Statistical Mechanics of the Combinatorial Synthesis and Analysis of Folding Macromolecules

Jeffery G. Saven*,† and Peter G. Wolynes

School of Chemical Sciences, University of Illinois at Urbana−Champaign, Urbana, Illinois 61801

Received: May 22, 1997; In Final Form: July 25, 1997[⊗]

Combinatorial chemistry techniques provide a promising route to the design of macromolecules that acquire predetermined folded conformations. A library of sequences based on a pool of different monomer types can be synthesized, where the sequences are partially designed so as to be consistent with a particular target conformation. The library is screened for folding molecules. The number of sequences grows rapidly with the length of the polymer, however, and both the experimental and computational tabulation of sequences become infeasible. For polymers and libraries of arbitrary size, we present a self-consistent, mean-field theory that can be used to estimate the number of sequences as a function of the energy in a target structure. The theory also yields the probabilities that each position in the sequence is occupied by a particular monomer type. The theory is tested using a simple lattice model of proteins, and excellent agreement between the theory and the results of exact enumerations are observed. The theory may be used to quantify particular design strategies and the facility of finding low-energy sequences for particular structures. The theory is discussed with an eye toward protein design and the mutability of particular residues in known proteins.

1. Introduction

Chemists have long been able to design and synthesize small compounds, even when the molecular mechanisms of such syntheses are not well-understood. Using such synthetic methods, researchers can craft new molecules and probe their chemical properties, thus shedding light on the mechanisms involved. For example, the physical organic chemist might make a series of compounds all based on a common theme in her search for quantifiable trends.[1] The chemist is guided by chemical theory both in her choice of molecules and in her interpretation of the experiments, experiments which in turn inform the theory. This symbiotic pairing of synthetic chemistry and chemical theory has proven fruitful in such areas as chemical bonding, conformational analysis, and reactivity. Clearly, our chemist is more likely to arrive at some sort of generalized understanding if she can access and analyze a large number of compounds. Facilitated by recent advances in managing chemical diversity, researchers can now synthesize, catalog, and assay huge libraries of molecules, where each member of the library is a different combination of a particular set of chemical building blocks. These "combinatorial chemistry" experiments are typically used to search for molecules having a desired property, i.e., the binding of a drug candidate to a receptor. By noting the numbers and common features of molecules with the desired property, researchers can also use these same experiments to reveal much about the underlying chemistry. Some researchers in protein chemistry have begun to take a similar combinatorial approach to the design of particular folded architectures. For most proteins, the polymer's three-dimensional structure is determined by its amino acid sequence. The protein chemist should be able then to design sequences that fold to a particular structure, once the physical chemistry of the folding is understood. While it is straightforward to make linear polymeric sequences, the protein chemist is troubled by a number of complications. Due to the size and complexity of proteins, many of the theoretical methods used to study small molecules, e.g., ab initio methods and complete conformational search, become infeasible. Simply synthesizing a representative number of sequences is a daunting task, since the number of possible peptide sequences is exponentially dependent on N, the number of residues. Choosing an appropriate target structure is also not without ambiguity. One recent study has found that although nearly half of all nonhomologous, globular proteins adopt one of nine families of folds, many protein folds have little similarity with any of the remaining protein structures.[2] In addition, in most proteins no single interaction dominates the folding energetics; hydrogen bonds, hydrophobic effects, van der Waals interactions, and steric packing effects all act in concert to stabilize the folded structure, but the relative magnitudes of their respective contributions are still controversial. Much has been done in the design of particular sequences, but as the size of the peptides in the library grows, the explicit tabulation of all sequences quickly becomes intractable, experimentally as well as computationally. Yet surveying the whole library can reveal trends as to what interactions are likely to destabilize as well as stabilize particular structures. We discuss in this report a workable theory that subtends all members of a given library. The theory can aid in the analysis and partial (combinatorial) design of proteins and other intramolecularly self-organizing molecules, and the theory can be used as a tool to winnow such libraries to a manageable size.

In the design of foldable protein sequences, we must first address why a particular amino acid sequence should fold to a unique structure. We consider this question from the viewpoint of the energy landscape theory of protein folding.[3−6] Amino acid sequences are heteropolymeric and typically comprise a large number of the available amino acids in an apparently random order. Upon nonspecific collapse of the random heteropolymer, incommensurate residues are likely to be brought into contact with one another. In addition, charged residues may be transiently buried in the interior of the globule, or hydrophobic residues may be exposed to solvent. For nearly all such random, collapsed states, not all of the interactions within the globule will be satisfied. The energy surface is said

† Present address: Department of Chemistry, University of Pennsylvania, Philadelphia, PA 19104.

⊗ Abstract published in *Advance ACS Abstracts,* September 15, 1997.

S1089-5647(97)01707-0 CCC: $14.00

to be frustrated, and the connectivity of the polymer chain is largely responsible for this frustration.[7] In addition, the number of collapsed conformational minima is large; it is exponential in the length of the chain. As in spin glasses, the lowest energy states for such random, frustrated systems need not be structurally similar.[8] At low temperatures the random heteropolymer may become trapped in any one of these low-energy minima. On the other hand, proteins are able to negotiate this rugged terrain and do fold reversibly to unique structures. At physiological temperatures, a protein must therefore have an additional feature to its energy landscape. The energy of the native (folded) state is sufficiently low such that it compensates for the free energy of the nonnative conformational states of the polymer. The free energy of the unfolded states is a function of the average energy of these conformations, the distribution of their energies, and their number (entropy). Due to the energetic ruggedness and the entropy of the nonfolded states, the energy of the folded structure must be much lower in energy than is typical for a collapsed, random heteropolymer of the same composition. The structure of the folded state is low in energy because it has a large number of favorable interactions, both between residues and between a particular residue type and its local environment. This energetic consistency between a protein sequence and its folded conformation has been termed the "principle of minimal frustration".[7,9]

A protein's amino acid sequence determines its structure, and at present many different protein topologies are known. Given this inherent plasticity of peptides, researchers have attempted to sculpt particular protein architectures by careful choice of the amino acid sequence. Some researchers interested in de novo protein design have focused on minimizing frustration, through hierarchical design schemes.[10] For example, in designing four helix bundles, workers have used peptide sequences that are known to form stable helixes. These helixes are then engineered to have complementary hydrophobic faces that interact upon aggregation. Appropriate linker sequences are added to connect the these helixes into one contiguous sequence. In one case, a disulfide bond connecting two helixes has been introduced to yield a conformationally defined, though small, two-helix peptide.[11] Using these methods, researchers have attempted to engineer helical bundles, α/β proteins, and β sheet proteins. In each case the global tertiary fold being designed is similar to a naturally occurring one, but recently, Dolgikh et al. have aimed for a truly novel protein structure.[12] The proteins so developed have many of the properties of natural proteins in that they are compact and have substantial amounts of secondary structure, but typically what forms is a molten globule state that has no well-defined tertiary fold.[13] Notwithstanding, these efforts appear very promising, and this de novo design of particular sequences is likely to lead to artificial proteins soon.

Alongside these experimental efforts, researchers have been tackling the issue of protein design from a computational viewpoint. Methods for scoring trial sequences are required, and researchers have developed a number of ways of quantifying which amino acid sequences might be compatible with a chosen structure. The chosen structure specifies a three-dimensional skeleton for the trial sequences. Richards and co-workers have determined what mutations are compatible with a given structure.[14,15] These authors used an atomistic approach in which they mutated side chain residues and then minimized the energy. These methods reveal much about the interactions between side chains and the effect of mutation on the packing of the individual residues. The computational demands of such methods, however, limit the number of sequences that may be considered. Other researchers have developed means of scoring trial sequences based upon the comparison of sequences that are known to fold to a particular structure in the Protein Data Bank (PDB).[16−18] Focusing primarily on backbone degrees of freedom, researchers have also considered simplified models of proteins, where reduced energy functions that involve effective residue−residue interaction energies have been used. The most commonly used energy functions are the "information-based" potentials, those that are derived statistically from the probability of occurrence of a particular residue contact pair in the PDB.[19−23] While much concern remains about the reliability, meaning, and accuracy of such potentials,[24−28] these potentials are at present the most useful reduced description for the free energies involved in folding. General issues regarding design using reduced descriptions have been addressed.[29−31] More concretely, some workers have used such potentials to verify design algorithms using both off-lattice,[32] and more commonly, lattice models of proteins.[33,34] Generally, such algorithms search for low-energy structures using a simple energy function, where mutations are randomly introduced along the chain. The search may be done by screening against a pool of known sequences[17] by using genetic algorithms[35,36] or by using Monte Carlo methods.[33,37−41] Other researchers have considered the polymerization of sequences on a lattice in the presence of an object to which the polymer binds as a route to designed ("imprinted") structures.[42] In one design scheme, the energy of the sequence is minimized at constant composition, under the assumption that the free energy of the unfolded ensemble of states is solely a function of the total numbers of each type of residue.[33] (If the composition were not constrained, a homopolymer would be the lowest energy sequence, which we know does not fold to a unique structure.) We must mention that a sequence with a low-energy in the target structure does not necessarily fold uniquely to that structure; that sequence may acquire a different conformation that is still lower in energy.[43] Researchers have addressed these issues by using alternate design criteria that take into account structures other than the target.[36−38,40] All these methods are limited, however, in that they rely upon the explicit tabulation of sequences and the subsequent evaluation of each sequence's energy in a particular structure (and in some studies nonnative structures). While much can be learned about the sequence−structure mapping,[44,45] with the exception of very simple systems, only a sparse sampling of sequence space of real proteins is possible.

Our knowledge of the principles of protein design remains incomplete, a fact which motivates the use of a combinatorial approach to protein design. By synthesizing large numbers of peptide sequences, researchers not only enhance their chance of discovering sequences that fold to a particular structure but they also stand to learn what properties foldable sequences in a library share. Combinatorial methods can reveal generalizable principles about the forces that stabilize protein structures. The design of specific sequences typically only probes portions of the sequence landscape for a chosen target structure. Given that many sequences can share a common structure, we seek to understand as large a tract of the sequence energy landscape as possible. In addition, in designing novel proteins, dictating the specific contacts between residues or the precise packing among residues may be unnecessary. This fact is underscored by the presence of peptides with a number of protein-like properties that have been isolated from random sequence libraries.[46,47] Combinatorial surveys have also been used to search for sequences that are compatible with a given structure. Kamtekar et al. have synthesized a library of peptide sequences, where a

binary patterning of hydrophobic and hydrophilic residues was chosen to be consistent with a four-helix bundle.[48] The sites on the interior (exterior) of the structure were chosen to be hydrophobic (hydrophilic). The precise amino acids at the these positions, however, was allowed to fluctuate. At each hydrophobic and hydrophilic position, five and six possible residue types, respectively, were permitted. Of a sampling of 48 sequences that were correctly expressed, 29 of these were protein-like in that they were resistant to proteases. Some of these sequences also folded to compact structures and had significant secondary structure. This partial design, based on a simple binary pattern, not only yielded molecules with protein-like properties but the experiment confirmed that "nonspecific" hydrophobic interactions, with concomitant secondary structure formation, are important determinants in specifying the folded structure.

The issue of which amino acids in a sequence are necessary to form a stable structure and which are not may be addressed by studying mutations in naturally occurring proteins. The Matthews group has synthesized many stable mutants of T4 lysozyme[49] and even a few stable double mutants.[50] The core residues of λ repressor can be very robust with respect to mutation; 70% of 125 combinations of the hydrophobic residues yield biologically active proteins. Axe et al. have found that 12 of the 13 core residues of barnase may be substituted and it still retains its biological activity,[51] thus indicating that simply maintaining a hydrophobic core is a dominant factor in the stability of the enzyme. Itzhaki et al. have found that chymotrypsin inhibitor 2 can support 100 different mutations.[52] In addition to laboratory mutagenesis, researchers have also considered the mutability of the residues in structures contained in the PDB.[53] Often, though not always, these conserved sites are crucial to the function or stability of the protein.

From the viewpoint of designing new proteins and understanding mutational variability using combinatorial libraries, the synthetic protein chemist would prefer to know a priori the number of sequences that are likely to fold to a given structure and the identities of those sequences—at least in some average sense. Given the number of possible sequences of even a moderately sized protein of 100 residues, $20^{100} \approx 10^{130}$, obtaining an understanding of even this library seems at first glance to be an insuperable task. Nonetheless, a serviceable theory for the distribution of the energies of different sequences in the target structure can guide the chemist in his design and interpretation of his experiments. Kamtekar et al. used qualitative chemical concepts to guide the design of their four-helix bundle library, but a quantitative theory has the potential to have even a stronger and more detailed predictive power.

