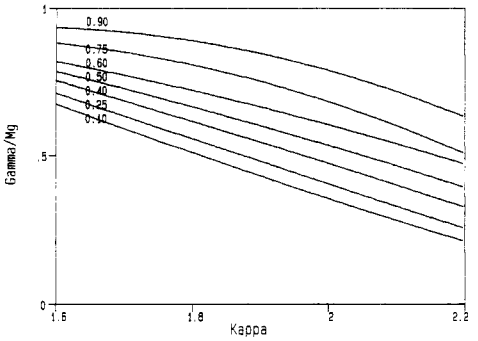

Figure 3. Plots of theoretical Γ_{Mg} versus κ at different N_1 values, using the values of parameters listed in Table III. The values of N_1 are indicated on the curves.

two ions. The dependence of the theoretical Γ_{Mg} on κ was calculated for N_1 values in the range 0.10–0.90. The resulting curves are shown in Figure 3. The curvature is large and negative at $N_1 > 0.6$ but is very small for $N_1 < 0.6$, eventually becoming positive at $N_1 < 0.3$. These results suggest that Γ_{Mg} versus κ will be linear enough for a test of this type for territorial binding when N_1 lies between 0.25 and 0.50. This condition was satisfied in the PG experiment using a 1:1 equivalent ratio of Mg^{2+} to Na^{2+} that did produce a linear relationship between Γ and κ.

Registry No. PG, 9046-38-2; Na, 7440-23-5; Mg, 7439-95-4.

References and Notes

(1) Anderson, C. F.; Record, M. T., Jr.; Hart, P. A. *Biophys. Chem.* **1978**, *7*, 301.
(2) Bleam, M. L.; Anderson, C. F.; Record, M. T., Jr. *Proc. Natl. Acad. Sci. U.S.A.* **1980**, *77*, 3085.
(3) Gustavsson, H.; Siegel, G.; Lindman, B.; Fransson, L. Å. *Biochim. Biophys. Acta* **1981**, *677*, 23.
(4) Manning, G. S. *Acc. Chem. Res.* **1979**, *12*, 443.
(5) Manning, G. S. *J. Chem. Phys.* **1969**, *51*, 924.
(6) Fuoss, R.; Katchalsky, A.; Lifson, S. *Proc. Natl. Acad. Sci. U.S.A.* **1951**, *37*, 579.
(7) Grasdalen, H.; Kvam, B. J. *Macromolecules* **1986**, *19*, 1913.
(8) Halle, B.; Wennerstrom, H.; Piculell, L. *J. Phys. Chem.* **1984**, *88*, 2482.
(9) Qian, C.; Asdjodi, M. R.; Spencer, H. G.; Savitsky, G. B. *Macromolecules* **1989**, *22*, 995.
(10) Dolar, D.; Peterlin, A. *J. Chem. Phys.* **1969**, *50*, 3011.
(11) Atkins, E. D. T.; Nieduszynski, I. A.; Mackie, W.; Parker, K. D.; Smolko, E. E. *Biopolymers* **1973**, *12*, 1865.

A Lattice Statistical Mechanics Model of the Conformational and Sequence Spaces of Proteins

Kit Fun Lau and Ken A. Dill*

Department of Pharmaceutial Chemistry, University of California, San Francisco, California 94143. *Received December 5, 1988; Revised Manuscript Received March 24, 1989*

ABSTRACT: We develop theory to explore the relationship between the amino acid sequence of a protein and its native structure. A protein is modeled as a specific sequence of H (nonpolar) and P (polar) residues, subject to excluded volume and an HH attraction free energy ϵ. Exhaustive exploration of the full conformational space is computationally possible because molecules are modeled as short chains on a 2D square lattice. We use this model to test approximations in a recent mean-field theory of protein stability. Also, exhaustive exploration permits us to identify the "native" state(s) in the model, the conformation(s) of global free energy minimum. We then explore the relationship between sequences and native structures by (i) further exhaustive exploration of the full space of all sequences, for short chains, and (ii) random selection of sequences, for longer chains, in some cases exploring exhaustively only the fully compact conformations. The model has the following properties. For small ϵ, the chains are unfolded. With increasing HH attraction, molecules with certain sequences fold to a state with relatively few conformations that have (i) low free energy, (ii) high compactness, (iii) a core of H residues, and (iv) substantial secondary structure. The potential of a molecule to fold to this state is predicted largely by the composition, but for intermediate compositions it depends also on the specific sequence of residues. Some folding sequences have multiple native states; those native structures are broadly distributed throughout the conformational space. However, a most interesting prediction is that, even with only the H and P discrimination among residues in this model, a folding sequence is most likely to have only a single native conformation, a predominance that increases with chain length.

Introduction

A problem of long-standing interest in biology has been that of predicting the three-dimensional structure of a globular protein from knowledge of its amino acid sequence. If the "thermodynamic hypothesis" of protein folding is correct, i.e., that the native structure of the globular molecule is that conformation which has the lowest free energy, then the native structure could be identified in principle simply by systematic evaluation of the free energy of every possible conformation. The problem is that this calculation is not yet practical using current force-field algorithms because the computer time required is far too great. The computer time scales with the number of conformations, which has an exponential dependence, a^n, on chain length n, where $a > 1$ is a constant that depends on chain flexibility and excluded volume.[1]

As a consequence, force-field studies at near-atomic resolution are presently limited to explorations of very small regions of conformational space and thus to problems involving dynamics,[2] binding, or catalytic mechanisms[3-4] wherein structural perturbations of the protein are small. Computer limitations currently preclude application of high-resolution methods to problems of (i) large conformational changes, (ii) predictions of thermodynamic properties of folding, which also require knowledge of the ensemble of unfolded reference states, or (iii) the prediction of the native structure from the primary amino acid sequence.

In the absence of high-resolution methods for large-scale exploration of conformational space, two alternative approaches have emerged to predict conformations of proteins from primary structures: (i) semiempirical methods and (ii) methods based on simplifications of conformational

0024-9297/89/2222-3986$01.50/0

space. Semiempirical methods make use of correlations observed in databases of known structural features such as nearest-neighbor pairwise residue conformations or radial distribution functions,[5-8] or they make use of extra constraining potentials to help achieve the known native structure.[9-18] Some limitations of some of these methods are well-known.[19-22]

Other methods involve considerable simplifications of the conformational space.[1,23-27] These methods also have limitations. One of them, a statistical mechanical treatment of stability,[1,27] does not attempt to address effects of the sequence of residues but only the effects of composition, for it is premised on averaging over the possible sequences. The method of Kolinski, Skolnick, and Yaris is also based on a lattice model to discretize the conformational space and uses simple potential functions.[23-25] The conformational space is then explored by Monte Carlo methods. An important finding of theirs is that certain secondary structural motifs arise naturally from these simple models. A principal conclusion from both of these models is that many of the general features of proteins can be predicted from simple and nonspecific interaction potentials. The model of Li and Scheraga[26] uses the most realistic potential function but so far appears to have been tested only on small peptides. Li and Scheraga also sample a discretized conformational space by Monte Carlo methods.

Our interest here is to explore the nature of the full conformational and sequence spaces of proteins. The methods cited above are therefore unsuitable for this purpose, since they explore, at most, a very few sequences, and since they sample conformational space only sparsely or through approximation. In the present work, we develop a third approach to the protein folding problem, in its simplest possible implementation. Our purpose is to explore the nature of the full conformational and sequence spaces of copolymeric chain molecules such as proteins. We develop a model that can explore every possible conformation of every possible sequence. This effort is in the same spirit as the use of model Hamiltonians, such as certain Ising models, which can be solved exactly through enumeration of every accessible state of the system. It follows from the remarks above, however, that such a model must be very simple and low-resolution. It also follows that the nature of the questions addressible by this model are different than those of the models described above. For example, what does conformational space look like? What does sequence space look like? Does the ability of the chain to fold into a compact state depend on the sequence of residues or only on the composition (the fraction of residues of each type), irrespective of their sequence? Are there "good" sequences and "bad" sequences, in terms of their potential to undergo a folding transition to a compact globular state with a hydrophobic core? How many minima of lowest energy are there in the conformational space? What is the energy distribution of the accessible minima in the conformational space?

We model a protein as a linear chain of n amino acids. Each amino acid can be either of two types: H (nonpolar) or P (polar). The fraction of the n residues that are of type H is Φ; hence the fraction of type P is $1 - \Phi$. A chain conformation is represented as a self-avoiding walk on a two-dimensional square lattice; an example of one compact conformation of one particular sequence is shown in Figure 1. The lattice simply serves as a tool to discretize the conformational space, i.e., to account for the freedom of the chain to have different backbone conformations and to account for the effects of excluded volume of any one

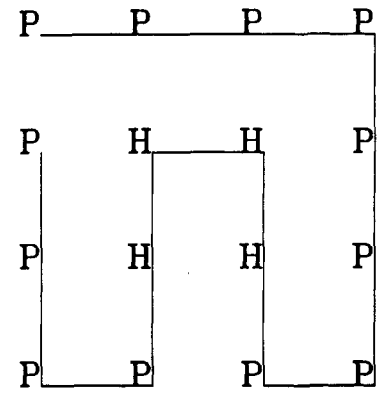

Figure 1. Model protein on a 2D square lattice with two types of residues, H and P.

chain segment by any other. Since each lattice site has $z = 4$ neighbors, the number of bond orientations for all internal chain segments is $z - 1 = 3$. Thus each amino acid is represented in this model as simply occupying one lattice site, connected to its chain neighbor(s), unable to occupy a site filled by another residue, and as indicated below, interacting with any other immediate spatially neighboring residue through an orientation-independent interaction energy.

