The effect of sequence on the conformational stability of a model heteropolymer in explicit water

Bryan A. Patel,[1] Pablo G. Debenedetti,[1,a)] Frank H. Stillinger,[2] and Peter J. Rossky[3]
[1]*Department of Chemical Engineering, Princeton University, Princeton, New Jersey 08544, USA*
[2]*Department of Chemistry, Princeton University, Princeton, New Jersey 08544, USA*
[3]*Department of Chemistry and Biochemistry, Institute for Theoretical Chemistry, University of Texas at Austin, Austin, Texas 78712, USA*

(Received 30 January 2008; accepted 25 March 2008; published online 1 May 2008)

We investigate the properties of a two-dimensional lattice heteropolymer model for a protein in which water is explicitly represented. The model protein distinguishes between hydrophobic and polar monomers through the effect of the hydrophobic monomers on the entropy and enthalpy of the hydrogen bonding of solvation shell water molecules. As experimentally observed, model heteropolymer sequences fold into stable native states characterized by a hydrophobic core to avoid unfavorable interactions with the solvent. These native states undergo cold, pressure, and thermal denaturation into distinct configurations for each type of unfolding transition. However, the heteropolymer sequence is an important element, since not all sequences will fold into stable native states at positive pressures. Simulation of a large collection of sequences indicates that these fall into two general groups, those exhibiting highly stable native structures and those that do not. Statistical analysis of important patterns in sequences shows a strong tendency for observing long blocks of hydrophobic or polar monomers in the most stable sequences. Statistical analysis also shows that alternation of hydrophobic and polar monomers appears infrequently among the most stable sequences. These observations are not absolute design rules and, in practice, these are not sufficient to rationally design very stable heteropolymers. We also study the effect of mutations on improving the stability of the model proteins, and demonstrate that it is possible to obtain a very stable heteropolymer from directed evolution of an initially unstable heteropolymer. © *2008 American Institute of Physics*. [DOI: 10.1063/1.2909974]

I. INTRODUCTION

Natural proteins possess a well-defined functional native conformation that is stable within a limited range of temperatures, pressures, and solvent conditions. A characteristic protein phase diagram showing the effects of pressure and temperature on the native state is given in Fig. 1 for the case of Staphylococcal nuclease.[1] This well-defined native structure is determined entirely by the protein's amino acid sequence, but a detailed understanding of how sequence determines structure and function remains elusive. Such knowledge would aid in the design of new proteins for a variety of applications, including pharmaceuticals, enzyme catalysis, and biomaterials. However, since there are 20 different naturally occurring amino acids, the number of possible sequences for a protein with N_{Mon} amino acids is $20^{N_{\mathrm{Mon}}}$. The majority of globular proteins have between 50 and 400 amino acids and, therefore, an average protein with 200 amino acids has more than $20^{200} \approx 10^{260}$ possible sequences available.[2] Because only a very small fraction of these possible sequences fold into functional native states,[3] finding a sequence to successfully fold into a desired native state is a significant design problem to overcome.[4,5]

Atomically detailed models have been extensively used to investigate the relationship between protein sequence and native state structure, but computational methods using these models cannot yet reliably predict structure from sequence. There are several challenges in this field, including the prohibitive computational cost of these calculations, the need for effective algorithms to search conformation space, and the development of accurate force fields to appropriately represent the molecular interactions between individual amino acids and their interaction with water. In addition to atomically detailed models, coarse-grained models are used as a complementary technique for the study of protein design. These models reduce the complexity of the protein's configuration space and sequence space to make a detailed investigation of the model protein's properties practical. Minimalist protein models have proven valuable in understanding protein sequence-structure relationships and important elements of folding.[6–20] Minimalist models typically employ single site representations of amino acid residues, and many reduce conformational complexity by restricting protein configurations to a lattice.[6–14] Many models use simplified descriptions of the intramolecular interactions based on the hydrophobicity of the amino acids, often employing a two-letter[7,9,18] or three-letter[15–17] amino acid alphabet. Dihedral angle potentials are also used in some off-lattice models to enforce the structural constraints of the polypeptide backbone,[15–19] a key determinant of a secondary structure.[21] In all of these minimalist models, solvent contributions are

a)Electronic mail: pdebene@princeton.edu.

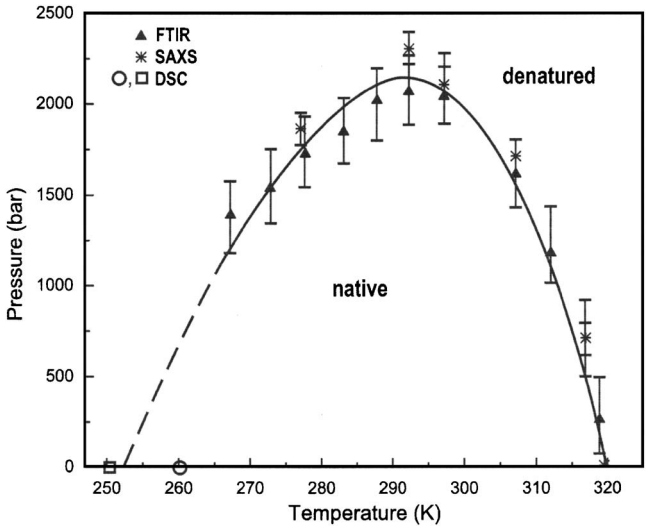

FIG. 1. Phase diagram of Staphylococcal nuclease from a combination of Fourier transform infrared spectroscopy, small angle X-ray scattering, and differential scanning calorimetry experiments. Adapted with permission (Ref. 1).

captured only through the use of effective potentials describing the mean forces between residues, which are implicitly solvent averaged.

One of the most basic of such minimalist approaches is the HP model,[7] a lattice heteropolymer model with hydrophobic (H) and polar (P) monomers. This two-letter partitioning reflects a basic property of biological proteins in which hydrophobic amino acids tend to be folded into the core of the protein in its native conformation to avoid exposure to water, while polar amino acids reside preferentially at the surface.[22] The HP model implicitly incorporates the effect of water only through an attraction between hydrophobic monomers. This interaction drives the formation of a hydrophobic core in the model protein's native state, which is defined as the lowest-energy state. Polar monomers are energetically neutral with respect to their surroundings, and so have no intrinsic preference to segregate in any region. While the HP model is a rough approximation of a protein, and the lack of secondary structural preferences limits its use as a predictive tool,[21] studies examining the design of protein native states and structures in these models have captured some critical properties of real proteins. For example, only a small fraction of HP sequences fold into unique native states.[23] HP sequences with low-degeneracy native states also have an optimal balance of H and P monomers, and increasing or decreasing the hydrophobicity destabilizes the native state.[24]

Experimental investigators have successfully used reduced protein alphabets to rationally design proteins with specific structures. Hecht and co-workers have taken advantage of recurring patterns of hydrophobic and polar amino acids common to secondary structural elements such as α-helices and β-sheets to synthesize large libraries of stable proteins.[25–29] Their technique generates a large body of distinct protein sequences with the same predesigned sequence in the reduced protein alphabet (e.g., PHPPHHPP), while each has a distinct sequence of amino acids in the 20-letter alphabet (e.g., KLNDLLED or KLQEMMKE).[29] They have had particular success in generating sequences of proteins that fold into four-helix bundles,[25,26] and in selecting these proteins for particular functions.[30,31] The α-helices are patterned in the form PHPPHHPPHPPHHP,[25] which is based on the periodicity of 3.6 residues per complete turn of the helix. Hydrophobic amino acids are placed along the protein chain according to this periodicity so that when the helix is folded it has a hydrophobic and a hydrophilic face. When the four helices in the bundle come together they are able to sequester their hydrophobic faces by forming a hydrophobic core. West *et al.* have also designed proteins with β-sheets based on an alternating pattern of PHPHPHPHP.[27] However, they were unable to synthesize proteins with nativelike states using this pattern, since the designed sequences tended to form nonnative fibrillar structures analogous to those observed in various degenerative neurological diseases such as Alzheimer's.[32] They then analyzed a large library of natural proteins, which revealed that these alternating patterns of H's and P's appear with less frequency than expected by random chance.[28] They concluded that the rarity of alternations is due to the potentially harmful effects of misfolded proteins containing these patterns and, consequently, is disfavored by natural selection. Other studies of patterns in protein libraries have focused on the physical locations of the patterns in the protein or the placement of the protein within the cell. Long blocks of hydrophobic amino acids are common in parallel β-sheets, which tend to be buried within the core of the protein, where hydrophobic amino acids preferentially reside.[33] Long blocks of hydrophobic amino acids are also far more common in membrane proteins, where there is little water present, than in aqueous proteins.[34]

Such insights provide clues about which patterns fold proteins into stable native states. Other studies have focused on correlating amino acid substitutions with increased stability to extreme conditions such as high temperature, high pressure, or low temperature.[35–39] These studies use proteins from extremophiles, bacteria specifically adapted to survive in harsh environments. These proteins are often very similar to proteins with the same function from normal bacteria, typically sharing the majority of their respective sequences.[35] Examining proteins with enhanced thermostability (stability at high temperatures) shows that only a few mutations are required to dramatically change the protein's stability,[40] and some guidelines for the interactions that confer thermostability have been put forward.[37,39] Proteins that are unusually stable to low temperature[36,41,42] or high pressure[43,44] have also been examined, but less data are available on the sequences or structures of these proteins. In spite of this work, it remains challenging to directly correlate sequence with overall thermodynamic properties of proteins, such as the temperature dependence of the free energy of unfolding.[39] A unified picture of how changes in sequence affects the protein folding phase diagram is still lacking, partly because experimental studies of cold and pressure denaturation are hindered by the difficulty of probing protein behavior at low temperatures and high pressures. Thermodynamic data for a wide range of temperature and pressures are only available for a small number of proteins.[1,45–49] Understanding how se-

quence affects all aspects of protein thermodynamics, especially low-temperature and high-pressure stability, would help to clarify the mechanisms and driving forces of these processes.

Because of the challenges to computationally determine the native state stability, little simulation work has been done to correlate sequence with stability. It is impractical to simulate a variety of sequences using atomically detailed models because of the intense computational effort required to obtain thermodynamic properties of the system. Coarse-grained models such as the HP model are more amenable to studying the effects of sequence but are limited in the properties that can be examined. The HP model shows only a broad unfolding transition upon increasing temperature, and because the model protein's native state is defined as the total system ground state, it does not cold-unfold upon decreasing temperature.[7] Thus, the HP model has only a single measure of stability, the unfolding temperature, which is analogous to a thermal denaturation temperature. A more detailed model that displays more proteinlike phase behavior (i.e., cold, pressure, and thermal denaturation) could offer a more realistic basis for studying sequence-stability correlations. With such a model, linking changes in the protein's sequence directly to perturbations of its native state stability, as represented in Fig. 1, could provide insight into designing stable protein sequences.

Here, we extend a recently developed lattice model of a hydrophobic homopolymer in explicit water that exhibits cold-, pressure-, and heat-induced protein unfolding,[50] reproducing the shape of the experimental phase diagram in Fig. 1. We incorporate sequence into the model with a two-letter alphabet of hydrophobic and polar amino acids. While we employ the same choice of an alphabet as the HP model, our model explicitly treats the hydrophobic effect through the enthalpy and entropy of the solvating water molecules. The flat histogram method used to simulate the model provides good sampling of rare protein configurations such as the native state. This approach yields the protein thermodynamics over a wide range of temperatures and pressures, but it does not address the kinetics of protein folding. In spite of the added complexity of including water explicitly, we are able to simulate a large body of heteropolymer sequences and relate aspects of the sequence to native state stability.

The outline of the paper is as follows. In Sec. II, we describe the details of the heteropolymer model. This is followed by an explanation of the flat histogram methods used to estimate the density of states and extract thermodynamic properties in Sec. III. We also develop the statistical analysis procedure required to extract meaningful trends from the simulation data. In Sec. IV, we discuss general observations on the phase diagram of model heteropolymers, and discuss how sequence affects stability through the use of statistical analysis and directed evolution studies. In Sec. V, we close by presenting the main conclusions and possible paths for further development of this work.

