genes

MDPI

Article

How Often Do Protein Genes Navigate Valleys of Low Fitness?

Erik D. Nelson * and Nick V. Grishin

Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, 6001 Forest Park Blvd., Room ND10.124, Dallas, TX 75235-9050, USA; grishin@chop.swmed.edu
* Correspondence: nelsonerikd@gmail.com

Received: 7 March 2019; Accepted: 2 April 2019; Published: 8 April 2019

Abstract: To escape from local fitness peaks, a population must navigate across valleys of low fitness. How these transitions occur, and what role they play in adaptation, have been subjects of active interest in evolutionary genetics for almost a century. However, to our knowledge, this problem has never been addressed directly by considering the evolution of a gene, or group of genes, as a whole, including the complex effects of fitness interactions among multiple loci. Here, we use a precise model of protein fitness to compute the probability $P(s, \Delta t)$ that an allele, randomly sampled from a population at time t, has crossed a fitness valley of depth s during an interval $[t - \Delta t, t]$ in the immediate past. We study populations of model genes evolving under equilibrium conditions consistent with those in mammalian mitochondria. From this data, we estimate that genes encoding small protein motifs navigate fitness valleys of depth $2Ns \gtrsim 30$ with probability $P \gtrsim 0.1$ on a time scale of human evolution, where N is the (mitochondrial) effective population size. The results are consistent with recent findings for Watson–Crick switching in mammalian mitochondrial tRNA molecules.

Keywords: molecular evolution; epistasis; fitness valley crossing; thermodynamic stability

1. Introduction

The effect of a mutation on the fitness of an organism usually depends on the genetic background, or context in which it occurs, a phenomenon known as epistasis [1,2]. Because of this, the fitness landscape of a gene, a group of genes, or an organism will contain many isolated peaks and valleys [3,4], resembling the energy landscape of a physical system such as a glass. Under selection pressure, a population tends to evolve along a path of steepest ascent in fitness until it arrives in the neighborhood of a local fitness peak; In order to escape a sub–optimal fitness peak, the population, or some part of the population, must traverse across a valley of lower fitness. How such transitions occur [1,4–8], and how they relate to adaptation [9] have remained subjects of active interest in evolutionary genetics for almost a century.

The most basic example of valley crossing is realized in the compensatory interaction of individually deleterious mutations at two genetic loci—for instance, as might result from the physical interaction between amino acids in a protein, or a pair of nucleotides in an RNA molecule [5]. The archetypal model of this situation consists of a pair of diallelic loci with initial and final states AB and A′B′ respectively; Mutations to A′ and B′ incur a fitness cost s relative to AB when introduced individually, but are neutral when introduced jointly. Kimura was the first to study this problem using the diffusion approach [5,10], and he found that deep fitness valleys could be crossed on a relatively short time scale if mutation rates are sufficiently large—specifically, when $2N\mu = 1$ where N is the population size and μ is the mutation rate per gene per generation. In this case, fitness valleys are navigated by a process known as stochastic tunneling [11], in which a small fraction of genes accumulate in an intermediate state, are compensated by a second mutation, and ultimately proceed to fixation—the intermediate acting as a kind of stepping stone [12]. The situation studied by Kimura closely resembles the process of Watson–Crick switching between favorably paired nucleotides in RNA stem sites, and in particular, switching in mammalian mitochondrial (mt) tRNA molecules where stochastic tunneling is significant. Meer et al. investigated this problem somewhat recently [13], and, using Kimura's model, they found that mammalian mt tRNA switches may navigate valleys of depth even as large as $2Ns \simeq 50$ (here, we assume that, for equal numbers of males and females, the effective population size for mitochondrial genes is one fourth the effective population size for nuclear genes [14,15]). In support of this result, Meer et al. obtain essentially the same estimate for $2Ns$ from the frequency (p) of disrupted Watson–Crick pairs using the relation $p = \mu/s$ for mutation–selection balance [16]. To put this number into context, it is at least ten times larger than would be expected if the same model had evolved by sequential fixation of deleterious and compensatory mutations (i.e., as would be expected when $\mu N \ll 1$ [17]).

While these estimates may be accurate, it is difficult to reconcile the evolutionary dynamics of folded biomolecules with two–locus models. Naturally evolving genes encoding proteins and RNA molecules are always faced with a complex spectrum of possible routes on their fitness landscapes, and it is these spectra that ultimately determine the rate for crossing valleys of a given depth. Even for tRNA molecules, compensation of disrupted Watson–Crick pairs seems to occur more often through complex, indirect mechanisms than through direct compensation to restore Watson–Crick pairing [18]. Proteins are more connected objects than RNA molecules (i.e., with more opportunities for epistatic interactions between loci), and the greater complexity of protein sequences is almost certain to present a more complex spectrum of possible routes to a protein gene in which valleys (ravines, etc.) are entered and exited in multiple steps (Figure A1).