In this paper, we present such a theory. Counting the number of sequences as a function of the energy is a task well-suited to statistical mechanics. We are further aided by the observation that the protein design problem has much in common with the Ising magnet of condensed matter theory.[33] A commonly used approach in condensed matter physics is mean-field theory, wherein we consider average local energies associated with each site that are determined by that site's local environment. Mean-field theory has seen extensive application to protein folding and the exploration of conformations,[54−57] but here we use it to quantify the characteristics of sequence space. Here the internal variables are not the conformational states of the monomers but, rather, the type of amino acid that is present at each sequence position. The theory can also accommodate restrictions imposed by the researcher, such as constraints on the number of each type of monomer and on the residue identities allowed at particular positions in the sequence. The theory provides us with a way to estimate not only the number of sequences for a given overall energy but also the probability that each sequence position is occupied by a given monomer type. The theory provides a convenient means to evaluate different combinatorial design strategies, where we can compare the effects of changing residue pattern and target structure. In this report, we present the model and compare it with results of an exactly solvable lattice polymer.

2. Theory for the Number and Composition of Sequences Compatible with a Chosen Structure

In this section we present a mean-field theory for counting the number of sequences compatible with a particular structure. The theory also yields the probability that each site is occupied by a particular type of monomer. Here we consider only the distribution of energies for different sequences in a chosen structure. Although simply finding the minimum energy sequence can be a misleading design strategy for some energy functions,[58] sequences that fold to the chosen structure must surely reside among those that are sufficiently low in energy. Thus the theory is useful in that, for a particular design strategy, it provides estimates of the number and average identities of the low energy sequences.

In our development of the theory, we neglect sampling issues. All possible sequences of a typical protein cannot be synthesized, e.g., a polypeptide of length of 100 residues has more than 10^{130} possible sequences. Thus any laboratory experiment must reflect a sampling of all possible realizations. In what follows, we assume that the space of all possible sequences is sampled uniformly, so that the probabilities presented below should mimic those seen experimentally. Since we are interested in characterizing the full sequence space, we neglect variation in the numbers of sequences due to incomplete sampling of the sequence space and non-uniform distribution of the monomers available at each sequence position that may be present in actual experiments.[48]

We outline the presentation of the theory. In section 2.1 we show how the probability that sequences are less than a certain energy may be calculated using the number of sequences as a function of the energy. Thus we need to determine a microcanonical entropy $S(E)$, which is generally an involved calculation. For one-body energy functions, however, we can write simple expressions for the entropy (section 2.2) that are valid for arbitrary (section 2.2.1) and fixed (section 2.2.2) total compositions. We review how effective one-body energies may be obtained using a simple mean-field theory in section 2.3. Last, we discuss the quantities of chemical interest that the theory yields, the most important of which are $S(E)$, the logarithm of the number of sequences, and the individual probability $w_i(\alpha)$ that a site i is occupied by a particular monomer type α.

2.1. Distribution of Energies Over Sequences. Here we let E be the energy of a sequence when it assumes the "folded" or target conformation. $\Omega_s(E)$ is the number of sequences having energy E in the chosen structure, and Ω_s is the total number of sequences. Both $\Omega_s(E)$ and Ω_s satisfy any constraints concerning the total number of each type of monomer or the allowed monomer identities at particular sites.

The *sequence entropy* $S(E)$ is the logarithm of the number of states having energy E in the folded structure.

$$S(E) = k_B \ln \Omega_s(E) \tag{1}$$

The sequence entropy $S(E)$ is defined in a way that is exactly analogous to Boltzmann's epitaphic equation for the entropy.

In calculating the number of sequences of a particular energy, we focus on estimating $S(E)$, exploiting the tools of statistical thermodynamics. For low energies, $S(E)$ an increasing function of E. As E increases, the frustration present in the folded structure increases. As the number of unfavorable interactions increases, there are more ways to distribute them, and the sequence entropy increases.

To find the probability of finding a sequence with energy less than E in the chosen target structure, we integrate the $\Omega_s(E)$ up to the chosen energy E.

$$f(E) = \int_{-\infty}^{E} dE' \, \Omega_s(E')/\Omega_s \tag{2}$$

Alternately, if the allowed energies are discrete, as they are in a lattice model, we sum over allowed energies $E' < E$.

$$f(E) = \sum_{E'=-\infty}^{E} \Omega_s(E')/\Omega_s \tag{3}$$

2.2. The Sequence Entropy for *S*(*E*) for One-Body Energy Functions. We now turn to the estimation of $S(E)$. Generally, for systems comprising many interacting sites, calculating the microcanonical entropy exactly is nontrivial. If the energy of the system is the sum of individual energies of each component, however, then equations for the entropy may be straightforwardly obtained. The ideal gas is one example. In our case, the interacting elements are the individual residues, as they occur the target structure. Effective energy functions, i.e., the profile scoring functions, have been developed by Bowie et al. that depend only on the identity and location of a residue in a particular structure.[16] Hence these types of energy functions are of a purely one-body form, for which the theory presented here is exact. For more complicated energy functions that involve many-body interactions, we can use the mean-field theory presented in section 2.3 to obtain a self-consistent effective one-body energy function. For such an energy function, the energy of a particular sequence in the chosen structure is given by

$$E = \sum_{i=1}^{N} \epsilon_i(\alpha_i) \tag{4}$$

Here N is the length of the polymer in monomer units, and ϵ_i is the effective one-body energy at site i in the structure. The sequence is denoted by an ordered list of monomer identities $\{\alpha_1 \ldots \alpha_N\}$, where α_i is the monomer type present at sequence position i. Since the conformation of the polymer is fixed, the index i labels both a particular monomer's (one-dimensional) position in the sequence and its (three-dimensional) position in the target structure. ϵ_i is a function of the monomer identity α_i present at site i.

The sequence entropy is obtained by maximizing the entropy $S(E)$ with respect to any unconstrained internal parameters. Since the energy may be written as a sum of individual one-body terms, the $S(E)$ may be expressed as

$$S(E)/k_B = -\sum_{i=1}^{N}\sum_{\alpha_i=1}^{m} w_i(\alpha_i) \ln w_i(\alpha_i) \tag{5}$$

where k_B is Boltzmann's constant, and m is the number of monomer types. In the case of peptides, m is just the number of different amino acids used in synthesizing the sequences. Here $w_i(\alpha)$ is the probability that residue type α is at position i in the structure. The sequence entropy $S(E)$ is maximized subject to the constraint that the total energy is conserved. We incorporate this constraint by restricting the value of the internal energy U such that $U = E$, where

$$U = \sum_{i=1}^{N}\sum_{\alpha_i=1}^{m} \epsilon_i(\alpha_i) w_i(\alpha_i) \tag{6}$$

An additional constraint is that the sum of the identity probabilities on each site is unity, i.e., each site must be occupied by at least one monomer

$$1 = \sum_{\alpha=1}^{m} w_i(\alpha) \tag{7}$$

If by design, monomer α is precluded from occupying site i, then we have additional constraints of the form

$$w_i(\alpha) = 0, \quad \text{if } \alpha \text{ is not allowed at site } i \tag{8}$$

There are $m - m_i$ such constraints for each sequence position i, where m_i is the number of allowed residue identities at position i.

As is done in conventional statistical thermodynamics,[59] the values of the individual identity probabilities are those that maximize the entropy, eq 5, subject to the constraints in eqs 6−8. We perform this maximization using the calculus of variations and introduce Lagrange multipliers for each of the constraints. The N multipliers that arise from the N constraints given in eq 7 are easily evaluated using the constraint conditions eq 7.[59]

2.2.1. Arbitrary Total Composition. Here we consider the case where the only restrictions on the total numbers of each type of amino acid in a sequence are those dictated by the allowed amino acids at each site (see eq 8). For such a system, after maximization of eq 5 subject to the constraints eqs 6−8, the sequence entropy may be written as

$$\frac{S(E)}{k_B} = \tilde{\beta} U + \sum_{i=1}^{N} \ln z_i \tag{9}$$

where $\tilde{\beta}$ is an unevaluated Lagrange multiplier, and

$$z_i = \sum_{\alpha_i=1}^{m} \xi_i(\alpha_i) \exp(-\tilde{\beta}\epsilon_i(\alpha_i)) \tag{10}$$

$\xi_i(\alpha_i)$ arises due to possible constraints on the allowed residue types at each sequence location (see eq 8), and it obeys

$$\xi_i(\alpha_i) = \begin{cases} 1 & \text{if residue } \alpha_i \text{ is allowed at site } i \\ 0 & \text{if not} \end{cases} \tag{11}$$

The site identity probabilities are given by a form that looks very much like a Boltzmann weight,

$$w_i(\alpha) = z_i^{-1} \xi_i(\alpha) \exp(-\tilde{\beta}\epsilon_i(\alpha)) \tag{12}$$

The Lagrange multiplier $\tilde{\beta}$ satisfies the constant energy constraint (see eq 6).

$$U = \sum_{i=1}^{N}\sum_{\alpha_i=1}^{m} \xi_i(\alpha_i) \frac{\epsilon_i(\alpha_i) \exp(-\tilde{\beta}\, \epsilon_i(\alpha_i))}{z_i} \tag{13}$$

The constant energy constraint eq 6 introduces the Lagrange

multiplier $\tilde{\beta}$. Using the analogy to statistical thermodynamics, we see that $\tilde{\beta}^{-1}$ has all the properties of an effective "temperature," the conjugate variable of the energy in thermodynamics, e.g., $k_B\tilde{\beta} = \partial S/\partial E$. As the energy increases, so does $\tilde{\beta}^{-1}$. Other researchers have referred to $\tilde{\beta}^{-1}$ as the "selection temperature."[29,31] In the language of thermodynamics, $\tilde{\beta}^{-1}$ is the temperature at which the average internal energy of the system is U. For large $\Omega_s(E)$, sequences sharing the same "effective temperature" $\tilde{\beta}^{-1}$ will have the same internal energy. For large $\tilde{\beta}^{-1}$ (high E near the maximum of $S(E)$), many different rearrangements of the monomer types are consistent with these high energies, and $S(E)$ is relatively insensitive to the precise location of each monomer type. At low effective temperatures $\tilde{\beta}^{-1}$ (low E), the requirements on the sequences having these low energies are severe, and the energy of a sequence in the chosen structure is acutely sensitive to the particular ordering of the monomer types.

We normalize the microcanonical entropy so that it yields the correct total number of possible sequences. For an arbitrary total composition, when no restrictions are placed on the number of each type of monomer,

$$\Omega_s = \prod_{i=1}^{N} m_i \tag{14}$$

Recall that m_i is the number of monomer identities allowed at position i. The sequence entropy is normalized so that

$$\int_{E_{min}}^{E_{max}} dE \exp\left(\frac{S(E)}{k_B}\right) = \Omega_s \tag{15}$$

Here E_{min} and E_{max} are the minimum and maximum allowed values for a given energy function.

2.2.2. Fixed Composition. Here we treat the case where the number of each type of monomer $n(\alpha)$ is the same for each sequence. This we refer to as the constraint of constant composition. The locations of the individual identities, however, can fluctuate and are constrained only by eq 8. Let $n(\alpha)$ be the total number of residues of monomer type α. Recall that $w_i(\alpha)$ is the probability that site i has residue type α. In applying this constraint, we see that for each sequence the sum of the individual identity probabilities over sequence positions must equal the number of monomers of that type.

$$n(\alpha) = \sum_{i=1}^{N} w_i(\alpha) \tag{16}$$

Recall that if the residue identity α is not allowed at site i, i.e., $\xi_i(\alpha) = 0$, then $w_i(\alpha) = 0$, as in eq 12.

