Quasi-species evolution maximizes genotypic reproductive value (not fitness or flatness)

Matteo Smerlak

Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany

Abstract

Growing efforts to measure fitness landscapes in molecular and microbial systems are motivated by a longstanding goal to predict future evolutionary trajectories. Sometimes under-appreciated, however, is that the fitness landscape and its topography do not by themselves determine the direction of evolution: under sufficiently high mutation rates, populations can climb the closest fitness peak (survival of the fittest), settle in lower regions with higher mutational robustness (survival of the flattest), or even fail to adapt altogether (error catastrophes). I show that another measure of reproductive success, Fisher's *reproductive value*, resolves the trade-off between fitness and robustness in the quasi-species regime of evolution: to forecast the motion of a population in genotype space, one should look for peaks in the (mutation-rate dependent) landscape of genotypic reproductive values—whether or not these peaks correspond to local fitness maxima or flat fitness plateaus. This new landscape picture turns quasi-species dynamics into an instance of non-equilibrium dynamics, in the physical sense of Markovian processes, potential landscapes, entropy production, etc.

Keywords: quasi-species, fitness landscape, reproductive value, NK model

Introduction

Darwinian micro-evolution is often conceptualized as the motion of populations in the space of all possible heritable types graded by their expected

*Declarations of interest: none

offspring number, the *fitness landscape* [1, 2, 3]. In Wright’s vivid words, the interaction of selection and variation enables populations to “continually find their way from lower to higher peaks” [4], in an open-ended exploration of possible forms that is the hallmark of biological evolution [5]. Thanks to modern sequencing technologies, fitness landscapes have now been measured in a variety of real molecular [6], viral [7] or microbial [8] systems. Thanks to these developments, the goal of *predicting* evolution no longer appears wholly out of reach [9, 10, 8, 11, 12]. In essence, if we know the topography of the fitness landscape—the location of its peaks, valleys, ridges, etc.—we should be able to compute how a population is likely to evolve next. Making such predictions from high-resolution fitness assays, possibly combined with inference methods [13, 14], is a central challenge of quantitative evolutionary theory.

In keeping with Wright’s metaphor, the quantitative analysis of empirical fitness landscapes [15, 16] has so far focused on locating fitness maxima and adaptive walks reaching them in the landscape [17, 18, 13]. This is legitimate in the strong selection regime of evolution [19], where populations are genetically homogeneous except for rare, fitness-increasing fixation events. But in large population subject to high mutational loads, e.g. in RNA viruses, these assumptions usually do not hold, and evolutionary dynamics is better studied within the framework of quasi-species theory [20], where new effects arise.

A hallmark of quasi-species evolution is that populations with different mutation rates experience the same fitness landscape differently. As is well known, populations whose mutation rate are too large can fail to adapt altogether, sometimes (though not always) following sharp “error threshold” phenomena [21, 22]. More subtly, mutational robustness tends to increase during neutral quasi-species evolution [23], and can sometimes outweigh reproductive rate as a determinant of evolutionary success (“survival of the flattest”) [24, 25]. In short, quasi-species evolution favours not just fitness, but some trade-off between fitness and robustness.

These phenomena raise fundamental questions regarding the *dynamical* analysis of fitness landscapes: When is flatter better than fitter? Where are the

Figure 1: Left: An NKp fitness landscape with 20 fitness peaks (inset) is more easily analyzed in terms of local maxima of the reproductive value $\mathbf{R}$, which are fewer than fitness peaks, with larger basins of attractions (size of points); see [26] for another illustration of this point. Here the mutation rate is $\mu = 0.1$, below the landscape's error threshold. Right: Solving the Crow-Kimura equation in this landscape shows that a monomorphic population with genotype 110 evolves towards the peak 179 through the basins of attraction of peaks 222 and 95, as dictated by the basin hopping graph of the $\mathbf{R}$ landscape. Trying to forecast this evolution from the bare fitness landscape $\mathbf{r}$ itself is hopeless in this case, especially since mean fitness *decreases* over time.

evolutionary attractors in a given landscape with multiple peaks and/or neutrality? What quantity do evolving populations optimize? Can we estimate the time scale before another attractor is visited? More simply, can we predict the future trajectory of an evolving population—say an influenza strain, or an intra-host HIV population—from its current location, the topography of its landscape, and the mutation rate?