The principal simplification implicit in the use of a 2D system as a model for molecules in three dimensions is simply in the lattice coordination number, z. In a 3D lattice model, there are more conformations per bond pair (i.e., $z - 1 = 5$ for a simple cubic lattice) and similarly there are more spatial nearest-neighbor residues with which one monomer can interact. In addition, excluded volume is a more severe constraint in 2D. These are the factors that contribute to the dimensionality dependence of the thermodynamic balance of forces (of conformational freedom versus contact interactions). Despite these quantitative differences, the qualitative physical behavior will be similar in 2D and 3D. The two principal advantages of the 2D model are that (i) the surface-to-volume ratio, a principal determinant of the physical behavior, of longer chains in 3D is the same as that for shorter chains in 2D and (ii), for a given chain length, the computational requirements are much less severe for the 2D model.

It is convenient to distinguish between pairs of monomers that are "connected neighbors", units j and $j + 1$ adjacent along the chain sequence, and pairs that are "topological neighbors", those that are adjacent in space (in contact) but are not adjacent in position along the sequence. We assume that every HH contact between topological neighbors has a contact free energy (divided by kT) equal to ϵ (<0) and every other interaction among neighbors (of any other pairwise combination of H, P, and solvent, S) has free energy equal to 0. In choosing the interaction potential this way, our zero-energy reference state is the very open state in which no topological contact is made. This interaction energy is among the simplest possible. However, the same interaction, used in conjunction with a mean-field approximation[1,27] predicts well the experimentally measured temperature[27] and solvent dependences of protein stability. We interchangeably refer to ϵ as a dimensionless energy or free energy, since we do not consider here the explicit temperature dependence of this contact interaction.

We define the "conformational space" and the "sequence space" as follows. The conformational space is the set of all the possible internal conformations of a molecule, due to all the different bond orientations. Thus in this model, the maximum size of the conformational space is ap-

proximately $(z-1)^{n-1}$. This simple estimate is a maximum because it double counts certain symmetric conformations and neglects excluded volume. Excluded volume is responsible for the reduction of the conformational space size to a^{n-1} where $a < z-1$.[1] In exhaustive simulation, we find that $a \approx 2.71$ when $z = 4$. The exact number of conformations accessible to a chain length of 10 on the 2D square lattice is found to be 2034. Of the entire conformational space of chain length $n = 10$, structures range from the open conformations, where no topological neighbors are made, to the most compact conformations, where the maximum number of topological neighbor contacts are made.

The sequence space is the set of all the possible sequences of H and P residues; in this model, the size of the sequence space is 2^n. Hence, for $n = 10$, there are 1024 different sequences.

Our aim is to calculate various properties of the conformational and sequence spaces, which we do as follows. For a particular sequence, we exhaustively enumerate the full conformational space and compute the energy for each configuration. In the process, we find the conformation(s) that is at the global minimum of free energy, which we refer to as the "native state(s)". (For sequences that do not have the potential to fold to compact structures, we still retain the term "native" state to refer to the conformations of lowest free energy, even though this terminology differs from that used in biochemistry, where the term "native" generally refers only to compact molecules.) For short chains, we then use this approach to find the native state for every possible sequence through exhaustive exploration of the sequence space. We compute various averages over the ensemble of all native states of all sequences in the sequence space.

The conformational enumeration process uses a depth-first algorithm, which seeks the longest branch of the directed graph representing all possible self-avoiding walks. The algorithm backtracks either when the full chain of a given length has been generated or there is a dead end due to an excluded-volume violation. The search stops when all nodes of the directed graph have been visited.

Exhaustive searches in both conformational space and sequence space are possible when the chain length is short ($n < 11$). To explore the properties of longer chains, we use (i) random sampling of the whole sequence space, with full search of the conformational space for each sequence, or (ii), for longer chains ($n \leq 30$), random sampling of the sequence space, with exhaustive search of only the maximally compact conformations, which is a small subset of all possible conformations. It is shown below that the set of maximally compact conformations nevertheless almost always contains the set of native conformations for chains that fold. Thus the latter subsearch produces native-state properties that do not deviate significantly from those obtained by the search of the full conformational space.

The maximally compact chain conformations are lattice walks in which every internal site is occupied by exactly one chain segment: no site is empty, and no site is filled by two chain segments. The maximally compact conformations have the largest number of topological neighbors, t_{max}, allowed by the chain length, where

$$t_{max} = n + 1 - P_{min}/2 \tag{1}$$

P_{min} represents the minimum surface the molecule can have. For the 2D square lattice, P_{min} is the perimeter of the smallest box that can contain all n residues. It is obvious that if k and m are the length and breadth of the rectangular box containing the chain, the perimeter $P = 2(k + m)$ will be minimum if $k = m$. If m is the smaller of the two integers, then the smallest box will be described by $m^2 < n \leq (m+1)^2$, and

$$\begin{aligned} P_{min} &= 4m + 2 \quad \text{for} \quad m^2 < n \leq m(m+1) \\ &= 4m + 4 \quad \text{for} \quad m(m+1) < n \leq (m+1)^2 \end{aligned} \tag{2}$$

The most compact conformations for a chain of length n are exhaustively enumerated by confining the lattice walks to this minimal rectangular box. The proof that this search is exhaustive for maximally compact conformations follows because any walk that crosses this rectangular boundary will have perimeter $> P_{min}$; hence t, the number of topological neighbors, will be smaller than t_{max}, the value given by eq 1.

The obvious advantage of this compact chain search is that the number of conformations which must be explored is reduced dramatically by a factor of approximately e^{-n}, and yet, as shown below, this search produces almost all the native conformations of folding molecules (i.e. the exceptions comprise a set of negligible size). This permits us to find the native states for molecules of lengths up to $n \approx 30$–36.

Stability and Denaturation

We first consider the process of collapse to a compact conformation. The simulations show that if the HH attraction is large, then chains with certain sequences of monomers become highly compact, with an internal core of H residues. With increasing HH attraction, the chains undergo a transition to this folded state from an ensemble of unfolded configurations. The nature of the collapsed state and the denaturation transition is described below.

The thermodynamic properties of a chain of fixed length n and given sequence of residues can be computed from the partition function, Z

$$Z = \sum_{m=0}^{s} g(m) e^{(s-m)\epsilon} \tag{3}$$

where $m = 0, 1, 2, \ldots, s$ is the number of HH topological contacts, and $g(m)$ is the degeneracy, the number of chain conformations with m HH topological contacts. ϵ (<0) is the HH contact energy divided by kT, Boltzmann's constant multiplied by temperature. The native conformation(s) represent the ground state(s) of the partition function.

We define the compactness, $\rho = (m + u)/t_{max}$, of any chain conformation in terms of u, the total number of topological neighbors of types HP and PP in the given chain conformation, m, the number of HH topological contacts, and t_{max}, the maximum possible number of topological neighbors of all types for the given chain length. Let the number of conformations that have m HH topological neighbor contacts and u other topological neighbor contacts be $G(m,u)$; it must be related to the degeneracy $g(m)$ by

$$g(m) = \sum_{u=0}^{t_{max}-m} G(m,u) \tag{4}$$

Thus the average compactness over all the chain conformations is

$$\langle \rho \rangle = Z^{-1} \sum_{m=0}^{s} \sum_{u=0}^{t_{max}-m} \frac{u+m}{t_{max}} G(m,u) \epsilon^{(s-m)\epsilon} \tag{5}$$

Using the following simple device, we can also compute the average compactness of only the native conformations. In the limit of infinite HH attraction, $\epsilon \to -\infty$, the maxi-

