The quasispecies for the Wright–Fisher model

Raphaël Cerf
DMA, École Normale Supérieure

Joseba Dalmau
CMAP, Ecole Polytechnique

February 27, 2017

Summary. We consider the classical Wright–Fisher model of population genetics. We prove the existence of an error threshold for the mutation probability, below which a quasispecies is formed. We show a new phenomenon, specific to a finite population model, namely the existence of a population threshold: to ensure the stability of the quasispecies, the population size has to be at least of the same order as the genome length. We derive an explicit formula describing the quasispecies.

Introduction. According to the Darwinian theory of evolution, mutation and selection are the fundamental forces governing the evolution of living creatures. The mutations occur during the reproduction process. A mutation on a fixed gene has a low probability, but for long genomes, it is very likely that several randomly chosen genes undergo a mutation event. The selection force is quantified by the Darwinian fitness, which measures the mean number of offspring of an individual. A fundamental problem is to develop a quantitative understanding of the statistical structure of a biological population in equilibrium. Such an equilibrium realizes a delicate balance between mutation and selection. In the absence of mutations, a typical stable situation occurs when the genomes of all the individuals are identical. Because of the mutations, genetic diversity is constantly reintroduced in the population and a typical stable situation occurs when the genomes of the population are very close to a specific well fit genotype, called the wild type or the master sequence. Hence the population looks like a cloud of mutants centered around the wild type. This kind of equilibrium was discovered within the framework of Eigen's model and was called a quasispecies[1,2]. Another fundamental discovery of Eigen is the existence of an error threshold. Namely, the quasispecies is stable only if the mutation probability is below a critical value, which scales as the inverse of the length of the genome. These notions had a profound impact on the understanding of molecular evolution[3]. It seems that some RNA viruses, like the HIV virus, evolve with a rather high mutation rate, which is adjusted to be close to an error threshold[4,5]. Some promising antiviral strategies consist in using mutagenic drugs that induce an error catastrophe[6,7].

The original goal of Eigen was to understand the first stages of life on Earth. Most presumably, the first living creatures were complex macromolecules. Eigen suggested that, at the macroscopic level, their evolution could be adequately described by a collection of chemical reactions. These reactions model the replication or the degradation of each type of macromolecule. The concentrations of each type of macromolecule obey a system of differential equations, derived from the laws of chemical kinetics. Thus Eigen's model is formulated for an infinite population and the evolution is deterministic. This creates a major obstacle if one wishes to extend the notions of quasispecies and error threshold to population genetics. Biological populations are finite, and even if they are large so that they might be considered infinite in some approximate scheme, it is not coherent to consider situations where the size of the population is much larger than the number of possible genotypes. Moreover, it has long been recognized that random effects play a major role in the genetic evolution of populations[8], yet they are ruled out from the start in a deterministic infinite population model. Therefore, it is crucial to develop a finite population counterpart to Eigen's model, which incorporates stochastic effects[2,9]. Here, we achieve this program within the framework of the classical Wright–Fisher model of population genetics. We prove the existence of an error threshold for the mutation probability, below which a quasispecies is formed. We show a new phenomenon, specific to a finite population model, namely the existence of a population threshold: to ensure the stability of the quasispecies, the population size has to be at least of the same order as the genome length. In addition, we derive an explicit formula describing the quasispecies. Our results hold also for the Moran model and are supported by computer simulations.

