Chapter 17

Statistical Physics of Evolutionary Trajectories on Fitness Landscapes

Michael Manhart and Alexandre V. Morozov*

Department of Physics and Astronomy & BioMaPS Institute for Quantitative Biology, Rutgers University, Piscataway, NJ 08854, USA

Random walks on multidimensional landscapes are important to many areas of science and engineering. In particular, properties of adaptive first-passage trajectories on fitness landscapes determine population fates and thus play a central role in evolutionary biology. The topography of fitness landscapes and its effect on evolutionary dynamics have been extensively studied in the literature. We will survey the current knowledge in this field, focusing on a recently developed systematic approach to characterizing path lengths, mean times, and other statistics of the first-passage path ensemble. This approach, based on general techniques from statistical physics, is applicable to landscapes of arbitrary complexity and structure. It is especially well-suited to quantifying the diversity of stochastic trajectories and repeatability of evolutionary events. We demonstrate this methodology using a biophysical model of protein evolution that describes how proteins maintain folding stability while evolving new functions.

1. Introduction

Random walks on networks are ubiquitous in nature. As an example, consider proteins, macromolecules that carry out a myriad of chemical and mechanical functions inside cells [1]. Each protein is a chain of amino acids chemically bonded to make a linear polypeptide [2], and the sequence of amino acids determines the protein fold — a compact 3D conformation which has the minimum free energy. Unlike random heteropolymers, naturally-occuring proteins have unique folds that they achieve robustly and, in many cases, rapidly (on the time scales of micro- or milliseconds),

*Corresponding author: morozov@physics.rutgers.edu

starting from arbitrary unfolded conformations [3].

Proteins are produced in the unfolded state inside the cell and have to fold before they can function. Protein conformations are often represented by sets of dihedral angles (torsion angles defined by three chemical bonds connecting four atoms [3]). Although in general the values of dihedral angles are continuous, they are typically discretized in protein structure prediction algorithms. In this case, protein folding can be viewed as a random walk on a network of conformation states with connectivity defined by the move set — a set of instructions for changing the dihedral angles in each step.

The network is very high-dimensional. For example, for a relatively small protein with $L = 100$ amino acids, 2 dihedral angles per amino acid, and 10° dihedral angle increments, there are 36^{200} possible conformations. With a simple move set that updates one angle at a time, each node is connected to $200 \times (36 - 1) = 7000$ neighbors. In such a large space, how can a protein reach its unique folded shape on reasonable time scales? This problem is known as the Levinthal paradox [4], and key to its resolution is the idea of the protein folding landscape [5, 6]. Each protein conformation has a free energy which is a function of the 200 dihedral angles, forming a free energy landscape over the network. The free energy values at a node and its neighbors determine rates of transition between nodes, e.g., according to the Metropolis algorithm [7]. This landscape is believed to have a global funnel shape, which allows the protein to find its folded structure efficiently through incremental moves, without searching the entire space [6].

This picture generalizes to many other search problems on networks in which each node, corresponding to a discrete (or discretized) state of the system, can be assigned a value of the objective function which sets the transition rates. As with protein folding, a major question is how the landscape topography and the move set determine the dynamics, especially first-passage processes. For example, an important quantity of interest is the mean first-passage time (e.g., to the global minimum on the protein folding landscape), which should be minimal in optimized algorithms [8].

The effect of landscape topography on dynamics is of particular importance in evolutionary theory, the study of how populations of organisms change over time through mutation and natural selection [9]. The genotype (genetic state) of an organism is represented by a sequence σ of letters drawn from an alphabet of size k. The sequence may represent nucleotides in genomic DNA ($\{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$, $k = 4$), amino acids in a protein ($k = 20$), or the binary presence/absence of a mutation at several genes across the genome ($k = 2$). Assuming a fixed number L of sites in each

sequence, the space of all k^L possible sequences represents a network with sequence nodes connected to each other if they differ by a mutation at a single site [10]. For simplicity we neglect recombination between sequences and insertion/deletion of sites which would redefine network connectivity and in some cases the total number of nodes.

Each sequence σ can be assigned a fitness value $\mathcal{F}(\sigma)$ that characterizes the reproductive success of an organism with that sequence. The exact definition of fitness can vary widely across different contexts, often depending on the construction of a model or what is observable in an experiment [11, 12]. Here we use a general theoretical definition of fitness as the relative probability an individual with that sequence will survive to reproduce [13]. This probabilistic definition of fitness is sometimes known as *multiplicative fitness*. In some circumstances it is more useful to consider $\log \mathcal{F}$, or *additive fitness*, which is related to growth rate.

Either case defines a fitness landscape or, more precisely, a genotypic fitness landscape [14]. Just as the folding landscape's structure is key to a protein's ability to reach its folded state efficiently, the fitness landscape is key to understanding how complex biological structures, such as bacterial flagella or the human eye, can arise through random, incremental mutations [10, 15]. Evolutionary adaptation, therefore, is represented by first-passage trajectories leading to local or global maxima on the fitness landscape. Characterizing the statistical properties of these first-passage trajectories is a major goal for evolutionary theory.

1.1. *Evolutionary dynamics*

In general, individuals in a population will have different sequences, occupying a distribution of points on the fitness landscape. However, in the limit $u \ll (LN \log N)^{-1}$ [16, 17], where u is the mutation rate (defined as the probability of mutation per site per generation) and N is an effective population size [13, 18], new mutations arise individually and either fix in the population or disappear from it on time scales that are short compared with the times between successive mutations [19–21]. Thus the population is monomorphic and, apart from short transition periods, occupies a single point in sequence space. The stochastic process of a new mutant appearing and fixing in the population is known as substitution, and the substitution rate from sequence σ to σ' is given by [19]

$$\langle \sigma' | \mathbf{W} | \sigma \rangle = N u \phi(s), \tag{1}$$

where Nu is the total number of new mutations in the population per site per generation, and $\phi(s)$ is the probability of a single σ' mutant fixing in a population of σ when the selection coefficient is $s = \mathcal{F}(\sigma')/\mathcal{F}(\sigma) - 1$ ($s > 0$ for beneficial mutations, $s < 0$ for deleterious ones).

The exact form of the fixation probability $\phi(s)$ depends on the underlying population dynamics. However, there are some common approximations valid in different asymptotic regimes [22]. For example, when $N \gg 1$ and $|s| \gg 1$, all beneficial mutations are essentially guaranteed to fix, while deleterious ones are guaranteed to be eliminated. Similar to zero-temperature Monte Carlo, the population can only undergo substitutions that increase fitness, and all allowed substitutions occur with the same rate Nu (since for $|s| \gg 1$, $\phi(s) \approx 1$ when $s > 0$ and $\phi(s) \approx 0$ when $s < 0$). Thus adaptation follows trajectories on the landscape along which fitness increases monotonically. This approximation is common for studying dynamics on model landscapes (see Sec. 1.5).

A different approximation holds when $N \gg 1$ but $N^{-1} \ll |s| \ll 1$. In this case, $\phi(s) \approx s$ for $s > 0$ and $\phi(s) \approx 0$ for $s < 0$ [13]. Thus deleterious mutations always get eliminated as before, but beneficial mutations fix at the rate Nus [23]. The true dynamics of real populations may not always fall into these special cases, instead involving more complex dynamics such as interference between multiple simultaneous mutations [24]. However, the simplified evolutionary dynamics described here are useful when our main objective is to qualitatively understand the role of the fitness landscape in constraining evolution over long time scales, rather than fine-grained details of short-time population dynamics.

1.2. *Epistasis*

The most basic aspect of fitness landscape topography is known as *epistasis*. Let the sequence σ be $\sigma^1\sigma^2\ldots\sigma^L$, where σ^μ is the letter at site $\mu \in \{1,\ldots,L\}$. In general the multiplicative fitness function $\mathcal{F}(\sigma)$ cannot be decomposed into a product of independent contributions from each site μ or, equivalently, the additive fitness $\log\mathcal{F}(\sigma)$ cannot be decomposed into a sum. This means that the fitness effect of a mutation at a given site may depend on the state of other sites. If this is true, the sites will be correlated, which can be thought of as a coupling between the sites. Mathematically this is reminiscent of a Hamiltonian for a system of interacting particles.

Epistasis is precisely this interactive coupling in the context of genotypic sequences. Following convention, we use additive fitness here ($\log\mathcal{F}$), and

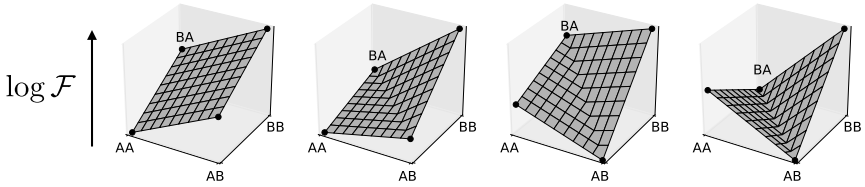

Fig. 1. The four qualitative types of epistasis for a two-letter, two-site model. From left to right: *no epistasis*, where each mutation has the same effect on additive fitness regardless of the state of the other site, yielding a linear landscape; *magnitude epistasis*, where the magnitude (but not the sign) of the additive fitness effect of a mutation depends on the other site; *sign epistasis*, where the sign of a mutation's fitness effect (beneficial or deleterious) depends on the other site; *reciprocal sign epistasis*, where multiple instances of sign epistasis can lead to local fitness maxima.

categorize types of epistasis according to the qualitative differences in the additive fitness effects of mutations. We summarize the four possible cases using a two-letter, two-site model in Fig. 1 in which sequence AA evolves into sequence BB, which has the highest fitness. In the first case on the left of Fig. 1, there is no epistasis: the fitness effect of $\mathsf{A} \to \mathsf{B}$ substitution at site 2 is the same regardless of the state of site 1, and vice versa. Thus the additive fitness can be decomposed into a sum of contributions from each site: $\log \mathcal{F}(\sigma) = \log \mathcal{F}_1(\sigma^1) + \log \mathcal{F}_2(\sigma^2)$, i.e., the additive fitness landscape is linear in sequence space. Under the aforementioned strong-selection evolutionary dynamics, both trajectories from AA to the global maximum at BB are accessible.