II. MODEL DESCRIPTION

The protein is modeled as a self-avoiding heteropolymer composed of hydrophobic (H) and polar (P) monomers on a two-dimensional (2D) square lattice. The protein is schematically shown in Fig. 2 by the connected black (H) and gray (P) beads, with covalently bonded monomers occupying nearest-neighbor sites on the lattice. The protein has no self-interaction aside from excluded volume effects. The only interaction of the protein with the water is through its indirect effect on water-water hydrogen bonding. The presence of hydrophobic and polar monomers produces different effects on the hydrogen bonding of the neighboring water molecules. The interactions described below generalize the implicit attraction of H monomers in the original HP model[7] to an explicit account of the effect of hydrophobic groups on the solvating water molecules.[50] In the original HP model, the H-P and P-P interaction is zero, so there was no energetic distinction between P monomers on the surface or in the interior of protein. With the introduction of an explicit solvent, an appropriate representation must be devised for the interaction of the polar monomers with the solvent and with each other. One alternative is to include an attraction between polar monomers and between polar monomers and water to mimic the presence of a hydrogen bond. However, in real proteins, polar side chains can also interact with the backbone of the protein, and in this model the backbone is not represented separately from the functional groups. In the present work, we take a simpler approach which nevertheless retains the heteropolymer character of the problem.

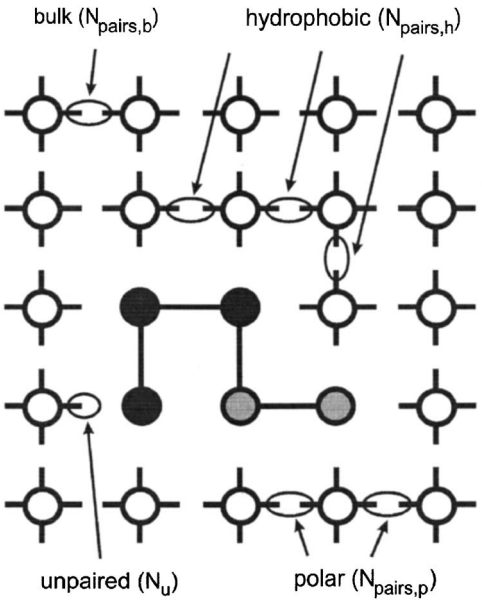

FIG. 2. Schematic of the model protein and water. The black circles are hydrophobic (H) monomers, the gray circles are polar (P) monomers, and the lines connecting them are covalent bonds. The white circles are water molecules, and the four arms on each water molecule are the hydrogen bonding arms. Examples of each of the four types of bonding arms are shown, along with the variables which count their number: bulk bonding arm pairs ($N_{pairs,b}$), hydrophobic bonding arm pairs ($N_{pairs,h}$), polar bonding arm pairs ($N_{pairs,p}$), and unpaired bonding arms (N_u). This figure shows a portion of the whole system and, in practice, a much larger box is used to prevent the protein from interacting with itself across the periodic boundary.

The water model is adapted from a lattice model developed by Sastry *et al.* to investigate the thermodynamics of supercooled water.[51] The model, which also included empty lattice sites, displays many of the signature anomalies of wa-

ter, including the isobaric density maximum,[51] and the increase upon isobaric cooling of the isothermal compressibility, isobaric heat capacity, and the magnitude of the thermal expansion coefficient.[52] Each water molecule occupies one site on the lattice, and every site not occupied by the protein is inhabited by water molecules. As shown in Fig. 2, the water molecules have four hydrogen bonding arms, each associated with a neighboring lattice site. The variable σ_{ij} represents the orientation of a bonding arm on water molecule i associated with the neighbor site j. There are q possible orientations for each bonding arm, and σ_{ij} can have values between 1 and q. Bonding arms on the same water molecule can adopt orientations independently of each other.

A hydrogen bond forms between two adjacent water molecules i and j when their bonding arm orientations satisfy the criterion $|\sigma_{ij}-\sigma_{ji}| \leq \lambda$. λ is the tolerance for hydrogen bonding, or alternatively the size of the range of acceptable bonding arm pair orientations. When the associated hydrogen bonding arms on neighboring water molecules satisfy the above criterion, the energy of the system is reduced by an amount J, the strength of a hydrogen bond. We treat the lattice as compressible in order to account for the lower local density associated with hydrogen bonding, and upon formation of a hydrogen bond, the total volume expands uniformly by an amount Δv. The total volume V is then determined by the total number of hydrogen bonds N_{HB} from the relation

$$V = V_0 + \Delta v N_{\mathrm{HB}}, \tag{1}$$

where V_0 is the volume of the lattice without hydrogen bonding.

We incorporate principles of hydrophobic hydration into the model by allowing the presence of the protein in the vicinity of water molecules to affect the parameters of hydrogen bonding, as in our earlier work.[50] Frank and Evans surmised that water molecules tend to order around nonpolar solutes,[53] avoiding orientations in which their hydrogen bonding arms point toward the hydrophobe. In such low-entropy configurations, these water molecules sample the distorted and weaker bonding structures present in bulk with comparatively less frequency and, therefore, form stronger hydrogen bonds in the first solvation shell than in bulk.[54,55] We refer to these phenomena as the entropic penalty and enthalpic bonus for hydrogen bonding around hydrophobic solutes, and model them by distinguishing between three classes of hydrogen bonds: those formed in bulk, those formed around hydrophobic monomers, and those formed around polar monomers. Quantities referring to one of these types of hydrogen bonds use the subscripts b for bulk, h for hydrophobic, and p for polar.

The criteria for determining whether a pair of bonding arms belongs in the bulk, hydrophobic, or polar classes are illustrated in Fig. 2. Each of these classes has a distinct hydrogen bonding criterion given by $|\sigma_{ij}-\sigma_{ji}| \leq \lambda_x$, where $x = b, h$, or p. A pair of water molecules is subject to the hydrophobic bonding arm tolerance λ_h if either member of the pair is adjacent to one or more H monomers. A bonding arm pair is subject to the polar bonding arm tolerance λ_p if either member of the pair is adjacent to one or more P monomers and *neither* member is adjacent to an H monomer. Finally, a bonding arm pair uses the bulk bonding arm tolerance λ_b if neither member of the pair is adjacent to any protein monomers.

The entropic penalty for hydrogen bonding around hydrophobic monomers arises from the relative values of the bonding arm tolerances λ_b, λ_p, and λ_h, since these parameters directly affect the fraction of orientations capable of bonding. There are q^2 possible values for the orientations of a pair of associated bonding arms on neighboring water molecules. The total number of orientations satisfying the bulk bonding criterion is $(2\lambda_b+1)q$ because if one of the arms adopts any of q possible orientations, its partner must assume one of $2\lambda_b+1$ orientations in order to form a hydrogen bond. If $\lambda_h < \lambda_b$, then there is an entropic cost for hydrogen bonding around H monomers, since there is a smaller fraction of orientation pairs that satisfy the hydrophobic bonding criterion $[(2\lambda_h+1)/q]$ than the bulk bonding criterion $[(2\lambda_b+1)/q]$.

We model the enthalpic bonus for hydrogen bonding around hydrophobic monomers by differentiating between the strengths of the three classes of hydrogen bonds. The bulk hydrogen bonds form with a base strength J, while hydrogen bonds around H and P monomers have strengths $J+J_{\mathrm{H}}$ and $J+J_{\mathrm{P}}$, respectively. When $J_{\mathrm{H}} > 0$, there is an enthalpic bonus for the hydrogen bonds around H monomers. While the model does allow for distinguishing the enthalpy and entropy of the hydrogen bonds around P monomers from the other classes of hydrogen bonds, in practice, we consider the entropic cost and enthalpic bonus of hydrogen bonding around the H monomers to be most important. In the calculations that follow we do not examine the effect of varying the hydrogen bonding properties of the water around P monomers, and use parameter choices $J_P=0$ and $\lambda_b=\lambda_p$. The Hamiltonian is then

$$\mathcal{H} = -J(N_{\mathrm{HB},b} + N_{\mathrm{HB},p}) - (J+J_{\mathrm{H}})N_{\mathrm{HB},h} \tag{2}$$

where $N_{\mathrm{HB},b}$ is the number of bulk hydrogen bonds, $N_{\mathrm{HB},p}$ is the number of hydrogen bonds around P monomers, and $N_{\mathrm{HB},h}$ is the number of hydrogen bonds around H monomers.

The ranges of potential parameters discussed above ensure that the solvation of hydrophobic monomers is consistent with the hydration thermodynamics of small hydrophobic solutes. For simplicity, the treatment of the "polar" monomers used here does not incorporate direct favorable interactions with the solvent. The P monomers are thus better described as being less hydrophobic than the H monomers, akin to amino acids such as glycine. For simplicity, we nevertheless refer to such residues as polar throughout this paper. In future studies, the effect of attractions between P residues and water, possibly including both hydrogen bonds as well as nondirectional attractions, should be investigated. We also note that in subsequent versions of the water-implicit HP model,[56] explicit attractions among hydrophobic and polar monomers were introduced. While these could be easily incorporated into the present framework, they are potentially confounding because these interactions represent water implicitly. In contrast, in our work, water is accounted for explicitly.

III. METHODS

A. Calculation of the density of states

To determine the folding properties of the model protein, we used a recently developed method to efficiently estimate the density of states of binary lattice systems.[57] The method separates the computation of the protein and water contributions to the density of states (DOS) and reduces the time required to accurately estimate the combined DOS. This separation is possible for systems when the degrees of freedom of the two components are *conditionally independent*. Here, we present the essential features of the method and extensions that are specific to this implementation.

A binary system with components 1 and 2 directly interacting has some potential

$$U(\xi_1,\xi_2) = U_1(\xi_1) + U_2(\xi_2) + U_i(\xi_1,\xi_2), \tag{3}$$

where U is the total potential energy, which is a function of the degrees of freedom of the two components, ξ_1 and ξ_2. The individual energies of the two components, U_1 and U_2, are only functions of their own respective internal degrees of freedom, while the interaction potential, U_i, is a function of both components' degrees of freedom. However, in many multicomponent systems, the interactions are short range and are dependent on the properties of the interfacial region. U_i can be expressed as a function of the degrees of freedom of the interface (ξ_i), which are a subset of the degrees of freedom of the two components. If the appropriate interfacial degrees of freedom can be found, the calculation of the properties of individual components can then be separated. For conditionally independent subsystems, the total DOS (Ω_t) can be subdivided according to the following relation:

$$\Omega_t(\xi_1,\xi_2;\xi_i) = \Omega_1(\xi_1;\xi_i)\Omega_2(\xi_2;\xi_i). \tag{4}$$

The notation $\Omega_1(\xi_1;\xi_i)$ refers to the DOS of component 1 as a function of its own internal degrees of freedom, given a specific set of interfacial degrees of freedom. Equation (4) allows us to express the total DOS in terms of the simpler component DOSs.

The model of protein and water presented here is well suited to separate computation because the interactions of the protein with the water do not extend past the first solvation shell. The interfacial degrees of freedom required to establish conditional independence can be simply described in terms of the properties of the first solvation shell waters: $N_{\text{pairs},h}$, $N_{\text{pairs},p}$, and N_u. $N_{\text{pairs},h}$ is the number of bonding arm pairs subject to the hydrophobic hydrogen bonding criterion, illustrated by the examples labeled "hydrophobic" in Fig. 2. $N_{\text{pairs},p}$ is the number of bonding arms pairs subject to the polar hydrogen bonding criterion, indicated by two examples labeled polar in Fig. 2. N_u is the number of unpaired bonding arms associated with a nearest-neighbor protein monomer, labeled "unpaired" in Fig. 2. These three variables contain all of the relevant information about the interfacial region required to separate the total DOS into its protein and water components. Note that the variables $N_{\text{pairs},h}$ and $N_{\text{pairs},p}$ measure the number of bonding arm *pairs*, regardless of whether or not a hydrogen bond is formed.

Since the protein has no self-interaction aside from excluded volume effects, we need only determine the degeneracy of protein configurations for each possible combination of the values of the interfacial degrees of freedom, $\Omega_{\text{prot}}(N_{\text{pairs},h},N_{\text{pairs},p},N_u)$, or $\Omega_{\text{prot}}(\xi_i)$ in shorthand notation. We use the Wang–Landau method[58] to estimate Ω_{prot} by simulating the protein *in vacuo*. Conventional Wang–Landau simulations perform a random walk in energy (U) in the range of possible energies with probabilities proportional to the reciprocal of the DOS, $1/\Omega(U)$. Instead, we perform a random walk in the interfacial degrees of freedom. In effect, this determines the degeneracy of protein configurations that would produce a set of solvation shell conditions if the water molecules were present. The simulation also determines which values of the interfacial degrees of freedom ξ_i are possible, since this is not known initially.