The Wright–Fisher model. A population of m individuals evolves under selection and mutation. The individuals are characterized by their genotype. Let us denote by ℓ the length of the genome of an individual. A genotype is a sequence of ℓ letters chosen among A, U, G, C. In the Wright–Fisher model, the generations do not overlap. Let us explain the mechanism to build the generation $n+1$ from generation n. We first select m individuals randomly with replacement from the generation n. The individuals are not chosen uniformly, but according to their Darwinian fitness. The probability to select an individual is equal to the fitness of the individual divided by the sum of the fitnesses of all the individuals in generation n. The m selected individuals undergo a reproduction phase and their m offspring constitute the generation $n+1$. Yet the reproduction mechanism is error–prone and gives rise to mutations. We suppose

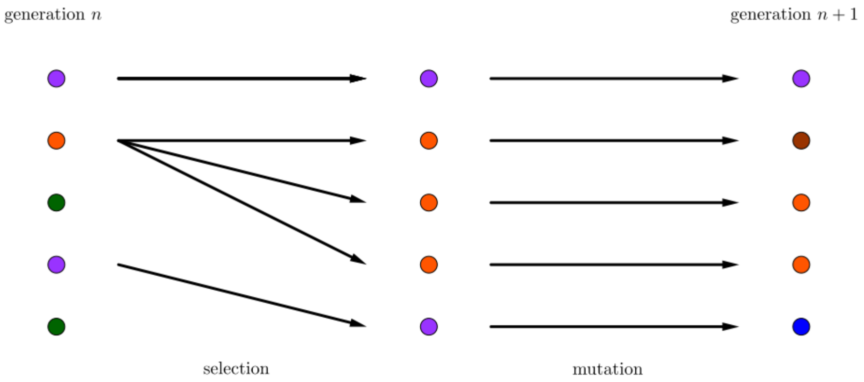

Figure 1: The transition mechanism of the Wright–Fisher model

that mutations occur independently on each gene of the genome and we denote by q the mutation probability. Whenever a mutation occurs, the final gene is chosen randomly with uniform probability among the three remaining genes in A, U, G, C. The average number of mutations during the reproduction cycle of an individual is then equal to $a = \ell q$. This model is already very difficult to analyze, especially on an arbitrary fitness landscape. We consider here the case of the sharp peak landscape. We suppose that there exists a specific genotype, denoted by w^*, and called the wild type or the Master sequence, which has a superior fitness $\sigma > 1$, while all other genotypes have fitness equal to 1. In the end, the population dynamics is governed by four parameters: m, ℓ, q, σ. We wish to understand the picture at equilibrium, after many generations, and we would like to answer the following questions. Does the population concentrate around the wild type? Will the wild type be lost? What is the probability of each of these scenarios?

Mutation and population threshold. The previous questions seem to be intractable for fixed values of the parameters m, ℓ, q, σ. We will consider the asymptotic regime of long genomes and large populations, which enables us to discover sharply contrasted pictures. In this asymptotic regime, we prove the existence of critical thresholds for the mutation probability and the population size. Namely, the equilibrium state of the population critically depends on the ratio m/ℓ and the product $a = \ell q$.

Theorem. There exists a function $\psi(a)$ such that the following holds.
1) If $m/\ell < \psi(a)$, then, at equilibrium, with probability going to 1 as m, ℓ go to ∞, the population does not contain the wild type.
2) If $m/\ell > \psi(a)$, then, at equilibrium, with probability going to 1 as m, ℓ go to ∞, the population contains a positive fraction of the wild type.
The function $\psi(a)$ is finite positive on $]0, \ln\sigma[$ and $\psi(\ln\sigma) = +\infty$.

This theorem is rigorously proved[10]. The proof involves classical probabilistic tools, namely the ergodic theorem, a renewal argument, lumping, coupling, large deviations estimates. Although we stated our result in the framework of the Wright–Fisher model, we proved it also for the Moran model[11], in which successive generations do overlap. The conclusions are essentially the same, only the function ψ is different. This leads us to believe that the result is quite robust and should hold for many variants of mutation–selection models. Two qualitative conclusions can be drawn from the two quantitative results of the theorem:
1) There exists an error threshold for the mutation probability, above which the system undergoes an error catastrophe: whatever large the population size is, the wild type is likely to be lost.
2) There exists also a population threshold: below the error threshold, the wild type is likely to be present in the population at equilibrium only if the population size is large enough.

