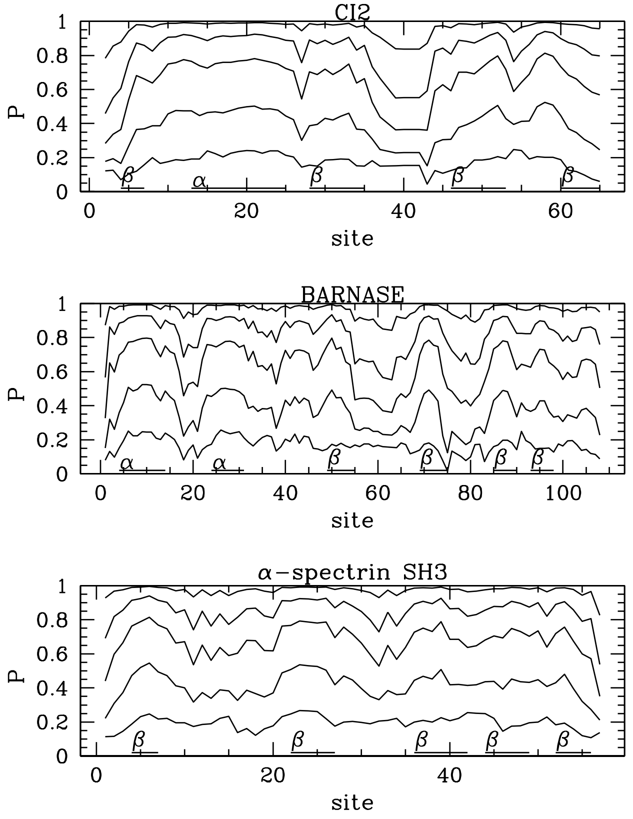

FIG. 2. Plot of P_i, the degree to which amino acid i is in a nativelike conformation, versus i for (a) 2ci2, (b) barnase, and (c) α-spectrin SH3. In ascending order the curves are calculated at $T = 2.0$, 1.5, 1.0, 0.5, and 0.35 (measured in units of the folding transition temperature). The bar at the bottom shows the secondary structure associated with amino acid i.

native interactions, we chose the strength of the peptide chain, K, by inspecting the behavior of the fraction of native contacts, Q, as a function of temperature. Q, which is often termed "native-state overlap," is defined as

$$Q = \frac{\sum'_{i,j} \Delta_{ij} p_{ij}}{\sum'_{i,j} \Delta_{ij}}, \tag{9}$$

where the prime denotes that the sum is not carried out over consecutive pairs, in order to exclude the effects of the peptide bond. We find that, almost irrespective of the length of the target proteins, a value of $K = 1/15$ yields $Q \approx 0.5$ at the folding transition, consistent with experimental findings [22] and previous observations [3,11,16,23,24].

While Q is a good global parameter to characterize the progress towards the native state in a folding process, it is useful to monitor the onset of native ordering at the level of individual residues. The quantity, P_i,

$$P_i = \frac{\sum'_{j} \Delta_{ij} p_{ij}}{\sum'_{j} \Delta_{ij}}, \tag{10}$$

provides an intuitive measure [9,25] of the degree to which amino acid i is in its nativelike conformation. Figure 2 shows the profiles of such environments for each amino acid in proteins CI2, barnase, and the α-spectrin domain of SH3 as a function of the temperature.

In agreement with the experimental findings on these heavily investigated proteins [26–28], and also with theoretical predictions [29–31], we observe that the secondary structures form at relatively high temperatures and condition subsequent folding events. Note the significant lack of ordering of the loop regions even at the folding transition temperature. β sheets are seen to form along one preferred direction, while the formation of α helices occurs from the ends. In general, the tendency of a site to reach its native environment increases with both its degree of burial and the locality of the contacts it forms. The intricate details of the figure reflect an incremental assembly process and also the complex interplay between the two effects mentioned above.

To further corroborate the validity of the proposed model in capturing the important folding steps we consider an application to an important enzyme, the protease of the HIV-1 virus (pdb code 1aid), which plays a vital role in the spreading of the viral infection. Through extensive clinical trials [15], it has been established that there is a well-defined set of sites in the enzyme that are crucial for developing, through suitable mutations, resistance against drugs and that play a crucial role in the folding process [16].

To identify the key folding sites we looked for contacts that are significantly formed above the folding transition temperature. Quantitatively we define the formation temperature of a contact as one at which a given $p_{i,j}$ takes

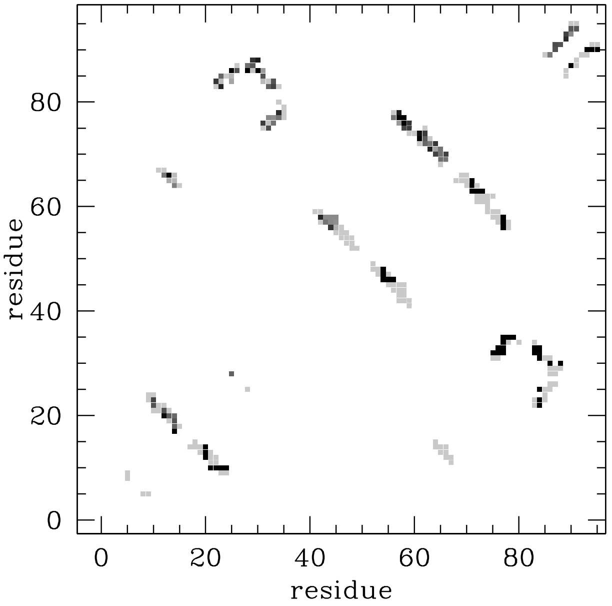

FIG. 3. Contact map of HIV-1 PR monomer. Upper-left triangular region: Contacts with a large (small) formation temperature are shown in dark (light) gray. Lower-right triangular region: contacts (not) involving one or more of the known key mutating sites are shown in dark (light) gray.