Are the large effects predicted by Meer et al. common in biomolecular evolution? To our knowledge, this kind of question has never been asked directly, by considering the problem of valley crossing for a protein or RNA molecule as a whole, including the complex effects of fitness interactions between multiple loci. Here, we simulate the evolution of a small protein motif using an exact fitness model that is simple enough to allow for adequate sampling of valley crossing statistics. We evolve populations of model genes by haploid Wright–Fisher dynamics across a range of mutation rates spanning the sequential fixation ($\mu N \ll 1$) and stochastic tunneling ($\mu N \geq 1$) regimes, and we record the mutational paths of all alleles in our populations. Using this data, we compute the probability $P(s, \Delta t)$ that an allele, randomly sampled from a population at time t, has crossed a fitness valley of depth s during a time interval $[t - \Delta t, t]$ in the immediate past. Surprisingly, we find that, even on the time scale of human evolution, genes encoding small protein motifs evolving under conditions consistent with mammalian mitochondria already navigate fitness valleys of depth $2Ns \gtrsim 30$ with probability $P \gtrsim 0.1$, in rough agreement with the estimate for Watson–Crick switching in mt tRNAs provided by Meer et al.

2. Methods

Epistatic effects play an essential part in protein evolution [19–24], and because these effects depend on the relative probabilities of conformations in protein ensembles [25–27], it is important to select a model in which the salient properties of protein ensembles are retained as much as possible. Ultimately, we found that we could obtain sufficient data for valley crossing statistics in a reasonable period of time using small lattice proteins. Lattice models have been used extensively in studies of protein folding and evolution, and the model we employ here is similar to one recently used to explore the effects of epistasis on the predictability of protein evolutionary pathways [25].

Below, we evolve lattice proteins under equilibrium conditions to maintain marginal stability in a specific folded (native) conformation. The stability of a protein is measured, as usual, by the free energy difference between the native conformation and the rest of the conformational ensemble,

$$\Delta G_N = E_N + \ln\left[-e^{-E_N} + \sum_{\gamma} e^{-E_\gamma}\right], \tag{1}$$

where E_γ is the energy of conformation γ, the subscript N denotes the native conformation, and factors of temperature are absorbed into the definition of energy; The energy of a conformation is determined from its amino acid contacts by empirical amino acid contact potentials [28] (as a result, energies are defined in units of $RT \simeq 0.6$ kcal/mol).

We assume that mis–folded proteins are non–functional, and otherwise toxic to an organism [29–31]. In this case, protein fitness can be defined by the probability of finding an individual protein folded in its native conformation [25,32],

$$P_N = 1 / \left[1 + e^{\Delta G_N}\right]. \tag{2}$$

However, since most naturally occurring proteins are only marginally (as opposed to maximally) stable [24], we decided to model fitness using a logistic function

$$w = 1 / \left[1 + e^{-k(P_N - 1/2)}\right] \tag{3}$$

where $k = 15$ (Supplementary Figure S1). Under this condition, evolved genes in our simulations typically encode proteins with $P_N > 0.75$ or, equivalently, $\Delta G_N < -1$ [33].

We evolve populations of protein genes by plain Wright–Fisher dynamics, with discrete generations, fixed population size, and no recombination [34]. In each generation, a Poisson random number of nucleotide sites with mean μN are selected at random; The sites are subjected to random mutation, the fitness values of mutant alleles are computed, and N offspring are selected from the population to form the next generation. The probability that an allele i survives to the next generation is $p'_i = w_i p_i \,/\, \sum_j w_j p_j$, where p_i is the frequency of allele i in the current generation [35].

In the absence of recombination, each allele in a population has a unique mutational history extending back to the origin of a simulation. To describe the statistics of valley crossing, we record the histories of all alleles, and we compute $P(s, \Delta t)$, the probability that an allele, randomly sampled from the population at time t, has crossed a fitness valley of depth s during the time interval $[t - \Delta t, t]$, where time is measured in generations. Depending on its length, an interval $[t - \Delta t, t]$ along the fitness history of an allele may contain several (perhaps nested) valleys of varying depth (Figure A1). However, rather than attempt to record each valley as an individual event, we simply define s as the maximum valley depth traversed along an interval (see Appendix A). A peculiar feature of this approach is that, for small values of s, we can choose an interval length Δt long enough that $P(s, \Delta t)$ is a decreasing function of Δt (i.e., since larger values of s are more likely to occur on longer time scales). However, below we will be concerned mainly with large values of s and time scales Δt that are far from this sort of turnover region.

