Supporting Information

C.F. Mugal, J.B.W. Wolf, I. Kaj

Why time matters: Codon evolution and the temporal dynamics of dN/dS

Basic properties of the Wright-Fisher diffusion process

Let $(W_t)_{t\geq 0}$ be a standard Brownian motion. The Markov diffusion process $(\xi_t)_{t\geq 0}$ defined as the unique strong solution of the stochastic differential equation

$$d\xi_t = \mu(\xi_t)\,dt + \sigma(\xi_t)\,dW_t, \quad t \geq 0,$$

starting at some point $\xi_0 = x$ of the open interval $(0,1)$ and with drift function $\mu(x)$ and variance function $\sigma^2(x)$ given by

$$\mu(x) = \gamma x(1-x), \quad \sigma^2(x) = x(1-x),$$

is known as the Wright-Fisher diffusion process with selection. This process arises in the limit of weak convergence as the diffusion approximation under time and space rescaling of the Wright-Fisher Markov model of selective reproduction, as the population size tends to infinity. We write $\mathbb{P}_x^\gamma$ for the probability measure and $\mathbb{E}_x^\gamma$ for the expectation of the process starting at x. The case $\gamma = 0$ is the neutral Wright-Fisher diffusion process, $\gamma > 0$ corresponds to positive selection and $\gamma < 0$ to negative selection. This and other versions of the Wright-Fisher diffusion process have become standard models in population genetics. See e.g. Karlin and Taylor (1981); Breiman (1992); Ewens (2004); Etheridge (2011) for detailed accounts of their mathematical properties. In the following we compile a listing of various properties and formulae for the Wright-Fisher diffusion with selection, which were used to obtain the results in the present article. These results are covered by the works already cited, with some exceptions for which we give additional references in what follows.

The scale function $S_\gamma(x)$ and speed function $m_\gamma(x)$ associated with the Wright-Fisher diffusion with selection parameter γ are defined by

$$S_0(x) = x, \quad S_\gamma(x) = \frac{1}{2\gamma}(1 - e^{-2\gamma x}), \quad \gamma \neq 0, \qquad m_\gamma(x) = \frac{e^{2\gamma x}}{x(1-x)}.$$

Since m_γ is not integrable near 0 or 1 it follows that both points $\{0,1\}$ are classified as exit boundary points, which are therefore accessible from the interior of the state space. Hence the diffusion can reach either of these boundaries but will stay at the point reached first. We call the time τ_0 required to reach 0 the extinction time and the time τ_1 to reach 1 the fixation time and denote by $\tau = \min(\tau_0, \tau_1)$ the resulting time to absorption. The corresponding exit measure is the fixation probability

$$q_\gamma(x) = \mathbb{P}_x^\gamma(\tau_1 < \tau_0) = \frac{1 - e^{-2\gamma x}}{1 - e^{-2\gamma}}, \quad \gamma \neq 0, \qquad q_0(x) = x,$$

where $\xi_0 = x$ is the initial state. By integration with respect to the Green function $G(x,y)$ defined by

$$G(x,y) = \begin{cases} 2q_\gamma(x)(S(1) - S(y))m(y), & 0 \leq x \leq y \leq 1 \\ 2(1 - q_\gamma(x))(S(y) - S(0))m(y), & 0 \leq y \leq x \leq 1, \end{cases}$$

it is possible to compute functionals of the form

$$\mathbb{E}_x^\gamma\Big[\int_0^\tau g(\xi_t)\,dt\Big] = \int_0^1 G(x,y)g(y)\,dy.$$

Hence

$$\mathbb{E}_x^\gamma\Big[\int_0^\tau g(\xi_t)\,dt\Big] = \frac{1-e^{-2\gamma x}}{1-e^{-2\gamma}}\int_x^1 \frac{1-e^{-2\gamma(1-y)}}{\gamma y(1-y)}g(y)\,dy + \frac{e^{-2\gamma x}-e^{-2\gamma}}{1-e^{-2\gamma}}\int_0^x \frac{e^{2\gamma y}-1}{\gamma y(1-y)}g(y)\,dy, \quad (1)$$

whenever the integrals on the right hand side are well-defined. In particular, for $x = 1/N$ and N large, if we first apply

$$q_\gamma(1/N) = \frac{1-e^{-2\gamma/N}}{1-e^{-2\gamma}} \sim \frac{\omega_\gamma}{N}, \quad \omega_\gamma = \frac{2\gamma}{1-e^{-2\gamma}}\ \gamma \neq 0, \quad \omega_0 = 1, \quad (2)$$

then after suitable integral approximations it follows for the case $g(y) = 1$ that

$$\mathbb{E}_{1/N}^\gamma[\tau] \sim \frac{2\ln N + 2}{N} + \frac{1}{N}\Big(2\gamma - \frac{\gamma^3}{9}\Big). \quad (3)$$