on the value 0.5. Our results are summarized in Fig. 3 (upper-left triangular region). As a comparison, in the lower-right triangular region of the same figure, we have highlighted the contacts involving the known mutating sites. Remarkably, among the top 20 contacts with the largest formation temperature there were 10 including one (or more) known mutating site. A straightforward calculation shows that the probability of observing this many successful matches by picking contacts at random is less than 2%, confirming that our model captures important aspects of the folding process with remarkable precision and reliability.

In summary, the self-consistent Gaussian approach provides a simple way of probing the equilibrium properties of proteins with the incorporation of amino-acid-specific interactions (when known) shorn away from the usual complications associated with imperfect or inadequately studied dynamics. A key advantage is that our approach is essentially analytic and the quantities of interest may be determined with arbitrary accuracy quite easily.

We are indebted to Alessandro Flammini for a careful reading of the manuscript. This work was supported by INFM, Murst Cofin99, and NASA.

[1] D. Baker, Nature (London) **405**, 39 (2000).

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

[3] C. Micheletti, J. R. Banavar, A. Maritan, and F. Seno, Phys. Rev. Lett. **82**, 3372 (1999).

[4] A. Maritan, C. Micheletti, and J. R. Banavar, Phys. Rev. Lett. **84**, 3009 (2000).

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

[6] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, Science **267**, 1619 (1995).

[7] N. Go and H. A. Scheraga, Macromolecules **9**, 535 (1976).

[8] V. Muñoz, E. R. Henry, J. Hofrichter, and W. A. Eaton, Proc. Natl. Acad. Sci. U.S.A. **95**, 5872 (1998).

[9] O. V. Galzitskaya and A. V. Finkelstein, Proc. Natl. Acad. Sci. U.S.A. **96**, 11299 (1999).

[10] E. Alm and D. Baker, Proc. Natl. Acad. Sci. U.S.A. **96**, 11 305 (1999).

[11] C. Clementi, H. Nymeyer, and J. N. Onuchic, J. Mol. Biol. **298**, 937 (2000).

[12] A. Flammini, J. R. Banavar, and A. Maritan, "*Energy Landscape in the Protein Folding Problem*" (to be published).

[13] The model does not include a hard-core repulsion between the amino acids. However, steric clashes are avoided to some extent because of the incorporation of the geometry of the self-avoiding native state.

[14] Recently, B. Erman and K. A. Dill, J. Chem. Phys. **112**, 1050 (2000), have demonstrated the viability of a pure Gaussian model for an attack of the folding problem.

[15] P. J. Ala *et al.,* Biochemistry **37**, 15 042 (1998).

[16] F. Cecconi, C. Micheletti, P. Carloni, and A. Maritan, Proteins **43**, 365 (2001).

[17] Alternatively, residues can be considered to be in contact if they contain heavy atoms that are within 4–5 Å. This alternative choice does not significantly alter the results discussed here (data not shown).

[18] M. Doi and S. F. Edwards, *The Theory of Polymer Dynamics* (Oxford University Press, Oxford, 1986).

[19] M. H. Hao and H. A. Scheraga, Physica (Amsterdam) **244A**, 124 (1997).

[20] S. E. Jackson, M. Moracci, N. Elmasry, C. M. Johnson, and A. R. Fersht, *et al.,* Biochemistry **32**, 11 259 (1993).

[21] H. Kaya and H. S. Chan, Phys. Rev. Lett. **85**, 4823 (2000); Proteins **40**, 637 (2000).

[22] K. M. Plaxco, K. T. Simons, and D. Baker, J. Mol. Biol. **277**, 985 (1998).

[23] M. Karplus and D. L. Weaver, Nature (London) **260**, 404–406 (1976); Protein Sci. **3**, 650 (1994).

[24] H. S. Chan and K. A. Dill, Proteins **30**, 2 (1998).

[25] T. Lazaridis and M. Karplus, Science **278**, 1928 (1997).

[26] L. S. Itzhaki, D. E. Otzen, and A. R. Fersht, J. Mol. Biol. **254**, 260 (1995).

[27] A. Matouschek, L. Serrano, and A. R. Fersht, J. Mol. Biol. **224**, 819 (1992).

[28] A. R. Viguera, L. Serrano, and M. Willmanns, Nat. Struct. Biol. **3**, 874 (1996).

[29] A. R. Fersht, *Structure and Mechanism in Protein Science* (W. H. Freeman and Company, New York, 1998).

[30] R. L. Baldwin and G. D. Rose, Trends Biochem. Sci. **24**, 26–33 (1999).

[31] T. Hoang and M. Cieplak, J. Chem. Phys. **113**, 8319 (2000).