Mutation and selection in age-structured populations

Evans, Steven N.
University of California
Department of Statistics
367 Evans Hall
Berkeley, CA 94720-3860, U.S.A.
evans@stat.berkeley.edu

Steinsaltz, David
University of Oxford
Department of Statistics
1 South Parks Road
Oxford, OX1 3TG, UNITED KINGDOM

Wachter, Kenneth W.
University of California
Department of Demography
2232 Piedmont Avenue
Berkeley, CA 94720-2120, U.S.A.
wachter@demog.berkeley.edu

1 Introduction

The progressive physical deterioration to which members of many species are subject with advancing age has long seemed to present a challenge to intuitions of natural selection. "It is indeed remarkable that after a seemingly miraculous feat of morphogenesis a complex metazoan should be unable to perform the much simpler task of merely maintaining what is already formed," wrote biologist G. Williams [11]. The modern theory of the evolution of ageing, to which Williams contributed significant insights, is generally considered to have begun with Peter Medawar's 1951 inaugural lecture at University College London delivered in 1951 [7], in which he called the origin of senescence "an unsolved problem in biology." His idea, called "mutation accumulation", presents senescence as a side-effect of the interplay between natural selection and mutation, a special version of his more general concept of "genetic load". Ongoing random mutation spews mostly deleterious changes into the genome. Since the only genetic "repair mechanism" is the death of the organism carrying the defect, there is perpetually an overhang of deaths not yet realised, stretching from the time of the initial mutation until all descendants have died from the effect of the allele. The more harmful a mutation (measured in reproductive success), the more rapidly appearances

Under the (admittedly naïve) assumption that there is a large class of potential mutations all recurring at identical rates, the prevalence of a given mutant allele in the population will be expected to track inversely the selective cost to individuals who carry that allele. The overall phenotypic impact of these accumulated mutant alleles will then reflect the disjunction between the impact we measure and the impact registered by natural selection. The application to senescence is immediate: If mutant alleles impose equivalent mortality cost, the one whose mortality tends to come earlier in life will be more costly from the perspective of natural selection. Carrying an allele that makes you prone to childhood leukaemia reduces your expected number of offspring more than would an increased susceptibility to Alzheimer's disease or lung cancer. There is little selective cost to being prone to a disease long after you have completed raising your children to adulthood, and quite likely have died

of some other cause. The cumulative effect on the age-pattern of morbidity and mortality should be broadly like what we observe: Increasing risk for all manner of ailments with increasing age, and a progressive failure of repair processes.

It has not been obvious how best to turn Medawar's conceptual model into something quantitative. An early quantitative model for mutation and selection is the well-known work of M. Kimura and T. Maruyama [6]. Our demographic formulations grow out of the work of Brian Charlesworth [2], [3]. This work has focused, for mainly technical reasons, on settings which are highly stylised in the following respects:

- The mortality impacts of a mutant allele is either focused on a single age, or on a simple range around a focal age.
- The mutations are uniformly distributed.
- The selective costs of distinct mutant alleles act additively.
- The mortality effects of distinct mutant alleles act additively.

The last two conditions are a crude mathematical version of the biological condition of *non-epistasis*, or non-interaction between the effects of different mutations. It is important to notice that there is a contradiction between these two conditions: As discussed in [9] and [4], the most plausible formula for the selective cost of carrying a collection of mutant alleles g is the change in the net reproductive ratio (NRR):

$$S(g) := \int f_x \, l_x(0) dx - \int f_x \, l_x(g) dx \tag{1}$$

The integrals are all taken over ages for which the integrands are non-zero. The function $l_x(g)$, the probability of survival to age x for members with genotype g, is derived from hazards that include increments from mutations in g. If g is comprised of components that increase mortality additively, the selective costs will clearly not be additive. This reflects the intuitive argument of Medawar: A mutant allele that kills an individual at age 90 becomes less selectively costly in the presence of another allele that might kill him at age 80.

More generally, since the disjunction between selection impacts and observable phenotypic effects on mortality is the crux of the Medawar theory, we should not wish to constrain the model in exactly this respect for the sake of mere mathematical convenience.

2 Informal description of the model

We repeat here the general description of the model that may be found in greater detail in [4].

The model has four fundamental ingredients:

- a complete, separable metric space $\mathcal{M}$ of loci;
- a finite Borel measure ν on $\mathcal{M}$ called the mutation measure because it describes the rate at which mutations occur in regions of the genome;
- the space $\mathcal{G}$ of integer–valued finite Borel measures on $\mathcal{M}$
- the selective cost function $S : \mathcal{G} \to \mathbb{R}_+$.