In the second case of Fig. 1, the additive fitness effect of $\mathsf{A} \to \mathsf{B}$ at site 2 differs in magnitude but not in sign depending on whether site 1 has A or B. This situation is known as *magnitude epistasis* [25, 26]. Magnitude epistasis does not completely block any trajectories, but affects quantitative aspects of dynamics such as adaptation times. Note that there are two kinds of magnitude epistasis: one in which the fitness benefit of a mutation is enhanced by the presence of other mutations (as shown in the second panel on the left of Fig. 1), and the "diminishing returns" case in which the fitness benefit is decreased by other mutations.

The third case of Fig. 1 shows how the $\mathsf{A} \to \mathsf{B}$ substitution at site 2 can have opposite effects on fitness depending on the state of site 1: it is deleterious if $\sigma^1 = \mathsf{A}$, but beneficial if $\sigma^1 = \mathsf{B}$. Since the sign of the fitness effect depends on the other site, the situation is known as *sign epistasis*. Sign epistasis can significantly affect accessibility of genotypes

on the landscape by blocking trajectories: under strong-selection dynamics, the trajectory $\mathsf{AA} \to \mathsf{AB} \to \mathsf{BB}$ is unavailable since it requires a deleterious substitution. When sign epistasis exists at multiple sites, it is known as *reciprocal sign epistasis*, as shown in the fourth case of Fig. 1. Reciprocal sign epistasis is a necessary condition for the existence of multiple local maxima [27]. As the examples in Fig. 1 show, epistasis underlies landscape ruggedness that can constrain evolutionary trajectories. Thus the existence and nature of epistasis is of prime interest in evolution.

1.3. *Measures of landscape ruggedness and accessibility*

Numerous measures have been proposed to quantify epistatic ruggedness of fitness landscapes and accessibility of evolutionary trajectories (summarized in [12]). One commonly-used measure is the number of local fitness maxima, which is indicative of the presence and type of epistasis [27]: more local maxima indicates a more rugged or epistatic landscape. For binary alphabets, deviations of the additive fitness function from linearity can be quantified by fitting a linear function and calculating the sum of squares of residuals, known as a roughness parameter [12]. A more local measure of ruggedness can be obtained by considering all pairs of sites and all pairs of possible letters at those sites, and then determining the sub-landscape for each combination like those shown in Fig. 1. Each sub-landscape can then be classified into one of the four categories of epistasis.

Other measures consider accessibility and other properties of the first-passage trajectories themselves, especially those leading to the global fitness maximum. For example, without epistasis all first-passage trajectories to the global maximum from any other point on the landscape are accessible under strong-selection dynamics, but with sign epistasis some trajectories become blocked. The distributions of first-passage trajectory times and lengths are important for understanding the effect of landscape ruggedness on adaptation [28–32].

1.4. *Repeatability of evolution and diversity of evolutionary trajectories*

Landscape ruggedness is especially relevant in its effect on the repeatability of evolution, a question of paramount importance in biology [33]. If "life's tape" could be replayed, would we see a completely different outcome because evolution is a largely stochastic phenomenon, or are accessible evolutionary trajectories so constrained that the outcome would have been the

same or recognizably similar [15]?

Discussing this question in general entails many issues, including environmental conditions, initial conditions, and details of population dynamics. One specific approach focuses on the diversity of first-passage trajectories leading from an ancestral state to a particular descendant state or set of states. One assessment of this diversity is simply counting the number of such trajectories. For example, Weinreich and co-workers found that only a small fraction of all possible trajectories from wild-type *E. coli* to a strain resistant to antibiotics was accessible to adaptation [34, 35]. In another approach, Koonin and co-workers devised a measure called mean path divergence [33, 36]:

$$\mathcal{D} = \sum_{\varphi_1 \neq \varphi_2} \Delta(\varphi_1, \varphi_2) p(\varphi_1) p(\varphi_2), \tag{2}$$

where the sum is over all pairs of distinct paths in an ensemble, $p(\varphi)$ is the probability of path φ, and $\Delta(\varphi_1, \varphi_2)$ is the path distance between φ_1 and φ_2. The path distance is defined as the average of the shortest Hamming distances between each sequence σ_1 on path φ_1 and all sequences on path φ_2, and vice versa:

$$\Delta(\varphi_1, \varphi_2) = \frac{1}{\mathcal{L}[\varphi_1] + \mathcal{L}[\varphi_2]} \left(\sum_{\sigma_1 \in \varphi_1} \operatorname*{argmin}_{\sigma_2 \in \varphi_2} d(\sigma_1, \sigma_2) + \sum_{\sigma_2 \in \varphi_2} \operatorname*{argmin}_{\sigma_1 \in \varphi_1} d(\sigma_2, \sigma_1) \right), \tag{3}$$

where $\mathcal{L}[\varphi]$ is the length (number of jumps) of path φ and $d(\sigma_1, \sigma_2)$ is the Hamming distance between σ_1 and σ_2. The divergence therefore captures not only how many paths are available, but weighs them by their proximity in sequence space. Other measures of path diversity, such as path entropy and the distribution of path lengths and times [32], are discussed in Secs. 2 and 3.

1.5. *Model and empirical landscapes*

A few simple models have traditionally dominated theoretical studies of fitness landscapes and evolutionary trajectories. These models serve as useful null hypotheses or limits of more complex scenarios; they generally consider sequences with binary alphabets ($k = 2$), in which case sequence space is a

unit hypercube. Without attempting to account for every landscape proposed in the literature, we will discuss and motivate several popular choices. In Kauffman's NK (or "LK") model [31, 37, 38], each of the L sites in the gene (or genes in the genome) interacts with K other sites chosen by random sampling. The additive fitness of genotype σ is given by

$$\log \mathcal{F}(\sigma) = \sum_{\mu=1}^{L} \log \mathcal{F}_\mu(\sigma^\mu, \sigma^{n_1(\mu)}, \ldots, \sigma^{n_K(\mu)}), \tag{4}$$

where $n_1(\mu), \ldots, n_K(\mu)$ are the randomly-chosen interaction partners of site μ. The single-site fitnesses $\mathcal{F}_\mu$ are obtained by sampling from a continuous distribution; each combination of 2^{K+1} possible states of the argument corresponds to an independent sampling. When $K = 0$, the NK landscape becomes fully additive and thus non-epistatic. Because in this limit the landscape is smooth and has a single maximum, it is sometimes called the "Mount Fuji" model [39]. The amount of epistasis, or landscape ruggedness, can be tuned by increasing K to the maximum value of $L - 1$. With $K = L - 1$, the fitnesses of different sequences are uncorrelated; this model is called the "House of Cards" [40] due to the unpredictable fitness effects of mutations. Realistically, closely-related genotypes should have correlated fitnesses, so this limit serves mainly as a null model. Many results are known for the NK landscape and its adaptive first-passage trajectories [28, 30, 37, 41, 42]. For example, in the $K = L - 1$ limit the average number of local maxima is $k^L/(L(k-1)+1)$ for any alphabet size k [12].

Another class of models starts from a non-epistatic landscape and adds a tunable amount of ruggedness to it. For example, in the "rough Mount Fuji" model [39], sequence σ_0 is arbitrarily picked as the global maximum and the fitness of sequence σ is given by

$$\log \mathcal{F}(\sigma) = \eta(\sigma) - \theta d(\sigma, \sigma_0), \tag{5}$$

where $d(\sigma, \sigma_0)$ is the Hamming distance between sequences σ and σ_0, θ is the parameter which controls the slope of the smooth part of the landscape, and $\eta(\sigma)$ is a zero-mean random variable sampled independently for each sequence σ. The ruggedness of the landscape is controlled by the ratio of θ and the standard deviation of the distribution from which the random variables $\eta(\sigma)$ are sampled.

The landscapes described above are dominated by selection. Another approach to evolution is based on the neutral theory, which postulates that

the majority of mutations have either no phenotypic effect (i.e., are selectively neutral) or are strongly deleterious and thus rapidly removed from the population [19]. This picture leads to evolution on a neutral network where all viable nodes have the same fitness [10, 43]. With some probability, a viable individual can acquire a lethal mutation from which it cannot recover, and will disappear from the population. Thus the population as a whole can only make transitions between viable, selectively-neutral states. Evolution on such a landscape is reminiscent of the percolation problem [44]: each node is assigned fitness 1 with probability p and fitness zero with probability $1 - p$, independent of the other nodes [31].

Due to the enormous number of sequences involved, the structure of fitness landscapes is difficult to probe experimentally. Typically, only a small number of sites is studied (approximately 4 to 9, summarized in [12]), and at those sites only a subset of all possible mutations is introduced, resulting in fitness measurements for tens or hundreds of different genotypes. In addition, because genotype survivability is not directly accessible in experiments, proxy measures of fitness are employed, such as growth rates and antibiotic resistance. Although these empirical studies can be used to probe the local structure of the landscapes, they are insufficient for analyzing the global properties of adaptive trajectories because adaptation may involve mutations outside of the experimentally-probed subset.

Nevertheless, many studies have attempted to characterize empirical landscapes in terms of their epistatic features, accessibility, and correspondence to theoretical models [12, 25, 31, 34–36, 45–47]. For example, magnitude and sign epistasis have been observed, as well as significant constraints on evolutionary trajectories [34, 35, 46, 47]. One general finding of such studies is that empirical landscapes include some epistasis, but are far from the House of Cards regime in which all fitness values are completely uncorrelated [25]. The emerging picture is closer to the rough Mount Fuji model, which includes a limited amount of epistasis around a mostly linear landscape [12].

2. Statistical physics of stochastic paths

Analytical treatments of evolutionary dynamics on fitness landscapes are typically restricted to uncorrelated or highly symmetric models, such as those described in Sec. 1.5. Simulations, meanwhile, can suffer from numerical inaccuracy and may be computationally expensive when rare events are considered. More systematic tools are necessary, especially tools that

directly address statistical properties of stochastic paths that are relevant for understanding the diversity of evolutionary pathways.

Physics and chemistry have long grappled with similar problems in the field of reaction rate theory [48], which studies rare transitions between metastable states that model phenomena ranging from protein folding [3] to chemical reactions [49]. In these systems, quantities of interest include not only mean first-passage times and reaction rates but also the spatial distribution of transition paths and identification of kinetic bottlenecks.

