found in cells much closer to the articular surface than normal. We conclude that Ihh, produced by differentiating chondrocytes, delays the differentiation of growth plate chondrocytes by stimulating the synthesis of perichondrial PTHrP, which then acts on the PTH/PTHrP receptor on chondrocytes.

REFERENCES AND NOTES

1. J. Potts Jr. *et al.*, in *Endocrinology*, L. DeGroot, Ed. (Saunders, Philadelphia, 1995), vol. 2, pp. 920–966.
2. A. E. Broadus and A. F. Stewart, in *The Parathyroids. Basic and Clinical Concepts*, J. P. Bilezikian, M. A. Levine, R. Marcus, Eds. (Raven, New York, 1994), pp. 259–294.
3. H. Jüppner *et al.*, *Science* **254**, 1024 (1991).
4. A. B. Abou-Samra *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 2732 (1992).
5. P. Ureña *et al.*, *Endocrinology* **133**, 617 (1993).
6. J. Tian, M. Smorgorzewski, L. Kedes, S. G. Massry, *Am. J. Nephrol.* **13**, 210 (1993).
7. M. Karperien *et al.*, *Mech. Dev.* **47**, 29 (1994).
8. K. Lee, J. D. Deeds, G. V. Segre, *Endocrinology* **136**, 453 (1995).
9. J. J. Orloff *et al.*, *Am. J. Physiol.* **262**, E599 (1992).
10. N. Inomata, M. Akiyama, N. Kubota, H. Jüppner, *Endocrinology* **136**, 4732 (1995).
11. S. Fukayama, A. H. Tashjian Jr., J. N. Davis, J. C. Chisholm, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 10182 (1995).
12. T. B. Usdin, C. Gruber, T. I. Bonner, *J. Biol. Chem.* **270**, 15455 (1995).
13. A. van de Stolpe *et al.*, *J. Cell Biol.* **120**, 235 (1993).
14. A. C. Karaplis *et al.*, *Genes Dev.* **8**, 277 (1994).
15. N. Amizuka, H. Warshawsky, J. E. Henderson, D. Goltzman, A. C. Karaplis, *J. Cell Biol.* **126**, 1611 (1994).
16. A. Vortkamp, K. Lee, B. Lanske, G. V. Segre, H. M. Kronenberg, C. Tabin, *Science* **273**, XXX (1996).
17. M. R. Capecchi, *ibid.* **244**, 1288 (1989).
18. Two genomic clones (λ-8 and λ-15) were isolated from a λ-DASH II 129/SvJ mouse liver genomic library (1×10^7 plaque-forming units per milliliter) with the coding sequence of the rat PTH/PTHrP receptor as a probe (*4*, *27*). For construction of the targeting vector, a 4.5-kb Bam HI fragment containing exon E1 of the PTH/PTHrP receptor gene was used as the 5′ flanking region and subcloned by blunt end ligation into a similarly treated Xho I site in the pPNT plasmid (*14*). A 5.6-kb Eco RI fragment containing the 3′ untranslated region of the PTH/PTHrP receptor gene, cloned into the Eco RI polylinker site of pPNT, was used as the 3′ flanking region. The resulting targeting vector was linearized at the unique Not I site in the pPNT backbone for electroporation of ES cells. Double selection was carried out, and resistant ES clones were analyzed by Southern (DNA) blot analyses with an external intronic 0.9-kb Sac I–Xho I fragment at the 5′ end of the gene. Positive clones (12%) were injected into recipient blastocysts to generate chimeras as described (*28*). Progeny of two independent clones yielded the identical phenotype. All animal experimentation followed institutional guidelines.
19. B. Lanske *et al.*, data not shown.
20. Twenty-five embryos from E6.5 to E9.5 were examined histologically and had normal appearing Reichert's membrane. Five E9.5 embryos were examined by in situ hybridization for PTH/PTHrP receptor mRNA (*29*). The hybridization signal was strong in four embryos but was completely absent from one, a presumed PTH/PTHrP (−/−) fetus. Immunohistochemical staining was carried out on paraffin sections by standard techniques. Hybridization was done with a rabbit antibody to α-laminin (dilution, 1/300; Gibco). To increase the sensitivity, we used a second antibody specific for the first one (swine antibody to rabbit immunoglobulin G complex; DAKO). Visualization of this complex was obtained by hybridization with the ABComplex (horseradish peroxidase; DAKO).
21. M. Karperien, P. Lanser, S. W. de Laat, J. Boonstra, L. H. K. Defize, *Int. J. Dev. Biol.*, in press.
22. F. Beck, J. Tucci, P. V. Senior, *J. Reprod. Fertil.* **99**, 343 (1993).
23. M. McLeod, *Teratology* **22**, 299 (1980).
24. Fresh tissues were obtained by cesarean sections from fetuses derived from heterozygous interbreeding at day 18.5 of gestation. Fetuses were fixed in 10% formalin–phosphate-buffered saline (PBS) (pH 7.2) and subsequently decalcified in neutral 40% EDTA. Various parts of the skeleton were embedded in paraffin, and 5- to 10-μm-thick sections were stained with hematoxylin and eosin.
25. E. Schipani, K. Kruse, H. Jüppner, *Science* **268**, 98 (1995).
26. Hind limbs of E16.5 fetuses were severed at midfemur, stripped of skin, and then placed on a filter paper (pore size, 0.8 μm) on a wire mesh in a Falcon organ culture dish, and 1 ml of BGJ_b medium (Gibco BRL) was added. The hind limbs, which lay at the air-fluid interface, were cultured at 37°C in a humidified atmosphere of 95% air–5% CO_2 with daily changes of medium. The hind limbs were treated with either 10^{-7} M human PTHrP (1–34), recombinant murine Shh (5 μg/ml) (*16*), or vehicle (BGJ_b medium–0.1% bovine serum albumin) alone from day 2 for 4 days. The samples were never exposed to serum. At the termination of culture, the hind limbs were fixed in 10% formalin-PBS, paraffin-embedded, and cut in serial sections.
27. X. F. Kong *et al.*, *Biochem. Biophys. Res. Commun.* **200**, 1290 (1994).
28. A. Bradley, in *Teratocarcinomas and Embryonic Stem Cells: A Practical Approach*, E. Robertson, Ed. (IRL, Washington, DC, 1987), pp. 113–152.
29. Complementary ^{35}S- labeled RNA probes were transcribed from the rat PTH/PTHrP receptor cDNA (*4*) and from the human Ihh cDNA (*16*). In situ hybridization was done as described (*8*).