As in section 2.2.1, we maximize the entropy $S(E)$ with respect to the constraints of constant energy (eq 6) and probability conservation (eq 7). We also include the constant composition constraints eq 16, each of which has a corresponding Lagrange multiplier we denote as the product $\tilde{\beta}\tilde{\mu}_\alpha$. The resulting microcanonical entropy is $S(E)$ is

$$S(E)/k_B = \tilde{\beta}U + \sum_{i=1}^{N} \ln z_i + \tilde{\beta}\sum_{\alpha=1}^{m} \tilde{\mu}_\alpha n(\alpha) \tag{17}$$

where

$$z_i = \sum_{\alpha=1}^{m} \xi_i(\alpha) \exp(-\tilde{\beta}(\epsilon_i(\alpha) + \tilde{\mu}_\alpha)) \tag{18}$$

The sequence identity probabilities become

$$w_i(\alpha) = z_i^{-1}\, \xi_i(\alpha) \exp(-\tilde{\beta}(\epsilon_i(\alpha) + \tilde{\mu}_\alpha)) \tag{19}$$

where z_i is given in eq 18.

Using the constant composition constraint in eq 16 we can write down a self-consistent equation that $\tilde{\mu}_\alpha$ must satisfy.

$$\tilde{\beta}\tilde{\mu}_\alpha = \ln\left[\sum_{i=1}^{N} z_i^{-1} \exp(-\tilde{\beta}\epsilon_i(\alpha))\right] - \ln n(\alpha) \tag{20}$$

We can again draw the analogy to statistical thermodynamics. As we discussed previously, $\tilde{\beta}^{-1}$ has the properties of an effective temperature. The constant composition constraints eq 16 require that the number of each monomer type is constant. The thermodynamic conjugate variables of the numbers of each component are their effective chemical potentials $\tilde{\mu}_\alpha$. The effective chemical potential $\tilde{\mu}_\alpha$ is an effective free energy per particle for monomer type α. For a specified overall composition, some monomer types may be more favorable than others on average. The effective chemical potentials readjust to maintain constant composition in such cases.

When the number of each monomer type is predetermined, the total number of sequences is given simply by the corresponding multinomial coefficient[59,60]

$$\Omega_s = \frac{N!}{\prod_{\alpha=1}^{m} n(\alpha)!} \tag{21}$$

where the reader will recall that N is the chain length in monomers, m is the total number of available residues, and $n(\alpha)$ is the number of residues of type α. Equation 21 holds when all m monomer types are allowed at each of the N sequence positions, i.e., $m_i = m$ for all i. The sequence entropy $S(E)$ in eq 17 is normalized according to eq 15, where eq 21 is used to define Ω_s.

2.3. Mean Field Treatment of the Molecular Energy Function for a Given Target Structure. In section 2.2, we presented the sequence entropy $S(E)$ and the sequence identity probabilities $w_i(\alpha)$ for arbitrary and constant compositions. The formalism applies when the energy may be written as a sum of one body terms. Though some authors have chosen to describe sequence−structure compatibility in terms of one-body functions[16] for which the theory is exact, the majority of energy functions involve two-body interactions in the form of a contact potential.[17,19,22] These are parametrized such that the probability that two residues are within a prescribed distance of one another yields an effective contact energy. Typically, such potentials are derived from the frequencies with which each type of contact occurs in a representative set of known protein structures. Some authors have also even included three-body interactions between residues.[17,61] The simple formalism presented below may be easily extended to energy functions involving arbitrary many-body interactions. For the sake of simplicity, we consider here only energy functions that involve a sum of one- and two-body terms.

The simple route to an effective one-body energy is based upon a well-known tool in condensed matter physics, that of a mean-field theory. The interactions among the components of a particular system make the system's statistical behavior nontrivial; the energy present at a particular site depends on the states, identities, and positions of the other components with which it interacts. In a simple form of mean-field theory,[62−65]

the fluctuating local energy at a particular position is replaced by the average energy of the interactions of the particle with its neighbors. The averaging is done over different realizations that are consistent with a particular set of thermodynamic conditions. We take a similar approach here. We note that many other researchers have applied the technique to biochemical problems, but usually with the intent of understanding conformational statistics.[54−57] Shakhnovich and Gutin have noted that protein design has many features in common with the random field Ising magnet,[33] and this approach motivated their design algorithm. Other physicists have developed design algorithms with this analogy in mind.[37,40] Though mean-field theory does poorly in the critical region for three-dimensional systems,[64,65] the critical region is not of interest in design problems where we are concerned primarily with configurations (sequences) of low energy or low "effective temperature." Furthermore, mean-field theory is known to improve as the dimensionality of the system increases, that is as each particle interacts with a larger number of the remaining particles. The average coordination number c_i in proteins can fluctuate and depends on the prescribed interaction radius r_c for two-body interactions, but in general it can be quite large: when $r_c = 12$ Å, as it does for some energy functions,[22,61] then $c_i \approx 20$. (In this crude example, r_c is the distance between alpha carbons, and we have used the PDB structures of myohemerythrin, myoglobin, and ribonuclease A).

The energy function most commonly applied to proteins includes one- and two-body terms. For a sequence whose site identities are specified by the configuration $\{\alpha_1 \ldots \alpha_N\}$, where α_i is the identity of site i, the energy of the sequence when it takes on the native conformation is

$$E = \sum_{i=1}^{N} \gamma_i(\alpha_i) + \sum_{i \neq j}^{N} \gamma_{ij}(\alpha_i,\alpha_j;r_{ij}) \tag{22}$$

where N is the total number of residues, $\gamma_i(\alpha_i)$ is an energy contribution due to the presence of residue type α_i at site i, and $\gamma_{ij}(\alpha_i,\alpha_j;r_{ij})$ is the interaction between sites i and j when their monomer types are α_i and α_j, respectively. Typically, γ_{ij} is a function of r_{ij} the distance of separation between sites i and j. The one-body term γ_i quantifies the propensity of monomer types to reside in a particular structural context.[16,22] For example, different amino acids will have different propensities to reside in an alpha helix or other secondary structure. Also included in γ_i are any propensities for monomers to be buried within the globule or exposed to solvent at the target structure's surface, the surface accessibility. The two-body term $\gamma_{i,j}$ can be used to quantify interresidue contact propensities,[19,20,22,23] as well as excluded volume interactions. For the sequence design problem discussed here, all the r_{ij} are determined by the chosen target conformation and do not fluctuate. Therefore we will suppress the dependence of γ_{ij} on r_{ij}.

We define $\epsilon_i(\alpha_i)$ as the local field (energy contribution) at site i when it is occupied by residue type α_i due to any local energy contribution plus that is due to its interaction with its neighbors.

$$\epsilon_i(\alpha) = \gamma_i(\alpha) + \sum_{j=1}^{N} \gamma_{ij}(\alpha,\alpha_j) \tag{23}$$

In our mean field treatment, we assume that $\epsilon_i(\alpha_i)$ can be replaced by its average value. The best choice for an effective one-body site energy is well-known and may be obtained using the Gibbs−Bogobiulov inequality[66]

$$\epsilon_i(\alpha) = \gamma_i(\alpha) + \left\langle \sum_{j=1}^{N} \gamma_{ij}(\alpha,\alpha_j) \right\rangle \tag{24}$$

The average denoted by $\langle \ldots \rangle$ is over all the possible identities of the neighboring residues at a given energy. The joint probability for a particular arrangement of monomer identities among site i's neighbors is simply the product of the individual site identity probabilities $w_j(\alpha_j)$. Since $w_j(\alpha_j)$ is normalized to unity (see eq 7), the average contribution due to the two-body term is

$$\left\langle \sum_{j=1}^{N} \gamma_{ij}(\alpha,\alpha_j) \right\rangle = \sum_{j=1}^{N} \sum_{\alpha_j=1}^{m} \gamma_{ij}(\alpha,\alpha_j) w_j(\alpha_j) \tag{25}$$

Recall that the γ_{ij} are nonzero only if residues i and j interact with one another. Therefore, the sum only has nonzero contributions from the neighbors of residue i. The resulting effective energy function is a sum of one-body terms $\epsilon_i(\alpha)$,

$$\epsilon_i(\alpha) = \gamma_i(\alpha) + \sum_{j=1}^{N} \sum_{\alpha_j=1}^{m} \gamma_{ij}(\alpha,\alpha_j) w_j(\alpha_j) \tag{26}$$

Equation 26 provides us with an expression for the effective local field in terms of the putative one- and two-body interactions γ_i and γ_{ij}. The reader will notice, however, that the $w_j(\alpha_j)$ also are functions of the $\epsilon_i(\alpha)$ (see eqs 12 and 19). Thus for a given energy function, we solve self-consistently for the weights $w_i(\alpha)$ and the effective site energies $\epsilon_i(\alpha)$. Note that, with this form of the effective site energy eq 26, the contribution due to two-body interactions in eq 6 is double counted.[54] Therefore, we must subtract one-half the value of this two-body energy contribution.

$$E = U - {}^1/_2 \sum_{i=1}^{N} \sum_{\alpha_i=1}^{m} \sum_{j=1}^{N} \sum_{\alpha_j=1}^{m} \gamma_{ij}(\alpha_i,\alpha_j) w_i(\alpha_i) w_j(\alpha_j) \tag{27}$$

where U is defined as in eq 6.

$$U = \sum_{i=1}^{N} \sum_{\alpha_i=1}^{m} w_i(\alpha_i)\, \epsilon_i(\alpha_i) \tag{28}$$

Alternately, we could have also subtracted a constant term from each of the energies $\epsilon_i(\alpha)$. Regardless, no correction is necessary for the individual site identity probabilities $w_i(\alpha)$.

Note that this development of an effective one-body energy function is similar in spirit to the quasi-chemical approach in the theory of the Ising magnet and binary alloys.[63,64,66] In the protein case considered here, however, all the sites are not equivalent. The equivalence among sites in the lattice Ising magnet simplifies the solution of the that problem, but it is the nonequivalence of the sites in the protein problem, the heterogeneity of the sequence and structure, that is paramount for the ability of the protein to fold to a unique structure.

2.4. Summary of Theory. We now summarize the implementation of the theory. In section 2.3 we presented how an effective one-body energy (eq 26) may be obtained from an energy function that involves a one-body (profile) and two-body (intermonomer contact) interactions. The effective site energies $\epsilon_i(\alpha)$ and site identity probabilities $w_i(\alpha)$ can be solved for self-consistently using eqs 26 and 12, or, if the composition is fixed, eqs 26 and 19. The microcanonical sequence entropy

$S(E)$ is then obtained using eq 9 or eq 17. In each case, the sequence entropy is normalized according to the total number of possible sequences (see eq 21). We may then use eq 2 to obtain the cumulative probability that a sequence having energy less than E in the target structure exists. In practice, we solve for the $\epsilon_i(\alpha_k)$ and $\tilde{\mu}_k$ for a given value of $\tilde{\beta}$, and then use eq 27 to determine the corresponding energy for these parameters. In thermodynamic terms, this is equivalent to choosing a temperature for the system $\tilde{\beta}^{-1}$ such that the internal energy is equal to the energy of interest. For a given target structure, the theory yields the probability $w_i(\alpha)$ that a site i in the sequence is occupied by the monomer type α at a given energy. The theory also provides the logarithm of the number of sequences $S(E)$ that have a particular energy in the target structure.

The theory may be used to calculate a quantity that reflects the mutational variability of each site. We will refer to this quantity s_i as the *local sequence entropy*.[45,53]

$$s_i = -\sum_{\alpha=1}^{m} w_i(\alpha) \ln w_i(\alpha) \tag{29}$$

where the sum is over the residue identities at site i. Depending upon the structure, some sites are more likely to have a particular identity than others. For a given value of the overall energy E, the local sequence entropy s_i is a measure of how likely mutations are at a particular residue position. Note that if only one residue type is permitted at site i, then $s_i = 0$. If all possible residue types are equally likely at site i ($w_i(\alpha) = 1/m_i$), then $s_i = \ln m_i$. Recall that m_i is the number of monomer identities allowed at site i. At a given total energy E, a site having a small value of s_i has an identity that is likely to be conserved across different sequences.

3. Results and Applications of the Model to Lattice Models of Proteins

The focus of this study is to illustrate the validity of the self-consistent mean-field theory presented in section 2. For natural protein structures, there is currently some controversy about the proper choice of energy functions,[27] and the number of available monomers (the amino acids) is large (20), implying that the number of possible sequences is gigantic (20^N). For these reasons, we initially apply the theory to a system that may be much more easily understood. We choose the well-studied 27-mer cubic lattice polymer.[34,67−70] In such lattice models, simple energy functions may be used for which some sequences exhibit "protein-like" behavior: many conformations are possible but one conformation is thermodynamically preferred. The simplicity of the model makes it easy to discuss quantities that appear in the theory in terms of particular structural features. Furthermore, if the number of allowed monomer types is sufficiently small, all possible sequences may be enumerated for any given target structure, and hence the lattice polymer provides an exactly solvable model that we can compare with the theory presented in section 2.