$\mathcal{M}$ is the collection of *loci* in the portion of the genome that is of interest to us. There is a distinguished reference wild type genotype, and each locus represents a "position" at which the

genotype of an individual may differ from that of the wild genotype. We allow the set $\mathcal{M}$ to be quite general and do not necessarily think of it as a finite collection of physical DNA base positions or a finite collection of genes. For example, the proposed explanation for the Gompertz mortality curve and mortality plateaus at extreme ages in Charlesworth [3] suggest taking $\mathcal{M}$ to be a class of functions from $\mathbb{R}_+$ to $\mathbb{R}_+$: the value of such a function at age $x \geq 0$ represents an additional increment to mortality at age x conferred by a mutation away from the wild type at this locus.

An element of $\mathcal{G}$ represents a "genotype", regarded as the set of loci at which there have been ancestral mutations away from the reference wild type. The null measure represents the wild genotype.

The state of the population at time $t \geq 0$ in our model is a probability measure P_t on $\mathcal{G}$. In the version with recombination this distribution will be a Poisson random measure, which is determined by its intensity measure, in general a locally-finite Borel measure on $\mathcal{M}$ and which is a finite measure when the mutation measure ν has finite total mass. For most results, we do assume that ν has finite total mass.

We specify the fitnesses of different genotypes by a *selective cost function* $S : \mathcal{G} \to \mathbb{R}_+$. The difference $S(g') - S(g'')$ for $g', g'' \in \mathcal{G}$ is the difference in the rate of sub-population growth between the sub-population of individuals with genotype g'' and the sub-population of individuals with genotype g'. We make the normalizing assumption $S(0) = 0$ and suppose that

$$S(g+h) \geq S(h), \quad g, h \in \mathcal{G}, \tag{2}$$

in line with our assumption above that genotypes with more accumulated mutations are less fit.

The genotype of an individual is specified by the set of loci at which there has been a mutation somewhere along the ancestral lineage leading to that individual. More precisely, a genotype is an element of the space $\mathcal{G}$ of integer–valued finite Borel measures on $\mathcal{M}$. An element of $\mathcal{G}$ is a finite sum $\sum_i \delta_{m_i}$, where δ_m is the unit point mass at the locus $m \in \mathcal{M}$. The measure $\sum_i \delta_{m_i}$ corresponds to a genotype that has ancestral mutations at loci $m_1, m_2, \ldots$. The wild type genotype is thus the null measure. We do not require that the loci $m_i \in \mathcal{M}$ be distinct. We thus allow several copies of a mutation. This is reasonable, since we are not identifying mutations with changes in nucleotide sequences in a one-to-one manner.

For example, if $\mathcal{M} = \{1, 2, \ldots, N\}$ for some positive integer N, then $\mathcal{G}$ is essentially the Cartesian product $\mathbb{N}^N$ of N copies of the nonnegative integers: A genotype is of the form $\sum_{j=1}^N n_j \delta_j$, indicating that an ancestral mutation is present n_j times at locus j, and we identify such a genotype with the nonnegative integer vector $(n_1, n_2, \ldots, n_N)$.

Since the population is infinite, and all that matters about an individual is the individual's genotype, the dynamics of the population are described by the proportions of individuals with genotypes that belong to the various subsets of $\mathcal{G}$. We are thus led to consider a family of probability measures P_t, $t \geq 0$, on $\mathcal{G}$, where $P_t(G)$ for some subset $G \subseteq \mathcal{G}$ represents the proportion of individuals in the population at time t that have genotypes belonging to G. Note that $P_t(G)$ also may be thought of as the probability that an individual chosen uniformly at random from the population will have genotype belonging to the set G. In other words, P_t is the distribution of a random finite integer-valued measure on $\mathcal{M}$. For example, if $\mathcal{M} = \{1, 2, \ldots, N\}$ and we identify $\mathcal{G}$ with the Cartesian product $\mathbb{N}^N$ as above, then $P_t(\{(n_1, n_2, \ldots, n_N)\})$ represents the probability that an individual chosen uniformly at random from the population will have n_j ancestral mutations at locus j for $j = 1, 2, \ldots, N$.

