Two modes of protein sequence evolution and their compositional dependencies

Ranjan V. Mannige*
Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, USA
(Received 14 March 2013; revised manuscript received 10 May 2013; published 24 June 2013)

Protein sequence evolution has resulted in a vast repertoire of molecular functionality crucial to life. Despite the central importance of sequence evolution to biology, our fundamental understanding of how sequence composition affects evolution is incomplete. This report describes the utilization of lattice model simulations of directed evolution, which indicate that, on average, peptide and protein evolvability is strongly dependent on initial sequence composition. The report also discusses two distinct regimes of sequence evolution by point mutation: (a) the "classical" mode where sequences "crawl" over free energy barriers towards acquiring a target fold, and (b) the "quantum" mode where sequences appear to "tunnel" through large energy barriers generally insurmountable by means of a crawl. Finally, the simulations indicate that oily and charged peptides are the most efficient substrates for evolution at the "classical" and "quantum" regimes, respectively, and that their respective response to temperature is commensurate with analogies made to barrier crossing in classical and quantum systems. On the whole, these results show that sequence composition can tune both the evolvability and the optimal mode of evolution of peptides and proteins.

DOI: 10.1103/PhysRevE.87.062714 PACS number(s): 87.14.et, 87.14.ef, 87.16.aj

I. INTRODUCTION

Important to both the history of life and the design of novel proteins is the unresolved question of how random peptide and arbitrary protein sequences are able to evolve into novel, structurally stable proteins [1]. While the field of protein *sequence* evolution is rich with understandings gleaned from half a century of research, little attention has been directed towards how sequence *composition* modulates evolvability. A recent study indicated that sequence composition may play an important role in the early beginnings of life [2], which further predicates a thorough exploration of the dependence of sequence evolvability on its composition.

This report explores how amino acid composition may tune the ability of a peptide (or protein) to evolve into a novel structure (and hence function). So far, it has been exceedingly difficult to provide such relationships using experiments and all-atom simulations, which is due to the enormity of both sequence space and the universe of structures available to any sequence. Because of these experimental and computational shortcomings, this report capitalizes upon the well-characterized cubic lattice model [3] rather than using all-atom models. Such lattice models, while losing complex molecular detail such as atom-resolution side-chain–side-chain interactions and secondary structure formations, have been shown previously to faithfully mimic protein behavior [3–8] while allowing for a more efficient sampling of both sequence and structure space. For example, variants of the model have already been used to explore folding cooperativity [9], criteria for designing and evolving stable folds [10], as well as crucial relationships between thermodynamic properties of the native state (e.g., "energy gap") and fast and cooperative folding kinetics [3,4], which is an important relationship for protein design algorithms today [11,12]. Relevant to this report, the cubic model was also used to show that, while the possibility that a random sequence may fold into stable structures is vanishing, some physiochemical factors may be sufficient to induce the emergence of stable folds [8,13].

Given the near-structural degeneracy of a random peptide's ground state, the question of how a random sequence may strengthen its lowest-energy (native) state is not of immediate relevance. However, of importance to *ab initio* protein fold invention [1] is the question of how a random peptide could "evolve" into an arbitrarily selected novel fold, assuming the fold's imminent importance in a biological or prebiological setting. The question asks how, given an arbitrarily selected fold (assumed to be of prebiological import), does composition modulate the transition to that fold from a random peptide (or even another protein)? Such a question involves directed evolution to a target fold, which is not a natural feature of evolution; however, such studies and simulations will allow us to use statistics in order to ask how evolutionarily accessible a fold is overall and how the starting sequence's composition may modulate that accessibility and evolvability.

These questions were addressed by performing millions of independent evolution simulations where target folds and starting sequence compositions are tuned. The simulations, described below, indicate that two types of protein folds exist—evolutionarily accessible or "good" folds and inaccessible or "bad" folds—each of which is dominantly accessible by distinct evolutionary mechanisms that are dependent on the fraction of oily (f_h; see Methods) and charged (f_c) residues within the starting peptide, respectively. To begin the results section, an example of a strong relationship evidenced between an initial peptide's oil content (f_h) and its propensity to evolve into an arbitrary "good" fold is discussed.

II. METHODS

A. Amino acid groupings

The amino acid groupings that will be referred to regularly are defined here. The set of amino acids defining oily (h), charged (c), and polar (p) amino acids are [FILV], [ERDK] and [STNQ], respectively (single letter codes are used). While

*rvmannige@lbl.gov

c is easy to define, as these are the only four residues that are charged in normal environments, the four most oily residues on the Kyte-Doolittle hydrophobicity scale [14] were chosen to populate h, partly to match the number of residues in c and partly because this metric was successfully used in another study [2]. Finally, the fractions of oily (h), charged (c), and polar (p) residues in any sequence are denoted as f_h, f_c, and f_p, respectively.

B. The model

To simulate peptide sequence-structure evolution, a previously well-characterized cubic lattice protein model was used, which, while lacking the nuances available to proteins such as atom-resolution interactions and secondary structure, is able to reiterate protein behavior and evolution [3–8]. The importance of this model lies in its relatively low computational load, due to the substantial but manageable number (N) of possible folds allowed within the structural ensemble, which allows this model to be used in millions of evolution simulations while varying target folds and sequence compositions (discussed below). This volume of studies is currently inaccessible by experiment and all-atom simulations.