We choose energy functions that are of a pure two-body (contact) form. In the theory presented in section 2, no simplifying approximations are made with regard to the one-body term in a given energy function (see eq 22). In addition, most of the energy functions that have been used for the lattice models involve only interresidue contacts. Therefore, in illustrating the theory, we consider energy functions that involve only two-body interactions, i.e., all the $\gamma_i(\alpha) = 0$. In the minimalist model we consider here, there are just two types of residues, H and P, which crudely mimic hydrophobic and hydrophilic residues in proteins.[5] Each energy function that we consider is of the form

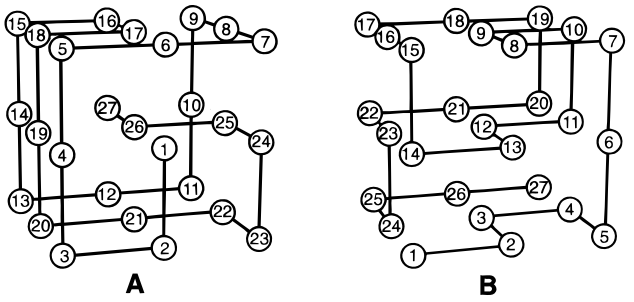

Figure 1. Compact structures of the 27-mer cubic lattice model. (**A**) The structure of Li et al. that is the conformational ground state of 3794 sequences.[70] (**B**) The ground state conformation of the 002 sequence of Socci and Onuchic.[71]

$$E = \sum_i \sum_{j>i} \gamma_{ij}(\alpha_i,\alpha_j) \tag{30}$$

where α_i, the amino acid identity, is either H or P. The γ_{ij} are nonzero only if the sequence positions i and j are nearest neighbors in the target structure and if i and j are not adjacent in the sequence ($|i - j| > 1$). In addition, the contact energies between dissimilar residues are symmetric $\gamma_{ij}(\mathrm{H,P}) = \gamma_{ij}(\mathrm{P,H})$. Thus the energy of a particular sequence is solely a function of the number of H−H, H−P, and P−P contacts that are made within the target structure. The two energy functions that we consider in sections 3.1 and 3.2 differ only in the relative strengths of these three types of contacts. In each case, the energy function is known to possess thermodynamically foldable sequences. That is, for some sequences, at sufficiently low temperatures one conformation has the dominant Boltzmann weight. Due to the discrete nature of both the energy functions and the lattice conformations, the exact total energy of each sequence may only take on specific discrete values.

We consider two compact lattice conformations of recent interest. For a specific choice of the contact energy function, a conformation recently highlighted by Li et al. (structure A in Figure 1) is the lowest energy conformation for 3794 sequences.[70] In contrast, some conformations are the lowest energy state of only a few sequences, or none at all. For a different choice of energy function, structure B in Figure 1 is the lowest energy conformation of just two sequences. One of these, the 002 sequence, has been studied extensively with respect to its folding kinetics and thermodynamics.[71,72]

In the remainder of this section, we compare the exact results with those of the theory, which takes as input only the target structure and a given energy function. The exact results are obtained by enumerating all sequences for each target conformation. With only two types of monomer, the total number of possible sequences for the 27-mer is large, $2^{27} = 134\,217\,728$, but is still easy to enumerate computationally.[67,70,71] For a given target structure, we group sequences according to their energy. In so doing, we obtain the exact $S(E)$ that may be compared with the self-consistent theory. From the explicit tabulation, we may also obtain the exact individual site identity probabilities $w_i(\alpha)$ and, using the $w_i(\alpha)$, the exact local sequence entropies s_i. We compare these three quantities with the values that are estimated by the theory.

3.1. Energy Function that Favors Buried Hydrophobic Residues. In this section we consider an energy function of the pure contact form in eq 30. Here we choose a form of the

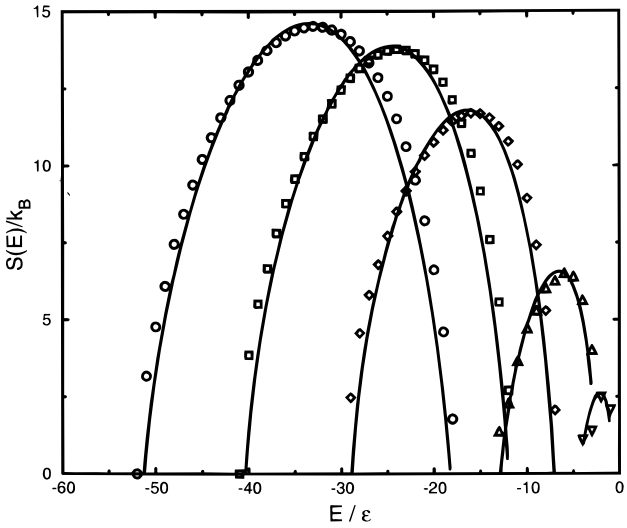

Figure 2. The sequence entropy $S(E)$ vs E for structure **A**. The energy function is that specified by eq 31. The heterogeneity of the sequences are varied by changing the ratio of polar (P) to hydrophobic (H) monomers. The following total P:H compositions were used: 14:13 (circles), 17:10 (squares), 20:7 (diamonds), 24:3 (triangles), and 26:1 (inverted triangles). The curves are the theoretical results for each composition.

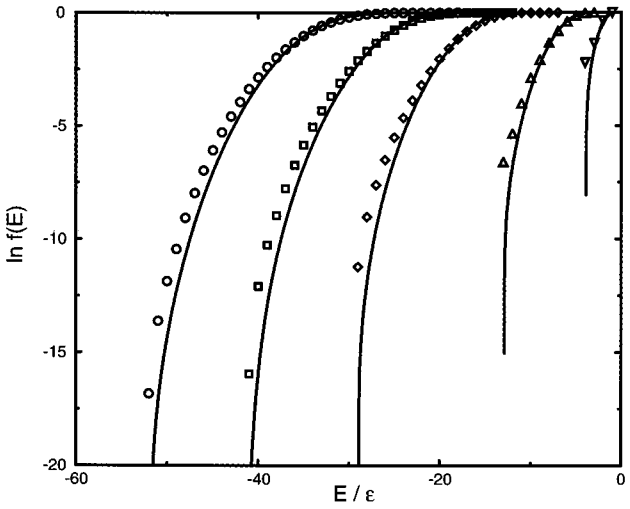

Figure 3. The fraction of sequences $f(E)$ having energy less than E for structure **A**. The energy function in eq 31 was used. The following total P:H compositions were used: 14:13 (circles), 17:10 (squares), 20:7 (diamonds), 24:3 (triangles), and 26:1 (inverted triangles). The curves are the theoretical results for each composition.

energy function that is very similar to one used recently by Li et al.[70]

$$\gamma_{ij}(\mathrm{H,H}) = -3\epsilon, \quad \gamma_{ij}(\mathrm{P,P}) = 0\epsilon, \quad \text{and } \gamma_{ij}(\mathrm{H,P}) = -\epsilon \qquad (31)$$

For this choice of the energy function, compact conformations have lower energies than extended ones. In addition, it satisfies the inequality $\gamma_{ij}(\mathrm{H,H}) < \gamma_{ij}(\mathrm{H,P}) < \gamma_{ij}(\mathrm{P,P})$, and the energy decreases for those sequences that have larger number of contacts involving H monomers. Hence, the energy function favors placing H monomers in the interior of a particular structure, since positions in the interior have the largest coordination numbers. Sequences with many H monomers in the interior are more likely to have large numbers of the lowest energy H−H contacts. Thus the energy function favors placing H monomers on the interior of a conformation, which conforms with the burial of hydrophobic residues seen in folded protein structures. This energy function also satisfies $\gamma_{ij}(\mathrm{H,H}) + \gamma_{ij}(\mathrm{P,P}) < 2\gamma_{ij}(\mathrm{H,P})$, which implies that dissimilar monomers favor segregation within the collapsed globule. In the study of Li et al., an effective H−H contact strength of $\gamma_{ij}(\mathrm{H,H}) = -2.3\epsilon$ was used, but these authors mention that their results were not sensitive to the precise value of $\gamma_{ij}(\mathrm{H,H})$, as long as the above inequalities are satisfied.

For structure **A**, we consider the number of sequences having a given energy in this target conformation. In Figure 2, we plot the sequence entropy $S(E)$ for five different overall monomer compositions. The heterogeneity of the sequences decreases as the ratio of polar to hydrophobic residues (P:H) increases. The compositions considered have the following values (P:H): 14:13, 17:10, 20:7, 24:3, and 26:1. In each case there is excellent agreement between the theory and the exact results. The predicted entropy $S(E)$ is slightly lower than the exact result for low and high energies, but it covers the same range of energies as the exact result with remarkable fidelity. The maximum value of $S(E)$ decreases with increasing homogeneity, since there are fewer sequences for the more homogeneous compositions (see eq 21). Note that, as the sequences gain increasing P content and become more like a homopolymer, the energies at the maxima of these distributions increases. With increasing P content, the number of stabilizing H−P and H−H contacts decreases. Furthermore, as the heterogeneity of the sequences decreases, the width of each distribution also decreases. The most heteropolymeric composition (14:13) spans the largest range of energy values. The corresponding P-homopolymer distribution (27:0) has zero width at $E = 0$ (not shown). In Figure 3 we present the cumulative probability $f(E)$ for each of the five chosen compositions (see eq 3). Since $f(E)$ is simply the integral of $S(E)$, close agreement with the exact result is again obtained. Note that at low energies, $S(E)$ drops rapidly as the energy decreases.

It is this low-energy region that is most likely to contain sequences that fold to the target structure. The identity probabilities $w_i(\alpha)$ reveal which residue types at each position are responsible for stabilizing these low energy sequences. In Figure 4, we plot the probability $w_i(\mathrm{P})$ that each site of conformation **A** is occupied by a polar residue for different energies. The P:H composition in this case is 14:13. By comparing with Figure 2, we see that the numbers of sequences at each energy are $\Omega_s(E = -50\epsilon) = 118$, $\Omega_s(E = -45\epsilon) = 27\,395$, and $\Omega_s(E = -45\epsilon) = 464\,062$. Note that at each of the energies, there is excellent agreement between the theory and the exact tabulation. At the lowest energy considered $E = -50\epsilon$, many of the sites are predominantly occupied by one type of residue; $w_i(\mathrm{P}) \approx 1$ or 0. For the given composition, there are severe restrictions on the allowed residues at each position if a sequence is to have this low energy in the target structure. As the energy is increased, these restrictions are relaxed, and a few high-energy contacts are present. The site identity probabilities take on values intermediate between 0 and 1. At the energy that maximizes $S(E)$ nearly all residue types are allowed at each position, subject only to the constraint of constant composition: $w_i(\mathrm{P}) \approx 14/27 \approx 0.52$ (not shown). On the low-energy side of $S(E)$, the residues that are most likely to remain hydrophobic with increasing energy are those in the "center" of this conformation, residues 1, 26, and 27 (see Figure 1). Each of these residues is coordinated with four other nonbonded neighbors, a coordination number that is larger than that of any of the other residues. For low-energy sequences, their high coordination numbers imply that these residues are likely to be hydrophobic (H). Hence $w_i(\mathrm{P})$ should be small. The energy function favors hydrophobes at sites of high coordination, which are likely to be buried within the structure.

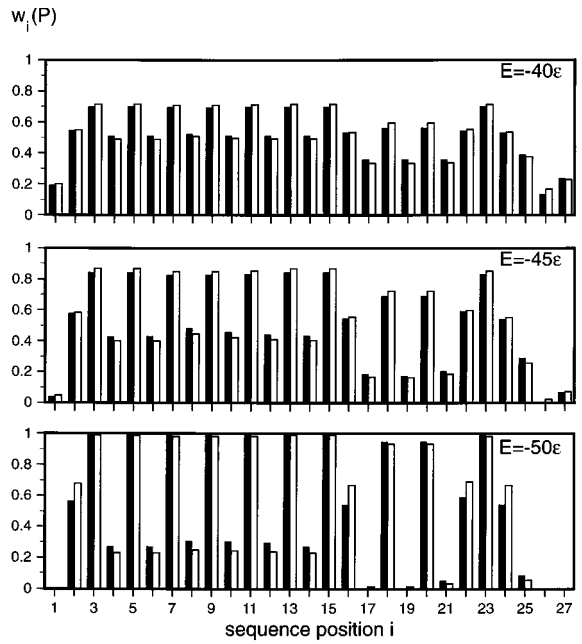

Figure 4. The sequence identity probability w_i(P) vs the sequence position i for the 14:13 (P:H) composition. Here structure **A** was used, and the energy function is that in eq 31. Shown are the results of exact tabulation (black) and the theory (white) for increasing energy, where the energies considered are on the low-energy side of $S(E)$ (see Figure 2).

