Problems on evolutionary dynamics for Gladys & Caroline

Jamie R. Blundell

January 2, 2019

These questions are aimed at giving a grounding on various concepts in evolutionary dynamics and population genetics. They are grouped, roughly, by theme and are rated from ❦ to ❦❦❦❦❦ from easy and short to more of a research problem. I would encourage you to check answers in any way you can e.g. by writing a simple simulation, looking at available data etc.

1. **Warm-up: a model of retinoblastoma ❦❦**

   Retinoblastoma is a childhood cancer of the retinal cells that is caused by losing both copies of the tumour suppressor gene RB1. In the heritable form of the disease ($\sim 40\%$ of cases) the child inherits one faulty gene from a parent, and then acquires the other mutation during development, diagnosis in these children is made on average before 1 years old. $\sim 60\%$ of cases are non-heritable in these cases both copies are lost during development and diagnosis is typically made before 3 years old. Approximately 1 in 100,000 children develop the disease before the age of 5.

   (a) How many cells make up the retina?

   (b) How many stem cell divisions occur back to the common ancestor of all retinal cells?

   (c) What is the per site somatic mutation rate per cell division in human cells?

   (d) What is the average size (bp) of a human gene (excluding introns)?

   (e) What is the target size (number of sites) for loss-of-function mutations for a typical human gene?

   (f) What is the probability that a child, who inherits one faulty copy of RB1, develops retinoblastoma?

   (g) In the heritable form of the cancer, it is common for multiple tumours to be found in a single eye ("unilateral"), and, in $\sim 25\%$ of cases, tumours are found in both eyes ("bilateral"). Does this agree with your model in (e)? What should the distribution of number of tumours in a given eye be?

   (h) Estimate the probability of developing retinoblastoma via the non-heritable route. How does this compare with the data?

   (i) Use mutation mutation rates to estimate the probability that a child inherits a faulty copy from a parent despite neither parent having been affected.

   (j) Estimate what the frequency the RB1 LoF allele will be at in the human population? Does this agree with data? How deleterious is the RB1 germ-line mutation?

2. **Fisher's fundamental theorem and evolution on standing variation ❦❦❦**

   Consider a trait (e.g. height) which is Gaussian distributed with mean $\bar{h}$ and standard deviation σ.

   (a) If the fitness[†], x, of an individual is proportional to their height above the current mean $x = s\left(h - \bar{h}\right)/\sigma$, show that the distribution evolves by being translated at a constant speed, v, and find an expression for v.

(b) In reality, for any finite number of individuals the true distribution cannot be Gaussian since there must be a maximum height g (and for that matter a minimum height) in the population. If there are N individuals in total, what is the distribution of g? How does this alter the conclusions from (a)?

(c) What happens to the rate of evolution and the distribution of heights in the long-time limit ($t \gg 1/s$)? What mechanisms generate diversity in the population?

[†] *Note fitness here is defined as growth rate i.e. in an interval of time δt individuals with fitness x will change in number by a factor $(1 + x\delta t) \approx \exp(x\delta t)$*

3. **An introduction to genetic drift ❦❦**

The number of offspring an individual contributes to the next generations is stochastic. Think of humans as an example: the average number of offspring per parent is close to 1, but there is substantial variation: some families have no children, others many (the variation is different for men and women, which is relevant in a later question). This variation is the cause of genetic drift. Consider an initial number of individuals, n_0, each of which contributes an (integer) number of offspring, X, to the next generation. X is drawn from a discrete probability distribution (e.g. a beta binomial, Poisson etc.) with mean 1 and variance $c \sim 1$.

(a) The number of cells after 1 generation is

$$n_1 = \sum_i^{n_0} X_i \tag{1}$$

What is the mean and variance of n_1?

(b) Show that the variance in n_t, $\mathtt{Var}(n_t)$, for early times is given approximately by cn_0t provided t is "small". When does this approximation break down and why?

