PHYSICAL REVIEW E **91**, 060701(R) (2015)

Anomalous diffusion in neutral evolution of model proteins

Erik D. Nelson* and Nick V. Grishin

Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, 6001 Forest Park Blvd., Room ND10.124, Dallas, Texas 75235-9050, USA

(Received 10 February 2015; published 1 June 2015)

Protein evolution is frequently explored using minimalist polymer models, however, little attention has been given to the problem of structural drift, or diffusion. Here, we study neutral evolution of small protein motifs using an off-lattice heteropolymer model in which individual monomers interact as low-resolution amino acids. In contrast to most earlier models, both the length and folded structure of the polymers are permitted to change. To describe structural change, we compute the mean-square distance (MSD) between monomers in homologous folds separated by n neutral mutations. We find that structural change is episodic, and, averaged over lineages (for example, those extending from a single sequence), exhibits a power-law dependence on n. We show that this exponent depends on the alignment method used, and we analyze the distribution of waiting times between neutral mutations. The latter are more disperse than for models required to maintain a specific fold, but exhibit a similar power-law tail.

DOI: 10.1103/PhysRevE.91.060701 PACS number(s): 87.14.E−, 87.23.Kg, 87.18.Cf, 87.10.Tf

In recent work, we investigated the evolution of small protein motifs using a simplified off-lattice heteropolymer model [1]. The model is analogous to a commonly used lattice model in which individual monomers interact as low-resolution amino acids [2] and evolves according to a Markov process in which sequences are subjected to replacements, insertions, and deletions, and are selected to fold reproducibly into ordered globules capable of supporting a small binding site against thermal fluctuations. The Markov process describes the gradual fixation of selectively neutral mutations in a population when the typical time to fixation or loss is smaller than the time between mutation events, and is commonly used to approximate population dynamics in evolutionary models [3]. Earlier, we found that polymers evolved by this process fold into soluble globules of similar length and complexity to small protein motifs. The folded states, or ensembles of the polymers are often less well ordered than those of small proteins, however, many of the results we obtain from the model are in good agreement with protein data [4–6]. In particular, rates for structural drift as a function of mutational distance and sequence identity agree closely with proteins when structural distance is measured using similar alignment methods. Here, we continue our analysis of structural change in the Markov model from the standpoint of a conventional diffusion problem [7–10].

Below, we first provide a summary our results and describe how they were obtained. The polymer model and the Markov process are described in Ref. [1] and in the Appendix to this Rapid Communication.

A trajectory, or *lineage* generated by the Markov process can be pictured as a series of flights between nodes of a neutral network [11], each node corresponding to a viable sequence, and each edge connecting a pair of sequences linked by a single (neutral) mutation. At any node visited along a lineage, the probability of obtaining a neutral mutation in a single iteration of the Markov process is a constant, independent of time, determined by the local connectivity of the network and the attempt frequency for each type of mutation. As a result, the waiting time distribution at each node of the network is binomial. However, because neutral connectivity varies from point to point in the network, the ticks of the "polymer clock" can become episodic [12]. Bastolla was the first to recognize the significance of this effect for proteins [13], and has argued that variations in neutral, or nearly neutral connectivity can account for the observed dispersion in mutation rates relative to the molecular clock (i.e., binomial, or Poisson) approximation [14]. Later, Wilke considered evolution of an explicit population of polymers to explore the range of validity of Bastolla's model and arrived at similar conclusions [15]. In both models, polymers were required to fold and maintain a specific structure. Here, both the length and the folded structures of the polymers are permitted to change.

