Structural evolution of proteinlike heteropolymers

Erik D. Nelson[*] and Nick V. Grishin
Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, 6001 Forest Park Boulevard, Room ND10.124, Dallas, Texas 75235-9050, USA

(Received 24 June 2014; revised manuscript received 5 October 2014; published 23 December 2014)

The biological function of a protein often depends on the formation of an ordered structure in order to support a smaller, chemically active configuration of amino acids against thermal fluctuations. Here we explore the development of proteins evolving to satisfy this requirement using an off-lattice polymer model in which monomers interact as low resolution amino acids. To evolve the model, we construct a Markov process in which sequences are subjected to random replacements, insertions, and deletions and are selected to recover a predefined minimum number of solid-ordered monomers using the Lindemann melting criterion. We show that polymers generated by this process consistently fold into soluble, ordered globules of similar length and complexity to small protein motifs. To compare the evolution of the globules with proteins, we analyze the statistics of amino acid replacements, the dependence of site mutation rates on solvent exposure, and the dependence of structural distance on sequence distance for homologous alignments. Despite the simplicity of the model, the results display a surprisingly close correspondence with protein data.

DOI: 10.1103/PhysRevE.90.062715 PACS number(s): 87.14.E−, 87.23.Kg, 87.18.Cf, 87.10.Tf

I. INTRODUCTION

Evolutionary change in proteins is the result of inherited alterations to genes selected to maintain a network of biochemical processes [1,2]. Most proteins that participate in these processes have been adapted to fold into ordered topologies [3], and their shapes reflect basic modes of genetic alteration in which conserved motifs are duplicated, combined, permuted, and edited [4,5], leading to emergent functional properties. The earliest proteins probably evolved from small cooperatively folding motifs of $\sim$30 amino acids [6–9], just large enough to stabilize chemically useful configurations of amino acids against thermal fluctuations [10]. Selection for folding and functional fidelity would tend to evolve mutationally stable structures that organize amino acids into energetically favorable patterns of solvent exposure [11–13]. It is obvious that these conditions would have had a strong influence on the development, or drift, of a motif in structure space [14–16]; however, it is difficult to obtain a thorough picture of this interplay from existing structure data. Many different polymer models have been used to describe protein evolution on this length scale [17–26]; however, the problems of structural drift and the dependence of mutation rates on local environment properties (e.g., on solvent exposure) have not been explored in a systematic way. Here, we employ a simple modification to one of these models to evolve soluble motifs of sufficient length and complexity to address these problems.

At the most basic level, a protein can be described as a flexible polymer, each monomer representing a common length of amino acids (the Kuhn segment length) which defines the length over which structural correlations persist in either direction along a typical chain of amino acids [27,28]. The interactions between segments of a protein can, in this way, be interpreted as lower resolution interactions between monomers, leading to a primitive model of a protein as an off-lattice chain of hydrophobic, hydrophilic, and charged beads. Recently, Lobkovsky and Koonin used this type of model to study protein evolutionary rates and have shown, using a physically based fitness criterion (folding probability), that such a model is capable of generating the universal form of the mesoscopic (whole protein) rate distribution [23]. Here, we consider a similar Langevin dynamics model in which pairs of monomers instead interact as low resolution amino acids via Miyazawa-Jernigan potentials (see the Appendix). As a result, monomers in the model are identified with the amino acids in a protein, while the structural correlations along amino acid chains are neglected. The polymers do not, of course, fold the same topologies as their protein counterparts; however, the ordered globules they do fold often contain small helical or strand structures due to the logic of the empirical potentials, and they are typically soluble (i.e., enclosed in a hydrophilic shell) for chain lengths $N \sim 30$ monomers or greater.

To evolve the model, we develop a Markov process [29] in which sequences are subjected to random replacements, insertions, and deletions and are selected to recover a solid-ordered nucleus of $\sim N/2$ monomers, sufficient to support a small binding site. The fidelity of this process is determined by folding $\mathcal{N} \gtrsim 100$ replicas of the mutated polymer on a parallel computer. The folding procedure consists of a series of temperature jumps which transfer the polymers between random coil, ordered globule, and melting temperatures roughly corresponding to the Gō model [30] for a small protein motif (see below). The structures recovered at the lower temperature are then collected, along with their $\mathcal{N}$ (energetically equivalent) mirror images, into an ensemble Γ, which is analyzed to determine the viability of the sequence.

In general, the energy landscape of a polymer replica can contain many deep energy basins. As the replicas are cooled they become trapped in these basins, so that Γ contains disparate clusters of structures. However, occasionally a sequence is encountered in which most or all of the replicas recover a single dominant energy basin (i.e., in each image space) corresponding to a narrow cluster of structures (Fig. 1). In order to select for this situation, we search for a structure $\mathbf{x}^\star \in \Gamma$ to

[*] nelsonerikd@gmail.com

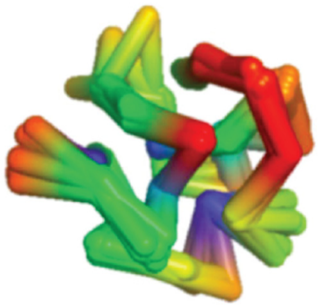

FIG. 1. (Color online) Sample of the ensemble, $\Delta\Gamma^{\star}$, corresponding to the dominant energy basin recovered by a typical viable sequence. The structures are obtained by folding many replicas of the sequence in parallel using methods described in the text. To indicate amino acid type, monomers are colored blue, light blue, blue-green, green, yellow, orange, and red, in order of increasing affinity to solvent.

represent the native ensemble of the mutated polymer, and we require that a significant fraction of the replicas recover structures close to $\mathbf{x}^{\star}$.

The reference structure $\mathbf{x}^{\star}$ plays a role analogous to the equilibrium (lattice) positions in a crystal in the usual formulation of the Lindemann parameter, which is defined as the root-mean-square displacement of a monomer from its equilibrium position. Ideally, we would select the reference structure to minimize these displacements using global structure alignments and the closest $\mathcal{N}$ structures in Γ. However, here it is necessary to allow for misfolding, and for weakly interacting (typically, uncharged hydrophilic) monomers on the surfaces of the globules. For this reason, we use a reductive procedure to align the structures [31] in which the most distant monomer pairings are removed iteratively until a nucleus of $2N/3$ optimally aligned pairs of monomers remains. These nuclear alignments are used to compute a nuclear Lindemann parameter, λ, for every structure $\mathbf{x}^{\mu} \in \Gamma$ using the closest $3\mathcal{N}/4$ remaining structures. The structure with minimal λ (in either image space) is selected as the reference structure, $\mathbf{x}^{\star}$. Finally, the multiple alignment with $\mathbf{x}^{\star}$ is used to compute Lindemann parameters, λ_j, for each monomer j individually, and the number of solid-ordered monomers is determined by the Lindemann criterion, $\lambda_j \gtrsim 0.15\, l$, where l is the length of a polymer link. If the number of solid-ordered monomers exceeds a specified value, the sequence is accepted; otherwise, it is rejected.