Transition state theory is a well-known approach to these problems; however, it relies on the existence and *a priori* identification of key transition states [48]. A more recent development has been transition path sampling [49–54], in which paths are directly sampled via Monte Carlo to estimate their statistical properties. Similar methods have been used in phylogenetic analysis of protein sequences [55–59]. These techniques are based on a finite sample of paths and do not provide natural cutoffs for the size of the sample, which may lead to noisy estimates of various path statistics. Another technique, called transition path theory [60–64], relies on explicit solutions to the Kolmogorov backward equation. This approach, though more systematic, does not directly address the diversity of paths.

Here we discuss a general statistical physics treatment of stochastic paths that provides many useful tools for analyzing evolutionary models and other stochastic processes [32]. A semi-Markov process (i.e., continuous-time random walk [65]) on the discrete state space $\mathcal{S}$ consists of jumps between states and continuous-time waiting within states; the jump process is memoryless, but the waiting process need not be. Thus the process is defined by a set of jump probabilities, $\langle\sigma'|\mathbf{Q}|\sigma\rangle$ for the jump $\sigma \to \sigma'$ ($\sigma, \sigma' \in \mathcal{S}$), and waiting time distributions $\psi_\sigma(t)$, which is the probability density of waiting exactly time t in state σ before jumping out. We assume that $\psi_\sigma(t)$ has finite mean $w(\sigma)$ for all $\sigma \in \mathcal{S}$. For fully Markov processes with memoryless waiting, $\psi_\sigma(t) = e^{-t/w(\sigma)}/w(\sigma)$; non-exponential $\psi_\sigma(t)$ may arise due to coarse-graining a Markov process [66–68]. Note that the space $\mathcal{S}$ equipped with the jump matrix $\mathbf{Q}$ defines a network with directed and weighted edges.

Define a trajectory as a path through state space $\varphi = \{\sigma_0, \sigma_1, \ldots, \sigma_\ell\}$ combined with a set of intermediate waiting times $\{t_0, t_1, \ldots, t_{\ell-1}\}$, where t_i is the waiting time in σ_i. We consider the trajectory finished once it reaches the final state σ_ℓ, and thus do not count the waiting time in that state. The probability functional of starting in the initial state σ_0 and

completing the path φ no later than time t is

$$\Pi[\varphi,t] = \pi(\sigma_0)\left(\prod_{i=0}^{\ell-1}\langle\sigma_{i+1}|\mathbf{Q}|\sigma_j\rangle\right) \times\left(\prod_{i=0}^{\ell-1}\int_0^\infty dt_i\ \psi_{\sigma_i}(t_i)\right)\Theta\left(t-\sum_{i=0}^{\ell-1}t_i\right), \quad (6)$$

where the first factor is the initial state probability $\pi(\sigma_0)$, the second is the product of jump probabilities, the third integrates over waiting times, and the fourth constrains the total waiting time to be less than t (Θ is the Heaviside step function). In the $t\to\infty$ limit we obtain the probability of the path φ for any duration,

$$\Pi_\infty[\varphi] = \pi(\sigma_0)\prod_{i=0}^{\ell-1}\langle\sigma_{i+1}|\mathbf{Q}|\sigma_i\rangle, \quad (7)$$

which is just the product of jump probabilities. In the time-dependent case, the Laplace transform of Eq. (6) results in a simpler expression through deconvolution [69]:

$$\tilde{\Pi}[\varphi,s] = \frac{\pi(\sigma_0)}{s}\prod_{i=0}^{\ell-1}\langle\sigma_{i+1}|\mathbf{Q}|\sigma_i\rangle\ \tilde{\psi}_{\sigma_i}(s), \quad (8)$$

where $\tilde{\psi}_{\sigma_i}(s)$ is the Laplace transform of $\psi_{\sigma_i}(t)$. For Markov processes, Eq. (8) becomes [70]

$$\tilde{\Pi}[\varphi,s] = \frac{\pi(\sigma_0)}{s}\prod_{i=0}^{\ell-1}\frac{\langle\sigma_{i+1}|\mathbf{Q}|\sigma_i\rangle}{1+sw(\sigma)}. \quad (9)$$

2.1. *Path ensemble averages*

The distribution $\Pi[\varphi,t]$ in principle contains all statistical information on a set of paths. However, direct analysis of this distribution is typically prohibitive due to the high dimensionality of path space. The simplest alternative entails taking averages of various path properties over this distribution. Let Φ be an ensemble of paths that defines some dynamical process; for example, this may be all first-passage paths from a set of initial states $\mathcal{S}_i$ to a set of final states $\mathcal{S}_f$. The partition function for this

ensemble is

$$\mathcal{Z}_\Phi(t) = \sum_{\varphi\in\Phi} \Pi[\varphi, t], \tag{10}$$

which represents the total probability of reaching $\mathcal{S}_f$ from $\mathcal{S}_i$ by time t via paths in Φ.

We define the following path functionals:

$$\begin{aligned} &\mathcal{L}[\varphi] = \text{length (number of jumps) of } \varphi, \quad && \mathcal{I}_\sigma[\varphi] = \begin{cases} 1 \text{ if } \sigma \in \varphi, \\ 0 \text{ otherwise,} \end{cases} \\ &\mathcal{T}[\varphi] = \sum_{i=0}^{\ell-1} w(\sigma_i), && \mathcal{T}_\sigma[\varphi] = \sum_{i=0}^{\ell-1} \delta_{\sigma,\sigma_i} w(\sigma_i), \end{aligned} \tag{11}$$

where δ is the Kronecker delta. We can now express various path statistics as averages of these functionals over the ensemble, conditioned on completing the process by time t. For example, the average time of paths is given by [53]

$$\bar{\tau}_\Phi(t) = \langle \mathcal{T}(t) \rangle_\Phi = \frac{1}{\mathcal{Z}_\Phi(t)} \sum_{\varphi\in\Phi} \mathcal{T}[\varphi]\Pi[\varphi, t]. \tag{12}$$

The distribution of path lengths is given by

$$\rho_\Phi(\ell, t) = \frac{1}{\mathcal{Z}_\Phi(t)} \sum_{\varphi\in\Phi} \delta_{\ell,\mathcal{L}[\varphi]}\Pi[\varphi, t], \tag{13}$$

from which the average length $\bar{\ell}_\Phi(t) = \langle \mathcal{L}(t) \rangle_\Phi$ and standard deviation of length $\ell^{\rm sd}_\Phi(t)$ are readily obtained.

Averages over state-dependent functionals can be used to characterize the spatial structure of paths. For example, the fraction of time paths spend in a state σ can be expressed as $\langle \mathcal{T}_\sigma(t) \rangle_\Phi / \bar{\tau}_\Phi(t)$; this is a normalized distribution over all states $\sigma \in \mathcal{S}$ and therefore it represents the density of states on the paths in the ensemble Φ. The quantity $\langle \mathcal{T}_\sigma(t) \rangle_\Phi / w(\sigma)$ gives the average number of visits to state σ. The probability that a path will visit a state σ at all is given by $\langle \mathcal{I}_\sigma(t) \rangle_\Phi$, which we will refer to as the density of paths in the ensemble Φ. We can also construct the two-point correlation function $\langle \mathcal{I}_{\sigma'}(t)\mathcal{I}_\sigma(t) \rangle_\Phi$, which gives the probability of paths passing through both σ and σ'.

In many cases we are interested in the time-independent versions of these quantities, i.e., statistical properties of paths taking any amount of time to finish. These can be obtained as the $t \to \infty$ limit of the above expressions, which amounts to replacing $\Pi[\varphi, t]$ (Eq. (6)) with $\Pi_\infty[\varphi]$ (Eq. (7)). We will denote these time-independent properties by simply omitting the time dependence, e.g., $\lim_{t\to\infty} \bar{\tau}_\Phi(t) = \bar{\tau}_\Phi$. The convergence of these limits depends on the path ensemble Φ. For example, if Φ includes all possible paths connecting the initial and final states, including those that visit the final state multiple times, then these limits generally diverge: typically, there is no finite average time or length for these paths. However, restriction to first-passage paths in Φ, as is our focus here, guarantees convergence.

This formalism also allows for development of path thermodynamics. The entropy of the path ensemble is given by

$$\begin{aligned} S_\Phi(t) &= -\frac{1}{\mathcal{Z}_\Phi(t)} \sum_{\varphi\in\Phi} \Pi[\varphi, t] \log\left(\frac{\Pi[\varphi, t]}{\mathcal{Z}_\Phi(t)}\right) \\ &= -\langle \log \Pi(t)\rangle_\Phi + \log \mathcal{Z}_\Phi(t). \end{aligned} \tag{14}$$

Indeed, if we define the path Hamiltonian to be

$$\mathcal{H}[\varphi, t] = -\log(\Pi[\varphi, t]), \tag{15}$$

(so that $\Pi[\varphi, t] = e^{-\mathcal{H}[\varphi,t]}$), we can express the path ensemble free energy as

$$F_\Phi(t) = \langle \mathcal{H}(t)\rangle_\Phi - S_\Phi(t) = -\log \mathcal{Z}_\Phi(t). \tag{16}$$

The partition function $\mathcal{Z}_\Phi(t)$ monotonically increases with time. Therefore the free energy $F_\Phi(t)$ monotonically decreases as $t \to \infty$, corresponding to equilibration of the path ensemble.

For recurrent processes (i.e., where the system will almost surely reach the final states eventually [67]), $\lim_{t\to\infty} \mathcal{Z}_\Phi(t) = \mathcal{Z}_\Phi = 1$, and hence equilibrium free energy is zero. In these cases, equilibrium path entropy is equal to the average Hamiltonian. If the ensemble Φ consists of only a single path with nonzero probability, its entropy is $S_\Phi = 0$. This situation may arise if a landscape is so constrained that only a single viable pathway exists between the initial and final states. In contrast, consider a purely random walk on a homogeneous network with γ nearest neighbors per node. The jump probability between any pair of neighboring nodes is thus γ^{-1}, so any

path φ has probability $\Pi_\infty[\varphi] = \gamma^{-\mathcal{L}[\varphi]}$, and the entropy of the ensemble is given by

$$S_\Phi = -\langle \log \Pi_\infty \rangle_\Phi = \bar{\ell}_\Phi \log \gamma. \tag{17}$$

Note that path entropy and the average path Hamiltonian scale with the average path length, which defines a notion of extensivity in the path ensemble. This is sensible if we think of a path as a gas of particles, where each jump in the path corresponds to a particle. The path ensemble, which includes paths of many lengths, therefore is equivalent to the grand canonical ensemble of the gas. In the case of the gas, extensive quantities like entropy and energy scale with the number of particles, and hence these quantities here scale with the path length.