In this model, 27-amino-acid peptides may fold into one of N collapsed cubic "folds," depending on their amino acid sequence. The sequence $\{\sigma_i\}$ is sourced from a 20-residue repertoire of naturally occurring amino acids, whose energy in configuration $\{r_i\}$ takes the form

$$E(\{r_i\},\{\sigma_i\}) = \frac{1}{2}\sum_{i,k}\mathbf{B}(\sigma_i,\sigma_k)\delta(r_i - r_k). \tag{1}$$

Here, $\mathbf{B}$ is a precompiled amino acid interaction potential matrix [15] and $\mathbf{B}(\sigma_i,\sigma_k)$ is the interaction energy between the types of the two amino acids σ_i and σ_k, assuming that they are in contact (denoted by δ). In the discussions, a collapsed structure $\{r_i\}$ is collectively referred to by its index j in the precompiled ensemble of possible collapsed structures (totaling N). Also, as a shorthand notation, E_j is the contact Hamiltonian of the sequence at hand in configuration j.

C. Simulating protein fold evolution

The evolution simulations utilized a well-characterized 27-residue cubic lattice model, whose amino acids are able to mutate to any of the 20 natural varieties [3]. Starting sequences are allowed to evolve via the Monte Carlo sampling of point mutations, with the probability of accepting or keeping a mutation equaling $\min(1, \exp[-\Delta E_{sep}/kT_{mc}])$. Here, ΔE_{sep} represents the change (upon mutation) in the *target energy separation*, i.e.,

$$E_{sep} = E_j - \langle E\rangle, \tag{2}$$

which is the sequence-dependent energy function that, when minimized, minimizes the target fold's energy E_j while maximizing the kinetic accessibility of j important to proteins (conversely, lower $|E_{sep}|$ indicates more shallow energy landscapes). Also, k is a constant and T_{mc} is the temperature of the simulation. Together, kT_{mc} scales the capacity for the mutating *sequence* to overcome energy barriers. As the energies concerned (E_{sep}) are physically sourced in kcal/mol, k was set to be the Boltzmann constant (1.9872041×10^{-3} kcal mol^{-1} K^{-1}) and $kT_{mc} = 0.59$ (i.e, $T_{mc} \approx 297$ K).

The evolution of a sequence is defined as complete when the Boltzmann probability P_j of folding into the target structure j exceeded 0.8. The Boltzmann probability of finding the sequence in fold j is

$$P_j = \frac{e^{-E_j/kT}}{\sum_i^N e^{-E_i/kT}}, \tag{3}$$

where k is the Boltzmann constant, T is the temperature of the system, E_j is the energy of the sequence in a conformation j, and N is the number of structures in the ensemble. In our simulations, kT is set to 0.59 (i.e., $T \approx 297$ K), and the number of structures in the ensemble (N) is kept at 10 000 (a fraction of the total number of collapsed structures) to maintain tractable computation speeds ($N = 10\,000$ is a small but still substantial fraction of the total number of the possible collapsed cubic forms [16]).

While $P_j > 0.8$ is a relatively lenient criterion for stability, it is a value that allows for computational tractability while ensuring the dominance of fold j in the evolved sequence's structural ensemble. Additionally, when structure j describes the sequence in its lowest energy (making j the ground or "native" structure), then P_j is also referred to as P_{Nat}. For reference, the free energy of obtaining structure j may be obtained from the Boltzmann probability P_j:

$$\Delta G_j = kT\ln\left[P_j^{-1} - 1\right]. \tag{4}$$

Finally, the number of mutations or steps (T_{P_j}) required to evolve a starting sequence to fold into an arbitrarily chosen fold j with high probability P_j was used as an inverse measure of sequence evolvability. As discussed above, the target Boltzmann probability is selected to be 0.8, and so our metric for "time" to complete an evolution run is $T_{0.8}$. While $T_{0.8}$ is expected to have strong dependence on its exact sequence, of interest to this report is the *average* behavior of $T_{0.8}$ ($\langle T_{0.8}\rangle$) with respect to the starting sequence's composition. Additionally, the average value of $\langle T_{0.8}\rangle$ over the entire range of sequence compositions ($\langle\langle T_{0.8}\rangle\rangle$) is a fold-specific metric (Fig. S6 [17]) and is useful in defining fold evolvability and fold "goodness" (Fig. S1). Particularly, as is seen below, good folds are defined as satisfying the criterion $\langle\langle T_{0.8}\rangle\rangle \leqslant 250$.

For any simulation trajectory, the number of intermediates (used in Fig. 3) is the total number of *distinct* ground states (folds) accessed by the evolving sequence before acquiring the target fold.

D. Composition as a measure of collapsedness

The average amino acid interaction energy B_0 is a useful metric for average "collapsedness" of a random sequence, since a lower (negative) value for B_0, due to its net favorable internal energy, indicates a higher tendency for a sequence to collapse on average [3].

For the superset of purely random sequences, given that each amino acid (and hence each pair of amino acids) in the interaction matrix will be equally encountered purely by chance, B_0 will be equal to the averaging of the elements of the interaction energy matrix $\mathbf{B}$, i.e., $B_0 = \langle\mathbf{B}(i,j)\rangle$. By constraining the fraction of $\mathbf{X}$ residues to f_x, one can calculate