The first conclusion is the counterpart for the Wright–Fisher model of the error threshold phenomenon in the Eigen model. The second conclusion shows that a new phenomenon occurs for finite populations, and that the ratio

$$\frac{m}{\ell} = \frac{\text{size of the population}}{\text{length of the genome}}$$

is crucial to determine the stable equilibrium. More precisely, a quasispecies can emerge only if

$$\frac{m}{\ell} > \min\psi\,, \qquad \ell q < \ln\sigma\,.$$

Hence, for the quasispecies to be a stable equilibrium, the mutation probability has to be of order the inverse of the genome length, while the size of the population has to be of the same order as the genome length.

The quasispecies formula. Another fundamental question concerning the quasispecies model is the following: when a quasispecies is formed, what does it look like? The first works on Eigen's model describe a quasispecies as a cloud of mutants centered around the master sequence. A quasispecies is formed at equilibrium, when the mutation probability is below the error threshold. To put in evidence the error threshold phenomenon, we have

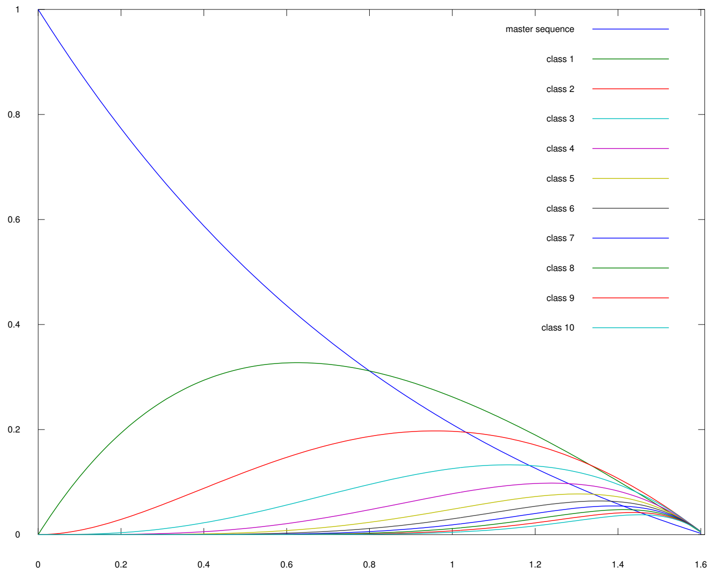

Figure 2: The quasispecies distribution as a function of a for $\sigma = 5$

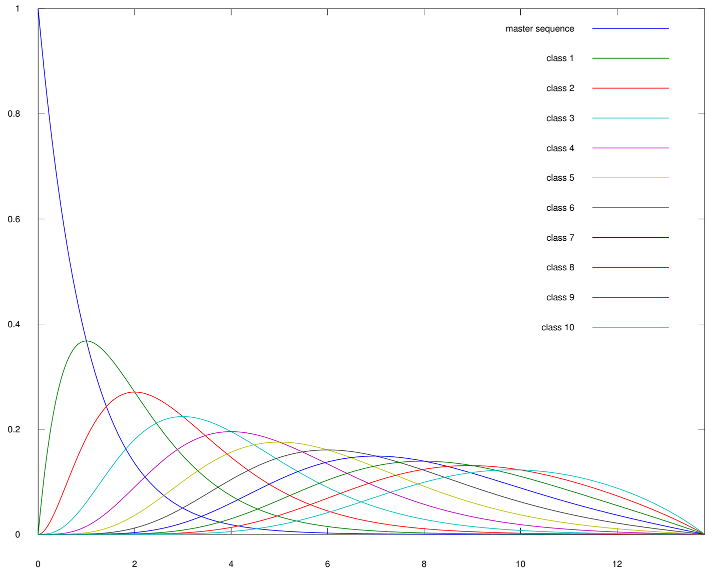

Figure 3: The quasispecies distribution as a function of a for $\sigma = 10^6$

to consider the asymptotic regime of long genomes. There are exactly 3ℓ mutants that are one mutation away from the wild type. Due to the symmetry of the mutation and selection mechanisms, all of them will have the same concentration at equilibrium. When the length of the genome is very long, the concentration of every single mutant becomes negligible. We collect together in a single class all the mutants that are one mutation away from the wild type, and we call this class the first mutant class. The second mutant class is built by collecting together all the mutants that are exactly two mutations away from the wild type. The third, fourth and following mutant classes are built in a similar way. The zeroth class is formed by the wild type alone. At equilibrium, when a quasispecies is formed, all these classes have positive concentrations. The concentration of the wild type is given by $\rho_0 = (\sigma e^{-a} - 1)/(\sigma - 1)$. More generally, the concentration of the kth mutant class is given by the formula