The Markov process is discussed in detail in Sec. II directly below. In simpler terms, this process selects sequences for which the root-mean-square distance between monomers in nuclear alignments with the reference structure, $\mathbf{x}^{\star}$, is typically less than $0.15\, l$ for more than $4\mathcal{N}/5$ of the replicas [24]. Evolved polymers exhibit sharp folding transitions similar to the "minimally frustrated" Gō model polymers studied by Jang and Zhou [30]. However, as in the Gō model, solid order is acquired very gradually below the transition so that the typical melting temperature of a nuclear monomer, $T^{\ddagger}$, is about one-third the transition temperature, $T^{\ddagger} \sim T^{\dagger}/3$. This situation is somewhat unrealistic from the standpoint of protein folding [32,33]; however, the model appears sufficient to describe the phenomena studied in this work.

Following Sec. II, we present and discuss our results. Our main objective in this paper is to compare the behavior of the model with the basic phenomenology of protein evolution. Below, we compute amino acid frequencies and replacement probabilities, the correlation between site mutation rates and exposed surface area, and the correlation between structural distance and sequence distance—all of which can be compared to protein data [34–39]. The level of agreement we obtain is somewhat surprising given the simplicity of the model and the neglect of the nucleotide coding sequence. To conclude, we briefly investigate the effect of the constraints on the length and complexity of evolved globules. Interestingly, we find that weaker constraints on the number of solid-ordered monomers tend to evolve larger and more thoroughly ordered globules that decrease in complexity over time [40]. We explore structural change along one such trajectory (a "propellerlike" motif) in the Supplemental Material [41].

II. METHODS

In this section we describe the Markov process and the schedule for mutations. The polymer model is described in the Appendix.

As noted above, the ensemble Γ for a mutant sequence is generated by folding $\mathcal{N} \gtrsim 100$ replicas of the mutated polymer on a parallel computer. The folding procedure equilibrates the polymers by Langevin dynamics at a series of temperatures

$$T_n = \left(\frac{T_1}{T_0}\right)^n T_0, \tag{1}$$

where $0 \leqslant n \leqslant 3$. The first temperature, $T_0 = 1.3\, T^{\star}$, corresponds to the random coil phase, where $T^{\star} = 302.15$ K is a reference temperature similar to the transition temperature of a viable sequence. The initial temperature jump transfers the system to the ordered-globule phase, similar to a folding trial, while the remaining jumps transfer the system to the solid-ordered phase. The amount of time allowed for equilibration at each temperature level is defined by the folding time estimate of Lin and Zewail [42],

$$\Delta t = N\left(\frac{3}{e}\right)^N \Delta t_0, \tag{2}$$

where Δt_0 denotes the typical time required for local changes to occur in the topology of a globule. The "sampling" time Δt_0 is set somewhat arbitrarily at 10 ps, which is a factor of 10 smaller than the time needed to thermally equilibrate a monomer in the polymer. (This number may seem rather small, but the entire folding process apparently corresponds to a weak form of kinetic control [23].) Finally, the structures recovered at $T_3 \sim T^{\star}/3$ are collected, along with their (energetically equivalent) mirror images in the ensemble Γ.

Each structure $\mathbf{x}^{\mu} \in \Gamma$ is considered as a possible reference structure, and the remaining structures, $\mathbf{x}^{\nu \neq \mu}$, are aligned to $\mathbf{x}^{\mu}$ by rigid rotation and translation [31]. Again, because the surfaces of the globules are liquid at T_3, we compute alignments using a smaller, nuclear set of monomers. Let $\mathbb{A}$ denote the set of monomer indices included in an alignment, initially including all indices (i.e., global alignment), and let

$\mathbb{A}^\star$ denote this nuclear group. The set $\mathbb{A}^\star$ is obtained by a reductive procedure, in which structures $\mathbf{x}^\mu$ and $\mathbf{x}^\nu$ are aligned to minimize the squared distance,

$$|\mathbf{x}^\mu - \mathbf{x}^\nu|^2_{\mathbb{A}} = \sum_{j\in\mathbb{A}} \left(\mathbf{x}^\mu_j - \mathbf{x}^\nu_j\right)^2, \tag{3}$$

and the index of the most distant monomer pairing is removed from $\mathbb{A}$ iteratively until $2N/3$ aligned pairs of monomers remain. To indicate this procedure, let $\|\mathbf{x}^\mu - \mathbf{x}^\nu\|$ denote the nuclear distance between structures and let $\|\mathbf{x}^\mu_j - \mathbf{x}^\nu_j\|$ denote nuclear-aligned distances between monomer positions.

The set of structures, $\mathbf{x}^{\nu\neq\mu}$, aligned to a structure $\mathbf{x}^\mu$ is then arranged in order of decreasing alignment quality (i.e., in order of increasing $\|\mathbf{x}^\mu - \mathbf{x}^\nu\|$). Let $\Delta\Gamma^\mu$ denote the first $3\mathcal{N}/4$ structures in this ordered set, and let $\mathbb{G}^\mu$ denote the corresponding set of structure indices. To define the fidelity of folding to $\mathbf{x}^\mu$, we compute *nuclear* Lindemann parameters,

$$\lambda(\mathbf{x}^\mu) = \left[\frac{2}{N\mathcal{N}} \sum_{\nu\in\mathbb{G}^\mu} \|\mathbf{x}^\mu - \mathbf{x}^\nu\|^2\right]^{1/2}. \tag{4}$$

The reference structure is then selected to minimize $\lambda(\mathbf{x}^\mu)$.