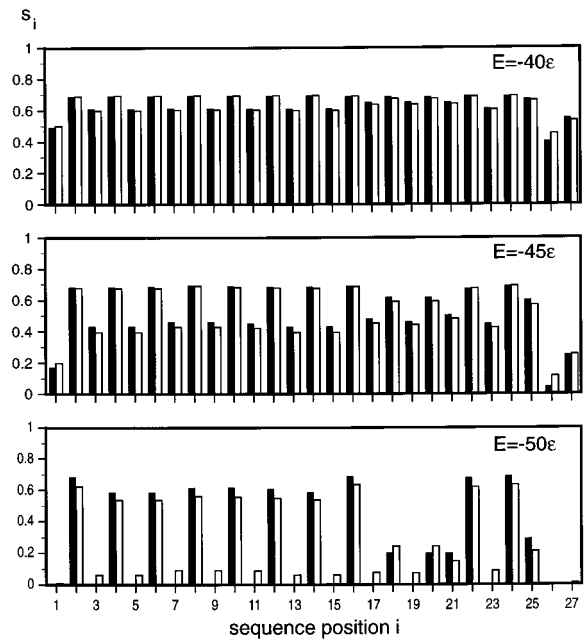

Figure 5. The local sequence entropy s_i vs the sequence position i for the 14:13 (P:H) composition. Here structure **A** was used, and the energy function is that in eq 31. Shown are the results of exact tabulation (black) and the theory (white) for increasing energy, where the energies considered are on the low-energy side of $S(E)$ (see Figure 2).

As the energy is increased, the probability that each of these residues is polar increases. At higher energies, polar residues can be accommodated at these positions. Residue 26, the one at the center of the cube, is the one that most retains its probability to be hydrophobic as the temperature is increased. Changing this residue from H to P will have the greatest impact on the energy of a particular sequence in this structure.

For models with only two monomer types, w_i(P) discloses which positions are conserved with increasing energy. For the general case of more than two types of monomers, however, s_i reveals which positions are highly conserved and which are more promiscuous with regard to their chemical identities. In Figure 5, we plot the local sequence entropy s_i for the 14:13 composition. For this simple case of two monomer types, there is no information contained in this plot that is not also present in Figure 4, since w_i(P) is not independent of w_i(H) (see eq 7). Since the theory does an excellent job of estimating the site identity probabilities w_i(P), from which s_i is calculated (see eq 29), the agreement of the theory with the exact results is again excellent. As would be expected for this energy function, the sequence position in the center of the cube $i = 26$ is the one with the lowest local sequence entropy, e.g., $s_{26} = 0.45$ at the energy $E = -40\epsilon$. Since the energy function (eq 31) "buries" hydrophobic residues, it is expected that the center residue should be H. There are other sites, however, that are conserved, i.e., that have low values of s_i, that we would not predict upon cursory inspection of the structure. The constraints of a low-energy and a given overall composition act in concert to dictate the identities of all residue positions except residues 2, 4, 6, 8, 10, 12, 14, 16, 22, and 24; each of these residues resides in the center of one of the edges of the cube and is coordinated to two other residues. Each of these positions is almost equally likely to be either H or P, as is indicated by their values of s_i which are very close to the maximum value of $s_i = \ln 2 \approx 0.693$. Those positions at the corners of the cube (3, 5, 7, 9, 11, 13, and 15), which each have but one neighbor, are constrained in their indentities at the lowest energy $E = -50\epsilon$. Their low coordination makes them the most likely places to locate the 14 polar residues that must be present for the given 14:13 overall composition. The identities of these two sets of residues are confirmed by the site indentity probabilities in Figure 4. As the energy is increased, the positions with the lowest values of s_i are those in the center (1, 26, and 27), which have the highest degree of coordination.

For other degrees of heterogeneity, we have determined both the identity probabilities w_i(P) and the local sequence entropies s_i (not shown). Similar trends are observed as for the 14:13 composition, and the theory is in very good agreement with the results of the exact tabulation.

For the energy function eq 31, structure **A** is a unique conformational minimum for 3794 sequences. Among this group of sequences, Li et al. calculated the probability that each sequence position is polar (P).[70] These authors placed no restriction on the total composition or conformational energy in their search for unique ground states. Nonetheless, we compare w_i(P) from that study with an identity probability calculated using the theory, where the composition and energy are fixed. Here we have chosen a single overall composition (14:13). The energy is $E = -47\epsilon$, which is on the low energy side of $S(E)$. At this energy, the theory predicts $\exp(S(E)) = 2265$ sequences have this energy (see Figure 2). The values of w_i(P) for these sequences, along with the results of Li et al., are presented in Figure 6. The theoretical values of w_i(P) for the 14:13 composition at these low energies has all the same trends as that seen among the sequences that are known to share structure **A** as their nondegenerate ground states. Sequence positions 1, 17, 19, 21, 25, 26, and 27 are those that have the smallest values of w_i(P) and hence are most likely to be hydrophobic (H). Sequence positions 3, 5, 7, 9, 11, 13, 15, 18, 20, and 23 are those most likely to be polar (P), since most of these residues lie at the corners of the cube. Similar trends concerning the identities of these squence positions are seen for other total compositions, where at each composition only the lowest energies of each distribution are considered. At each composition, sequences that are simply low in energy conform to the same identity patterning seen among those sequences that

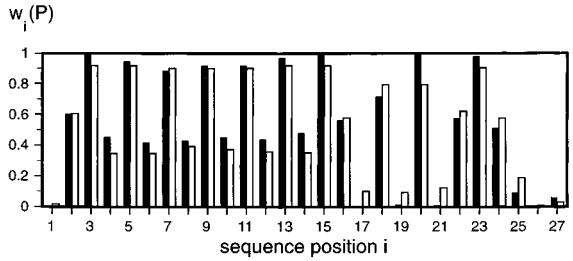

Figure 6. The sequence identity probability w_i(P) vs the sequence position i. Shown are the exact enumeration results for w_i(P) as calculated by Li et al.[70] over sequences for which structure **A** is a unique conformational ground state (black). Using this same conformation (structure **A**) and a similar energy function (eq 31), the theoretical results (white) for the 14:13 (P:H) composition at $E/\epsilon = -47$ are also shown.

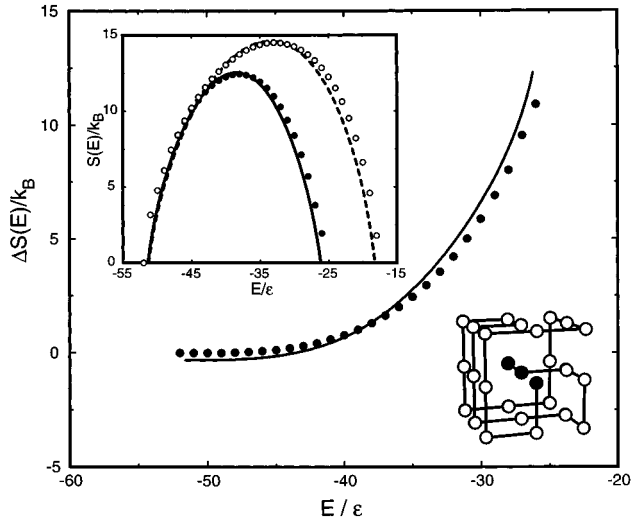

Figure 7. The exact (solid circles) and theoretical (solid curve) results for the difference of the sequence entropies $\Delta S(E) = S(E) - S_p(E)$. Shown in the inset, are the exact (filled circles) and theoretical (solid curve) results for the sequence entropy $S_p(E)$. Also shown in the inset are the exact (open circles) and theoretical (dashed curve) results for $S(E)$ in the absence of the patterning constraints at positions 1, 26, and 27. In these plots, the total composition is held fixed, where the P:H monomer ratio is 14:13. For S_p, the identities of sequence positions 1, 26, and 27 in structure **A** are constrained to be hydrophobic (H). Residues 1, 26, and 27 are black in the structure shown. The energy function in each case is that of eq 31.

are known to share structure **A** as their unique ground state. These observations emphasize the notion that, in the context of combinatorial design, sequences of fixed total composition in the low energy part of $S(E)$ are those most likely to behave like proteins.

For the 14:13 compostion, the conservation of hydrophobicity at postions 1, 26, and 27 suggests a simple design strategy: make only sequences that have H at these positions, and the fraction of sequences having low energy should increase. This approach is similar in practice to the binary patterning based upon a four-helix bundle topology used by the Hecht group.[48] The theory may be used to assay the viability of such design strategies that involve the patterning of particular residue types. We can constrain the indentities of these residues using the identity constraints as in eq 8. In implementing this design strategy, we simply introduce the following constraints:

$$w_1(\mathrm{P}) = w_{26}(\mathrm{P}) = w_{27}(\mathrm{P}) = 0 \qquad (32)$$

We use the theory to estimate the number of conformations and the site identity probabilities. In Figure 7 we compare the exact and theoretical results for $S(E)$. Since low-energy sequences contain hydrophobic residues at these positions anyway (see Figure 4), there is little change in the number of low energy sequences as a result of these constraints. What does change however is the number of high energy sequences ($E > -40\epsilon$), in which there is a great diminution. Hence, for this design strategy, not only is the total number of sequences reduced but a larger fraction of this total have low energies in the target structure ($f(E = -40\epsilon) = 0.34$) than that in the case where the identities of these residues are unconstrained ($f(E = -40\epsilon) = 0.058$). There is excellent agreement between the theory and the exact result when we consider the difference in the two entropies in Figure 7, and it is objects of this type that we are most likely to calculate in comparing two different design strategies.

Constraining monomer identities can also reveal correlations among the identities at the various sequence positions. In Figure 8, the same residues are held fixed (1, 26, and 27) as in Figure 7, but here the individual site probabilities w_i(P) are plotted for three different energies. By way of comparison, we also present the identity probabilities for the case where there is no identity patterning. Note that many of the residues are insensitive to restrictions on the identities of positions 1, 26, and 27, especially at the lowest energies $E = -45\epsilon$ and $E = -50\epsilon$. At the highest energy considered $E = -40\epsilon$, w_i(P) decreases slightly at positions 3, 5, 7, 9, 11, 13, 15, and 23−those positions with the least degree of coordination. On the other hand, the remaining unconstrained sites in the sequence are all more likely to be polar than in the case where no identity constraints are imposed; this is especially so for positions 17, 19, 21, and 25, positions at the centers of four of the faces of the cube. Recall that in the unconstrained case, these residues were very likely to be H at low energies (see Figure 4). The increased values of w_i(P) for these residues is most likely due to the constant composition constraint. Since the number of sites available to P monomers has decreased overall, w_i(P) increases at these four sites. In each case the theory has all the same trends as is seen in the exact results.