The evolutionary process P_t may be motivated as the limit of a discrete-time process in the space of probability measures on $\mathcal{G}$, progressing in time intervals $1/n$ (where n is a scaling constant which will be taken to ∞ in the limit) by iterating some combination of Mutation, Selection, and Recombination. Starting from a probability distribution Q on $\mathcal{G}$, a single step of these transformations works as follows:

Mutation A randomly chosen genotype from Q is added to a randomly chosen element from the mutation distribution, which is a Poisson random measure with intensity ν/n.

Selection Q is reweighted in proportion to the weight function $e^{-S(g)/n}$.

Recombination A random subset $A \subset \mathcal{M}$ is chosen, according to a probability measure $\mathcal{R}$ on $\mathcal{B}(\mathcal{M})$ (the Borel sets on $\mathcal{M}$), and two independent random "parents" are selected. The offspring receives the A portion of one parent's genotype, and the $A^{\complement}$ portion of the other parent's genotype. That is, g and g' are selected independently from Q. The offspring genotype is then $g_A + g'_{A^{\complement}}$.

3 Main results

3.1 Mutation and selection

If there is mutation and selection, but no recombination, then the limiting evolution equation for P_t is

$$\frac{d}{dt}P_t\Phi = P_t\left(\int_{\mathcal{M}} [\Phi(\cdot+\delta_m)-\Phi(\cdot)]\ \nu(dm)\right) - P_t(\Phi\cdot[S-P_tS]). \tag{3}$$

This is the model introduced and analyzed at length in [8] using the Feynman-Kac formula. When $\mathcal{M}=\{1,2,\ldots,N\}$, (3) is a special case of the classical system of ordinary differential equations for mutation and selection in continuous-time – see Section III.1.2 of [1] for the derivation of an analytic solution that agrees with the one that arises from a Feynman-Kac analysis.

Two principal results from [8] are the following:

Theorem 3.1 *Let $\tilde{\Pi}$ be the Poisson random measure on $\mathcal{M}\times\mathbb{R}_+$ with intensity measure $\nu\otimes$Lebesgue, and $Z_t := \int_{\mathcal{M}\times[0,t]} \delta_m\, d\tilde{\Pi}((m,u))$. Suppose that there is a positive T such that $\exp\left(-\int_0^t S(Z_u)du\right) S(Z_t)$ has finite expectation for all $t\in[0,T)$. Then Equation (3) has a solution on $[0,T)$, given by*

$$P_t\Phi = \frac{\mathbb{E}\left[\exp\left(-\int_0^t S(Z_u)\,du\right)\Phi(Z_t)\right]}{\mathbb{E}\left[\exp\left(-\int_0^t S(Z_u)\,du\right)\right]}. \tag{4}$$

We order the points put down by $\tilde{\Pi}$ in order of their arrival times $\tau(1),\tau(2),\ldots$ and (solely in the context of this section) write $Y_n := Z_{\tau(n)}$.

Theorem 3.2 *Suppose $\nu(\mathcal{M})<\infty$ and P_0 puts unit mass at the null state 0, with $S(0)=0$. Then the solution (4) may be written as $P_t\Phi = \tilde{P}_t\Phi/\tilde{P}_t\mathbf{1}$, with $J_n = \mathbf{1}_{g(\mathcal{M})=n}$, the indicator function of the set with $g(\mathcal{M})=n$:*

$$\tilde{P}_tJ_n\Phi = \nu(\mathcal{M})^n e^{-\nu(\mathcal{M})t}\mathbb{E}\left[(S(Y_1)\ldots S(Y_n))^{-1}\,H_{t,n}\Phi(Y_n)\right]. \tag{5}$$

Here $H_{t,n}$ is a conditional probability defined in terms of independent unit-rate exponential variables $Z_1, Z_2, \ldots$ by the formula

$$H_{t,n} = \mathbb{P}\left\{\sum Z_j/S(Y_j) < t \quad | Y_1,\ldots Y_n\right\} \tag{6}$$

If $\sum \nu(\mathcal{M})^n\mathbb{E}[((S(Y_1)\ldots S(Y_n))^{-1}]$ is finite, P_t converges in distribution as t goes to infinity. If the sum is infinite, P_tJ_n goes to zero for all n.