To describe the polymer clock, we compute the probability, $P(\mathcal{T} \geqslant \tau)$, of waiting times $\mathcal{T} \geqslant \tau$ both for individual lineages and for combinations of lineages such as the star phylogeny [16] depicted in Fig. 1. We find that $P(\mathcal{T} \geqslant \tau)$ typically follows a power law:

$$P(\mathcal{T} \geqslant \tau) \simeq \frac{1}{1 + (\tau/\tau_m)^{\beta}} \tag{1}$$

with exponents $\beta \gtrsim 1.3$ well into the episodic regime, where the index of dispersion (variance divided by the mean) of the distribution function,

$$P(\mathcal{T}) = -\frac{\partial}{\partial \tau} P(\mathcal{T} \geqslant \tau), \tag{2}$$

is infinite. Using Eq. (2), we fit Eq. (1) to the data reported by Bastolla and obtain a similar result: Although the scale parameter, τ_m, differs in our model (our results are more disperse than those of Bastolla), the distribution function decays by a similar power of τ/τ_m.

Next, we consider structural change along lineages rationalized as flights between folded structures coordinated by the neutral network [7,8]. In each iteration of the Markov process, a structure ensemble Γ is generated by folding $\mathcal{N} \sim 100$ polymer replicas on a parallel computer (see Appendix). From this ensemble, a smaller ensemble, $\Delta\Gamma^{\star}$, consisting

*nelsonerikd@gmail.com

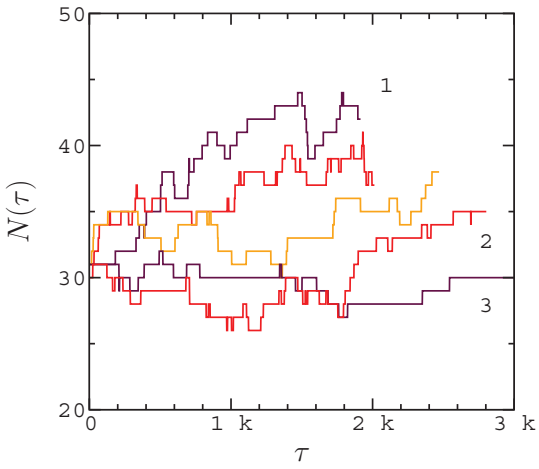

FIG. 1. (Color online) Polymer length, $N(\tau)$, for several lineages initiated from the same sequence. τ is the number of iterations of the Markov process (mutation attempts). Lineage 1 is evolved under the constraint $\delta N \geqslant 15$, lineage 2 under the constraint $\delta N \geqslant N/2$, and lineage 3 under the constraint $\delta N \geqslant 16$, where δN is the number of solid-ordered monomers in the folded ensemble of a polymer as determined by the Lindemann melting criterion (see Appendix). Unmarked lineages are evolved under the constraint $\delta N \geqslant 15$.

of $3\mathcal{N}/4$ replicas is selected to define the dominant energy basin recovered by the replicas. This procedure also selects a *reference fold*, $\mathbf{x}^\star$, closest to the center of the ensemble $\Delta\Gamma^\star$. Each neutral mutation is then viewed as a flight between reference folds, analogous to a simple random walk.

Let $\mathbf{x}(\tau)$ denote the coordinates of the reference fold in a particular lineage after τ neutral mutations (for clarity, we drop the superscript on the reference fold in the discussion below). Because indels are permitted, polymers at distant points along a lineage can have different lengths. For this reason, it is necessary to establish a correspondence, or homology between the monomers at τ and τ' in order to compare the structures $\mathbf{x}(\tau)$ and $\mathbf{x}(\tau')$. To establish monomer homology, we construct a complete alignment of the sequences along each lineage. Each alignment results in an array of (gapped) sequences, $\mathbf{s}(\tau)$, of equal length (i.e., including gap positions). A pair of monomers in structures $\mathbf{x}(\tau)$ and $\mathbf{x}(\tau')$ are considered homologous when their positions in sequences $\mathbf{s}(\tau)$ and $\mathbf{s}(\tau')$ are aligned. Let $\mathbb{A}(\tau,\tau')$ denote the set of homologous positions in a pair of homologous sequences (i.e., sequences from the same lineage). To measure the distance between $\mathbf{x}(\tau)$ and $\mathbf{x}(\tau')$, we compute a structural alignment by rotation, translation, and reflection to minimize the mean-square distance (MSD) between monomers in a *subset* of homologous sequence positions $\mathbb{Q}(\tau,\tau') \subset \mathbb{A}(\tau,\tau')$. The distance between $\mathbf{x}(\tau)$ and $\mathbf{x}(\tau')$ is defined as the resulting root-mean-square distance (RMSD).