At this point, it would be natural to use the parameter $\lambda(\mathbf{x}^\star)$ to accept or reject a sequence, for example, using the Lindemann criterion $\lambda \leqslant 0.15\,l$. However, here we want to control the number of solid-ordered monomers, and since λ averages the displacements of monomers at different sites, it is somewhat inadequate for this purpose. To limit the displacements individually, we compute *monomeric* Lindemann parameters,

$$\lambda_j = \left[\frac{4}{3\mathcal{N}} \sum_{\nu\in\mathbb{G}^\star} \left\|\mathbf{x}^\star_j - \mathbf{x}^\nu_j\right\|^2\right]^{1/2}. \tag{5}$$

To determine the number of solid-ordered monomers, it simply remains to specify a threshold value of λ_j for solid order. Normally, this value is considered constant, e.g., so that a condition like $\lambda_j \gtrsim 0.15\,l$ can be used uniformly [30]. However, here the lengths of the polymers are changing, which affects the inherent accuracy of the alignment procedure [43]. To account for this effect, we define the melting point threshold by a function $\lambda^\ddagger(N)$ that scales with the radius of gyration of a collapsed polymer [44],

$$\frac{\lambda^\ddagger(N)}{\gamma} = -4.54 + 2.36\left(\frac{2N}{3}\right)^{1/3}. \tag{6}$$

Except for the factor of 2/3, the right-hand side of this expression is identical to the similarity threshold for protein alignments suggested by Maiorov and Crippen [44]. The factor of 2/3 accounts for the number of monomers used in nuclear alignments, and the parameter γ is selected so that $\lambda^\ddagger(N) = 0.16\,l$ for polymers of length $N = 30$. Finally, a monomer is considered solid ordered when $\lambda_j \leqslant \lambda^\ddagger$.

The mutation schedule consists of the following operations: First, a mutation type is selected, with probability 0.9 for replacements, 0.05 for insertions, and 0.05 for deletions. Next, an amino acid type is selected at random, and a random location along the current sequence is selected to apply the mutation (here, insertions are allowed at the ends of the sequence). The mutation is then applied to a copy of the current sequence, which is used to define the energy function for the polymer replicas. If the number of solid-ordered monomers exceeds a specified limit, the mutant sequence is accepted; otherwise it is rejected. The number of replicas, $3\mathcal{N}/4$, used to determine the reference structure in Eq. (4) was adjusted to obtain an acceptance rate of between 5% and 10%.

Below, we use the symbol τ to denote the number of time steps (mutation attempts) accumulated along a trajectory.

III. RESULTS

In this section, we examine statistical properties of the Markov model and compare our results to proteins. The results in this work are based on 14 trajectories consisting of about 5×10^3 mutation events in total. All trajectories start from sequences that were originally designed into one of several randomly selected, crumpled homopolymers using the protein engineering method [45] and then evolved to satisfy a specific constraint over a period of about 1×10^2 mutations. Below, sequences are evolved to maintain either a fixed number $\delta N \geqslant 15$ or a fixed fraction $\delta N \geqslant N/2$ of solid-ordered monomers, where the number of solid-ordered monomers, δN, is determined from the condition $\lambda_j \leqslant \lambda^\ddagger$. Half of the trajectories are evolved under each condition. To study longer, more cooperatively folding polymers, sequences with more than a single cysteine monomer are rejected. A number of alternate conditions were studied, and our methods are a product of these studies. In particular, we found that sequences evolved under "nuclear" constraints such as $\lambda \leqslant 0.15\,l$ or $\lambda \leqslant \lambda^\ddagger$ tend to "evaporate" or decay into smaller chain lengths. By limiting the displacements of monomers individually, similar to the biological requirement of maintaining a specific "active site," we consistently obtain sequences with lengths typical of a small motif.

We begin by comparing the amino acid statistics of the model to those of proteins. Because we neglect the nucleotide coding sequence, and because the globules are much smaller than typical protein domains, we do not expect close agreement with protein statistics. For this reason, we simply compare our results to the early data of Dayhoff *et al.* [34]. Our data are obtained from exact alignments of the sequences along each trajectory. Unless otherwise noted, figures represent the combined result of all 14 trajectories in our sample. Similar results are obtained using fewer trajectories.

Figures 2–4 provide a comparison of amino acid frequencies $p(\nu)$, mutabilities $m(\nu)$, and replacement probabilities, $p(\mu,\nu) = \mathrm{A}_{\mu\nu}/\sum_{\nu\neq\mu}\mathrm{A}_{\mu\nu}$, where $\mathrm{A}_{\mu\nu}$ is the number of replacements of amino acid type μ by amino acid type ν as defined by Dayhoff *et al.* [34]. In each figure, amino acid labels are arranged from left to right in order of increasing affinity to solvent (decreasing hydrophobicity).

The model frequencies reflect the roughly spherical structure of small soluble globules, in which hydrophilic (surface) monomers tend to outnumber hydrophobic (core) monomers. The largest departures from protein frequency data appear to be linked to the neglect of the coding sequence. For example, leucine, which is encoded by six codons, is underestimated by the model, while tryptophan (one codon) is overestimated. In other cases, departures from protein data seem to relate

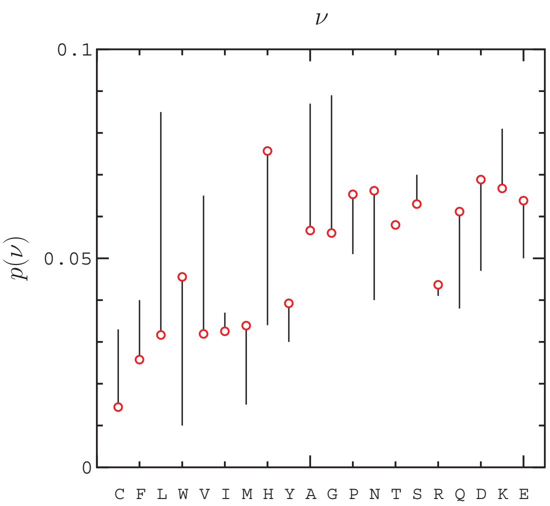

FIG. 2. (Color online) Comparison of model and empirical amino acid frequencies, $p(\nu)$. Model results are indicated by circles. Differences between model and empirical results are indicated by one-sided error bars. Amino acid labels on the lower axis are arranged in order of increasing affinity to solvent (see Appendix).