2.2. *Numerical algorithm*

The factorized form of the path probability distribution functional (Eqs. (7)–(9)) permits efficient calculation of path ensemble averages via a recursive algorithm [32]. Here for simplicity we consider the time-independent case, and thus assume that Φ consists of first-passage paths to guarantee convergence of path averages. Let $|\sigma\rangle$ be the vector with 1 at position σ and zero otherwise, and let $|\pi\rangle = \sum_\sigma \pi(\sigma)|\sigma\rangle$ be the vector of initial state probabilities. For each jump ℓ and intermediate state σ, we calculate $P_\ell(\sigma) = \langle\sigma|\mathbf{Q}^\ell|\pi\rangle$, the total probability of all paths that end at σ in ℓ jumps; $T_\ell(\sigma)$, the total average time of all such paths; and $\Gamma_\ell(\sigma)$, the total entropy of all such paths. These quantities obey the following recursion relations:

$$\begin{aligned}
P_\ell(\sigma') &= \sum_{\text{nn } \sigma \text{ of } \sigma'} \langle\sigma'|\mathbf{Q}|\sigma\rangle P_{\ell-1}(\sigma), \\
T_\ell(\sigma') &= \sum_{\text{nn } \sigma \text{ of } \sigma'} \langle\sigma'|\mathbf{Q}|\sigma\rangle \left[T_{\ell-1}(\sigma) + w(\sigma) P_{\ell-1}(\sigma)\right], \\
\Gamma_\ell(\sigma') &= \sum_{\text{nn } \sigma \text{ of } \sigma'} \langle\sigma'|\mathbf{Q}|\sigma\rangle \left[\Gamma_{\ell-1}(\sigma) - \log\langle\sigma'|\mathbf{Q}|\sigma\rangle P_{\ell-1}(\sigma)\right],
\end{aligned} \tag{18}$$

where $P_0(\sigma) = \pi(\sigma)$ and $T_0(\sigma) = \Gamma_0(\sigma) = 0$ for all $\sigma \in \mathcal{S}$, and the sums run over all nearest neighbors (nn) σ of σ'. The final states $\sigma \in \mathcal{S}_f$ are treated as absorbing to ensure that only first-passage paths are counted. This procedure can be considered a generalization of the exact-enumeration

algorithm of [71]. Path ensemble averages are then given by

$$\mathcal{Z}_\Phi = \sum_{\ell=1}^{\infty} \sum_{\sigma \in \mathcal{S}_f} P_\ell(\sigma), \qquad \rho_\Phi(\ell) = \frac{1}{\mathcal{Z}_\Phi} \sum_{\sigma \in \mathcal{S}_f} P_\ell(\sigma), \tag{19}$$

$$\bar{\tau}_\Phi = \frac{1}{\mathcal{Z}_\Phi} \sum_{\ell=1}^{\infty} \sum_{\sigma \in \mathcal{S}_f} T_\ell(\sigma), \qquad S_\Phi = \frac{1}{\mathcal{Z}_\Phi} \sum_{\ell=1}^{\infty} \sum_{\sigma \in \mathcal{S}_f} \Gamma_\ell(\sigma).$$

We can similarly calculate state-dependent quantities such as $\langle \mathcal{I}_\sigma \rangle_\Phi$ and $\langle \mathcal{T}_\sigma \rangle_\Phi$. The two quantities to be recursively updated are $P_\ell(\sigma';\sigma)$, the total probability of all paths currently at σ' at jump ℓ that have visited σ at least once previously, and $T_\ell(\sigma';\sigma)$, the total average time that all such paths have spent in σ. These obey the following recursion relations:

$$P_\ell(\sigma';\sigma) = \begin{cases} \displaystyle\sum_{\text{nn } \sigma'' \text{ of } \sigma'} \langle \sigma' | \mathbf{Q} | \sigma'' \rangle P_{\ell-1}(\sigma'';\sigma), & \sigma' \neq \sigma, \\ P_\ell(\sigma), & \sigma' = \sigma, \end{cases} \tag{20}$$

$$T_\ell(\sigma';\sigma) = \sum_{\text{nn } \sigma'' \text{ of } \sigma'} \langle \sigma' | \mathbf{Q} | \sigma'' \rangle [T_{\ell-1}(\sigma'';\sigma) + \delta_{\sigma,\sigma''} w(\sigma'') P_{\ell-1}(\sigma'';\sigma)],$$

with the initial conditions $P_0(\sigma';\sigma) = T_0(\sigma';\sigma) = 0$ for all $\sigma, \sigma' \in \mathcal{S}$, $\sigma \neq \sigma'$ ($P_0(\sigma;\sigma) = \pi(\sigma)$, $T_0(\sigma;\sigma) = 0$). Averages are then expressed as

$$\langle \mathcal{I}_\sigma \rangle_\Phi = \frac{1}{\mathcal{Z}_\Phi} \sum_{\ell=1}^{\infty} \sum_{\sigma' \in \mathcal{S}_f} P_\ell(\sigma';\sigma), \qquad \langle \mathcal{T}_\sigma \rangle_\Phi = \frac{1}{\mathcal{Z}_\Phi} \sum_{\ell=1}^{\infty} \sum_{\sigma' \in \mathcal{S}_f} T_\ell(\sigma';\sigma). \tag{21}$$

Furthermore, we can calculate mean path divergence that characterizes the spatial diversity of the paths in Φ:

$$\mathcal{D}_\Phi = \sum_{\ell=1}^{\infty} \sum_{\sigma,\sigma' \in \mathcal{S}} d(\sigma,\sigma') P_\ell(\sigma) P_\ell(\sigma'), \tag{22}$$

where $d(\sigma,\sigma')$ is a distance metric on $\mathcal{S}$. This definition is distinct from that proposed in [33, 36] (Eq. (2)) in that it dynamically calculates distances between points on paths as they propagate, rather than comparing the minimal distance between complete paths. Thus for a path that revisits

some states multiple times, the divergence with a path that travels through the same set of states without revisiting any of them will be zero according to Eq. (2), but nonzero with the definition in Eq. (22).

This algorithm allows for very general definitions of the path ensemble Φ without having to explicitly enumerate all paths. For instance, Φ can include paths that begin and end at arbitrary sets of states, or are prohibited from passing through arbitrary sets of intermediate states. The time complexity of the algorithm is $\mathcal{O}(\gamma N\Lambda)$ for $\mathcal{Z}_\Phi$, $\rho_\Phi(\ell)$, $\bar{\tau}_\Phi$, S_Φ, and $\mathcal{O}(\gamma N^2\Lambda)$ for $\langle\mathcal{I}_\sigma\rangle_\Phi$, $\langle\mathcal{T}_\sigma\rangle_\Phi$, $\mathcal{D}_\Phi$, where γ is the average number of nearest neighbors, N is the number of states visited by paths in Φ, and Λ is the cutoff path length. The cutoff Λ scales with network size N in the same way as the average path length $\bar{\ell}_\Phi$; for simple random walks, it is known that

$$\Lambda \sim \bar{\ell}_\Phi \sim \begin{cases} N^{d_w/d_f}, & d_w \geq d_f \quad \text{(compact exploration)}, \\ N, & d_w < d_f \quad \text{(non-compact exploration)}, \end{cases} \tag{23}$$

where d_w is the dimension of the walk and d_f is the fractal dimension of the space [72, 73]. Therefore, the algorithm scales as

$$\mathcal{O}(\gamma N\Lambda) = \begin{cases} \mathcal{O}(\gamma N^{1+d_w/d_f}), & d_w \geq d_f, \\ \mathcal{O}(\gamma N^2), & d_w < d_f, \end{cases} \tag{24}$$

automatically accounting for the sparseness of network connections. This scaling compares favorably with standard linear algebra algorithms, which in general require $\mathcal{O}(N^3)$ operations [74] to solve the backward equation [62, 63].

2.3. *Evolution on a neutral network*

As a simple application of this approach, we consider a population evolving on a neutral network [31]. In the space of all sequences of length L and with an alphabet of size k, we assign each sequence fitness 1 with probability p or fitness zero with probability $1-p$. The subset of fit states connected to each other forms a neutral network; there can be several disconnected neutral networks in each landscape realization. All jumps between neighboring fit states occur at the same rate, and waiting times are Markovian. We choose $L = 8$ and a binary alphabet $\{\mathsf{A}, \mathsf{B}\}$ ($k = 2$), which gives $2^8 = 256$ total nodes in the network, and we consider the ensemble Φ of first-passage paths from the sequence AAAAAAAA to the sequence BBBBBBBB.

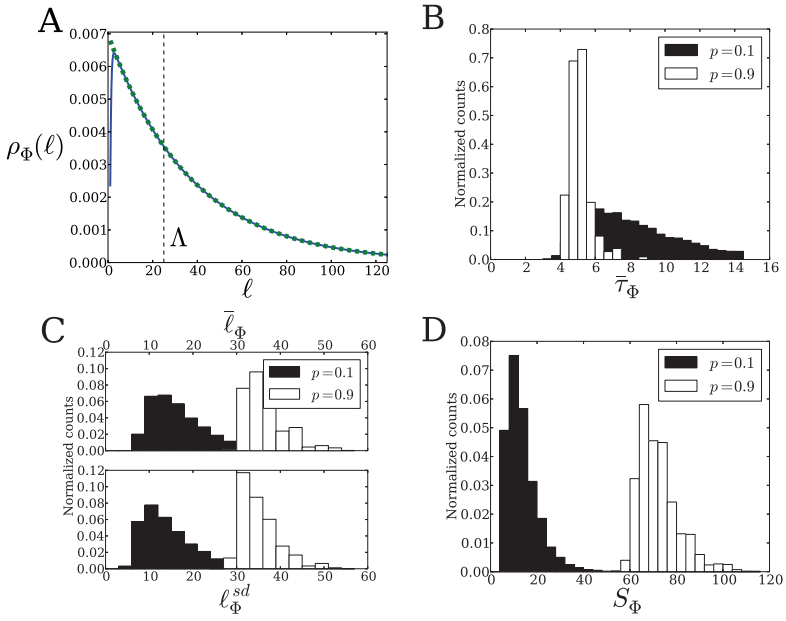

Fig. 2. (Color online) First-passage path ensemble statistics in a neutral network model. (A) The path length distribution $\rho_\Phi(\ell)$ (solid, blue) and exponential fit (dashed, green) in the interval $[\Lambda - 5, \Lambda]$ for $\Lambda = 25$ in a single realization of a neutral network with $p = 0.9$. (B) Distribution of mean path times $\bar{\tau}_\Phi$, (C) distribution of mean path lengths $\bar{\ell}_\Phi$ and standard deviations of path lengths $\ell_\Phi^{\rm sd}$, and (D) distribution of path entropies S_Φ for $p = 0.1$ and $p = 0.9$. All quantities in (A)-(D) are per site. Histograms in (B)-(D) are generated from 10^4 successful random realizations of the neutral network for each value of p; a realization is considered successful if both initial and final states are included in a single connected network.