In addition to its use in evaluating design strategies, the theory may also be used to estimate the number of sequences that possess low energies in a chosen target structure. Structures that are low in energy for a large number of sequences may be highly "designable," in that they may be the thermodynamic ground state of many sequences. The theory may be used to assess the "designability" of particular structures. We must note, however, that it is well-known that simply minimizing the energy by varying a sequence's identity does not necessarily guarantee that the sequence will fold to that structure.[58] The low-energy regime of $S(E)$ is likely to be the richest in foldable sequences, however, and the theory presented here can identify such regions. Using the energy function in eq 31, we compare $S(E)$ for the two structures A and B (see Figure 1) in Figure 9. Note that structure **A** has more low energy sequences than structure B. This is seen more clearly by looking at the difference in the two entropies in Figure 9. As seen in Figure 7, the theory does an excellent job of estimating sequence entropy differences. The larger number of low-energy sequences means that there is likely to be a larger number sequences that fold to this structure. It has been suggested that a conformation containing substructural elements that are easily stabilized should be the folded state of many sequences.[73,74] With its core of highly coordinated sites, structure **A** is an example of such a conformation that is "easily optimized".[44] The number of foldable sequences, however, need not be proportional to the number of low-energy sequences. The free energy of the remaining conformations F_u of each sequence is

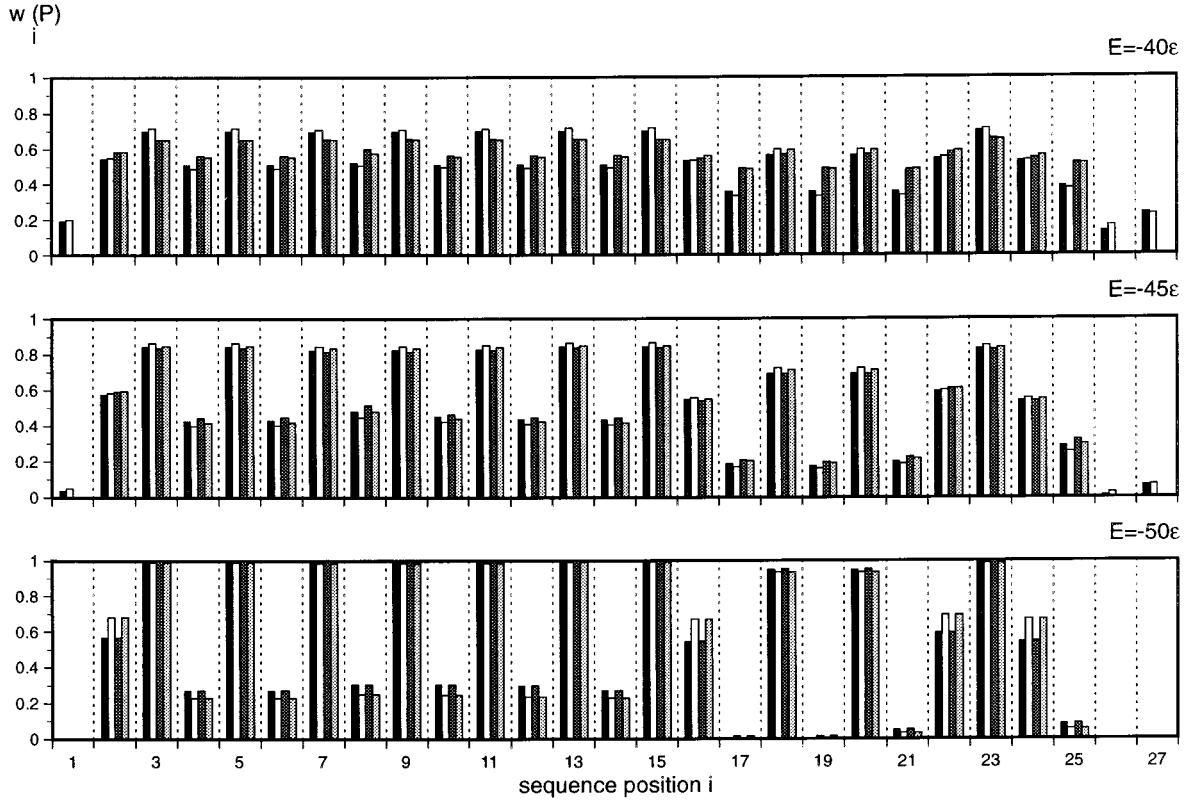

Figure 8. The probability w_i(P) that each site is polar (P) vs the sequence position *i*. The structure **A** is used (see Figure 1), and the energy function is that of eq 31. The total composition is held fixed, where the P:H monomer ratio is 14:13. The identities of sequence positions 1, 26, and 27 are constrained to be hydrophobic (H). The exact (dark gray) and theoretical (light gray) values, respectively. For comparison, the exact (black) and theoretical (white) values for the probabilities w_i(P) in the absence of such identity patterning constraints are also shown.

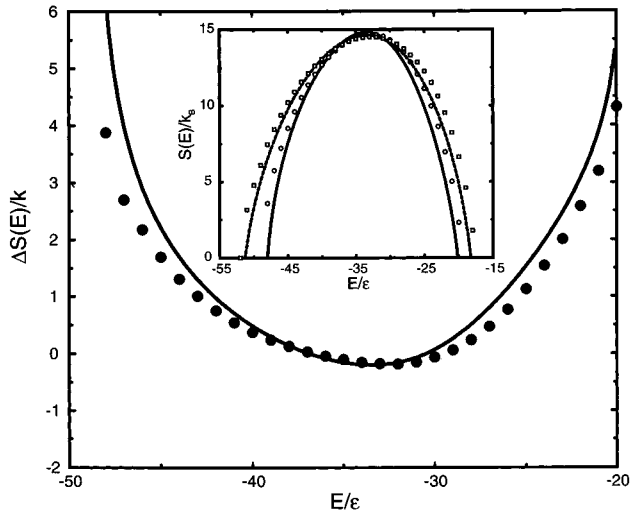

Figure 9. The exact (solid circles) and theoretical (solid curve) results for the difference of the sequence entropies $\Delta S(E) = S_A(E) - S_B(E)$ of structures **A** and **B**. Shown in the inset are the sequence entropies $S(E)$ vs E for the structures **A** and **B** (see Figure 1). Both the exact (structure **A**, squares; structure **B**, circles) and theoretical (structure **A**, dashed; structure **B**, solid curve) results are shown. The energy function is that of eq 31. The total composition is held fixed, where the P:H monomer ratio is 14:13.

as yet unspecified. Some of the low-energy sequences may not have an energy below F_u, may have other structures energetically degenerate with the target structure, or may assume structures lower in energy than the target. All this notwithstanding, it seems likely that having a greater pool of low-energy sequences from which to choose should increase the probability of finding sequences that fold to the target conformation. In light of the results for the 14:13 composition, the theory is able to quantitatively account for the large number of low-energy states of structure A (for the energy function eq 31), and hence provide insight into why structure A may be so "designable."

Last, we consider the similarity between different sequences at a particular energy. For the simple binary model considered here, we define a sequence similarity measure q_s of two different sequences $\alpha = \{\alpha_1 \ldots \alpha_N\}$ and $\alpha' = \{\alpha'_1 \ldots \alpha'_N\}$

$$q_s(\alpha,\alpha') = \sum_{i=1}^{N} \sigma(\alpha_i)\sigma(\alpha'_i) \tag{33}$$

where σ are the binary (spin variable) representations of the two residue identities

$$\sigma(\alpha) = \begin{cases} 1 & \text{for } \alpha = \text{P} \\ -1 & \text{for } \alpha = \text{H} \end{cases} \tag{34}$$

If two sequences are identical, then $q_s = N$; if their sequence indentities are perfectly anticorrelated, $q_s = N$. At each value of the energy E, there is distribution $P_s(q_s)$ that quantifies how likely each value of the overlap q_s is. We have used the exact tabulation to calculate $P_s(q_s)$ (see Figure 10). Note that the distribution of overlaps is peaked at large values of q_s for a low energy $E = -51\epsilon$. As the energy increases, $P_s(q_s)$ broadens and shifts toward $q_s = 0$. At higher energies, the pool of available sequences becomes increasingly diverse and the average similarity between any two sequences decreases. Note that $P_s(q_s)$ shifts and broadens smoothly with increasing energy. The sequence energy landscape for this structure and energy function appear "smooth" or "funnel-like"[44,68,75] in that the distribution of sequence similarities contains a single maximum that shifts smoothly toward increasing similarity with decreasing energy.

3.2. Energy Function where γ_{ij}(P,P) = γ_{ij}(H,H). We now consider a different form of the energy function, but one for

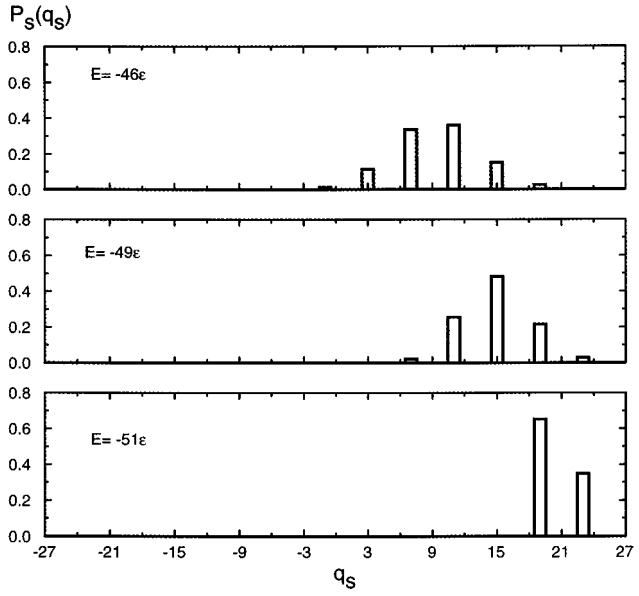

Figure 10. The sequence overlap probability $P_s(q_s)$ as a function of the sequence overlap q_s. $P_s(q_s)$ is plotted for increasing values of the energy that sequences have when they assume the structure **A**. The energies considered are on the low energy side of $S(E)$ (see Figure 2). The energy function is that of eq 31.

which foldable sequences are known.[71] The energy function is of the contact form in eq 30, where the $\gamma_{ij}(\alpha_i,\alpha_j)$ obey

$$\gamma_{ij}(\mathrm{H,H}) = -3\epsilon, \quad \gamma_{ij}(\mathrm{P,P}) = -3\epsilon, \quad \text{and } \gamma_{ij}(\mathrm{H,P}) = -\epsilon \tag{35}$$

Recall that in each case i and j must be nonbonded nearest neighbors, for $\gamma_{ij}(\alpha_i,\alpha_j)$ is 0 otherwise. As in the previous energy function, the energy is measured in units of the H−P contact strength ϵ. Energy functions of this form favor collapsed structures and are known to possess foldable sequences.[71] There are two sequences for which the structure **B** in Figure 1 has no high-energy (H−P) contacts.[71,72] A symmetry is present between the H−H and P−P contact energies, and this symmetry means that there may be two very dissimilar low-energy sets of sequences, a situation that is analogous to the low-energy spin up and spin down states seen in the absence of a magnetic field in Ising magnets.[64] While much can be learned about the kinetics and thermodynamics of particular sequences using such an energy function, the symmetry of the energy terms does seem unrealistic from a chemical perspective. For a given conformation of a polypeptide sequence, it seems unlikely that for that same conformation, a sequence where the hydrophobic and hydrophilic residues are exchanged will have a similar energy. Nonetheless, we briefly present the application of the theory using such an energy function.

The target structure in this case is structure **B** in Figure 1. In Figure 11, $S(E)$ is plotted for three different compositions, where the ratio (P:H) of P to H residues takes on the values 14:13, 20:7, and 24:3. In each case, the theory is close to the exact tabulation results, although the agreement is not as close as seen in Figure 2. For different compositions, the $S(E)$ all approximately coincide at very low energies. This convergence is due to the symmetry of the P−P and H−H interactions. For a high degree of heterogeneity (14:13), there are few structures that minimize the number of high energy (H−P) contacts. Most sequences will have many H−P contacts in the target conformation. As the sequence compositions become more homopolymeric, the number of high-energy (H−P) contacts decreases. The range of energies these more homogeneous sequences may

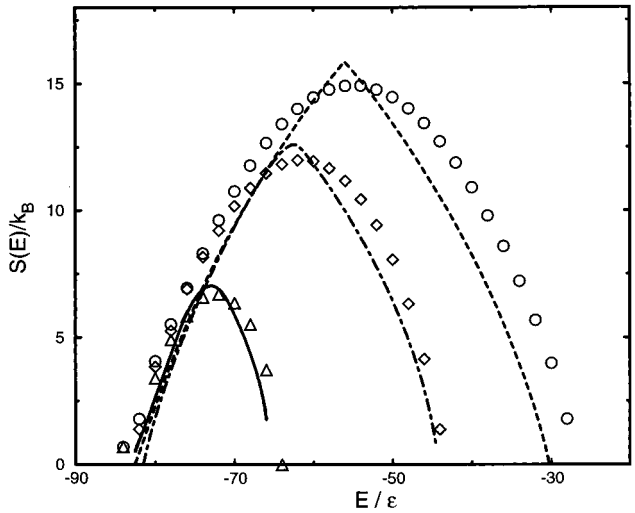

Figure 11. The sequence entropy $S(E)$ vs E for structure **B**. The energy function in eq 35 was used. Three different values of the overall heterogeneity, as specified by the P:H ratio, are presented (exact, theory): 14:13 (circles, dashed curve), 20:7 (diamonds, dot−dashed curve), and 24:3 (triangles, solid curve).

acquire narrows and shifts to lower energy. This does not imply, of course, that those sequences that are nearly homopolymers are more likely to fold to the target conformation, for the free energy of the unfolded conformations F_u also shifts to lower energy with increasing homogeneity, and the free energy difference $E_f - F_u$ decreases. For this energy function, there is better agreement between theory and the exact results as the sequences become more homogeneous. With increasing homogeneity of the overall composition, the mean field approximation becomes more accurate.

