Computational Characterization of the Sequence Landscape in Simple Protein Alphabets

M. Scott Shell, Pablo G. Debenedetti, and Athanassios Z. Panagiotopoulos
Department of Chemical Engineering, Princeton University, Princeton, NJ 08544

ABSTRACT **We characterize the "sequence landscapes" in several simple, heteropolymer models of proteins by examining their mutation properties. Using an efficient flat-histogram Monte Carlo search method, our approach involves determining the distribution in energy of all sequences of a given length when threaded through a common backbone. These calculations are performed for a number of Protein Data Bank structures using two variants of the 20-letter contact potential developed by Miyazawa and Jernigan [Miyazawa S, Jernigan WL. Macromolecules 1985;18:534], and the 2-monomer HP model of Lau and Dill [Lau KF, Dill KA. Macromolecules 1989;22:3986]. Our results indicate significant differences among the energy functions in terms of the "smoothness" of their landscapes. In particular, one of the Miyazawa-Jernigan contact potentials reveals unusual cooperative behavior among its species' interactions, resulting in what is essentially a set of phase transitions in sequence space. Our calculations suggest that model-specific features can have a profound effect on protein design algorithms, and our methods offer a number of ways by which sequence landscapes can be quantified. Proteins 2006;62:232–243.**

Key words: proteins; thermodynamics; statistical mechanics; landscapes; phase transitions; Monte Carlo; flat-histogram

INTRODUCTION

Proteins possess a number of distinctive properties with respect to synthetic polymers: they fold into well-defined structures, exhibit highly specific interactions with other molecules, and have often evolved to operate within very specific environmental conditions.[1] These attributes make proteins potentially able to be programmed or designed for specific technological ends. Indeed, one of the driving forces behind theoretical and computational efforts to model proteins is our need for a comprehensive methodology for protein engineering, since the *de novo* design of these molecules has far-reaching ramifications in applications ranging from materials assembly to targeted drug delivery.

Because of the prohibitive computational cost associated with studying a fully atomistic description of proteins, an important body of work has addressed the design problem using minimalist heteropolymer models of proteins.[2–5] In these coarse-grained models, each amino acid is represented by a single bead on a chain. A simple intra-bead energy function approximates the effective pairwise interactions experienced between various amino acids with the solvent degrees of freedom averaged out. Perhaps the simplest kind of energy function one can use in such models is a so-called contact potential, which assigns an amino acid-specific energy to pairs of residues that lie within some cutoff distance of each other (a typical value being 6.5 Å[6]). One of the most studied examples of this sort is the HP model of Lau and Dill in which there are only two types of monomers, hydrophobic and hydrophilic.[7] The HP model captures the bare features of hydrophobic collapse: hydrophobic monomers that are in contact make a negative contribution to the overall energy of the system, with no other contributions from hydrophobic-hydrophilic or hydrophilic-hydrophilic contacts.

Unsurprisingly, the particular alphabet of amino acids available for sequence design can have a significant effect on the number and kinds of protein structures that can be encoded. The HP model, for example, might be considered a poor design alphabet; the majority of HP sequences do not possess a unique, nondegenerate ground state.[8] Several authors suggest that by moving to larger alphabets, such as the more realistic 20-letter contact potential developed by Miyazawa and Jernigan,[6] the "designability" of conformations increases,[8,9] although exceptions have been noted.[10] A fundamental question that must be addressed is therefore the dependence of the collection of all possible sequences on the particular amino acid alphabet used. That is, how does the *sequence landscape* differ from one monomer set to another?

In the present work, we present a flexible algorithm[11] for the characterization of protein sequence landscapes, and we demonstrate its application to several simplified but routinely studied models of proteins. Specifically, we examine the Miyazawa-Jernigan (MJ) and HP interaction potentials as applied to selected Protein Data Bank (PDB)[12] structures. For each potential, our algorithm enables the precise determination of the distribution along any order

Grant sponsor: Fannie and John Hertz Foundation; Grant sponsor: Department of Energy; Grant numbers: DE-FG02-87ER13714 (to P.G.D.), DE-FG02-01ER15121 (to A.Z.P.)

Correspondence to: M. Scott Shell, Department of Chemical Engineering, Princeton University, Princeton, New Jersey 08544. E-mail: shell@princeton.edu

Received 16 May 2005; Revised 14 July 2005; Accepted 24 July 2005

Published online 11 November 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20714

parameter of all possible sequences of a given length. For this initial study, we choose energy as the order parameter of interest, so that we calculate the distribution in energy of all sequences of a particular alphabet when threaded through each PDB backbone structure. Use of the energy as the order parameter permits a connection to the equilibrium mutation properties of proteins, that is, the equilibrium ensemble of sequences existing at a given temperature for a fixed backbone conformation. Such mutation behavior is of relevance to protein design algorithms that make random mutations along a backbone structure, and thus our results help clarify the nature of the sequence landscape explored by these methods. We find that, for one version of the MJ model, the landscape is quite complex, possessing what amounts to first-order phase transitions in sequence space, which may lead to substantial hysteresis in design algorithms. For the other models investigated, we do not find such transitions or hysteresis. Thus our results indicate that the sequence landscape of these simple models can possess nontrivial differences that are important to consider in the development of protein design methods.

The goal of protein design is to create a molecule that performs a desired function. Since protein structure and function are closely coupled, this goal can be equivalently stated as that of identifying the amino acid sequences that fold into a target configuration. For the simple heteropolymer protein models of the kind discussed here, the design problem can therefore be framed in the following way: given the three-dimensional structure of the protein backbone, and hence the relative positions of all its "beads," find the sequences of amino acids that have the target configuration as their ground-state (i.e., the configuration with lowest energy). Rigorously speaking, this task requires an exhaustive search of both configurational and sequence spaces. For each trial sequence, the energies of all possible configurations must be evaluated to determine whether the target structure is indeed the ground state, making sequence design a formidable computational task. However, algorithms that focus on one aspect of this problem, the search of sequence space, have emerged.[13–21] One such method was introduced by Shakhnovich and Gutin,[13] and by Pande et al.[15] This approach, loosely termed "sequence annealing," consists of minimizing the energy of a given protein by allowing mutations in the primary sequence while holding the backbone fixed. These mutations are performed using a Monte Carlo scheme by which proposed modifications are accepted or rejected based on a Boltzmann criterion:

$$P_{\mathrm{acc}} = \min[1,\exp(-\Delta E/k_B T_{\mathrm{des}})] \qquad (1.1)$$

where P_{acc} is the probability with which a proposed mutation is accepted, ΔE is the difference in energy between the proposed and current sequence, k_B is Boltzmann's constant, and T_{des} is the so-called design temperature. T_{des} is essentially held at a low but finite value so as to permit a simulated annealing-type minimization of the sequence energy. This method entails the additional constraint that the composition of the protein is held fixed, that is, the total number of units of each amino acid type is constant throughout the mutations and only "swap" moves between pairs of residues are permitted. This is a necessary condition that helps ensure that the energy minimization procedure is equivalent to maximizing the probability that the target structure is the ground state of the sequence.

The sequence annealing algorithm is based on an approximate form for the protein free energy, motivated by principles of heteropolymer theory.[22] One limitation of the approach is that, prior to running the algorithm, one must select the fixed amino acid composition of the design procedure by some means. Improved algorithms have emerged, including a derivative by Shakhnovich and coworkers using the "Z-score" for ranking sequences by their ability to fold into target structures.[23] First introduced by Bowie et al.,[24] the Z-score measures the difference in energy between the target configuration and the average energy of all other compact structures, expressed in terms of the standard deviation of the latter. Thus, a very low (negative) Z-score indicates that the target configuration is situated at an energy far below that of all others, which increases the probability that it is the ground state of the system. By using the Z-scores of protein sequences instead of the energy in Eq. 1.1, the constraint of fixed composition in the Monte Carlo search is no longer necessary. This improved approach has met considerable success in designing short sequences (~36 residues) that fold to target structures on a cubic lattice.[16]

The common feature of many of these protein design algorithms is a Monte Carlo search in sequence space at fixed backbone configuration. This approach works particularly well with coarse-grained heteropolymer models, since the task of having to arrange side chains after a mutation is avoided altogether. Moreover, heteropolymer-based algorithms are typically a necessary first component of multiscale protein design approaches, in which a formidable design problem is tackled in successive steps of increasing atomic detail. An important factor affecting the performance of these algorithms is the sequence landscape of a protein—that is, how sequence characteristics like energy and Z-score change as one moves from one mutation to another. It has been suggested that the sequence landscape is unfrustrated,[22,25] which means that the system is not likely to become trapped in metastable, local minima during the annealing procedure. However, little direct characterization of the sequence landscape in this context has been performed, and so it is not entirely clear how different amino acid alphabets affect the design problem.

