PNAS BRIEF REPORT | EVOLUTION STATISTICS OPEN ACCESS

Unpredictable repeatability in molecular evolution

Suman G. Das[a,1] and Joachim Krug[a]

Edited by Eugene Koonin, NIH, Bethesda, MD; received June 2, 2022; accepted August 12, 2022

The extent of parallel evolution at the genotypic level is quantitatively linked to the distribution of beneficial fitness effects (DBFE) of mutations. The standard view, based on light-tailed distributions (i.e., distributions with finite moments), is that the probability of parallel evolution in duplicate populations is inversely proportional to the number of available mutations and, moreover, that the DBFE is sufficient to determine the probability when the number of available mutations is large. Here, we show that when the DBFE is heavy-tailed, as found in several recent experiments, these expectations are defied. The probability of parallel evolution decays anomalously slowly in the number of mutations or even becomes independent of it, implying higher repeatability of evolution. At the same time, the probability of parallel evolution is non-self-averaging—that is, it does not converge to its mean value, even when a large number of mutations are involved. This behavior arises because the evolutionary process is dominated by only a few mutations of high weight. Consequently, the probability varies widely across systems with the same DBFE. Contrary to the standard view, the DBFE is no longer sufficient to determine the extent of parallel evolution, making it much less predictable. We illustrate these ideas theoretically and through analysis of empirical data on antibiotic-resistance evolution.

parallel evolution | distribution of fitness effects | predictability of evolution | antibiotic resistance

The repeatability and predictability of evolution are important questions in the field of evolutionary biology. In 1990, Stephen Jay Gould famously mused about "replaying life's tape" (1). In subsequent years, the topic of parallel evolution has become a major subject of empirical research (2–4), and theoretical questions concerning the probability of parallel evolution within the mathematical theory of population genetics have also attracted substantial attention (5–7). Here, the questions are focused mostly on changes at the level of genetic sequences. According to a common definition (6, 7), parallel evolution is said to occur when the exact same mutation is substituted in replicate populations. It is in this strong sense that we shall use the term "parallel evolution" here. The computation of the probability of parallel evolution is often set in a simplified scenario (5–7), where an asexual population evolves by strong selection and weak mutation (SSWM), which applies for moderately large population size and low mutation rate (more details are in *SI Appendix*). The evolutionary process starts with a homogeneous population and n possible beneficial mutations that can occur. Denoting by r_i the substitution rate of the i-th mutation, $i = 1, 2, \ldots, n$, the probability that the i-th mutation will be the first to fix is $W_i = r_i/(\sum_j r_j)$. Therefore, the probability that k replicate populations will all fix the same mutation is given by $P_k = \sum_i W_i^k$. Although we focus on the repeatability of the first substitution event, this measure can be generalized to address evolutionary trajectories with several mutations (see ref. 2 and further discussion in *SI Appendix*).

Within the SSWM approximation, the substitution rate of a mutation is the product of its mutation rate and fixation probability, $r_i = \mu_i \pi_i$, and the repeatability measure is affected equally by heterogeneities in μ_i and π_i (4). However, because empirical information on mutation-rate heterogeneity is sparse, theoretical studies have commonly assumed that all mutations occur at the same rate. If, in addition, the selection coefficients are small, $s_i \ll 1$, it follows that $r_i \sim \pi_i \sim s_i$, and the probability of parallel evolution is given by

$$P_k = \sum_i \frac{s_i^k}{\left(\sum_j s_j\right)^k}. \qquad [1]$$

We will adopt this simplification throughout, but emphasize that our theoretical results hold equally well for the full expression of P_k with r_i in place of s_i, provided that the distribution of the substitution rates r_i is heavy-tailed in the sense specified below.

The selection coefficients follow a distribution denoted by $P_s(s)$, which we refer to as the distribution of beneficial fitness effects (DBFE) (8). The mean probability of parallel evolution is $\langle P_k \rangle$, where $\langle \cdot \rangle$ denotes the average with respect to the DBFE. The extreme-value theory (EVT) hypothesis of fitness effects (see *SI Appendix* for details) predicts that the DBFE belongs to one of three classes of

Author affiliations: [a]Institute for Biological Physics, University of Cologne, D-50937 Cologne, Germany

Author contributions: J.K. designed research; S.G.D. and J.K. performed research; S.G.D. analyzed data; and S.G.D. and J.K. wrote the paper.