The density of states is not known *a priori*, but is initially set to $\Omega_{\text{prot}}=1$ for all possible configurational states. The DOS is then gradually refined during the simulation. Trial moves from an old configuration (o) with interfacial properties $\xi_{i,o}$ to a new configuration (n) with interfacial properties $\xi_{i,n}$ are accepted with probability

$$p_{\text{acc}}(o \to n) = \min\left[1, \frac{\Omega_{\text{prot}}(\xi_{i,o})}{\Omega_{\text{prot}}(\xi_{i,n})}\right]. \tag{5}$$

When a state with interfacial properties ξ_i is visited during the simulation, the corresponding bin in the DOS estimate is updated by multiplying the current value by a modification factor f, i.e., $\Omega_{\text{prot}}(\xi_i) \to \Omega_{\text{prot}}(\xi_i)f$. Initially, the modification factor is set to $f_0 = e^1 \approx 2.718\,28$ to ensure efficient sampling of all possible protein configurations. A histogram counting the frequency of visits to each configurational state $h(\xi_i)$ is updated after each trial move in the simulation, i.e., $h(\xi_i) \to h(\xi_i)+1$. The simulation continues until $h(\xi_i)$ is sufficiently flat to ensure that there is a relatively even sampling of configurational states. Here, we consider the histogram of visited states to be sufficiently flat if every bin of $h(\xi_i)$ is at least 80% of the average histogram value $\langle h(\xi_i)\rangle$. To refine the precision of the DOS estimate, the modification factor is reduced to $f_{\text{new}} = \sqrt{f_{\text{old}}}$ upon satisfying the flat histogram condition. The histogram $h(\xi_i)$ is then reset to zero and a new iteration begun. The simulation continues until the histogram of visited states is again sufficiently flat, and the modification factor is again reduced according to the same schedule. This iterative refinement of the DOS is repeated until f is less than $\exp(10^{-7})$.

We use a different approach to compute the water density of states Ω_w. The water DOS is composed of configurational ($\Omega_{w,c}$) and orientational ($\Omega_{w,o}$) components. $\Omega_{w,c}$ for a fully occupied lattice is independent of the degrees of freedom of the system and is simply the number of ways of placing N_w water molecules onto N_w lattice sites, or $N_w!$. $\Omega_{w,o}$ is dependent on the interfacial degrees of freedom, and its form can be adapted from Eq. (4) to the notation $\Omega_{w,o}(\xi_w;\xi_i)$. This quantity can be exactly computed because the orientations of the bonding arms on an individual water molecule vary independently of each other. Based on the hydrogen bonding criteria described above, there are four

possible kinds of hydrogen bonding arms, as shown in Fig. 2. The three types of paired bonding arms, $N_{\text{pairs},b}$, $N_{\text{pairs},h}$, and $N_{\text{pairs},p}$, can be further subdivided into those that have formed hydrogen bonds and those that have not. For example, the total number of bulk bonding arm pairs is the sum of the number of bulk hydrogen bonds ($N_{\text{HB},b}$) and the number of nonbonding bulk pairs ($N_{\text{NHB},b}$), or $N_{\text{pairs},b}=N_{\text{HB},b}+N_{\text{NHB},b}$. If we define ξ_w to include $N_{\text{HB},b}$, $N_{\text{HB},h}$, and $N_{\text{HB},p}$, then specifying ξ_w and ξ_i describes the orientational state of the water molecules completely.

The DOS of each of the four kinds of bonding arms can be analytically calculated from expressions developed in detail in Ref. 57. The orientational DOS of the three types of bonding arm pairs (bulk, polar, or hydrophobic) follow the same general form,

$$\Omega_{w,x}(N_{\text{HB},x},N_{\text{NHB},x}) = \frac{(N_{\text{pairs},x})!}{N_{\text{HB},x}!N_{\text{NHB},x}!} \times q^{N_{\text{pairs},x}}(2\lambda_x+1)^{N_{\text{HB},x}} \times (q-2\lambda_x-1)^{N_{\text{NHB},x}}, \quad (6)$$

where the subscript $x=b,h$, or p depends on the class of bonding arm pairs. The orientational DOS of the unpaired bonding arms is

$$\Omega_{w,u}(N_u) = q^{N_u}. \quad (7)$$

Because the protein simulation determines which combinations of the variables in ξ_i are possible, we can then determine the range of the values of the variables in ξ_w to compute the water orientational DOS. For a given set of interfacial conditions, we know the value of $N_{\text{pairs},h}$, $N_{\text{pairs},p}$, and N_u. Specifying $N_{\text{pairs},h}$ and $N_{\text{pairs},p}$ places an upper limit on the values of $N_{\text{HB},h}$ and $N_{\text{HB},p}$, since $N_{\text{pairs},h}=N_{\text{HB},h}+N_{\text{NHB},h}$. $N_{\text{pairs},b}$ can then be calculated by breaking down the total number of bonding arms into each of the four classes

$$4N_w = 2N_{\text{pairs},b} + 2N_{\text{pairs},h} + 2N_{\text{pairs},p} + N_u. \quad (8)$$

The value of $N_{\text{pairs},b}$ places an upper limit on the value of $N_{\text{HB},b}$, since $N_{\text{pairs},b}=N_{\text{HB},b}+N_{\text{NHB},b}$. The water orientational DOS can then be calculated as a product of the orientational DOS of each of the types of bonding arms, or

$$\Omega_{w,o}(\xi_w;\xi_i) = \Omega_{w,b}(N_{\text{HB},b},N_{\text{NHB},b})\Omega_{w,h}(N_{\text{NB},h},N_{\text{NHB},h}) \times \Omega_{w,p}(N_{\text{HB},p},N_{\text{NHB},p})\Omega_{w,u}(N_u). \quad (9)$$

Note that the values of $N_{\text{HB},b}$, $N_{\text{HB},h}$, and $N_{\text{HB},p}$ can vary independently of each other because the orientations of bonding arms on an individual water molecule fluctuate independently.

Thus, for a protein configurational state given by $N_{\text{pairs},h}$, $N_{\text{pairs},p}$, and N_u, we can determine the upper and lower limits of the values of the variables $N_{\text{HB},b}$, $N_{\text{NHB},b}$, $N_{\text{HB},p}$, $N_{\text{NHB},p}$, $N_{\text{HB},h}$, and $N_{\text{NHB},h}$. Then, for every possible combination of the preceding variables which describe the orientational state of water, we use Eqs. (6) and (7) to calculate each of the components of the water orientational DOS. Applying Eq. (9) gives us the total orientational DOS for water for each of the possible specifications of water's hydrogen bonding state associated with a protein configurational state. Adapting Eq. (4) to our protein and water system allows for the computation of the total DOS:

$$\Omega_t(\xi_w;\xi_i) = \Omega_{\text{prot}}(\xi_i)\Omega_{w,o}(\xi_w;\xi_i)\Omega_{w,c}. \quad (10)$$

Repeating the procedure for all possible protein configurational states (or all possible values of the variables in ξ_i) yields the total DOS for all configurational and orientational states of the water and protein. This quantity can be used to calculate the thermodynamics of the system.

In order to extract the protein properties from the simulation data, we must convert the DOS into more useful quantities. Because there are fluctuations in both internal energy and volume in the simulation, we can reweight the DOS in the isobaric-isothermal ensemble. For a given pressure P and temperature T, the probability of a state j, specified by the water and protein degrees of freedom included in ξ_w and ξ_i, is

$$p_j(P,T) = \frac{\Omega_t(\xi_w;\xi_i)e^{-\beta U(\xi_w,\xi_i)-\beta PV(\xi_w,\xi_i)}}{\Delta(P,T)}, \quad (11)$$

where $\beta=1/k_BT$ and Δ is the isobaric-isothermal partition function. The internal energy and volume can be calculated from ξ_w and ξ_i for each state using Eqs. (2) and (1). Because we do not specify the values of the model parameters during the determination of the total DOS, the simulation data can be analyzed for any parameter set. This is an important advantage of our method.

The protein native state is identified as the configuration in which the protein is maximally compact and has the smallest surface area of hydrophobic groups exposed to the solvent. In the native state configuration, the protein has formed a core of hydrophobic monomers, with polar monomers preferentially residing at the surface of the protein. The change in the free energy upon folding ΔG can be calculated from the probability of occupying the native state p_n by using the equilibrium relation of the two-state model of protein folding

$$\Delta G(P,T) = G_{\text{native}} - G_{\text{denatured}} = RT\ln\left[\frac{1-p_n(P,T)}{p_n(P,T)}\right]. \quad (12)$$

The transition between the native and denatured states occurs when $\Delta G(P,T)=0$ or, equivalently, when $p_n(P,T)=0.5$.

B. Sequence pattern analysis

Inspired by the experimental studies of binary patterning in secondary structure,[25,26] we investigate which patterns of H's and P's may confer stability on heteropolymer sequences in our model. To accomplish this task, we adapt an existing technique from genomic analysis that is used for finding conserved DNA sequences.[59] We apply this analysis to determine if the distributions of various patterns in very stable sequences are random. We begin with the hypothesis that there is no correlation between sequence and stability. Therefore, we assume that any variation in the appearance of patterns in very stable sequences that departs from what is expected from a randomly generated set of sequences is due

entirely to random chance. The alternative hypothesis is that the appearance (or absence) of certain patterns in very stable sequences is indeed significant, and that some correlations between those patterns and protein stability exist. In the explanation that follows, we demonstrate how we perform the statistical analysis to determine that certain patterns are indeed important in very stable sequences.

To begin, we must quantify stability in a way that can be simply measured from the simulation results. Three natural measures of protein stability are evident in the P-T phase diagram. One option is the thermal stability ΔT, which we define as the range of temperatures at pressure $P=0$ where the protein is in its native configuration. This is the difference between the thermal denaturation temperature T_H and the cold denaturation temperature T_C, or $\Delta T=T_H-T_C$. Another option is the pressure stability P_{max}, which is quantified by the maximum stable pressure of the native protein. Finally, the aggregate stability A is defined as the area enclosed by the protein native state phase boundary and the x axis in Fig. 1. To determine which sequences are "very stable," we set a threshold value of one of these metrics (i.e., $\Delta T^\dagger$ for thermal stability). Sequences with values of ΔT greater than $\Delta T^\dagger$ are included in the subset of very stable sequences, while the remainder are excluded. In lieu of setting arbitrary values for the threshold parameters, $\Delta T^\dagger$, $P^\dagger_{max}$, and $A^\dagger$ are selected such that the subset of very stable sequences always contains the top 10% of sequences simulated at a given protein size and composition. For example, for the set of 16-mers with 50% hydrophobicity, the subset of very stable sequences includes the $n_{sample}=64$ sequences with the highest ΔT values out of the full set of 644 sequences simulated.

We then examine the frequencies of each pattern of H's and P's to determine whether they are more or less common in very stable sequences than expected by random chance. For a particular pattern, such as PHPP, we scan among the subset of very stable sequences to count the average frequency of the pattern $\bar{x}$. For larger proteins ($N_{Mon}>25$), we would compare the very stable sequences to the properties of a random sample of protein sequences with the same size and composition. However, for the 16- and 20-mers studied here, it is possible to enumerate the complete population of sequences with a given size and overall composition. We then calculate the mean (μ) and standard deviation (σ) of the frequency of a particular pattern in the full population.

If the frequency of a pattern in the population of all sequences is normally distributed, we can apply a standard Z-test to determine if deviations from random chance in the subset of very stable sequences are statistically significant. This procedure mirrors the conventional method for a Z-test, as described in standard statistics textbooks.[60] The Z-score of a pattern is calculated as $Z=(\bar{x}-\mu)/(\sigma/\sqrt{n_{sample}})$, and is compared to the standard normal probability distribution to get a p-value. This p-value is the probability that the deviation of the pattern frequency in the very stable sequences from the population is due to random chance. For very low p-values, our hypothesis that any deviation is due to random chance will break down, and we can conclude that the appearance or the absence of the pattern is statistically significant. A cutoff of $p=0.025$ is used to determine which patterns are statistically significant.

In practice, the Z-test is applied to distributions that are either normal or at least approximately normal. The approximation of normality holds well for the distributions of small pattern sizes examined here, but begins to break down for larger patterns. Patterns of five monomers are at the threshold of what can be tested for these smaller proteins due to the limited frequency of observing larger patterns in the sequences simulated. Visual inspection of the population distribution of patterns of more than five monomers show that they are irregular and do not resemble normal distributions. To check that the results presented here are not heavily affected by the assumption of a normal distribution, we also used a nonparametric test, the Wilcoxon rank-sum test.[60] The rank-sum test does not require an assumption about the shape of the population distribution, and any inconsistencies between the results of the Z-tests and the rank-sum tests would indicate shortcomings in our analysis. However, for all the statistical tests applied in this study, both methods come to the same conclusions about statistically significant patterns

TABLE I. Example 16-mer and 20-mer sequences with their thermal stability (ΔT), pressure stability (P_{max}), and aggregate stability (A) given in dimensionless units for model parameters $J_H/J=0.05$, $\lambda_b=\lambda_p=3$, $\lambda_h=0$, $q=70$, and $\Delta v/v_0=0.35$. The ranks listed next to each stability measure are the position of that sequence when ranked among sequences of the same size and composition in order of most stable to least stable for that stability measure.