3. Results

To obtain data for $P(s, \Delta t)$ in a reasonable period of time, we limit our study to chains with at most 16 amino acids (802,075 conformations unrelated by symmetry [36]). The fitness landscapes of longer chains will clearly differ, however, we expect that within reason, results for somewhat longer chains (e.g., 32 amino acids) will be similar, given proper adjustments to the mutation rate per gene μ (see below).

We simulate protein evolution for different chain lengths ($L \leq 16$), native folds (Supplementary Materials Figures S2 and S5), population sizes ($N \leq 10^3$), and mutation rates ($\mu N \leq 2$). In each situation, we conduct replicate simulations in parallel on multiple processors of a high performance computer [37]. Each processor begins with a monomorphic population constructed from N copies of a gene encoding a randomly selected amino acid sequence. Each population is then equilibrated until an allele reaches fixation with $P_N > 0.75$. After this point, alleles are sampled at random from a population every $64N$ generations and their histories are recorded. For the most computationally intensive problems (i.e., for the largest chain lengths and mutation rates), we are able to generate 10^5 samples in about ten days using 128 processors. For chains with 12 amino acids (15,037 conformations), samples can be obtained much more rapidly, and results for chains of 12 and 16 amino acids are actually very similar (computer code and sample data are available from the authors on request).

Since effective population size varies substantially across mammalian species, it is important to ask whether simulations conducted for a particular population size can be used to estimate $P(s, \Delta t)$ for larger (or smaller) populations evolving at the same overall rate μN, as expected from diffusion theory [38]. To answer this question, we compared the scaled distributions $P(Ns > x, \Delta t)$ for population sizes $N = 100$, 200, 500 and 1000. For a fixed mutation rate μN, we find that plots of $P(Ns > x, \Delta t)$ roughly collapse to the same curve. As a result, we can estimate $P(s, \Delta t)$ for realistic populations using a much smaller population size, which greatly reduces the amount of time spent on the simulations. The reason for this is fairly simple–the local structure of the fitness landscape, as measured by e.g., the distribution of fitness effects or the probability of compensatory neutral mutations, scales in a similar way with population size; As population size increases, the landscape in the neighborhood of an evolved sequence becomes less rugged in proportion to N.

Results of this exploration are described in Figures 1 and 2. In Figure 1, we plot the distributions of beneficial and deleterious fitness effects, $P(\Delta w > x)$ and $P(-\Delta w > x)$, respectively. Each pair of plots corresponds to a simulation for one of the population sizes listed above (the width of a plot increases with decreasing N). The results describe proteins with 12 amino acids folding to the native conformation in Supplementary Materials Figure S2, and the overall mutation rate in each simulation is $\mu N = 1$.

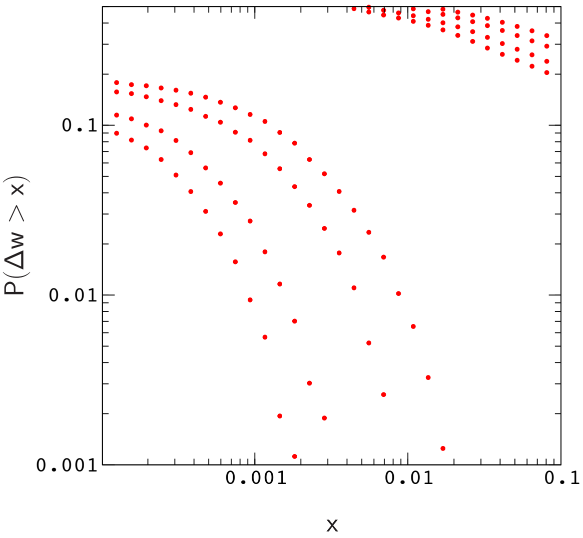

Figure 1. Distribution of beneficial fitness effects, $P(\Delta w > x)$. The distribution of deleterious fitness effects, $P(-\Delta w > x)$, is partially included in the upper right corner of the figure for reference. The plots are generated by randomly mutating evolved sequences sampled from simulations with $N = 100$, 200, 500, and 1000 (the width of a plot increases with decreasing N). If the data for Δw in each plot is rescaled by the appropriate factor of N, the distributions $P(N\Delta w > x)$ roughly collapse to a single curve. For $N = 1000$, about 47 percent of the mutations are strongly deleterious ($N\Delta w < -5$), about 28 percent are nearly neutral ($-1 < N\Delta w < 1$), and about 0.7 percent are beneficial ($N\Delta w > 1$), consistent with results obtained by Tamuri et al. [39] for mammalian mitochondrial proteins (Tamuri et al. use logarithmic fitness differences in their work, however, this distinction can be neglected when $\ln(1 + x) \simeq x$ [40]).