Below, we consider two basic methods to select the set $\mathbb{Q}(\tau,\tau')$: In method (i), the set $\mathbb{Q}(\tau,\tau')$ is selected from $\mathbb{A}(\tau,\tau')$ iteratively to minimize the distance between $\mathbf{x}(\tau)$ and $\mathbf{x}(\tau')$, but the cardinality of $\mathbb{Q}(\tau,\tau')$ (i.e., the number of positions ultimately compared to measure MSD) is held constant along a lineage. This method is similar to the "core alignment" procedure used by Illergard *et al.* [6] and Chothia and Lesk [5] to align protein domains, and generates similar results for RMSD as a function of evolutionary distance (i.e., neutral mutations) and sequence identity if the number of monomers compared in alignments is comparable to the number of ordered monomers required in folded ensembles [1]. A more sophisticated version of this method is outlined in Ref. [17]. In method (ii), a single set of "tracer" positions is selected for the entire lineage from the set $\mathbb{M} = \bigcap_{\tau=1}^{2\mathcal{T}} \mathbb{A}(0,\tau)$, for comparison with theoretical models [7,8]. Here, $\mathbb{M}$ consists of all positions that are conserved (i.e., have not encountered a deletion) along an interval of length $2\mathcal{T}$, where $\mathcal{T}$ is about half the length of typical lineage (see below). The set $\mathbb{Q}$ is defined as the set of conserved, hydrophobic and charged positions in the ancestral sequence, $\mathbf{s}(0)$, which typically occupy the nucleus of compared monomers in ensembles $\Delta\Gamma^\star$. An alignment of structures by method (ii) along lineage 3 is provided in the Supplemental Material file [18].

To describe structural diffusion along a lineage, we compute averages over sub-paths,

$$\langle \Delta x^2 \rangle(n) = \frac{1}{\mathcal{T}} \sum_{\tau=0}^{\mathcal{T}-1} \Delta x^2(\tau, \tau+n), \tag{3}$$

where $\Delta x^2(\tau,\tau+n)$ is the mean-square distance between $\mathbf{x}(\tau)$ and $\mathbf{x}(\tau+n)$ computed by either method, $n \leqslant \mathcal{T}$ denotes the length of a sub-path measured in neutral mutations, and $\mathcal{T} = 100$ corresponding to about three mutations per monomer. For a number of lineages studied in this work, $\langle \Delta x^2 \rangle$ can be fit by a power law [19],

$$\langle \Delta x^2 \rangle \simeq A + B\, n^\alpha \tag{4}$$

over a substantial part of the interval $n \leqslant \mathcal{T}$, however, more often $\langle \Delta x^2 \rangle$ exhibits a mixture of behaviors. To typify structural drift for a particular method, we average over groups of lineages

$$\overline{\langle \Delta x^2 \rangle} = \frac{1}{L} \sum_{l=1}^{L} \langle \Delta x^2 \rangle_l, \tag{5}$$

where L is the number of lineages considered. Below, we average over the lineages depicted in Fig. 1, and those evolved under the constraints $\delta N \geqslant 15$ and $\delta N \geqslant N/2$, respectively, where δN is the number of solid-ordered monomers in the folded ensemble of a sequence (see Appendix). Averages over groups of lineages are fit closely by Eq. (4), however, the exponents obtained for each method are quite different, and both depart from case of normal diffusion considered in theoretical models [7,8] : For method (i), structural change is super-diffusive, with exponents $\alpha \sim 1.6$, suggesting Lévy-like behavior. [19]. For method (ii), structural change is sub-diffusive, with exponents $\alpha \sim 0.5$, similar to kinetic diffusion of small proteins [20,21].