Figure 2A shows $\rho_\Phi(\ell)$ for a single realization of this model with $p = 0.9$. The exponential tail of $\rho_\Phi(\ell)$ is a universal feature of first-passage processes on finite spaces [72]; other path statistics, such as the average time $\bar{\tau}_\Phi(\ell)$ of paths up to length ℓ, also show asymptotic behavior that is exponential for long paths. We can use this feature to determine the cutoff path length Λ for the algorithm: Λ is set at a length such that $\rho_\Phi(\ell)$ and $\bar{\tau}_\Phi(\ell)$ are close to exponential in a region around Λ. Then one need only consider paths with $\ell < \Lambda$ and infer the contributions of all longer paths from an exponential fit to the tail, which considerably improves the efficiency of the algorithm. This procedure takes advantage of the fact that information about longer paths is already contained in the structure of shorter paths; the longer

paths are built on the shorter paths by adding loops. The maximum length Λ of the shorter paths that must be explicitly calculated depends on the chemical distance between the initial and final states and the lengths over which the landscape is correlated. This essentially implements a numerical renormalization scheme on the ensemble of paths [70].

In Fig. 2B,C,D we show distributions of the mean path time $\bar{\tau}_\Phi$, mean path length $\bar{\ell}_\Phi$, path length standard deviation $\ell_\Phi^{\rm sd}$, and path entropy S_Φ for multiple realizations of the neutral network with high and low values of p. We see that long paths are likely in these models: dozens of substitutions can occur at each site before the final state is reached. The larger size of the neutral network for $p = 0.9$ allows longer paths on average than for $p = 0.1$. However, the mean time of paths for the larger neutral network is usually smaller (Fig. 2B), since the increased connectivity of the network leads to shorter waiting times at individual nodes. Larger p leads to substantially more diversity of paths and path lengths, as expected due to the increased size and connectivity of the network (Fig. 2C,D). Note that the distributions of $\bar{\ell}_\Phi$ and $\ell_\Phi^{\rm sd}$ in Fig. 2C are very similar, owing to the nearly exponential distribution of $\rho_\Phi(\ell)$ in this model (cf. Fig. 2A).

In an unconstrained sequence space, the number of nearest neighbors is $\gamma = L(k-1)$, and the average path length $\bar{\ell}_\Phi$ scales as $N = k^L$ (Eq. (23)). According to Eq. (17), the entropy of paths in sequence space is

$$S_\Phi = \bar{\ell}_\Phi \log L(k-1) \sim k^L \log L(k-1). \tag{25}$$

When $p = 0.9$ the neutral network is nearly the size of the entire sequence space, and these results hold approximately. Indeed, we see that $\bar{\ell}_\Phi$ and S_Φ differ by roughly a factor of $\log L(k-1) \approx 2.1$ (Fig. 2C,D).

3. Biophysics of protein evolution

We now consider more realistic models of evolution based on protein biophysics, where the fitness landscape depends on protein folding stability and energetics of intermolecular interactions [3]. Many recent studies have focused on how proteins evolve under the constraint of maintaining thermodynamic stability of their folded state [75–82]. Suppose that an organism encodes a particular protein that is folded with probability $1/(1 + e^{\beta E_f})$, where E_f is the free energy of folding (i.e., the free energy difference between folded and unfolded states of the protein) and $\beta = 1.7$ (kcal/mol)$^{-1}$ is inverse room temperature. The protein contributes multiplicative fitness

1 if it is folded and $f_0 < 1$ if it is unfolded. Then the total fitness averaged over all proteins in an organism is given by

$$\mathcal{F}(E_f) = \frac{1 + f_0 e^{\beta E_f}}{1 + e^{\beta E_f}}. \tag{26}$$

Equation (26) states that robust protein folding confers a fitness advantage; in the extreme case of a protein essential to the organism, $f_0 = 0$ and so $\lim_{E_f \to +\infty} \mathcal{F}(E_f) = 0$. Some studies simplify this idea further by assuming that the folding energy E_f need only be below a particular threshold E_f^{thr}; below that threshold all proteins are adequately stable and equivalent in fitness [75, 77, 81]. Mathematically,

$$\mathcal{F}(E_f) = \Theta(E_f^{\text{thr}} - E_f), \tag{27}$$

where Θ is the Heaviside step function. This model is equivalent to the zero-temperature limit of Eq. (26). Similar approaches based on protein-DNA binding energies have also been used to study evolution of gene regulation [20, 83–87].

An extension of this model considers both protein stability and function [32], which we take to be the ability to bind a target molecule such as another protein (e.g., in a signaling pathway). Let E_b be the free energy of binding relative to the chemical potential of the target molecule, so that the probability of binding is $1/(1 + e^{\beta E_b})$. We assume that the protein contributes fitness 1 if it both folds and binds, and $f_0 < 1$ otherwise [88]. Then fitness averaged over all proteins in an organism is given by

$$\mathcal{F}(E_f, E_b) = \frac{1 + f_0(e^{\beta E_f} + e^{\beta E_b} + e^{\beta(E_f + E_b)})}{1 + e^{\beta E_f} + e^{\beta E_b} + e^{\beta(E_f + E_b)}}. \tag{28}$$

The folding and binding energies depend on the amino acid sequence σ. Many proteins have only a small number of residues at the binding interface that contribute the majority of the binding affinity; these are known as "hotspot" residues [89]. We assume that there are L such residues and that they make additive contributions to the total folding and binding free energies [90]:

$$E_f(\sigma) = E_f^0 + \sum_{\mu=1}^{L} \epsilon_f(\mu, \sigma^\mu), \quad E_b(\sigma) = E_b^0 + \sum_{\mu=1}^{L} \epsilon_b(\mu, \sigma^\mu), \tag{29}$$

where E_f^0, E_b^0 are overall offsets and $\epsilon_f(\mu, \sigma^\mu), \epsilon_b(\mu, \sigma^\mu)$ are the folding and binding energy contributions of amino acid σ^μ at position μ. The offset E_f^0 is a fixed contribution to the folding energy from all other residues in the protein, which we assume to be perfectly adapted. We sample ϵ_f's from a Gaussian distribution with mean 1.25 kcal/mol and standard deviation 1.6 kcal/mol, consistent with computational studies showing the mutational effects on stability are universally distributed [91].

We randomly assign a sequence $\sigma_{\rm bb}$ to be the best-binding sequence. Since binding hotspot residues typically have a minimum penalty of 1–3 kcal/mol for mutations away from the wild-type amino acid [92] (this requirement is used to define which residues make up the hotspot), we set $\epsilon_b(\mu, \sigma^\mu_{\rm bb}) = 0$ for all $\mu = \{1, \ldots, L\}$, and sample the other ϵ_b's from an exponential distribution defined in the range of $(1, \infty)$ kcal/mol, with mean 2 kcal/mol. This distribution is consistent with alanine-scanning experiments which probe energetics of amino acids at the binding interface [93]. Here we consider $L = 5$ hotspot residues and a reduced alphabet of $k = 8$ amino acids grouped by physico-chemical properties, resulting in $8^5 = 32768$ possible sequences. Different choices of these parameters can be considered, but they appear to have little effect on the overall qualitative features of the model.

We consider a population of individuals whose genomes encode a protein of interest. In each individual, the sequence of the protein is perfectly adapted to binding an original target molecule. Then the population is subjected to a selection pressure which favors binding a new target. This situation is common in directed evolution experiments which attempt to evolve new protein functions in a laboratory [94]. To model such experiments, we sample one set of ϵ_f's and two sets of ϵ_b's (one for each target), while E_f^0 and E_b^0 are assumed to be fixed. This procedure defines two fitness landscapes, $\mathcal{F}_1$ and $\mathcal{F}_2$, through Eq. (28); the entire population begins at the sequence with the global maximum on $\mathcal{F}_1$ and proceeds to adapt to a new global or local maximum on $\mathcal{F}_2$. We assume the strong-selection evolutionary dynamics as described in Sec. 1.1: any beneficial mutation that arises is guaranteed to fix in the population. Thus the substitution rate from σ to σ' is $\langle\sigma'|\mathbf{W}|\sigma\rangle = Nu$ if $\mathcal{F}(\sigma') > \mathcal{F}(\sigma)$ and zero otherwise, and the mean waiting time for sequence σ is $w(\sigma) = (\sum_{\rm nn\ \sigma'\ of\ \sigma}\langle\sigma'|\mathbf{W}|\sigma\rangle)^{-1} = (Nub(\sigma))^{-1}$, where $b(\sigma)$ is the number of beneficial mutations available from σ. Therefore the jump probability is $\langle\sigma'|\mathbf{Q}|\sigma\rangle = \langle\sigma'|\mathbf{W}|\sigma\rangle w(\sigma)$, which equals $1/b(\sigma)$ for a beneficial mutation and zero for a deleterious one. Note that in this limit the results are independent of f_0 and the population mutation rate