$$\rho_k \,=\, (\sigma e^{-a} - 1)\frac{a^k}{k!}\sum_{i\geq 1}\frac{i^k}{\sigma^i}\,.$$

The sequence $(\rho_k)_{k\geq 0}$ is a probability distribution on the non–negative integers, $\mathcal{Q}(\sigma, a)$, which we call the quasispecies distribution with parameters σ and a[12,13]. Recall that σ represents the selective advantage of the master sequence, and $a = \ell q$ represents the mean number of mutations per genome per generation. Thus, the population size m is a fundamental parameter in order to decide whether a quasispecies can form or not, but if it does form, the distribution of the quasispecies is independent of the population size. The figures $2, 3$ show the concentrations of the different mutant classes as a function of a, for fixed values of σ. Similar curves have been obtained in the framework of Eigen's model[2]. The curves obtained in these previous works were generated by simulating Eigen's system of differential equations. The quasispecies distribution provides now an exact formula for generating the same curves. The formula for the quasispecies distribution possesses an extremely rich combinatorial structure, and it is linked to several classical mathematical sequences of numbers, like the Eulerian numbers and the Stirling numbers of the second kind.

Computer simulations. These results are supported by simulations. Figures $4, 5, 6$ show the fraction of the Master sequence in the equilibrium population as a funtion of both $a = \ell q$ and m/ℓ. In figure 4, the ratio m/ℓ is fixed and we vary $a = \ell q$, while in figure 5, the parameter $a = \ell q$ is fixed and we vary m/ℓ. In figure 6, we vary simultaneously m/ℓ and $a = \ell q$, and we obtain a two–dimensional surface. The programs are written in C with the help of the GNU scientific library and the graphical output is generated with the help of the Gnuplot program. The number of generations in a simulation run was adjusted empirically in order to stabilize the output

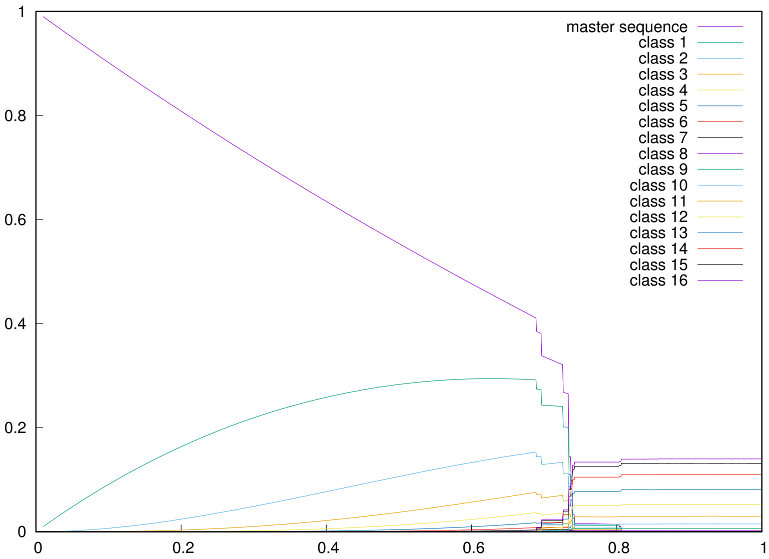

Figure 4: Density of the first mutant classes as a function of $a = \ell q$, for the Wright–Fisher model with parameters $\sigma = 2$, $\ell = 32$, $m/\ell = 3$