To generate data for Figure 1, we sampled the landscape around evolved genes using a simple procedure that mimics error–prone polymerase chain reactions [41]; The procedure begins from a large number of copies of an evolved gene. A Poisson random number of random mutations are applied to each copy, and the results are sorted by the number of mis–sense mutations (i.e., neglecting back mutations). For each simulation, we randomly selected 10^2 evolved sequences. Each evolved sequence was then used to generate 10^4 random single amino acid mutants, for a total of 10^6 mutants per plot.

As is evident by closer inspection of Figure 1, the probability of a beneficial (or deleterious) mutation with effect $\Delta w > x$ decreases almost linearly with increasing population size; The scaled distributions $P(N\Delta w > x)$ for different population sizes roughly collapse onto a single curve. A similar result is obtained for the distribution of compensatory neutral double mutants $P(s > x)$ (Supplementary Materials Figure S3). The collapse is shown explicitly for the valley crossing probability $P(s > x, \Delta t)$ in Figure 2.

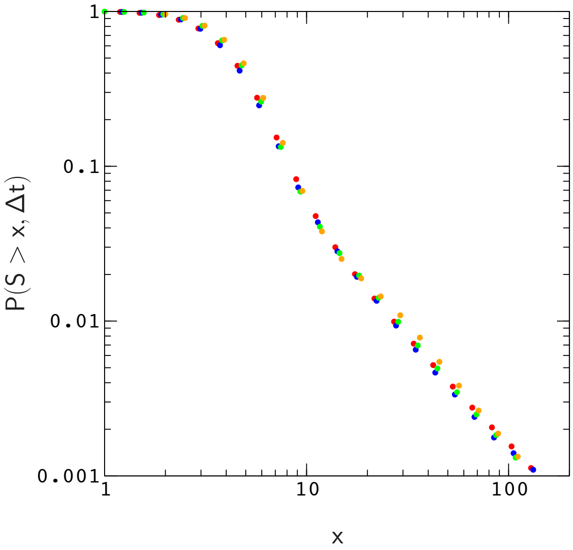

Figure 2. Probability, $P(S > x, \Delta t)$, that an allele, randomly sampled at time t, has crossed a valley of depth $S = 2Ns$ in the interval $[t - \Delta t, t]$ for $\Delta t = 128N$ and $\mu N = 1$. The data describe the same model as in Figure 1. Plots for population sizes $N = 100$, 200, 500, and 1000 are colored red, blue, green, and orange, respectively.

Given this result, we now restrict our attention to populations with $N = 200$ and proteins with 16 amino acids. To compare our results to those of Meer et al., we require the site mutation rate in our model to agree with the pedigree rate for the control region in human mitochondria used in their estimate for mammalian mt tRNA molecules—about 1×10^{-6} per site per year. Assuming a typical length of about 80 nucleotides for tRNA molecules, a generation time of 20 years, and a (mitochondrial) effective population size of $N = 2500$, we arrive at an overall mutation rate of $\mu N \simeq 4$ for human mt tRNA genes. To obtain the same site mutation rate for protein genes with 48 nucleotides (16 amino acids), we need an overall mutation rate of about $\mu N \simeq 2$.

We plot $P(S > x, \Delta t)$ versus Δt for this situation in Figure 3, where $S = 2Ns$. The range of the plot, $\Delta t \leq 128N$, roughly corresponds to the time scale of human evolution (about six million years). Over this time scale, $P(S > x, \Delta)$ is roughly linear in Δt for $x \gtrsim 10$. The time scale for mammalian evolution is much longer (on the order of tens of millions of years), however, on human time scales, the frequencies of large events in our model are already approaching the ten percent levels observed for Watson–Crick switching events in mammalian mt tRNA phylogenies [13]. For example, the probability of sampling an allele that has crossed a fitness valley of depth $S > 9.1$ for $\Delta t = 128N$ is about 0.36, the probability for $S > 17.8$ is about 0.07, and the probability for $S > 27.8$ is already about 0.03.