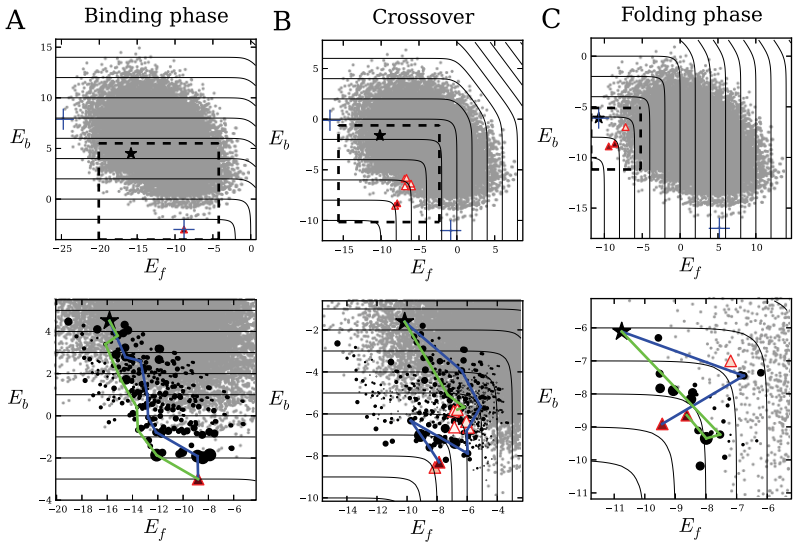

Fig. 3. (Color online) Three realizations of the fitness landscape. The offsets E_f^0 and E_b^0 are different for each realization, but ϵ_f's and the two sets of ϵ_b's (one for $\mathcal{F}_1$ and another for $\mathcal{F}_2$) are the same. (A) Binding phase, with $E_f^0 = -17$ kcal/mol and $E_b^0 = -3$ kcal/mol. (B) Crossover regime, with $E_f^0 = -9$ kcal/mol and $E_b^0 = -11$ kcal/mol. (C) Folding phase, with $E_f^0 = -3$ kcal/mol and $E_b^0 = -17$ kcal/mol. Top panels of (A)-(C) show the global distribution of all $8^5 = 32768$ sequences in energy space according to $\mathcal{F}_2$, where the blue crosses indicate the best-folding (σ_{bf}) and best-binding (σ_{bb}) sequences, red triangles indicate local fitness maxima on $\mathcal{F}_2$ (shaded according to their commitment probabilities), and black stars indicate the initial state for adaptation (sequence with global maximum on $\mathcal{F}_1$). Black lines are contours of constant fitness $\mathcal{F}_2$. In the bottom panels of (A)-(C) only the regions of energy space accessible to APs are shown; these regions are outlined by dashed lines in the top panels. Example APs are shown in blue and green; black circles indicate intermediate states along APs, sized proportional to the AP density $\langle \mathcal{I}_\sigma \rangle_{\mathrm{AP}}$; small gray circles are sequences inaccessible to APs.

Nu only affects the overall time scale. The path ensemble consists of all adaptive paths (APs), which are first-passage paths leading from the initial state to a local maximum on $\mathcal{F}_2$, with fitness monotonically increasing along each path. In Fig. 3 we show three realizations of $\mathcal{F}_2$ with examples of APs.

We focus on the generic properties of these landscapes, averaged over multiple realizations of ϵ_f and ϵ_b (Fig. 4). Varying E_f^0 and E_b^0 reveals two qualitatively different phases of adaptation. When E_f^0 is low and E_b^0 is high (see Fig. 3A for an example), adaptation is in the *binding phase*, i.e., the

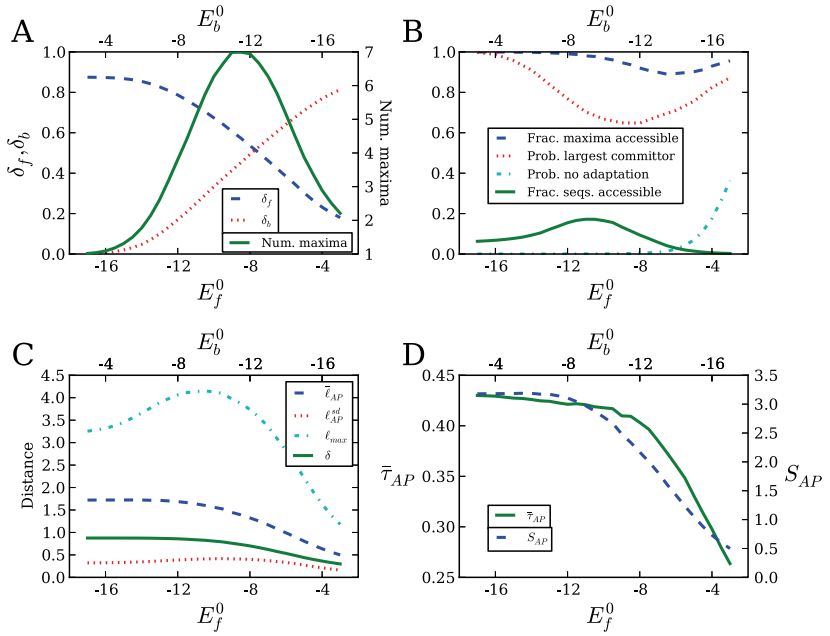

Fig. 4. (Color online) Statistics of fitness landscapes and adaptive paths averaged over multiple landscape realizations. All quantities in (A), (C), and (D) are per-residue (except the number of local maxima). (A) Average number of local fitness maxima (solid, green), average Hamming distance δ_f (number of mutations) between the maxima and the best-folding sequence $\sigma_{\rm bf}$ (dashed, blue), and average Hamming distance δ_b between the maxima and the best-binding sequence $\sigma_{\rm bb}$ (dotted, red) for the parameter subspace $E_f^0 + E_b^0 = -20$ kcal/mol. Note that the average distance between two random sequences is $1 - 1/k = 0.875$, where $k = 8$ is the size of the alphabet. (B) Fraction of local fitness maxima accessible from the initial state (dashed, blue), fraction of all landscape realizations in which the global maximum has the largest commitment probability (committor) among all local maxima (dotted, red), probability that the initial sequence starts at a local maximum resulting in no adaptation (dashed and dotted, cyan), and fraction of all sequences accessible to APs (solid, green). (C) Mean AP length $\bar{\ell}_{\rm AP}$, standard deviation $\ell_{\rm AP}^{\rm sd}$ of APs, average Hamming distance δ between the initial state and the final states, and average length $\ell_{\rm max}$ of the longest APs connecting the initial state with the final states. (D) Path ensemble entropy $S_{\rm AP}$ (dashed, blue) and the mean time of paths $\bar{\tau}_{\rm AP}$ (solid, green), in units of inverse population mutation rate $(Nu)^{-1}$. The probability of no adaptation in (B) is an average over 2×10^4 landscape realizations; all other data points are averages over 5×10^3 realizations, and realizations with no adaptation are excluded.

need to bind the new target molecule dominates evolutionary dynamics. In this phase, there is typically a single local fitness maximum which coincides with the best-binding sequence $\sigma_{\rm bb}$ (Fig. 4A). In contrast, when E_f^0 is high

and E_b^0 is low (see Fig. 3C for an example), adaptation is in the *folding phase*, where evolution is mostly constrained by the need to maintain or increase folding stability. In this case there are also few local maxima and they tend to be close in sequence space to the best-folding sequence σ_{bf} (Fig. 4A). Between these phases there is a crossover regime, where folding and binding compete more equally in shaping the landscape and adaptive dynamics (see Fig. 3B for an example). The crossover regime has the most epistasis, as indicated by the number of local maxima, the accessibility of those maxima, and the fraction of fitness landscape realizations in which the global maximum has the largest commitment probability (Fig. 4A,B). The differences in the landscape structure in the binding and folding phases lead to substantial differences in adaptive dynamics. In particular, APs are longer and take more time in the binding phase compared to the folding phase; they are also more diverse (Fig. 4C,D). Initial and final states in the binding phase are separated by longer Hamming distances (Fig. 4C). In the folding phase, there is an appreciable probability that no adaptation occurs at all, since the initial state may coincide with one of the local maxima (Fig. 4B).

This model reproduces several important features of molecular evolution observed in experimental studies. First of all, we see that adaptive dynamics involve tradeoffs between folding and binding as frequently observed in directed evolution experiments [94–96], even though mutational effects on folding and binding energies are uncorrelated [97]. This evolutionary coupling between folding and binding is introduced through nonlinearities in the fitness function $\mathcal{F}(E_f, E_b)$ (Eq. (28)), which contains both magnitude and sign epistasis. We note that although these landscapes are generated from randomly-drawn parameters ϵ_f and ϵ_b, similar to many classical model landscapes (Sec. 1.5), the protein landscapes studied here are highly correlated [98]: fitness values of k^L sequences are determined by $2Lk$ ϵ_f and ϵ_b parameters. Indeed, the average number of local maxima on a House of Cards landscape for the same sequence space ($L = 5$ and $k = 8$) is ≈ 910, while the protein landscape has on average no more than 7 (Fig. 4A). Thus, this model is far less epistatic than completely uncorrelated landscapes [38]. This more moderate level of epistasis is consistent with previous analyses of empirical fitness landscapes [12, 25, 36].

There are other features of the model that correspond to experimental observations. In the folding phase, APs tend to be short and no adaptation may occur if the old global maximum on $\mathcal{F}_1$ coincides with a new local maximum on $\mathcal{F}_2$ (Fig. 4B). This lack of adaptation is sometimes observed

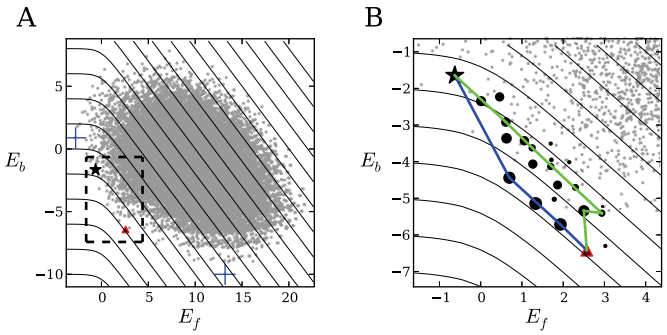

Fig. 5. (Color online) A realization of the fitness landscape for the three-state model (Eq. (30)) in the crossover regime, with $E_f^0 = 5$ kcal/mol and $E_b^0 = -10$ kcal/mol. As in Fig. 3, $k = 8$ and $L = 5$, resulting in 8^5 sequences. Note that ϵ_f's and the two sets of ϵ_b's (one for $\mathcal{F}_1$ and another for $\mathcal{F}_2$) are the same as in Fig. 3. Black lines are contours of constant fitness $\mathcal{F}_2$ (Eq. (30)). All symbols are identical to those in Fig. 3.

in experiments in which a protein already exhibiting some affinity for the new ligand cannot increase it any further [94]. Natural proteins are often found to have only marginal folding stability [76]; the model presented here shows that marginal stability can arise purely from the evolutionary tradeoff between binding and folding in the crossover regime, unlike previous hypotheses that explain it with mutational entropy [78] or a fitness function that explicitly disfavors hyperstable proteins [76].