Notice that a finite-population model without recombination along the same lines would be vulnerable to Mueller's ratchet, the process in which the fittest classes of genotypes in an asexual population can be successively lost through drift, carrying the population to extinction. A thorough discussion of the ratchet and the associated advantages of sex and recombination with references is found in [1], pages 303–308. In the face of Mueller's ratchet, a finite-population version of the model in [8] would not be viable. This observation underscores the importance of incorporating recombination, which avoids the ratchet by allowing fittest classes to be reconstituted in every generation.

3.2 Adding recombination

If P_0 is the distribution of a Poisson random measure, then mutation preserves this property. On the other hand, epistatic selection drives the population distribution away from Poisson, while increasing rates of recombination push it towards Poisson. Thus, when all three processes operate and we consider a limiting regime where recombination acts on a much faster time scale than selection and recombination, we expect asymptotically that if the initial condition P_0 is the distribution of a Poisson random measure on $\mathcal{M}$, then P_t will also be the distribution of a Poisson random measure for all $t > 0$.

Of course, in order that recombination eventually will poissonise any initial distribution, the segregating sets A must be sufficiently generic. More specifically, it must be the case that there is a positive probability that the segregating set and its complement will both intersect a generic non-null set in sets that each have positive mass. The following condition (with $\lambda = \mu P$) will be key to establishing establishing quantitative bounds on the rate with which recombination applied to the distribution P converges to $\Pi_{\mu P}$:

Given a (symmetric) recombination measure $\mathcal{R}$ and a finite measure λ on $\mathcal{M}$, we say that the pair $(\mathcal{R}, \lambda)$ is *shattering* if there is a positive constant α such that for any Borel set A,

$$\begin{aligned} \lambda(A)^3 &\leq \alpha \left[\lambda(A)^2 - 2 \int \lambda(A \cap R)^2 \, d\mathcal{R}(R) \right] \\ &= 2\alpha \int \lambda(A \cap R) \lambda(A \cap R^c) \, d\mathcal{R}(R) \end{aligned} \tag{7}$$

Supposing this is true, we may think of P_t as a process ρ_t of Poisson intensities, which are finite measures on the space $\mathcal{M}$ of loci. We write X^π for a Poisson random measure on $\mathcal{M}$ with intensity measure π. We expect then that ρ_t should satisfy the evolution equation

$$\rho_t(dm) = \rho_0(dm) + t\,\nu(dm) - \int_0^t \mathbb{E}\left[S(X^{\rho_s} + \delta_m) - S(X^{\rho_s}) \right] \, \rho_s(dm) \, ds. \tag{8}$$

We define the rigorous counterpart of (8) in [4], and establish the existence and uniqueness of solutions.

Theorem 3.3 *Fix a mutation measure $\nu \in \mathcal{H}^+$, the nonnegative subset of the Banach space of finite signed Borel measures on $\mathcal{M}$ equipped with the norm Wasserstein norm. Let $S : \mathcal{G} \to \mathbb{R}_+$ be a selective cost function that satisfies the conditions*

- *$S(0) = 0$,*
- *$S(g) \leq S(g + h)$ for all $g, h \in \mathcal{G}$,*
- *for some constant K, $\left|S(g) - S(h)\right| \leq K \left\|g - h\right\|_{\text{Was}}$, for all $g, h \in \mathcal{G}$.*

Then equation (8) has a unique solution for any $\rho_0 \in \mathcal{H}^+$.

Furthermore, we show there that our dynamical equation is a limit of a sequence of standard discrete generation, mutation-selection-recombination models. We define Q_k to be the result of applying k rounds of mutation, selection, and recombination to an initial distribution Q_0 on $\mathcal{G}$.

Theorem 3.4 *Let $(\rho_t)_{t \geq 0}$ be the measure-valued dynamical system of (8) whose existence is guaranteed by Theorem 3.3. Suppose that the selective cost function S satisfies the hypotheses of Theorem 3.3, and in addition that the pairs $(\mathcal{R}, \nu)$ and $(\mathcal{R}, \rho_0)$ consisting of the recombination measure and the mutation measure are shattering, and the initial measure Q_0 is equivalent to its Poissonization $P_0 := \Pi_{\rho_0}$, with $\log dQ_0/dP_0$ being Lipschitz.*

Then, for any $T > \epsilon > 0$, letting Π_ρ be the Poisson random measure with intensity ρ,

$$\lim_{n\to\infty} \sup_{\epsilon\le t\le T} \left\|\Pi_{\rho_t} - Q_{\lfloor tn\rfloor}\right\|_{\mathrm{Was}} = 0.$$

If, in addition, the initial measure $Q_0 = P_0$ is Poisson, then this equation holds for $\epsilon = 0$.