A sample of our results for individual lineages are provided in Figs. 2–4. The data included in these figures are obtained from a set of 14 lineages generated in Ref. [1]. Half of the lineages were generated under each of the two constraints on δN noted above. Each lineage begins from a sequence folding to one of five distinct structures.

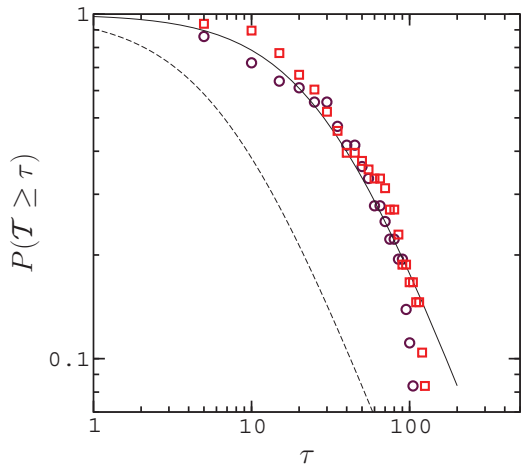

FIG. 2. (Color online) Distribution of waiting times, $P(\mathcal{T} \geqslant \tau)$, for lineage 1 (circles) and lineage 2 (squares) in Fig. 1. The solid line is a fit to the data for lineage 1 using Eq. (1). The dashed line is fit of Eq. (1) to the data reported by Bastolla using Eq. (2). The exponents of both fits are $\beta \simeq 1.2$.

Figure 2 plots the distribution, $P(\mathcal{T} \geqslant \tau)$, for lineage 1 (circles) and lineage 2 (squares) in Fig. 1. The solid line is a fit to the data for lineage 1 according to Eq. (1) in the region $\tau \leqslant 100$. The exponent of the fit is $\beta \simeq 1.2$. The dashed line describes the distribution $P(\mathcal{T} \geqslant \tau)$ obtained by fitting $P(\mathcal{T})$ to the data reported by Bastolla [13] which leads to essentially the same exponent. The exponents for individual lineages vary from about $\beta \simeq 1.0$ to about $\beta \simeq 1.7$. Each lineage includes between 6 and 12 mutations per monomer, exceeding the typical "lifetime" of a protein in Ref. [6]. For combinations of lineages (such as those in Fig. 1) we obtain $\beta \simeq 1.3$.

Figure 3 plots the average (circles) and the width (squares) of the distribution for $\Delta x^2(\tau, \tau + n)$ along lineage 3 computed according to method (i). The number of monomers participating in structural alignments is $N_{\|} = 16$, or about half the typical length of a polymer. The solid line is a fit to the data according to Eq. (4) with exponent $\alpha \simeq 1.0$, indicating normal diffusion. However, structural flights along lineage 3 are episodic, similar to fluctuations within folded ensembles, punctuated by longer flights corresponding to more collective changes in structure (Fig. 4). This result is somewhat typical, and for combinations of lineages, the distribution of flight lengths is Lévy-like, resembling a Gaussian with an extended tail [22].

Finally, Fig. 5 plots the average (circles) and the width (squares) of the distribution for $\Delta x^2(\tau, \tau + n)$ along lineage 1 computed by method (ii). The solid line is a fit to the data for $n \leqslant \mathcal{T}$ according to Eq. (4) yielding an exponent $\alpha \simeq 0.5$. The number of monomers compared in structural alignments is typically $N_{\|} \sim 20$. Here, structural flights are larger and more erratic than those in Fig. 4 since the pattern of ordered, nuclear monomers can change along a lineage. For

