Protein Structures and Optimal Folding from a Geometrical Variational Principle

Cristian Micheletti,[1] Jayanth R. Banavar,[2] Amos Maritan,[1] and Flavio Seno[3]
[1]*International School for Advanced Studies (SISSA) and INFM, Via Beirut 2-4, 34014 Trieste, Italy*
and The Abdus Salam Centre for Theoretical Physics, Trieste, Italy
[2]*Department of Physics and Center for Materials Physics, 104 Davey Laboratory, The Pennsylvania State University,*
University Park, Pennsylvania 16802
[3]*INFM–Dipartimento di Fisica, Via Marzolo, 8, 35100 Padova, Italy*
(Received 16 November 1998)

A novel approach, validated by an analysis of barnase and chymotrypsin inhibitor, is introduced to elucidate the paramount role played by the geometry of the protein backbone in steering the folding to the native state. It is found that native states of proteins, compared with compact artificial backbones, have an exceedingly large number of conformations with a given amount of structural overlap with them; moreover, the density of overlapping conformations, at a given overlap, of unrelated proteins of the same length are nearly equal. These results suggest an extremality principle underlying protein evolution, which, in turn, is shown to be possibly associated with the emergence of secondary structures. [S0031-9007(99)08834-1]

PACS numbers: 87.15.By, 64.60.Cn, 87.10.+e

The rapid and reversible folding of proteins into their thermodynamically stable native state [1] is accompanied by a huge reduction in conformational entropy [2,3]. Evidence has been accumulating for an achievement of the entropy reduction through a folding funnel favoring the kinetic accessibility of the native state [4–8]. Some fundamental questions remain unanswered. What makes proteins special compared to random heteropolymers? What guides the folding of a protein? Is it the sequence that is fundamental or its native structure?

In this Letter, we examine these issues and focus on the special role played by the native structure of proteins, with no input of information regarding amino acid sequences. The study is carried out through a novel theoretical probe of the conformation space of proteins: a measure of the density of overlapping conformations (DOC) having a given overlap or percentage of contacts in common with a fixed native structure. We show with studies on chymotrypsin inhibitor (reference 2ci2 of the Brookhaven Protein Data Bank) and barnase (1a2p) that the DOC provides key information on the folding pathway. An analysis of the DOC for real protein structures and for artificially generated decoy ones suggests that an extremal principle is operational in nature, which maximizes the DOC at intermediate overlap, providing a large basin of attraction [5–7,9,10] for the native state and promoting the emergence of secondary structures.

Our study consists of determining the number of structures with a given structural similarity to a putative native state. The structural similarity between the native structure and another one is defined as the percentage of native contacts in the alternative conformation. It is well known that such a measure is a good coordinate characterizing the folding process [11–13]. As is customary, two residues are defined to be in contact if the distance between their C_α atoms is less than 6.5 Å. In an unbiased study, conformations that differ slightly should not be considered distinct. To avoid this problem, we coarse-grain the configurational degrees of freedom by adopting a discretization approach [14] where the C_α's occupy sites on a suitably oriented fcc lattice (of edge 3.8 Å). This discretization does not distort the peptide angles, and the position of the coarse-grained C_α's differ from the true ones by typically less than 1 Å root mean square deviation [15]. For proteins of about 100 residues, the contact maps [16] of the real and fcc coarse-grained contact maps are virtually identical.

The generation of conformations was carried out using a standard Monte Carlo procedure (see, e.g., Refs. [13,17]) which allows one to move simultaneously up to 7 randomly chosen C_α's to unoccupied fcc sites. Each conformation is required to satisfy certain constraints of steric overlap and peptide geometry. These constraints (any two nonconsecutive residues cannot be closer than 4.65 Å due to excluded volume effects and the peptide bond is not stretched beyond 5.37 Å) were determined after carrying out an fcc coarse-graining of several proteins of intermediate length ($\approx$100 residues) and enforced in the generation of protein-like conformations.