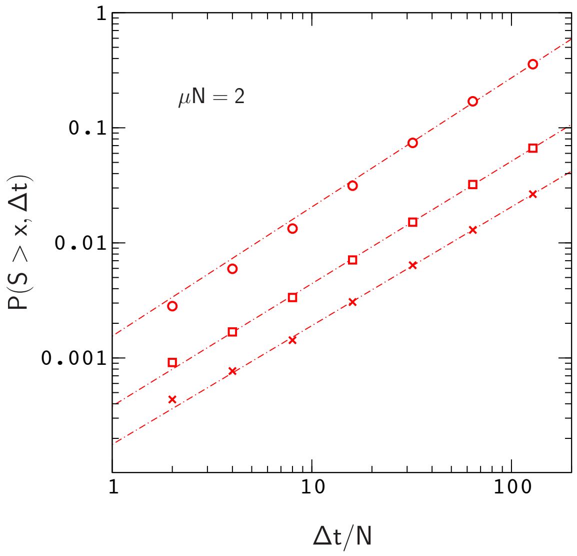

Figure 3. $P(S > x, \Delta t)$ versus Δt for $x \simeq 9.1$ (circles), 17.8 (squares) and 27.8 (crosses). The results describe proteins with 16 amino acids folding to the native conformation in Figure S5. Each data point is determined from over 10^5 allele histories. The range of the plot, $\Delta t \leq 128N$, corresponds to the time scale of human evolution (about six million years). The dashed lines (power law fits to the data) are very close to linear, increasingly so for larger values of x (see Appendix A for more details).

Clearly, these numbers will continue to increase for larger values of Δt and larger mutation rates μN (Figure 4). In addition, $P(S > x, \Delta)$ will also increase with chain length since, for a constant site mutation rate, the mutation rate per gene μ is proportional to chain length. If we assume that the fitness landscapes of proteins with 16 amino acids can be used to represent the landscapes of larger chains, then e.g., doubling the mutation rate per gene will have the same effect as doubling the chain length to obtain a protein encoded by the same amount of genetic material as a tRNA molecule or small protein motif [42]. This is not an unreasonable assumption, since, as we have noted earlier (see the captions to Figure 1 and Figure S3), the local structures of fitness landscapes in the model, as measured by the scaled distribution functions $P(N\Delta w > x)$ and $P(Ns > x)$, are already similar to those inferred from real proteins with much longer sequences. In this case, extrapolating from the data in Figure 4, we find that the probability of sampling an allele that has crossed a valley of depth $S > 27.8$ over a time interval $\Delta t = 128N$ increases to about $P \simeq 0.1$. Thus, even neglecting the anticipated increase in $P(S > x, \Delta)$ for $\Delta t > 128N$, the results for small protein motifs are already consistent with those of Meer et al.

As a final note, it is important to remark that the deepest valleys navigated by alleles in our simulations actually correspond to events in which a deleterious mutation is, to a major extent, compensated by a mutation back to a similar amino acid type at the same site (see Supplementary Materials Figure S6). It is also worth noting that local fitness peaks sampled at various points in our simulations (by steepest ascent in fitness, starting from a randomly sampled sequence) are often separated by deep fitness valleys, or ravines.

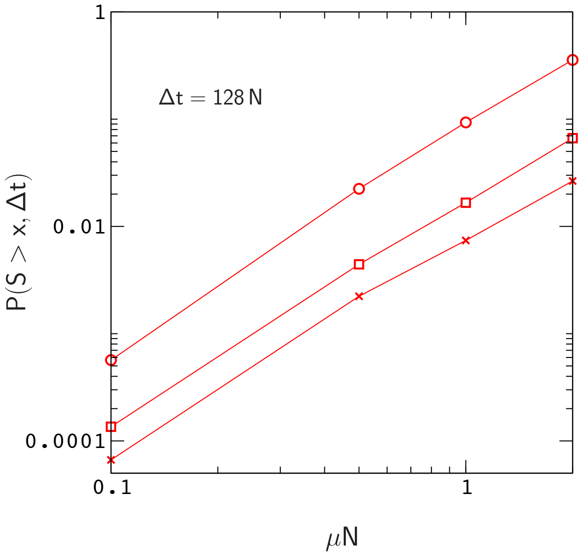

Figure 4. $P(S > x, \Delta t)$ versus μN for $x \simeq 9.1$ (circles), 17.8 (squares) and 27.8 (crosses). The results describe the same simulation data as Figure 3.

4. Discussion

Does the model provide an accurate picture of fitness valley crossing for small protein motifs? This question is very difficult to answer due to the extreme complexity of protein fitness landscapes, and the unknown effects of varying host genetic backgrounds experienced by protein genes. However, from basic principles, it appears that the model is roughly accurate: Local features of protein fitness landscapes, such as the distribution of fitness effects, can be inferred from protein sequences by fitting a population dynamics model to branches of a phylogenetic tree [39] such that background effects are accounted for by allowing the parameters of the model to vary among branches. Tamuri et al. have used this type of approach to estimate the distribution of fitness effects for mammalian mitochondrial proteins [39], and our results for $P(N\Delta w > x)$ are in good agreement with their data (Figure 1). Given the similar nature of fitness conditions in each system (i.e., that proteins are also polymers required to fold into a specific shape in order to function), it is seems reasonable to expect that the topographies of model fitness landscapes resemble those of small protein motifs.