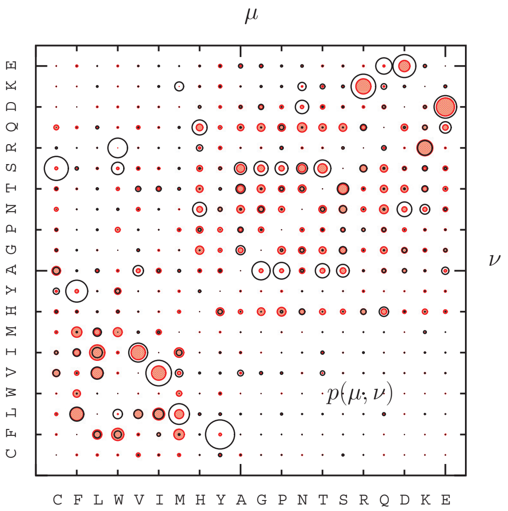

FIG. 4. (Color online) Comparison of model and empirical amino acid replacement probabilities, $p(\mu,\nu)$. The plot describes the replacement of amino acids of type μ by amino acids of type ν with probabilities indicated by circle radii. Solid red circles indicate model values; open black circles indicate empirical values computed by Dayhoff *et al.* (correlation coefficient ∼0.55).

to the lengths of the polymers, or to the lack of explicit secondary structure. For example, histidine is frequently recruited to bridge between hydrophobic and hydrophilic (mutually repulsive) regions of the flexible globules; in proteins, the partitioning of such regions is aided by secondary structure formation. Conversely, in proteins, glycine is often recruited to maintain flexibility in tight turns between secondary structures. In the model, glycine simply functions as a weakly hydrophobic monomer.

Interestingly, we obtain relatively good agreement in Fig. 4 for amino acid replacement probabilities (correlation coefficient ∼0.55). In recent work, Hormoz [46] obtained a similar level of agreement using a theoretical approach based on the protein engineering method using the empirical amino acid frequency distribution [34]. Together, these comparisons suggest that the environments of individual amino acids which determine the replacement probabilities are linked to generic folding principles.

To examine these environments, we compute mutability as a function of the level of solidlike order and exposed surface area. Figure 5 provides a plot of mutability, $m(\lambda)$, as a function of the monomeric Lindemann parameter λ_j. Here, $m(\lambda)$ is proportional to the number of mutations to

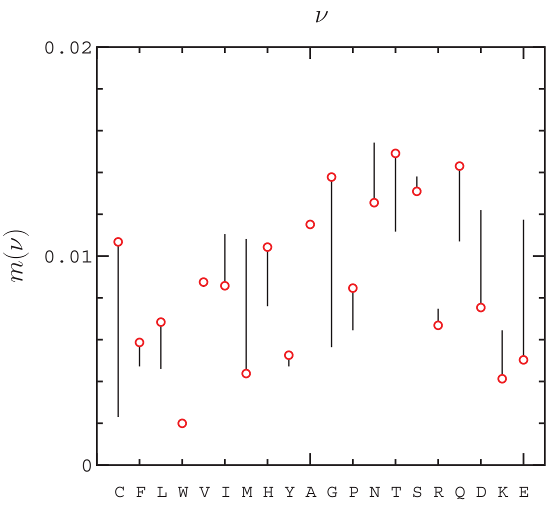

FIG. 3. (Color online) Comparison of model and empirical amino acid mutabilities, $m(\nu)$. Model results are indicated by circles. Differences between model and empirical results are indicated by one-sided error bars. Model results are scaled to the empirical mutability for alanine.

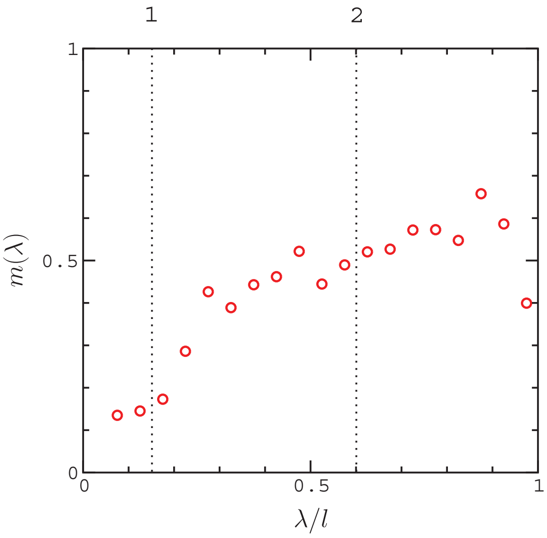

FIG. 5. (Color online) Mutability, $m(\lambda)$, of monomers with Lindemann parameter $\lambda_j = \lambda$. Dotted lines roughly indicate the transition (1) from solid to liquid and (2) from liquid to disordered phases within folded globules.

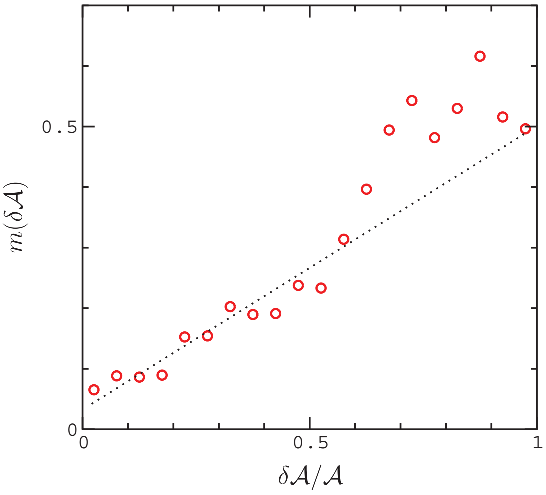

FIG. 6. (Color online) Mutability, $m(\delta\mathcal{A})$, of monomers with exposed surface area $\delta\mathcal{A}_j = \delta\mathcal{A}$. The exposed area of an isolated monomer is $\mathcal{A} = 4\pi\varrho^2$, where $\varrho = 0.65\,l$. The dashed line is a linear fit to the data in the region $\delta\mathcal{A}/\mathcal{A} < 0.7$ corresponding to one or more cross-chain (sphere) contacts.

monomers with Lindemann parameter $\lambda_j = \lambda$ divided by the number of time steps (attempts) that monomers with $\lambda_j = \lambda$ are exposed to mutation. The vertical scale in Fig. 5 is arbitrary. Dotted lines roughly indicate the transition (1) from solid to liquid and (2) from liquid to disordered regions in the folded globules. The plot is roughly linear below the melting line, increases more rapidly after the melting line is crossed, and reaches a plateau approaching the disorder line, beyond which sampling is inaccurate. As a result, the melting criterion roughly separates the mutation rates into liquidlike and solidlike phases analogous to the structural phases of monomers in polymer globules [30] and amino acid residues in proteins [47].