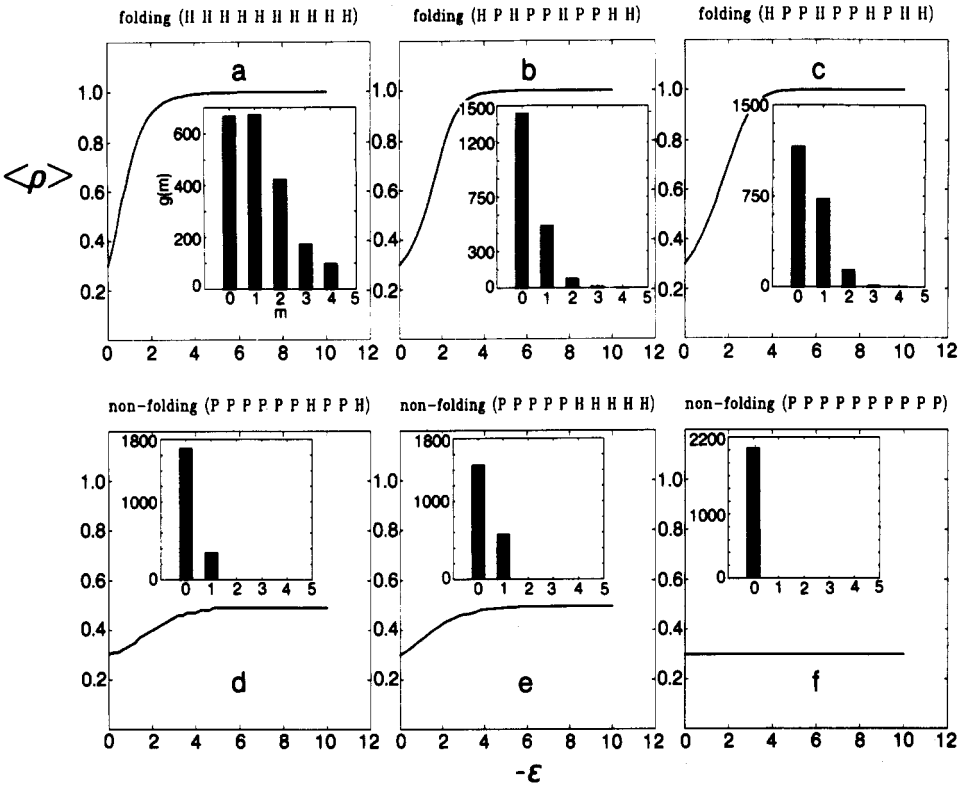

Figure 2. The average compactness $\langle\rho\rangle$ as a function of ϵ, the energy of HH contacts, for six different sequences: (a) HHHHHHHHHHH (Φ = 1); (b) HPHPPHPPHH (Φ = 0.5); (c) HPPHPPHPHH (Φ = 0.5); (d) PPPPPPHPPH (Φ = 0.2); (e) PPPPPHHHHH (Φ = 0.5); (f) PPPPPPPPPP (Φ = 0).

mum term of the partition function is only that of the native states,

$$Z_\infty = g(s) \tag{6}$$

Also

$$\lim_{\epsilon \to -\infty} \sum_{m=0}^{s} G(m,u)e^{(s-m)\epsilon} = G(s,u) \tag{7}$$

Hence the average compactness of the native state is defined by substitution of eq 6 and 7 into eq 5

$$\langle\rho\rangle_{ns} = \lim_{\epsilon \to -\infty} \langle\rho\rangle = Z_\infty^{-1} \sum_{u=0}^{t_{max}-s} \frac{u+s}{t_{max}} G(s,u) \tag{8}$$

The average compactness of the native state is a useful measure of the "folding potential" of the molecule, the degree to which the molecule can configure to a compact state of low free energy. For example, any sequence for which $\langle\rho\rangle_{ns} = 1$ represents a molecule whose conformations of lowest free energy are also maximally compact (i.e., those conformations with the maximum number of HH contacts also have the maximum total number of intrachain contacts). On the other hand, any sequence for which $\langle\rho\rangle_{ns} < 1$ represents a molecule that has at least some conformations of lowest free energy that are not maximally compact. These sequences have less folding potential insofar as the average compactness over the ensemble of native states will be smaller.

A chain with folding potential is simply one that can configure to a compact conformation of low contact free energy. The condition that $\epsilon \to -\infty$ is equivalent to accounting only for the contact interactions, with the driving force from the conformational entropy equal to zero. Hence the folding potential does not reflect the balance of forces for stability of the molecule. The stability (i.e., the free energy of folding) and other thermodynamic quantities are derivable from the partition function, Z, in the standard way. Similarly the average compactness over all the conformations is obtained from eq 5 instead of eq 8. These quantities, stability and average compactness, which depend on the balance of forces, are strongly dependent on chain length.[1] For example, short peptide chains do not fold under ordinary solution conditions because their compact cores would have too few HH contacts to overcome the conformational entropy.[1] Therefore simulations of the balance of forces for short chain peptides, which do not fold, would not be a good model for longer chain proteins, which do. On the other hand, the "folding potential" is a measure of whether there is any achievable conformation in which the lowest energy states are compact. For this property, a short-chain model is a useful predictor of the behavior of longer chains.

The dependence of average compactness, $\langle\rho\rangle$, on ϵ for six sequences is shown in Figure 2, along with the density of states for each sequence. For many sequences, increasing ϵ, the strength of the HH attraction, leads to increased average compactness. Many of the exceptions are sequences that have no or few H residues (Figure 2f). The density of states, $g(m)$, determines the ability of the sequence to fold and the sharpness of the transition. Those sequences that permit folding to the maximally compact state $\langle\rho\rangle_{ns} = 1$ must have at least a few conformations of low energy, whereas nonfolding sequences do not.

It is clear that some sequences permit folding to the maximally compact state, $\langle\rho\rangle_{ns} = 1$, but others do not. Figure 2d,e shows two nonfolding sequences. The inability to achieve the compact states follows from the impossibility of simultaneously forming many HH contacts and many total topological contacts. Thus the lowest energy states of nonfolding molecules consist of a large number of conformations of low compactness.

Some sequences fail to fold because of the composition, i.e., they have too few H residues, irrespective of their

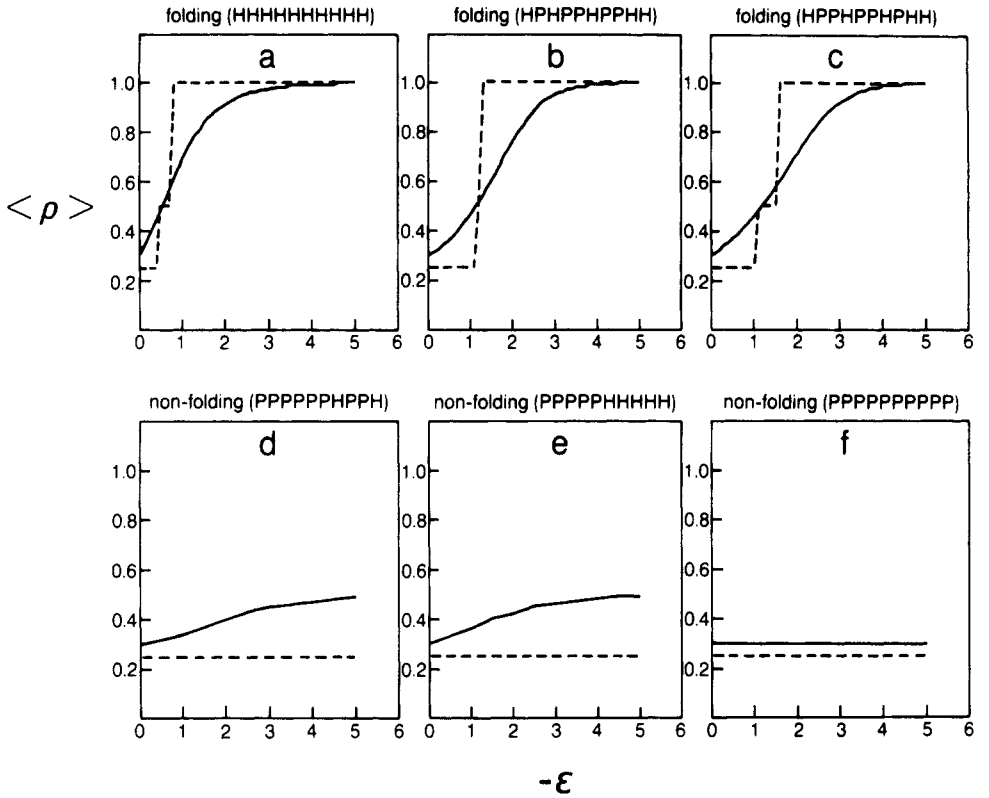

Figure 3. Comparison between $\langle\rho\rangle$ and ρ^* as a function of ϵ (—) $\langle\rho\rangle$, (- - -) ρ^*] for six different sequences as in Figure 2.

position along the chain. Others fail because of the specific sequence of the residues, even when the number of H residues is sufficient. As an example of the former, Figure 2d shows a case of $\Phi = 0.2$, for which the maximum number of HH contacts possible is one. When that contact is formed however, the chain is not very compact; it has many configurations consistent with that constraint. This is a problem with the composition since changing the position of the H residues in that sequence does not increase the folding potential. On the other hand, Figure 2b,c,e shows three sequences with different folding propensities, all of which have the same composition. The sequence shown in Figure 2e does not fold; it cannot achieve maximum density because it has a string of P residues at one end. In that case, even a very high value of the HH attraction will never contribute to the folding of the P tails. On the other hand, if P residues are interspersed in the sequence, they can be "carried along for the ride" as spacers between the H residue contacts.

Whereas the continuous curves in Figure 3 show the average densities $\langle\rho\rangle$ of these molecules, the dashed curves show, for comparison, the behavior of the "maximum term" ρ^*, i.e., only the most probable state. ρ^* is the density that gives rise to the largest term of the sum P_ρ leading to the partition function Z. Z can be written as

$$Z = \sum_\rho P_\rho$$

$$P_\rho = \sum_{i=1}^{n_\rho} e^{-E_i/kT}$$

where n_ρ is the number of conformations having density ρ and E_i is the energy of the ith conformation having density ρ.