The authors declare no competing interest.

[1] To whom correspondence may be addressed. Email: sdas3@uni-koeln.de.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2209373119/-/DCSupplemental.

Published September 19, 2022.

distributions: The Weibull and Gumbel classes contain distributions with finite moments, whereas the Fréchet class contains distributions with power-law tails (and therefore diverging moments). In the last case, the asymptotic form of the DBFE is

$$P_s(s) \sim \frac{A}{s^{1+\alpha}}, \qquad [2]$$

with a scale parameter $A > 0$ and the tail exponent $\alpha > 0$. The cases of the Gumbel and Weibull extreme-value distributions have been explored in some detail for $k = 2$ (5–7). The Fréchet EVT class had been conjectured to be relatively unimportant biologically (7), but several subsequent studies (9–11) have uncovered signatures of heavy-tailed distributions of fitness effects (see *SI Appendix* for further details on tails of empirical DBFEs). In the realistic case of a large number of available beneficial mutations n, the statistics of P_k for heavy-tailed distributions of the form Eq. **2** are markedly different from those of light-tailed distributions, as will be shown below.

Results

The number n of beneficial mutations that can occur in a population varies widely across organisms and environments, but it is likely to be large. For bacterial populations, one may conservatively estimate that there are several thousand beneficial mutations (*SI Appendix*). We therefore focus on P_k in the large-n regime. The simplest computation is in the limiting case of neutral variation, where all selection coefficients are identical. Since all mutations are equally likely to be the first to fix, $P_k = 1/n^{k-1}$ (which is exact for all n). Specifically, the probability of parallel evolution in two replicates is $P_2 = 1/n$. Using this observation, one can define, for any system, the quantity $n_e = P_2^{-1}$ as the effective number of mutations that dominate the dynamics of fixation (further comments in *SI Appendix*). It is similar to the notion of the effective number of reproducing lineages studied in ref. 12 in the context of family size distributions.

The most commonly studied class of DBFEs is where all the moments are finite. In this case, the numerator and denominator in each term in Eq. **1** become uncorrelated as $n \to \infty$ and the distribution of $(\sum_i s_i^m)/n$ becomes sharply centered around the moment $\langle s^m \rangle$ of $P_s(\cdot)$. Therefore, for large n, we have

$$P_k \simeq \frac{\langle s^k \rangle}{\langle s \rangle^k} n^{-(k-1)}. \qquad [3]$$

Notice that, so far, we have omitted the angular brackets around P_k since it converges to the mean value in the limit of large n, as shown by the highly localized distribution of P_k in Fig. 1*A* (red dashed curve). Such quantities are described as "self-averaging." For the particular case $k = 2$, we have $P_2 \sim \frac{1}{n}$, which is the characteristic decay in self-averaging systems. It was shown in ref. 6 that, for an exponential distribution, $\langle P_2 \rangle = 2/(n+1)$ (see *SI Appendix* for a general expression for $\langle P_k \rangle$). Our focus here, however, is on heavy-tailed distributions with tails of the form Eq. **2**. When $k < \alpha$, Eq. **3** continues to hold. Particularly for $\alpha > 2$, P_2 still decays as $\sim 1/n$ (see *SI Appendix*, Fig S1*A*). However, when $k > \alpha$, the moment $\langle s^k \rangle$ diverges, and Eq. **3** no longer holds. We can now break the analysis down into two cases.