30. We thank C. Tabin for valuable collaboration and helpful reading of the manuscript, T. Doetschmann for reagents, E. Samson for technical assistance with histology, and M. Mannstadt for help and technical expertise. Supported by National Institutes of Health grants DK 47038 and DK 47237. B.L. was supported in part by a fellowship of the Max-Kade Foundation.

10 April 1996; accepted 20 June 1996

Emergence of Preferred Structures in a Simple Model of Protein Folding

Hao Li, Robert Helling,* Chao Tang,† Ned Wingreen

Protein structures in nature often exhibit a high degree of regularity (for example, secondary structure and tertiary symmetries) that is absent from random compact conformations. With the use of a simple lattice model of protein folding, it was demonstrated that structural regularities are related to high "designability" and evolutionary stability. The designability of each compact structure is measured by the number of sequences that can design the structure—that is, sequences that possess the structure as their nondegenerate ground state. Compact structures differ markedly in terms of their designability; highly designable structures emerge with a number of associated sequences much larger than the average. These highly designable structures possess "proteinlike" secondary structure and even tertiary symmetries. In addition, they are thermodynamically more stable than other structures. These results suggest that protein structures are selected in nature because they are readily designed and stable against mutations, and that such a selection simultaneously leads to thermodynamic stability.

Natural proteins fold into specific compact structures despite the huge number of possible configurations (*1*). For most single-domain proteins, the information coded in the amino acid sequence is sufficient to determine the three-dimensional (3D) folded structure, which is the minimum free-energy structure (*2*). Protein sequences must undergo selection so that they fold into unique 3D structures. Because folding maps sequences to structures, it is relevant to ask whether selection principles also apply to structures that have evolved in nature. Protein structures often exhibit a high degree of regularity—for example, secondary structures such as α helices and β sheets and tertiary symmetries—that is absent from random compact structures. What is the origin of these regularities? Does nature select special structures for design? What are the underlying principles that govern the selection of structures?