Longer chains with more developed core structures, and more restrictive fitness conditions such as binding to proteins within a larger domain, may lead to qualitatively different results due to the potential for larger compensatory effects [27], and the suppression of substitution rates in the core and binding interface regions. In addition, much larger mutation rates can occur in micro–organisms such as viruses [43]. and our results suggest that, under these conditions, fitness valley crossing will be more pronounced. RNA viruses such as HIV–1 are also subject to high rates of recombination [44,45], which may interfere with valley crossing [5,6]. Because the computational cost of our simulations increases in proportion to the number of mutations (i.e., the number of fitness calculations), a full study of the model for virus proteins may be challenging. However, we hope to address these problems in future work.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-4425/10/4/283/s1. Figure S1: Plot of the logistic fitness function defined in Equation (3); Figure S2: Native fold studied in Figures 1 and 2; Figure S3 : Plot of $P(s > x)$ for the native fold studied in Figures 1 and 2; Figure S4: Plot of $P(S > x)$ for an inferred Potts model of cytochrome c oxidase subunit 2; Figure S5: Native fold studied in Figures 3 and 4; Figure S6: Plot of $P(S > x, \Delta t)$ for the subset of deleterious mutations compensated by a mutation back to a similar amino acid type at the same site.

Author Contributions: Conceptualization, E.D.N.; Funding acquisition, N.V.G.; Investigation, E.D.N.; Software, E.D.N.; Supervision, N.V.G.

Funding: This research was funded in part by a grant from the National Institutes of Health GM127390 to NVG.

Acknowledgments: It is a pleasure to thank Daniel Moser for help with the computations, and two anonymous reviewers for helpful suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

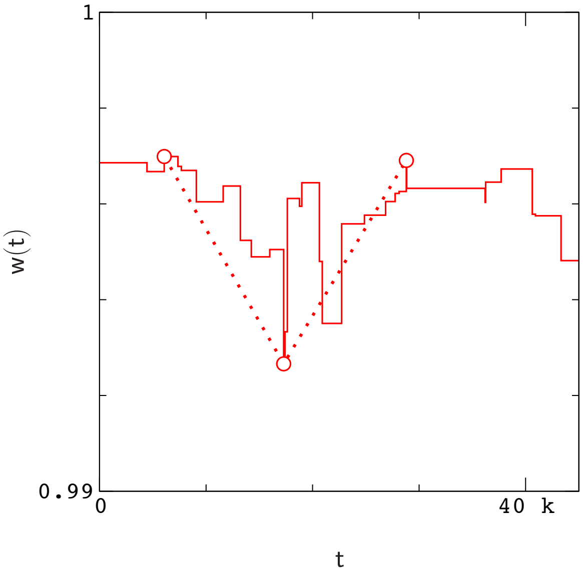

Figure A1. Fitness history of an allele between adjacent fixation events (solid line) sampled from a simulation with $N = 10^3$ and $\mu N = 1$. Circles and dashed lines indicate the configuration of points which maximize the valley depth $s = \mathbf{min}\{w_i - w_j, w_k - w_j\}$.

To compute the maximum valley depth traversed along an interval (Figure A1), we consider all possible placements of three points, w_i, w_j and w_k, where $i < j < k$ denote the occurrence times of mutations. For a given placement of points, valley depth s is defined as the smaller of the two fitness differences $w_i - w_j$ or $w_k - w_j$. The maximum valley depth can then be expressed as,

$$\mathbf{max}\, s = \mathbf{max}_{i<j<k} \left[\mathbf{min}\{w_i - w_j, w_k - w_j\} \right] . \tag{A1}$$

To avoid complicating our expressions, we use the plain symbols s and $S = 2Ns$ to denote maximum valley depth in $P(s > x, \Delta t)$ and $P(S > x, \Delta t)$.

The structure of $P(S > x, \Delta t)$ can be explained roughly as follows: As we noted earlier, an interval $[t - \Delta t, t]$ along a fitness trace may contain several distinct valleys of varying depth. Valleys of large depth are rare on the time scale of human evolution and short lived. As a result, when x is large, doubling Δt doubles the probability that a valley with depth greater than x will occur within an interval $[t - \Delta t, t]$, and $P(S > x, \Delta t)$ increases linearly with Δt. This will be true as long as Δt is not too large or too small; For large enough Δt, $P(S > x, \Delta t)$ will begin to saturate (i.e., $P \to 1$), at which point the slope of the curve, $\partial P(S > x, \Delta t)/\partial \Delta t$, tends to zero, and linearity is lost. Conversely, for small enough Δt, the typical duration of an event (valley of depth greater than x) will begin to exceed Δt, and again $\partial P(S > x, \Delta t)/\partial \Delta t$ will begin to change. This change will depend both on x and the topography of valleys with depth greater than x, however, we have not explored this issue in detail.