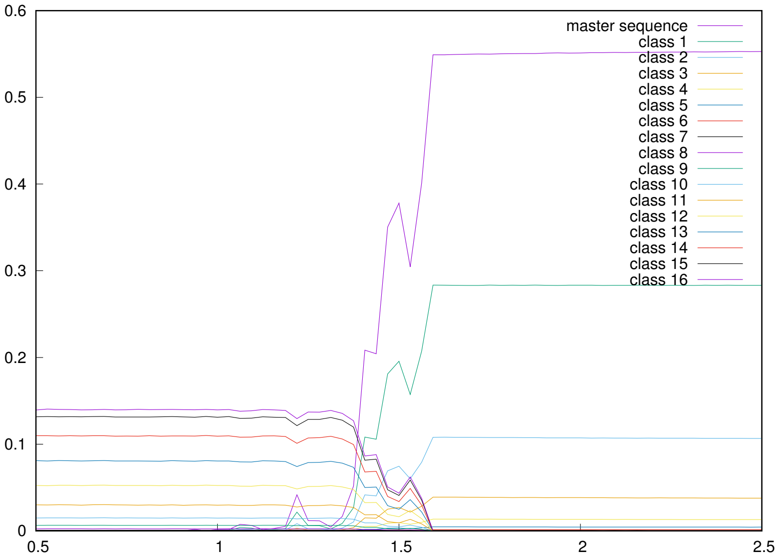

Figure 5: Density of the first mutant classes as a function of m/ℓ, for the Wright–Fisher model with parameters $\sigma = 2$, $\ell = 32$, $a = \ell q = 0.5$

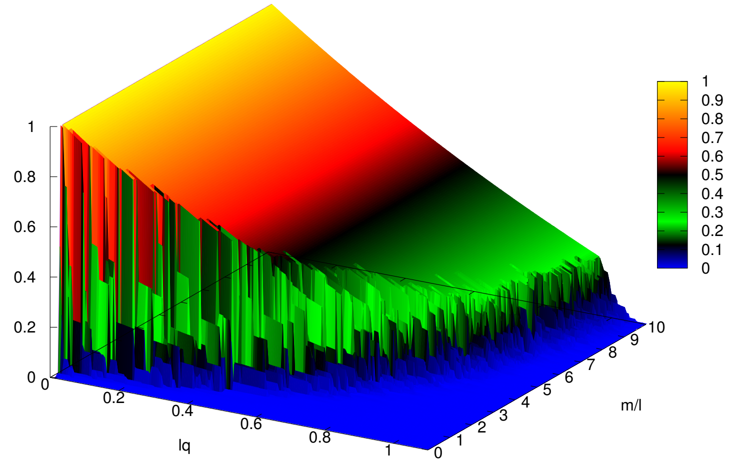

Figure 6: Density of the wild type as a function of ℓq and m/ℓ, for the Wright–Fisher model with parameters $\sigma = 2$, $\ell = 24$, q, m varying

within a reasonable amount of time. Typically, in a simulation of the model with parameters ℓ, m, the number of generations is taken to be $100\,000 \times 2^{\max(\ell,m)}$ multiplied by a factor between 1 and 100. The good news is that, already for small values of ℓ, the simulations are very conclusive.