In this work, we take a global view of sequence landscapes and ask: how are sequences distributed in the full landscape (i.e., no compositional constraints), and how does this distribution depend on the available amino acid alphabet? We attempt to understand this problem by computing, exactly, the number of sequences threaded through a fixed backbone, as a function of their energy. Of direct relevance to protein design, this approach elucidates the distribution of sequences underlying the fixed-backbone sequence design algorithms described above, and hence, can quantify the space of sequences which these

algorithms search. Crucially, as we explain below, we perform this calculation avoiding the computationally intractable $O(20^N)$ enumeration of sequences. Moreover, the distributions we calculate can be expressed as entropy functions and have a straightforward thermodynamic analogy: they relate to the hypothetical scenario in which the backbone configuration of a protein is fixed, but each residue freely mutates and attains thermal (chemical) equilibrium with the surrounding environment. In this scenario, the temperature mediates the observed ensemble of sequences, in which individual sequences carry a probability proportional to exp $(-E/k_BT)$. At infinite temperature, no sequence is preferred, and this ensemble consists of entirely random sequences. In contrast at very low temperatures, only those sequences whose amino acid composition yields the lowest energies are observed (typically homopolymers). Clearly, this scenario is not fully representative of real mutation thermodynamics in proteins, since it neglects both backbone rearrangements and the kinetics of mutations. We therefore regard it primarily as an insightful and computationally tractable perspective for characterizing different monomer alphabets. Throughout this article, we will refer back to the hypothetical fixed-backbone mutation scenario as an important mode of interpretation.

In contrast to protein design algorithms, our method does not involve simulations at fixed temperature, as one might anticipate from the mutation scenario just described. Instead for each candidate backbone of length N, we perform a *statistical* counting of the sequences so that we can reconstruct, exactly, the distribution $\Omega(E)$ of the number of sequences Ω with energy E. As shown below, this enables one to extract properties over the entire temperature range from a single simulation. Previously, Shakhnovich and coworkers calculated these kind of distributions in several candidate proteins; however, their simulations were performed at constant overall amino acid composition and they utilized an indirect thermodynamic integration approach to extract Ω.[26] Our work is most similar to the recent efforts of Elber and coworkers, who have also developed a statistical sequence counting procedure.[27] That group's interest has been in the so-called "evolutionary temperature" of particular proteins,[25] that is, the temperature T, which maximizes the probability that the original, native sequence appears as one of the mutations in the fixed-backbone scenario. Our work differs from Elber's in two respects: we are interested in the complete temperature dependence of sequence mutation properties, and the effects of alphabet choice on such properties; and we employ a different algorithm (developed by us) that offers considerable advantages in terms of flexibility of implementation and propagation of statistical error.

This article is organized as follows. The Methods section reviews the models and computational algorithms employed. The Results section presents the mutation calculations and offers an analysis of the differences between the models. Finally, we comment on the significance of this work for protein design algorithms and the possibilities for future application of our approach.

METHODS

The heteropolymer protein models we consider here are completely characterized by the positions $\mathbf{r}^N = \{\mathbf{r}_1,\mathbf{r}_2,\ldots,\mathbf{r}_N\}$ and species $s^N = \{s_1,s_2,\ldots,s_N\}$ of each amino acid bead. The energy of a particular configuration and sequence is given by the following contact-potential Hamiltonian:

$$E(\mathbf{r}^N,s^N) = \sum_{i<j} B(s_i,s_j)\Delta(\mathbf{r}_i,\mathbf{r}_j) \qquad (2.1)$$

where the summation runs over all pairs of monomers, the matrix B gives the energetic contribution of a contact between amino acids of type s_i and s_j, and Δ is one if residues i and j form a contact and zero otherwise. For the models studies here, two amino acids form a contact if their mutual separation distance is less than 6.5 Å.[29] This value is chosen based on the Miyazawa-Jernigan study[6] (discussed below).

Three contact potentials are the subject of our current investigation. These are represented by the matrices $B(s_i, s_j)$ appearing in the Hamiltonian in Eq. 2.1. One is the HP model of Lau and Dill,[7] which consists of only two monomers but has been studied in great detail as a minimalist model of protein folding.[2,26] For this contact potential, $B = -1$ for hydrophobic-hydrophobic contacts and $B = 0$ otherwise. The other two potentials are 20-letter codes and stem from the work of Miyazawa and Jernigan.[6] These authors used experimentally determined structures for 42 globular proteins of 100–500 residues in length to calculate frequencies of contacts between each pair of the 20 natural amino acids. The quasichemical approximation was invoked to compute effective pairwise contact energies from the relative frequency of occurence of pair contacts. Miyazawa and Jernigan produced two approximations for $B(s_i, s_j)$ of relevance to this work. The first, which we term MJa (upper half of Table V in Ref. 6), captures the effects of hydrophobicity and corresponds to the free energy of "reaction" when two solvated residues come in contact with one another. This is represented schematically by the reaction

$$x\text{-}w + y\text{-}w \overset{B(x,y)}{\longleftrightarrow} x\text{-}y + w\text{-}w \qquad (2.2)$$

where x and y are residue types and w is water. The other matrix, MJb (Table VI in Ref. 6), is essentially an averaged version of MJa that measures the tendency of pairs of specific amino acids to be in contact relative to other amino acids, rather than relative to solvent:

$$x\text{-}r + y\text{-}r \overset{B(x,y)}{\longleftrightarrow} x\text{-}y + r\text{-}r \qquad (2.3)$$

where r is the average residue. Numerous studies have investigated the properties of the MJa and MJb matrices, including their behavior in protein design problems.[8–10, 26, 30–33] Although Miyazawa and Jernigan have since refined their original contact potentials,[34,35] we

chose the original calculations in order to make contact with the above-mentioned studies. Moreover, although more biologically accurate potential energy functions may exist, the three we examine here are arguably the most-studied in computational protein design and are frequently used to benchmark new algorithms. It is therefore fundamentally important that these contact potentials be fully characterized and compared. As our study will demonstrate, significant model differences exist among them that can greatly affect protein design algorithms.

In this work we primarily examine results for ribonuclease A (Protein Data Bank code: 1c0c), because its behavior is typical among the proteins we have studied. For the MJb potential, which exhibits unique behavior (discussed below), we examine 51 additional proteins. Of these, 47 are randomly selected proteins of lengths 20–200 residues[28] from the list of Elber and coworkers.[27] The Elber catalog contains roughly 4000 proteins found in the Protein Data Bank,[12] whose member structures are collectively representative of all the structural motifs in the complete PDB. In addition, we also select 4 commonly studied proteins: trypsin inhibiter (5pti), lysozyme (132l), myoglobin (1mbo), and human carbonic anhydrase (1ca2). For all of these proteins, the backbone configuration is determined by taking the center position (geometric) of the side-chain heavy atoms.

The main part of our calculations proceeds as follows. For a particular backbone $\mathbf{r}^N$, we desire to calculate the distribution in energy of all sequences s^N possible with a given alphabet (consisting of 2 monomers for HP or 20 for MJa and MJb). Let this distribution be given by $\Omega_{\text{seq}}(E)$, which provides the number of sequences which have energy E and is implicitly specific to the particular $\mathbf{r}^N$ considered. This distribution will have a maximum and it must approach zero as the energy tends toward positive and negative infinity. This constraint emerges because the total number of sequences is finite, and so $\Sigma_E \Omega_{\text{seq}}(E) = q^N$, where q is the number of monomers in the alphabet. In analogy with other thermodynamic systems, the maximum of the distribution is pronounced and on the order of the total number of sequences, which implies that Ω_{seq} spans enormous orders of magnitudes. For example, a 100-residue protein using a 20-letter alphabet produces 20^{100} sequences such that $\Omega(E)$ will vary between $O(1)$ and roughly $O(10^{130})$.