References

1. Phillips, P.C. Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* **2008**, *9*, 855–867. [CrossRef]
2. Starr, T.N.; Thornton, J.W. Epistasis in protein evolution. *Protein Sci.* **2016**, *25*, 1204–1218. [CrossRef]
3. Wright, S. The roles of mutation, inbreeding, cross–breeding and selection in evolution. In *Proceedings of the Sixth International Congress of Genetics*; Jones, D.F., Ed.; Brooklyn Botanic Garden: Menasha, WI, USA, 1932; pp. 356–366.
4. Johnson, N. Sewall Wright and the Development of Shifting Balance Theory. *Nat. Educ.* **2008**, *1*, 52.
5. Kimura, M. The role of compensatory neutral mutations in molecular evolution. *J. Genet.* **1985**, *64*, 7–19. [CrossRef]
6. Weinreich, D.M.; Chao, L. Rapid evolutionary escape by large populations from local fitness peaks is likely in nature. *Evolution* **2005**, *59*, 1175–1182. [CrossRef]
7. Van Nimwegen, E.; Crutchfield, J.P. Metastable evolutionary dynamics: crossing fitness barriers or escaping via neutral paths? *Bull. Math. Biol.* **2000**, *62*, 799–848. [CrossRef]
8. Burton, O.J.; Travis, J.M.J. The Frequency of fitness peak shifts is increased at expanding range margins due to mutation surfing. *Genetics* **2008**, *179*, 941–950. [CrossRef]
9. Arias, M.; le Poul, Y.; Chouteau, M.; Boisseau, R.; Rosser, N.; Thery, M.; Llaurens, V. Crossing fitness valleys: Empirical estimation of a fitness landscape associated with polymorphic mimicry. *Proc. R. Soc. B* **2016**, *283*, 20160391. [CrossRef]
10. Kimura, M. Diffusion models in population genetics. *J. Appl. Probab.* **1964**, *1*, 177–232. [CrossRef]
11. Iwasa, Y.; Michor, F.; Nowak, M.A. Stochastic tunnels in evolutionary dynamics. *Genetics* **2004**, *166*, 1571–1579. [CrossRef]
12. Covert, A.W., III; Lenski, R.E.; Wilke, C.O.; Ofria, C. Experiments on the role of deleterious mutations as stepping stones in adaptive evolution. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, E3171–E3178. [CrossRef] [PubMed]
13. Meer, M.V.; Kondrashov, A.S.; Artzy-Randrup, Y.; Kondrashov, F.A. Compensatory evolution in mitochondrial tRNAs navigates valleys of low fitness. *Nature* **2010**, *464*, 279–283. [CrossRef]
14. Osada, N.; Akashi, H. Mitochondrial–nuclear interactions and accelerated compensatory evolution: Evidence from the primate cytochrome c oxidase complex. *Mol. Biol. Evol.* **2012**, *29*, 337–346. [CrossRef]
15. Charlesworth, B. Effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* **2009**, *10*, 195–205. [CrossRef]
16. Haldane, J.B.S. The effect of variation on fitness. *Am. Nat.* **1937**, *71*, 337–349. [CrossRef]
17. McCandlish, D.M.; Stoltzfus, A. Modeling evolution using the probability of fixation: History and implications. *Q. Rev. Biol.* **2014**, *89*, 225–252. [CrossRef] [PubMed]
18. Kern, A.D.; Kondrashov, F.A. Mechanisms and convergence of compensatory evolution in mammalian mitochondrial tRNAs. *Nat. Genet.* **2004**, *36*, 1207–1212. [CrossRef]
19. Breen, M.S.; Kremena, C.; Vlasov, P.K.; Notredame, C.; Kondrashov, F.A. Epistasis as the primary factor in molecular evolution. *Nature* **2012**, *490*, 535–538. [CrossRef]
20. McCandlish, D.M.; Rajon, E.; Shah, P.; Ding, Y.; Plotkin, J.B. The role of epistasis in protein evolution. *Nature* **2013**, *497*, E1–E2. [CrossRef]
21. Breen, M.S.; Kremena, C.; Vlasov, P.K.; Notredame, C.; Kondrashov, F.A. Breen et al. reply. *Nature* **2013**, *497*, E2–E3. [CrossRef]
22. Starr, T.N.; Flynn, J.M.; Mishra, P.; Bolon, D.N.A.; Thornton, J.W. Pervasive contingency and entrenchment in a billion years of Hsp90 evolution. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 4453–4458. [CrossRef]
23. Otwinowski, J.; McCandlish, D.M.; Plotkin, J.B. Inferring the shape of global epistasis. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E7550–E7558. [CrossRef]
24. Posfai, A.; Zhou, J.; Plotkin, J.B.; Kinney, J.B.; McCandlish, D.M. Selection for protein stability enriches for epistatic interactions. *Genes* **2018**, *9*, 423. [CrossRef]
25. Sailer, Z.R.; Harms, M.J. Molecular ensembles make evolution unpredictable. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 11938–11943. [CrossRef]
26. Noivirt-Brik, O.; Unger, R.; Horovitz, A. Analyzing the origin of long–range interactions in proteins using lattice models. *BMC Struct. Biol.* **2009**, *9*, 4. [CrossRef]