In order to minimize the effects of correlation between successively generated structures, we typically discarded 50 elementary moves before accepting each new conformation. A newly generated conformation was accepted with the usual Metropolis rule according to the change in the Boltzmann weight: e^{Δ/K_BT}, where Δ is the change in contact overlap and T is a fictitious temperature. By choosing T appropriately, one can readily generate conformations with a desired average contact overlap, $\bar{q}$. At a given temperature, the true number of structures with overlap q is proportional to the number of conformations with overlap q obtained in the simulation multiplied by the Boltzmann weight. On undoing the Boltzmann bias, it is possible to recover the true density of conformations in a region around $\bar{q}$. In order to obtain the density of conformations for all values of overlap, we performed *2500* Monte Carlo samplings at different decreasing temperatures and then used standard deconvolution procedures [18]. Overall, for each distinct value of the overlap, more than 1000 structures were sampled. We have confirmed that the DOC curves are independent of the starting conformation and that the "folding" DOC obtained starting from a random conformation and cooling agrees to better than 3% (in logarithmic scale) with the "unfolding" DOC obtained starting from the target structure and increasing the temperature.

We begin with the backbones of the chymotrypsin inhibitor and barnase. We generated 2500 structures with a not too large overlap [19] (≈40%) for each of them. It turned out that the most frequent contacts shared by the native conformation of 2ci2 with the others involved the helical residues 30–42 (see top of Fig. 1). Contacts involving such residues were shared by 56% of the sampled structures. On the other hand, the rarest contacts pertained to interaction between the helix and β strands and between the β strands themselves. This is in excellent agreement with the studies of Fersht *et al.* [21,22], which observed the formation of the helix at early stages of the folding [23]. A different behavior (see bottom of Fig. 2) was found for barnase, where, again, for overlap of ≈40%, we find many contacts pertaining to the nearly complete formation of helix 1 (residues 8–18), a partial formation of helix 2, and bonds between residues 26–29 and 29–32 as well as several nonlocal contacts bridging the β strands, especially residues 51–55 and 72–75. This picture is fully consistent with the experimental results obtained in Ref. [24]. This provides a sound *a posteriori* justification that the main features of the folding of a protein can be followed from a study of the DOC. Remarkably, the method discussed above relies entirely on structure-related properties and suggests that the main features of the folding funnel are determined by the geometry of the "bare" backbone, while the finer details, of course, depend on the specific well-designed sequence.

We now turn to an analysis of three proteins of length 51 (1hcg, 1hja, and 1sgp) which have nearly the same number of native contacts (≈83). For each structure, we calculated the DOC with the constraint that the total number of contacts in the alternative structures do not exceed the number of contacts in the native state by more than 10% to avoid excessive compactness. To assess whether the DOC associated with naturally occurring proteins had special features, we generated three decoy conformations of the same length and number of contacts, but with different degrees of short- and long-range contacts (in sequence separation). These decoys (subject to the aforementioned "physical constraints") were generated with a simulated annealing procedure to find the structure with the highest overlap with a target contact matrix. By tuning the number of short-range versus long-range entries in the target random contact matrix, we generated three structures with different degrees of compactness and local geometrical regularity.

The logarithmic plots of the DOC are shown in Fig. 3. A striking feature of the curves is that, for intermediate overlap, the DOC of the real proteins is enormously larger than that of the decoys and suggests that naturally

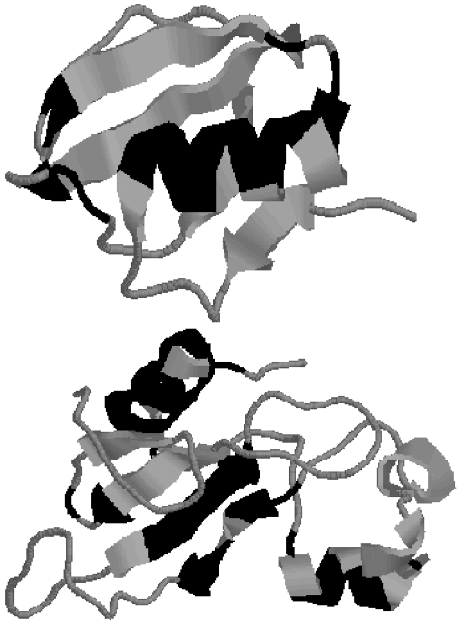

FIG. 1. Ribbon plot (obtained with RASMOL) of 2ci2 (top) and barnase (bottom). The residues involved in the 12 (16) most frequent contacts of alternative structures with overlap ≈40% with the native conformations are highlighted in black. The majority of these coincide with contacts that are formed at the early stages of folding.