Rather than work with Ω_{seq} directly, it is natural to examine its logarithm. This permits an analogy with Boltzmann's equation, in which one defines a sequence entropy:

$$S_{\text{seq}}(E) = k_B \ln \Omega_{\text{seq}}(E) \tag{2.4}$$

where k_B is Boltzmann's constant. The sequence entropy differs substantially from the familiar entropy of protein thermodynamics; in the latter case, the relevant Ω counts the number of *configurations* of a fixed sequence that have a given energy, rather than the number of sequences of fixed configuration. Still, the sequence entropy retains a physical significance if one considers the hypothetical situation described in methods in which a protein backbone is held fixed, but its amino acids are allowed to mutate. In this scenario, the mutations affect the energy of the protein, via the Hamiltonian in Eq. 2.1. This means that, if thermalized (i.e., connected to a heat bath), the temperature of the system will mediate the observed ensemble of mutated sequences, each sequence having probability $\exp(-H/k_BT)$. This ensemble can be characterized by its average energy, or by any other moments of the distribution in energy of its sequences:

$$\langle E^n \rangle = \frac{1}{Q(T)} \sum_E E^n \cdot \Omega_{\text{seq}}(E) e^{-E/k_BT} = \frac{1}{Q(T)} \sum_E E^n \cdot e^{S_{\text{seq}}(E)/k_B - E/k_BT} \tag{2.5}$$

where the partition function $Q(T)$ equals

$$Q(T) = \sum_E \Omega_{\text{seq}}(E) e^{-E/k_BT} = \sum_E e^{S_{\text{seq}}(E)/k_B - E/k_BT} \tag{2.6}$$

Thus, given the sequence entropy function it is possible to calculate the temperature dependence of the average sequence energy from these expressions using $n = 1$. It is also useful to consider these calculations in simpler form, when Ω_{seq} and S_{seq} can be approximated as continuous functions. Although technically both stem from a discrete distribution owing to the finite number of sequences q^N, the extremely high number density of sequences within any finite energy range permits their treatment as continuous. Consequently, the sum in Eq. 2.6 will be dominated by one particular energy, E^*, whose value is obtained by maximizing the term in the exponential. This analysis yields the familiar thermodynamic expression

$$\frac{dS_{\text{seq}}}{dE}(E^*) = \frac{1}{T} \tag{2.7}$$

In other words, for any given energy, the slope of the sequence entropy curve at that point gives the associated reciprocal temperature. By associated temperature, we mean the temperature at which sequences of that energy are spontaneously produced in the mutation scenario. These ideas are identical to those in elementary statistical mechanics texts, with the only distinction being that sequence or compositional degrees of freedom are being thermalized, rather than configurational ones. Moreover, the situation we have described is relevant to protein design algorithms that entail such fixed-backbone mutations.

To determine the sequence entropy for each protein, we use a flat-histogram algorithm developed by us,[11] which is a derivative of the Wang-Landau method.[36,37] Generally speaking, this algorithm is capable of calculating entropies and free energies directly and works by continuously adapting an initial estimate for these functions based on the instantaneous state of the simulation. The current implementation is as follows. For each backbone, we start with an initial random sequence. The algorithm proceeds by making successive random changes to the sequence, which consist of 50 percent point mutations and 50 percent swaps of residues at two random locations. For each

proposed change, the mutation from initial state A to final state B is accepted or rejected based on the following criterion:

$$P_{\text{acc}} = \min[1,\ \exp(S_{\text{seq}}(E_A)/k_B - S_{\text{seq}}(E_B)/k_B)] \quad (2.8)$$

where E_A and E_B are the initial and final energies, respectively. This criterion ensures that the probability of any one mutated sequence A is proportional to $1/\Omega_{\text{seq}}(E_A)$. Of course, we do not know the sequence entropy at the start of the simulation, and therefore what appears in this expression is the current running estimate of the entropy function. The initial estimate is simply $S_{\text{seq}} = 0$ everywhere. Over the course of the simulation, the entropy is systematically refined by observing the following: when the algorithm has converged upon the true sequence entropy to within an additive constant, on average each energy level will be visited with equal probability, owing to the acceptance probability in Eq. 2.8. That is to say, if the sequence entropy has converged and the simulation proceeds for a very large number of steps, a histogram of energies of all the mutated sequences taken during that period will appear flat. Thus one can initiate a feedback mechanism whereby the sequence entropy is systematically modified until a flat histogram is observed.

This feedback mechanism is the Wang-Landau component of the algorithm we employ— at the end of each Monte Carlo step, we increment the entropy at the current energy by a small amount g. When each energy level is visited equally, this procedure serves only to shift the entire S_{seq} curve upward. Such shifting is an irrelevant change because it only affects the unknown additive constant, which can always be determined by the normalization condition $\Sigma_E \Omega_{\text{seq}}(E) = q^N$. The parameter g starts off at a value of 1, and during the course of the simulation is tuned in stages to smaller values, in order to resolve the entropy to greater precision and to asymptotically achieve proper sampling of the mutations. Specifically, g is changed by $g \leftarrow 0.5g$ when each energy level is visited at least 20 times during the interval when g has its current value. The simulation is complete when g falls below a value of 10^{-10}.

To accelerate the convergence of this procedure, we have supplemented the Wang-Landau component with an additional, more robust estimator of the sequence entropy based on "transition matrix" estimators.[38–40] Essentially, we use recorded frequencies of move proposals between all of the *pairs* of energy levels to generate a more accurate entropy estimate, which periodically replaces the Wang-Landau calculation. To do this, we add 1 to a matrix entry $C(E_A, E_B)$ every time a move is proposed from energy E_A to E_B. At the start of the calculation, C is initialized to zero for all E_A and E_B. After numerous moves, the C matrix provides an estimate for the macroscopic move proposal probabilities, $T_{\text{prop}}(E_A \to E_B) \approx C(E_A, E_B)/\Sigma_{E_C} C(E_A, E_C)$, which in turn yield relative values of the entropy:

$$S_{\text{seq}}(E_A)/k_B - S_{\text{seq}}(E_B)/k_B = \ln \frac{T_{\text{prop}}(E_B \to E_A)}{T_{\text{prop}}(E_A \to E_B)} \quad (2.9)$$

A derivation of Eq. 2.9 is available in Ref. 11; it will be sufficient to say here that it is a consequence of detailed balance. Given the matrix C, therefore, the entire sequence entropy can be reconstructed using all instances of this expression for each pair of energy levels. In the flat histogram algorithm, transition probability statistics are automatically calculated once $g < 10^{-5}$ and are used to generate new entropy estimates at every change in the value of g. The complete algorithmic procedure is described in greater detail in Ref. 11.

Our method is a statistical (Markovian) evaluation of the sequence entropy. The statistical route to determining the sequence entropy is essentially the only possible approach for proteins of any modest length, since a direct enumeration of every sequence requires the evaluation of an enormous and intractable number, 20^N, of energies. Previous studies have also employed a Monte Carlo approach to sequence enumeration, starting with a pioneering investigation of sequence entropies by Shakhnovich.[26] Unlike the present study, however, Shakhnovich's work examined the sequence entropy under the constraint of constant amino acid composition, in order to assess the sequence landscapes explored by the annealing design algorithm.[13,15] Furthermore, that work utilized a thermodynamic integration technique to determine the entropy, rather than a direct (microcanonical) counting approach that we employ here. Thermodynamic integration yields the true microcanonical entropy only in the infinite-system limit, and it is not capable of resolving curvatures in the entropy responsible for phase transitions.

Our numerical procedure is most similar to an algorithm recently developed by Elber and coworkers.[27] In the Elber "telescoping" method, the entire energy spectrum is subdivided into very small slices, and the calculations entail determining the number of random mutations that keep sequence energies within each slice. By combining results from all of the slices, the sequence entropy is then reconstructed. A limiting factor in the Elber approach is that the slices must be sufficiently narrow in order to attain reasonable statistical accuracy. In contrast, our method can consider the entire energy spectrum at once, without affecting the statistical accuracy of the calculated results. This is because the flat-histogram approach ensures that each energy level is visited with equal frequency. Along the same lines, our method offers significant implementation flexibility, because the range of energy explored by a single processor can be fixed arbitrarily while maintaining a high statistical quality of results. This feature is particularly important for parallelized calculations (see comment below).