Figure 6 provides a plot of the mutability, $m(\delta\mathcal{A})$, as a function of monomeric exposed surface area, $\delta\mathcal{A}_j$. Here, $m(\delta\mathcal{A})$ is proportional to the number of mutations to monomers with $\delta\mathcal{A}_j = \delta\mathcal{A}$ divided by the number of time steps (attempts) that monomers with $\delta\mathcal{A}_j = \delta\mathcal{A}$ are exposed to mutation. To compute $\delta\mathcal{A}_j$, the reference structure of the polymer is viewed as a set of interpenetrating spheres of radius ϱ, and the exposed area of a monomer is defined as the part of its surface not enclosed by any other sphere [48]. The exposed area is computed by coating each sphere with a very large number of equally spaced points [49], and the fractional area, $\delta\mathcal{A}/\mathcal{A}$, is estimated as the number of exposed points divided by the total number of points on the surface of a monomer. To generate Fig. 6, we select a sphere radius $\varrho = 0.65\,l$ to obtain complete burial for monomers in the interior of the globules. The dotted line is a fit to the data in the region $\delta\mathcal{A}/\mathcal{A} < 0.7$, roughly corresponding to monomers with one or more cross-chain contacts. Within this region, $m(\delta\mathcal{A})$ is linearly correlated with $\delta\mathcal{A}$ (correlation coefficient ~0.95), consistent with mutation rates in proteins [35,36].

Next, we examine structural change using alignments of "homologous" sequences along the Markov trajectories to establish correspondence between monomers in structural alignments. In Fig. 7, we plot the typical distance, $\mathcal{D}(n)$, between structures as a function of mutational distance, n. To compute $\mathcal{D}(n)$, we interpret the Markov process as a sequence of random flights [50–52] connecting pairs of structures $\mathbf{x}^\star(\tau)$ and $\mathbf{x}^\star(\tau')$ recorded at times τ and $\tau' \geqslant \tau$ just preceding mutation events. The distance $\mathcal{D}(n)$ is the average over end-to-end distances $\mathcal{D}(\tau,\tau')$ for multiple flights with exactly $n(\tau,\tau') = n$ mutations. In Fig. 7, the average is restricted to trajectories evolved under the condition $\delta N \geqslant 15$. The dotted line is a linear fit to the data (circles) following the linear correlation observed by Illergard *et al.* [38] for "core alignments" of homologous proteins. To approximate the alignment procedure used in that work, distances are computed by the reductive method in Eq. (3). The reduction is continued to a constant 16 aligned monomers, and $\mathcal{D}(\tau,\tau')$ is defined as the resulting root-mean-square distance between structures. On average, most of the sites participating in the alignment of a structure $\mathbf{x}^\star(\tau)$ with $\mathbf{x}^\star(\tau')$ are preserved in later alignments of $\mathbf{x}^\star(\tau)$ with structures $\mathbf{x}^\star(\tau'' \gg \tau')$ (Fig. 8). As the number of sites required in alignments is increased, the data can be described accurately by a power law, $\mathcal{D}(n) \simeq A + Bn^\alpha$, with exponent $\alpha < 1$ (not shown).

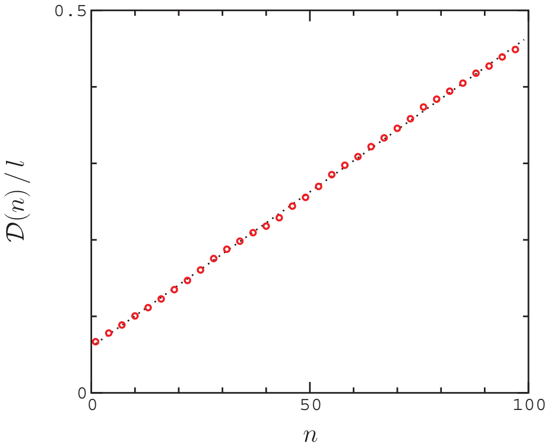

FIG. 7. (Color online) Aligned distance between structures, $\mathcal{D}(n)$, as a function of mutational distance, n. The dotted line is a linear fit to the data (points are plotted every fourth mutation for clarity in the figure).

In Fig. 9, we plot $\mathcal{D}(\tau,\tau')$ as a function of the percentage of nonidentical amino acids, $q(\tau,\tau')$. Here, the data points represent averages over paths with $q(\tau,\tau') = q$. The solid line is an exponential fit to the data following the empirical result of Chothia and Lesk for homologous proteins [37]. The dotted line is a power-law fit, $\mathcal{D}(q) \simeq A + B\,q^\alpha$, with exponent $\alpha > 1$. As the number of sites required in alignments is increased, the quality of the power-law fit is maintained, while the exponential fit deteriorates. The plots in Figs. 7 and 9 are in close agreement with those in Refs. [38] and [37]. A more complete account of these subjects will be provided in future work [53].

To conclude, we briefly explore the growth and complexity of evolved globules. As indicated above, both conditions, $\delta N \geqslant 15$ and $\delta N \geqslant N/2$, evolve soluble globules; however, the less restrictive constraint, $\delta N \geqslant 15$, leads to more interesting results, so we again focus on this situation below.

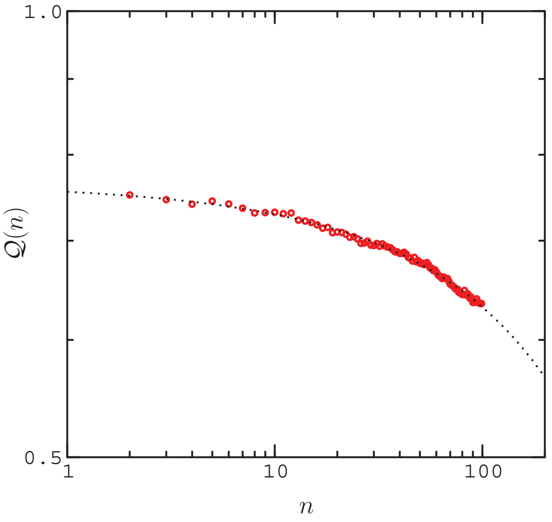

FIG. 8. (Color online) Fraction of sites, $\mathcal{Q}(n)$, in the alignment of sequential structures, $\mathbf{x}^\star(\tau)$ and $\mathbf{x}^\star(\tau')$, preserved in later alignments of $\mathbf{x}^\star(\tau)$ with structures $\mathbf{x}^\star(\tau'')$. Data points represent averages over paths with $n(\tau,\tau'') = n$. The dotted line is a power-law fit to the data.