The result from (b) can be used as the basis for a very crude, but useful result. After $t \sim n/c$ generations of drift an initial number of n cells have a substantial chance (say $\sim 1/2$) of having fluctuated to extinction ($n = 0$) and a substantial chance (say $\sim 1/2$) of having fluctuated to $2n$.

(c) Apply this rule recursively to the lineage that starts from a single cell ($n_0 = 1$) and show that the probability the lineage remains alive after t generations is:

$$P(n > 0, t) \approx \frac{1}{1 + ct} \tag{2}$$

(d) For lineages that remain alive, show that their approximate size after t generations is $n \sim ct$. Explain in plain English why a higher variance in offspring, c, makes a lineage less likely to survive, but larger if it does so?

4. **An aside on generating functions ❦❦❦**

For a discrete probability distribution P_n or continuous density $\rho(n)$ the moment generating function is defined as:

$$M(\phi) = \sum_n P_n e^{-n\phi} = \int \rho(n) e^{-n\phi} dn \tag{3}$$

.

(a) Show that the moment generating function has the same information contained in it as the probability distribution i.e. that from $M(\phi)$ one can recover $\rho(n)$

(b) What is the significance of $M(\phi \to \infty)$?

(c) For a random variate, X, if the moment generating function

$$\langle e^{-X\phi} \rangle = \exp(a_0 + a_1\phi + a_2\phi^2 ...) \tag{4}$$

show that $a_0 = 0$, $a_1 = \langle X \rangle$, $a_2 = \langle X^2 \rangle - \langle X \rangle^2$. What is the meaning of a_3?

(d) If X is drawn from a Bernoulli distribution with success probability p, what is its moment generating function $\langle e^{-X\phi} \rangle$?

(e) What is the moment generating function for $S = \sum_i^N X_i$

(f) By expressing the mgf for S as a power series in ϕ show that for large N, show that the distribution is well approximated by a Gaussian (an example of the central limit theorem) and determine how many standard deviations from the mean this limit breaks down. How does this scale with N?

(g) Show that the distribution can be recovered by performing the inverse Laplace transform via the Bromwich integral

$$P(S) = \int_{\gamma - iY}^{\gamma + iY} M(\phi) e^{S\phi} d\phi \tag{5}$$

where γ is to the right of all the poles and $Y \to \infty$. Use the saddle point method to calculate an approximate expression for $P(S)$, comparing this with the exact expression (Binomial).

5. The birth-death process and its generating function ❦❦❦❦

Consider n cells which divide at rate B and die at rate D. We are interested in calculating the distribution of clone sizes, $\rho(n)$ that results from this as a function of time.

(a) In an infinitesimal δt and write down a stochastic equation relating $n(t)$ to $n(t + \delta t)$.

(b) By calculating the generating function $\langle e^{n(t+\delta t)\phi(t+\delta t)} \rangle$ show that the dynamics of $\phi(t)$ are given by:

$$\partial_\tau \phi = s\phi - c\phi^2 \tag{6}$$

where $2c = B + D$, $s = B - D$ and $\tau = T - t$, where T is the current time and thus runs backward in time.

(c) By solving this and using the initial condition $n(t = 0) = n_0$, show that the generating function for the distribution of cells at time t is given by

$$M(\phi) = \exp\left(-\frac{a\phi}{b\phi + 1}\right) \tag{7}$$

where $a = n_0 e^{st}$ and $b = \tilde{n} = (e^{st} - 1)/(s/c)$

(d) What is the extinction probability?

(e) For $b \gg a$ which is equivalent to $n_0 \ll c/s$, the argument is always small. By small expanding, and inverting, show that the clone-size distribution is well approximated by:

$$\rho(n) = \delta(n)\left(1 - \frac{e^{st}}{\tilde{n}}\right) + \frac{e^{st}}{\tilde{n}^2} \exp\left(-n/\tilde{n}\right) \tag{8}$$

where $\tilde{n} = (e^{st} - 1)/(s/c)$. This is an important and useful result!