In addition to generating the sequence entropy functions, we also record the average fractional occurrence of each amino acid (number of occurrences of a given amino acid along the protein backbone divided by the number of amino acid residues in the protein) for all sequences of a given energy. This information is readily determined from our simulations by maintaining a histogram in energy for each species and incrementing the corresponding bins at every Monte Carlo step. Using these results, one can determine the temperature-mediated composition of the mutation ensemble:

$$\langle x_i \rangle = \frac{1}{Q(T)} \sum_E x_i(E) \cdot e^{S_{\mathrm{seq}}(E)/k_B - E/k_B T} \qquad (2.10)$$

where $x_i(E)$ is the average fraction of residue i in all sequences of energy E (what we measure), and $\langle x_i \rangle$ is the same quantity as a function of temperature.

Before proceeding to the results of our investigation, we comment on some of the logistical aspects of this calculation. One issue concerns the initial setup of the unknown sequence entropy function. In order to determine the minimum and maximum energy levels, we first run a set of very short Monte Carlo simulations for each protein using the acceptance criterion in Eq. 1.1 with the reciprocal temperature at large positive and negative values, respectively. This procedure finds sequences of very low and high energies, and from these results, we establish an energy range over which to calculate the sequence entropy. Because the entropy is peaked at intermediate energies, it is not necessary to find the global extrema of the energy range in order to satisfy the normalization condition, $\Sigma_E \Omega_{\mathrm{seq}}(E) = q^N$, to good accuracy.

For the actual computations the energy ranges are finely discretized into a large number of bins, each of width 1 in units of the HP and MJ potentials. Our sequence entropy calculations then proceed in the automated fashion described above, with an average convergence time of $\tau \approx (7.2 \cdot 10^{-4}\ \mathrm{h}) N^{1.9}$ on a single 2.4 GHz Intel Xeon processor for MJb, where N is the number of amino acid residues along the backbone. For the other alphabets, which do not possess phase transitions as described below, convergence is much faster. We also note that for the longer proteins, the real-time requirement can be reduced substantially by dividing the total energy range into a small number of overlapping windows and parallelizing the calculations. Although we do not elaborate on this strategy here, details of such a procedure can be found in Ref. 41.

RESULTS

A typical result from our calculations is displayed in Figure 1, which contains the sequence entropy for ribonuclease A backbone using each of the three contact potentials. Several observations stem from our results. First, the maximum of the sequence entropy is significantly lower for the HP model, with a value of $82.4k_B$ versus $\sim 370k_B$ for the MJ models. This is not surprising, as the total number of possible sequences in the former is only 2^N, which yields a maximum possible entropy of $(124 \ln 2)k_B \sim 86.0k_B$ for a chain of length 124. Second, each curve is fairly asymmetric, the maxima occurring at high energies relative to the average of the energy range. Asymmetry in the HP model is straightforward to rationalize: the minimum energy sequence in this case is a completely hydrophobic chain with each contact contributing -1. In contrast, the maximum energy sequence with $E = 0$ is degenerate and can consist of both hydrophobic and hydrophilic monomers provided there are no hydrophobic-hydrophobic contacts. The energy at the maximum roughly corresponds to the average energy one would obtain for random sequences of half hydrophobic and half

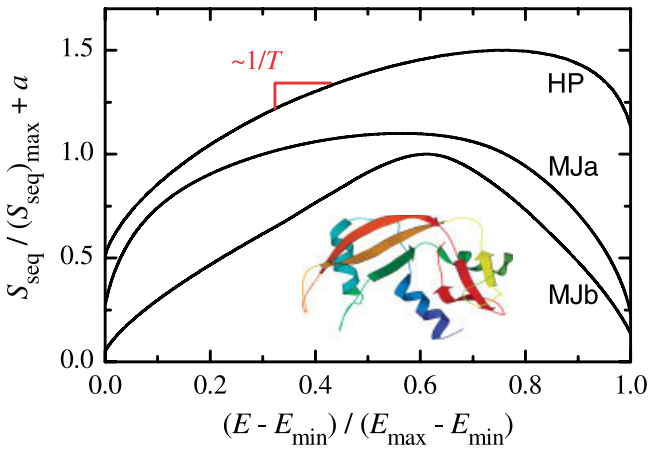

Fig. 1. Scaled sequence entropy functions, S_{seq}, for ribonuclease A using the HP, MJa, and MJb contact potentials. The quantity $\exp[S_{\mathrm{seq}}(E)/k_B]$ gives the number of sequences with energy E when threaded through the ribonuclease A backbone. For each curve, the energy range has been nondimensionalized by the minimum and maximum studied energies. The entropy axis has also been scaled by its maximum value, and in addition, shifted for clarity between the three curves. The parameters $\{E_{\mathrm{min}}, E_{\mathrm{max}}, (S_{\mathrm{seq}})_{\mathrm{max}}/k_B, a\}$ for each of the potentials are, in units of the respective potentials: HP $\{-286.5, -0.5, 82.4, 0.5\}$, MJa $\{-1951.3, 26.2, 366.5, 0.1\}$, MJb $\{-301.0, 201.0, 368.8, 0.0\}$. A schematic representation of ribonuclease A is also shown. [Color Figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

hydrophilic monomers. Because in these random sequences the monomers are randomly dispersed, each contact will contribute on average -0.25 to the total energy, thus placing the entropy maximum above the midpoint of the energy range. This line of reasoning also provides some hint as to asymmetry in the MJ models; the minimum value of both of these B matrices occurs for a self interaction (Phe-Phe in MJa, Cys-Cys in MJb), whereas the maximum contact energy results from unlike pairs (Cys-Pro in MJa, His-Met in MJb).

It bears mentioning the comparison of our sequence entropy calculations with those of Shakhnovich.[26] Those results, performed at constant composition and using a thermodynamic integration technique, suggest that the functional form of the the sequence entropy is quadratic in energy, independent of protein backbone and regardless of the amino acid alphabet used (see Fig. 2 of Ref. 26). Shakhnovich attributes the emergence of this universal form to a self-averaging heteropolymer principle in the constant-composition case. This is clearly not the case when the entire sequence landscape is considered (i.e., when there are no compositional constraints), given the asymmetry and variance in the curvature of the sequence entropy functions represented in Figure 1. Thus considering sequences of all compositions does not entail the same universality principle, at least when viewed over the entire range of sequence energies. For comparison with the earlier studies, we have also performed sequence entropy calculations for ribonuclease A at constant composition. The methodological procedure is identical to our previous simulations, except that only bead swap moves are permitted. Figure 2 presents these calculations along with quadratic fits to the data. Although the asymmetry of these

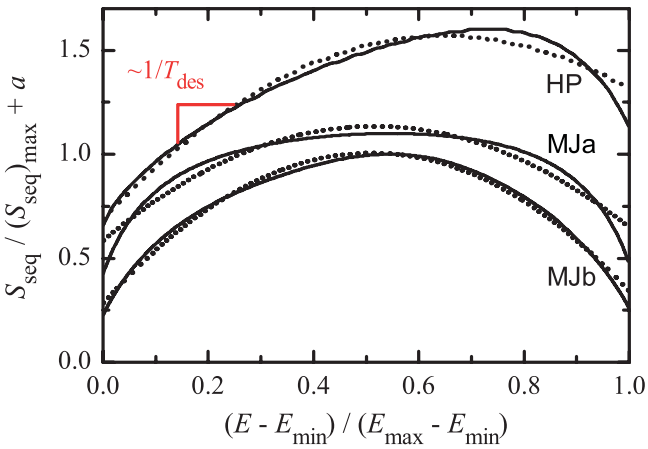

Fig. 2. Scaled sequence entropy functions calculated at constant composition for ribonuclease A. The quantity $\exp[S_{seq}(E)/k_B]$ gives the number of sequences with energy E and the same amino acid composition as ribonuclease A when threaded through its backbone. The scaling is analogous to Figure 1, with parameters for $\{E_{min}, E_{max}, (S_{seq})_{max}/k_B, a\}$ as follows: HP {−108.5, −0.5, 73.6, 0.6}, MJa {−948.1, −449.4, 313.1, 0.1}, MJb {−105.1, 95.3, 314.5, 0.0}. Quadratic fits to each sequence entropy are shown as dotted lines. [Color Figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

results is less pronounced than that of the unconstrained landscape in Figure 1, it is clear that none of the models yields a purely Gaussian distribution of sequences, even at constant composition. The sequence entropy of the MJb potential is the closest to a quadratic form, with the most marked deviations stemming from the HP result. This suggests that the existence of a Gaussian distribution of sequence energies still depends on the nature of the particular model being studied. For the remainder of this study, however, we will focus on the properties of the unconstrained landscapes, which are necessarily the focus in bottom-up protein design.