No.	Sequence	ΔT	Rank	P_{max}	Rank	A	Rank
16.1	H_8P_8	0.080 77	1	2.101	2	0.087 79	2
16.2	$P_5HPH_3PH_3PH$	0.079 22	2	2.105	1	0.088 62	1
16.3	$PH_2P_2HPHP_3H_3PH$	0.078 02	3	2.015	5	0.083 67	3
16.4	$(P_2H_2)_4$	0.076 90	7	1.982	8	0.080 89	8
16.5	$HPHPH_2PHP_2HPHPHP$	0.058 20	354	2.050	3	0.068 64	123
16.6	$P_4H_8P_4$	0.040 86	493	0.792	527	0.018 73	522
20.1	$H_{10}P_{10}$	0.074 54	1	1.905	1	0.075 54	1
20.2	$P_5H_2P_5H_8$	0.073 15	2	1.859	2	0.072 46	2
20.3	$PHP_2HP_7H_8$	0.072 73	3	1.853	4	0.072 20	3
20.4	$PHP_4H_4P_2H_2P_3H_3$	0.072 71	4	1.854	3	0.071 86	4
20.5	$P_4H_3PHP_3H_4P_2H_2$	0.069 56	6	1.755	8	0.065 74	6

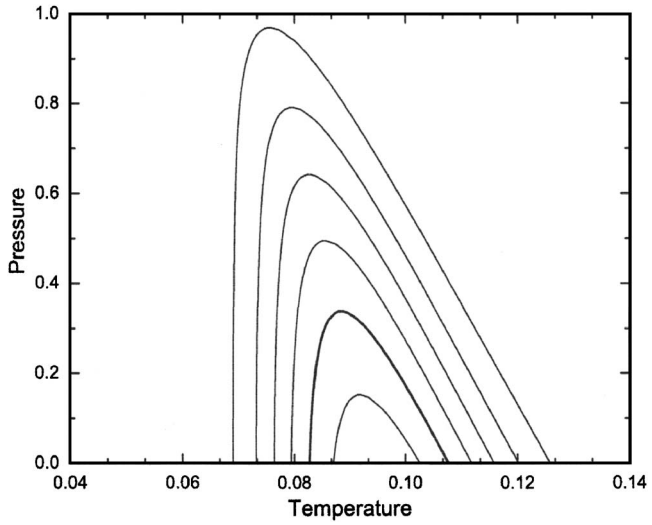

FIG. 3. The phase diagram of a 16-mer heteropolymer denoted 16.4, with sequence $(P_2H_2)_4$, for parameter values of $J_H/J=0.1$, $\lambda_b=\lambda_p=2$, $\lambda_h=0$, $q=50$, $\Delta v/v_0=0.35$. The inner line marks the region within which the probability of observing the native state is 60% or greater. In the same way, the other lines mark the regions within which the native state probabilities are greater than 50% (bold), 40%, 30%, 20%, and 10% (outermost).

over 90% of the time, validating the use of the Z-test. In Sec. IV, the results of the pattern analysis include only the Z-test data, while complete information on both the Z-test and the Wilcoxon rank-sum test are included in the Supplemental Materials.[61]

IV. RESULTS

A. General model properties

Figure 3 shows the phase diagram for a 16-mer heteropolymer with sequence $(P_2H_2)_4$ (referred to as sequence 16.4). Table I collects the sequences and properties of the heteropolymers discussed in this section, including sequence 16.4. The phase diagram in Fig. 3 is representative of a heteropolymer that folds into a stable native state at positive pressures. The native state structure of sequence 16.4 stable at low pressures and intermediate temperatures is depicted in Fig. 4(a). As described in Sec. III A, the model protein in its native state has a compact configuration with a hydrophobic core. This structure minimizes the exposed surface area of the hydrophobic monomers, reducing the number of interfacial hydrogen bonds around hydrophobic monomers that must pay an entropic cost to form. Increasing temperature causes the protein to unfold, bringing more hydrophobic monomers into contact with the solvent. At higher temperature, the thermal energy can overcome the entropic cost of forming entropically penalized hydrogen bonds next to hydrophobic monomers. There is a large ensemble of conformations sampled in the thermally denatured state, and several of these configurations are shown in Fig. 4(c) for sequence 16.4. Lower temperatures favor the formation of a cold-denatured state, in which the protein remains compact but exposes additional hydrophobic monomers to the solvent in order to form more of the enthalpically favorable hydrogen bonds around hydrophobic monomers. Figure 4(b) shows the configuration of the protein cold-denatured state, in which the protein's hydrophobic core is disrupted. Experimentally, cold-denatured states are often characterized as collapsed configurations with more structure than the thermally denatured states, which constitute an ensemble of unfolded and partially folded configurations.[62,63] Figures 4(b) and 4(c) show that the model protein approximates this behavior well through the ensemble of thermally denatured configurations and the compact cold-denatured configuration lacking a hydrophobic core.

Figure 3 also shows that the slope of the cold-denaturation curve is infinite at $P=0$. The cause becomes clear when the properties of the native and cold-denatured states are considered. Both structures are compact, and the water around them forms the same number of hydrogen bonds. Because the system volume, as defined in Eq. (1), is only a function of the number of hydrogen bonds, both the cold-denatured and native states have the same volume.

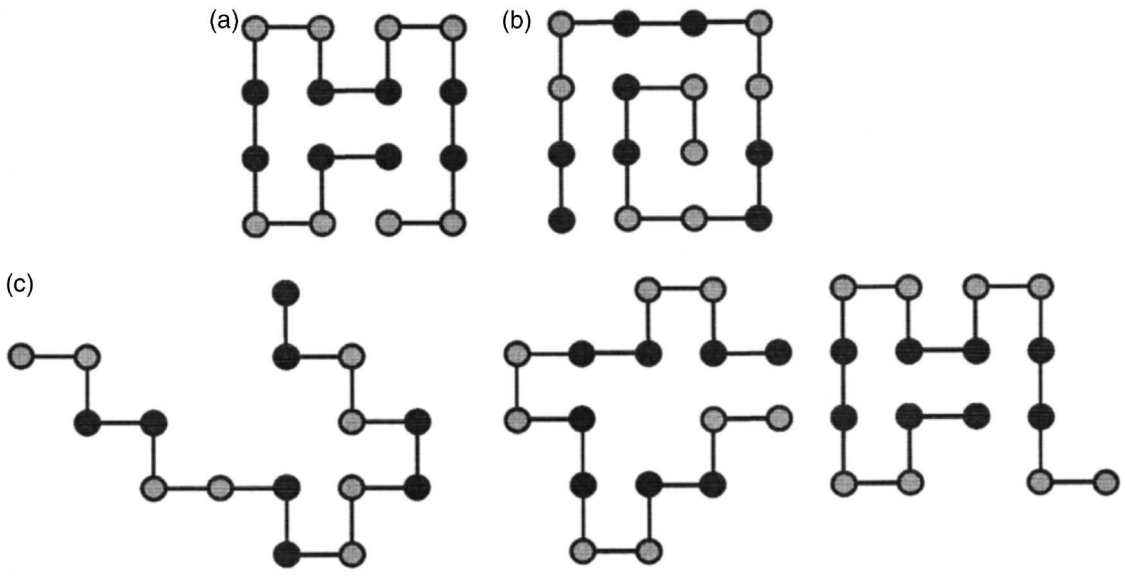

FIG. 4. Representative configurations for sequence 16.4 in the (a) native state, (b) cold-denatured state, and (c) thermally denatured ensemble of states.

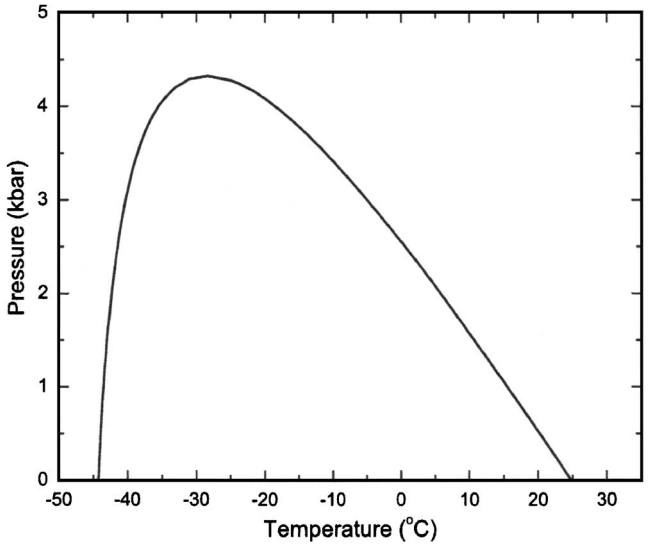

FIG. 5. Contour of 50% native state probability for sequence 16.4 with temperature and pressure converted into dimensional quantities using $J=23$ kJ/mol and $v_0=18$ cm^3/mol for parameter values $J_H/J=0.1$, $\lambda_b=\lambda_p=2$, $\lambda_h=0$, $q=50$, $\Delta v/v_0=0.35$.

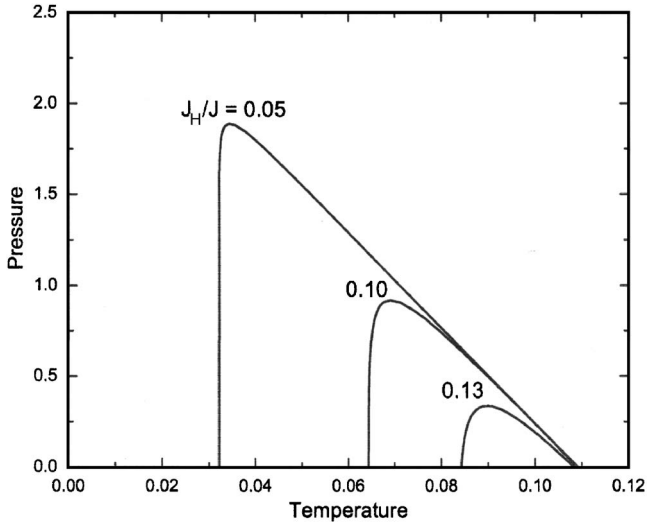

FIG. 6. Contours of 50% native state probability for sequence 16.4 for varying values of the enthalpic bonus J_H/J. The other model parameters remained constant at $\lambda_b=\lambda_p=3$, $\lambda_h=0$, $q=70$, and $\Delta v/v_0=0.35$.

Since these two states are the only significant states at $P=0$, the volume does not change upon cold-unfolding, or $\Delta V_u(T_C)=0$. The Clapeyron equation describes the slope of a phase boundary as

$$\frac{dP}{dT}=\frac{\Delta S}{\Delta V}, \tag{13}$$

where ΔS and ΔV are the changes in entropy and volume upon undergoing a phase transition. Applying the Clapeyron equation to the cold-unfolding transition in our model protein, we observe that the slope of cold-denaturation curve will always be infinite since there is no difference in the volume of the cold-denatured and native states. The positive slope of the cold-denaturation curve at higher pressures reflects the fact that other noncompact denatured states are also observed at these conditions, leading to nonzero volume changes upon unfolding. The slope remains very high even at pressures approaching $P_{\max}$, indicating that the volume difference between the native and denatured states is still very small and that the low-pressure cold-denatured state remains important at high pressures.

Using physically chosen scaling factors of 23 kJ/mol for the strength of a hydrogen bond (J) and 18 cm^3/mol for the molar volume of a lattice site ($v_0=V_0/N_{\text{sites}}$), we can convert the phase diagram of the model heteropolymer into physical units. Figure 5 shows the phase diagram of sequence 16.4 with potential parameters $\lambda_b=\lambda_p=2$, $\lambda_h=0$, $J_H/J=0.1$, and $\Delta v/v_0=0.35$. The model phase diagram has the same general shape as the phase diagram of Staphylococcal nuclease from Fig. 1, with minor differences in the ranges of stable temperatures and pressures. The comparison between the model protein and Staphyloccal nuclease is not based on a direct correspondence between their respective sequences or compositions. Rather, we emphasize that the model heteropolymers exhibit denaturation in the same regions of temperature and pressure, as experimentally observed in biological proteins. Even without incorporating effects such as internal protein hydrogen bonding, electrostatics, disulfide bonds or side-chain packing, the model protein retains the general thermodynamic characteristics of real proteins with respect to pressure and temperature.