This is of value for testing an approximation common to some Monte Carlo and mean-field models. The results show that the phase transition behavior is very different if the maximum term is used to approximate the true average. In particular, whereas the average shows a gradual change in density, the maximum term shows a first-order transition for some folding sequences and no transition for the nonfolding sequences (see Figure 3). It is quite interesting that in other cases, however, there is a relatively sharp transition to an intermediate state, then a further transition to the folded state (see Figure 3a,c). Although we have not done a complete statistical analysis, it appears that considerably more of the folding sequences show the intermediate state than the first-order transition. Of course, we do not know if the difference between the maximum term method and the true average would persist for longer chains or for molecules in three dimensions.

The folding process can alternatively be characterized by several different physical properties in addition to compactness. First is the energy, averaged over the ensemble of all conformations. Since the contact energy is $m\epsilon kT$ where m is the number of HH contacts and ϵ is the dimensionless interaction energy between each pair of HH contacts, $\langle m\rangle$ is proportional to the average energy of the molecule and is given by

$$\langle m\rangle = Z^{-1} \sum_{m=0}^{s} m g(m) e^{(s-m)\epsilon} \tag{9}$$

Second, folding is characterized by a decrease in the average number of effectively accessible conformations; i.e., the partition function Z decreases as the compactness of the molecule increases. Third, folded molecules are characterized by a core of H residues, buried away from solvent contact. The fraction x is a measure of the degree to which a compact chain molecule has a solvophobic core

$$x = n_{hi}/n_i \tag{10}$$

where n_{hi} is the number of hydrophobic residues in the interior and n_i is the total number of interior residues (i.e., residues that are completely surrounded by other residues). The average degree to which the H residues are partitioned

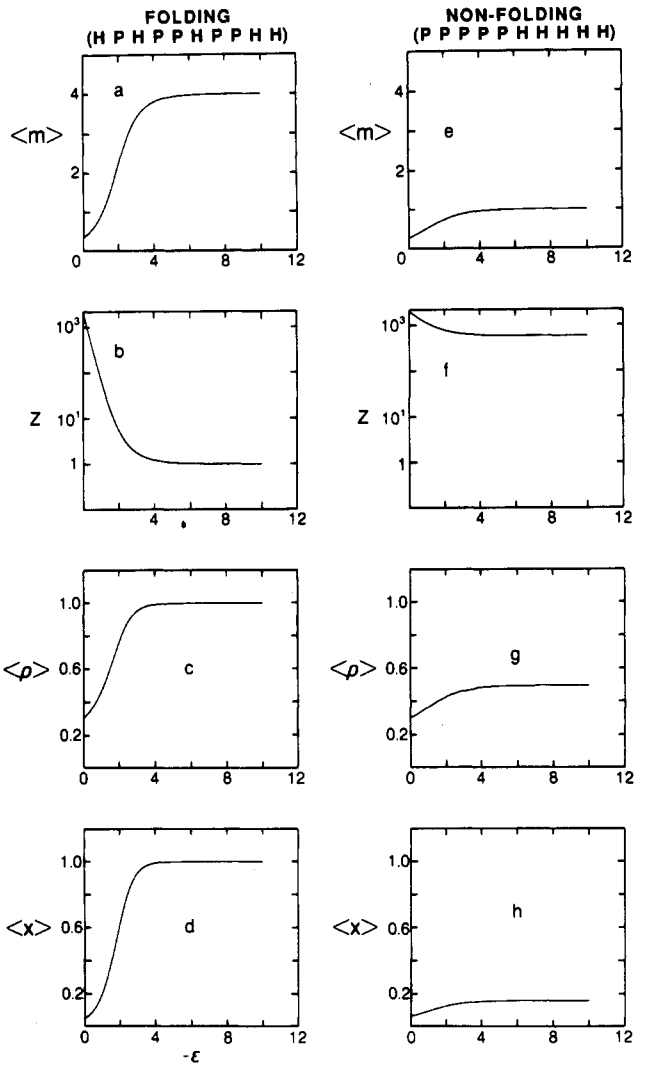

Figure 4. Ensemble averages of folding (left) (a)–(d) and nonfolding sequences (right) (e)–(h) as a function of ϵ, the energy of HH contacts: (a), (e) energy/ϵ, $\langle m \rangle$; (b), (f) number of effectively accessible states, Z; (c), (g) compactness, $\langle \rho \rangle$; (d), (h) core distribution, $\langle x \rangle$.

into a core, over the ensemble of accessible conformations, is given by

$$\langle x \rangle = Z^{-1} \sum_{i=1}^{N} x \epsilon^{(s-m_i)\epsilon} \tag{11}$$

where m_i is the number of HH contacts in the ith conformation and N is the total number of conformations in the full conformational space. We can also compute the average core distribution of only the native conformations by

$$\langle x \rangle_{ns} = \frac{1}{g(s)} \sum_{j=1}^{g(s)} x_j \tag{12}$$

The ensemble averages of these quantities are shown in Figure 4 for folding and nonfolding sequences as a function of ϵ. These properties tend to be correlated. In general a "folding" sequence tends to lead to a relatively small number (sometimes one) of low free energy compact conformations with an H core, whereas a "nonfolding" sequence tends not to have any of these attributes.

The Nature of the Native States

In this section our purpose is to explore the relationship between the amino acid sequences and the native structures of the model chains. For each sequence, we find the native state(s) of lowest free energy, as described above. We then perform this search for every one of the 1024 sequences of the sequence space of $n = 10$. We generate all possible sequences by starting with the sequence HH...H (00...0) and incrementing a binary counter by one unit with each step to the right until the sequence PP...P (11...1) is reached. Then the sequences are sorted by composition. We number the sequences in order of decreasing number of H residues, from sequence 1, which is comprised of all H residues, to sequence 1024, which is comprised of all P. The first question we have asked is, to what extent can the folding potential be predicted by composition alone, the fraction of H-type residues in the chain, irrespective of their sequence? The composition versus sequence number is shown in Figure 5; the number of sequences of each composition, the width of each step in Figure 5, is given simply by the binomial distribution. This figure is simply the composition reference for the following figures.

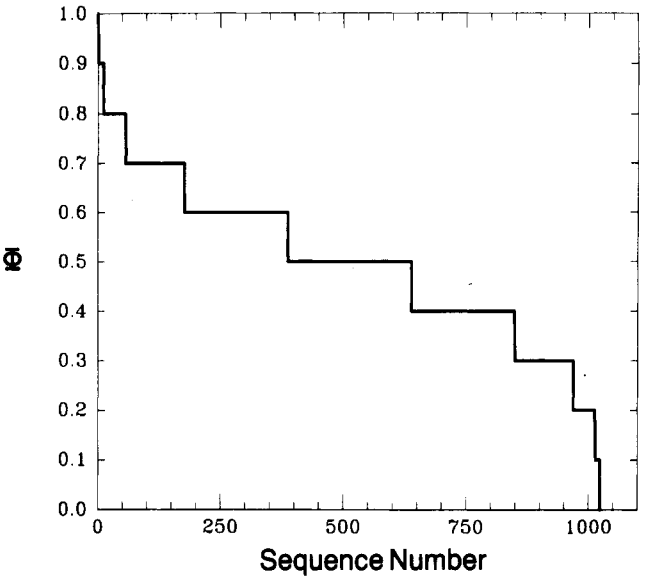

Figure 5. Composition of sequences ($n = 10$) for the purpose of interpreting Figure 6.

Figure 6 shows the properties of the native state(s) as a function of sequence. The properties shown are the number of HH topological contacts (s), which is also a measure of the native state energy, the average compactness of the native state(s) ($\langle \rho \rangle_{ns}$), the native state degeneracy ($g(s)$), and the average fraction ($\langle x \rangle_{ns}$) of the interior filled by hydrophobic residues.

The effect of the composition on the native state properties is evident from Figure 6. On average, larger Φ leads to more native states with high compactness, low energy, small degeneracy, and a highly hydrophobic core. Thus a large component of the folding potential is just due to the composition, the fraction of residues that are H. However, it is also clear that there are sequences with high H content that do not fold (downward spikes in Figure 6a,b,d and upward spikes in Figure 6c, and see below).