Case I. The moderately heavy-tailed case occurs when $\alpha > 1$; in this case, $\langle s \rangle$ is finite, but higher moments corresponding to $k > \alpha > 1$ diverge. For $k > \alpha$, the asymptotic behavior of $\langle P_k \rangle$ is

$$\langle P_k \rangle \simeq C_k n^{-(\alpha-1)}, \qquad [4]$$

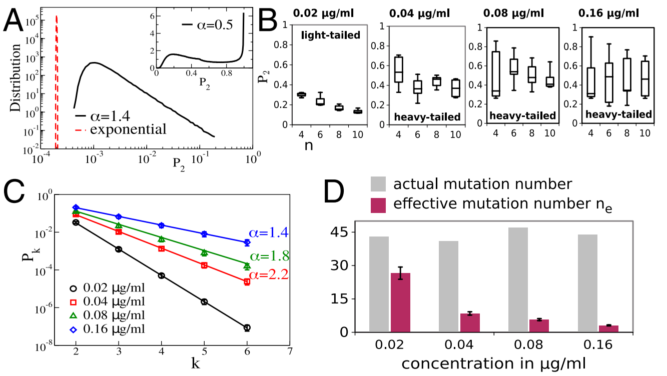

Fig. 1. (*A*) The black curve is the numerically sampled distribution of P_2 for $\alpha = 1.4$, and *Inset* shows the same for $\alpha = 0.5$; we used 10^6 realizations and $n = 10^4$ mutations. The dashed red curve is the distribution of P_2 for an exponential distribution of selection coefficients and $n = 10^4$. (*B*) This and the following panels analyze data from the study based on mutant screening reported in ref. 9, which determined the selection coefficients for several resistance-conferring mutations in TEM-1 β-lactamase. Here, we numerically estimate the distribution of P_2 from the selection coefficients reported in ref. 9. The dataset at each cefotaxime concentration was randomly split into subsets of size n in order to obtain distributions of P_2 as a function of n. The box plots show median, quartiles, and extreme values. (*C*) The P_k were obtained from the entire dataset at each concentration, and Eq. **5** was used to infer α. (*D*) The effective mutation number $n_e = 1/P_2$ has been computed and compared with the actual number of mutations in the available dataset at each concentration.

where the constant $C_k = A\Gamma(k-\alpha)\Gamma(\alpha)/(\Gamma(k)\langle s \rangle^\alpha)$. Note that $\langle P_k \rangle$ decays with an exponent less than $k-1$; therefore, the mean probability of parallel evolution is asymptotically much larger than in the case of light-tailed DBFEs. The scaling $n^{-(\alpha-1)}$ in Eq. **4** was first reported in ref. 9 and recently derived independently in ref. 12 in a different context. In particular, we see that when $1 < \alpha < 2$, $\langle P_2 \rangle$ decays anomalously—i.e., with an exponent < 1—in contrast to $P_2 \sim n^{-1}$, as in the light-tailed case (*SI Appendix*, Fig S1*A*).

It is important to point out that P_k does not become sharply centered around its mean value when $k > \alpha$, which can be shown as follows. The m-th moment is given by $\langle P_k^m \rangle = \langle P_{km} \rangle$ (*SI Appendix*). The value of $\langle P_{km} \rangle$ can be read off from Eq. **4** by replacing k by km. Thus, all moments are of the same order $n^{-(\alpha-1)}$. In particular, we notice that for $1 < \alpha < 2$, $\langle P_2^2 \rangle / \langle P_2 \rangle^2 \sim n^{\alpha-1}$. For self-averaging systems (which obey Eq. **3** for all k), this ratio goes asymptotically to one, and the SD vanishes relative to the mean (as seen in the red dashed curve in Fig. 1*A*). In contrast, here, we see that the SD diverges relative to the mean, implying a broad distribution for P_2, as illustrated in Fig. 1*A*. This non-self-averaging effect arises because the sum $P_k = \sum_i W_i^k$ is dominated by the largest weight $W_{\max}^k$ (12, 13). According to EVT, the largest selection coefficient scales as $n^{1/\alpha}$, implying that $W_{\max} \sim n^{1/\alpha - 1}$. Therefore, the scale of typical P_k is

$$P_k \sim n^{k(\frac{1}{\alpha}-1)}, \qquad [5]$$

for $k > \alpha$, which is asymptotically smaller than $\langle P_k \rangle$, as given by Eq. **4**. In fact, most of the weight is concentrated near the typical value, and the much higher mean is obtained from values of P_k that are much rarer, but have much higher magnitude.