We vary the parameters determining the strength of the enthalpic bonus and entropic penalty to examine in greater detail how the forces stabilizing the native state influence the shape of the phase diagram. Figure 6 shows the effect of varying the enthalpic bonus J_H on the phase diagram of sequence 16.4. The stability of the native state at low temperature decreases as J_H increases because the cold-denatured state is relatively more stabilized. Increasing J_H strengthens the hydrogen bonds solvating hydrophobic monomers, and the cold-denatured state exposes more hydrophobic monomers to the solvent than the native state. Above $J_H=0.14$, a compact native state with a hydrophobic core as shown in Fig. 4(a) is no longer stable at positive pressures. However, we find that a minimum value of $J_H \geqslant 0$ is necessary to form a stable native state, and without an enthalpic bonus, the protein does not fold into the native state.

The effect of varying the entropic penalty for forming hydrogen bonds around hydrophobic monomers is shown in Fig. 7. We tune the strength of the entropic penalty by changing the values of λ_b and λ_p while fixing the value of λ_h to zero. By increasing λ_b and λ_p relative to λ_h, we increase the entropy of hydrogen bonds formed in bulk and around polar monomers relative to the entropy of hydrogen bonds formed around hydrophobic monomers. q also varies so that the fraction of total orientation pairs suitable for hydrogen bonding in bulk, $(2\lambda_b+1)q/q^2$ remains fixed. We observe that an entropic penalty is necessary to stabilize the native state relative to other protein configurations. Figure 7 shows that for a minimal entropic penalty of $\lambda_b=\lambda_p=1$, we observe a stable native state at positive pressures. Increasing the entropic penalty by increasing λ_b and λ_p further stabilizes the protein native state at the expense of the cold-denatured state. As noted above, the native state exposes fewer hydrophobic

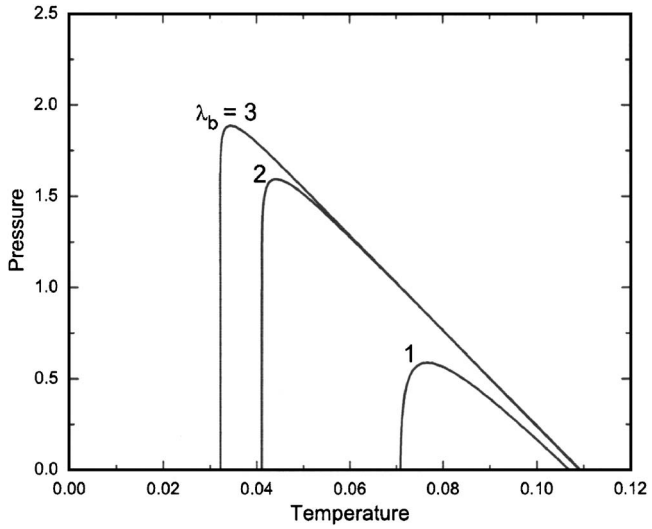

FIG. 7. Contours of 50% native state probability for sequence 16.4 for varying values of the relative entropic penalty for hydrogen bonding around hydrophobic monomers. The parameter values used were $\lambda_h=0$ and changing $\lambda_b(=\lambda_p)$. To maintain the same bulk water thermodynamics, the total number of water orientations q increases so that the fraction of bonding orientations for a pair of bonding arms [i.e., $(2\lambda_b+1)q/q^2=(2\lambda_b+1)/q$] is kept constant at 0.1. The other model parameters remained constant at $J_H/J=0.05$, $\Delta v/v_0=0.35$.

monomers to the solvent than the cold-denatured state. Increasing the entropic penalty reduces the entropy of the cold-denatured state more than that of the native state, which pushes the transition between the two states to a lower temperature.

Figures 6 and 7 confirm that the model parameters have the same basic effect on the model heteropolymer as on the model hydrophobic homopolymer previously studied.[50] It is significant that while the homopolymer model showed cold-denaturation at ambient pressures for a limited range of values of J_H and λ_b, the heteropolymer model shows a much wider range of model parameters that display this phenomenon. In the heteropolymer model, as long as $J_H>0$ and $\lambda_b,\lambda_p>\lambda_h$, the protein will cold-unfold at ambient pressure. It is a novel feature of the heteropolymer model that if conditions favor a stable native state at ambient pressure, then a lower-enthalpy ground state structure also exists, which is stable at lower temperatures. From consideration of the structures in Fig. 4, we can see that if there is a native state configuration in which a minimum number of hydrophobic monomers are exposed, then there must also be a cold-denatured configuration in which the maximum number of hydrophobic monomers are exposed. In contrast, the protein unfolds into a fully extended configuration in the cold-denatured state in the homopolymer model.[50]

The thermodynamic properties of sequence 16.4 provide a general description of the stability of a characteristic sequence. To probe the effect of sequence on stability, we performed simulations of a large number of randomly generated heteropolymer sequences: 10% of all possible unique sequences of 16-mers of 37.5%, 50%, and 62.5% H composition (6, 8, and 10 H monomers, respectively); and 1% of all possible unique sequences of 20-mers of 50% composition. We limit the investigation to unique sequences, excluding

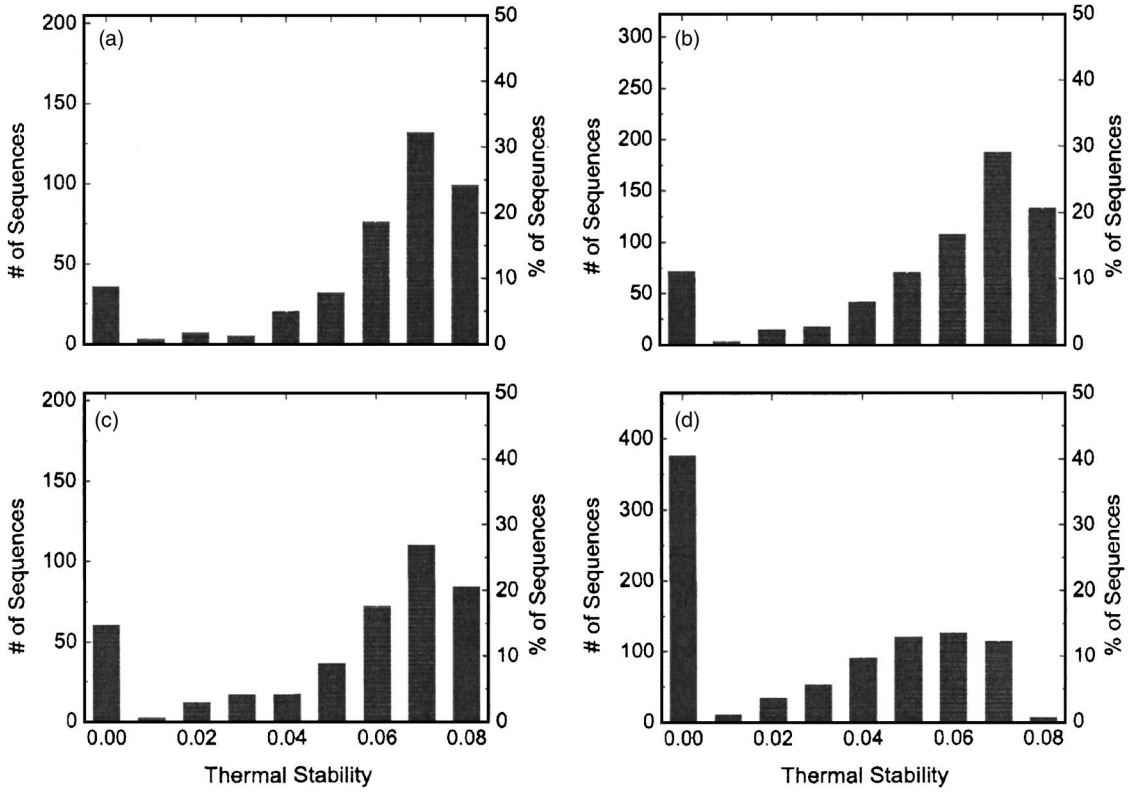

FIG. 8. Distributions of the range of thermal stability ΔT of randomly generated sequences for four different sets of sizes and H composition: (a) 16-mers at 37.5%, (b) 16-mers at 50%, (c) 16-mers at 62.5%, and (d) 20-mers at 50%. The number of sequences with thermal stability in the interval $0<\Delta T<0.01$ in dimensionless units is shown by the height of the bar marked 0.01. The left axis shows the total number of sequences in each interval of thermal stability, while the right axis shows that number relative to the total number of simulated sequences in that set. The model parameters values are $J_H=0.05$, $\lambda_b=\lambda_p=3$, $\lambda_h=0$, $q=70$, and $\Delta v/v_0=0.35$.

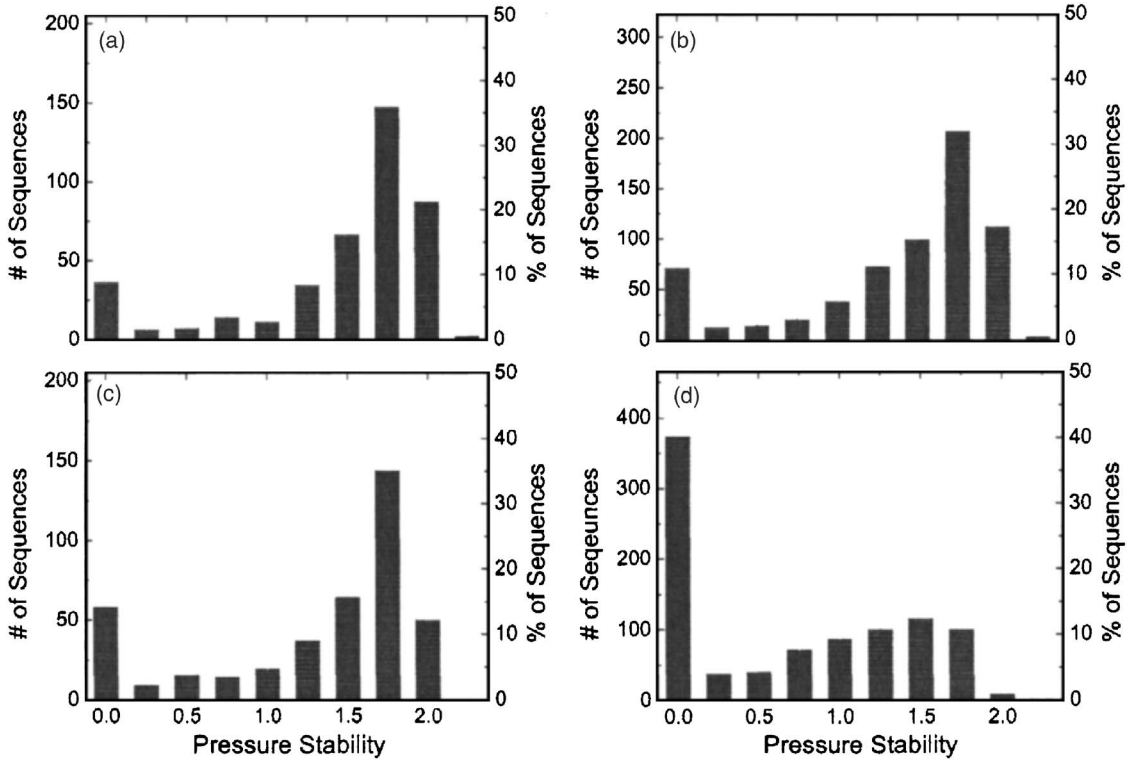

FIG. 9. Distributions of the pressure stability P_{max} of randomly generated sequences for four different sets of sizes and H composition: (a) 16-mers at 37.5%, (b) 16-mers at 50%, (c) 16-mers at 62.5%, and (d) 20-mers at 50%. The model parameter values are $J_H=0.05$, $\lambda_b=\lambda_p=3$, $\lambda_h=0$, $q=70$, and $\Delta v/v_0=0.35$.

sequences that are reflections of each other. It is impractical to simulate all of the sequences in each set, as the total number of unique 16-mers with 37.5%, 50%, and 62.5% composition is 4032, 6470, and 4032, respectively, and there are 92 504 unique 20-mers of 50% hydrophobicity. For these sampled subsets, the distributions of the thermal stabilities of the simulated proteins are given in Fig. 8, while the distributions of the pressure stabilities are given in Fig. 9. In all cases, the sequences show a bimodal distribution in stability, with a large number of unstable sequences and a large number of stable sequences distributed around a comparatively high mean stability. These distributions demonstrate that stability varies greatly depending on its sequence, even among sequences with the same size and composition. Figures 8 and 9 also show that the fraction of stable sequences decreases upon increasing from 16 to 20 protein monomers. However, the distributions do not clearly show the dependence of stability on composition.