In Ref. [26] I introduced a general framework to coarse-grain and forecast intermittent dynamics in random media. This Letter aims to highlight its applicability to quasi-species evolution, and to explain its connection with (i) the concept of *reproductive value* introduced by R. Fisher in relation to age-structured populations [27], and (ii) the *retrospective process* discussed in Ref. [28] in the context of mutation-selection balance. Out of these connections emerges a new landscape picture of microevolution, in which large evolving populations climb peaks in the landscape of genotypic reproductive values—not fitness peaks.

Theory

Quasi-species theory [20] describes the evolution of large molecular or viral populations using ordinary differential equations. To each genotype i we associate a density $x_i(t)$ that changes over time through selection and mutation. In the Crow-Kimura formulation [29], we have a non-linear system of first-order equations

$$\dot{x}_i = (r_i - \bar{r}(x))x_i + \mu \sum_j \Delta_{ij}x_j \quad (1)$$

where Δ is the Laplacian matrix of the mutation graph ($\Delta_{ij} = 1$ if i and j can be reached through a mutation, Δ_{ii} is minus the number of mutants of i , and $\Delta_{ij} = 0$ else), μ the mutation rate (assumed symmetric for simplicity), r_i is the Malthusian fitness of i , and $\bar{r}(x) = \sum_j r_j x_j$ is the population mean fitness. Other formulations are possible, e.g. Eigen's model where mutation arise as replication events [21]. The key point is that populations are described as a heterogeneous mutant cloud rather than a single consensus sequence. Such a description is necessary for RNA viruses such as influenza or HIV whose mutation rates are so high that any population consists of highly diverse genomes [30].

The quasi-species equation (1) can be linearized by dropping the non-linear mean-fitness term $\bar{r}(x)$, whose sole effect is to maintain normalization of densities. In this way (1) becomes similar to the parabolic Anderson model introduced by Zel'dovich *et al.* as a model of intermittent dynamics [31]; indeed, one salient feature of quasi-species dynamics is its punctuated nature, with populations localizing in regions of the fitness landscapes for long transients before suddenly jumping to remote areas in bursts of rapid adaptation [32].

In [26] I argue that population-dynamical equations of the form $\dot{\mathbf{x}} = \mathbf{x}M$ (with $\mathbf{x} = (x_i)_i$ and $M = \mu\Delta + r$ in the Crow-Kimura model¹) are best studied in terms of a related Markov process, generated by the Markov infinitesimal

¹In general M is an irreducible, essentially non-negative matrix.

generator

$$Q_{ij} = R_i^{-1}(M_{ij} - \Lambda\delta_{ij})R_j. \quad (2)$$

Here $\mathbf{R} = (R_i)$ is the right eigenvector of M corresponding to the eigenvalue Λ of M with largest real part, normalized so that $\sum_i R_i = 1$; in particular $\mathbf{R}$ depends on the mutation graph and rate. The process (2), which determines $\mathbf{x}$ as much as the original equation $\dot{\mathbf{x}} = \mathbf{x}M$, may be interpreted as follows.

Firstly, the component R_i of the right dominant eigenvector measures the contribution of a type i to the asymptotic population $\mathbf{x}(t)$ with $t \rightarrow \infty$: after a long time, the number of descendants of an ancestor of type i is proportional to R_i . In this sense, R_i is a generalization of Fisher’s *reproductive value*, a notion he originally introduced in age-structured populations as the answer to the question “To what extent will persons of this age, on the average, contribute to the ancestry of future generations?” [27] In the context of quasi-species dynamics, i is a genotype and R_i measures the relative number of i -clones in a lineage started from an i -ancestor. (This should not be confused with the *quasi-species distributions*, given by the left dominant eigenvector $\mathbf{L} = (L_i)$ of M and describing the composition of the population at mutation-selection equilibrium; $\mathbf{L}$ and $\mathbf{R}$ only coincide if M is symmetric.)