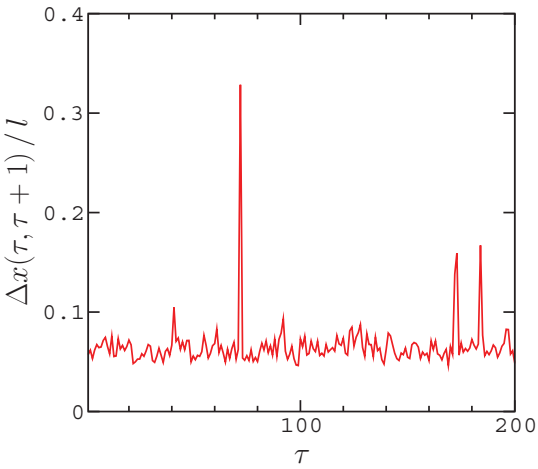

FIG. 4. (Color online) Length of structural flights, $\Delta x(\tau, \tau + 1)$, along lineage 3 computed by method (i). The longest flight in the figure is the result of a single amino acid replacement.

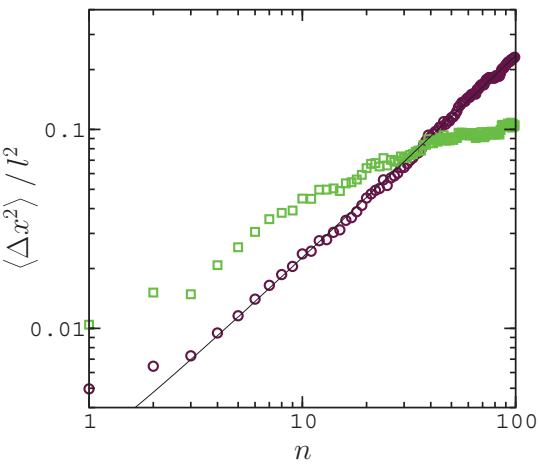

FIG. 3. (Color online) Average (circles) and width (squares) of the distribution of aligned distances, $\Delta x^2(\tau, \tau + n)$, along lineage 3 computed by method (i). The solid line is a fit to the data using Eq. (4). The exponent of the fit is $\alpha \simeq 1.0$.

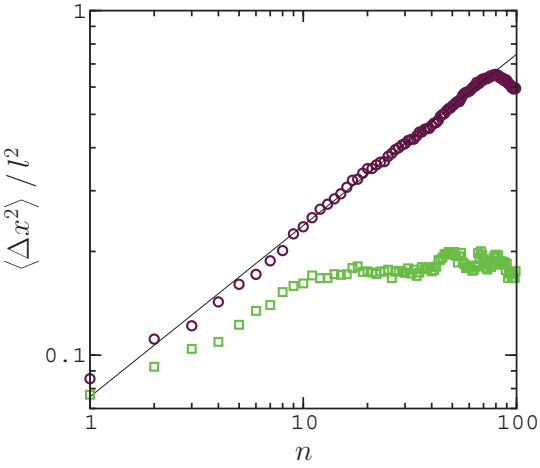

FIG. 5. (Color online) Average (circles) and width (squares) of the distribution of aligned distances, $\Delta x^2(\tau, \tau + n)$, along lineage 1 in Fig. 1 computed by method (ii). The solid line is a fit to the data for $n \leqslant 80$ according to Eq. (4). The exponent of the fit is $\alpha \simeq 0.5$.

this reason, we also examined a variant of method (ii) in which tracer positions are instead determined by the earlier sequence in an alignment, with conservation required over the next $n \leqslant \mathcal{T}$ neutral mutations. For this variant, we obtain similar exponents, $\alpha \simeq 0.5$, and similar agreement with Eq. (4).

To summarize, our results suggest that the evolution of small protein motifs is both temporally and structurally episodic. For method (i), in which compared monomers are selected to optimize MSD, structural change is, on average, super-diffusive. Here, individual lineages consist of many small structural changes, comparable to fluctuations within folded ensembles, punctuated by larger, more collective changes involving the entire nucleus. For method (ii), in which compared monomers are defined by the ordered nucleus of an ancestral sequence, structural change is sub-diffusive, with an exponent similar to small proteins. This result appears to be a general consequence of structural comparison through a fixed set of monomer positions. For example, if the set of compared monomers is defined locally (i.e. by the earlier sequence in an alignment), structural drift remains sub-diffusive.