Figure 10 plots polymer length $N(\tau)$ for several trajectories evolved under this condition, where τ is the time measured in mutation attempts. The paths in Fig. 10(a) start from a single structure, which was first evolved from the most designable structure in our sample of crumpled homopolymers. The paths in Fig. 10(b) stem from three different target structures that were more difficult to design. Most paths evolved under the condition $\delta N \geqslant 15$ are able to explore a size range of $35 \leqslant N \leqslant 45$ monomers. Interestingly, following the rapid increase in length along path 1 in Fig. 10(a), a significant fraction of polymers far exceeds this constraint, sometimes containing as many as $3N/4$ ordered monomers, reminiscent of the critical nucleation of a liquid droplet. (We explore structural evolution along this path in the Supplemental Material [41].) By contrast, all paths evolved under the more restrictive condition $\delta N \geqslant N/2$ (not shown) are localized in an envelope of about $26 \leqslant N \leqslant 32$ monomers.

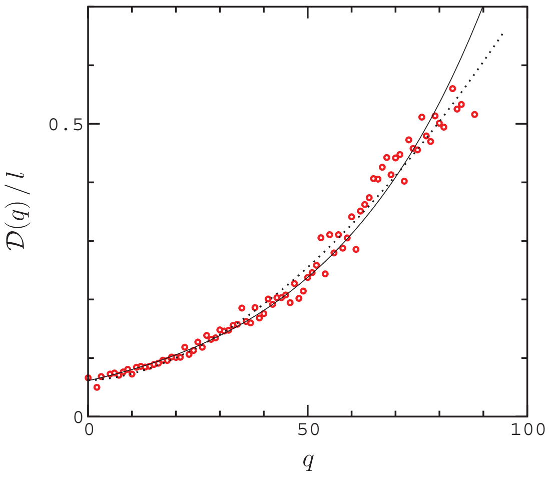

FIG. 9. (Color online) Aligned distance between structures, $\mathcal{D}(q)$, as a function of the percentage of nonidentical amino acids, q. The data used to compute $\mathcal{D}(q)$ are the same as in Fig. 7. The solid line is an exponential fit to the data; the dotted line is a power-law fit.

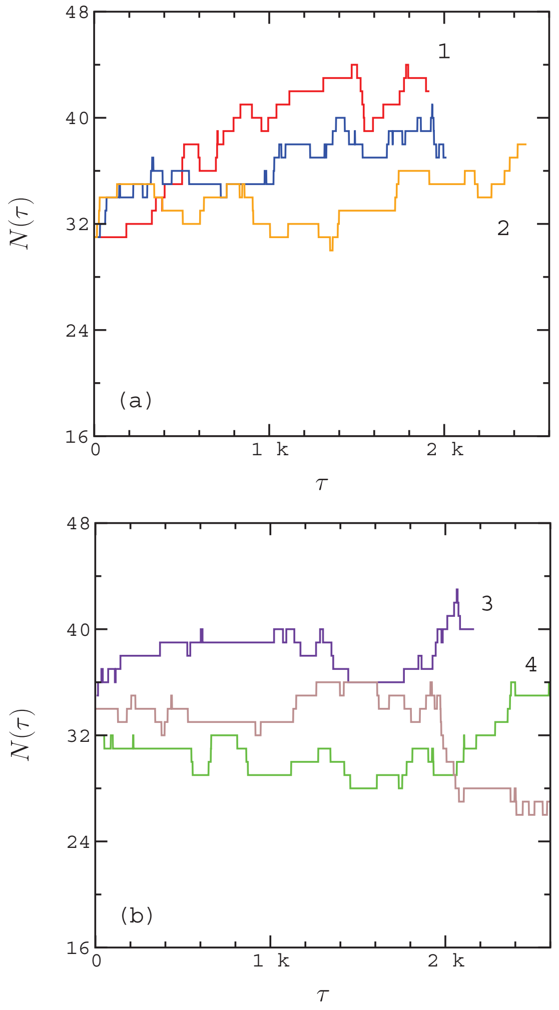

FIG. 10. (Color online) Evolution of polymer length, $N(\tau)$, for sequences evolved under the condition $\delta N \geqslant 15$. The lower axis, τ, in each panel denotes elapsed time along a trajectory measured in mutation attempts. (a) The system evolves from a single sequence which stems from a designable structure. (b) The system evolves from multiple sequences which stem from less designable structures. Paths are numbered for later reference in Fig. 11.

To measure structural complexity, we compute the average length of loops formed in folded globules, or “contact order” [54]:

$$\langle \ell(\mathbf{x}) \rangle = \sum_{i,j \geqslant i+2} |i-j|\theta(2\varrho - |\mathbf{x}_i - \mathbf{x}_j|) \Bigg/ \sum_{i,j \geqslant i+2} \theta(2\varrho - |\mathbf{x}_i - \mathbf{x}_j|), \qquad (7)$$

where $\theta(x)$ is the Heaviside step function and 2ϱ defines the range for cross-chain contacts between monomers. Low contact order typically indicates structures that fold more hierarchically and efficiently. In Fig. 11, we plot $\langle \ell(\tau) \rangle$ along four of the trajectories in Fig. 10. (Two of the paths are omitted for clarity in the figure.) Each data point denotes the average

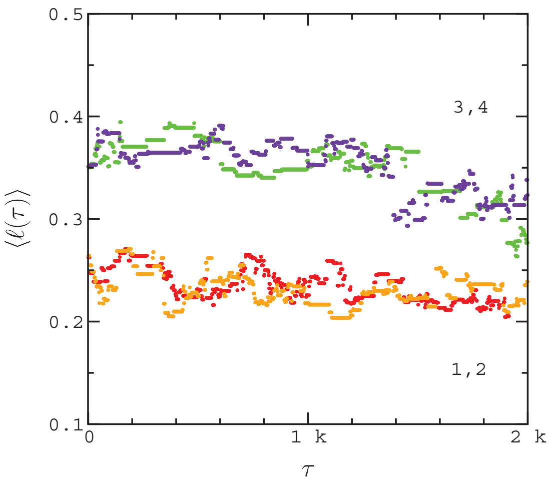

FIG. 11. (Color online) Evolution of complexity (average loop length), $\langle \ell(\tau) \rangle$, for sequences evolved under the condition $\delta N \geqslant 15$. Each point represents an average over a small group of structures in $\Delta\Gamma^\star(\tau)$ closest to $\mathbf{x}^\star(\tau)$. Path colors and numbers correspond with Fig. 10.

over a small group of structures closest to $\mathbf{x}^\star(\tau)$. All paths in exhibit a trend of decreasing contact order, consistent with the result of Debes *et al.* for resurrected protein domains [40]. By contrast, paths generated under the condition $\delta N \geqslant N/2$ are usually more erratic, punctuated by jumps in $\langle \ell(\tau) \rangle$ in which disordered loops or end segments are suddenly detached or incorporated into the ordered part of the globule.