6. Introducing fitness and establishment ❦❦❦

What if some members of a population have, on average, more offspring than others? These difference are called “fitness differences” and are the basis for adaptation. Consider a population of N identical individuals, in which a new mutation arises, which is initially present in a single individual: $n(t = 0) = 1$. This mutation confers a “’fitness advantage”, s, meaning the average number of offspring begat by individuals carrying the mutation is $1 + s$.

(a) If there were no genetic drift and no constraint on the total number of individuals, how long would it take for this mutation to reach a size of N?

(b) What is the total number of individuals in the population as a function of time?

(c) In order to stop the total population from growing indefinitely it is common to constrain the population size, N, to remain fixed. Show that this can be achieved by considering the fitness-advantage over a “mean-fitness”, $\bar{x}$, such that the mutant lineage grows at a rate $s - \bar{x}(t)$. Find an expression for $\bar{x}(t)$ and plot the trajectory the mutant takes to fixation, $n(t)$.

Note: this is a simple example of how evolutionary dynamics can be non-linear. The population size constraint introduces non-linearities and stops growth from being perfectly exponential.

Consider now that there is both drift and selection and that the abundance of the mutant is n.

(d) What is the characteristic timescale for selection τ_s? (i.e. number of generations for selection to increase the abundance of the mutant by a substantial factor)

(e) What is the characteristic timescale for drift τ_d? (i.e. number of generations for drift to increase the abundance of the mutant by a substantial factor)

(f) Explain the significance of the parameter combination ns and discuss the behaviour of the mutation when $n \ll 1/s$ and $n \gg 1/s$

(g) Using the results from (f) determine the probability that a mutation reaches its establishment size $1/s$.

(h) How long does it typically take for a mutation that is destined to establish to do so?

(i) For mutations that establish, what is their size over time $n(t)$ and how long does it take for them to sweep through the population (ie. to reach a size $\sim N$)

7. Successive-sweeps vs clonal interference ❦❦

Consider a population of N identical individuals. An individual acquires mutations at a rate of μ per generation and mutations each confer a fitness advantage s.

(a) At what rate do mutations — that are destined to establish — enter the population?

(b) What is the average waiting time for the first mutation that will establish?

(c) What is the distribution of these waiting times for the first mutation that will establish?

(d) What is the distribution of waiting times for the kth mutation that will establish?

(e) Recall that the “sweep time” for a beneficial mutation is $\tau = (1/s)\ln(Ns)$. Successive sweeps occur when the time between sweeps is much longer than the sweep time itself. Show that successive sweeps will occur when

$$N\mu \ln(Ns) \ll 1 \tag{9}$$

and sketch what the dynamics of mutations looks like for a time period involving multiple sweeps. What parameter combination is most important in determining whether we are in a successive sweep regime?

(f) Graphically, or otherwise, show that when $N\mu \ln(Ns) \gg 1$ the dynamics are characterised by both "clonal interference" (different mutations arising on the wild-type competing concurrently) and "multiple mutations" (clones with different numbers of beneficial mutations expanding concurrently)

8. **Timescales of drift, selection and mutation ❦❦❦**

Consider a population of N individuals and a single site in their genome, which carries allele `a`. This can mutate to the variant allele `A` at rate μ with the variant conferring a selective advantage s. The timescale for a variant, present in n individuals, to change frequency by a significant factor (i.e. a factor of $\sim e$) via drift is $\tau_d \sim n$, via positive selection is $\tau_s \sim 1/s$, and, via mutation is $\tau_m \sim 1/\mu$. If the site is initially unmutated i.e. `a` is present in all N individuals, describe and sketch in as quantitative way as possible[†] what the dynamics of `A` look like when

(a) $N \ll 1/s \ll 1/\mu$
(b) $N \ll 1/\mu \ll 1/s$
(c) $1/s \ll N \ll 1/\mu$
(d) $1/s \ll 1/\mu \ll N$
(e) $1/\mu \ll 1/s \ll N$
(f) $1/\mu \ll N \ll 1/s$