Secondly, the process (2) can be interpreted as a *retrospective process* generating the transitions between types observed along successful lineages of a branching process with mean matrix M . This process is built as follows: sample an individual at some late time from the equilibrium quasi-species distribution $\mathbf{L}$, and trace its ancestry back in time along its genealogical tree. Along this backward path ancestors have genotypes $i_1, i_2, \dots$. The retrospective process (2) generates these transitions forward in time [33, 34].²

Thirdly, because (2) generates a Markov process, we can study it using the same tools as other non-equilibrium systems in physics or biology. In particular, we can picture (2) as describing a random walk (with unit diffusivity) in the

²Refs. [33, 34] used the process (2) to characterize mutation-selection equilibria; here, by contrast, I use it to describe the dynamics of a quasi-species away from equilibrium.

potential $U_i = -2 \log R_i$, characterize the distance to equilibrium using the relative entropy with respect to the Gibbs distribution $\pi_i = R_i L_i$ (the equilibrium distribution for (2)), or measure the irreversibility of evolutionary dynamics with path entropies $s(\omega) = \log \text{Prob}(\omega) / \text{Prob}(\omega')$, where $\omega = (i_0, i_1, \dots)$ is a path of the retrospective process and ω' the time-reversed trajectory.

Perhaps more importantly, this new landscape picture provides a solution to the prediction problem in quasi-species dynamics: to find the region of genotype space where a population will go from its current location i , find the closest $\mathbf{R}$ peak α . Once the population enters the basin of attraction of that peak, it will concentrate there for a long time—until it jumps to another peak surround by another basin of attraction. In other words, coarse-grained evolutionary trajectories can be defined by aggregating the densities within the basins of attraction of reproductive value maxima, and transitions (punctuations) forecasted in terms of neighboring basins and lowest saddles between them.

Illustration

To illustrate the predictive power of this principle in a model fitness landscape exhibiting both ruggedness and neutrality, I considered an instance of the NKp landscape [35], in which the fitness of a binary sequence of length N derives from N additive contributions whose values depend on the alleles at K linked loci, with a probability p that each contribution is exactly zero.³ The landscape in Fig. 1, chosen with $N = 8$, epistasis parameter $K = 6$ and neutrality parameter $p = 0.7$, has 20 local maxima and an error threshold at $\mu_c \simeq 0.2$. Comparing the basin hopping graphs of the fitness landscape $\mathbf{r}$ and of the genotypic reproductive value $\mathbf{R}$ reveals that most of the complexity of the former is spurious. Moreover, coarse-grained evolutionary trajectories, described by the basin frequencies, are consistent with the succession of transitions predicted

³Specifically, the NKp model defines the fitness of a length- N binary string $i = (s_1, \dots, s_N)$ as $r_i = \sum_{m=1}^N f_m(s_m, \dots, s_{m+K}) b_m(s_m, \dots, s_{m+K}) / N$ where the fitness components f_m and b_m are independent with $U(0, 1)$ and Bernoulli(p) distributions, respectively.

from the basin hopping graph [36] of $\mathbf{U}$: a population initially concentrated around the genotype 110 (a global fitness maximum) will evolve towards the flatter genotype 179 via the basins of 222 and 95. This simple example highlights the usefulness of studying quasi-species dynamics in terms of reproductive values: while forecasting the coarse-grained dynamics through $\mathbf{R}$ is straightforward, the same task becomes exceedingly impractical from the peak, valley and plateau structure of the fitness landscape $\mathbf{r}$.