IV. SUMMARY

To summarize, polymers evolved under the weaker constraint $\delta N \geqslant 15$ tend to evolve larger and more thoroughly ordered structures, apparently due to the greater availability of fortuitous mutations (i.e., greater connectivity in the space of viable sequences). It is difficult to tell how strongly this result depends on the designability of the initial structure from the small sample studied above. However, our results are consistent with the idea that "forced" or directed polymer evolution leads to slower relaxation [20,55].

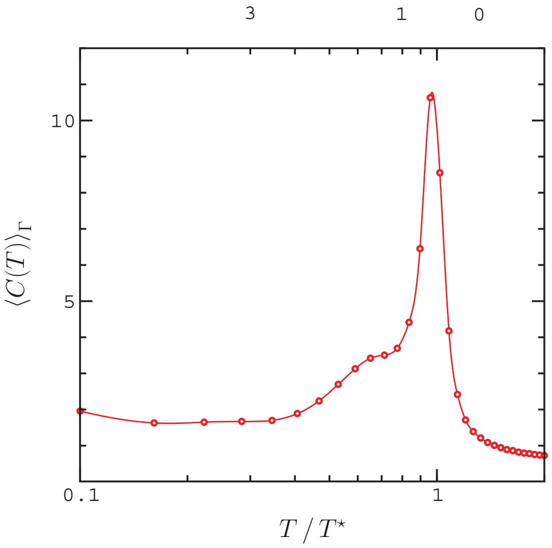

FIG. 12. (Color online) Replica average specific heat function, $\langle C(T) \rangle_\Gamma$, for an evolved sequence of length $N = 35$ monomers. Numbers along the top edge of the figure indicate equilibration temperatures in Eq. (1).

Although the phenomenology of the model bears a striking resemblance to proteins, the lack of explicit secondary structure presents certain difficulties in comparing the model to protein data. Typically, the procedures used to relate protein structures in phylogenetic studies first depend on the identification of common regions of secondary structure. As a result, the alignment procedures used by Chothia and Lesk and Illergard *et al.* cannot be applied literally to the model. However, the fact that we can reproduce the same behavior using a similar method suggests that this behavior may be a generic feature of proteinlike polymers [in particular, once structures are compared by a reductive alignment method like that in Eq. (3)]. Protein mutational data are, of course, a product of evolving populations of genes, not a Markovian chain. However, the fact that mutabilities in the model exhibit the same dependence on exposure as mutation rates in proteins suggests this behavior is linked to the fidelity of folding proteinlike polymers. Finally, we remark that although it may have been simpler to constrain the formation of a binding site explicitly, the methods developed above make it possible to explore the interplay between folding and functional requirements in more realistic models that include the genetic code.

ACKNOWLEDGMENTS

The authors would like to thank Liqiang Wang and Shenghua Wan for help with the computations.

APPENDIX

The model is a low resolution map of an amino acid chain onto an idealized polymer via Miyazawa-Jernigan potentials. Adjacent monomers along the polymers are linked by a potential of the form

$$U^{\text{link}}(r) = \frac{\kappa}{2}(r - l)^2, \tag{A1}$$

where r is the distance between monomers, l is the equilibrium length of a link, and κ is a constant (see below). The sequence dependent potentials are based on the Morse function [44],

$$\mu(r) = \exp[-2\alpha(r - l)] - 2\exp[-\alpha(r - l)]. \tag{A2}$$

To construct the potentials, we separate the Morse function into components,

$$\mu^{r<l}(r) = \mu(r)\theta(l - r) \tag{A3}$$

and

$$\mu^{r\geqslant l}(r) = \mu(r)[1 - \theta(l - r)], \tag{A4}$$

where $\theta(l - r)$ is the Heaviside step function. The potentials for attractive and repulsive monomer pairings are then defined as

$$U^{\epsilon'<0}(r) = \epsilon\mu^{r<l}(r) + (\epsilon + \epsilon')\theta(l - r) - \epsilon'\mu^{r\geqslant l}(r) \tag{A5}$$

and

$$U^{\epsilon'\geqslant 0}(r) = \epsilon\mu^{r<l}(r) + \epsilon\theta(l-r) + \epsilon'\exp[-\alpha(r-l)], \tag{A6}$$

respectively. Each potential consists of an excluded volume part, modulated by the constant ϵ, and a sequence dependent part, modulated by the parameter ϵ'. The parameter ϵ' takes on different values $\epsilon\,\mathrm{M}_{\mu\nu}/\overline{\mathrm{M}}$ depending on the type of interaction, where $\overline{\mathrm{M}} = \sum_{\mu,\nu\geqslant\mu}|\mathrm{M}_{\mu\nu}|/210$ is the average strength of an interaction and $\mathrm{M}_{\mu\nu}$ denotes the scaled Miyazawa-Jernigan matrix developed by Betancourt and Thirumalai [56]. The diagonal elements of $\mathrm{M}_{\mu\nu}$ define the hydrophobicity scale used to arrange amino acid labels in the figures.

To simulate folding, we integrate the Langevin equation using the method of van Gunsteren and Berendsen [57] with monomer mass $m = 1\times10^{-22}$ g, friction coefficient $\gamma = 50$ ps^{-1}, and integration time step $\Delta t = 0.005$ ps, leading to diffusive kinetics. The potentials are defined by the parameters $l = 1$, $\kappa = 120$, $\alpha = 8$, and $\epsilon = 3/2$ with length in angstroms, and energy in units of $k_BT^\star$, where k_B is Boltzmann's constant and $T^\star = 302.15$ K. The constant ϵ is selected to locate the specific heat peaks of viable sequences near $T^\star$. Figure 12 plots the replica average specific heat, $\langle C(T)\rangle_\Gamma$, for an evolved sequence, where [30]

$$C(T) = \frac{\langle E^2\rangle - \langle E\rangle^2}{(k_BT)^2}, \tag{A7}$$

E is the energy of a replica, and braces, $\langle\rangle$, denote time averages. Numbers along the top edge of the plot indicate equilibration temperatures in Eq. (1).

[1] D. de Juan, F. Pazos, and A. Valencia, Nat. Rev. Genet. **14**, 249 (2013).

[2] E. V. Koonin, Nucleic Acids Res. **37**, 1011 (2009).