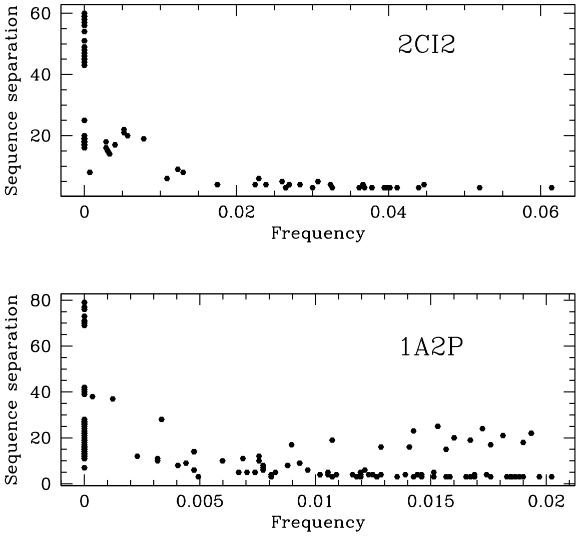

FIG. 2. Distribution of sequence separation of contacts commonly found in the conformations that overlap with the native state structures of 2ci2 and 1a2p. The most frequent contacts for 2ci2 have a small sequence separation (3–4) and pertain to helix formation. The 1a2p case shows a very different behavior with several contacts with very large sequence separation.

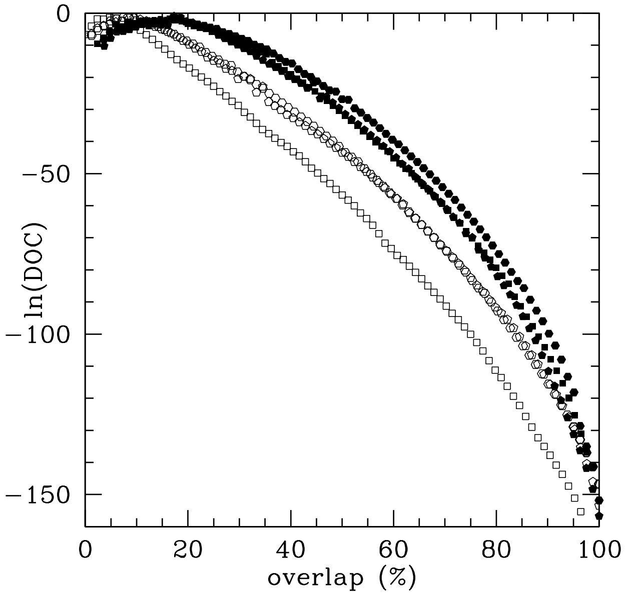

FIG. 3. Density of overlapping conformations for proteins for 1sgp (filled squares), 1hja (filled pentagons), and 1hcg (filled hexagons). Curves for artificial decoy structures are denoted by the open symbols.

occurring conformations have a much larger number of entryway structures than random compact conformations. Furthermore, for very high values of the overlap, the steepness of the protein curves is much larger than those of the decoys, showing that the reduction in the conformational entropy is also higher. This implies the existence of a funnel with a very large basin and steep walls. Another significant feature is the good collapse of the protein curves. We verified that this feature also obtains for 1bd0 and 2pk4 which each have 80 residues and 140 and 146 contacts, respectively. A simple explanation for the curve collapse could be that the DOC of real proteins is "extremal," in that it is close to the maximum possible value for intermediate values of the overlap.

The importance of the locality of contacts for folding kinetics was highlighted recently by Plaxco *et al.* [25] who found a correlation between folding rate and contact order, defined as the average sequence separation of contacts normalized to the total number of contacts and sequence length. With reference to Fig. 3, the contact order values for proteins 1hcg, 1hja, and 1sgp are 0.139, 0.214, and 0.204, respectively. For the decoy structures, they are 0.424, 0.222, and 0.179 for the curves denoted by open squares, pentagons, and hexagons, respectively. The structure with an unusually high contact order has the lowest DOC curve and optimal sequences designed on it (or equivalently a Go-like model [11]) would be expected to exhibit slow folding dynamics [26] in accord with the findings of Plaxco *et al.* [25].