Here we describe results from a simple model of protein folding that suggest some answers to these questions. We focus on the properties of each individual compact structure by determining the sequences that have the given structure as their nondegenerate ground state. We show that the number of sequences (N_S) associated with a given structure (S) differs from structure to structure and that preferred structures emerge with N_S values much larger than the average. These preferred structures are "proteinlike," with secondary structures and symmetries, and are thermodynamically more stable than other structures.

Our results are derived from a minimal model of protein folding, which we believe captures the essential components of the

NEC Research Institute, 4 Independence Way, Princeton, NJ 08540, USA.

*Present address: Second Institute for Theoretical Physics, DESY/University of Hamburg, Hamburg, Germany. †To whom correspondence should be addressed. E-mail: tang@research.nj.nec.com

problem. In this model, a protein is represented by a self-avoiding chain of beads placed on a discrete lattice, with two types of beads used to mimic polar (P) and hydrophobic (H) amino acids (3). A sequence is specified by a choice of monomer type at each position on the chain, $\{\sigma_i\}$, where σ_i could be either H or P, and i is a monomer index. A structure is specified by a set of coordinates for all the monomers, $\{r_i\}$. The energy of a sequence folded into a particular structure is given by short-range contact interactions

$$H = \sum_{i<j} E_{\sigma_i\sigma_j}\Delta(r_i - r_j) \qquad (1)$$

where $\Delta(r_i - r_j) = 1$ if r_i and r_j are adjoining lattice sites but i and j are not adjacent in position along the sequence, and $\Delta(r_i - r_j) = 0$ otherwise. Depending on the types of monomers in contact, the interaction energy will be E_{HH}, E_{HP}, or E_{PP}, corresponding to H-H, H-P, or P-P contacts, respectively (Fig. 1).

This simple model has some justification in nature. The major driving force for protein folding is the hydrophobic force (4). The tendency of amino acids to avoid water drives proteins to fold into a compact shape with a hydrophobic core, and such a force is effectively described by a short-range contact interaction. Although 20 different amino acids exist in nature, quantitative analysis of natural-protein data reveals that they fall into two distinct groups (H and P) according to their affinities for water (5). Experimental evidence also indicates that certain proteins can be designed by specification of only this HP pattern of the sequence (6).

We chose the interaction parameters in Eq. 1 to satisfy the following physical constraints: (i) compact shapes have lower energies than any noncompact shapes; (ii) H monomers are buried as much as possible, which is expressed by the relation $E_{PP} > E_{HP} > E_{HH}$, which lowers the energy of configurations in which H residues are hidden from water; and (iii) different types of monomers tend to segregate, which is expressed by $2E_{HP} > E_{PP} + E_{HH}$. Conditions (ii) and (iii) were derived from our analysis of the real-protein data contained in the Miyazawa-Jernigan matrix of interresidue contact energies between different types of amino acids (5).

We have studied the model on a 3D cubic lattice and on a 2D square lattice. We focus on the "designability" of each compact structure. Specifically, we count the number of sequences (N_S) that have a given compact structure (S) as their unique ground state, which requires identification of the minimum-energy compact conformations of each sequence. Because all compact structures have the same total number of contacts, we can freely shift and rescale the interaction energies, leaving only one free parameter. Throughout this paper, we chose $E_{HH} = -2.3$, $E_{HP} = -1$, and $E_{PP} = 0$, which satisfy conditions (ii) and (iii) above. The results are insensitive to the value of E_{HH} as long as both these conditions are satisfied.