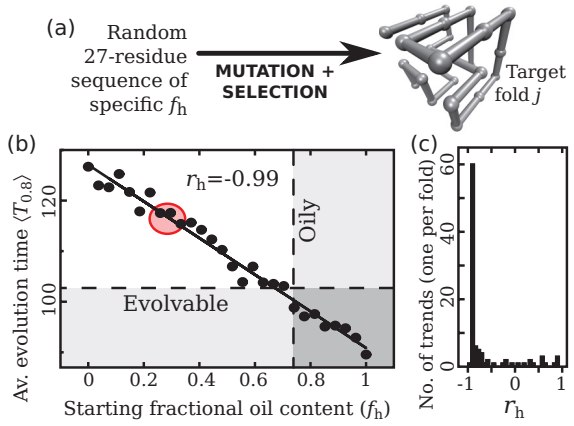

FIG. 1. (Color online) Oily peptides are highly evolvable. 27-residue sequences are evolved in steps of mutation and selection to fold into an arbitrary cubic "target" fold j (a) with high Boltzmann probability of 0.8 (see Methods). One "set" of simulations consists of peptides of varying oil content ($f_h = \{0/27, 1/27, \ldots, 27/27\}$) that are each evolved to an arbitrarily chosen target fold j. The results, shown explicitly for one set (b) and summarized for 100 distinct sets (c), indicate a dominantly negative correlation (ascertained by negative Spearman coefficient r_h) between starting f_h and average time (number of steps) taken to complete the evolution ($\langle T_{0.8} \rangle$). The lines describing "evolvable" and "oily" in panel b are placed to visually indicate that relatively more oily peptides are more evolvable. These simulations present a clear relationship between a sequence's starting oil content and its average evolvability. Each value (dark, filled circle) in panel (b) is obtained from 100 starting sequences evolved 50 times independently (simulations starting with pre-evolved proteins are averaged in the large, pink (light gray) shaded circle; see text). The residues (in single-letter code) considered to be oily in this study are [FILV]. However, using expanded definitions for oily (e.g., adding M, W, and C to [FILV]) does not change the trend found in panel (b) (Fig. S4).

the effective average amino acid interaction energy, $B_0(f_x)$, which must be a summation of probability-weighted averages of sections of the potential energy matrix $\mathbf{B}$; i.e.,

$$B_0(f_x) = f_x^2 \times \langle \mathbf{B}(i,j) \rangle_{i,j \in \mathbf{X}} + 2 f_x (1 - f_x) \times \langle \mathbf{B}(i,j) \rangle_{i \in \mathbf{X}, j \in \mathbf{X}'} + (1 - f_x)^2 \times \langle \mathbf{B}(i,j) \rangle_{i,j \in \mathbf{X}'}. \quad (5)$$

This relationship follows given that f_x and $(1 - f_x)$ are the probabilities of randomly picking amino acids from set $\mathbf{X}$ and its complement $\mathbf{X}'$.

III. RESULTS

A. Evolvability is composition dependent

The initial experiments consisted of evolving 27-residue lattice peptides of random sequence (sourced from a 20-residue alphabet) to fold well into a particular maximally collapsed (cubic) form. For every possible fractional oil content (f_h), evolution simulations were performed on 100 distinct random sequences (each repeated 50 times), from which the average "time" taken to complete—$\langle T_{0.8} \rangle$—was recorded (see Methods). The relationship [shown in the black-filled circles in Fig. 1(b)] between the starting oil content of a random peptide and its $\langle T_{0.8} \rangle$ is very strong (Spearman correlation coefficient $r_h \approx -0.99$), which indicates a strong positive dependence of a starting peptide's evolvability ($\propto \langle T_{0.8} \rangle^{-1}$) and its oil content f_h (also, the slope of the relationship steepens when employing more "realistic" structural ensembles in the simulation; see Fig. S2). Evolvability's strong dependence on oil content is dominantly reiterated in 100 independent data sets, with 81% and 63% of the trends having $r_h \leqslant 0$ and $r_h \leqslant -0.9$, respectively, indicating a near-universal and favorable dependence of evolvability on oil content.

Surprisingly, even the evolvability of pre-evolved lattice *proteins* that attain other folds with $P_{i \neq j} \geqslant 0.99$ [filled circle, red online, in Fig. 1(b)] appears to adhere to the trend evident in random peptides (see Fig. S3 for more independent examples). This universal adherence of evolvability to composition indicates a scenario where well-folding proteins may often be lower in fold evolvability than oily random peptides, which, as a trend, may prove to be useful in understanding the origination of the protein fold repertoire [1,2]. Current work on this matter is under way.

Composition of polar and charged residues also appears to affect evolvability to varying extents. In a manner identical to Figs. 1(b) and 1(c), for each randomly chosen target fold, sets of evolution experiments were performed with varying starting oil [FILV], charge [ERDK], and polar [STNQ] residue contents, which resulted in three Spearman correlation coefficients r_h, r_c, and r_p per fold, each relating the dependence of evolution time $\langle T_{0.8} \rangle$ and starting fractional content of oils (f_h), charges (f_c), and polar residues (f_p), respectively.

Figure 2(a) describes the distribution of r's, which indicates that highly polar content actually undermines fold evolution (given that all r_p's are greater than zero). While such behavior indicates a universal evolutionary "drag" for polar peptides (given the universally slow relative speed of evolution), such a term might be misleading, since in fact polar peptides sample the largest and the most diverse number of intermediates during their evolution [high f_p peptides sample 55.4% $\pm$ 0.5 standard deviation (s.d.) and 60.6% $\pm$1.4 s.d. more intermediates compared to high-f_c and high-f_h peptides, respectively; Table S1); it appears that while polar peptides are able to sample a large portion of the free energy landscape, they do not easily "fall" into the desired structural minimum, potentially due to the lack of enough self-energy in any particular configuration. This behavior may be compared analogously to a "low-mass golf ball" in a golf course, which potentially gets near the target hole (topology), but never easily falls in due to lack of enough "weight" (self-energy). The next focus will be on substrates that hasten the evolution to a target fold.