Case II. For the severely heavy-tailed case $0 < \alpha < 1$, all integer moments of s diverge. It was shown in ref. 13 that for a power-law distribution with $0 < \alpha < 1$,

$$\langle P_k \rangle \simeq \frac{\Gamma(k-\alpha)}{\Gamma(k)\Gamma(1-\alpha)}, \qquad [6]$$

in the limit of large n. Specifically, the average probability of parallel evolution in two replicates is $\langle P_2 \rangle \simeq 1 - \alpha$. Note that the asymptotic form in Eq. **6** is independent of n, and, thus, we have the striking result that the probability of parallel evolution remains finite, even in the limit of an infinite number of available alternative mutations. In the present case, all moments of P_2 are of $O(1)$, and, therefore, P_2 is non-self-averaging. This is visible in the wide distribution of P_2, as shown in the numerically sampled plot in Fig. 1 *A*, *Inset*. Similar non-self-averaging effects are familiar in the physics of disordered systems (ref. 13 and references therein) and in probability theory (14).

While the moderately ($\alpha > 1$) and severely ($\alpha < 1$) heavy-tailed cases display somewhat different behavior, we note that both Eqs. **4** and **6** give rise to the recursion relation

$$\frac{\langle P_{k+1} \rangle}{\langle P_k \rangle} = 1 - \frac{\alpha}{k} \quad \text{for} \quad k \geq 2, \qquad [7]$$

which, therefore, holds for the entire range $0 < \alpha < k$. The result is independent of n and of all features of the underlying distribution, except the tail exponent α. It is, therefore, suitable for extracting α from empirical data; however, the disadvantage is that the averages require large datasets. Eq. **7** easily yields an approximate solution for large k, $\langle P_k \rangle \sim 1/k^{\alpha}$, shown in *SI Appendix*, Fig. S1*B*. The slow decay of $\langle P_k \rangle$ with k contrasts with the exponential decay of the typical P_k, as given by Eq. **5**.

The theoretical results discussed so far are valid in the limit of large n. Nonetheless, we will show that signatures of non-self-averaging effects can be discovered in limited empirical datasets. For this purpose, we use data on selection coefficients associated with antibiotic-resistance evolution reported in ref. 9. In this study, the fitness effects of 48 beneficial mutations in the resistance enzyme TEM-1 β-lactamase were reported for *Escherichia coli* growing at four different concentrations of the antibiotic cefotaxime (see *SI Appendix* for further details of the experiment). An analysis based on EVT indicated that the DBFE is light-tailed for the lowest concentration and heavy-tailed for the three higher concentrations, although large uncertainties were associated with the exponents in the latter case (9). We evaluate the statistics of P_2 for the four different concentrations (see *SI Appendix* for further details). Fig. 1*B* shows P_2 as a function of n. For the lowest concentration, P_2 is seen to be small with a small dispersion, and it decreases with n, consistent with our expectation. For the three higher concentrations, the values of P_2 are larger and have a large dispersion, which is consistent with heavy-tailed distributions. There is no discernible decrease with n. However, due to the relatively small values of n and the modest size of the datasets, it is not possible to distinguish this from a slow decrease with n. In Fig. 1*C*, we have plotted P_k as a function of k. Note that the distinction between the typical and mean values (12) has important implications here. Due to the limited size of the data, we have not used the recursion relation Eq. **7** to infer α. Instead, for each concentration, we have used the entire set of selection coefficients to create a single sample value of P_k, which is expected to be of the typical scale given by Eq. **5**. Using this, we estimate the exponent α, which is seen to progressively decrease with increasing concentration, indicating an increasingly heavy-tailed distribution. Thus, stronger selection pressures amplify the differences between fitness effects of beneficial mutations, leading to a broader distribution. Nonetheless, we should mention that inferred power-law exponents should be treated with some caution, since these can be sensitive to experimental errors or methods of analysis (9, 11). What is clear from Fig. 1*C*, however, is that the behavior of P_k is at least a good qualitative indicator of the dispersion of selection coefficients. We have also computed and plotted the effective mutation number n_e in Fig. 1*D*. The trend is, again, seen to be as predicted by theory. At the lowest concentration, n_e is relatively large and close to $(n+1)/2$ (where n is the actual number of mutations), consistent with an exponential distribution of selection coefficients. The effective mutation number decreases progressively with increasing concentration and indicates a slower-than-exponential tail.

