Genomics and bioinformatics

BIO-463

Prof. Anne-Florence Bitbol
EPFL

2 Population genetics: natural selection and genetic drift

2.1 Deterministic description: large populations, large number of mutants and large fitness advantage

For simplicity, let us consider a population undergoing deterministic exponential growth.

Let us consider two types of organisms A and B. Let us choose a continuous time description, and assume that they are growing exponentially, with respective division rates $r = 1 + s$ and 1. Note that this means that the growth rate of B is taken as reference and defines the unit of time. Denoting their respective numbers by A and B, we have:

$$\begin{cases} \frac{dA}{dt} = (1+s)A\,, \\ \frac{dB}{dt} = B\,, \end{cases} \tag{1}$$

Introducing the fraction x of A organisms as $x = A/(A+B)$, we can focus on the two variables $N = A + B$ and x instead of considering A and B. From Eq. 1 we obtain

$$\begin{cases} \frac{dN}{dt} = (1+s)A + B = N(1+sx)\,, \\ \frac{dx}{dt} = \frac{1}{N}\frac{dA}{dt} - \frac{A}{N}\frac{1}{N}\frac{dN}{dt} = (1+s)x - x(1+sx) = sx(1-x)\,, \end{cases} \tag{2}$$

and thus the fraction x of A organisms satisfies

$$\frac{dx}{dt} = sx(1-x)\,. \tag{3}$$

This deterministic differential equation describes the evolution of mutant fraction x, in the limits of large population sizes, large mutant numbers and relative fitness advantage s. Its solution reads:

$$x(t) = \frac{x_0 e^{st}}{1 + x_0(e^{st} - 1)}\,, \tag{4}$$

with $x_0 = x(0)$.

2.2 Wright-Fisher model

Introduction. Fluctuations coming from finite size and individual entities appearing or disappearing are very important at the scale of a population of microorganisms. Here, they may regard the number of microorganisms in the population, and/or the number of microorganisms having a given genotype. Let us study in more detail how the latter number evolves due to birth and death events in a simple model called the Moran model. We will assume that the population size N is held fixed and that there are two types of individuals at the beginning, namely type A (e.g. mutants) and type B (e.g. wild-type organisms). Eventually, after a sufficient number of generations, because N is finite, all individuals will be descended from just one single ancestor, and thus one of the two types will take over (we say that it fixes) and the other one will disappear. If there is no fitness difference between the two types, any variation in the proportion of mutant organisms will just arise from finite size fluctuations associated to birth and death events. In population genetics, this process is known as

genetic drift. In the presence of fitness differences, both natural selection and genetic drift will come into play.

The Wright-Fisher model is a population model at constant size N where non-overlapping generations are considered. Consider that there are two types of haploid individuals, wild-types with fitness 1 and mutants with fitness $1+s$ (where we make no assumption on the sign of s), and denote by x_n the fraction of mutants at generation n in the population. (Note that traditionally one considers diploid individuals with sexual reproduction, but here we consider haploid ones with asexual reproduction.) In the Wright-Fisher model, N individuals are randomly sampled with replacement using a binomial law with proportion x'_n in order to form generation $n+1$, where x'_n accounts both for the fraction x_n and for the fitness of each type:

$$x'_n = \frac{(1+s)x_n}{(1+s)x_n + 1 - x_n} = \frac{(1+s)x_n}{1+sx_n} \,. \tag{5}$$

Now N individuals are sampled using a binomial law with proportion x'_n. Concretely, we sample a number k_{n+1} of mutants comprised between 0 and K according to

$$P(k_{n+1}) = \binom{N}{k_{n+1}} (x'_n)^{k_{n+1}} \left(1 - x'_n\right)^{N - k_{n+1}} \,, \tag{6}$$

and we complete the new generation population by adding $K - k_{n+1}$ wild-type individuals.

2.2.1 Branching process approximation with Poisson law

Consider a well-mixed population of size N in the Wright-Fisher model, starting from the state where there is 1 mutant and $N-1$ wild-type organisms. The probability p of extinction of the mutant can be calculated using a branching process approach [1]. For this, we approximate the number of mutants k_{n+1} sampled to make generation $n+1$ by a Poisson distribution. Under the Wright-Fisher model, k_{n+1} actually follows a binomial distribution with parameters N and x'_n in Eq. 5. If $N \gg 1$, $x'_n \ll 1$ and Nx'_n is of order 1, then this binomial distribution is well approximated by a Poisson distribution with mean Nx'_n. In particular, at the first generation, with $x_1 = 1/N \ll 1$, the relevant Poisson distribution has mean

$$\lambda = Nx'_1 = N\frac{(1+s)x_1}{1+sx_1} = \frac{1+s}{1+s/N} = 1 + s - \frac{s}{N} + O\left(\frac{s^2}{N^2}\right) \,, \tag{7}$$

which is indeed of order 1. Let us assume in addition that $|s| \ll 1$ and $N|s| \gg 1$: then, $|s|/N \ll s^2$ and we can write

$$\lambda = 1 + s + o\left(s^2\right) \,. \tag{8}$$

The extinction of the mutant lineage can occur in the following ways:

- No mutant is sampled to constitute the first generation. This occurs with probability $e^{-\lambda}$.
- One mutant is sampled to constitute the first generation, but then it gets extinct. This extinction has probability p because we are in the exact same situation as at generation 0 (1 mutant). Thus, this occurs with probability $p\,\lambda e^{-\lambda}$.
- Two mutants are sampled to constitute the first generation, but then both of them get extinct. *If their lineages can be considered independent* (this is a key assumption in the branching process formalism, and it is acceptable if the population is large enough and the mutant fraction remains small enough while the fate of the mutant is set), these two extinctions occur together with probability p^2. Thus, this scenario occurs with probability $p^2\,\lambda^2 e^{-\lambda}/2$.
- Similarly, the case where n mutants are sampled to constitute the first generation and then they get extinct has probability $p^n\,\lambda^n e^{-\lambda}/n!$.

Summing over all these distinct cases, we obtain an equation on p:

$$p = \sum_{n=0}^{\infty} p^n \frac{\lambda^n e^{-\lambda}}{n!} = e^{-\lambda} \sum_{n=0}^{\infty} \frac{p^n \lambda^n}{n!} = e^{-\lambda} e^{\lambda p} = \exp\left[\lambda\left(p-1\right)\right] \,. \tag{9}$$

Note that in fact $n \leq N$, and that the assumption that lineages are independent requires $n \ll N$. However, when n increases substantially above $\sim \lambda$, the associated probability becomes extremely small, so these terms are negligible, and including them (up to $n \to \infty$) has the advantage of giving a nice expression. Eq. 9 can be interpreted as $p = g(p)$ where g is the generating function of the Poisson distribution: the extinction probability p is a fixed point of this generating function. Indeed, the generating function of the Poisson distribution reads

$$g(x) = \sum_{n=0}^{\infty} x^n P(n) = \sum_{n=0}^{\infty} x^n \frac{\lambda^n e^{-\lambda}}{n!} = e^{-\lambda} \sum_{n=0}^{\infty} \frac{x^n \lambda^n}{n!} = e^{-\lambda} e^{\lambda x} = \exp\left[\lambda\left(x-1\right)\right] . \tag{10}$$

If $\lambda \leq 1$, corresponding to $s \leq 0$ (see Eq. 8), the only solution of Eq. 9 is $p = 1$, and extinction is certain. This illustrates that this approach is not taking drift into account properly, and really focuses on $N \gg 1$. Let us thus focus on the regime where $s > 0$, $s \ll 1$, $N \gg 1$ and $Ns \gg 1$. We then expect the extinction probability to be close to 1. Let us expand it perturbatively in powers of s:

$$p = 1 - as + \frac{b}{2} s^2 + o(s^2) , \tag{11}$$

where a and b are unknown prefactors. To determine a, let us use Eq. 9, injecting the expansions of λ and p from Eqs. 8 and 11;

$$\begin{aligned} 1 - as + \frac{b}{2} s^2 + o(s^2) &= \exp\left[\left(1 + s + o(s^2)\right)\left(-as + \frac{b}{2} s^2 + o(s^2)\right)\right] \\ &= \exp\left[-as + \frac{b}{2} s^2 - as^2 + o(s^2)\right] = 1 - as + \frac{b}{2} s^2 - as^2 + \frac{a^2}{2} s^2 + o(s^2) . \end{aligned} \tag{12}$$