Discussion

A widely shared understanding of the role of mutations in evolution has them feeding raw material to the fitness-maximizing sieve of natural selection. But when mutation rates are high, as they are in *e.g.* RNA viruses [37] and likely were in early life [20], evolutionary success requires more than the discovery of a high-fitness mutant genotype: the mutants of the new mutant must also have relatively high fitness, *i.e.* the mutant type must be mutationally robust. Fisher’s notion of reproductive value combines fitness and flatness into a derived landscape that determines evolutionary trajectories across the spectrum of mutation rates.

The importance of reproductive values in evolutionary dynamics—in particular, its role as the target of selection—has been emphasized repeatedly since Fisher, see *e.g.* [38, 39, 40]. Usually, though, reproductive values are defined with respect to classes other than genotype: sex, age, etc. The conclusion of this work is that reproductive values defined with respect to genotypes themselves (and parametrized by mutation rates) help us coarse-grain, predict and generally study quasi-species dynamics (*e.g.* using physics-inspired methods) in ways that the bare fitness landscape cannot.

Funding for this work was provided by the Alexander von Humboldt Foundation in the framework of the Sofja Kovalevskaja Award endowed by the German Federal Ministry of Education and Research.

References

References

- [1] P. F. Stadler, Fitness landscapes, in: Biological Evolution and Statistical Physics, Springer Berlin Heidelberg, 2002, pp. 183–204. doi:10.1007/3-540-45692-9_10.
URL https://doi.org/10.1007%2F3-540-45692-9_10
- [2] H. A. Orr, Fitness and its role in evolutionary genetics, Nature Rev. Genet. 10 (8) (2009) 531–539. doi:10.1038/nrg2603.
URL <https://doi.org/10.1038%2Fnrg2603>
- [3] I. Fragata, A. Blanckaert, M. A. D. Louro, D. A. Liberles, C. Bank, Evolution in the light of fitness landscape theory, Trends Ecol. Evol. 34 (1) (2019) 69–82. doi:10.1016/j.tree.2018.10.009.
URL <https://doi.org/10.1016%2Fj.tree.2018.10.009>
- [4] S. Wright, The roles of mutation, inbreeding, crossbreeding and selection in evolution, Proceedings of the Sixth International Congress on Genetics 1 (1932) 356–366.
- [5] H. P. de Vladar, M. Santos, E. Szathmary, Grand Views of Evolution, Trends in Ecology & Evolution 32 (5) (2017) 324–334. doi:10.1016/j.tree.2017.01.008.
URL <https://doi.org/10.1016%2Fj.tree.2017.01.008>
- [6] C. Blanco, E. Janzen, A. Pressman, R. Saha, I. A. Chen, Molecular Fitness Landscapes from High-Coverage Sequence Profiling, Annu. Rev. Biophys. 48 (1) (2019) 1–18. doi:10.1146/annurev-biophys-052118-115333.
URL <https://doi.org/10.1146%2Fannurev-biophys-052118-115333>
- [7] P. T. Dolan, Z. J. Whitfield, R. Andino, Mapping the Evolutionary Potential of RNA Viruses, Cell Host Microbe 23 (4) (2018) 435–446. doi:10.1016/j.chom.2018.03.012.
URL <https://doi.org/10.1016%2Fj.chom.2018.03.012>