B. Distinct evolutionary regimes

At first glance, both high oil content and high charge content of a starting sequence appear to be important to the evolution of all novel folds, with both r_h and r_c displaying dominantly negative values [Fig. 2(a)]. However, charge content is less significant in "practical" evolvability, as its importance lies mostly in hastening the evolution of folds that are evolutionarily difficult to approach in the first place [Fig. 2(d)], while oily peptides hasten the evolution of more approachable and evolvable folds [Fig. 2(c)] marked by

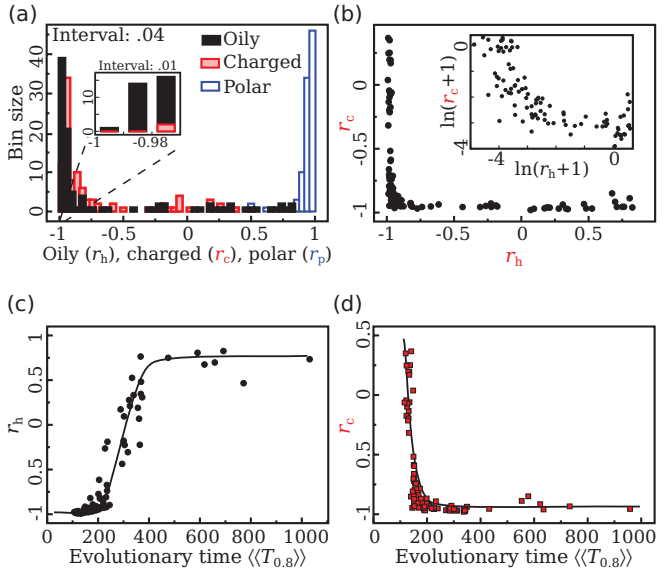

FIG. 2. (Color online) Large-scale studies of the effect of initial composition vs evolvability. For each amino acid class $x \in$ {h[oily],c[charged],p[polar]}, correlation coefficients were obtained [r_x's; similar to the r_h in Fig. 1(b)] for a hundred target folds. The distribution of r_x's (a) indicates that evolution is hastened with the increase in oils and charges (i.e., for most folds, r_h and r_c tend towards -1), while polar groups are actually counterproductive to the evolution process (with $r_p \geqslant 0$; also see Fig. S5). Remarkably, oils and charges modulate evolvability at different fold regimes, given the anticorrelation between fold-specific r_h's and r_c's (b; Spearman $r = -0.82$, p-val $= 1.42 \times 10^{-25}$), with oily peptides being most useful in "good" designs that are evolutionary most approachable ($\langle\langle T_{0.8}\rangle\rangle \leqslant 250$; (c), while evolution of "difficult folds" appears to depend mostly on high charge content (d).

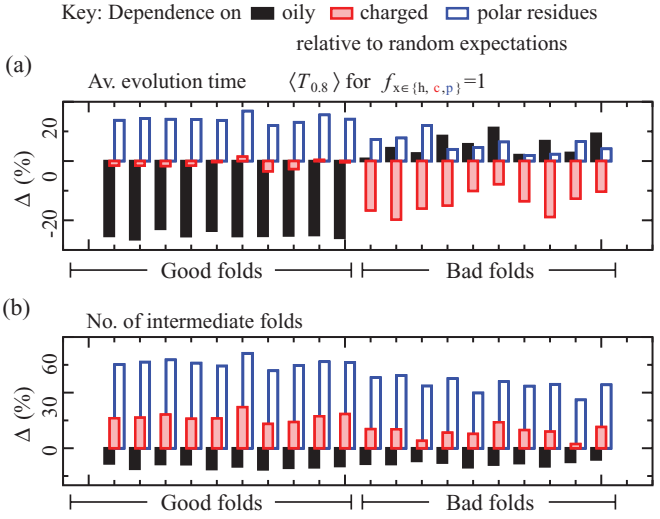

FIG. 3. (Color online) Oily peptides access the most direct paths and intermediates to the target fold. Listed in the abscissa of each panel are the 10 best and worst folds, defined by the average time ⟨⟨$T_{0.8}$⟩⟩ taken for a random peptide (i.e., ⟨$T_{0.8}$⟩'s averaged over all $f_{x\in\{h,c,p\}}$'s) to evolve into a target fold. The ordinate shows three evolutionary propensities per fold for peptides of maximal oils (black), charges (red online), and polar residues (blue online) when compared to unbiased random peptides (Δ; positive and negative values indicate values that are higher and lower than random results, respectively). As expected from Fig. 2, oils are most important for evolution of good folds (a), while charges are most important for the evolution of bad ones. Surprisingly, the number of structural intermediates encountered by oily peptides while approaching a target fold (b) indicates that while oily peptides provide the most "direct route" (lowest intermediates) to the target, the transition barriers to reach the target are too high for bad folds.

⟨⟨$T_{0.8}$⟩⟩ $\leqslant$ 250 (Fig. S1). Indeed, the capacities to evolve from oily versus charged peptides into any particular target fold are anticorrelated [Fig. 2(b); Spearman $r = -0.82$, p-val $= 1.4 \times 10^{-25}$]. These results indicate two distinct regimes of evolvability, each affected mostly by either oils or charges, respectively. For more on the criterion for distinguishing good from bad folds, please refer to Fig. S1.