Similarly, for the case $g(y) = y$ we have the large N approximation

$$\mathbb{E}_{1/N}^\gamma\Big[\int_0^\tau \xi_t\,dt\Big] \approx \frac{\omega_\gamma}{N}\int_0^1 \frac{1-e^{-2\gamma(1-y)}}{\gamma(1-y)}\,dy \approx \frac{\omega_\gamma}{N}\frac{e^{-2\gamma}-1+2\gamma+2\gamma^2}{2\gamma^2}, \quad (4)$$

and for $g(y) = y(1-y)$ an approximation based on (1) yields

$$\mathbb{E}_{1/N}^\gamma\Big[\int_0^\tau \xi_t(1-\xi_t)\,dt\Big] \approx \frac{\omega_\gamma}{N}\int_0^1 \frac{1-e^{-2\gamma(1-y)}}{\gamma}\,dy \approx \frac{\omega_\gamma}{N}\frac{e^{-2\gamma}-1+2\gamma}{2\gamma^2}. \quad (5)$$

The process conditioned on fixation. Let $\mathbb{P}_x^{*\gamma}$ and $\mathbb{E}_x^{*\gamma}$ be the distribution and expectation of the Wright-Fisher diffusion conditioned on the event of fixation, $\tau_1 < \infty$. Then

$$\mathbb{P}_x^\gamma(\tau < t) = \mathbb{P}_x^\gamma(\tau_1 < t) + \mathbb{P}_x^\gamma(\tau_0 < t) = \mathbb{P}_x^{*\gamma}(\tau_1 < t)q_\gamma(x) + \mathbb{P}_x^{*\gamma}(\tau_0 < t)(1-q_\gamma(x)). \quad (6)$$

The drift and variance functions for the conditioned process are

$$\mu^*(x) = \mu(x) + \frac{s(x)}{S(x)}\sigma^2(x) = \gamma x(1-x)\frac{1+e^{-2\gamma x}}{1-e^{-2\gamma x}}$$

and

$$\sigma^{2*}(x) = \sigma^2(x) = x(1-x).$$

Again one can find the speed and scale functions as well as the Green function, which gives in particular

$$\mathbb{E}_x^{*\gamma}(\tau_1) = \frac{(e^{-2\gamma x}-e^{-2\gamma})}{\gamma(1-e^{-2\gamma x})(1-e^{-2\gamma})}\int_0^x \frac{(1-e^{-2\gamma y})^2}{y(1-y)e^{-2\gamma y}}\,dy + \int_x^1 \frac{1-e^{-2\gamma(1-y)}}{\gamma y(1-y)}\frac{1-e^{-2\gamma y}}{1-e^{-2\gamma}}\,dy,$$

see Karlin and Taylor (1981), Ch 15, (9.9).

For $x = 1/N$, the first term on the right side is of the order $2/N$ and vanishes asymptotically in comparison with the second term, which yields the approximation

$$\mathbb{E}_{1/N}^{*\gamma}(\tau_1) \approx \int_0^1 \frac{(1-e^{-2\gamma y})(1-e^{-2\gamma(1-y)})}{y(1-y)\gamma(1-e^{-2\gamma})}\,dy = 2 - \frac{1}{9}\gamma^2 + \frac{7}{675}\gamma^4 + O(\gamma^6). \quad (7)$$

The neutral model. For the neutral Wright-Fisher model with $\gamma = 0$ one has $q_0(x) = x$,

$$\mathbb{E}_x^0\Big[\int_0^\tau g(\xi_t)\,dt\Big] = x\int_x^1 \frac{2}{y}g(y)\,dy + (1-x)\int_0^x \frac{2}{1-y}g(y)\,dy, \tag{8}$$

and

$$\mathbb{E}_x^0(\tau) = -2x\ln x - 2(1-x)\ln(1-x), \tag{9}$$

in agreement with (3). Conditional on fixation in 1 the process satisfies the stochastic differential equation

$$\xi_t = \xi_0 + \int_0^t (1-\xi_s)\,ds + \int_0^t \sqrt{\xi_s(1-\xi_s)}\,dW_s.$$

Since the Ito integral has expectation zero the expected value $m_t(x) = \mathbb{E}_x[\xi_t|\tau_1 < \tau_0]$ is the solution $m_t(x) = 1-(1-x)e^{-t}$ of the ordinary differential equation $m_t'(x) = 1 - m_t(x)$, $m_0(x) = x$. By symmetry, we obtain

$$\mathbb{E}_x[\xi_t|\tau_1 < \tau_0] = 1-(1-x)e^{-t}, \quad \mathbb{E}_x[\xi_t|\tau_0 < \tau_1] = xe^{-t}. \tag{10}$$

By Kimura Kimura (1955, 1970), conditional on fixation,

$$\mathbb{P}_x^{*0}(\tau_1 > t) = (1-x)\sum_{i=1}^{\infty}(1+2i)(-1)^{i-1}H([i-1,i+2],[2],x)e^{-i(i+1)t/2},$$

where H is the hypergeometric function. Taking $x \to 0$, $H([i-1,i+2],[2],0) = 1$ for all $i \geq 1$. Hence,

$$\mathbb{P}_0^{*0}(\tau_1 > t) = \sum_{i=1}^{\infty}(1+2i)(-1)^{i-1}e^{-i(i+1)t/2}, \quad t \geq 0, \tag{11}$$

and

$$G_0(t) = \mathbb{E}_0^{*0}(\min(\tau_1,t)) = \sum_{i=1}^{\infty}\frac{2(1+2i)(-1)^{i-1}}{i(i+1)}(1-e^{-i(i+1)t/2}). \tag{12}$$

Also, $\mathbb{E}_0^{*0}(\tau_1) = \lim_{t\to\infty} G_0(t) = 2$.