- [8] J. A. G. de Visser, J. Krug, Empirical fitness landscapes and the predictability of evolution, *Nature Rev. Genet.* 15 (7) (2014) 480–490. doi:10.1038/nrg3744.
URL <https://doi.org/10.1038/nrg3744>
- [9] D. M. Weinreich, Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins, *Science* 312 (5770) (2006) 111–114. doi:10.1126/science.1123539.
URL <https://doi.org/10.1126/science.1123539>
- [10] A. E. Lobkovsky, E. V. Koonin, Replaying the Tape of Life: Quantification of the Predictability of Evolution, *Front. Genet.* 3 (2012).
URL <https://doi.org/10.3389/fgene.2012.00246>
- [11] M. Lässig, V. Mustonen, A. M. Walczak, Predicting evolution, *Nature Ecol. Evol.* 1 (3) (feb 2017). doi:10.1038/s41559-017-0077.
URL <https://doi.org/10.1038/s41559-017-0077>
- [12] J. A. G. M. de Visser, S. F. Elena, I. Fragata, S. Matuszewski, The utility of fitness landscapes and big data for predicting evolution, *Heredity* 121 (5) (2018) 401–405. doi:10.1038/s41437-018-0128-4.
URL <https://doi.org/10.1038/s41437-018-0128-4>
- [13] C. Bank, S. Matuszewski, R. T. Hietpas, J. D. Jensen, On the (un) predictability of a large intragenic fitness landscape, *Proceedings of the National Academy of Sciences* 113 (49) (2016) 14085–14090.
- [14] X. Ding, Z. Zou, C. L. Brooks III, Deciphering protein evolution and fitness landscapes with latent space models, *Nature communications* 10 (1) (2019) 1–13.
- [15] D. M. McCandlish, Visualizing fitness landscapes, *Evolution* 65 (6) (2011) 1544–1558. doi:10.1111/j.1558-5646.2011.01236.x.
URL <https://doi.org/10.1111/j.1558-5646.2011.01236.x>

- [16] I. G. Szendro, M. F. Schenk, J. Franke, J. Krug, J. A. G. M. de Visser, Quantitative analyses of empirical fitness landscapes, *J. Stat. Mech.* 2013 (01) (2013) P01005. doi:10.1088/1742-5468/2013/01/p01005. URL <https://doi.org/10.1088/1742-5468/2013/01/p01005>
- [17] S. Kauffman, S. Levin, Towards a general theory of adaptive walks on rugged landscapes, *Journal of Theoretical Biology* 128 (1) (1987) 11–45. doi:10.1016/s0022-5193(87)80029-2. URL [https://doi.org/10.1016/s0022-5193\(87\)80029-2](https://doi.org/10.1016/s0022-5193(87)80029-2)
- [18] J. Franke, A. Klözer, J. A. G. M. de Visser, J. Krug, Evolutionary Accessibility of Mutational Pathways, *PLoS Comput. Biol.* 7 (8) (2011) e1002134. doi:10.1371/journal.pcbi.1002134. URL <https://doi.org/10.1371/journal.pcbi.1002134>
- [19] G. Sella, A. E. Hirsh, The application of statistical physics to evolutionary biology, *Proc. Natl. Acad. Sci. USA* 102 (27) (2005) 9541–9546. doi:10.1073/pnas.0501865102. URL <https://doi.org/10.1073/pnas.0501865102>
- [20] M. Eigen, J. McCaskill, P. Schuster, The Molecular Quasi-Species, *Adv. Chem. Phys.* 75 (1989) 149–263. doi:10.1002/9780470141243.ch4. URL <https://doi.org/10.1002/9780470141243.ch4>
- [21] M. Eigen, Selforganization of matter and the evolution of biological macromolecules, *Sci. Nat.* 58 (10) (1971) 465–523. doi:10.1007/bf00623322. URL <https://doi.org/10.1007/bf00623322>
- [22] E. Domingo, P. Schuster, What Is a Quasispecies? Historical Origins and Current Scope, in: *Current Topics in Microbiology and Immunology*, Springer International Publishing, 2015, pp. 1–22. doi:10.1007/82_2015_453. URL https://doi.org/10.1007/82_2015_453