C. Exploring the distinct evolutionary mechanisms

The distinct regimes of evolution espoused by oily and charged peptides [Figs. 2(b)–2(d)] indicate distinct molecular mechanisms of evolution. By studying the effect of peptide evolution in the 10 best and worst folds, an attempt is made to dissect these molecular mechanisms. Figure 3 presents the evolutionary properties of the starting sequences (of maximum $f_{x\in\{h,c,p\}}$) for 10 of the best and worst target folds indicated by the lowest and highest ⟨⟨$T_{0.8}$⟩⟩'s, respectively. As expected from Fig. 2, oily starting peptides are the best substrates for good folds, while charged peptides are the best substrates for bad ones [Fig. 3(a)]. Interestingly, the number of intermediate folds required by oily peptides to reach the target fold [Fig. 3(b)] is consistently lower than average for both good *and* bad target topologies (raw data in Table S1), which indicates that, in all scenarios, oily peptides sample the most expeditious paths—shortest paths with intermediates that are structurally most similar to each other (Fig. S7)—in protein structural space (characterized by the "protein domain universe graph" or PDUG [18]), while the relatively much higher number of intermediate topologies sampled by highly charged peptides indicate a random-jump move in intermediate structure space until the target topology is obtained as its native state.

D. "Crawling" versus "tunneling" (Fig. 4)

This section extends the previous discussions relating the dynamics of evolutionary barrier crossing to the quantum and classical mechanics of energy barrier crossing [19–21]. Two pictures emerge when considering the evolution of a sequence from structure A to structure B. First, in the classical picture, the sequence overcomes an energy barrier (associated with breaking interactions that stabilize A) by hopping or crawling through often dynamic [22] intermediates ($AB^{\ddagger}$) until B is exclusively obtained. Alternatively, in the "quantum mechanical" picture, the sequence could jump or "tunnel" more drastically and blindly over large swaths of structural space, thereby eventually finding structure B. While the classical method is the canonical one, tunneling has also been discussed in reference to genetic recombination [20] and is related to surmounting a large sequence-sampling-dependent *entropic* barrier to a final extremum fitness and functionality [23]. This section describes how evolution by oily and charged

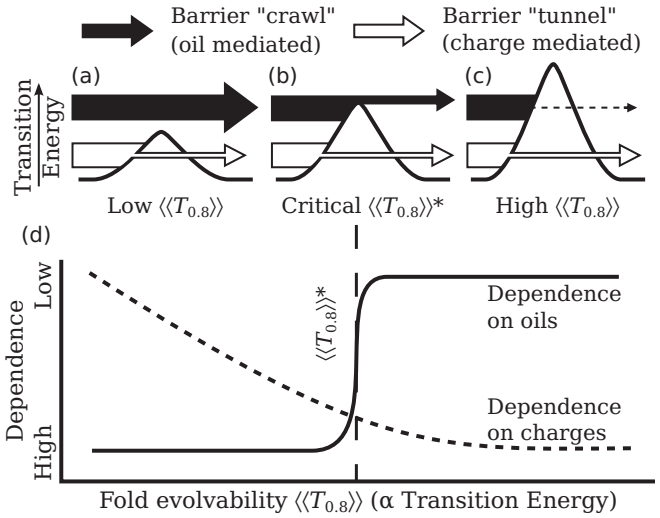

FIG. 4. An interpretation of Figs. 2(c) and 2(d). Panels (a)–(c) describe three fold-specific evolutionary transition barriers associated with three evolvabilities or $\langle\langle T_{0.8}\rangle\rangle$'s (low, critical, high). The thick, horizontal arrows indicate the evolutionary "transition rate" (black-filled) and "transmission rate" (white-filled) that indicate distinct (oil-dependent) crawling and (charge-dependent) tunneling mechanisms, respectively. Given the probabilistic nature of tunneling, the increase in the average required number of steps $\langle\langle T_{0.8}\rangle\rangle$ will result in a monotonic and steady increase in the evolutionary contribution of substrates employing such a mechanism [(d), dashed line; ordinate axis is flipped to resemble that in Fig. 2(c)]. Conversely, the dependence on substrates that utilize the "crawl" mechanism will diminish swiftly and sigmoidally as the critical transition barrier (indicated by a critical $\langle\langle T_{0.8}\rangle\rangle^*$) is crossed [(d), solid line]. The two scenarios appear to be applicable to the evolution of charged [Fig. 2(c)] and oily peptides [Fig. 2(d)], respectively.

sequences may be related to classical "crawls" and quantum mechanical "tunnels" through energy barriers, respectively.

The comparison of the two modes of evolution (from two classes of starting peptides) to crawling over and tunneling through a cumulative evolutionary barrier is potentially useful in explaining the following behaviors at the two regimes. For oily starting peptides, a classical crawl over evolutionary barriers is indicated by (i) a sigmoidal descent in the utility of oily peptides beyond a particular threshold evolutionary barrier "height" [i.e., at a critical threshold $\langle\langle T_{0.8}\rangle\rangle^*$, r_h abruptly departs the negative regime; see Figs. 2(c) and 4(d), solid line], and (ii) the increase of the critical threshold value $\langle\langle T_{0.8}\rangle\rangle^*$ with increasing evolutionary temperatures (simulations run at $kT_{mc} = 0.59$ and 0.79 are presented and compared in Figs. S8 and S9). Both these properties are expected for mechanisms dependent upon the classical crawling (or "hopping") over evolutionary barriers. Conversely, for charged peptides, tunneling through evolutionary barriers is indicated by (i) an *increase* in the utility of charged peptides with increasing energy barriers [$r_c \to -1$ as $\langle\langle T_{0.8}\rangle\rangle \to \infty$; Figs. 2(c) and 4(d), dashed line], which occurs in a more gradual (nonsigmoidal) fashion, and (ii) the decrease in the utility of this putative tunneling method with the increase of evolutionary temperature (see Figs. S8 and S9 for simulations at two kT_{mc}'s). In these ways, evolution from oily and charged peptides appears to