The effect of composition on the native properties of sequences is shown more explicitly in Figure 7. We have found that even with the same composition, there is a wide distribution of energy, compactness, degeneracy, and fraction of H residues in the core. The spread is greatest for the middle range of composition, $\Phi = 0.5$, and is the smallest at the two ends. Ensemble averages over sequences with the same composition are shown in Figure 7 as a function of composition. With increasing H content, the average compactness and H localization into the core

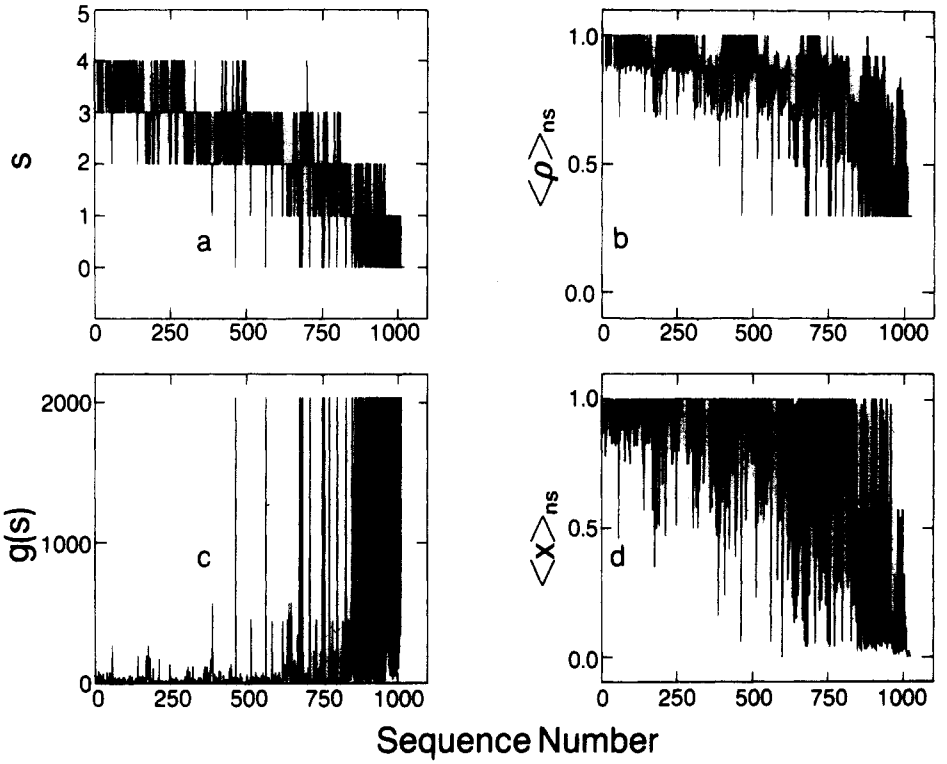

Figure 6. Native-state properties as a function of sequence number, $n = 10$: (a) free energy/ϵkT (s); (b) average compactness ($\langle\rho\rangle_{ns}$); (c) degeneracy ($g(s)$); (d) average core distribution ($\langle x\rangle_{ns}$).

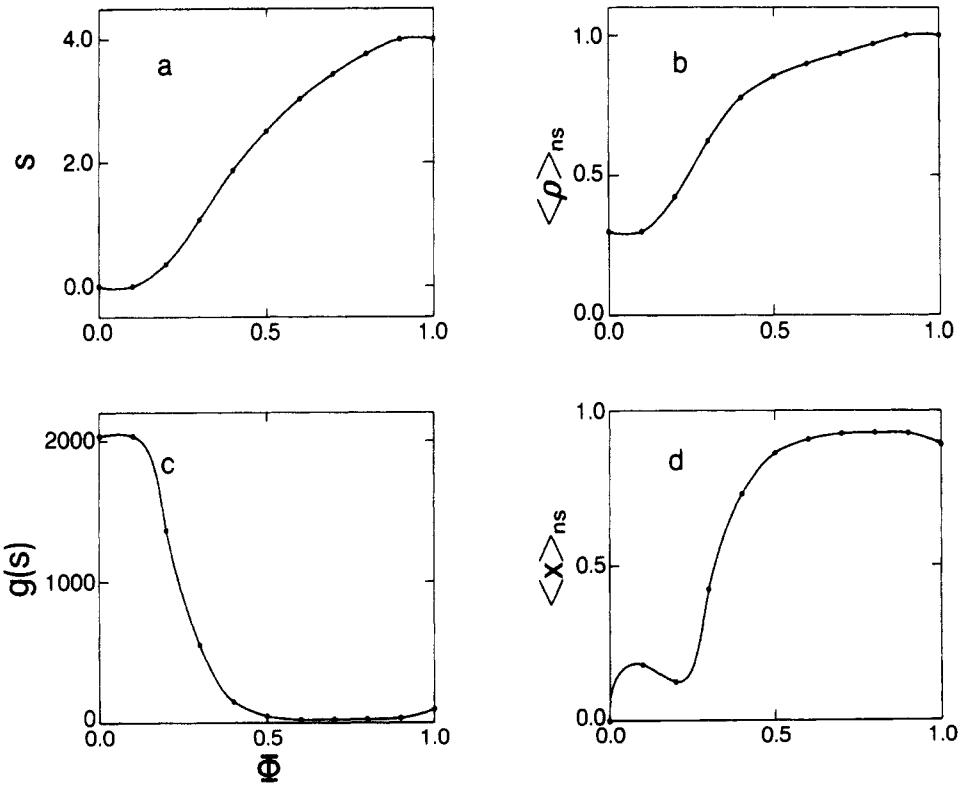

Figure 7. Native-state ensemble averages over sequences of fixed composition as a function of composition, $n = 10$.

increase, and the average energy and degeneracy decrease. At the maximum H content, the degeneracy increases slightly because there are many ways to configure a homopolymer to have low energy: for $\Phi = 1$, there are 98 conformations that have maximum density, $\rho = 1$.

The properties of the native states tend to be correlated so that sequences can be simply divided into those which are "good" and those which are "bad" in terms of their folding potential. Good sequences have native states with all the following properties: low free energy, high compactness, core of H residues, and low degeneracy. In other studies,[28,29] it is shown that these 2D lattice proteins also have significant amounts of secondary structure, lattice equivalents of helix, and parallel and antiparallel sheet topologies. All the other sequences, the bad sequences, tend not to have any of these properties.

Figure 8 shows the distribution of the native-state properties over the full sequence space. Of the 1024 se-

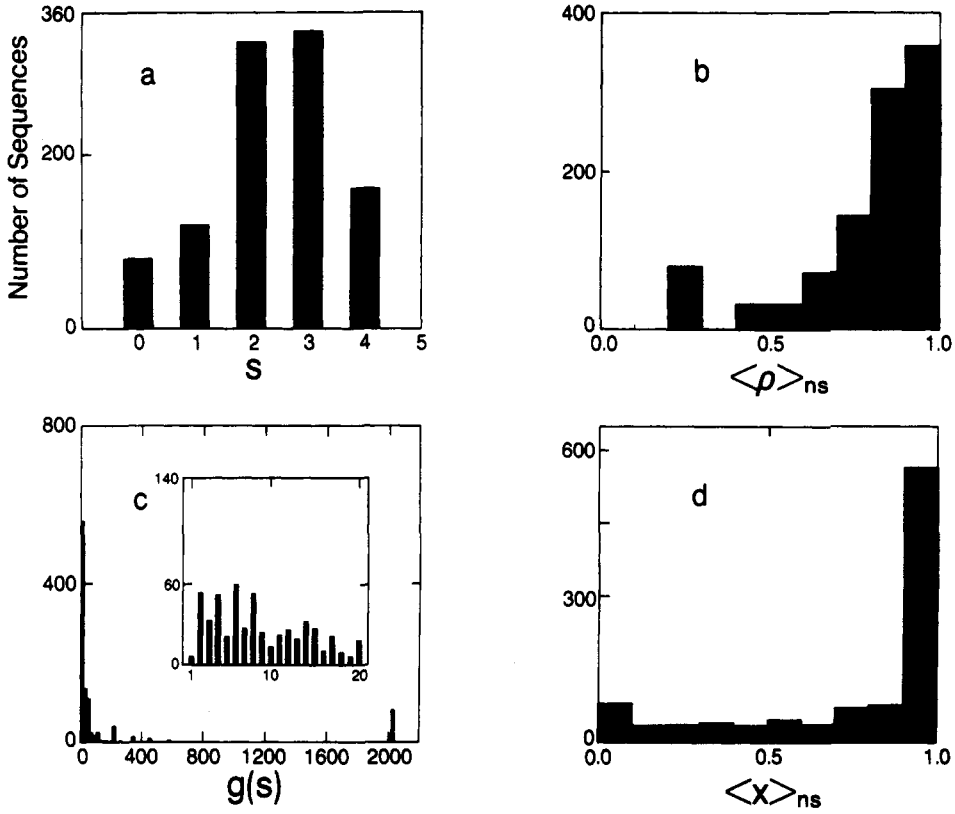

Figure 8. Distribution of the native-state properties for the full sequence space, $n = 10$.

quences for $n = 10$, the number of different sequences that can fold to maximum compactness is 259 and there are 349 sequences that are in the range $0.9 < \langle\rho\rangle_{ns} \leq 1.0$. The most probable native state for all the sequences is maximally compact with core completely occupied by H residues. Only a very few sequences have a single native state (see Figure 8c); most sequences have less than 20 native conformations. This may depend on the use of only two types of residue, H and P, in the model; if there were 20 types of residues, as in real proteins, we anticipate that the histogram of degeneracies might show a much larger population with only one or a few conformations.