A group of measures we have not yet considered are those based on shared contacts [23]. A potential advantage of this approach is that the elements compared to measure distance (i.e., contacts) are purely determined by the folded ensembles of the sequences, and can be measured (for example) using the Etters-Kaelberer parameter [24]. In addition, protein sequence alignments can be constructed using a reliable algorithm based on alignment of internal contact patterns [25], which provides for a more consistent comparison of the model with protein data. We intend to explore this problem in future work, including the genetic code, and more realistic functional constraints to model explicit binding of a target molecule.

The authors would like to thank Ugo Bastolla for helpful suggestions during the review of this work.

APPENDIX

The polymer model is a freely jointed chain of point monomers which interact according to spherically symmetric pair potentials. The potentials for unit strength attractive and repulsive interactions are plotted in Fig. 6. The strengths and forms of the potentials for a particular sequence are determined by empirical amino acid contact energies [26,27]. The polymers evolve kinetically by Langevin dynamics with parameters adjusted to obtain diffusive kinetics and room temperature folding transitions.

The viability of a sequence 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 replicas between random coil, ordered globule, and melting temperatures for a typical viable sequence. The structures recovered at the lower temperature are collected, along with their $\mathcal{N}$ (energetically equivalent) mirror images into an ensemble, Γ, and the ensemble is analyzed using the Lindemann melting criterion [28,29].

In general, the energy landscape of a polymer can contain many deep energy basins. As the replicas are cooled they can 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. In order to select for this situation, we search for a structure $\mathbf{x}^\star \in \Gamma$ to represent the native ensemble of the mutated polymer, and we require that a significant fraction of the replicas fold into structures that are close to $\mathbf{x}^\star$.

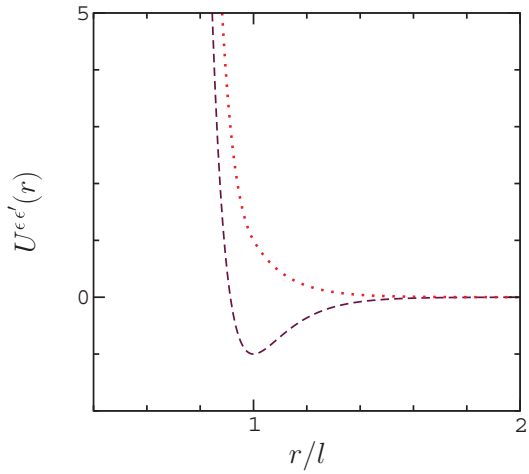

FIG. 6. (Color online) Potential functions, $U^{\epsilon\epsilon'}(r)$, for cross-chain interactions at unit core strength, $\epsilon = 1$, unit attraction, $\epsilon' = -1$ (dashed line) and unit repulsion, $\epsilon' = 1$ (dotted line). The polymers are linked by stiff quadratic potentials of equilibrium length l. In the model, the parameter ϵ' is determined by empirical amino acid contact energies [26,27].

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. Here, the Lindemann parameter measures the typical distance of a monomer in a structure $\mathbf{x}^\mu \in \Gamma$ aligned to $\mathbf{x}^\star$ from its corresponding monomer in $\mathbf{x}^\star$. Ideally, we would want to select the reference structure to minimize the average displacement of every monomer over the closest $\mathcal{N}$ replica structures in Γ. However, here it is necessary to allow for misfolding, and for weakly interacting (typically, uncharged hydrophilic) monomers which occupy the surfaces of the globules and are disordered in folded ensembles. For this reason, we use a reductive procedure to align the structures in which the most distant monomer pairings are removed, iteratively, until a nucleus of $2N/3$ optimally aligned pairs of monomers remain. 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. This process results in an ensemble $\Delta\Gamma^\star$ aligned to a structure $\mathbf{x}^\star$ yielding a minimal value of λ (i.e., in each image space). This alignment is then 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, $\lambda_j \lesssim 0.15\,l$ where l is the length of a polymer link. If the number of solid-ordered monomers, δN, exceeds a specified value, the sequence is accepted—otherwise, it is rejected.