Identifying terms in this expansion in powers of s, we obtain $a = 2$. Therefore, $p = 1 - 2s + O(s^2)$. Because this is the extinction probability of the mutant, its fixation probability is

$$\rho = 1 - p = 2s + O(s^2) . \tag{13}$$

2.2.2 Diffusion approximation

From Eq. 6, we find that the number k_{n+1} of mutant organisms in generation $n + 1$ has mean

$$\langle k_{n+1} \rangle = N x'_n , \tag{14}$$

and variance

$$\Delta k^2_{n+1} = N x'_n (1 - x'_n) . \tag{15}$$

Therefore the mean and variance of the fraction $x_{n+1} = k_{n+1}/N$ of mutant organisms in generation $n + 1$ read

$$\langle x_{n+1} \rangle = x'_n , \tag{16}$$

and

$$\Delta x^2_{n+1} = \frac{x'_n (1 - x'_n)}{N} . \tag{17}$$

In other words, in one generation, x is shifted by an average of

$$M(x_n) \equiv \langle \delta x_n \rangle = \langle x_{n+1} \rangle - x_n = x'_n - x_n = \frac{(1+s)x_n}{1 + s x_n} - x_n = \frac{s x_n (1 - x_n)}{1 + s x_n} \approx s x_n (1 - x_n) , \tag{18}$$

where the mean denoted by $\langle . \rangle$ is over the possible values of $\delta x_n = x_{n+1} - x_n$ at a given x_n, and where we have assumed $|s| \ll 1$ in the last expression. In addition,

$$V(x_n) \equiv \langle (\delta x_n)^2 \rangle - \langle \delta x_n \rangle^2 = \Delta x^2_{n+1} = \frac{x'_n (1 - x'_n)}{N} = \frac{(1+s) x_n (1 - x_n)}{N (1 + s x_n)^2} \approx \frac{x_n (1 - x_n)}{N} , \tag{19}$$

where we have assumed $|s| \ll 1$ again (at the last step).

Fixation probability. In the regime $N \gg 1$, it can be shown (see e.g. Ref. [2] p. 423 & p. 371) that the fixation probability $\rho(x)$ of the mutant type starting from a fraction x of mutants satisfies the following differential equation:

$$\frac{V(x)}{2}\frac{d^2\rho}{dx^2} + M(x)\frac{d\rho}{dx} = 0\,, \tag{20}$$

where $M(x)$ and $V(x)$ are given in Eqs. 18 and 19. In addition, $\rho(0) = 0$ and $\rho(1) = 1$. Therefore,

$$\rho(x) = \frac{\int_0^x \exp\left(-\int_0^u \frac{2M(p)}{V(p)}dp\right)du}{\int_0^1 \exp\left(-\int_0^u \frac{2M(p)}{V(p)}dp\right)du} = \frac{1-e^{-2Nsx}}{1-e^{-2Ns}}\,. \tag{21}$$

In particular, starting from one mutant, $x = 1/N$, we have

$$\rho(1/N) = \frac{1-e^{-2s}}{1-e^{-2Ns}}\,. \tag{22}$$

Note that the result in Eq. 21 is most often given for a population of N diploid organisms with sexual reproduction, in which case both factors 2 are replaced by 4 (but then, starting from 1 mutant gene gives an initial frequency $x = 1/(2N)$).

Limiting behaviors. Let us analyze the limiting behaviors of Eq. 22. First, if $s \gg 1$, we find that $\rho_1 \to 1$: fixation becomes certain. However, such highly beneficial mutations are rare (note also that this is in fact outside of the domain of validity of the diffusion approximation). Second, if $|s| \ll 1/N$,

$$\rho(1/N) = \frac{1-(1-2s+O(s^2))}{1-(1-2Ns+O(N^2s^2))} \sim \frac{1}{N}\;: \tag{23}$$

we recover the neutral regime for such small fitness differences. Such mutations are said to be effectively neutral. Third, if $N \gg 1$ while $s > 0$ but $s \ll 1$ and $Ns \gg 1$, then

$$\rho(1/N) = \frac{1-(1-2s+O(s^2))}{1-e^{-2Ns}} \sim 2s\,, \tag{24}$$

which is much larger than the neutral result but still much smaller than one. This result is the same as in the branching process regime above. We see that weakly beneficial mutants are in fact unlikely to take over. This is because of the importance of fluctuations while mutants are in small numbers. Fourth, if $N \gg 1$ while $s < 0$ but $|s| \ll 1$ and $N|s| \gg 1$, then

$$\rho(1/N) = \frac{1-(1-2s+O(s^2))}{1-e^{-2Ns}} \sim -2se^{2Ns} \to 0\,, \tag{25}$$

which indicates that the fixation of deleterious mutations is exponentially suppressed.

Starting from several mutants. If we start from a fraction $x > 1/N$ of mutants, the probability of mutant fixation reads (see Eq. 21):

$$\rho(x) = \frac{1-e^{-2Nsx}}{1-e^{-2Ns}}\,. \tag{26}$$

If $Ns \gg 1$, then we can approximate it by

$$\rho(x) \approx 1 - e^{-2Nsx}\,. \tag{27}$$

The exponential term here becomes small if $x > 1/(Ns)$. This means that a sufficiently beneficial mutant satisfying $s \gg 1/N$ is very likely to fix if it reaches a fraction $x > 1/(Ns)$ in the population. Extinctions of beneficial mutants thus happen early, while mutants are in small numbers, due to fluctuations.