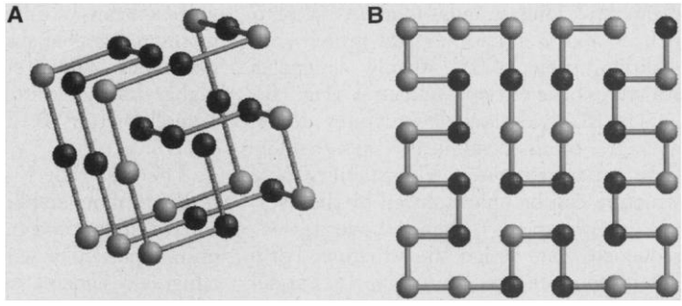

Fig. 1. Structures with the largest value of N_S for 3D 3 by 3 by 3 (**A**) and 2D 6 by 6 (**B**) systems. Each sequence represents one of the N_S possible sequences. Black beads indicate H residues; gray beads indicate P residues. Two beads are considered to be in contact if they are nearest neighbors but are not connected by the backbone.

For the 3D case, we analyzed a chain composed of 27 monomers. We considered all the structures that form a compact 3 by 3 by 3 cube. There are a total of 51,704 such structures unrelated by rotational, reflection, or reverse-labeling symmetries. For a given sequence, the ground state structure is determined by calculation of the energies of all compact structures. We completely enumerated the ground states of all 2^{27} possible sequences, showing that 4.75% of the sequences have unique ground states. As a result of this complete enumeration, we obtained all possible sequences that "design" a given structure—that is, that have that structure as their unique ground state. Thus, N_S is a measure of the designability of a given structure, and this information is available for all compact structures.

Compact structures were found to differ markedly in terms of their designability. There are structures that can be designed by a large number of sequences, and there are "poor" structures that can be designed by only a few or even no sequences. For example, the top structure can be designed by 3794 different sequences ($N_S = 3794$), whereas there are 4256 structures for which $N_S = 0$. The number of structures with a given N_S value decreases monotonically (with small fluctuations) as N_S increases (Fig. 2A). For structures that contribute to the long tail of the distribution, $N_S \gg \overline{N_S} = 61.72$, where $\overline{N_S}$ is the average number. We refer to these structures as "highly designable." The distribution differs markedly from the Poisson distribution that would result if the compact structures were statistically equivalent. For a Poisson distribution with $\overline{N_S} = 61.72$, the probability of finding even one structure with $N_S > 120$ is 1.76×10^{-6}.

Highly designable structures exhibit certain secondary structures that are absent from random compact structures. We examined the compact structures with the 10 largest N_S values and found that all have parallel running lines folded in a regular manner (Fig. 1A). These structures contain eight or nine strands (three amino acids in a row), whereas the average structure has only 5.4 strands.

To ensure that the above results were not artifacts of small size (the 3 by 3 by 3 cube), we also performed systematic studies of size dependence in two dimensions (the study of larger structures in three dimensions is not practical because of the limitations of computing power). We studied systems of sizes 4 by 4, 5 by 5, 6 by 5, and 6 by 6 on a 2D square lattice. For systems of sizes 6 by 5 and 6 by 6, a random sampling of sequences was performed (7). Comparison of systems of different sizes requires appropriate rescaling of the axes. We chose bin sizes for N_S to be proportional to $\overline{N_S}$, and rescaled the number of structures by a factor proportional to the total number of structures. For the 6 by 5 and 6 by 6 cases, we ensured that the random sampling of sequences produced a reliable distribution by doubling the number of sequences until a fixed distribution was achieved.

The systems of different sizes in two dimensions all showed the same qualitative behavior as that apparent in three dimensions. In each instance, there are highly designable structures that stand out. For the 6 by 5 and 6 by 6 systems, for which the total numbers of structures are sufficiently large to produce smooth distributions, the two distributions showed virtually identical shapes (Fig. 2B). The tail of the 2D distribution could be fitted by an exponential function (Fig. 2B, inset). In contrast, the falloff of the tail in the 3D distribution is slightly less than exponential.

Similarly to the 3D case, the highly designable structures in two dimensions also exhibit secondary structures. In the 2D 6 by 6 system, as the surface-to-interior ratio approaches that of real proteins, the highly designable structures often have bundles of

pleats and long strands, reminiscent of α helices and β strands in real proteins; in addition, some of the highly designable structures have tertiary symmetries (Fig. 1B).