- [23] E. van Nimwegen, J. P. Crutchfield, M. Huynen, Neutral evolution of mutational robustness, *Proc. Natl. Acad. Sci. USA* 96 (17) (1999) 9716–9720. doi:10.1073/pnas.96.17.9716.
URL <https://doi.org/10.1073/pnas.96.17.9716>
- [24] C. O. Wilke, J. L. Wang, C. Ofria, R. E. Lenski, C. Adami, Evolution of digital organisms at high mutation rates leads to survival of the flattest, *Nature* 412 (6844) (2001) 331–333. doi:10.1038/35085569.
URL <https://doi.org/10.1038/35085569>
- [25] F. M. Codoñer, J.-A. Darós, R. V. Solé, S. F. Elena, The Fittest versus the Flattest: Experimental Confirmation of the Quasispecies Effect with Subviral Pathogens, *PLoS Pathog.* 2 (12) (2006) e136. doi:10.1371/journal.ppat.0020136.
URL <https://doi.org/10.1371/journal.ppat.0020136>
- [26] M. Smerlak, Intermittency as metastability: a predictive approach to evolution in disordered environments, *arXiv preprint arXiv:2009.03608* (2020).
- [27] R. A. Fisher, *The genetical theory of natural selection.*, Clarendon Press, 1930. doi:10.5962/bhl.title.27468.
URL <https://doi.org/10.5962/bhl.title.27468>
- [28] Y. Wang, Variational characterization of multitype galton-watson branching processes, *Acta Mathematicae Applicatae Sinica, English Series* 30 (1) (2014) 27–36.
- [29] J. F. Crow, M. Kimura, *An Introduction to Population Genetics Theory*, Harper and Row, New York, 1970.
- [30] E. Domingo, C. Perales, Viral quasispecies, *PLOS Genet.* 15 (10) (2019) e1008271. doi:10.1371/journal.pgen.1008271.
URL <https://doi.org/10.1371/journal.pgen.1008271>
- [31] Y. B. Zel'dovich, S. A. Molchanov, A. A. Ruzmaikin, D. D. Sokolov, Intermittency in random media, *Soviet Physics Uspekhi* 30 (5) (1987) 353–369.

doi:10.1070/pu1987v030n05abeh002867.

URL <https://doi.org/10.1070%2Fpu1987v030n05abeh002867>

- [32] Y. Bakhtin, M. I. Katsnelson, Y. I. Wolf, E. V. Koonin, Punctuated equilibrium as the default mode of evolution of large populations on fitness landscapes dominated by saddle points in the weak-mutation limit, arXiv preprint arXiv:2007.10228 (2020).
- [33] J. Hermisson, O. Redner, H. Wagner, E. Baake, Mutation–selection balance: ancestry, load, and maximum principle, *Theoretical population biology* 62 (1) (2002) 9–46.
- [34] E. Baake, H.-O. Georgii, Mutation, selection, and ancestry in branching models: a variational approach, *Journal of mathematical biology* 54 (2) (2007) 257–303.
- [35] L. Barnett, Ruggedness and neutrality—the NKp family of fitness landscapes, in: *ALIFE Proceedings of the sixth international conference on Artificial life*, MIT Press Cambridge, MA, USA, 1998.
- [36] M. Kucharik, I. L. Hofacker, P. F. Stadler, J. Qin, Basin Hopping Graph: a computational framework to characterize RNA folding landscapes, *Bioinformatics* 30 (14) (2014) 2009–2017. doi:10.1093/bioinformatics/btu156.
URL <https://doi.org/10.1093%2Fbioinformatics%2Fbtu156>
- [37] J. W. Drake, J. J. Holland, Mutation rates among RNA viruses, *Proc. Natl. Acad. Sci. USA* 96 (24) (1999) 13910–13913. doi:10.1073/pnas.96.24.13910.
URL <https://doi.org/10.1073%2Fpnas.96.24.13910>
- [38] P. D. Taylor, Allele-frequency change in a class-structured population, *The American Naturalist* 135 (1) (1990) 95–106.
- [39] A. Grafen, A theory of fisher’s reproductive value, *Journal of mathematical biology* 53 (1) (2006) 15–60.

- [40] A. Gardner, The genetical theory of multilevel selection, *Journal of evolutionary biology* 28 (2) (2015) 305–319.