Sequence position $i = 12$ is at the center of the cube for structure B in Figure 1. Due to this residue's high coordination, we expect the identities of the monomers throughout the sequence to be most sensitive to the identity of this central residue. In Figure 12 we constrain the identity of sequence position 12 to be either H or P and plot the identity probability $w_i(\mathrm{P})$ for sets of sequences on the low-energy side of $S(E)$ for the 14:13 composition. The theory again shows good agreement with the exact results. The relative error is greater than that seen in Figure 4, but the theory accurately accounts for which residue type is most likely at each position. Overall, in each case, there is a mirrorlike symmetry in the identity probabilities, i.e., if $w_i(\mathrm{P})$ is large for $\alpha_{12} = \mathrm{H}$, then $w_i(\mathrm{P})$ is small for $\alpha_{12} = \mathrm{P}$ and vice versa. The symmetry between H−H and P−P interactions results in low-energy sequences that have little similarity. In the absence of the constant composition constraint, a given sequence could be "inverted" by exchanging the identities of the P and H residues, and the energy of the resulting sequence in the target structure would remain unchanged.

Using the exact tabulation, we can further characterize the ruggedness in the energy landscape by calculating $P_s(q_s)$. Here again structure **B** is the target, and eq 35 is the energy function. In Figure 13, we present $P_s(q_s)$ vs q_s for different energies. Note that at the lowest energies $E = -84\epsilon$ and $E = -82\epsilon$, q_s is clustered around large, positive values; the sequences at these energies are similar to one another. At $E = -80\epsilon$, however, a second peak in $P_s(q_s)$ appears for negative values of q_s. The peak for positive q_s and that for negative q_s represent groups of sequence pairs that are correlated and anticorrelated, respectively. Two sequences are anticorrelated if the sequences positions that are most likely P in one sequence are likely to be H in the other, and vice versa. For the energy function eq 35, the landscape has two deep minima that compete at low energies.

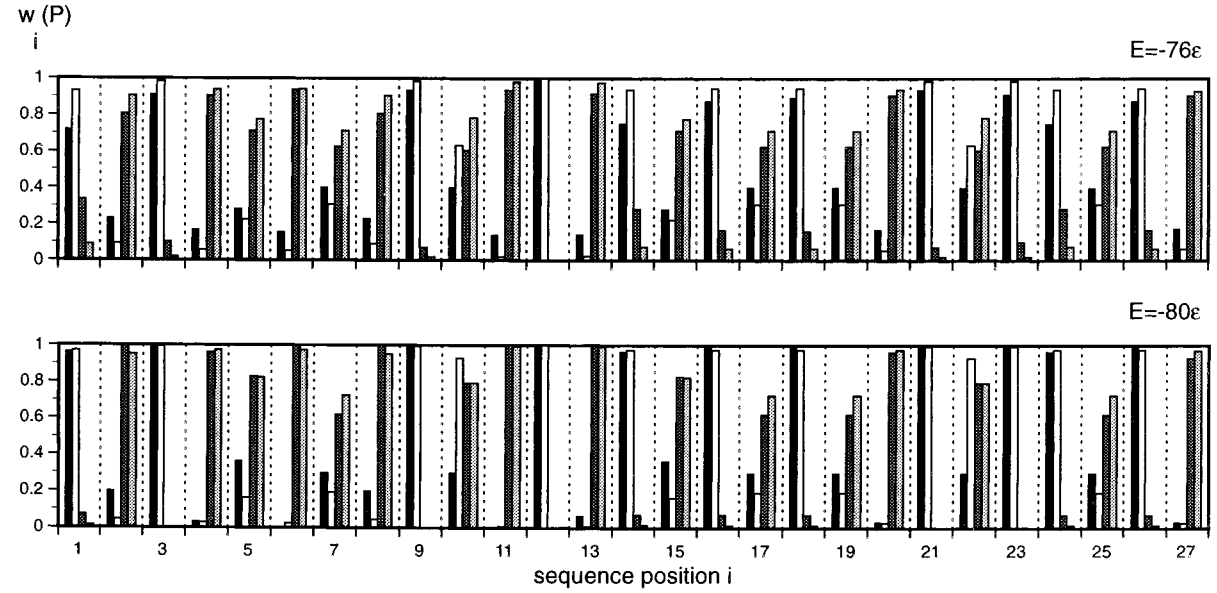

Figure 12. The site identity probabilities w_i(P) vs i for two cases where site $i = 12$ is constrained to be either P or H. When site 12 is P, both the exact (black) and theoretical (white) are shown. Also shown are the exact (dark gray) and theoretical (light gray) results when site 12 is H. Here the P:H composition is 14:13, and the energy function is that of eq 35.

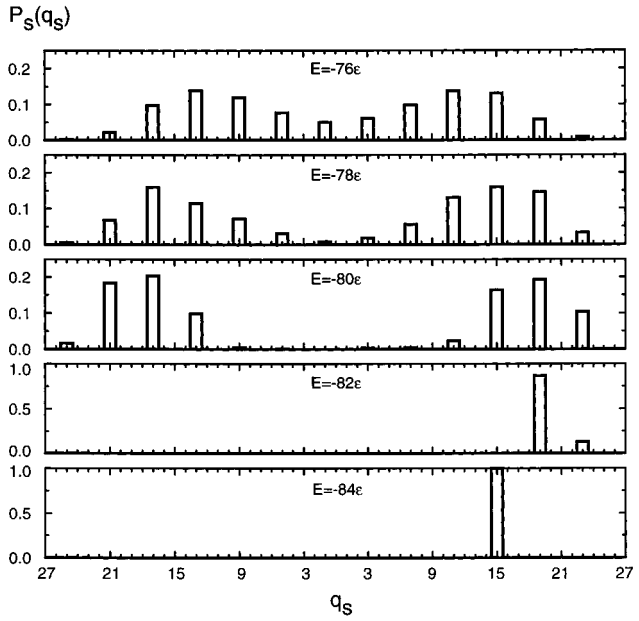

Figure 13. The sequence overlap probability $P_s(q_s)$ as a function of the sequence overlap q_s. $P_s(q_s)$ is plotted for different values of the energy that sequences have when they assume structure **B**. The energy function is that of eq 35. The energies considered are on the low-energy side of $S(E)$ (see Figure 11).

As we mentioned previously, the origin of these two minima arises because of the symmetry of the energy function (see eqs 30 and 35.

4. Discussion of the Implications of the Theory

The theory allows us to estimate the number of foldable sequences, as well as the "mutability" of particular residues and correlations between residue mutations. The theoretical results are in very good agreement with the exact results presented here. The mean-field theory provides an efficient way to estimate the number of sequences that have a certain energy in the chosen structure. This is especially useful for large macromolecules such as proteins, where exact enumeration is only possible for instances with limited numbers of variable residues or with very small monomer pools.

The theory presented here does not treat a number of issues that are undoubtedly present in the design specific sequences. The theory presented does not account for the possibility that some sequences may fold to structures other than the chosen folded structure. In addition, many of the protein design experiments done so far have yielded proteins that exhibit a large degree of ordering but which do not fold to a unique structure. These proteins have many of the same characteristics of molten globule states.[76] In addressing these experiments, we will need to consider the probability of folding not to a single structure but to a set of structures that are all within some chosen degree of similarity with the target structure. We reserve this for future work.

Using a thermodynamic functional $S(E)$ that treats the protein as a collection of sites that are effectively noninteracting is a severe approximation. Just as in theories of the liquids, however, we can develop higher order functionals which account for multibody correlations between monomer identities.[77] Such embellishments may improve the accuracy of the theory. Regardless, the exact results presented in this report are in strikingly good agreement with the simple version of the theory presented in section 2.

In discussing the significance of the theory to the folding problem, we made use of the assumption that the distribution of the conformational energies of the nonfolded states is dependent only on composition. The energetics of the unfolded states of the polymer, however, need not be independent of the order of the sequence. At low temperatures, this "constant composition assumption" implies that the energies of these low energy unfolded conformations are essentially the same across sequences having the same overall composition. The energies of the lowest energy unfolded conformations are said to be self-averaging across different sequences. In many cases the assumption is useful,[28,33,41] but more generally, it neglects a number of important effects. A sequence having a low-energy in the target structure need not fold uniquely to that structure. The sequence may acquire a alternate conformations having comparable or lower energies.[43,58] In proteins, correlations between the target or folded structure and the ensemble of unfolded conformations are undoubtedly present. This is due in part to the fact that the molten globule, which is an ensemble of incompletely folded conformations, can have substantial

amounts of native ordering. Modest mutations can substantially affect a molten globule's structure[78] and stability.[79] The conformational constraints of the individual monomers may preclude one unfolded sequence from assuming conformations that are accessible to other sequences. In the unfolded ensemble, the peptide backbone of proteins may readjust into stable conformations, e.g., by readjusting so as to minimize exposure of hydrophobic residues to solvent. This type of conformational amenability is likely to fluctuate from sequence to sequence due to the conformational preferences of the monomers, especially if some form of partial ordering is present. Nonetheless, the theory presented in this paper for quantifying the sequence energetics of a particular structure is an important step toward a statistical understanding the compatibility between sequence and structure in foldable polymers.

Most of the potentials used in computational studies of protein folding are statistically derived. The theory of combinatorial design perhaps makes the best use of these types of potential. In some sense, these potentials are attempts to make the wisdom of the database quantitative. In so doing, researchers average over many structures and sequences. A statistical theory of partial design using these potentials may be more appropriate than their use to design particular sequences. The design of *particular* sequences using these potentials is tenuous, since the accuracy of these potentials is still controversial, and the search for particular sequences is very sensitive to the energy function. On the other hand, when used in combination with a statistical theory of sequence ensembles, information-based potentials are likely to yield useful estimates of the *average* identities of sequences having a chosen energy and conformation. The theory presented here may also use statistical energy functions to filter the sequence space, and the effluent identity probabilities can then be used to specify the optimal monomer set for each site in the sequence. More detailed energy functions or design algorithms may then be used to trickle down the sequence energy landscape in the search for local minima that represent foldable sequences. More refined potentials are undoubtedly necessary for the design of specific sequences, since specific bonding between residues and the complementary packing of side chains are likely prerequisites for obtaining structures with all the conformational specificity of natural proteins.[80,81]

In addition to aiding in the design of sequences that fold to a particular target, the theory should allow us to assay the designability of particular structures. One suggested determinant in the designability of particular structures is symmetry. A large degree of inexact symmetry may be present in a structure that is more designable than others, i.e., a structure that is the ground state for a large number of sequences. Li et al. have noted some crude symmetries in their most designable structures and have suggested that these symmetries are paramount.[70] It is unclear, however, how to extrapolate their results on small lattice systems to actual proteins. The issue of symmetry has been addressed from a more general viewpoint, and it has been suggested that the symmetries observed in proteins arise from physical principles analogous to those that guide clusters of small molecules into symmetric forms.[82,83]

We close this section with a conjecture on the form of the sequence energy landscape for structures and energy functions where foldable sequences are common. Recall that for structure **A**, the distribution of the similarity of sequences $P_s(q_s)$ widens and shifts as the energy increases, but does not jump sharply and is not bimodal. As the energy increases, the number of sequences increases, and this ensemble of sequences becomes gradually more diverse. This smooth increase in the number of sequences, or the sequence entropy $S(E)$ with energy means that the sequence landscape has a funnel- or basin-like topography.[44] Such a sequence space topography may describe why some structures can support a large number of mutations of their individual residues. Evolutionary dynamics in such a funnel should lead rather quickly to low-energy sequences, in analogy with Brownian motion in a basin-like potential. Many native structures may be near the bottom of the funnel, which may help explain why many mutations in native proteins are destabilizing.[84] Here we have discussed the evolutionary effect of mutations only in terms of their effect on protein stability. In evolving organisms, a number of other issues must be considered including the effects of these mutations on the rate of folding in vivo and on the function of the protein, both of which affect may affect the viability of an organism. Clearly, thermostability of some form, however, is a prerequisite for a functional protein, and native state stability must be considered in the evolution of proteins.