Proteins whose sequences lead to only a single native structure may have significant biological advantage over sequences that lead to ambiguity of the native structure. In the model, there are six sequences that have singly degenerate native states; all the rest have more than one native state. For those sequences that lead to multiple native states, we now ask, how similar or different are the native conformations of a given sequence? To measure the differences in conformation, we must define a "metric" or distance measure. The choice of a proper distance measure is partly arbitrary; apart from a few requirements, there are no deeper principles upon which one can be chosen. The most common distance measure for comparing geometrical structures of molecules is that of the sum of squared distances taken pairwise; this is the so-called maximum likelihood best estimator, provided the errors are uncorrelated and are Gaussian-distributed with zero mean. This method is useful for comparing two very closely similar structures. However, for our purpose of comparing structures that can be very different, we are better served by a measure that treats the conformation of every bond, rather than the square of each distance, on equal footing.

If each bond pair b_i is represented by 0 (collinear) + 1 (right turn) − 1 (left turn), then a chain conformation is represented by the vector $\mathbf{v} = [b_1, b_2, ..., b_i, ... b_{n-2}]$ (see Figure 9). This has the useful feature that the mirror image conformation is represented simply by $-\mathbf{v}$. A convenient distance measure $d(\mathbf{v}_1, \mathbf{v}_2)$ between 2 conformations is

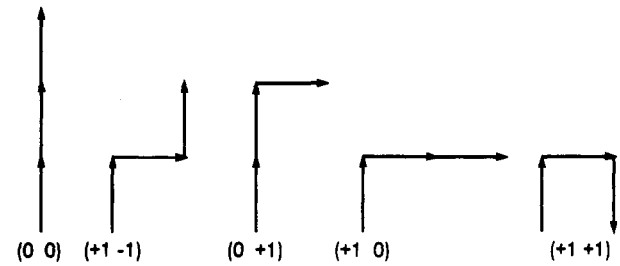

Figure 9. Distance measure of conformational similarity. Representation of bond conformations for eq 13–15.

$$d(\mathbf{v}_1, \mathbf{v}_2) = \min [c(\mathbf{v}_1 - \mathbf{v}_2), c(\mathbf{v}_1 + \mathbf{v}_2)] \quad (13)$$

where c represents the operation of summing the absolute values of the elements. In this way, conformational distance is a measure of how many bond angles are different, weighted by how different they are. Taking the minimum of the two arguments in eq 13 is a simple way of accounting for the identity of a conformation and its mirror image and would be inappropriate for molecules in 3D, where chirality is important. In that case, the appropriate measure would be

$$d(\mathbf{v}_1, \mathbf{v}_2) = c(\mathbf{v}_1 - \mathbf{v}_2) \quad (14)$$

When there are multiple native states for a given sequence, we compute their average pairwise distance

$$\langle d \rangle = \frac{2}{g(s)(g(s) - 1)} \sum_{i=1}^{g(s)} \sum_{j>i}^{g(s)} d_{ij} \quad (15)$$

This distance measure, eq 13 and 14, can be used to compare any two conformations in the conformational space,

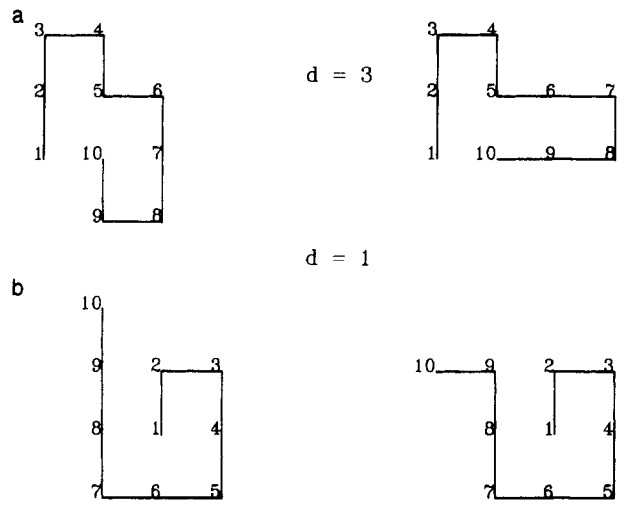

Figure 10. Examples of the distance measure of conformational dissimilarity.

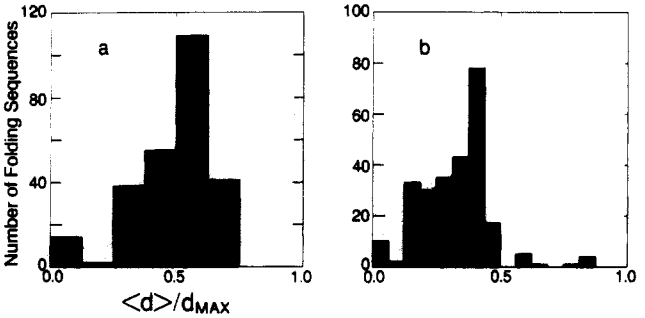

Figure 11. Conformational differences $\langle d \rangle / d_{max}$ for multiple native structures of a given sequence for all sequences with $\langle \rho \rangle_{ns} = 1$, $n = 10$. d_{max} is the maximum possible difference between two conformations: (a) d is defined in eq 13 and $d_{max} = 8$; (b) d is defined in eq 14 and $d_{max} = 16$.

provided only that they have the same chain length. The smaller the value of d, the greater the conformational similarity. This measure satisfies the requirement that the distance between identical conformations is zero. Figure 10 shows two examples of the distance measure; the top two conformations differ by a distance $d = 3$ and the bottom two by $d = 1$.

Figure 11a shows the degree of structural similarity among the different native states of the molecules which have $\langle \rho \rangle_{ns} = 1$, with d defined in eq 13. Figure 11b shows the same experiment but with d as defined in eq 14. Both distance measures lead to the same general conclusion. We find that different native states are structurally very different, on average. Pairwise comparison shows that it is most probable that two different native structures will differ in the conformations of 25–50% of the bonds. One might imagine two possible extreme views of the nature of conformational space. On the one hand, most of the lowest energy configurations could be clustered in a single small region of conformational space. The other view, supported by these model simulations, is that energy wells for the lowest energy conformations are quite broadly distributed throughout conformational space.

Since the most interesting sequences are those that can have very compact native conformations, we take a closer look at the 259 sequences that have $\langle \rho \rangle_{ns} = 1$. We have studied their composition, native state energy, degeneracy, and the average fraction of H residues in the core. The lowest possible composition that can achieve compactness is $\Phi = 0.3$. The worst native energy is 2ϵ, as compared to 0 for the whole ensemble. The maximum degeneracy of these sequences is 98, which is the number of maximally compact conformations. There is a slight tendency with decreasing Φ toward fewer HH contacts and fewer native conformations in the compact states.

One interesting prediction arises from chain end effects. Residues at the chain ends can have a maximum of three topological neighbors, whereas residues in the interior of the chain sequence can only have two. Hence it is advantageous for a sequence to have an H residue at each chain end, where, in the best of situations, it can contact three other H residues. This prediction is confirmed by the simulations. While only 25% of all possible sequences will have H residues at both ends, 90% of all folding sequences have H residues at both ends. One consequence of this is that ends tend to be found adjacent to each other more often than would be expected by chance. Of course for these short chains, end effects will be more important than for longer molecules. Similarly, it is more advantageous for an H residue to be in the interior of the globular core of the folded protein than at the surface, since it can have more topological neighbors inside. By definition, monomers at the surface must have some contacts with solvent, reducing the possible number of intrachain contacts for those residues. This difference between surface and interior residues is the driving force for the protein to form an H core in folded molecules. Although the general trends for H core formation are clear from Figure 6d, these chains are sufficiently short that many cores are not completely sequestered from contact with the solvent.

The Effects of Chain Length

Computational limitations prevent us from exhaustive exploration of both the full conformational and sequence spaces for chains of length much greater than $n = 10$, the case considered in some detail in the preceding sections. Since such chains are extremely short, it is important to know how these results generalize for longer chains. In the present section, we describe methods and results which show that the principal features of conformational and sequence spaces are relatively little changed at least up to about $n < 30$. For chains of length ≤ 20, it is possible to explore the full conformational space for a given sequence, but this can be done for only a relatively small fraction of sequence space. Therefore, in some cases below, we choose randomly a small population of sequences (200) and exhaustively explore the full conformational space for each sequence. For chains still longer, $n < 30$, we can exhaustively explore the full space of all the maximally compact conformations for a given sequence, and in these cases we also randomly sample a small subset of all possible sequences.