Are the dynamics in each case actually different? In each case, can any of the effects be effectively ignored? [†] *Consider how often each independent mutation is generated and the fate of these once generated.*

9. **Asexual fitness wave: successive sweeps regime ❦❦**

Consider a population of N cells, each of which accumulates beneficial mutations, of fitness effect s at a rate μ. Assume that fitness effects combine additively such that a clone with k mutations has a fitness of ks, and that the parameter $N\mu \ll 1$ and $\mu/s \ll 1$

(a) What is the steady-state speed of fitness increase

$$v = \frac{d(\text{fitness})}{dt} \tag{10}$$

in this successive sweeps regime?

(b) Sketch the typical fitness trajectory. What determines the scale of the fluctuations in this speed? i.e. what determines the width of the distribution of many fitness trajectories?

(c) List natural populations / real scenarios where successive sweeps are likely the dominant mode of evolution.

10. **Asexual fitness wave: clonal interference regime ❦❦❦❦**

Consider the same set up as in the previous question except that now $N\mu \gg 1$. Initially all N cells are wildtype i.e. have no mutations. We will first consider a deterministic approximation (ignoring drift) and see where and why this breaks down before solving the stochastic problem.

(a) Show that the size of the single-mutant fitness class, under the deterministic approximation is given by

$$\partial_t n_1 = s n_1 + N\mu \tag{11}$$

by solving this, show that $n_1(t) = (N\mu/s)(e^{st} - 1)$.

(b) Extend this argument for additional fitness classes and show that the deterministic approximation the total size of all mutant classes added together would grow as

$$\sum_k n_k \approx N \exp\left[\left(\frac{\mu}{s}\right) e^{st}\right] - 1 \tag{12}$$

(c) When does the kth fitness class reach a size $\sim N$?

(d) When, and for which fitness classes does the deterministic approximation break down and why?

(e) By considering the classes for whom the deterministic approximation holds, show that an approximation for the initial speed of the fitness wave is

$$v \sim s^2 \frac{\log(Ns)}{(\log(s/\mu))^2} \tag{13}$$

(f) Now consider the full model with drift included and a fixed total number of cells N. In steady state the fitness wave (both the nose and the body) must move at some constant speed v, which we would like to determine. Suppose the fittest class in the population at $t = 0$ is q classes above the mean i.e. has fitness advantage qs and has just established. Show that it grows as

$$n_q = \frac{e^{qst - vt^2/2}}{qs} \tag{14}$$

(g) By considering when this will will become the most abundant class in the population show that the speed of the "body" of the wave is related to the lead, q via

$$v_{\text{body}} \approx s^2 \frac{q^2}{2\ln(Nqs)} \tag{15}$$

(h) By calculating when the newly established class is likely to give rise to the next established class show that the speed of the nose is given by:

$$v_{\text{nose}} \approx s^2 \frac{q}{\ln(qs/\mu)} \tag{16}$$

(i) By setting $v_{\text{nose}} = v_{\text{body}}$ show that

$$q \sim \frac{2\ln(Ns)}{\ln(s/\mu)} \qquad \text{and} \qquad v \sim s^2 \frac{2\ln(Ns)}{\ln(s/\mu)} \tag{17}$$

Compare this to the speed calculated above.

11. **Dominant balance**

12. **Site frequency spectra: constant feeding from simple heuristics**

In many contexts (e.g. the earliest stages of healthy tissues accumulating cancer-driver mutations as in clonal haematopoiesis) mutations at a specific site k are generated at a roughly constant rate $\theta_k = N\mu_k \ll 1$ but where there are many distinguishable sites such that $\sum_k \theta_k \gtrsim 1$. We want to understand the clone size distribution that results from this process given these "single mutants" have a fitness advantage s.