The highly designable structures are, on average, thermodynamically more stable than other structures. The stability of a structure can be characterized by the average energy gap ($\overline{\delta_S}$), averaged over the N_S sequences that design the structure. For a given sequence, the energy gap (δ_S) is defined as the minimum energy required to change the ground-state structure to a different compact structure. For the 3D 3 by 3 by 3 structures, there is a marked correlation between N_S and $\overline{\delta_S}$ (Fig. 3). Highly designable structures have average gaps much larger than those of structures with small N_S values, and there is a sudden jump in $\overline{\delta_S}$) for structures with $N_S \approx 1400$. The number of structures with large gaps is 60. The abrupt jump in $\overline{\delta_S}$ is somewhat unexpected compared with the smooth distribution of N_S. Such an abrupt transition provides a means of differentiating the special, highly designable structures from the ordinary ones. According to this distinction, highly designable structures constitute only a small fraction (0.12%) of all the compact structures.

The fact that highly designable structures are more stable than other structures can be understood qualitatively by considering a particular sequence associated with a highly designable structure, S. A mutation of the sequence may change the energy of the structure S as well as those of the competing structures. If the gap is large, it is less probable that the energies of the competing structures will shift below that of the structure S. Thus, the structure S is likely to remain as the ground state of the mutant. Therefore, a large gap is likely to correlate with a large number of sequences that design the structure.

An important approach in studying real protein structures is to assess mutation effects and homologous sequences (sequences related by a common ancestor) (8). In our simple model, we refer to the N_S different sequences that design the same structure as "homologous." Analysis of mutation patterns of the homologous sequences for highly designable structures revealed phenomena similar to those observed in real proteins. For example, sequences with no apparent similarities (with different types of monomer at more than half of the sites) can design the same structure. Furthermore, some sites are highly mutable, whereas others are highly conserved. The conserved sites for a given structure are generally those sites with the smallest or largest number of sides exposed to water. Figure 4 shows the probability, P_P of finding a P monomer at a particular site, calculated for the structure with the largest N_S for the 3D 3 by 3 by 3 and the 2D 6 by 6 systems (Fig. 1). For the 3D case, there are sites that are perfectly conserved with $P_P = 0$ and $P_P = 1$.

The mutation pattern for the 2D 6 by 6 system shows additional characteristics:

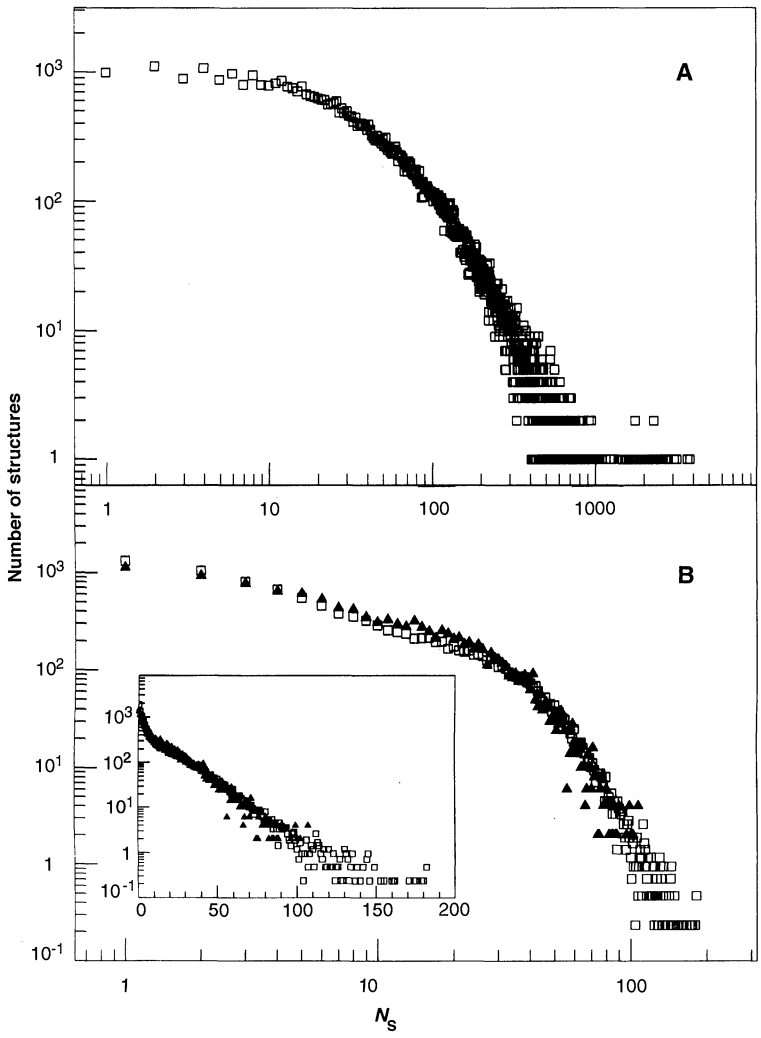

Fig. 2. (**A**) The number of structures with a given N_S for a 3D 3 by 3 by 3 system. (**B**) The number of structures with a given N_S for 2D 6 by 5 (▲) and 6 by 6 (□) systems. (Inset) Same data in a semi-log plot.

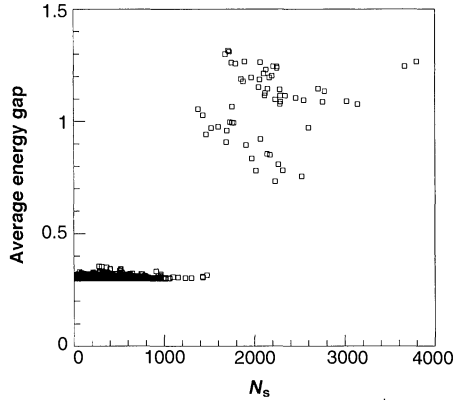

Fig. 3. Average energy gap of 3D 3 by 3 by 3 structures plotted against N_S of the structures.

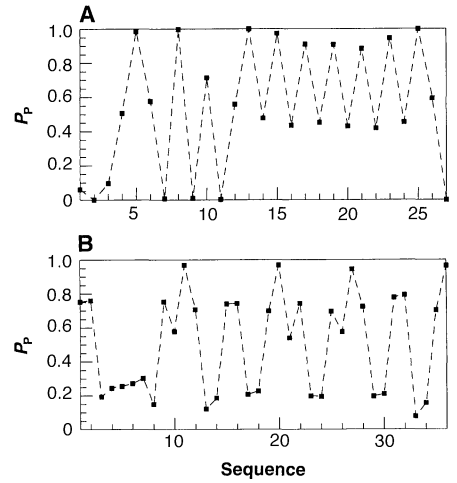

Fig. 4. Probability of finding a P monomer at a particular site, calculated for the structure with the largest N_S for 3D 3 by 3 by 3 (**A**) and 2D 6 by 6 (**B**) systems.

The pleated regions have alternating arrangements (with period 4) of H and P types, the region of long folded strands is essentially all H type, and the region connecting these substructures (similar to turns in proteins) contains predominantly P type monomers.

Mutation patterns can also be characterized by calculating the entropy of the homologous sequences for a given structure from

$$\text{Entropy} = \sum_{\text{sites}}[-P_P\ln(P_P) + (P_P - 1)\ln(1 - P_P)] \quad (2)$$

Given only the knowledge of entropy, an estimate of the number of all possible homologous sequences can be made, N_{est} = exp (entropy), assuming that mutation of each site is independent. We calculated N_{est} for all the structures and compared it with the exact value N_S. For the highly designable structures, N_{est} is a good order-of-magnitude estimate for N_S. For example, for the top structure in the 3D 3 by 3 by 3 case, $N_{est}/N_S \approx 3.5$. However, for the less designable structures, N_{est} greatly overestimates N_S. The large deviation begins at $N_S \approx 1400$, at the boundary between large-gap and small-gap structures. Thus, for highly designable structures, the mutations are roughly independent, whereas they are highly correlated for other structures.

Although our results for the 3D system were derived for small structures (3 by 3 by 3), we believe similar results hold for much larger structures. Evidence to this effect is provided by recent studies of design in larger structures by Yue and Dill (9) with a similar model. In a few examples studied by these researchers, they found that sequences with a small ground-state degeneracy corresponded to structures with certain proteinlike secondary structures and tertiary symmetries. In light of our results, we believe that such proteinlike structures are the highly designable structures with large N_S values. This interpretation is different from that of Yue and Dill, who suggest that minimal degeneracy is sufficient to produce proteinlike secondary structures and tertiary symmetries. Our results show that it is possible to find sequences that uniquely design even "poor" structures. It is the requirement that many sequences design a particular structure that leads to proteinlike secondary structures and tertiary symmetries.