The basic model described here can be generalized to account for many other aspects of protein evolution. For example, we can incorporate chaperone-assisted folding [99] by modifying E_f^0 or the distribution of ϵ_f's, and include "folding hotspots" away from the binding interface, which may acquire stabilizing mutations as a buffer against destabilizing but function-improving mutations at the interface [94, 95]. Neutral and weakly-selected mutations can be incorporated as well by using substitution rates from more complex population genetics models [18, 21], although we expect non-adaptive substitutions to play little role on short time scales.

Another important case is binding-mediated stability, in which binding stabilizes an otherwise disordered protein [100]. In this case, instead of considering folding and binding to be independent as in Eq. (28), which leads to four possible protein states, we assume that the protein can only bind when it is folded. This results in three possible protein states (excluding the bound-and-unfolded state), yielding a fitness function

$$\mathcal{F}(E_f, E_b) = \frac{1 + f_0(e^{\beta E_b} + e^{\beta(E_f+E_b)})}{1 + e^{\beta E_b} + e^{\beta(E_f+E_b)}}, \tag{30}$$

where E_f, E_b are again defined by Eq. (29). In this model a protein may have high fitness even if $E_f > 0$, as long as the protein is stabilized by binding ($E_f + E_b < 0$). The methodology described here can be straightforwardly applied to this new fitness function. When E_f^0 is low, adaptive dynamics resemble the binding phase of the four-state model (Eq. (28)). However, when $E_f^0 > 0$, the dynamics enter a crossover regime, in which adaptation changes folding and binding energies simultaneously (Fig. 5). Unlike the four-state model (Eq. (28)), there is no folding-dominated phase.

In this chapter we have summarized many aspects of evolutionary trajectories on fitness landscapes, focusing especially on landscape topography and the statistical properties of adaptive first-passage trajectories. We have also described a general statistical physics-based methodology for studying stochastic paths on arbitrary landscapes and networks. This approach can be widely applied to first-passage problems in physics, chemistry, biology, and engineering, including protein folding, transport and search in complex media, stochastic phenotypes, and cell-type differentiation. Here we have emphasized its utility in exploring evolutionary problems, which can often be modeled as random walks on fitness landscapes and where the diversity and reproducibility of evolutionary paths is a central issue. The path-based methodology is well-suited for providing intuitive path statistics in problems whose complexity and high dimensionality make direct visualizations impossible.

References

1. P. Nelson, *Biological Physics: Energy, Information, Life*. W.H. Freeman and Company, New York, USA (2007).
2. T. E. Creighton, *Proteins: Structures and Molecular Properties*. W.H. Freeman and Company, New York, USA (1992).
3. A. V. Finkelstein and O. Ptitsyn, *Protein Physics: A Course of Lectures*. Academic Press, London, UK (2002).
4. C. Levinthal, Are there pathways for protein folding?, *J. Chim. Phys.* **65**, 44 (1968).
5. J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Funnels, pathways, and the energy landscape of protein folding: A synthesis, *Proteins: Str. Func. Genet.* **21**, 167–195 (1995).
6. K. A. Dill and J. L. MacCallum, The protein-folding problem, 50 years on, *Science*. **338**, 1042 (2012).

7. N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, Equation of state calculations by fast computing machines, *J. Chem. Phys.* **21**, 1087–1092 (1953).
8. D. Tolkunov and A. V. Morozov, Single temperature for Monte Carlo optimization on complex landscapes, *Phys. Rev. Lett.* **108**, 250602 (2012).
9. C. R. Darwin, *The Origin of Species*. J. Murray, London (1859).
10. J. Maynard Smith, Natural selection and the concept of a protein space, *Nature*. **225**, 563–564 (1970).
11. H. A. Orr, Fitness and its role in evolutionary genetics, *Nat. Rev. Genet.* **10**, 531–539 (2009).
12. I. G. Szendro, M. F. Schenk, J. Franke, J. Krug, and J. A. de Visser, Quantitative analyses of empirical fitness landscapes, *J. Stat. Mech.* p. P01005 (2013).
13. J. Gillespie, *Population Genetics: A Concise Guide*. The Johns Hopkins University Press, Baltimore, USA (2004).
14. S. Wright, The roles of mutation, inbreeding, crossbreeding and selection in evolution, *Proc. 6th Int. Cong. Genet.* **1**, 356–366 (1932).
15. S. J. Gould, *Wonderful Life: The Burgess Shale and the Nature of History*. W. W. Norton and Company, New York, USA (1990).
16. N. Champagnat, A microscopic interpretation for adaptive dynamics trait substitution sequence models, *Stochastic Process. Appl.* **116**, 1127–1160 (2006).
17. N. Champagnat, R. Ferrire, and S. Mlard, Unifying evolutionary dynamics: from individual stochastic processes to macroscopic models, *Theor. Popul. Biol.* **69**, 297–321 (2006).
18. J. F. Crow and M. Kimura, *An Introduction to Population Genetics Theory*. Harper and Row, New York (1970).
19. M. Kimura, *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK (1983).
20. M. Lässig, From biophysics to evolutionary genetics: statistical aspects of gene regulation, *BMC Bioinformatics*. **8**, S7 (2007).
21. M. Manhart, A. Haldane, and A. V. Morozov, A universal scaling law determines time reversibility and steady state of substitutions under selection, *Theor. Popul. Biol.* **82**, 66–76 (2012).
22. J. Wakeley, The limits of theoretical population genetics, *Genetics*. **169**, 1–7 (2005).
23. J. H. Gillespie, Molecular evolution over the mutational landscape, *Evolution*. **38**, 1116–1129 (1984).
24. M. M. Desai and D. S. Fisher, Beneficial mutation-selection balance and the effect of linkage on positive selection, *Genetics*. **176**, 1759–1798 (2007).
25. M. Carneiro and D. L. Hartl, Adaptive landscapes and protein evolution, *Proc. Natl. Acad. Sci. USA*. **107**, 1747–1751 (2010).
26. D. M. Weinreich, R. A. Watson, and L. Chao, Sign epistasis and genetic constraint on evolutionary trajectories, *Evolution*. **59**, 1165–1174 (2005).
27. F. J. Poelwijk, S. Tanase-Nicola, D. J. Kiviet, and S. J. Tans, Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes,

J. Theor. Biol. **272**, 141–144 (2011).
28. H. Flyvbjerg and B. Lautrup, Evolution in a rugged fitness landscape, *Phys. Rev. A.* **46**, 6714–6723 (1992).
29. A. Traulsen, Y. Iwasa, and M. A. Nowak, The fastest evolutionary trajectory, *J. Theor. Biol.* **249**, 617–623 (2007).
30. S. Kryazhimskiy, G. Tkačik, and J. B. Plotkin, The dynamics of adaptation on correlated fitness landscaps, *Proc. Natl. Acad. Sci. USA.* **106**, 18638–18643 (2009).
31. J. Franke, A. Klozer, J. A. de Visser, and J. Krug, Evolutionary accessibility of mutational pathways, *PLoS Comput. Biol.* **7**, e1002134 (2011).
32. M. Manhart and A. V. Morozov, Path-based approach to random walks on networks characterizes how proteins evolve new functions, *Phys. Rev. Lett.* **111**, 088102 (2013).
33. A. E. Lobkovsky and E. V. Koonin, Replaying the tape of life: quantification of the predictability of evolution, *Front. Gene.* **3**, 246 (2012).
34. D. M. Weinreich, N. F. Delaney, M. A. DePristo, and D. L. Hartl, Darwinian evolution can follow only very few mutational paths to fitter proteins, *Science.* **312**, 111–114 (2006).
35. F. J. Poelwijk, D. J. Kiviet, D. M. Weinreich, and S. J. Tans, Empirical fitness landscapes reveal accessible evolutionary paths, *Nature.* **445**, 383–386 (2007).
36. A. E. Lobkovsky, Y. I. Wolf, and E. V. Koonin, Predictability of evolutionary trajectories in fitness landscapes, *PLoS Comput. Biol.* **7**, e1002302 (2011).
37. S. A. Kauffman and E. D. Weinberger, The NK model of rugged fitness landscapes and its application to maturation of the immune response, *J. Theor. Biol.* **141**, 211–245 (1989).
38. S. Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution.* Oxford University Press, New York (1993).
39. T. Aita, H. Uchiyama, T. Inaoka, M. Nakajima, T. Kokubo, and Y. Husimi, Analysis of a local fitness landscape with a model of the rough Mt. Fuji-type landscape: application to prolyl endopeptidase and thermolysin, *Biopolymers.* **54**, 64–79 (2000).
40. J. F. C. Kingman, A simple model for the balance between selection and mutation, *J. Appl. Probab.* **15**, 1–12 (1978).
41. L. Altenberg. NK fitness landscapes. In eds. T. Bäck, D. Fogel, and Z. Michalewicz, *Handbook of Evolutionary Computation*, pp. B2.7:5–B2.7:10. IOP Publishing Ltd and Oxford University Press (1997).
42. D. R. Rokyta, C. J. Beisel, and P. Joyce, Properties of adaptive walks on uncorrelated landscapes under strong selection and weak mutation, *J. Theor. Biol.* **243**, 114–120 (2006).
43. E. van Nimwegen, J. P. Crutchfield, and M. Huynen, Neutral evolution of mutational robustness, *Proc. Natl. Acad. Sci. USA.* **96**, 9716–9720 (1999).
44. D. Stauffer and A. Aharony, *Introduction to Percolation Theory.* Taylor and Francis, London (1994).
45. F. Mammano, V. Trouplin, V. Zennou, and F. Clavel, Retracing the evo-