Recalling the connection between sequence entropy and the hypothetical fixed-backbone mutation scenario, we return to Figure 1, where the proportional relationship between the slope of these curves and the reciprocal temperature is illustrated. To determine the average energy of thermalized, mutated sequences, one simply finds the point on the sequence entropy curve where the slope matches the value of the inverse temperature. This brings to light an interesting feature: because each curve possesses a maximum, each also contains a high-energy region with negative slope and hence corresponding negative temperature. This implies that these energies would only be explored by spontaneous mutations of the system for a negative environmental temperature. Because it is the inverse temperature that is actually relevant, negative temperatures are in some sense greater than positive ones and are "hotter" because they correspond to higher-energy states. Although negative temperatures are a laboratory curiosity,[42] any system with bounded degrees of freedom (e.g., two-level energies) will exhibit an entropy-energy curve that includes a negatively sloped region (consider, e.g., the Ising model). The entropy maximum itself corresponds to infinite temperature, and as one moves to lower energies in these curves, the associated temperature decreases until its value is nearly zero at the low end of the energy range.

Consideration of the energy-temperature relationship suggested by the curves in Figure 1 reveals qualitatively different behavior in the MJb model. For the HP and MJa contact potentials, the ensemble-averaged energy steadily and continuously decreases as the temperature moves from high positive values toward absolute zero. This is evident from the fact that the slope of these sequence entropy curves gradually increases as one moves to lower energies; the curvature is consistently negative. In contrast, the MJb results contain a "flat" region of near-zero curvature. The implication in the mutation scenario is the following: at the temperature corresponding to the slope of this flat region, the mutations fluctuate over the entire range of energy for which the sequence entropy exhibits quasi-linear behavior. This temperature has a special significance—it is a point at which the ensemble of observed sequences has the maximum possible variance in energy. Consequently, we will use the phrase "temperature of maximum variance" and notation T_{MV}. This terminology is also beneficial because it provides a strict recipe for the calculation of the temperature T_{MV}. By appealing to Eq. 2.5, one finds the temperature which maximizes the quantity $\langle E^2 \rangle - \langle E \rangle^2$. In the case of ribonuclease A, this temperature turns out to be 0.92 in reduced units.

Upon closer examination, the quasi-linear behavior of the sequence entropy for MJb includes subtle features that are reminiscent of a first-order phase transition. By constructing the free energy analogue $A(E;T) = E - TS_{seq}(E)$, one can visualize this phase transition. The top panel of Figure 3 shows the sequence free energy as a function of energy, at the temperature of maximum variance. As is the usual case with free energies, A attains a minimum at thermodynamic equilibrium. Interestingly, at T_{MV} the free energy contains three minima separated by small free energy barriers on the order of $0.2–0.4\, k_BT$. In a thermodynamic sense, each one of these minima corresponds to a distinct ensemble or "phase" of sequences, with the global minimum being the dominant stable phase and the other two existing as metastable states. By increasing or decreasing the temperature slightly, the relative depth of each can be changed and either of the metastable phases can be made the global free energy minimum. Thus two phase transitions occur around T_{MV}, each separating adjacent free energy minima. It should be noted that these are not true phase transitions in the strictest sense, because the systems are of finite size. Our calculations for the 51 additional proteins discussed below indicate that the magnitude of these free energy barriers is positively correlated with chain length, scaling roughly as $0.01Nk_BT$, although the data's scatter prevents a definitive assignment of any one particular scaling law (results not shown). For the present discussion, our use of "phase transition" for the finite systems studied refers to the existence of multiple free energy minima.

An issue that emerges is the nature and origin of these phase transitions. Generally speaking, the prerequisite

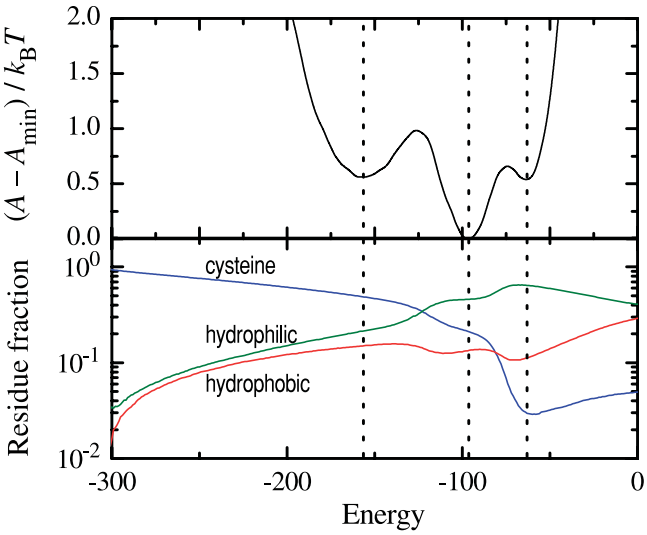

Fig. 3. Top: sequence free energy for the ribonuclease A backbone using the MJb potential. Here, $A(E) = E - TS_{seq}(E)$ and the y-axis is normalized by its minimum value. The reduced temperature shown is 0.92. Bottom: sequence composition for selected amino acid groups as a function of energy. The strongly hydrophobic and hydrophilic groups include the amino acids (F,I,V,L,M,W) and (K,Q,D,E,N,P,R,S), respectively. The amino acids (A,G,H,T,Y), which are intermediate in the hydrophobicity scale of Rose et al.,[43] are omitted.

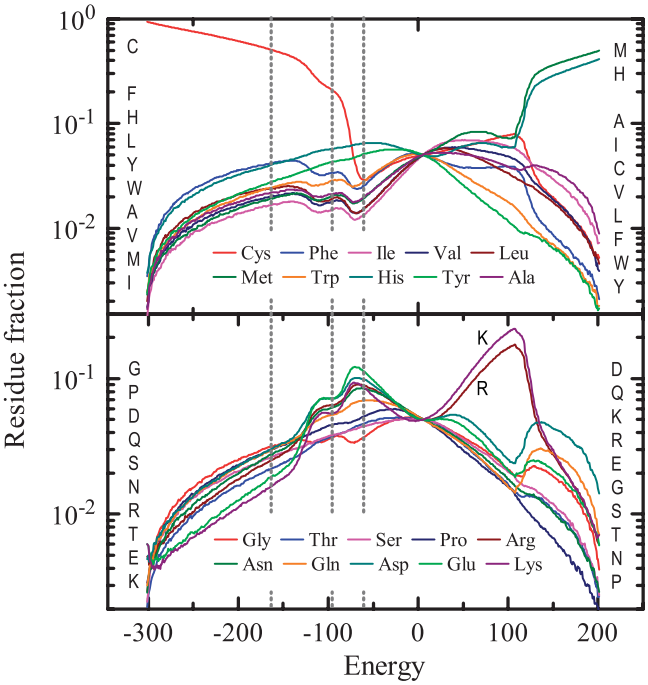

Fig. 4. Composition of sequences for the ribonuclease A backbone using the MJb potential. Each of the curves shows $x_i(E)$, the average fraction of residue i averaged over all sequences of energy E. Amino acids are listed in order of most to least hydrophobic, according to the scale of Rose et al.[43] Two panels are used for clarity, with strongly hydrophobic residues appearing in the top one. The vertical lines correspond to the free energy minima in Figure 3.