27. Nelson, E.D.; Grishin, N.V. Long–range epistasis is mediated by structural change in a model of ligand binding proteins. *PLoS ONE* **2016**, *11*, e0166739. [CrossRef]
28. Miyazawa, S.; Jernigan, R.L. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term for simulation and threading. *J. Mol. Biol.* **1996**, *256*, 623–644. [CrossRef]
29. Serohijos, A.W.R.; Rimas, Z.; Shakhnovich, E.I. Protein biophysics explains why highly abundant proteins evolve slowly. *Cell Rep.* **2012**, *2*, 249–256. [CrossRef]
30. Serohijos, A.W.R.; Lee, S.Y.R.; Shakhnovich, E.I. Highly abundant proteins favor more stable 3D structures in yeast. *Biophys. J.* **2013**, *104*, L01–L03. [CrossRef]
31. Drummond, D.A.; Wilke, C.O. Mistranslation–induced protein misfolding as a dominant constraint on coding–sequence evolution. *Cell* **2008**, *134*, 341–352. [CrossRef]
32. Goldstein, R.A. The evolution and evolutionary consequences of marginal thermostability in proteins. *Proteins* **2011**, *79*, 1396–1407. [CrossRef]
33. Bloom, J.D.; Silberg, J.J.; Wilke, C.O.; Drummond, D.A.; Adami, C.; Arnold, F.H. Thermodynamic prediction of protein neutrality. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 606–611. [CrossRef]
34. Gillespie, J.H. *Population Genetics*; Johns Hopkins University Press: Baltimore, MD, USA, 2004.
35. Tachida, H. Molecular evolution in a multisite nearly neutral mutation model. *J. Mol. Evol.* **2000**, *50*, 69–81. [CrossRef]
36. Slade, G. Self–avoiding walks. *Math. Intel.* **1994**, *16*, 29–35. [CrossRef]
37. UT Southwestern Medical Center BioHPC. Available online: https://portal.biohpc.swmed.edu/content/ (accessed on 7 March 2019).
38. Felsenstein, J. Theoretical Population Genetics. 2016. See Box 3, p. 340. Available online: https://http://evolution.gs.washington.edu/pgbook/pgbook.pdf (accessed on 7 March 2019).
39. Tamuri, A.U.; dos Reis, M.; Goldstein, R.A. Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation–selection models. *Genetics* **2012**, *190*, 1101–1115. [CrossRef] [PubMed]
40. Nelson, E.D.; Grishin, N.V. Inference of epistatic effects in a key mitochondrial protein. *Phys. Rev. E* **2018**, *97*, 062404. [CrossRef]
41. Sarkisyan, K.S.; Bolotin, D.A.; Meer, M.V.; Usmanova, D.R.; Mishin, A.S.; Sharonov, G.V.; Ivankov, D.N.; Bozhanova, N.G.; Baranov, M.S.; Soylemez, O.; et al. Local fitness landscape of the green fluorescent protein. *Nature* **2016**, *533*, 397–401. [CrossRef]
42. Jäger, M.; Zhang, Y.; Bieschke, J.; Nguyen, H.; Dendle, M.; Bowman, M.E.; Noel, J.P.; Gruebele, M.; Kelly, J.W. Structure–function–folding relationship in a WW domain. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 10648–10653. [CrossRef] [PubMed]
43. Wilke, C.O. Molecular clock in neutral protein evolution. *BMC Genet.* **2004**, *5*, 25. [CrossRef]
44. Bonhoeffer, S.; Chappey, C.; Parkin, N.T.; Whitcomb, J.M.; Petropoulos, C.J. Evidence for positive epistasis in HIV–1. *Science* **2004**, *306*, 1547–1550. [CrossRef]
45. Moradigaravand, D.; Kouyos, R.; Hinkley, T.; Haddad, M.; Petropoulos, C.J.; Engelstädter, J.; Bonhoeffer, S. Recombination accelerates adaptation on a large–scale empirical fitness landscape in HIV–1. *PLoS Genet.* **2014**, *10*, e1004439. [CrossRef] [PubMed]