Heuristics. The proof of the theorem is rather long. However the heuristics behind it are quite simple. In the finite population model, the number of copies of the master sequence fluctuates with time. Suppose that the process starts with a population of size m containing exactly one master sequence. The master sequence is likely to invade the whole population and become dominant. Then the master sequence will be present in the population for a very long time without interruption. We call this time the **persistence** time of the master sequence. The destruction of all the master sequences of the population is quite unlikely, nevertheless it will happen and the process will eventually land in the neutral region consisting of the populations devoid of master sequences. The process will wander randomly throughout this region for a very long time. We call this time the **discovery** time of the master sequence. Because the cardinality of the possible genotypes is enormous, the master sequence is difficult to discover, nevertheless the mutations will eventually succeed and the process will start again with a population containing exactly one master sequence. If, on average, the discovery time is much larger than the persistence time, then, by the ergodic theorem, the equilibrium state will be totally random, while a quasispecies will be formed if the persistence time is much larger than the discovery time. The crucial problem is to estimate the persistence time and the discovery time of the master sequence. For the persistence time, we rely on a classical computation from mathematical genetics. Suppose we start with a population containing $m-1$ copies of the master sequence and another non master sequence. The non master sequence is very unlikely to invade the whole population, yet it has a small probability to do so, called the fixation probability. If we neglect the mutations, standard computations yield that, in a population of size m, if the master sequence has a selective advantage of $\sigma > 1$, the fixation probability of the non master sequence is roughly of order $1/\sigma^m$. Now the persistence time can be viewed as the time needed for non master sequences to invade the population. This time is approximately equal to the inverse of the fixation probability of the non master sequence, that is of order σ^m. For the discovery time, there is no miracle: before discovering the master sequence, the process is likely to explore a significant portion of the genotype space, hence the discovery time should be of order 4^ℓ. These simple heuristics indicate that the persistence time depends on the selection drift, while the discovery time depends on the spatial entropy. Suppose that we send m, ℓ to ∞ simultaneously. If the discovery time is much larger than the persistence time, then the population will be neutral most of the time and the fraction of the master sequence at equilibrium will be null. If the persistence time is much larger than the discovery time, then the population will be invaded by the master sequence most of the time and the fraction of the master sequence at equilibrium will be positive. This leads to an interesting feature, namely the existence of a critical population size for the emergence of a quasispecies. For chromosomes of length ℓ, a quasispecies can be formed only if the population size m is such that the ratio m/ℓ is large enough. In order to go further, we must put the heuristics on a firmer ground and we should take the mutations into account when estimating the persistence time. The main problem is to obtain finer estimates on the persistence and discovery times. We cannot compute explicitly the laws of these random times, so we will compare the Wright–Fisher model with simpler processes. In the non neutral populations, we shall compare the process with a jump process $(Z_n)_{n\geq 0}$ on $\{\,0,\dots,m\,\}$, which approximates the number of copies of the master sequence present in the population. (In the cas of the Moran model, the process $(Z_n)_{n\geq 0}$ on $\{\,0,\dots,m\,\}$ was already introduced by Nowak and Schuster[14]). We analyze the dynamics of this process with the help of the Freidlin–Wentzell theory of random perturbations of dynamical systems. We obtain that

$$\textbf{persistence time} \sim \exp\big(m\phi(a)\big)\,.$$

In the neutral populations, we shall replace the process with a random walk on $\mathcal{A}^\ell$, or equivalently an Ehrenfest process $(Y_n)_{n\geq 0}$ on $\{\,0,\dots,\ell\,\}$. The value Y_n represents the distance of the walker to the master sequence. A celebrated theorem of Kac[15] from 1947, which helped to resolve a famous paradox of statistical mechanics, yields that

$$\textbf{discovery time} \sim 4^{\ell}\,.$$

The critical curve is then obtained by equating the persistence time and the discovery time. It is certainly well known that the population dynamics depends on the population size[9]. Van Nimwegen, Crutchfield and Huynen[16] show that an important parameter is the product of the population size and the mutation rate. The nature of the dynamics changes radically depending on whether this product is small or large[17,18]. Van Nimwegen and Crutchfield[19] observe and discuss the transition from the quasispecies regime for large populations to the disordered regime for small populations. **Conclusion.** We have shown rigorously the existence of an error threshold phenomenon for the sharp peak landscape and the Wright–Fisher model. We have derived an explicit formula for the distribution of the quasispecies. Compared to the Eigen model, a new phenomenon occurs for a stochastic population model with finite size: to ensure the stability of the quasispecies, the population size has to be at least of order of the inverse of the genome length. Thus, even in the very simple framework of the Wright–Fisher model on the sharp peak landscape, cooperation is necessary to achieve the survival of the wild type.