Wright-Fisher diffusion with returns from the boundary. Following Maruyama Maruyama (1977), we now modify the boundary behavior of the diffusion process. When the process hits the boundary point 0 then it remains in 0 during a random time which is exponentially distributed with rate κ, after which it returns to the interior of the state space by a jump to a fixed point x_0. From this time on the process starts executing a new path of (ξ_t) with $\xi_0 = x_0$. If the process hits the upper boundary point 1 it immediately jumps to the lower boundary 0 where it holds for an exponential duration and then jumps to x_0, as in the previous case. More general versions of this type of modified process have been studied systematically in e.g. Peng and Li (2013), where it is called diffusion with holding and jumping boundary. It is clear that the modified process no longer get trapped at exit points but rather has a steady state on the semi closed interval $[0,1)$. The unique stationary distribution is given by

$$\mu(0) = \frac{1}{1+\kappa\int_0^1 G(x_0,y)\,dy}, \quad \mu(y)\,dy = \frac{\kappa\,G(x_0,y)\,dy}{1+\kappa\int_0^1 G(x_0,y)\,dy}, \quad 0 < y < 1. \tag{13}$$

The site frequency spectrum. In this paragraph we let L independent copies of the Wright-Fisher diffusion process with return from the boundary represent the status of a sequence of L consecutive codon sites. The state of each component is the frequency of a mutating derived allele at the site and the boundary 0 is the clonal state that corresponds to a single ancestral allele throughout the population.

To comply with the choice of parameters in the main article, we take $x_0 = 1/N$ and $\kappa = N\theta/L$ in (13) and suppose that both N and L are large. Then by (2) and (3),

$$\mu(0) = \frac{1}{1 + 2\theta \log(N)/L}, \quad \mu(y)\, dy \approx \frac{\theta/L}{1 + 2\theta \log(N)/L}\, \omega_\gamma\, \frac{1 - e^{-2\gamma(1-y)}}{\gamma y(1-y)}\, dy, \quad 0 < y < 1.$$

For our application the measure $\mu(y)\, dy$ has the interpretation of a "site frequency spectrum" and $\mu(0)$ is the probability that the site is clonal. This is in agreement with Eq. (31) of Evans et al. (2007) and Eq. (9.23) of Ewens (2004). It follows that the number of polymorphic sites has a binomial distribution $\mathrm{Bin}(L, 1 - \mu(0))$, hence approximately a Poisson distribution with mean $2\theta \log(N)$. Thus,

$$\text{the fraction of polymorphic sites in a sequence of length } L \sim \frac{2\theta \log N}{L}, \tag{14}$$

and if this measure is sufficiently small then in a typical site the expected frequency is

$$\int_0^1 y\mu(y)\, dy \approx \frac{\theta}{L}\, \omega_\gamma \int_0^1 \frac{1 - e^{-2\gamma(1-y)}}{\gamma(1-y)}\, dy \tag{15}$$

and the expected heterozygosity

$$\int_0^1 2y(1-y)\mu(y)\, dy \approx \frac{\theta}{L}\, \omega_\gamma \int_0^1 \frac{1 - e^{-2\gamma(1-y)}}{\gamma}\, dy, \tag{16}$$

where the integrals in the above expressions are both finite.

References

Breiman L. 1992. Probability. SIAM: Classics in Applied Mathematics.

Etheridge A. 2011. Some mathematical models from population genetics. Springer-Verlag.

Evans SE, Shvets Y, Slatkin M. 2007. Non-equilibrium theory of the allele frequency spectrum. Theor Popul Biol. 71:109–119.

Ewens WJ. 2004. Mathematical population genetics: 1. theoretical introduction, second ed. Springer-Verlag, Berlin.

Karlin S, Taylor HM. 1981. A second course in stochastic processes. Academic Press.

Kimura M. 1955. Solution of a process of random genetic drift with a continuous model. Proc Natl Acad Sci USA. 41:144–150.

Kimura M. 1970. The length of time required for a selectively neutral mutant to reach fixation through random frequency drift in a finite population. Genet Res. 15:131–133.

Maruyama T. 1977. Stochastic problems in population genetics. Springer-Verlag.

Peng J, Li WV. 2013. Diffusions with holding and jumping boundary. Sci China Math. 56:161–176.