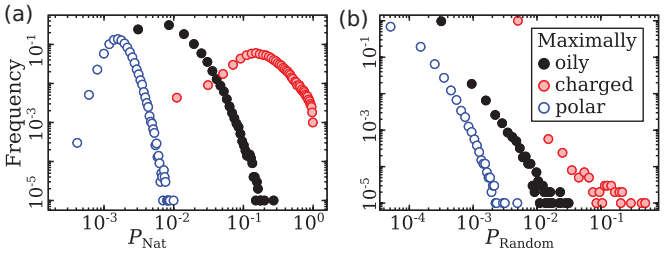

FIG. 5. (Color online) Charged peptides are optimal at evolutionary "tunneling." 100 000 random sequences of maximal f_h [black; see key in panel (b)], f_c (red online), and f_p (blue online) were produced, whose histogram distributions of Boltzmann probabilities of attaining the ground state (P_{Nat}) and a randomly chosen state from the ensemble (P_{Random}) are shown in panels (a) and (b), respectively. It is evident that although the chances are low (note the log-log scale), random charged peptides display greater stabilities for both ground state (native) folds *and* random folds, indicating a rugged folding landscape for charged peptides with the greatest potential for describing thermodynamically stable non-native structures.

suitably reflect the behavior of evolution via the surmounting of an evolutionary barrier by (classical) crawls and (quantum mechanical) tunnels, respectively.

E. Charged peptides' tunneling potential

The utility of charged peptides in tunneling is potentially due to their pronounced ability to encounter relatively high Boltzmann probabilities of assuming both their ground states [$P_{j=Native}$ or P_{Nat}; Fig. 5(a)] and other random folds [P_{Random}; Fig. 5(b)]. Given that such probabilities are very low but nonvanishing in value, their effect is only expected to be visible or prominent when a relatively high number of evolutionary steps are required and taken ($\langle\langle T_{0.8}\rangle\rangle > 250$).

F. Oily peptides' crawling potential

This section discusses the properties of oily peptides that may enable them to efficiently "crawl" over evolutionary barriers.

1. *Structural plasticity*

The ability to select the most expedient path (as close to the allowed "shortest path") among all possible evolutionary routes requires that the most expedient *first* intermediate must be "accessible" to the sequence. This is similar to the concept of a dynamic intermediate discussed in context of protein *function* evolution [22,24,25]. Figure 6(a), which is an average energy diagram of the 100 lowest-energy structures in the ensemble, shows that high f_h peptides do, as expected, have numerous transient structures thermodynamically accessible at reasonable temperatures; i.e., the sequence is structurally plastic. A related metric, P_{Nat}—which in low values indicates the structural degeneracy ("plasticity") of a sequence—indicates that high-f_h peptides (along with high-f_p peptides) have a number of folds that are energetically close to the ground-state value [Fig. 6(b)]. This indicates that oily peptides have a large number of structural intermediates to choose from during their evolutionary journeys.

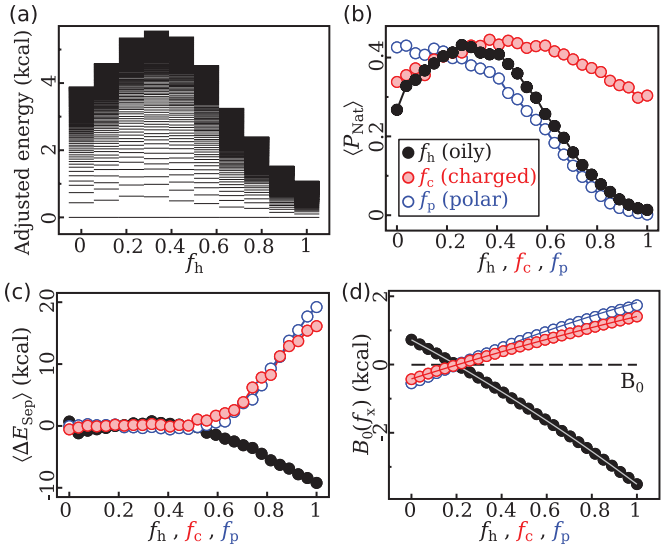

FIG. 6. (Color online) Oily peptides are structurally degenerate [panels (a) and (b)], have high mutational allowance (c), and are dominantly collapsed (d). Panel (a) represents the 100 lowest-energy levels (skewed so $E_0 = 0$) averaged for 100 random sequences for $f_h = 0 \ldots 1$. This result is reiterated when studying the average Boltzmann probability of occupying the "native" state P_{Nat} [panel (b); lower P_{Nat}'s indicate higher structural degeneracy]. Panel (c) shows that oily peptides uniquely have high mutational allowance, which is inversely related to ΔE_{sep} [panel (c); Eq. (2)]. Finally, panel (d) shows that only oily peptides dominantly collapse in proteinlike environments [negative average self-energy, B_0—obtained from Eq. (5), solid lines, and enumeration of 1000 peptides per $f_{x\in\{h,c,p\}}$, solid circles—indicates average collapsedness; changing the source of the pairwise potential matrix B does not change this result; Fig. S10].

2. *Mutational allowance*

Second, to efficiently navigate evolutionary barriers, a sequence must have the capacity to be regularly mutated with nearly neutral or advantageous outcomes, a quality that may be called high "mutational allowance." To assay a sequence's mutational allowance, an inversely related metric—the change, upon mutation, in the energy separation of the sequence [ΔE_{sep}; Eq. (2)]—was utilized, which indicates that oily peptides are the only class of peptides that display high mutational allowance [Fig. 6(c)].