Table II gives the average values of several measures of stability among each of the four sets of simulated sequences. The 16-mers demonstrate that native state stability decreases with increasing hydrophobicity for all of the measures. With additional hydrophobic monomers, the protein cannot as effectively isolate the hydrophobic monomers in the core in the native state configuration and instead exposes more of them to water. Thus, proteins with higher average hydrophobicity have a reduced native state enthalpy and entropy because the additional exposed hydrophobic monomers create more solvating hydrogen bonds that are stronger, but have fewer available hydrogen bonding orientations. However, the enthalpy of the cold-denatured state decreases more than that of the native state with additional H monomers, reducing the native state's stability to low temperature. The lower entropy of the native state weakens its stability to high temperature, since the entropy of the thermally denatured states is not similarly reduced. The entropic cost of these additional H monomers does not equally affect the thermally denatured states and the native state because some of the solvating hydrogen bonds subject to the entropic cost are broken in the thermally denatured states. The net effect is that the native state stability is reduced at high and low temperatures through two different mechanisms. In contrast, proteins with more polar monomers have a higher average stability because water molecules can behave around polar monomers as

TABLE II. Average values of the properties of large sets of simulated sequences for model parameters $J_H/J=0.05$, $\lambda_b=\lambda_p=3$, $\lambda_h=0$, $q=70$, and $\Delta v/v_0=0.35$. %H is the percent hydrophobicity of the set of sequences.

N_{Mon}	%H	T_H	T_C	ΔT	P_{max}	A
16	37.5	0.092 02	0.037 19	0.054 83	1.353	0.047 66
16	50.0	0.089 40	0.037 67	0.051 74	1.262	0.043 64
16	62.5	0.086 12	0.036 83	0.049 29	1.198	0.041 34
20	50.0	0.057 09	0.030 07	0.027 02	0.628	0.018 51

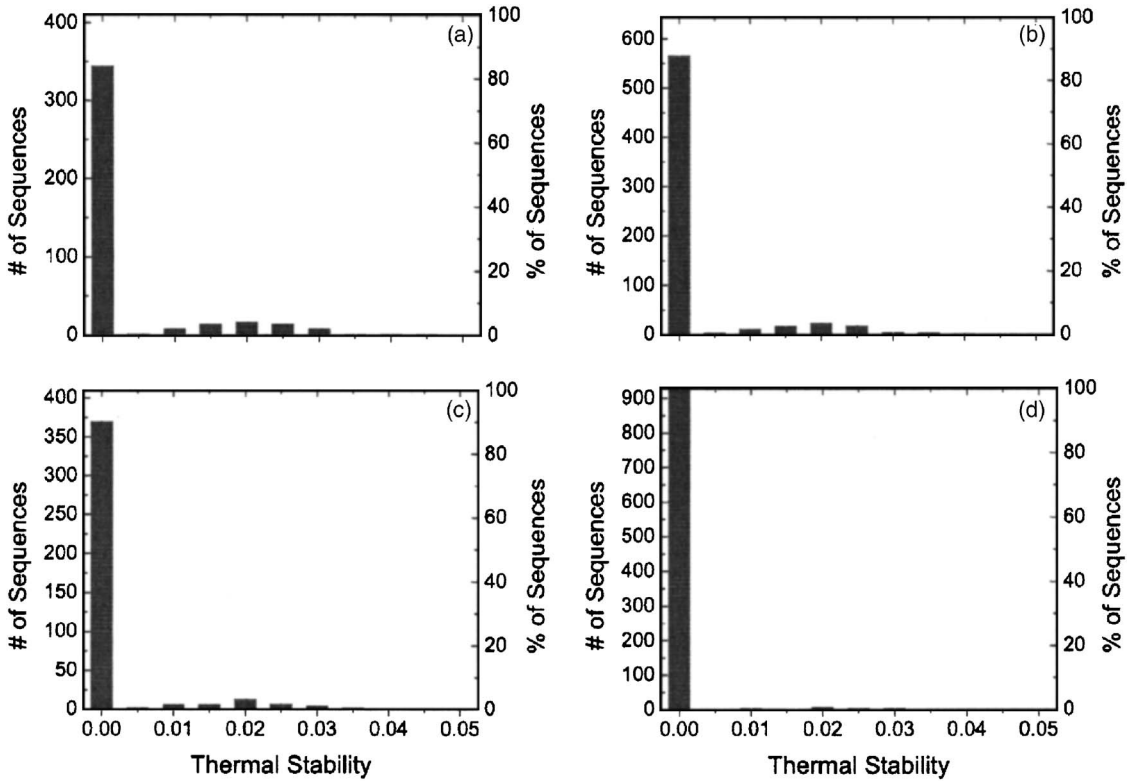

FIG. 10. Distributions of the range of thermal stability ΔT of randomly generated sequences for four different sets of sizes and H composition: (a) 16-mers at 37.5%, (b) 16-mers at 50%, (c) 16-mers at 62.5%, and (d) 20-mers at 50%. The model parameter values are $J_{\mathrm{H}}=0.1$, $\lambda_b=\lambda_p=2$, $\lambda_h=0$, $q=50$, and $\Delta v/v_0=0.35$.

they do in bulk. Table II also confirms that 20-mers are less stable than the 16-mers of the same composition. This indicates that the destabilizing effect of adding two hydrophobic monomers is not balanced by the addition of two polar monomers, which are effectively neutral in their interaction with water. It is likely that this destabilizing trend would continue to larger protein sizes, given the lack of any additional protein-protein stabilizing interactions in the present model.

Experimentally, only a small fraction of randomly generated sequences are expected to fold into functional native states.[3,13] In the case of our 20-mers, 60% of the sequences fold into stable native states at positive pressures, a majority of the sequences. However, the fraction of sequences which fold successfully depends greatly on the choice of parameters. Figure 10 shows that for parameter values $\Delta v/v_0=0.35$, $J_{\mathrm{H}}/J=0.1$, $\lambda_b=\lambda_p=2$, $\lambda_h=0$, and $q=50$, a much smaller fraction of proteins fold stably. This follows logically from Figs. 6 and 7, which show that increasing the enthalpic bonus and reducing the entropic penalty destabilize the na-

TABLE III. Statistically significant patterns between two and five monomers in length from the set of very stable 16-mers with 50% composition. The frequent patterns appear more often than expected by random chance in the top 10% most stable simulated sequences, while the infrequent patterns appear less often than expected by random chance.

ΔT		$P_{\max}$		A	
Frequent	Infrequent	Frequent	Infrequent	Frequent	Infrequent
HH	HP	None	HHPHP	HH	HP
PP	HPH			PP	PHP
HHH	PHP			PPP	HPHP
PPP	HPHP			HHHH	HHPHP
HHHH	HPPH			PPPP	HPHPH
HPPP	HHPHP			HHHHP	PHPHP
PPPP	HPHPH			PPPPP	
HHHHH	HPHPP				
HHHHP	HPPHP				
HPPPP	PHPHP				
PHPPP					
PPPPP					

TABLE IV. Statistically significant patterns from the set of very stable 16-mers with 37.5% composition.

ΔT		P_{max}		A	
Frequent	Infrequent	Frequent	Infrequent	Frequent	Infrequent
HH	HP	HH	HP	HH	HP
PP	HPH	PP	HPHP	PP	HPP
HHH	HPP	HHH	HPHPP	HHH	PHP
PPP	PHP	PPP	PHPHP	HHP	HPHP
HHHH	HPHP	HHHH		PPP	HPPH
HHHP	HPPH	HHHP		HHHH	PHPP
PPPP	PHPP	PPPP		HHPP	HPHPP
HHHHH	HPHPP	HHHHH		PPPP	HPPHP
HHHHP	HPPHP	HHHHP		HHHHH	PHPHP
HHHPP	PHPHP	HHHPH		HHHHP	PHPPP
HHPPP	PHPPP	PPPPP		HHPPP	PPHPP
HPPPP				PPPPP	
PPPPP					

tive state. Using this parameter set, only 5 out of the 928 simulated 20-mers have stable native states. These five 20-mers that are stable for this choice of potential parameters remain the five most stable sequences for other values of parameters, including those used in Figs. 8 and 9, and the relative stability of sequences in a set does not depend on the choice of parameters. In the following calculations, we will continue to use the parameter set employed in Figs. 8 and 9 because of the larger number of stable sequences available under those conditions. This allows us to better discriminate between the properties of the very stable sequences and the less stable sequences.

Sequences 16.1–16.5 and 20.1–20.5 listed in Table I are some of the most stable sequences in the sets of 16-mers and 20-mers with 50% composition. Table I shows the thermal, pressure, and aggregate stability of these sequences, along with the rank of each sequence when compared to other sequences of the same size and composition. Interestingly, the most stable sequences of each size were H_8P_8 and $H_{10}P_{10}$, diblocks of H's and P's. These were not randomly generated sequences but instead were specifically selected to see how they compared to the randomly generated set. However, in both cases, the diblocks were only slightly more stable than the most stable randomly generated sequences of those sizes, sequences 16.2 and 20.2. The factors influencing the model protein's stability are not obvious from inspection of Table I. While blocks of H's and P's seem to be important contributors to protein stability, it is not clear by visual inspection of Table I that this is the only contributor. In the next section, we apply the statistical analysis explained in Sec. III B to probe the connection between sequence and stability more quantitatively.

B. Pattern analysis

For reference, the Supplemental Materials[61] contain the complete pattern analysis results, including the Z-scores and *p*-values of each statistical test applied to every pattern. Tables III–VI present the final results of the pattern analysis on each of the four sets of simulated sequences. Each table lists the patterns which are very frequent and very infrequent in the subset of very stable sequences, for each of the three stability metrics, ΔT, P_{max}, and A. The pattern analysis of thermal stability demonstrates that blocks of H's or P's are more frequent in very thermally stable sequences. This confirms the previous observation that the diblock sequences 16.1 and 20.1 are among the sequences with the greatest stability, and that there is some feature inherent in long blocks which favors greater protein stability. Alternations in sequence such as the pattern HPHPH are very infrequent in the very thermally stable sequences. However, Table III shows that the significant patterns observed vary depending

TABLE V. Statistically significant patterns from the set of very stable 16-mers with 62.5% composition.

ΔT		P_{max}		A	
Frequent	Infrequent	Frequent	Infrequent	Frequent	Infrequent
PP	HP	PP	HP	HH	HP
PPP	HPH	PPP	HPH	PP	HPH
HHHH	PHP	HHHH	PHP	HHH	PHP
PHHP	HHHP	PPPP	HHHP	PPP	HHHP
PPPP	HPHP	HHHHH	HPHP	HHHH	HPHP
HHHHH	HHHPH	HPPPP	HHPPH	PPPP	HHHPH
HPHHP	HHPHP	PPPPP	HHPHP	HHHHH	HHPHP
HPPPP	HPHPH		HPHPH	HPPPP	HPHPH
PPPPP			PHHHP	PPPPP	PHHHP

TABLE VI. Statistically significant patterns from the set of very stable 20-mers with 50% composition.

ΔT		P_{max}		A	
Frequent	Infrequent	Frequent	Infrequent	Frequent	Infrequent
HH	HP	HH	HP	HH	HP
PP	HPH	PP	HPH	PP	HHP
HHH	PHP	HHH	HHPH	HHH	HPH
PPP	HHPH	PPP	HPHP	PPP	PHP
HHHH	HPHP	HHHH	HHPHP	HHHH	HHPH
HPPP	HHPHP	HPPP	HPHHP	HPPP	HPHP
PPPP	HPHHP	HHHHH	HPHPH	PPPP	HHPHP
HHHHH	HPHPH	HHHHP	HPHPP	HHHHH	HPHHP
HHHHP	HPHPP	PPHPP	PHHHP	HHHHP	HPHPH
HHPPP		PPPPP		PPHPP	HPHPP
PPPPP				PPPPP	PHHHP

on the measure of stability used. For the set of 16-mers with 50% composition, the importance of specific patterns is clearer among the thermally stable sequences than among the pressure-stable sequences. 22 patterns are statistically significant in their appearance or absence in the thermally stable sequences, while only one pattern is significant in the pressure-stable sequences. This distinction is evident from Table I as well, where sequence 16.5 ranks third in pressure stability but 354th in thermal stability. The analysis gives no indication that any particular pattern is important for the pressure stability of these proteins, even patterns which differ from those observed in the thermally stable subset. Since the aggregate stability incorporates aspects of both pressure and temperature stability, the results of the patterns analysis for aggregate stability in Table III are intermediate between those for pressure and temperature stability.