References

1. Eigen, M. Self-organization of matter and the evolution of biological macromolecules. *Naturwissenschaften* **58**(10), 465–523 (1971).

2. Eigen, M., McCaskill, J., and Schuster, P. The molecular quasi-species. *Advances in Chemical Physics* **75**, 149–263 (1989).

3. Domingo, E. Quasispecies theory in virology. *Journal of Virology* **76**(1), 463–465 (2002).

4. Domingo, E., Biebricher, C., Eigen, M., and Holland, J. J. *Quasispecies and RNA virus evolution: principles and consequences.* Landes Bioscience, Austin, Tex., (2001).

5. Tripathi, K., Balagam, R., Vishnoi, N. K., and Dixit, N. M. Stochastic simulations suggest that hiv-1 survives close to its error threshold. *PLOS Computational Biology* **8**(9), 1–14 (2012).

6. Anderson, J. P., Daifuku, R., and Loeb, L. A. Viral error catastrophe by mutagenic nucleosides. *Annual Review of Microbiology* **58(1)**, 183205 (2004).

7. Crotty, S., Cameron, C. E., and Andino, R. RNA virus error catastrophe: Direct molecular test by using ribavirin. *Proceedings of the National Academy of Sciences* **98(12)**, 68956900 (2001).

8. Kimura, M. *The Neutral Theory of Molecular Evolution.* Cambridge University Press, (1985 (reprint)).

9. Wilke, C. Quasispecies theory in the context of population genetics. *BMC Evolutionary Biology* **5**, 1–8 (2005).

10. Cerf, R. Critical population and error threshold on the sharp peak landscape for the Wright-Fisher model. *Ann. Appl. Probab.* **25**(4), 1936–1992 (2015).

11. Cerf, R. Critical population and error threshold on the sharp peak landscape for a Moran model. *Mem. Amer. Math. Soc.* **233**(1096), vi+87 (2015).

12. Cerf, R. and Dalmau, J. The distribution of the quasispecies for a Moran model on the sharp peak landscape. *Stochastic Process. Appl.* **126**(6), 1681–1709 (2016).

13. Dalmau, J. The distribution of the quasispecies for the Wright-Fisher model on the sharp peak landscape. *Stochastic Process. Appl.* **125**(1), 272–293 (2015).

14. Nowak, M. A. and Schuster, P. Error thresholds of replication in finite populations. Mutation frequencies and the onset of Muller's ratchet. *Journal of theoretical Biology* **137 (4)**, 375–395 (1989).

15. Kac, M. Random walk and the theory of brownian motion. *American Mathematical Monthly* **54**(7), 369–391 (1947).

16. Nimwegen, E. V., Crutchfield, J. P., and Huynen, M. Neutral evolution of mutational robustness. *Proc. Natl . Acad. Sci . USA* **96**, 9716–9720 (1999).

17. Sumedha, Martin, O. C., and Peliti, L. Population size effects in evolutionary dynamics on neutral networks and toy landscapes. *Journal of Statistical Mechanics: Theory and Experiment* **05**, P05011 (2007).

18. Elena, S. F., Wilke, C. O., Ofria, C., and Lenski, R. E. Effects of population size and mutation rate on the evolution of mutational robustness. *Evolution* **61(3)**, 666–74 (2007).

19. van Nimwegen, E. and Crutchfield, J. Metastable evolutionary dynamics: Crossing fitness barriers or escaping via neutral paths? *Bulletin of Mathematical Biology* **62**, 799–848 (2000).