3. *Collapsibility*

Finally, peptides that are dominantly collapsed would have a higher chance of transiently describing folds within their accessible structural ensemble. It has already been shown [3] that the average interaction energy B_0, obtainable by integrating the elements of the amino acid interaction matrix **B**, may be used as an indication of the average polymer's tendency to collapse (with lower or negative values of B_0 indicating greater propensity to collapse). Equation (5) describes B_0 as a function of the fraction of **X** residues (f_x) in the peptide, which shows how collapsibility may be tuned, on average, by varying the fraction of oily (f_h), charged (f_c), and polar (f_p) residues. Both theory [Fig. 6(d); solid lines] and sequence enumeration (circles) indicate that oily random peptides are the only class of peptides that are expected to be dominantly collapsed (Fig. S10 shows that this result is invariant of the choice of pairwise residue potentials [15,26]); while some charged sequences *may* collapse, most are not expected to do so in protein environments.

Oily peptides are the only class studied that possess all three properties important in efficient barrier crawling—structural plasticity, mutational amenability, and collapsibility—which makes them effective in the "accessible" fold regime. Charged peptides, lacking these abilities, depend on their high $\langle P_{Random} \rangle$ [or "tunneling potential"; Fig. 5(b)] to eventually tunnel into the vicinity of the target free energy minimum, upon which downhill adaptation would proceed; the requirement of high sampling for tunneling ensures that this method is only efficient in the bad fold regime. Finally, polar sequences, on account of displaying both the inability to crawl and low $\langle P_{Random} \rangle$, are left with the distinction of being the worst substrate in both good and bad fold regimes.

G. Is fold invention by tunneling really possible?

It is important to note that the increased evolutionary "tunneling potential" observed here for charged peptides is possibly exaggerated, particularly due to the following three points: (i) the frustrated nature of a *collapsed* charged polymer, (ii) the relatively low probability that a charged polymer will collapse in the first place [Fig. 6(d)], and (iii) the limited size of the available structural ensemble ($N = 10\,000$). The true practicality of utilizing charges for "tunneling into" evolutionarily inaccessible folds remains to be realized. Oily peptides, which do not depend on a peptide's vanishing chances of naïvely obtaining a low Boltzmann probability of folding, may not be affected by these caveats.

H. Alternate universes affect fold accessibility

Although deciphering what makes a "good" fold (a fold with low $\langle\langle T_{0.8} \rangle\rangle$) is not the focus of this report, it is interesting to note that while properties inherent to a protein fold and topology are important in determining its designability and evolvability [27–32], the relationship of a particular fold in context of the universe of possible folds (indicated by its position in the PDUG [18]) is also expected to be important (as alluded to previously [33]). This was ascertained by repopulating a new and randomized 9900-structure ensemble (from the superset of 103 346 possible collapsed forms [16]; see Methods) with the 100 target folds studied in Fig. 2 and reperforming all evolution experiments. Interestingly, while the global trends in Fig. 2 are reiterated [Figs. S11(a)–S11(c)], the evolvability of individual folds from the two protein universes do not correspond well [Spearman r between corresponding r_x's ≈ 0.4; Figs. S11(d) and S12; Table S2]. This indicates that while the distribution of good and bad folds are the same in both protein fold universes (ensembles), the "social standing" or goodness of each individual fold is contingent upon its position in each fold universe (PDUG).

I. Other effects and modes of evolution

It should be noted that the studies discussed here are on isolated peptides (at infinite dilution), and the addition of other effects such as molar concentration dependence (crowding) and change in salt conditions were not taken into consideration. For example, higher-order relationships, such as the propensity of oily peptides to aggregate, should be considered. While aggregation is potentially detrimental to biological systems [34], results from folding studies in confined or hydrophobic environments are ambiguous, with detrimental [35,36], stabilizing [37–40], and topology-specific [41] examples. Extensive work to overlay such aspects onto single-molecule studies remains to be performed.

Also, while oily peptides are more evolvable on average, "successful" sequences must possess *some* polar and charged residues that would partake in (potentially enzymatic) activity and allow for their recruitment into a biological system. The "sweet spot" of oil content that allows for optimal recruitment and evolvability is unknown.

This report focused primarily on point mutations as a mode of evolution, which is partially due to the historical focus on this mode of evolution and partially due to the immutability of the lattice model protein length. However, numerous other modes of sequence evolution exist (such as recombination [42–45]), which may modify the effective extent of sequence sampling during protein and peptide evolution. The effects of such moves on the regimes of evolution studied here are left to future studies.

J. Further discussions

In this report, using composition-comprehensive lattice model simulations, meaningful relationships between a peptide's average evolvability (into an arbitrary fold) and its amino acid composition (with respect to oils, charges, and polar groups) are established. Particularly, random oily peptides—those peptides with high fractional [FILV] content or f_h—were established as the most effective at evolving into most random target folds, and charged peptides (peptides with high fractional [ERDK] content or f_c) were found to be most effective at evolving into random target folds that are more evolutionary inaccessible (by means of a distinct "tunneling mechanism"). The mechanisms for their respective superior evolvabilities were shown to be distinct (Figs. 2 and 3), which shows how distinct regimes of protein structure and function may be approached using distinct mechanisms and substrates.

Finally, the results in this paper were obtained from more than 80 million independent evolutionary simulations, which reiterate the utility of lattice models in such exploratory endeavors, as experiments or all-atom simulations are as yet incapable of approaching such volumes. However, with the knowledge of such comprehensive studies, further experimental and atomistic investigations into the utility of amino acid composition as an accelerant for evolution would be useful from (i) the practical perspective of how one would better design a novel protein and (ii) the more philosophical perspective of how the first complete protein repertoire may have originated from random peptides.