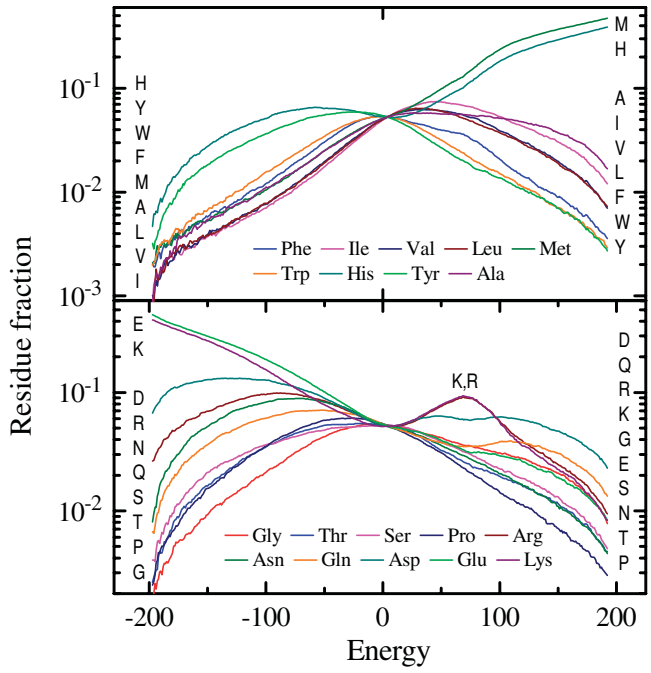

Fig. 5. Composition of sequences for the ribonuclease A backbone using the MJb potential with cysteine excluded from the amino acid alphabet. See Figure 4 for further explanation.

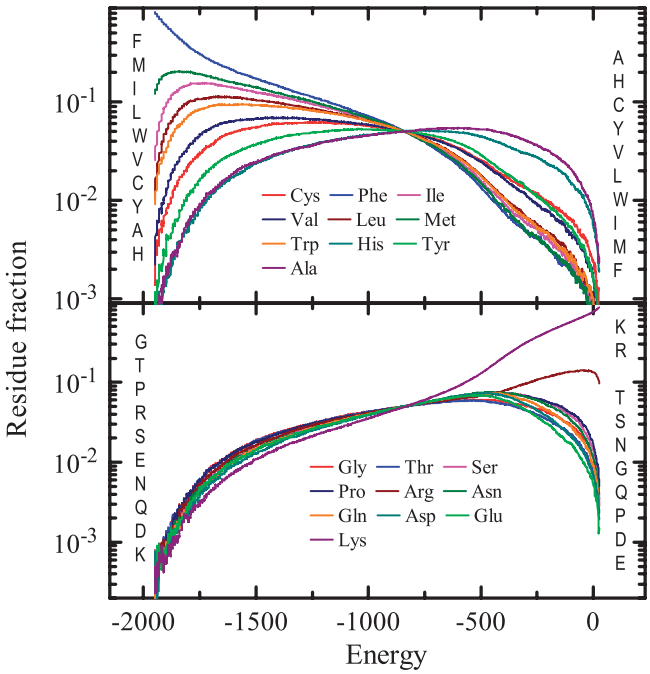

Fig. 6. Composition of sequences for the ribonuclease A backbone using the MJa potential. See Figure 4 for further explanation.

for a first-order phase transition is the existence of collective behavior or cooperativity within a system's degrees of freedom. For example, a gas condenses into a liquid when sufficient pressure is applied because the collective energetic benefit of molecular close-contact in the condensed phase outweighs the entropic cost of condensation. Similarly, the cooperative and energetically favorable arrangement of molecules into a lattice structure gives rise to a crystallization transition. In the present results, the degrees of freedom are not configura-

tional in nature, but rather compositional. Therefore, the cooperativity that gives rise to these sequence phase transitions must emerge because of differences in the interactions between various groups of species. Note that the sequence entropy for MJb in the composition-constrained case (Fig. 2) does not possess a "flat" region. This implies that the fractional residue constraint is sufficient to disrupt the cooperativity responsible for the phase transition. Thus, the phase transition in the unconstrained landscape is likely to be characterized by the composition of sequences as one moves across the energy spectrum.

Figure 4 shows how sequence composition varies with energy. Here we have calculated the function $x_i(E)$, which gives the fraction of the N monomers of type i, averaged over all sequences of energy E. Thus, the 20 values $x_i(\langle E \rangle)$ give the average sequence composition for a given temperature T and its average energy $\langle E \rangle (T)$. Referring to the location of the free energy minima in Figure 3, one can then identify the differences in amino acid composition that distinguish the associated phases. The lowest energy phase, which is explored at low temperatures, entails sequences that are dominated by cysteine residues. That the lowest-energy sequences are nearly cysteine homopolymers is not unexpected, because the cysteine-cysteine interactions are the most favorable in MJa (and also for the potentials used in Elber and co-workers' studies[27]). However, the fact that a first-order transition separates predominantly cysteine sequences from more heterogeneous ones is quite striking.

Figure 4 implies that the upper two free energy minima correspond to phases of increasing hydrophilic content, at the expense of the hydrophobic residues and cysteine. In fact, it is quite remarkable that the strongly hydrophobic (C,F,I,V,L,M,W) and hydrophilic (K,Q,D,E,N,P,R,S) amino acids naturally segregate in this way, when viewed along the hydrophobicity scale of Rose et al.[43] With regards to the free energy minima, one can define three classes of amino acids in addition to cysteine; those which exhibit a sharp increase in content as one moves from the lowest to highest minimum (appearing as small "bumps" in Fig. 3 at the location of the dotted lines), those that exhibit a similarly sharp decrease, and those whose variation over this range is not substantially pronounced. The resulting clustering of amino acids closely parallels the Rose scale, with the strongly hydrophilic and hydrophobic residues corresponding to the first and last groups, respectively. The general behavior of each cluster of residues can be demonstrated by combining multiple composition curves. The bottom panel of Figure 3 presents the composition analysis in this way and clearly demonstrates the phases' differences in terms of the fractional participation of each group. Interestingly, the strongly hydrophilic groups dominate the sequences at low energy. This seemingly unusual phenomenon emerges because the MJb matrix is normalized relative to the average residue-residue interaction and reflects the favorable hydrophilic-hydrophilic clustering on the exterior of folded proteins. Recall that the only residue-specific information present in this study is the set of contact energies in the MJb potential. Therefore, this matrix of information naturally contains a type of cooperativity between each class of amino acids, borne out in the energetics of their interaction. The emergence of such classes in contact matrices bas been studied before—see Ref. 30, for example, which employed an eigenvector analysis. However, the observation of this kind of behavior in the form of a phase transition is unexpected.

These results suggest that cysteine plays a special role in the phase transition scenario. This likely stems from its relatively large self-interaction energy, which in turn results from the prevalence of disulfide bonds in real protein systems. The MJ matrices clearly do not provide an accurate portrayal of disulfide bonds in nature, because they permit a cysteine residue to be bonded to multiple other cysteines. This potentially exaggerates the extent of the cysteine-rich region in the sequence entropy of MJb. It is therefore desirable to elucidate the exact nature of cysteine's role in creating sequence phase transitions. To investigate this question, we have performed an identical calculation of sequence entropy for the MJb potential, but with a 19-letter alphabet in which cysteine is absent. Figure 5 shows the result of this calculation. Without cysteine, much less interesting behavior is observed and, in fact, the phase transitions disappear (results not shown). This suggests that, in MJb, cysteine acts as a partitioning agent that serves to enhance energetic differences between the other hydrophobic and the hydrophilic amino acids. Because cysteine is dominant in low-energy, low-temperature sequences, it is an effective mediator for the kinds of residues that enter as the temperature is increased.

It is worth mentioning, briefly at least, that several other amino acids appear to possess unique behavior in the composition plots of Figure 4. Namely, the pairs of residues (M,H) and (K,R) stand out in the high-energy, negative temperature region of the sequence entropy. The dominance of these pairs emerges from their high, positive interaction energies, which are the largest and second largest values in the entire MJb matrix, respectively. The existence of a maximum in the fractional composition of K and R is likely due to an underlying clustering of amino acids as well. One hint at this is the relatively weak interaction of K and R with M and H, which requires that K and R be "swapped out" for M and H in going toward the high-energy extreme of the sequence distribution. Still, the biological significance of such clustering in the high-energy region is questionable, because the positive total energy indicates that a compact protein structure of the type examined is unstable with respect to extended configurations.