The 16-mers with 50% composition have the sharpest contrast between the pattern analysis results for thermal- and pressure-stable proteins. Tables IV–VI show that the distinction breaks down at different compositions and different protein sizes. For these sets, there is a greater correspondence between thermally stable and pressure-stable proteins, evident in the percentage of the subset of most thermally stable sequences which also belong to the subset of most pressure stable sequences. Only 66% of the very thermally stable 16-mers of 50% hydrophobicity show this correspondence with the pressure-stable subset, while 86% of the very thermally stable 20-mers also belong to the pressure-stable subset. In these other cases, all measures of stability show that the very stable sequences have blocks of H's and P's more frequently than expected and have alternations of H's and P's less frequently than expected.

The reasons for the frequency of blocks of H's and P's and the lack of alternating patterns among stable sequences are apparent upon consideration of the model geometry and physics. The driving force for folding of the model protein is to limit exposure of hydrophobic monomers to the solvent through the formation of a hydrophobic core. It is difficult to form a hydrophobic core when the protein has long sections of alternating H's and P's because of the geometry of the 2D square lattice. For example, a fully alternating sequence, $(HP)_8$, lacks the protein properties characterizing the other heteropolymer sequences. At low pressures, it does fold into a compact configuration, but this structure does not have a hydrophobic core. Because there is no hydrophobic core to disrupt, this sequence does not have the distinction between the native and cold-denatured configurations of sequence 16.4 shown in Figs. 4(a) and 4(b). Therefore, the fully alternating sequence does not cold-unfold at low temperatures, and the non-native compact structure remains stable. This extreme example indicates that alternations of H's and P's reduce the differences between the native and cold-denatured protein configurations and weaken the stability of the native state. In contrast, long blocks of H's and P's allow for greater flexibility in the configurations which can form a hydrophobic core. This increases the difference between the number of hydrophobic monomers exposed in the native and cold-denatured states, improving the stability of the native state. This flexibility provided by blocks of H's and P's also increases the configurational degeneracy of the native state relative to that of denatured states, leading to a greater native state stability.

In practice, it is difficult to use the results of the pattern analysis alone to design new sequences. Aside from the diblock heteropolymers of sequences 16.1 and 20.1, it is difficult to rationally design a sequence that would be among the top 10% of stable sequences for any of the three stability parameters. From the pattern analysis results, we would expect sequence 16.6, $P_4H_8P_4$, to be among the most stable, given its long blocks of H's and P's. In fact, sequence 16.6 is in the bottom quartile of all sequences simulated with that size and composition. This demonstrates that while the pattern analysis shows general trends in the data, it does not provide simple absolute rules for designing new sequences.

As noted in Sec. I, significant patterns in naturally occurring proteins are heavily dependent on their location in the protein and their presence in the secondary structural elements. Because the model proteins do not reproduce the secondary structure of biological proteins, we should not expect a direct correspondence between the important patterns in our model and those in the naturally occurring proteins. Indeed, since the native conformations for model heteropolymers do not possess structural subunits resembling α-helices, we do not observe the patterns used to design, for example,

TABLE VII. Results of four generations of directed evolution beginning with an unstable initial 16-mer of 50% composition, for parameters $J_H/J=0.05$, $\lambda_b=\lambda_p=3$, $\lambda_h=0$, $q=70$, and $\Delta v/v_0=0.35$. The sequence, percent hydrophobicity (%H), and properties of the best mutant at each generation are given below. Three different selection criteria were used to determine the best mutant at each generation: the cold denaturation temperature T_C, the thermal denaturation temperature T_H, and the maximum stable pressure P_{max}. After two generations of mutations selecting for T_H, none of the mutants improved upon the previous generation's best sequence.

	Generation	Sequence	%H	T_C	T_H	P_{max}
	0	$PHPH_4PH_2P_3HP_2$	50	Unstable	Unstable	0
T_C	1	$PHPH_4PH_2P_2H_2P_2$	56.25	0.036 06	0.108 69	1.747 18
	2	$PHPH_4PH_2P_2HP_3$	50	0.033 15	0.106 83	1.827 78
	3	$PHPH_4PH_2P_2HPHP$	56.25	0.032 42	0.106 36	1.864 55
	4	$PH_6PH_2P_2HPHP$	62.5	0.032 32	0.108 71	1.881 36
T_H	1	$PHPH_4PH_2P_2H_2P_2$	56.25	0.036 06	0.108 69	1.747 18
	2	$H_2PH_4PH_2P_2H_2P_2$	62.5	0.035 85	0.108 86	1.757 27
P_{max}	1	$PHPH_4PH_2P_2H_2P_2$	56.25	0.036 06	0.108 69	1.747 18
	2	$PHPH_4PH_2P_2HP_3$	50	0.033 15	0.106 83	1.827 78
	3	$PH_6PH_2P_2HP_3$	56.25	0.032 63	0.108 95	1.866 44
	4	$PH_6PH_2P_2HPHP$	62.5	0.032 32	0.108 71	1.881 36

four-helix bundles. Furthermore, we would not expect the periodicity of H's and P's in helixlike structures on a 2D square lattice to favor the same patterns as observed in biological proteins. Some features of binary patterning from biological proteins are observed in the model heteropolymers, including the prevalence of long hydrophobic blocks in buried segments.[33] While model hydrophobic blocks do not belong to parallel β-sheet structures like in naturally occurring proteins, the role the of blocks of H's in stabilizing the hydrophobic core and the protein is the same. While the lack of alternating H's and P's in the present model heteropolymers is due to features of the model geometry, these patterns lead to the formation of ultrastable non-native states. This is similar, in principle, to the formation of non-native fibrils from the same patterns in biological proteins.[27]

C. Directed evolution

Directed evolution is experimentally used to optimize proteins for particular functions through successive generations of random mutation and selection for specific traits.[64] It is instructive to briefly explore an analog for the present model. Here, the method is adapted to optimize our model proteins by improving their stability with respect to one of the three properties: cold denaturation temperature T_C, heat denaturation temperature T_H, and maximum pressure stability P_{max}. The procedure begins with a starting sequence, and mutant sequences are then created by performing single point mutations at each monomer position. Thus, beginning with a 16-mer heteropolymer, there are 16 mutants, each with a mutated monomer at one of the positions along the chain. A point mutation is defined as switching the mutated monomer's hydrophobicity, from H to P or vice versa. Each of these mutant sequences are simulated to determine their pressure and temperature stabilities. At the end of one round, three sequences are selected, the most stable for each of the three different stability measures. A round of mutations followed by the selection of the three best mutants is referred to as a generation. The cycle of mutation and selection is repeated for up to four generations and the resulting optimized sequences and their properties are examined.

Table VII shows the results of multiple rounds of mutating a 16-mer, selecting for each of the three measures. Additional data for the properties of the mutants that were not selected for future rounds of mutation are available in the Supplemental Materials.[61] The directed evolution began with a sequence that did not fold into a stable native state at positive pressures, and only one point mutation was required to obtain a successful folder. Although the procedure called for the selection of three different mutants based on T_H, T_C, and P_{max}, in practice, the most stable mutant for one measure was often the same as for another measure. After the first round of mutations, the sequence with the lowest cold-denaturation temperature also had the highest thermal denaturation temperature and the highest pressure stability. However, the selection for T_H reached a dead-end after two generations because there were no single-monomer mutants that would improve the high-temperature stability of the sequence. By the end, the procedure selecting for T_H produced a different sequence than that produced by selecting for T_C and P_{max}. In fact, selecting for T_C produced the same final sequence yielded by selecting for P_{max}, although through different intermediate point mutations. The sequence produced after four rounds of optimizing for T_C and P_{max} would rank among the top 10% in stability of sequences at its size and composition, demonstrating that the procedure applied here works effectively to find very stable sequences.

Figure 11 compares the properties of all of the mutants in each generation by showing the effect of the directed evolution of the 16-mer from Table VII on P_{max}. The majority of the first-generation mutants were improvements since these changes produced 10 stable mutants out of 16 possible. In subsequent rounds, only a few mutants improved upon the previous generation's most stable sequence, and these improvements are small in contrast to the leap in pressure stability from the initial sequence to the first generation. Desta-

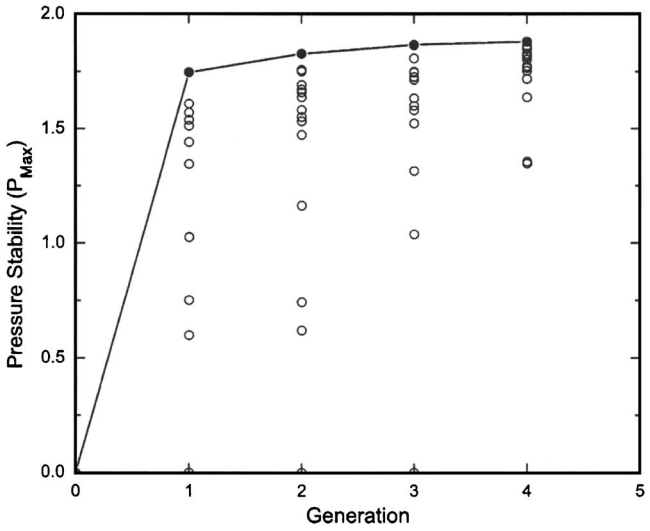

FIG. 11. Directed evolution of initial sequence $PHPH_4PH_2P_3HP_2$ through four generations of mutation and selection for optimal pressure stability. The black circles are the best mutants at each generation that are used for subsequent rounds of mutation, and the line shows the improvement in pressure stability of the selected sequence. The empty circles show the pressure stability of the other mutants not selected at each generation for comparison with the best mutant.

bilizing mutations are still possible after multiple rounds, since the second and third generations also produce unstable mutants. No unstable mutants are produced in the fourth generation, indicating that the mutation process by then has reached a point where it is difficult to drastically destabilize the protein.

Recall that Fig. 8 suggests that there are two major groups of sequences separated by a wide gap in stability. If this is the case, then the number of sequence modifications separating the two groups is small. In the case of the mutations studied here, only one monomer difference can change a sequence from the unstable population to the stable population. However, within the population of stable sequences the monomer changes separating sequences with average stability from those with the highest stability is likely complex. For example, in the process of mutating and selecting for P_{max} in Table VII, at the third generation we obtained a sequence with a higher T_H than the sequence yielded by selecting directly for T_H. This demonstrates that the optimization of a single parameter by selecting the best mutant at each generation does not necessarily provide a direct path to the optimal sequence.

Table VIII shows the results of another directed evolution of a 20-mer, beginning with 50% hydrophobicity. The initial sequence is again unstable, and only one monomer change is required to create a sequence with a stable native state. The directed evolution of the 20-mer also runs into a dead-end after three generations and no further mutations improve stability for any measure. Optimizing for T_C and P_{max} yielded the same final sequence and followed the same path of mutations, further demonstrating the connection between the two metrics in the model, as noted earlier. Aside from the cold-denatured state, the maximum stable pressure is also affected to a lesser degree by the denatured structures observed at very high pressures. These are likely the cause of the minor deviations between the paths for optimizing T_C and P_{max} in Table VII.

Upon inspection of the selected sequences, it appears that the directed evolution does elongate blocks of P's in the case of the 20-mer or H's in the case of the 16-mer. However, just as in the pattern analysis, this elongation of blocks is not universally favored for the most stable sequences. The final sequence at the end of the selection for P_{max} in Table VII has actually added an alternation HPH at the end of the chain instead of creating a block of 9 H's in its middle because the former mutation was more favorable. Moreover, while a fourth generation of mutants of the 20-mer could create a 10-monomer block of P's, this mutant is, in fact, less stable than the best sequence created after three generations.

V. CONCLUSIONS

We have presented a model for a heteropolymer in water that folds into a stable native state with a hydrophobic core, and captures the phenomena of cold, pressure, and thermal denaturation. The model proteins exhibit rich phase behavior with great variability in the stability of the native state with sequence. Sequences are generally either unstable or very stable, with few marginally stable sequences intermediate between the two groups. Analyzing the patterns significant in very stable sequences indicates that blocks of H's and P's are preferred, while alternations of H's and P's are disfavored. These observations have the character of guidelines or clues

TABLE VIII. Results of four generations of directed evolution beginning with an unstable 20-mer of 50% composition, for parameters $J_H/J=0.05$, $\lambda_b=\lambda_p=3$, $\lambda_h=0$, $q=70$, and $\Delta v/v_0=0.35$. The selection for T_C, and P_{max} proceeded along identical paths. After three generations, none of the subsequent mutations improved protein stability for any of the metrics.