lutionary pathways of human immunodeficiency virus type 1 resistance to protease inhibitors: virus fitness in the absence and in the presence of drug, *J. Virol.* **74**, 8524–8531 (2000).
46. A. I. Khan, D. M. Dinh, D. Schneider, R. E. Lenski, and T. F. Cooper, Negative epistasis between beneficial mutations in an evolving bacterial population, *Science*. **332**, 1193–1196 (2011).
47. H.-H. Chou, H.-C. Chiu, N. F. Delaney, D. Segrè, and C. J. Marx, Diminishing returns epistasis among beneficial mutations decelerates adaptation, *Science*. **332**, 1190–1192 (2011).
48. P. Hänggi, P. Talkner, and M. Borkovec, Reaction rate theory: fifty years after kramers, *Rev. Mod. Phys.* **62**, 251–341 (1990).
49. P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, Transition path sampling: Throwing ropes over rough mountain passes, in the dark, *Ann. Rev. Phys. Chem.* **53**, 291–318 (2002).
50. C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler, Transition path sampling and the calculation of rate constants, *J. Chem. Phys.* **108**, 1964–1977 (1998).
51. C. Dellago, P. G. Bolhuis, and P. L. Geissler, Transition path sampling, *Adv. Chem. Phys.* **123**, 1 (2003).
52. G. Hummer, From transition paths to transition states and rate coefficients, *J. Chem. Phys.* **120**, 516–523 (2004).
53. B. Harland and S. X. Sun, Path ensembles and path sampling in nonequilibrium stochastic systems, *J. Chem. Phys.* **127**, 104103 (2007).
54. T. More, A. Walczak, and F. Zamponi, Transition path sampling algorithm for discrete many-body systems, *Phys. Rev. E.* **85**, 036710 (2012).
55. D. M. Robinson, D. T. Jones, H. Kishino, N. Goldman, and J. K. Thorne, Protein evolution with dependence among codons due to tertiary structure, *Mol. Biol. Evol.* **20**, 1692–1704 (2003).
56. N. Rodrigue, N. Lartillot, D. Bryant, and H. Philippe, Site interdependence attributed to tertiary structure in amino acid sequence evolution, *Gene*. **347**, 207–217 (2005).
57. N. Rodrigue, H. Philippe, and N. Lartillot, Assessing site-interdependent phylogenetic models of sequence evolution, *Mol. Biol. Evol.* **23**, 1762–1775 (2006).
58. S. C. Choi, A. Hobolth, D. M. Robinson, H. Kishino, and J. L. Thorne, Quantifying the impact of protein tertiary structure on molecular evolution, *Mol. Biol. Evol.* **24**, 1769–1782 (2007).
59. N. Rodrigue, C. L. Kleinman, H. Philippe, and N. Lartillot, Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons, *Mol. Biol. Evol.* **26**, 1663–1676 (2009).
60. W. E and E. Vanden-Eijnden, Toward a theory of transition paths, *J. Stat. Phys.* **123**, 503–523 (2006).
61. P. Metzner, C. Schütte, and E. Vanden-Eijnden, Illustration of transition path theory on a collection of simple examples, *J. Chem. Phys.* **125**, 084110 (2006).
62. P. Metzner, C. Schütte, and E. Vanden-Eijnden, Transition path theory for

Markov jump processes, *Multiscale Model. Simul.* **7**, 1192–1219 (2009).
63. F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations, *Proc. Natl. Acad. Sci. USA*. **106**, 19011–19016 (2009).
64. W. E and E. Vanden-Eijnden, Transition-path theory and path-finding algorithms for the study of rare events, *Ann. Rev. Phys. Chem.* **61**, 391–420 (2010).
65. G. H. Weiss, *Aspects and Applications of the Random Walk*. North Holland, Amsterdam (1994).
66. J. Klafter and R. Silbey, Derivation of the continuous-time random-walk equation, *Phys. Rev. Lett.* **44**, 55 (1980).
67. S. Redner, *A Guide to First-Passage Processes*. Cambridge University Press, Cambridge (2001).
68. C. Maes, K. Netočný, and B. Wynants, Dynamical fluctuations for semi-markov processes, *J. Phys. A: Math. Theor.* **42**, 365002 (2009).
69. O. Flomenbom and J. Klafter, Closed-form solutions for continuous time random walks on finite chains, *Phys. Rev. Lett.* **95**, 098105 (2005).
70. S. X. Sun, Path summation formulation of the master equation, *Phys. Rev. Lett.* **96**, 210602 (2006).
71. I. Majid, D. ben-Avraham, S. Havlin, and H. E. Stanley, Exact-enumeration approach to random walks on percolation clusters in two dimensions, *Phys. Rev. B.* **30**, 1626–1628 (1984).
72. E. M. Bollt and D. ben-Avraham, What is special about diffusion on scale-free nets?, *New J. Phys.* **7**, 26–47 (2005).
73. S. Condamin, O. Bénichou, V. Tejedor, R. Voituriez, and J. Klafter, First-passage times in complex scale-invariant media, *Nature*. **450**, 77–80 (2007).
74. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, second edn. Cambridge, Cambridge (1992).
75. J. D. Bloom, J. J. Silberg, C. O. Wilke, D. A. Drummond, C. Adami, and F. H. Arnold, Thermodynamic prediction of protein neutrality, *Proc. Natl. Acad. Sci. USA*. **102**, 606–611 (2005).
76. M. A. DePristo, D. M. Weinreich, and D. L. Hartl, Missense meanderings in sequence space: a biophysical view of protein evolution, *Nat. Rev. Genet.* **6**, 678–687 (2005).
77. J. D. Bloom, S. T. Labthavikul, C. R. Otey, and F. H. Arnold, Protein stability promotes evolvability, *Proc. Natl. Acad. Sci. USA*. **103**, 5869–5874 (2006).
78. K. B. Zeldovich, P. Chen, and E. I. Shakhnovich, Protein stability imposes limits on organism complexity and speed of molecular evolution, *Proc. Natl. Acad. Sci. USA*. **104**, 16152–16157 (2007).
79. J. D. Bloom, Z. Lu, D. Chen, A. Raval, O. S. Venturelli, and F. H. Arnold, Evolution favors protein mutational robustness in sufficiently large populations, *BMC Biology*. **5**, 29 (2007).
80. J. D. Bloom, P. A. Romero, Z. Lu, and F. H. Arnold, Neutral genetic

drift can alter promiscuous protein functions, potentially aiding functional evolution, *Biology Direct*. **2**, 17 (2007).
81. J. D. Bloom, A. Raval, and C. O. Wilke, Thermodynamics of neutral protein evolution, *Genetics*. **175**, 255–266 (2007).
82. S. Bershtein, K. Goldin, and D. S. Tawfik, Intense neutral drifts yield robust and evolvable consensus proteins, *J. Mol. Biol.* **379**, 1029–1044 (2008).
83. A. M. Sengupta, M. Djordjevic, and B. I. Shraiman, Specificity and robustness in transcription control networks, *Proc. Natl. Acad. Sci. USA*. **99**, 2072–2077 (2002).
84. U. Gerland and T. Hwa, On the selection and evolution of regulatory DNA motifs, *J. Mol. Evol.* **55**, 386–400 (2002).
85. J. Berg and M. Lässig, Stochastic evolution of transcription factor binding sites, *Biophysics (Moscow)*. **48**, S36–S44 (2003).
86. J. Berg, S. Willmann, and M. Lässig, Adaptive evolution of transcription factor binding sites, *BMC Evolutionary Biology*. **4**, 42 (2004).
87. V. Mustonen, J. Kinney, C. G. Callan, and M. Lässig, Energy-dependent fitness: A quantitative model for the evolution of yeast transcription factor binding sites, *Proc. Natl. Acad. Sci. USA*. **105**, 12376–12381 (2008).
88. S. Mayer, S. Rüdiger, H. C. Ang, A. C. Joerger, and A. R. Fersht, Correlation of levels of folded recombinant p53 in *escherichia coli* with thermodynamic stability *in vitro*, *J. Mol. Biol.* **372**, 268–276 (2007).
89. T. Clackson and J. A. Wells, A hot spot of binding energy in a hormone-receptor interface, *Science*. **267**, 383–386 (1995).
90. L. Serrano, A. G. Day, and A. R. Fersht, Step-wise mutation of barnase to binase. a procedure for engineering increased stability of proteins and an experimental analysis of the evolution of protein stability, *J. Mol. Biol.* **233**, 305–312 (1993).
91. N. Tokuriki, F. Stricher, J. Schymkowitz, L. Serrano, and D. S. Tawfik, The stability effects of protein mutations appear to be universally distributed, *J. Mol. Biol.* **369**, 1318–1332 (2007).
92. A. A. Bogan and K. S. Thorn, Anatomy of hot spots in protein interfaces, *J. Mol. Biol.* **280**, 1–9 (1998).
93. K. S. Thorn and A. A. Bogan, Asedb: a database of alanine mutations and their effects on the free energy of binding in protein interactions, *Bioinformatics*. **17**, 284–285 (2001).
94. J. D. Bloom and F. H. Arnold, In the light of directed evolution: Pathways of adaptive protein evolution, *Proc. Natl. Acad. Sci. USA*. **106**, 9995–10000 (2009).
95. N. Tokuriki and D. S. Tawfik, Stability effects of mutations and protein evolvability, *Curr. Opin. Struct. Biol.* **19**, 596–604 (2009).
96. X. Wang, G. Minasov, and B. K. Shoichet, Evolution of an antibiotic resistance enzyme constrained by stability and activity tradeoffs, *J. Mol. Biol.* **320**, 85–95 (2002).
97. N. Tokuriki, F. Stricher, L. Serrano, and D. S. Tawfik, How protein stability and new functions trade off, *PLoS Comput. Biol.* **4**, e1000002 (2008).
98. L. D. Bogarad and M. W. Deem, A hierarchical approach to protein molec-

ular evolution, *Proc. Natl. Acad. Sci. USA*. **96**, 2591–2595 (1999).
99. S. L. Rutherford, Between genotype and phenotype: protein chaperones and evolvability, *Nat. Rev. Genet.* **4**, 263–274 (2003).
100. C. J. Brown, A. K. Johnson, A. K. Dunker, and G. W. Daughdrill, Evolution and disorder, *Curr. Opin. Struct. Biol.* **21**, 441–446 (2011).