Before describing the effects of chain length, it is important to test whether the native states found by search of only the maximally compact conformations differ to any significant degree from those found by search of the full conformational space. Of course, if all the native states (conformations of lowest free energy) are maximally compact for a given sequence, then this procedure will produce exactly the same set of native states as the full conformational search. However, if a sequence has some native states that are not maximally compact, then this search will miss those conformations. Hence this procedure is premised on the proposition that at least most native states are maximally compact. The results of the search over only maximally compact conformations for N = 10 are shown in Figure 12. Comparison of Figures 8 and 12 shows that the distribution of energies, degeneracies, and localization of H to the core are very similar for the maximally compact and full conformational searches. No distribution of

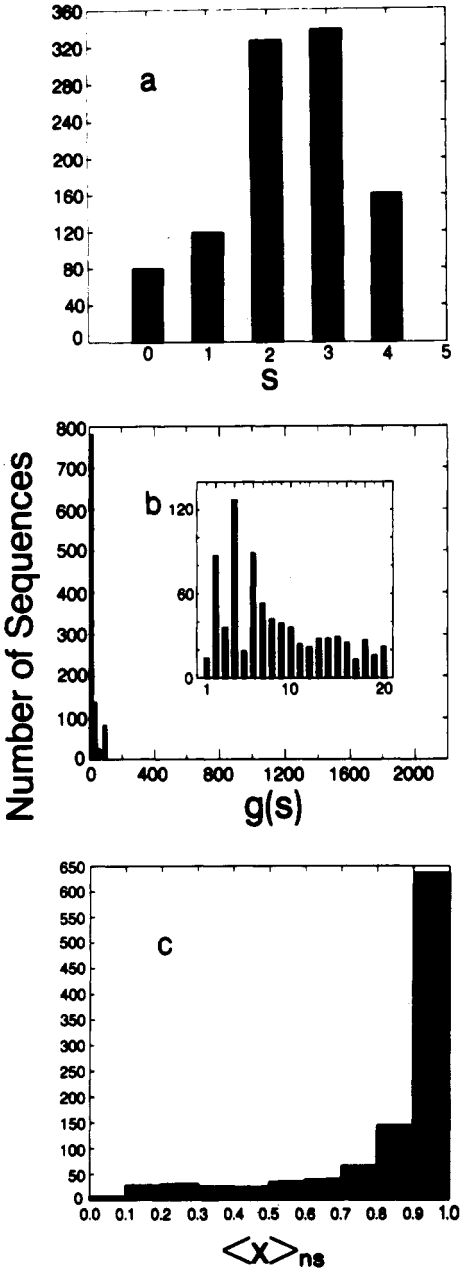

Figure 12. Distribution of the native-state properties, obtained by exhaustive searching of only the most compact conformations, for the full sequence space, $n = 10$.

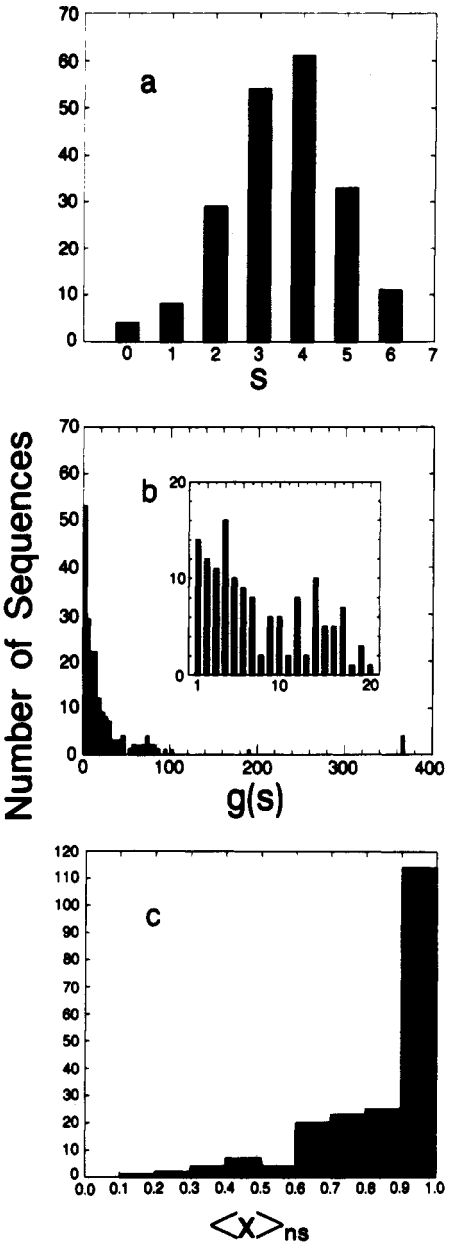

Figure 13. Same as Figure 12, but for 200 random sequences and $n = 13$.

compactness appears in Figure 12, since this is fixed, of course, by the nature of the search over only maximally compact conformations. As expected, however, the maximally compact search does not well characterize native states of sequences that are poor folders, highly degenerate, and low-density conformations. In those cases, the number of conformations is underestimated, and their density is overestimated by the maximally compact search. Even these differences diminish significantly, however, if we broaden the search from just the conformations that are maximally compact, $u + m = t_{max}$, to those that are nearly maximally compact, $u + m \geq t_{max} - 1$ or $u + m \geq t_{max} - 2$.

Figures 13–15 show the distributions of native-state energies, degeneracies, and H localization in the core, for 200 sequences each of chain lengths 13, 18, and 24, in maximally compact searches. In general, we observe that these distributions of native-state properties are not particularly sensitive to chain length. The distributions of free energy are least sensitive: the most probable native states have energies about half the maximum possible value for the chain lengths. With increasing chain length, the most probable value of $\langle x \rangle_{ns}$ decreases, since the size of the core increases, while the most probable composition remains fixed at $\Phi = 0.5$. Hence in larger molecules, a smaller fraction of the sequences will be able to fully fill a core with H residues. Perhaps most interesting is that increasing chain length tends to lead to a greater fraction of folding sequences that are singly degenerate; i.e., each one has only a single native conformation. One might speculate that increased chain length may thereby have biological advantage inasmuch as it is likely to be advantageous for a real protein to have less conformational ambiguity of native structures.

A Test of Mean-Field Model Approximations

The theory presented elsewhere for protein stability[1,27] is based on a mean-field Bragg–Williams approximation for the number of HH contacts as a function of the density and organization of compact copolymeric chain molecules. The present simulations, which are not subject to that approximation, can be used to test its validity.

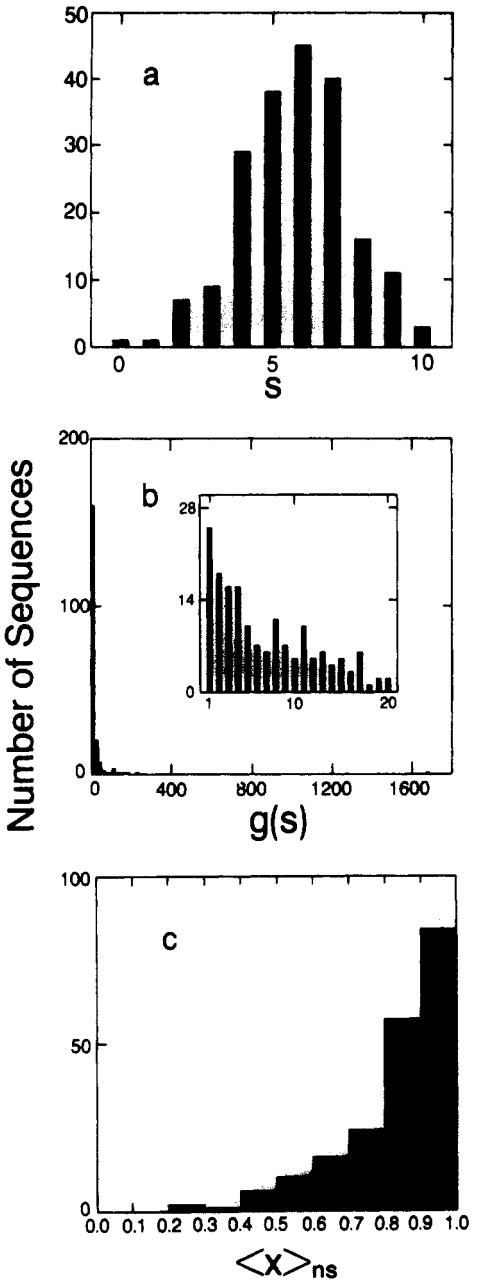

Figure 14. Same as Figure 13, but for $n = 18$.

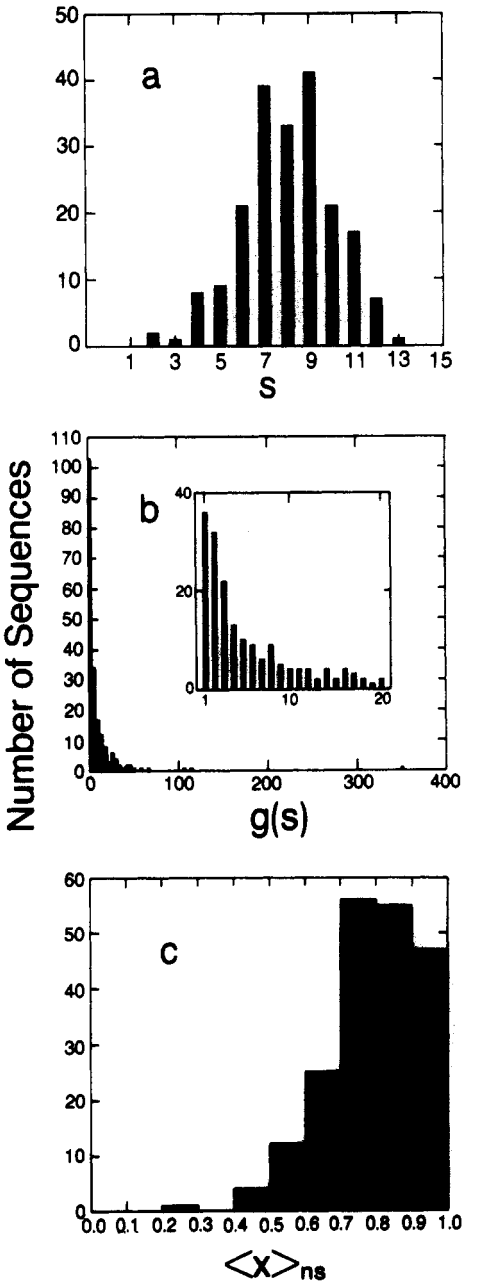

Figure 15. Same as Figure 13, but for $n = 24$.