The data presented in this work are obtained from 14 lineages of length ~ 3–5×10^2 mutations generated under the constraints $\delta N \geqslant 15$–16 and $\delta N \geqslant N/2$ in Ref. [1]. The attempt frequencies for random monomer replacements,

insertions, and deletions are adjusted to approximate protein data; For this set of lineages, amino acid replacements are accepted at a rate of about 0.05–0.10 per mutation attempt. Indels are accepted at about one-tenth this rate.

[1] E. D. Nelson and N. V. Grishin, Phys. Rev. E **90**, 062715 (2014).
[2] E. I. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994).
[3] D. M. McCandlish and A. Stoltzfus, Q. Rev. Biol. **89**, 225 (2014).
[4] D. C. Ramsey, M. P. Scherrer, T. Zhou, and C. O. Wilke, Genetics **188**, 479 (2011).
[5] C. Chothia and A. M. Lesk, EMBO J. **5**, 823 (1986).
[6] K. Illergard, D. H. Ardell, and A. Elofsson, Proteins **77**, 499 (2009).
[7] A. M. Gutin and A. Y. Badretdinov, J. Mol. Evol. **39**, 206 (1994).
[8] N. V. Grishin, J. Mol. Evol. **45**, 359 (1997).
[9] G. Tiana, N. V. Dokholyan, R. A. Broglia, and E. I. Shakhnovich, J. Chem. Phys. **121**, 2381 (2004).
[10] Tiana *et al.* have also investigated structural drift using a lattice model similar to that in Ref. [2].
[11] P. Schuster and P. F. Stadler, Complexity **8**, 34 (2003).
[12] J. H. Gillespie, Proc. Natl. Acad. Sci. U.S.A. **81**, 8009 (1984).
[13] U. Bastolla, H. E. Roman, and M. Vendruscolo, J. Theor. Biol. **200**, 49 (1999).
[14] F. J. Ayala, Proc. Natl. Acad. Sci. U.S.A. **94**, 7776 (1997).
[15] C. O. Wilke, BMC Genetics **5**, 25 (2004).
[16] M. Kimura, *The Neutral Theory of Evolution* (Cambridge University Press, New York, 1983).
[17] D. L. Theobald and D. S. Wuttke, Bioinformatics **21**, 2171 (2006).
[18] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.91.060701 for structural alignments along lineage 3 using method (ii).
[19] B. B. Mandelbrot, Complexity **14**, 55 (2008).
[20] A. E. Garcia, R. Blumenfeld, G. Hummer, and J. A. Krumhansl, Physica D **107**, 225 (1997).
[21] T. Neusius, I. Daidone, I. M. Sokolov, and J. C. Smith, Phys. Rev. Lett. **100**, 188103 (2008).
[22] B. B. Mandelbrot, Intl. Econ. Rev. **1**, 79 (1960).
[23] A. Pascual-Garcia *et al.*, Proteins **78**, 181 (2010).
[24] Y. Zhou, M. Karplus, K. D. Ball, and R. S. Berry, J. Chem. Phys. **116**, 2323 (2001).
[25] A. S. Konagurthu, J. C. Whisstock, P. J. Stuckey, and A. M. Lesk, Proteins **64**, 559 (2006).
[26] M. R. Betancourt and D. Thirumalai, Prot. Sci. **8**, 361 (1999).
[27] S. Miyazawa and R. L. Jernigan, J. Mol. Biol. **256**, 623 (1996).
[28] H. Jang, C. K. Hall, and Y. Zhou, Biophys. J. **82**, 646 (2002).
[29] R. A. la Violette and F. H. Stillinger, J. Chem. Phys. **83**, 4079 (1985).