Figure 6 shows the composition analysis for the full MJa potential. Recall that the MJa sequence entropy does not contain a "flat" region, and hence does not exhibit phase transition behavior of the kind just described. The residue fraction curves in this case do not reveal any unexpected patterns: the energy dependence of each x_i is strongly correlated with amino acid i's hydrophobicity. Very hydrophobic residues are dominant at low energies and rare at high ones, and the opposite is true for the most hydrophilic

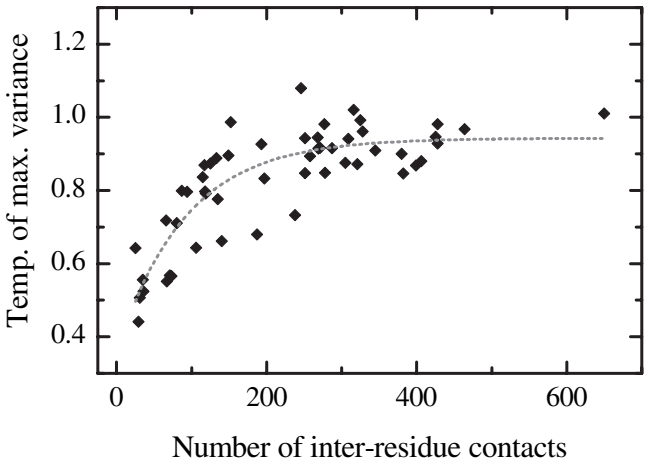

Fig. 7. Temperature of maximum energy variance as a function of number of residue-residue contacts, for 52 proteins[28] using the MJb potential. This temperature corresponds to the inverse slope of the "flat" region in each sequence entropy curve. The dotted line is an exponential fit and yields a large-protein asymptotic temperature of 0.93 (dimensionless units). If the original temperature scale proposed by Miyazawa and Jernigan is employed,[6] in which a temperature of 1.0 is equivalent to 300 K, then this asymptotic temperature is roughly 283 K.

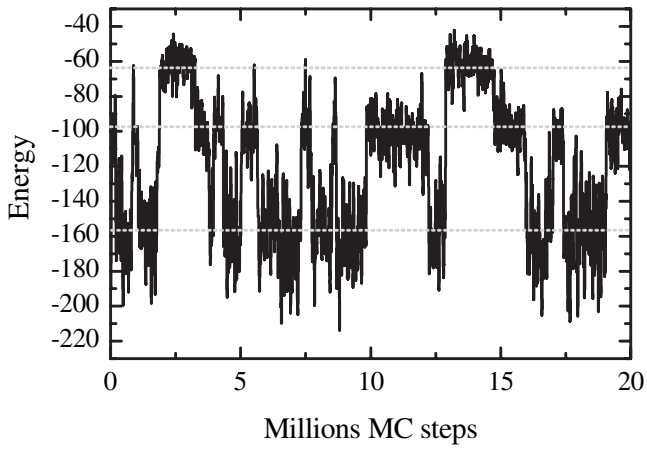

Fig. 8. Evolution of an annealing-type simulation of ribonuclease A using the MJb potential. The plot depicts the instantaneous potential energy as a function of number of Monte Carlo steps. The temperature is 0.92 in MJ units, which corresponds to the point of equilibrium between three sequence phases. Significant hysteresis is observed in moving between each of the phases, with a time scale on the order of a million steps. The dotted lines are a guide to the eye to help distinguish the phases.

species. Thus no significant cooperativity exists in MJa, and in this sense it possesses a very smooth sequence landscape. Because MJa provides a measure of the free energy of solvation, whereas MJb is renormalized to measure residue rearrangement, it is likely that the renormalization averaging brings out greater distinctions between the amino acids. One expects this to be the case to a certain extent, because the "rearrangement" perspective more directly pits species against each other.

So far we have only considered calculations for ribonuclease A, because these results are typical of those observed with several other proteins. The most significant quantitative differences among all proteins investigated here occur in the temperature of maximum variance, T_{MV}, for the MJb case. This is depicted in Figure 7 for the proteins we have studied and is shown as a function of the number of residue contacts. There is large degree of variance in T_{MV}, particularly for very small proteins, but the trend suggests that it levels off to a reduced value of 0.93 as the protein size increases. If one employs the original temperature scale used by Miyazawa and Jernigan,[6] where reduced unit temperature equals 300 K, this asymptotic temperature is roughly 283 K. Although in the present work phase transitions arise only as a consequence of the use of a specific contact potential (MJb), should they be found to occur in nature such phase transitions would provide a mechanism for the optimization of amino acid sequences. That is, fluctuations are naturally enhanced near phase transitions and would facilitate the search for biologically useful proteins.

What is the implication of sequence phase transitions on protein design algorithms? With the multiple first-order phase transitions found here, it becomes apparent that metastability can play a significant role. That is, the multiple free energy minima at T_{MV} are separated by free energy barriers; therefore, crossing these local maxima is in some sense a rare event. To demonstrate this property, we have performed an annealing-type simulation of ribonuclease A using MJb at its T_{MV}. We use the same Monte Carlo moves as in the flat histogram calculations, employing Eq. 1.1 for the acceptance criterion instead, although we do not employ compositional constraints. Figure 8 shows the progression of the sequence energy as a function of time for this simulation. The three sequence phases are indeed evident in this representation; the system typically spends a large number of steps in local energy fluctuations around one of the free energy minima, with occasional hops to one of the other two. The "time" scale for these transitions is nontrivial: roughly 1 million Monte Carlo moves. This suggests that, with contact potentials of this variety, short design simulations of less than roughly a million steps may not be sufficient to completely explore the sequence landscape. In these cases, the protein can become trapped in a particular phase/ensemble of sequences for durations on the order of the simulation length.

CONCLUSIONS

The "sequence landscape" corresponding to a particular amino acid alphabet, that is to say the collection of possible sequences and its dependence on fitness metrics, is of relevance to sequence-searching protein design algorithms. In "rugged" landscapes, backbone mutations encounter significant barriers on their progression to an "equilibrium" state; however, such state may be defined by the algorithm at hand. In contrast, "smooth" landscapes permit a rapid convergence to a free energy minimum or an optimum sequence. A large number of methods exist for protein design by performing fixed-backbone searches of sequence space, and thus these landscapes are one of the primary determinants of their performance.

In the present work, we have investigated the sequence landscapes of three simple, heteropolymer models of proteins. Our results underscore the significance of model-specific effects in protein design. We have only characterized these landscapes through one metric, the total intraresidue energy, but our calculations reveal nontrivial differences when different alphabets are employed. Two of the contact potentials we have studied, the 2-species HP and 20-species MJa models, yield "smooth" sequence energy landscapes. As one performs an unrestricted sequence annealing simulation, these models yield continuous and rapidly converging behavior. In contrast, the MJb potential gives rise to discrete effects in a similar setting. We have found that the MJb sequence landscape can be characterized as having multiple phase transitions, which are a consequence of strong cooperativity amongst groups of hydrophilic and hydrophobic species in the alphabet. The presence of these transitions may inhibit rapid convergence in design algorithms because transitions between phases require barrier-crossing events. In this sense, one might term these landscapes "rugged," although we caution that this is distinct from "frustrated" or glassy behavior. In possessing a phase transition, the MJb landscapes can be associated with a characteristic temperature. Other notable temperatures have been found to emerge in sequence landscapes, in particular by Shakhnovich and coworkers[44] and Elber and coworkers.[27] The comparison of all three results in the context of a biological interpretation remains an interesting subject for future study.

Our work suggests other avenues of future study. In our calculations, we only briefly considered the case when compositional constraints are imposed on the progression through sequence landscapes. Such constraints are a hallmark of early sequence annealing algorithms,[13,15] and it would be useful to examine their effects in greater detail. In particular, we found that the degree of deviation from a Gaussian distribution of sequence energies at constant composition is substantial for some models. Future studies might examine Gaussian deviation in a wider range of contact potentials and under various kinds of constraints, such as constant fraction of specific amino acids, fixed balance between hydrophobic and hydrophilic residues, or maximum presence of particular species. Understanding the degree to which such constraints make the sequence landscape more rugged or more smooth would be of great interest.