Although the detailed structures of real proteins are determined by many factors—for example, hydrogen bonding and the shapes of the amino acids—our results from the simple model suggest that there is a principle of design and evolutionary stability that should play a crucial role in the selection of protein structures; that is, real protein structures must be highly designable and mutable. Because highly designable structures are also more stable, such a selection principle solves the problem of thermodynamic stability simultaneously. From an evolutionary point of view, highly designable structures are more likely to have been chosen through random selection of sequences in the primordial age, and they are stable against mutations.

Our proposed principle of selection based on designability and mutability should have important corollaries in prediction and design of protein structure. If, in fact, nature selects only highly designable structures, then structure prediction algorithms should limit the search of the conformational space to these special structures, which might constitute only a tiny fraction of the total number of possible structures. Indeed, certain structural motifs occur frequently in the protein structure data bank (*10*), and it has been estimated that only ~1000 possible folds exist in the structures of natural proteins (*11*). A relatively successful algorithm for structure prediction has been developed with the use of the templates from known protein structures (*12*). Our study lends theoretical support to such an approach. Further improvement depends on finding practical ways to identify highly designable structures.

An important question concerns the kinetic accessibility of these structures, and whether there are other selection principles imposed by kinetics. Studies have examined structure selection based on folding kinetics (*13*). It is likely that our highly designable structures also fold more readily because of the large gap in their excitation spectrum (*14*). We have performed successful preliminary folding simulations for some highly designable structures (*15*). A more systematic study of kinetics, including ordinary structures, is under way.

REFERENCES AND NOTES

1. T. E. Creighton, Ed., *Protein Folding* (Freeman, New York, 1992).
2. C. Anfinsen, *Science* **181**, 223 (1973).
3. K. A. Dill, *Biochemistry* **24**, 1501 (1985); K. F. Lau and K. A. Dill, *Macromolecules* **22**, 3986 (1989).
4. W. Kauzmann, *Adv. Protein Chem.* **14**, 1 (1959). For a recent review, see K. A. Dill, *Biochemistry* **29**, 7133 (1990).
5. H. Li, C. Tang, N. Wingreen, "Dominant driving force for protein folding—a result from analyzing the statistical potential" (preprint, cond-mat/9512111; http://xxx.lanl.gov/).
6. S. Kamtekar, J. M. Schiffer, H. Xiong, J. M. Babik, M. H. Hecht, *Science* **262**, 1680 (1993).
7. Complete enumerations for chains of shorter length (up to 18 monomers) have been performed previously, with a focus on average properties of sequences and structures [K. F. Lau and K. A. Dill, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 638 (1990); H. S. Chan and K. A. Dill, *J. Chem. Phys.* **95**, 3775 (1991); *Proteins* **24**, 335 (1996).
8. T. Alber *et al.*, *Nature* **330**, 41 (1987); J. F. Reidhaar-Olson and R. T. Sauer, *Science* **241**, 53 (1988).
9. K. Yue and K. Dill, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 146 (1995).
10. C. A. Orengo, D. T. Jones, J. M. Thornton, *Nature* **372**, 631 (1994).
11. C. Chothia, *ibid.* **357**, 543 (1992).
12. D. T. Jones, W. R. Taylor, J. M. Thornton, *ibid.* **358**, 86 (1992).
13. S. Govindarajan and R. A. Goldstein, *Biopolymers* **36**, 43 (1995); V. I. Abkevich, A. M. Gutin, E. I. Shakhnovich, *J. Mol. Biol.* **252**, 460 (1995).
14. A. Sali, E. Shakhnovich, M. Karplus, *Nature* **369**, 248 (1994).
15. R. Melin, H. Li, N. Wingreen, C. Tang, in preparation.

16. We thank W. Bialek, M. Hecht, A. Libchaber, G. McLendon, and Y. Meir for helpful discussions.

19 March 1996; accepted 4 June 1996