5. Summary

We have presented a self-consistent theory that, for a fixed target conformation, may be used to characterize the energetics of sequence ensembles of any size. In this report, we present the theory and compare its predictions with results from an exactly enumerable system. The theory gives excellent estimates of $S(E)$, the logarithm of the number of sequences having energy E in a chosen structure. At a chosen energy, the theory gives excellent estimates of $w_i(\alpha)$, the probability that each sequence position is occupied by a particular type of monomer. The theory may be used to evaluate design strategies for particular target protein architectures. The theory may also be used to compare the "designability" of different structures, that is, how the number of low-energy sequences varies between structures. Last, the theory makes scanning large ranges of sequence space fast and feasible, which is currently not possible with either computational or experimental methods.

Acknowledgment. This material is based upon work supported in part by the National Science Foundation under grant CHE-93-01474 awarded to J.G.S. in the form of a Postdoctoral Fellowship. The authors also thank the Critical Research Initiatives Program of the University of Illinois and the Donors of the Petroleum Research Fund (administered by the American Chemical Society) for additional support. This work was completed while P.G.W. was a Scholar-in-Residence at the Fogarty International Center of the National Institutes of Health.

References and Notes

(1) Leffler, J. E.; Grunwald, E. *Rates and Equilibria of Organic Reactions*; John Wiley and Sons: New York, 1963.

(2) Orengo, C. A.; Jones, D. T.; Thornton, J. M. *Nature* **1994**, *372*, 631−634.

(3) Wolynes, P. G.; Onuchic, J. N.; Thirumalai, D. *Science* **1995**, *267*, 1619−1620.

(4) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. *Proteins* **1995**, *21*, 167−195.

(5) Dill, K. A.; Bromberg, S.; Yue, K.; Fiebig, K. M.; Yee, D. P.; Thomas, P. D.; Chan, H. S. *Protein Sci.* **1995**, *4*, 561−602.

(6) Shakhnovich, E. I. *Folding Des.* **1996**, *1*, R50−R54.

(7) Bryngelson, J. D.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 7524−7528.

(8) Mezard, M.; Parisi, E.; Vivasaro, M. A. *Spin Glass Theory and Beyond*; World Scientific Press: Singapore, 1986.

(9) Bryngelson, J. D.; Wolynes, P. G. *Biopolymers* **1990**, *30*, 177−188.

(10) Bryson, J. W.; Betz, S. F.; Lu, H. S.; Suich, D. J.; Zhou, H. X.; O'Neil, K. T.; DeGrado, W. F. *Science* **1995**, *270*, 935−941.

(11) Kuroda, Y. *Protein Eng.* **1995**, *8*, 97−101.

(12) Dolgikh, D. A.; Kirpichnikov, M. P.; Ptitsyn, O. B.; Chemeris, V. V. *Molec. Biol.* **1996**, *30*, 149−156.

(13) Betz, S.; Raleigh, D.; DeGrado, W.; Lovejoy, B.; Anderson, D.; Ogihara, N.; Eisenberg, D. *Folding Des.* **1996**, *1*, 57−64.
(14) Ponder, J. W.; Richards, F. M. *J. Mol. Biol.* **1987**, *193*, 775−791.
(15) Hellinga, H. W.; Richards, F. M. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 5803−5807.
(16) Bowie, J. U.; Lüthy, R.; Eisenberg, D. *Science* **1991**, *253*, 164−170.
(17) Godzik, A.; Kolinski, A.; Skonick, J. *J. Mol. Biol.* **1992**, *227*, 227−238.
(18) Brenner, S. E.; Berry, A. *Protein Sci.* **1994**, *3*, 1871−1882.
(19) Miyazawa, S.; Jernigan, R. L. *Macromolecules* **1985**, *18*, 534−552.
(20) Sippl, M. J. *J. Mol. Biol.* **1990**, *213*, 859−883.
(21) Friedrichs, M. S.; Goldstein, R. A.; Wolynes, P. G. *J. Mol. Biol.* **1991**, *222*, 1013−1034.
(22) Goldstein, R.; Luthey-Schulten, Z. A.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 9029−9033.
(23) Miyazawa, S.; Jernigan, R. L. *J. Mol. Biol.* **1996**, *256*, 623−644.
(24) Bryngelson, J. D. *J. Chem. Phys.* **1994**, *100*, 6038−6045.
(25) Pande, V. S.; Grosberg, A. Y.; Tanaka, T. *J. Chem. Phys.* **1995**, *103*, 9482−9491.
(26) Rooman, M. J.; Wodak, S. J. *Protein Eng.* **1996**, *8*, 849−858.
(27) Thomas, P. D.; Dill, K. A. *J. Mol. Biol.* **1996**, *257*, 457−469.
(28) Pereira de Araújo, A. F.; Pochapsky, T. C. *Folding Des.* **1996**, *1*, 299−314.
(29) Ramanathan, S.; Shakhnovich, E. *Phys. Rev. E* **1994**, *50*, 1303−1312.
(30) Pande, V. S.; Grosberg, A. Y.; Tanaka, T. *Phys. Rev. E* **1995**, *51*, 3381−3392.
(31) Pande, V. S.; Grosberg, A. Y.; Tanaka, T. *Macromolecules* **1995**, *28*, 2218−2227.
(32) Sasai, M. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 8438−8442.
(33) Shakhnovich, E. I.; Gutin, A. M. *Protein Eng.* **1993**, *6*, 793−800.
(34) Govindarajan, S.; Goldstein, R. A. *Biopolymers* **1995**, *36*, 43−51.
(35) Jones, D. T. *Protein Sci.* **1994**, *3*, 567−574.
(36) Sun, S.; Brem, R.; Chan, H. S.; Dill, K. A. *Protein Eng.* **1995**, *8*, 1205−1213.
(37) Deutsch, J. M.; Kurosky, T. *Phys. Rev. Lett.* **1996**, *76*, 323−326.
(38) Abkevich, V. I.; Gutin, A. M.; Shakhnovich, E. I. *Folding Des.* **1996**, *1*, 221−230.
(39) Morrissey, M. P.; Shakhnovich, E. I. *Folding Des.* **1996**, *1*, 391−405.
(40) Seno, F.; Vendruscolo, M.; Maritan, A.; Banavar, J. R. *Phys. Rev. Lett.* **1996**, *77*, 1901−1904.
(41) Hinds, D. A.; Levitt, M. *J. Mol. Biol.* **1996**, *258*, 201−209.
(42) Pande, V. S.; Grosberg, A. Y.; Tanaka, T. *J. Chem. Phys.* **1994**, *101*, 8246−8257.
(43) Richardson, J. S.; Richardson, D. C.; Tweedy, N. B.; Gernert, K. M.; Quinn, T. P.; Hecht, M. H.; Erickson, B. W.; Yan, Y. B.; Mcclain, R. D.; Donlan, M. E.; Surles, M. C. *Biophys. J.* **1992**, *63*, 1186−1209.
(44) Govindarajan, S.; Goldstein, R. A. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 3341−3345.
(45) Shakhnovich, E.; Abkevich, V.; Ptitsyn, O. *Nature* **1996**, *379*, 96−98.
(46) Davidson, A. R.; Sauer, R. T. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 2146−2150.
(47) LaBean, T. H.; Kauffman, S. A.; Butt, T. R. *Molecular Diversity* **1995**, *1*, 29−38.
(48) Kamtekar, S.; Schiffer, J. M.; Xiong, H.; Babik, J. M.; Hecht, M. H. *Science* **1993**, *262*, 1680−1685.
(49) Matthews, B. W. *Annu. Rev. Biochem.* **1993**, *62*, 139−160.
(50) Baldwin, E.; Xu, J.; Hajiseyedjavadi, O.; Baase, W. A.; Matthews, B. W. *J. Mol. Biol.* **1996**, *259*, 542−559.
(51) Axe, D. D.; Foster, N. W.; Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 5590−5594.
(52) Itzhaki, L. S.; Otzen, D. E.; Fersht, A. R. *J. Mol. Biol.* **1995**, *254*, 260−288.
(53) Sander, C.; Schneider, R. *Proteins* **1991**, *9*, 56−68.
(54) Finkelstein, A.; Reva, B. *Nature* **1991**, *351*, 497−499.
(55) Lee, C. *J. Mol. Biol.* **1994**, *236*, 918−939.
(56) Koehl, P.; Delarue, M. *Curr. Opin. Struct. Biol.* **1996**, *6*, 222−226.
(57) Reva, B. A.; Finkelstein, A. V.; Rykunov, D. S.; Olson, A. *J. Proteins* **1996**, *26*, 1−8.
(58) Yue, K.; Fiebig, K. M.; Thomas, P. D.; Chan, H. S.; Shakhnovich, E. I.; Dill, K. A. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 325−329.
(59) McQuarrie, D. A. *Statistical Mechanics*; Harper & Row: New York, 1976.
(60) Feller, W. *An Introduction to Probability Theory and Its Applications*; John Wiley & Sons: New York, 1957; Vol. 1.
(61) Koretke, K. K.; Luthey-Schulten, Z.; Wolynes, P. G. *Protein Sci.* **1996**, *5*, 1043−1059.
(62) Bethe, H. A. *Proc. R. Soc. London, Ser A* **1935**, *150*, 552−575.
(63) Hill, T. L. *Statistical Mechanics*; McGraw-Hill: New York, 1956.
(64) Huang, K. *Statistical Mechanics*, 2nd ed.; John Wiley & Sons, Inc.: New York, 1987.
(65) Kubo, R.; Toda, M.; Hishitsume, N. *Statistical Physics II: Nonequilibrium Statistical Mechanics*, 2nd ed.; Springer Series in Solid-State Sciences; Springer-Verlag: Berlin, 1991; Vol. 30.
(66) Callen, H. B. *Thermodynamics and an Introduction to Thermostatistics*, 2nd ed.; John Wiley & Sons: New York, 1985.
(67) Shakhnovich, E.; Gutin, A. *J. Chem. Phys.* **1990**, *93*, 5967−5971.
(68) Leopold, P. E.; Montal, M.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 8721−8725.
(69) Unger, R.; Moult, J. *J. Mol. Biol.* **1996**, *259*, 988−994.
(70) Li, H.; Helling, R.; Tang, C.; Wingreen, N. *Science* **1996**, *273*, 666−669.
(71) Socci, N. D.; Onuchic, J. N. *J. Chem. Phys.* **1994**, *101*, 1519−1528.
(72) Socci, N. D.; Onuchic, J. N. *J. Chem. Phys.* **1995**, *103*, 4732−4744.
(73) Finkelstein, A. V.; Gutin, A. M.; Badretdinov, A. Y. *FEBS Lett.* **1993**, *325*, 23−28.
(74) Finkelstein, A. V.; Badretdinov, A. Y.; Gutin, A. M. *Proteins* **1995**, *23*, 142−150.
(75) Onuchic, J. N.; Wolynes, P. G.; Luthey-Schulten, Z.; Socci, N. D. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 3626−3630.
(76) Kuwajima, K. *Proteins* **1989**, *6*, 87−103.
(77) Morita, T.; Hiroike, K. *Prog. Theor. Phys.* **1961**, *25*, 537−578.
(78) Lin, L.; Pinker, R. J.; Forde, K.; Rose, G. D.; Kallenbach, N. R. *Nature Struct. Biol.* **1994**, *1*, 447−452.
(79) Uchiyama, H.; Perezprat, E. M.; Watanabe, K.; Kumagai, I.; Kuwajima, K. *Protein Eng.* **1995**, *8*, 1153−1161.
(80) Richards, F. M.; Lim, W. A. *Q. Rev. Biophys.* **1993**, *26*, 423−498.
(81) Bahar, I.; Jernigan, R. L. *Folding Des.* **1996**, *1*, 357−370.
(82) Lingård, P.-A.; Bohr, H. *Phys. Rev. Lett.* **1996**, *77*, 779−782.
(83) Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 14249−14255.
(84) Goldenberg, D. P. Mutational analysis of protein folding and stability. In Protein Folding; Creighton, T. E., Ed.; W. H. Freeman: New York, 1992.