A ubiquitious feature of protein structures is the existence of secondary-structure motifs [27,28] which have characteristic signatures in the contact maps, such as bands parallel to the diagonal (α helices and parallel β sheets) or orthogonal to it (antiparallel β sheets). We have carried out some simple investigations to assess whether a correlation exists between the extremality of the DOC curve and the emergence of secondary-structure-like motifs. We considered a space of contact maps [16], within which each of the residues interacted with the same number of other residues, n_c (typically $n_c = 5$, as in the average case of a protein with about 100 residues and a cutoff distance of 6.5 Å). This space contains maps corresponding to both real structures and unphysical ones. Furthermore, to mimic the effects of the rigidity and geometry of the peptide bond, we disallowed contacts between residue i and the four neighboring residues along the sequence $i \pm 2$, $i \pm 1$.

In this context, the maximization of the density of states corresponds to finding the target matrix with the highest number of matrices sharing a given fraction of its contacts. Although it is difficult to solve this problem, for arbitrary values of the overlap, it is relatively easy to generate matrices with an overlap close to the maximum value, $\bar{q}_{\max}$ (for a $L \times L$ matrix, $\bar{q}_{\max} = Ln_c$). To enumerate all matrices with overlap $\bar{q}_{\max} - 2$, one first identifies a pair of nonzero entries in the target matrix $\bar{m}$: $\bar{m}_{ij} = \bar{m}_{kl} = 1$. Then it is necessary to check whether entries $\bar{m}_{il}, \bar{m}_{kj}$ are both "free" (i.e., equal to zero) and do not correspond to forbidden contacts (e.g., between i and $i + 1$). If this is so, the old pair of entries (and their symmetric counterpart) are set to zero, and the new ones to 1. By considering, in turn, all possible pairs of nonzero entries one can generate all matrices of overlap $\bar{q}_{\max} - 2$. Then, by performing a simulated annealing in contact-map space one can isolate the map having the highest number of matrices with overlap $\bar{q}_{\max} - 2$.

We carried out our calculations for values of L around 60. The optimal matrices exhibit clustering reminiscent of α helices and β sheets, as shown in the upper triangular region of Fig. 4. A more quantitative measurement of the secondary-structure content of the optimal matrices can be obtained by considering the correlation functions, $g_{\pm}(x) = \sum_i m_{i,i\pm x}$, which show peaks in correspondence with the sequence separation of residues involved in α helices and parallel β sheets (g_+) or antiparallel β sheets (g_-). A typical plot of the correlation functions for an optimal map of length 60 and for the protein 3ebx (length 62) are shown in Fig. 5. The similarity of the plots is striking, particularly because, in both cases, the height of the peaks in g_+ decreases with sequence separation, unlike the situation with g_-.

In summary, a novel approach is used to elucidate the key role played by the geometry of the protein backbone in providing a large basin of attraction to the native state. Strikingly, by studying the conformational entropy of a backbone it is possible to identify the peptide regions which come in contact at early stages of folding with no detailed information on the sequences that are housed in the target fold. Our results are consistent with the recent findings on the folding nucleus of Ref. [29] and of Ref. [30]

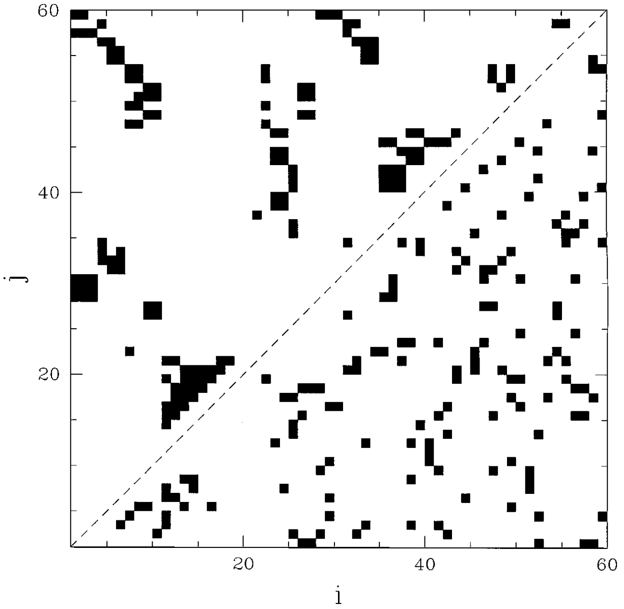

FIG. 4. The upper (lower) triangular region shows a target contact matrix with $L = 60$ that has a large (intermediate) number of contact maps with an overlap of $\bar{q}_{\max} - 2$ contacts.

in which it was shown that, compared to an arbitrary fold, the native state of a protein is in the proximity (in structure space) of many more low-energy structures, obtained by perturbing the original conformation. Our results are suggestive of an extremality principle underlying the selection of naturally occurring folds of proteins which, in turn, is shown to be possibly associated with the emergence of secondary structures.