	Generation	Sequence	%H	T_C	T_H	P_{max}
	0	$P_2HP_2H_4P_2H_2P_4H_3$	50	Unstable	Unstable	0
T_C, P_{max}	1	$P_2HP_2H_4P_2H_2P_5H_2$	45	0.033 96	0.104 59	0.617 00
	2	$P_2HP_2H_4P_2H_2P_6H$	40	0.033 24	0.104 75	0.628 93
	3	$P_2HP_2H_4P_2HP_7H$	35	0.032 80	0.105 06	0.636 51
T_H	1	$P_2HP_2H_4P_2H_2P_5H_2$	45	0.033 96	0.104 59	0.617 00
	2	$PH_2P_2H_4P_2H_2P_5H_2$	40	0.033 94	0.104 82	0.619 24
	3	$PH_2P_2H_4P_2HP_6H_2$	45	0.034 27	0.104 86	0.614 54

rather than design principles. Directed evolution of protein sequences showed that it is not difficult to stabilize and optimize an unstable sequences through random mutation, but that searching for sequences through single point mutations has limitations on finding the maximally stable sequence.

This model offers many opportunities for future studies of heteropolymers. The treatment of polar residues and how they interact with each other and with water is one aspect that remains unexplored. Also, the two directed evolution studies performed here barely scratch the surface of the exploration of the evolutionary landscape of the model proteins. Several investigators have examined the evolutionary landscape of compact lattice protein models by mutating the protein sequence while holding the backbone fixed, searching for the lowest energy state of a structure.[24,65–68] However, fixing the protein backbone in these studies restricts the protein from finding other configurations which have the same or lower energies than the designed configuration.[68] Instead of using energy as a substitute for native state stability, our model could be used to connect each sequence to one of the stability metrics used here $(\Delta T, P_{\max}, A)$ while allowing for a complete search of conformation space for the true native state. Although it would be computationally intensive, it is possible to simulate the complete set of 6470 unique 16-mers with 50% composition to determine each sequence's stability. The evolutionary landscape of the model 16-mer could then be exactly determined and analyzed to determine its shape and ruggedness. Peaks in the landscape would represent families of related, very stable sequences, and simulation results could determine how many peaks exist and how far apart they are. The sequence space could also be analyzed as a complex network with sequences representing nodes and a series of mutations representing the links between nodes.[69] This analysis could then be used to compute the average number of mutations required to proceed from an unstable sequence to the a very stable sequence. Furthermore, since simulation data can be reanalyzed for different choices of potential parameters with little additional effort, the effect of parameter choice on the shape of the sequence landscape could also be characterized. We also mention the possibility of designability studies exploring sequence and configuration space with the explicit water model as an interesting avenue for future research.[56]

There are differences between the significant patterns observed in our model and those observed in the naturally occurring proteins. These discrepancies exist because the heteropolymer model lacks secondary structure, and patterning of hydrophobic and hydrophilic amino acids in biological proteins is strongly influenced by the secondary structure. Secondary structure could be incorporated into the model given the appropriate lattice geometry and meaningful protein-protein interactions. Lattice approximations of α-helices have been identified for several different lattice geometries.[9,70,71] These models use torsional potentials or configuration-specific protein-protein interactions to favor the formation of these structures. Application of the pattern analysis developed here to a modified version of our heteropolymer model might yield stricter rules for designing proteins with greater stability and show more correspondence with biological proteins.

ACKNOWLEDGMENTS

We thank Shona Patel, Scott McAllister, and Christopher Bristow for many helpful discussions throughout the course of this investigation. P.G.D. and P.J.R. gratefully acknowledge the support of the National Science Foundation [Collaborative Research in Chemistry Grants CHE0404699 (P.G.D.) and CHE0404695 (P.J.R.)], the U.S. Department of Energy, Division of Chemical Sciences, Geosciences, and Biosciences, Office of Basic Energy Sciences, Grant No. DE-FG02-87ER13714 (P.G.D.), and the R.A. Welch Foundation [No. F0019 (P.J.R.)]. We also acknowledge the Texas Advanced Computing Center (TACC) at the University of Texas at Austin for high performance computing resources.

[1] R. Ravindra and R. Winter, ChemPhysChem **4**, 359 (2003).
[2] T. Creighton, *Proteins: Structures and Molecular Properties*, 3rd ed. (W.H. Freeman, New York, 1993).
[3] A. R. Davidson and R. T. Sauer, Proc. Natl. Acad. Sci. U.S.A. **91**, 2146 (1994).
[4] S. G. Kang and J. G. Saven, Curr. Opin. Chem. Biol. **11**, 329 (2007).
[5] P. S. Shah, G. K. Hom, S. A. Ross, J. K. Lassila, K. A. Crowhurst, and S. L. Mayo, J. Mol. Biol. **372**, 1 (2007).
[6] N. Go and H. Taketomi, Proc. Natl. Acad. Sci. U.S.A. **75**, 559 (1978).
[7] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
[8] J. Skolnick, A. Kolinski, and R. Yaris, Proc. Natl. Acad. Sci. U.S.A. **86**, 1229 (1989).
[9] P. Pokarowski, K. Droste, and A. Kolinski, J. Chem. Phys. **122**, 214915 (2005).
[10] A. Sali, E. Shakhnovich, and M. Karplus, Nature (London) **369**, 248 (1994).
[11] A. Sali, E. Shakhnovich, and M. Karplus, J. Mol. Biol. **235**, 1614 (1994).
[12] C. J. Camacho and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **90**, 6369 (1993).
[13] D. K. Klimov and D. Thirumalai, Proteins **26**, 411 (1996).
[14] N. D. Socci, J. N. Onuchic, and P. G. Wolynes, Proteins **32**, 136 (1998).
[15] J. D. Honeycutt and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **87**, 3526 (1990).
[16] J. D. Honeycutt and D. Thirumalai, Biopolymers **32**, 695 (1992).
[17] S. Brown, N. J. Fawzi, and T. Head-Gordon, Proc. Natl. Acad. Sci. U.S.A. **100**, 10712 (2003).
[18] D. K. Klimov, M. R. Betancourt, and D. Thirumalai, Folding Des. **3**, 481 (1998).
[19] J. Karanicolas and C. L. Brooks, Protein Sci. **11**, 2351 (2002).
[20] T. Head-Gordon and S. Brown, Curr. Opin. Struct. Biol. **13**, 160 (2003).
[21] B. Honig and F. E. Cohen, Folding Des. **1**, R17 (1996).
[22] G. D. Rose, A. R. Geselowitz, G. J. Lesser, R. H. Lee, and M. H. Zehfus, Science **229**, 834 (1985).
[23] R. Schiemann, M. Bachmann, and W. Janke, J. Chem. Phys. **122**, 114705 (2005).
[24] H. S. Chan and K. A. Dill, J. Chem. Phys. **95**, 3775 (1991).
[25] S. Kamtekar, J. M. Schiffer, H. Y. Xiong, J. M. Babik, and M. H. Hecht, Science **262**, 1680 (1993).
[26] H. Y. Xiong, B. L. Buckwalter, H. M. Shieh, and M. H. Hecht, Proc. Natl. Acad. Sci. U.S.A. **92**, 6349 (1995).
[27] M. W. West, W. X. Wang, J. Patterson, J. D. Mancias, J. R. Beasley, and M. H. Hecht, Proc. Natl. Acad. Sci. U.S.A. **96**, 11211 (1999).
[28] B. M. Broome and M. H. Hecht, J. Mol. Biol. **296**, 961 (2000).
[29] M. H. Hecht, A. Das, A. Go, L. H. Bradley, and Y. N. Wei, Protein Sci. **13**, 1711 (2004).
[30] D. A. Moffet, L. K. Certain, A. J. Smith, A. J. Kessel, K. A. Beckwith, and M. H. Hecht, J. Am. Chem. Soc. **122**, 7612 (2000).
[31] Y. N. Wei and M. H. Hecht, Protein Eng. Des. Sel. **17**, 67 (2004).
[32] P. T. Lansbury, Proc. Natl. Acad. Sci. U.S.A. **96**, 3342 (1999).
[33] Y. Mandel-Gutfreund and L. M. Gregoret, J. Mol. Biol. **323**, 453 (2002).
[34] R. Schwartz and J. King, Protein Sci. **15**, 102 (2006).

[35] R. Jaenicke, Eur. J. Biochem. **202**, 715 (1991).
[36] G. Feller, F. Payan, F. Theys, M. X. Qian, R. Haser, and C. Gerday, Eur. J. Biochem. **222**, 441 (1994).
[37] S. Chakravarty and R. Varadarajan, Biochemistry **41**, 8152 (2002).
[38] I. N. Berezovsky, W. W. Chen, P. J. Choi, and E. I. Shakhnovich, PLOS Comput. Biol. **1**, 322 (2005).
[39] A. Razvi and J. M. Scholtz, Protein Sci. **15**, 1569 (2006).
[40] R. Ladenstein and G. Antranikian, Adv. Biochem. Eng./Biotechnol. **61**, 37 (1998).
[41] N. J. Russell, Adv. Biochem. Eng./Biotechnol. **61**, 1 (1998).
[42] D. Georlette, V. Blaise, T. Collins, S. D'Amico, E. Gratia, A. Hoyoux, J. C. Marx, G. Sonan, G. Feller, and C. Gerday, FEMS Microbiol. Rev. **28**, 25 (2004).
[43] D. J. Hei and D. S. Clark, Appl. Environ. Microbiol. **60**, 932 (1994).
[44] F. Simonato, S. Campanaro, F. M. Lauro, A. Vezzi, M. D'Angelo, N. Vitulo, G. Valle, and D. H. Bartlett, J. Biotechnol. **126**, 11 (2006).
[45] S. A. Hawley, Biochemistry **10**, 2436 (1971).
[46] A. Zipp and W. Kauzmann, Biochemistry **12**, 4217 (1973).
[47] G. Panick, G. J. A. Vidugiris, R. Malessa, G. Rapp, R. Winter, and C. A. Royer, Biochemistry **38**, 4157 (1999).
[48] M. W. Lassalle, H. Yamada, and K. Akasaka, J. Mol. Biol. **298**, 293 (2000).
[49] H. Lesch, H. Stadlbauer, J. Friedrich, and J. M. Vanderkooi, Biophys. J. **82**, 1644 (2002).
[50] B. A. Patel, P. G. Debenedetti, F. H. Stillinger, and P. J. Rossky, Biophys. J. **93**, 4116 (2007).
[51] S. Sastry, P. G. Debenedetti, F. Sciortino, and H. E. Stanley, Phys. Rev. E **53**, 6144 (1996).
[52] L. P. N. Rebelo, P. G. Debenedetti, and S. Sastry, J. Chem. Phys. **109**, 626 (1998).
[53] H. S. Frank and M. W. Evans, J. Chem. Phys. **13**, 507 (1945).
[54] P. J. Rossky and D. A. Zichi, Faraday Symp. Chem. Soc. **17**, 69 (1982).
[55] K. A. T. Silverstein, A. D. J. Haymet, and K. A. Dill, J. Chem. Phys. **111**, 8000 (1999).
[56] H. Li, R. Helling, C. Tang, and N. Wingreen, Science **273**, 666 (1996).
[57] B. A. Patel, P. G. Debenedetti, and F. H. Stillinger, J. Phys. Chem. A **111**, 12651 (2007).
[58] F. G. Wang and D. P. Landau, Phys. Rev. Lett. **86**, 2050 (2001).
[59] O. Elemento and S. Tavazoie, Adv. Genome Biol. **6**, R18 (2005).
[60] D. Montgomery and G. Runger, *Applied Statistics and Probability for Engineers* (Wiley, New York, 1994).
[61] See EPAPS Document No. E-JCPSA6-128-001818 for Z and Wilcoxon test data, and for stability data on all sequences investigated in the directed evolution studies. For more information on EPAPS, see http://www.aip.org/pubservs/epaps.html.
[62] D. P. Nash and J. Jonas, Biochemistry **36**, 14375 (1997).
[63] D. P. Nash and J. Jonas, Biochem. Biophys. Res. Commun. **238**, 289 (1997).
[64] F. H. Arnold, Acc. Chem. Res. **31**, 125 (1998).
[65] D. J. Lipman and W. J. Wilbur, Proc. R. Soc. London, Ser. B **245**, 7 (1991).
[66] E. I. Shakhnovich, Folding Des. **3**, R45 (1998).
[67] Y. Xia and M. Levitt, Proc. Natl. Acad. Sci. U.S.A. **99**, 10382 (2002).
[68] R. Wroe, E. Bornberg-Bauer, and H. S. Chan, Biophys. J. **88**, 118 (2005).
[69] R. Albert and A. L. Barabasi, Rev. Mod. Phys. **74**, 47 (2002).
[70] P. Romiszowski and A. Sikorski, Physica A **336**, 187 (2004).
[71] Y. Y. Chen, Q. Zhang, and J. D. Ding, J. Chem. Phys. **124**, 184903 (2006).