ACKNOWLEDGMENTS

I thank Alana Canfield, Ron Hills, and the two reviewers for providing input to the manuscript. The lattice simulations were performed on the Odyssey computing cluster hosted by Harvard University's Faculty of Arts and Sciences, and access was provided by Eugene Shakhnovich. This work was funded in part by the Mannige family and the NIH grant GM068670 to Professor Eugene Shakhnovich (www.nih.gov).

[1] C. Chothia, J. Gough, C. Vogel, and S. A. Teichmann, Science **300**, 1701 (2003).
[2] R. V. Mannige, C. L. Brooks, and E. I. Shakhnovich, PLoS Comput. Biol. **8**, e1002839 (2012).
[3] E. Shakhnovich, Chem. Rev. **106**, 1559 (2006).
[4] A. Šali, E. Shakhnovich, and M. Karplus, J. Mol. Biol. **235**, 1614 (1994).
[5] G. Tiana, R. A. Broglia, and E. I. Shakhnovich, Proteins **39**, 244 (2000).
[6] E. J. Deeds, N. V. Dokholyan, and E. I. Shakhnovich, Biophys. J. **85**, 2962 (2003).
[7] K. B. Zeldovich, I. N. Berezovsky, and E. I. Shakhnovich, J. Mol. Biol. **357**, 1335 (2006).
[8] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, Pac. Symp. Biocomput. 27 (1997), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.221.8673.
[9] E. I. Shakhnovich and A. M. Gutin, Proc. Natl. Acad. Sci. USA **90**, 7195 (1993).
[10] E. I. Shakhnovich and A. M. Gutin, Nature (London) **346**, 773 (1990).
[11] K. W. Plaxco, D. S. Riddle, V. Grantcharova, and D. Baker, Curr. Opin. Struct. Biol. **8**, 80 (1998).
[12] O. Alvizo, B. D. Allen, and S. L. Mayo, Biotechniques **42**, 31 (2007).
[13] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. USA **93**, 839 (1996).
[14] J. Kyte and R. F. Doolittle, J. Mol. Biol. **157**, 105 (1982).
[15] S. Miyazawa and R. L. Jernigan, J. Mol. Biol. **256**, 623 (1996).
[16] E. Shakhnovich and A. Gutin, J. Chem. Phys. **93**, 5967 (1990).
[17] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.87.062714 for additional figures and discussions.
[18] E. J. Deeds, B. Shakhnovich, and E. I. Shakhnovich, J. Mol. Biol. **336**, 695 (2004).
[19] W. Ebeling, A. Engel, B. Esser, and R. Feistel, J. Stat. Phys. **37**, 369 (1984).
[20] Y. Cui, W. H. Wong, E. Bornberg-Bauer, and H. S. Chan, Proc. Natl. Acad. Sci. USA **99**, 809 (2002).
[21] K. Jain and J. Krug, J. Stat. Mech. (2005) P04008.
[22] N. Tokuriki and D. S. Tawfik, Science **324**, 203 (2009).
[23] E. van Nimwegen and J. P. Crutchfield, Bull. Math. Biol. **62**, 799 (2000).
[24] L. C. James and D. S. Tawfik, Trends Biochem Sci **28**, 361 (2003).
[25] I. Yadid, N. Kirshenbaum, M. Sharon, O. Dym, and D. S. Tawfik, Proc. Natl. Acad. Sci. USA **107**, 7287 (2010).

[26] R. I. Dima, G. Settanni, C. Micheletti, J. R. Banavar, and A. Maritan, J. Chem. Phys. **112**, 9151 (2000).
[27] H. Li, R. Helling, C. Tang, and N. Wingreen, Science **273**, 666 (1996).
[28] S. Govindarajan and R. A. Goldstein, Proc. Natl. Acad. Sci. USA **93**, 3341 (1996).
[29] E. I. Shakhnovich, Fold Des. **3**, R45 (1998).
[30] J. L. England and E. I. Shakhnovich, Phys. Rev. Lett. **90**, 218101 (2003).
[31] L. Meyerguz, C. Grasso, J. Kleinberg, and R. Elber, Structure **12**, 547 (2004).
[32] B. E. Shakhnovich, E. Deeds, C. Delisi, and E. Shakhnovich, Genome Res. **15**, 385 (2005).
[33] N. V. Dokholyan, B. Shakhnovich, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. USA **99**, 14132 (2002).
[34] F. Chiti and C. M. Dobson, Annu. Rev. Biochem. **75**, 333 (2006).
[35] A. Dhar, A. Samiotakis, S. Ebbinghaus, L. Nienhaus, D. Homouz, M. Gruebele, and M. S. Cheung, Proc. Natl. Acad. Sci. USA **107**, 17586 (2010).
[36] A. C. Miklos, M. Sarkar, Y. Wang, and G. J. Pielak, J. Am. Chem. Soc. **133**, 7116 (2011).
[37] D. K. Eggers and J. S. Valentine, J. Mol. Biol. **314**, 911 (2001).
[38] H.-X. Zhou and K. A. Dill, Biochemistry **40**, 11289 (2001).
[39] F. Takagi, N. Koga, and S. Takada, Proc. Natl. Acad. Sci. USA **100**, 11367 (2003).
[40] D. Thirumalai, D. K. Klimov, and G. H. Lorimer, Proc. Natl. Acad. Sci. USA **100**, 11195 (2003).
[41] L. Javidpour and M. Sahimi, J. Chem. Phys. **135**, 125101 (2011).
[42] M. H. Schierup and J. Hein, Genetics **156**, 879 (2000).
[43] C. M. Thomas and K. M. Nielsen, Nat. Rev. Microbiol. **3**, 711 (2005).
[44] M. Syvanen, Annu. Rev. Genet. **28**, 237 (1994).
[45] W. F. Doolittle, Science **284**, 2124 (1999).