4 Haldane's Principle

Given that it involves computing an expected value for a quite general Poisson process, equation (8) may look rather forbidding. However, for certain reasonable choices of selective costs the integral can be evaluated explicitly, leading to a simpler and more intuitive system. Some such examples may be found in [4], [9], and [10].

There are also interesting general conclusions that may be drawn. We mention here just one, reproducing in brief a more lengthy discussion in [9].

Working with genetic models without age structure, [5], page 341, announced that "... the loss of fitness to the species depends entirely on the mutation rate and not at all on the effect of the gene upon fitness of the individual carrying it ..." He found the sum total of mutation rates at different sites to be approximately equal to the resulting decrement in the logarithm of fitness, a measure of "genetic load" essentially equivalent to our selective cost.

Can this principle be generalised to the setting of age-specific mortality effects? The simplest version of our model which might be relevant would have mutations imposing point-mass increments to mortality at fixed ages, with the effect at age a tuned by a parameter $\eta(a)$. It is shown that if an equilibrium exists, the aggregate population hazard maintained by mutation-selection balance at equilibrium does not depend on the sizes $\eta(a)$ of the mutational effects. The aggregate population hazard, we note, is lower at advanced ages than the expected hazard, since the aggregate hazard for the population at age x averages only those individuals who have actually survived to age x. Survivors to advanced age carry, on average, lower mutational loads than a typical member of the population.

This result generalises a property of linear approximate models, but in a surprising direction. Scaling up the size of the effect of a mutation increases the selective pressure against it and reduces its expected frequency. In the linear setting, the expected hazard, that is, the average hazard averaged across the population, is insensitive to $\eta(a)$. Doubling $\eta(a)$ halves the expected frequency $\rho(a)$ and leaves the expected hazard $\eta(a)\rho(a)$ unchanged. In our full non-linear setting, it is not the expected hazard but the aggregate population hazard that comes out to be invariant to changes in $\eta(a)$.

REFERENCES (RÉFERENCES)

[1] Reinhard Bürger. *The mathematical theory of selection, recombination, and mutation.* John Wiley, Chichester, New York, 2000.

[2] Brian Charlesworth. *Evolution in age-structured populations.* Cambridge University Press, Cambridge, 1994.

[3] Brian Charlesworth. Patterns of age-specific means and genetic variances of mortality rates predicted by the mutation-accumulation theory of ageing. *J. Theor. Bio.*, 210(1):47–65, 2001.

[4] Steven N. Evans, David Steinsaltz, and Kenneth W. Wachter. A mutation-selection model with recombination for general genotypes. Available at arXiv:q-bio/0609046., 2011.

[5] J. B. S. Haldane. The effect of variation on fitness. *American Naturalist*, 71:337–49, 1937.

[6] Motoo Kimura and Takeo Maruyama. The mutational load with epistatic gene interactions in fitness. *Genetics*, 54:1337–1351, 1966.

[7] Peter Medawar. *An unsolved problem in biology: An inaugural lecture delivered at University College, London, 6 December, 1951.* H. K. Lewis and Co., London, 1952.

[8] David Steinsaltz, Steven N. Evans, and Kenneth W. Wachter. A generalized model of mutation-selection balance with applications to aging. *Adv. Appl. Math.*, 35(1):16–33, 2005.

[9] Kenneth W. Wachter, Steven N. Evans, and David R. Steinsaltz. The age-specific forces of natural selection and walls of death. Technical Report 757, Department of Statistics, University of California at Berkeley, 2008. Available at `http://arxiv.org/abs/0807.0483`.

[10] Kenneth W. Wachter, David R. Steinsaltz, and Steven N. Evans. Vital rates from the action of mutation accumulation. *Journal of Population Ageing*, page Currently only published online, 2010.

[11] George C. Williams. Pleiotropy, natural selection, and the evolution of senescence. *Evolution*, 11:398–411, December 1957.

Acknowledgements – SNE supported in part by grants DMS-04-05778 and DMS-09-07630 from the National Science Foundation (U.S.A.). DRS supported by grant K12-AG00981 from the National Institute on Aging (U.S.A.) and a Discovery Grant from the National Science and Engineering Research Council (Canada). KWW supported by grant P01-008454 from the National Institute on Aging (U.S.A.) and by the Miller Institute for Basic Research in Science at U.C. Berkeley.