(a) By using the clone size distribution for a single mutant, show that the clone size distribution of the constant feeding process must be approximately given by

$$\rho(f)df = \frac{\theta}{f} e^{-f/\phi} df \tag{18}$$

13. **Site frequency spectra: constant feeding more formally from generating functions ❦❦❦❦**

    (a) Show that the generating function for the clone size distribution for mutations generated at rate θ is

    $$M(\phi) = \exp\left(-\theta \int_0^t \frac{\phi e^{st'}}{\phi \tilde{n} + 1}\right) dt' = \exp\left(-\theta \ln(\tilde{n}\phi + 1)\right) \tag{19}$$

    (b) By performing the inverse Laplace tranform exactly using theresidue theorem show that

    $$\rho(\nu) = 1/\Gamma(\theta) e^{-\nu}/\nu^{1-\theta} \tag{20}$$

    with $\nu = n/\tilde{n}$.

    (c) Transform this into a distribution of "establishment times " defined by $\nu = e^{-s\tau}$ and carefully sketch this distribution (on log y-axis scales) for $\theta \gg 1$, $\theta \sim 1$, and, $\theta \ll 1$.

    (d) By considering the relative sizes of subsequent mutations calculate how many independent clones contribute to the expanding fitness class when $\theta \ll 1$ and $\theta \gg 1$.

14. **Site frequency spectra: time dependent feeding**

15. **The Luria-Delbruck experiment and PCR errors**

16. **Inferring human demography from SFS data**

17. **Heterozygosity in humans**

18. **Mitochondrial Eve and Y-chromosomal Adam**

19. **Distribution of fixed beneficial mutations**

20. **Deleterious mutations and Muller's ratchet**

21. **Balancing deleterious and beneficial mutations**

22. **Sex Ratios**

    A not very well known fact is that the probability of giving birth to a boy versus a girl (in humans) is not actually 50-50, but rather 51.6-48.4 in favour of boys (see for example gov.uk national statistics from 2012-2016). One possible reason for this is that men are more likely to die before reproductive age (see UK death rates broken down by age and sex). The chance of a male dying before age 35 is d_b ≈0.021 whereas for a female it is d_g ≈0.013.

    (a) Imagine the probability of giving birth to a boy vs a girl were 50-50, but that a larger fraction of boys die before reproductive age, such that the fraction of boys in the population at reproduction is $f < 0.5$. Show that

    $$f \approx \frac{1 - d_b}{2 - d_b - d_g} \tag{21}$$

(b) Assuming equal mating preference and monogamous coupling, show that a mother who gives birth to a boy with probability b will have a fitness advantage of

$$1 + s = \frac{2b(1-f)/f + 2(1-b)f/(1-f)}{(1-f)/f + f/(1-f)} \tag{22}$$

which results in a stable fixed point: i.e. if $f < 0.5$, then those with $b > 0.5$ have the advantage, which increases the numbers of boys and hence pushed the numbers at reproductive age back to 50-50. This is an example of “Fishers Principle”.

(c) How good of a theory is this for the sex ratio of humans? What assumptions might be invalidated?

23. **A model of viral infections**

24. **The Distribution of Fitness Effects**

25. **On the evolution of sex**

26. **Human-Chimp divergence ❦❦❦166**

There are $\sim$ 30 million substitutions between humans and chimps. When did the last common ancestor of humans and chimps live? By considering the rates at which neutral and beneficial mutations accumulate, estimate what fraction of these fixed differences are adaptive.

27. **Linkage disequlibirum**

28. **Hardy-Weinberg equilibrium**

29. **Linkage blocks**

30. **Soft vs Hard sweeps**

31. **A model of genetic hitchhiking**

32. **Wright-Fisher vs Moran models**

33. **Ewans sampling formula**

34. **Kingman coalescence**

35. **Bolzthauser-Snitzmann coalescent**

36. **Crossing a fitness valley**

37. **The birth-death process and relation to Fokker-Plank equation.**

38. **The route taken by deleterious mutations to fixation, and the effect of conditioning**

39. **Gene surfing and range expansions**

40. **Fisher's exact test**

41. **The distribution of surnames in the UK**