Conclusion. The Wright-Fisher model shows us that a neutral mutant in a finite population can either take over or disappear with fixed probabilities, and this dramatic evolution of the frequency of the mutant type is due to fluctuations associated to finite size and birth and death events. Even a rather fit mutant can disappear due to these fluctuations.

The average time to fixation can be studied using similar methods [3, 4]. The fixation probability and the average fixation time are first passage properties (in the particular case of absorbing states). These quantities are intensely studied for random walks.

2.3 Serial dilution: linking abstract models to evolution experiments

Evolution experiments are often conducted in batch culture with serial transfers, also called serial passage. Let us consider a model with serial dilutions such that the population of initial size K (bottleneck) undergoes deterministic exponential growth for a time t and then K individuals are selected randomly from the grown population to form the next bottleneck, and so on. As before, we assume that there are two types of individuals, wild-types with fitness 1 and mutants with fitness $1+s$. Now these fitnesses represent deterministic growth rates. Consider that the sampling upon the dilution step follows a binomial law, as in the Wright-Fisher model.

Assume that at generation n the initial fraction of mutants (among the K founding individuals of this generation) is x_n. After growth, the fraction of mutants reads

$$x'_n = \frac{x_n e^{st}}{1 + x_n(e^{st}-1)} \,. \tag{28}$$

Introducing

$$\sigma = e^{st} - 1 \,, \tag{29}$$

we get

$$x'_n = \frac{(1+\sigma)x_n}{1+\sigma x_n} \,, \tag{30}$$

which has the exact same form as Eq. 5 but with σ standing in for s. To form the next bottleneck, mutants are sampled according to the binomial law $\mathcal{B}(K, x'_n)$. Concretely, we sample a number k_{n+1} of mutants comprised between 0 and K according to

$$P(k_{n+1}) = \binom{K}{k_{n+1}} (x'_n)^{k_{n+1}} (1-x'_n)^{K-k_{n+1}} \,, \tag{31}$$

and we complete the new bottleneck population by adding $K - k_{n+1}$ wild-type individuals.

As mutants are sampled using a binomial law with proportion x'_n and number of samplings K, we can calculate the mean shift and variance of the mutant fraction from generation n to generation $n+1$ exactly as we did before for the Wright-Fisher model. Then we can employ the same Kolmogorov backward equation as in the Wright-Fisher model, yielding the following formula for the fixation probability $\rho(x)$ of the mutant type starting from a fraction x of mutants:

$$\rho(x) = \frac{1-e^{-2K\sigma x}}{1-e^{-2K\sigma}} \,. \tag{32}$$

As above, this holds under the assumptions that $K \gg 1$ and $|\sigma| \ll 1$. Eq. 29 shows that if $|\sigma| \ll 1$ then $|s|t \ll 1$ and $\sigma \approx st$. Thus, Eq. 32 finally becomes

$$\rho(x) = \frac{1-e^{-2Kstx}}{1-e^{-2Kst}} \,. \tag{33}$$

under the assumptions $K \gg 1$ and $|s|t \ll 1$. This result matches that of Eq. 21 obtained in the Wright-Fisher model, but with st replacing s and K replacing N. In particular, starting from one mutant at a bottleneck, $x = 1/K$, we have

$$\rho(1/K) = \frac{1-e^{-2\sigma}}{1-e^{-2K\sigma}} = \frac{1-e^{-2st}}{1-e^{-2Kst}} \,, \tag{34}$$

still under the assumptions $K \gg 1$ and $|s|t \ll 1$. Note that for $t = 1$ we recover the formula obtained in Wright-Fisher model with $N = K$.

References

[1] J. B. Haldane. A mathematical theory of natural and artificial selection. V. Selection and mutation. *Camb. Philos. Soc.*, 23:838–844, 1927.

[2] J. F. Crow and M. Kimura. *An Introduction to Population Genetics Theory.* Blackburn, 2009.

[3] W. J. Ewens. *Mathematical Population Genetics.* Springer-Verlag, 1979.

[4] A. Traulsen and C. Hauert. Stochastic evolutionary game dynamics. In H.-G. Schuster, editor, *Reviews of Nonlinear Dynamics and Complexity*, volume II. Wiley-VCH, 2009.