Conclusions

Parallel phenotypic evolution often proceeds through distinct genotypic pathways. Here, we have shown that heavy-tailed DBFEs can substantially enhance the probability of parallel evolution even at the genotypic level. However, we also find that this probability varies widely across large, independent samples generated from the same heavy-tailed DBFE. This makes it harder to generalize the degree of repeatability from one model system to other, closely related ones. On the flip side, the evolutionary process is dominated by a few mutations of high weight, making evolution of the system more repeatable and, therefore, more predictable. For example, in the study (15) on antibiotic-resistance evolution, the authors found that the mutation of highest effect (which also features in the heavy-tailed distributions reported in ref. 9 and discussed above) occurred in the majority of multiple-replicate experiments. The full implications of these ideas in the context of natural populations remain to be elucidated.

Data, Materials, and Software Availability. All study data are included in the article and/or *SI Appendix*. The codes for numerical simulations are available on GitHub (https://github.com/meetsumand/Parallel-evolution) (16).

ACKNOWLEDGMENTS. S.G.D. and J.K. acknowledge support by the Deutsche Forschungsgemeinschaft (German Research Foundation) within Collaborative Research Centre 1310 Predictability in Evolution. We thank Arjan de Visser for helpful comments on the manuscript.

1. S. J. Gould, *Wonderful Life: The Burgess Shale and the Nature of History* (WW Norton & Company, New York, 1990).
2. J. A. G. M. de Visser, J. Krug, Empirical fitness landscapes and the predictability of evolution. *Nat. Rev. Genet.* **15**, 480–490 (2014).
3. J. F. Storz, Causes of molecular convergence and parallelism in protein evolution. *Nat. Rev. Genet.* **17**, 239–250 (2016).
4. S. F. Bailey, F. Blanquart, T. Bataillon, R. Kassen, What drives parallel evolution? How population size and mutational variation contribute to repeated evolution. *BioEssays* **39**, 1–9 (2017).
5. J. H. Gillespie, A simple stochastic gene substitution model. *Theor. Popul. Biol.* **23**, 202–215 (1983).
6. H. A. Orr, The probability of parallel evolution. *Evolution* **59**, 216–220 (2005).
7. P. Joyce, D. R. Rokyta, C. J. Beisel, H. A. Orr, A general extreme value theory model for the adaptation of DNA sequences under strong selection and weak mutation. *Genetics* **180**, 1627–1643 (2008).
8. T. Bataillon, S. F. Bailey, Effects of new mutations on fitness: Insights from models and data. *Ann. N. Y. Acad. Sci.* **1320**, 76–92 (2014).
9. M. F. Schenk, I. G. Szendro, J. Krug, J. A. G. M. de Visser. Quantifying the adaptive potential of an antibiotic resistance enzyme. *PLoS Genet.* **8**, e1002783 (2012).
10. C. Bank, R. T. Hietpas, A. Wong, D. N. Bolon, J. D. Jensen, A Bayesian MCMC approach to assess the complete distribution of fitness effects of new mutations: Uncovering the potential for adaptive walks in challenging environments. *Genetics* **196**, 841–852 (2014).
11. M. Foll *et al.*, Influenza virus drug resistance: A time-sampled population genetics perspective. *PLoS Genet.* **10**, e1004185 (2014).
12. H.-S. Niwa, Reciprocal symmetry breaking in Pareto sampling. arXiv [Preprint] (2022). https://arxiv.org/abs/2202.04865 (Accessed 10 February 2022).
13. B. Derrida, From random walks to spin glasses. *Physica D* **107**, 186–198 (1997).
14. J. Pitman, M. Yor, The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25**, 855–900 (1997).
15. T. van Dijk, S. Hwang, J. Krug, J. A. G. M. de Visser, M. P. Zwart, Mutation supply and the repeatability of selection for antibiotic resistance. *Phys. Biol.* **14**, 055005 (2017).
16. S. G. Das, Codes to generate numerical data in "Unpredictable repeatability in molecular evolution." GitHub. https://github.com/meetsumand/Parallel-evolution. Deposited 3 September 2022.