[3] V. N. Uversky, Biochim. Biophys. Acta **1834**, 932 (2013).

[4] N. V. Grishin, J. Struct. Biol. **134**, 167 (2001).

[5] M. Wang and G. Caetano-Anolles, Structure **17**, 66 (2009).

[6] J. Soeding and A. N. Lupas, Bioessays **25**, 837 (2003).

[7] A. N. Lupas, C. P. Ponting, and R. B. Russell, J. Struct. Biol. **134**, 191 (2001).

[8] N. Berezovsky and E. N. Trifonov, Comp. Funct. Genomics **3**, 525 (2002).

[9] A. Goncearenco and N. Berezovsky, Bioinformatics **27**, 2368 (2011).

[10] A. Fernandez, Phys. Chem. Chem. Phys. **1**, 4347 (1999).

[11] J. Miller, C. Zeng, N. S. Wingreen, and C. Tang, Protein Struct. Funct. Genet. **47**, 506 (2002).

[12] E. G. Emberly *et al.*, Protein Struct. Funct. Genet. **47**, 295 (2002).

[13] A. Fernandez, *Transformative Concepts for Drug Design: Target Wrapping* (Springer, Heidelberg, 2010).

[14] S. Krishna and N. V. Grishin, Bioinformatics **21**, 1308 (2005).

[15] L. N. Khinch and N. V. Grishin, Curr. Opin. Struct. Biol. **12**, 400 (2002).

[16] I. Coluzza *et al.*, PLoS One **7**, e34228 (2012).

[17] D. M. Taverna and R. M. Goldstein, Biopolymers **53**, 1 (2000).

[18] E. Bornberg-Bauer and H. S. Chan, Proc. Natl. Acad. Sci. U.S.A. **96**, 10689 (1999).

[19] Y. Cui, W. H. Wong, E. Bornberg-Bauer, and H. S. Chan, Proc. Natl. Acad. Sci. U.S.A. **99**, 809 (2002).

[20] J. D. Bloom, C. O. Wilke, F. H. Arnold, and C. Adami, Biophys. J. **86**, 2758 (2004).

[21] K. B. Zeldovich and E. I. Shakhnovich, Annu. Rev. Phys. Chem. **59**, 105 (2008).

[22] K. B. Zeldovich, P. Chen, B. E. Shakhnovich, and E. I. Shakhnovich, PLoS One **3**, e139 (2007).

[23] A. E. Lobkovsky and E. V. Koonin, Proc. Natl. Acad. Sci. U.S.A. **107**, 2983 (2010).

[24] A. E. Lobkovsky and E. V. Koonin, PLoS Comput. Biol. **7**, e1002302 (2011).

[25] A. R. Khokhlov and P. G. Khalatur, Curr. Opin. Mat. Sci. **8**, 3 (2004).

[26] A. V. Berezkin, P. G. Khalatur, A. R. Khokhlov, and P. Reineker, New J. Phys. **6**, 44 (2004).

[27] A. Y. Grosberg and A. R. Khokhlov, *Statistical Physics of Macromolecules* (American Institute of Physics, Woodbury, NY, 1994).

[28] V. Receveur-Brechot and D. Durand, Curr. Protein Pept. Sci. **13**, 55 (2012).

[29] M. Kimura, *The Neutral Theory of Evolution* (Cambridge University Press, New York, 1983).

[30] H. Jang, C. K. Hall, and Y. Zhou, Biophys. J. **82**, 646 (2002).

[31] P. Liu, D. K. Agrafiotis, and D. L. Theobald, J. Comput. Chem. **31**, 1561 (2009).

[32] H. Kaya and H. S. Chan, Proteins **40**, 637 (2000).

[33] M. R. Ejtehadi, S. P. Avall, and S. S. Plotkin, Proc. Natl. Acad. Sci. U.S.A. **101**, 15088 (2004).

[34] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt, *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Washington, DC, 1972), pp. 345–352.

[35] D. C. Ramsey, M. P. Scherrer, T. Zhou, and C. O. Wilke, Genetics **188**, 479 (2011).

[36] E. A. Franzosa and X. Xia, PLoS One **7**, e46602 (2012).

[37] C. Chothia and A. M. Lesk, EMBO J. **5**, 823 (1986).

[38] K. Illergard, D. H. Ardell, and A. Elofsson, Proteins **77**, 499 (2009).

[39] Z. Zhang, Y. Wang, L. Wang, and P. Gao, PLoS One **5**, e14316 (2010).

[40] C. Debes, M. Wang, G. Caetano-Anolles, and F. Grater, PLoS Comput. Biol. **9**, e1002861 (2012).

[41] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.90.062715 for images of the folded structures, $\mathbf{x}^\star(\tau)$, and ensembles, $\Delta\Gamma^\star(\tau)$, selected along the trajectory of a "propellerlike" motif.

[42] M. M. Lin and A. H. Zewail, Proc. Natl. Acad. Sci. U.S.A. **109**, 9851 (2012).

[43] V. N. Maiorov and G. M. Crippen, J. Mol. Biol. **235**, 625 (1994).

[44] A. Milchev, W. Paul, and K. Binder, J. Chem. Phys. **99**, 4786 (1993).

[45] E. I. Shakhnovich and A. M. Gutin, Proc. Natl. Acad. Sci. U.S.A. **90**, 7195 (1993).

[46] S. Hormoz, Sci. Rep. **3**, 2919 (2013).

[47] Y. Zhou, D. Vitkup, and M. Karplus, J. Mol. Biol. **285**, 1371 (1999).

[48] R. Lustig, Mol. Phys. **59**, 195 (1986).

[49] M. Tegamark, Astrophys. J. Lett. **470**, L81 (1996).

[50] N. V. Grishin, J. Mol. Evol. **45**, 359 (1997).

[51] A. M. Gutin and A. Ya. Badretdinov, J. Mol. Evol. **39**, 206 (1994).

[52] J. H. Gillespie, Proc. Natl. Acad. Sci. U.S.A. **81**, 8009 (1984).

[53] E. D. Nelson and N. V. Grishin (unpublished).

[54] D. N. Ivankov *et al.*, Protein Sci. **12**, 2057 (2003).

[55] P. Šulc, A. Wagner, and O. C. Martin, J. Bioinf. Comput. Biol. **8**, 1027 (2010).

[56] M. R. Betancourt and D. Thirumalai, Protein Sci. **8**, 361 (1999).

[57] W. F. van Gunsteren and H. J. C. Berendsen, Mol. Phys. **45**, 637 (1982).