This work was supported by grants from INFM, NASA, NATO, and The Petroleum Research Fund administered by The American Chemical Society.

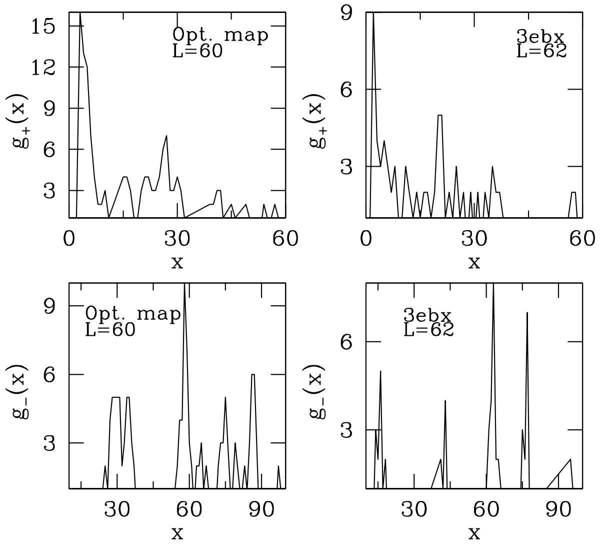

FIG. 5. Correlation functions (see text) for an optimal target matrix of length 60 and for a protein of length 62 taken from the protein data bank.

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

[2] M. Karplus and D. L. Weaver, Nature (London) **260**, 404–406 (1976); Protein Sci. **3**, 650–668 (1994).

[3] O. B. Ptitstyin, FEBS Lett. **285**, 176–181 (1991).

[4] H. S. Chan and K. A. Dill, J. Chem. Phys. **99**, 2116–2127 (1994).

[5] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524–7528 (1987).

[6] P. E. Leopold, M. Montal, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **89**, 8721–8725 (1992).

[7] J. N. Onuchic, P. G. Wolynes, Z. Luthey-Schulten, and N. D. Socci, Proc. Natl. Acad. Sci. U.S.A. **92**, 3626–3630 (1995).

[8] H. Nymeyer, A. E. Garcia, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **95**, 5921–5928 (1998).

[9] K. A. Dill and H. S. Chan, Nat. Struct. Biol. **4**, 10–19 (1997).

[10] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins: Struct. Funct. Genet. **21**, 167–195 (1995).

[11] N. Go, Macromolecules **9**, 535 (1976).

[12] C. J. Camacho and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **90**, 6369–6372 (1993).

[13] A. Sali, E. Shakhnovich, and M. Karplus, Nature (London) **369**, 248–251 (1994).

[14] D. G. Covell and R. Jernigan, Biochemistry **29**, 3287 (1990).

[15] B. H. Park and M. Levitt, J. Mol. Biol. **249**, 493–507 (1995).

[16] M. Levitt, J. Mol. Biol. **104**, 59–107 (1976).

[17] A. Kolinski and J. Skolnick, J. Chem. Phys. **97**, 9412 (1992).

[18] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **63**, 1195 (1989).

[19] The degree of the overlap may be benchmarked against the typical overlap of any two compact conformations, which is about 10%–20% (consistent with the unfolding simulations of Ref. [20]). A value of $q \approx 20\%$ thus corresponds to a "random coil" state.

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

[21] A. R. Fersht, Proc. Natl. Acad. Sci. U.S.A. **92**, 10 869 (1995).

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

[23] E. Shakhnovich, V. Abkevich, and O. Ptitstiyn, Nature (London) **379**, 96 (1996).

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

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

[26] P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **94**, 6170 (1997); S. S. Plotkin, J. Wang, and P. G. Wolynes, J. Chem. Phys. **106**, 2932 (1997).

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

[28] H. Li, R. Helling, C. Tang, and N. Wingreen, Science **273**, 666 (1996).

[29] N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, and E. I. Shakhnovich, cond-mat/9812284.

[30] D. Shortle, K. T. Simons, and D. Baker, Proc. Natl. Acad. Sci. U.S.A. **95**, 11 158 (1998).