Another extension concerns the use of other characteristics of the sequence landscape. In this case, one seeks to characterize the distribution of sequences along some order parameter other than energy. The most obvious candidate is the Z-score,[24] which measures the number of standard deviations in energy the target structure falls below all other configurations. The Z-score can be determined relatively fast in simulation under certain theoretical approximations.[45] The flat histogram simulation in this case would proceed in a nearly identical fashion to the methods described above, except that the Z-score would be examined in place of the energy E. The calculated function $S_{\mathrm{seq}}(Z)$ would then provide information about the distribution of sequences in terms of their ability to contain a target structure as the native state. The presence of discrete or phase-transition behavior in these results might signify the conservation of motifs in well-designed sequences and would have significant implications for design algorithms that attempt to minimize the Z-score. Sharp changes of the ensemble-averaged Z-score with design temperature in a sequence mutation study by Dokholyan and Shakhnovich already bolster this possibility.[44] Most importantly, the flat-histogram Z-score approach would provide a quantitative measure of designability without exhaustive enumeration of conformations: assuming that sequences with Z less than some critical value Z_{crit} are able to fold into the target structure, the expression $\exp[S_{\mathrm{seq}}(Z_{\mathrm{crit}})/k_B]$ provides a direct estimate of their number. We are currently pursuing this approach.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of the Fannie and John Hertz Foundation and of the Department of Energy, Division of Chemical Sciences, Geosciences, and Biosciences, Office of Basic Energy Science (grants DE-FG02-87ER13714 to PGD and DE-FG02-01ER15121 to AZP.) We also thank Ned Wingreen for insightful discussions and Leonid Meyerguz and Ron Elber for providing their list of proteins.

REFERENCES

1. Malacinski GM. Essentials of molecular biology. Boston: Jones and Bartlett; 2003. xix, 491 p.
2. Dill KA, Bromberg S, Yue KZ, Fiebig KM, Yee DP, Thomas PD, Chan HS. Principles of protein-folding — a perspective from simple exact models. Protein Sci 1995;4(4):561–602.
3. Wolynes PG, Onuchic JN, Thirumalai D. Navigating the folding routes. Science 1995;267(5204):1619–1620.
4. Pande VS, Grosberg AY, Tanaka T. Statistical mechanics of simple models of protein folding and design. Biophys J 1997;73(6): 3192–3210.
5. Shakhnovich EI. Theoretical studies of protein-folding thermodynamics and kinetics. Curr Opin Struct Biol 1997;7(1):29–40.
6. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal-structures — Quasi-chemical approximation. Macromolecules 1985;18(3):534–552.
7. Lau KF, Dill KA. A lattice statistical-mechanics model of the conformational and sequencespaces of proteins. Macromolecules 1989;22(10):3986–3997.
8. Buchler NEG, Goldstein RA. Effect of alphabet size and foldability requirements on protein structure designability. Proteins 1999; 34(1):113–124.
9. Fan K, Wang W. What is the minimum number of letters required to fold a protein? J Mol Biol 2003;328(4):921–926.
10. Li H, Tang C, Wingreen NS. Designability of protein structures: a lattice-model study using the Miyazawa-Jernigan matrix. Proteins 2002;49(3):403–412.
11. Shell MS, Debenedetti PG, Panagiotopoulos AZ. An improved Monte Carlo method for direct calculation of the density of states. J Chem Phys 2003;119(18):9406–9411.
12. Berman HM, Westbrook Z, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
13. Shakhnovich EI, Gutin AM. Engineering of stable and fast-folding sequences of model proteins. Proc Natl Acad Sci USA 1993;90(15): 7195–7199.
14. Shakhnovich EI. Proteins with selected sequences fold into unique native conformation. Phys Rev Lett 1994;72(24):3907–3910.
15. Pande VS, Grosberg AY, Tanaka T. Thermodynamic procedure to synthesize heteropolymers that can renature to recognize a given target molecule. Proc Natl Acad Sci USA 1994;91(26):12976–12979.

16. Abkevich VI, Gutin AM, Shakhnovich EI. Impact of local and nonlocal interactions on thermodynamics and kinetics of protein-folding. J Mol Biol 1995;252(4):460–471.
17. Sun SJ, Brem R, Chan HS, Dill KA. Designing amino acid sequences to fold with good hydrophobic cores. Protein Eng 1995;8(12):1205–1213.
18. Deutsch JM, Kurosky T. New algorithm for protein design. Phys Rev Lett 1996;76(2):323–326.
19. Seno F, Vendruscolo M, Maritan A, Banavar JR. Optimal protein design procedure. Phys Rev Lett 1996;77(9):1901–1904.
20. Morrissey MP, Shakhnovich EI. Design of proteins with selected thermal properties. Fold Des 1996;1(5):391–405.
21. Betancourt MR, Thirumalai D. Protein sequence design by energy landscaping. J Phys Chem B 2002;106(3):599–609.
22. Pande VS, Grosberg AY, Tanaka T. Heteropolymer freezing and design: towards physical models of protein folding. Rev Mod Phys 2000;72(1):259–314.
23. Gutin AM, Abkevich VI, Shakhnovich EI. Evolution-like selection of fast-folding model proteins. Proc Natl Acad Sci USA 1995;92(5): 1282–1286.
24. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known 3-dimensional structure. Science 1991;253(5016):164–170.
25. Shakhnovich EI, Gutin AM. A new approach to the design of stable proteins. Protein Eng 1993;6(8):793–800.
26. Shakhnovich EI. Protein design: a perspective from simple tractable models. Fold Des 1998;3(3):R45–R58.
27. Meyerguz L, Grasso C, Kleinberg J, Elber R. Computational analysis of sequence selection mechanisms. Structure 2004;12(4): 547–557.
28. The proteins are the following, from shortest to longest: 1egp_B, 1ckk_B, 1ce0_A, 1roo, 1clg_A, 2bpa_3, 1bgk, 1mkc_A, 1huc_A, 1yrn_A, 2occ_K, 1fak_I, 1bpt, 1ssr, 1jen_B, 1bun_B, 1dax, 1pse, 1scj_B, 1dgr_V, 1fow, 1b4r_A, 1tnm, 1ihf_A, 1roe, 1hks, 1jwe_A, 1n72_A, 1c2n, 2a8v_A, 2fvw_H, 1jia_A, 1aks_A, 1cmo_A, 5msf_A, 3nul, 1dqg_A, 487d_I, 1an7_A, 1nbe_D, 1alw_A, 1jxp_A, 1br9, 1fak_T, 1qn8_A, 1ef6_B, 1qbk_C, 1ill_G.
29. We have considered both cases in which energetic contributions from nearest-neighbors along the protein backbone are either included or excluded from the total Hamiltonian. The results of the two approaches are qualitatively similar and entail only slight quantitative differences, which become less pronounced for the larger proteins. The data presented in this discussion corresponds to the former version.
30. Li H, Tang C, Wingreen NS. Nature of driving force for protein folding: a result from analyzing the statistical potential. Phys Rev Lett 1997;79(4):765–768.
31. Wang J, Wang W. A computational approach to simplifying the protein folding alphabet. Nat Struct Biol 1999;6(11):1033–1038.
32. Betancourt MR, Thirumalai D. Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Protein Sci 1999;8(2):361–369.
33. Papoian GA, Wolynes PG. The physics and bioinformatics of binding and folding — an energy landscape perspective. Biopolymers 2003;68(3):333–349.
34. Miyazawa S, Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol 1996;256(3): 623–644.
35. Miyazawa S, Jernigan RL. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. Proteins 1999;34(1):49–68.
36. Wang FG, Landau DP. Efficient, multiple-range random walk algorithm to calculate the density of states. Phys Rev Lett 2001;86(10):2050–2053.
37. Wang FG, Landau DP. Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram. Phys Rev E 2001;64(5):056101.
38. Smith GR, Bruce AD. A study of the multi-canonical Monte Carlo method. J Phys A 1995;28(23):6623–6643.
39. Fitzgerald M, Picard RR, Silver RN. Monte Carlo transition dynamics and variance reduction. J Stat Phys 2000;98(1–2):321–345.
40. Wang JS, Swendsen RH. Transition matrix Monte Carlo method. J Stat Phys 2002;106(1–2):245–285.
41. Shell MS, Debenedetti PG, Panagiotopoulos AZ. Flat histogram dynamics and optimization in density of states simulations of fluids. J Phys Chem B 2004;108(51):19748–19755.
42. Ramsey NF. Thermodynamics and statistical mechanics at negative absolute temperature. Phys Rev 1953;103:20.
43. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH. Hydrophobicity of amino-acid residues in globular-proteins. Science 1985;229(4716):834–838.
44. Dokholyan NV, Shakhnovich EI. Understanding hierarchical protein evolution from first principles. J Mol Biol 2001;312(1):289–307.
45. Mirny LA, Shakhnovich EI. How to derive a protein folding potential? A new approach to an old problem. J Mol Biol 1996; 264(5):1164–1179.