In the mean-field model, the protein is considered to be configured into a spherical core surrounded by a concentric exterior shell, one monomer unit in thickness. For the purpose of testing here, we consider the two-dimensional equivalent protein, of a circular molecule, in which a fraction

$$f_e = \frac{2r + 1}{r^2} \tag{16}$$

of the residues occupies the exterior annular ring and the remaining fraction of residues

$$f_i = 1 - f_e \tag{17}$$

fills a core. The radius of the compact molecule is given by

$$r = (n/\pi)^{1/2} \tag{18}$$

The behavior of a protein will depend on whether there are more H residues than could fill a core, $\Phi > f_i$ (usually the case for small stable globular proteins), or whether there are fewer H residues than could fill a core, $\Phi \leq k_i$.

According to the model (see eq 27 in ref 1), when $\Phi \geq f_i$, in the limit of $\epsilon \to -\infty$, x, the fraction of the core sites that are filled by H residues, is

$$x = 1 - \delta/f_i \tag{19}$$

where $\delta \ll 1$ is

$$\delta = \frac{f_i(1 - \Phi)}{\Phi - f_i} \exp\left\{-\epsilon\left[1 - \sigma\frac{\Phi - f_i}{f_e}\right]\right\} \tag{20}$$

and where $1 - \sigma$, the fraction of surface of each exterior residue that is exposed to solvent, is taken to be equal to 0.5.[1]

In the other case, $\Phi \leq f_i$, in the limit of $\epsilon \to -\infty$

$$x = (\Phi - \delta)/f_i \tag{21}$$

where $\delta \ll 1$ is

$$\delta = \frac{f_e\Phi}{1 - \Phi - f_e} \exp\left\{-\frac{\epsilon\Phi}{f_i}\right\} \tag{22}$$

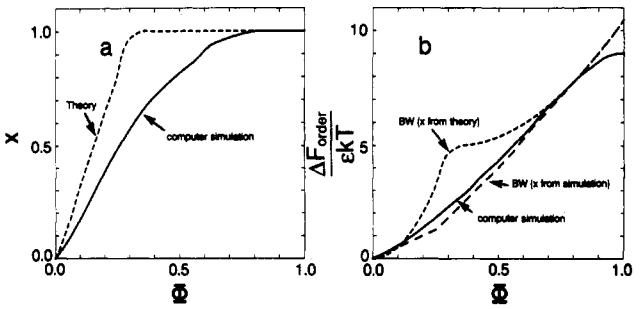

Figure 16. Comparison between simulation results and theory for $n = 16$: (a) core distribution, x; and (b) $\Delta F_{order}/\epsilon kT$.

The free energy of rearrangement or "ordering" of the H residues between interior and exterior sites in the compact molecule is given in the limit of $\epsilon \to -\infty$ through use of the Bragg–Williams approximation as

$$\frac{\Delta F_{order}}{\epsilon kT} = \left(\frac{z-2}{2}\right) n[f_i x^2 + \sigma f_e \theta^2] \qquad (23)$$

where $z = 4$ for the 2D square lattice and

$$\theta = (\Phi - f_i x)/f_e \qquad (24)$$

(ΔF_{order} is ΔF_{Π} in the earlier notation[1]).

For x and $\Delta F_{order}/\epsilon kT$ as a function of composition, the results of the simulations and the mean-field model are compared in Figure 16, for $n = 16$. It is clear from Figure 16b that if x is taken from the simulations and is subtituted into the Bragg–Williams-approximated free energy expression, eq 23, then the agreement with the simulations is very good. Hence the mean-field approximation well characterizes the native states of the ensemble of all sequences of given composition. (Deviations above $\Phi = 0.9$ can be attributed to the corners in the square-lattice model; these differences vanish for longer chain lengths and are thus unimportant.) The mean-field model, however, overestimates the freedom of the chain to distribute its residues freely between interior and exterior sites. According to the mean-field model, any residue of the chain has equal access to either region of the protein, hence increasing Φ when there are few H residues in the chains leads to the filling of the core until it is essentially full, then additional H residues locate at exterior sites. This leads to sharp changes in the slope at $\Phi = f_i$ in Figure 16a. The simulations, however, show that added H residues at low Φ are not quite so free to choose to enter the core; this is due to restrictions imposed by the chain connectivity. The comparisons made here are likely to be a worse-case analysis, since excluded volume is more restrictive of this freedom in this 2D test case than it would be if our simulations could test the mean-field model where it has been more appropriately applied, in 3D and for longer chains. Hence these differences between the mean-field model and the simulations should be expected to diminish if the simulation tests could be applied to those more realistic situations.

Acknowledgment. We thank the NIH, the DARPA University Research Initiative program, and the Pew Scholars Foundation in the Biomedical Sciences for support. We also thank Darwin Alonso and Dr. Hue Sun Chan for helpful discussions and Dr. Richard Rodgers for assistance with the Sun 3 computer system and the ⟨PLOT79⟩ plotting routines.[30]

References and Notes

(1) Dill, K. A. *Biochemistry* **1985**, *24*, 1501.
(2) Karplus, M.; McCammon, J. A. *Annu. Rev. Biochem.* **1983**, *52*, 263.
(3) Rao, N. R.; Singh, U. C.; Bash, P. A.; Kollman, K. A. *Nature* **1987**, *328*, 551.
(4) McCammon, J. A. *Science* **1987**, *238*, 486.
(5) Chou, P. Y.; Fasman, G. D. *Biochemistry* **1974**, *13*, 222.
(6) Krigbaum, W. R.; Lin, S. F. *Macromolecules* **1982**, *15*, 1135.
(7) Cohen, F. E.; Abarbanel, R. M.; Kuntz, I. D. *Biochemistry* **1983**, *22*, 4894.
(8) Cohen, F. E.; Abarbanel, R. M.; Kuntz, I. D. *Biochemistry* **1986**, *25*, 266.
(9) Levitt, M.; Warshel, A. *Nature* **1975**, *25*, 3694.
(10) Levitt, M. *J. Mol. Biol.* **1976**, *104*, 59.
(11) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *106*, 421.
(12) Kuntz, I. D.; Crippen, G. M.; Kollman, P. A.; Kimelman, D. *J. Mol. Biol.* **1976**, *106*, 983.
(13) Taketomi, H.; Ueda, Y.; Go, N. *Int. J. Pept. Protein Res.* **1975**, *7*, 445.
(14) Go, N.; Taketomi, H. *Proc. Natl. Acad. Sci. U.S.A.* **1978**, *75*, 559.
(15) Ueda, Y.; Taketomi, H.; Go, N. *Biopolymers* **1978**, *17*, 1531.
(16) Go, N.; Taketomi, H. *Int. J. Pept. Protein Res.* **1979**, *13*, 235.
(17) Go, N.; Taketomi, H. *Int. J. Pept. Protein Res.* **1979**, *13*, 447.
(18) Clore, G. M.; Bruner, A. T.; Karplus, M.; Gronenborn, A. M. *J. Mol. Biol.* **1986**, *191*, 523.
(19) Hagler, A. T.; Honig, B. *Proc. Natl. Acad. Sci. U.S.A.* **1978**, *75*, 554.
(20) Thornton, J. *Nature* **1988**, *335*, 10.
(21) Rooman, M. J.; Wodak, S. J. *Nature* **1988**, *335*, 45.
(22) Qian, N.; Sejnowski, T. J. *J. Mol. Biol.* **1988**, *202*, 865.
(23) Kolinski, A.; Skolnick, J.; Yaris, R. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 7267.
(24) Kolinski, A.; Skolnick, J.; Yaris, R. *Biopolymers* **1987**, *26*, 937.
(25) Skolnick, J.; Kolinski, A.; Yaris, R. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 5057.
(26) Li, Z.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 6611.
(27) Dill, K. A.; Alonso, D. O. V.; Hutchinson, K. *Biochemistry*, in press.
(28) Chan, H. S.; Dill, K. A. *J. Chem. Phys.* **1989**, *90*, 492.
(29) Chan, H. S.; Dill, K. A. *Macromolecules*, in press.
(30) Beebe, N. H. F.; Rodgers, R. P. C., manuscript in preparation.