Introductory seminar on "Mathematical Population Genetics" Summer term 2019

Prof. Joachim Hermisson
Dr. Sylvain Mousset

Next class: Wednesday, May 15th 2019.
Prepare exercises 14–18 (part 4).

List of exercises

1 Hardy Weinberg Law

Exercise 1 Blood types

The ABO blood types can (in the simplest case) be coded by the three alleles A, B, and O at a single locus, where A and B are dominant over O (genotypes AO and BO respectively have phenotypes $[A]$ and $[B]$).

1. Denote with p_A p_B and p_O the frequencies of these alleles. Calculate, under the assumption of Hardy-Weinberg equilibrium, the relative frequencies of the blood types A, B, AB and O (phenotypes $[A]$, $[B]$, $[AB]$, and $[O]$). Why exactly these?
2. Let R_A, R_B, R_{AB} and R_O denote the observed frequencies of the respective blood types within a population. From this calculate the allelic frequencies. Is it possible to do this without Hardy-Weinberg? How could one test wether the population is in Hardy-Weinberg equilibrium?

Exercise 2 Dynamics of phenotype frequencies under assortative mating

Mating can be non-random with respect to some traits, respectively genes. We speak of assortative mating when similar individuals are more likely to mate with each other than expected by chance. Consider the following simple but instructive example. The two alleles A_1 and A_2 occur at a gene locus. Furthermore, the genotype A_1A_1 has the same phenotype as A_1A_2, which is different from the phenotype of A_2A_2. Thus A_1 is dominant. Let ρ denote the proportion of individuals that mate assortatively, *i.e.* only with individuals of the same phenotype. This proportion is assumed to be identical for both phenotypes. Denote the (ordered) genotype frequencies with P_{ij} and the allelic frequencies with p and q, $(p + q = 1)$.

1. The contribution of those individuals that mate randomly to the pool of individuals with genotype A_1A_1, A_1A_2 and A_2A_2 in the next generation is then $(1-\rho)p^2$, $(1-\rho)2pq$ and $(1-\rho)q^2$ respectively (why?).
2. Calculate the contribution of the assortatively mating individuals of genotype A_2A_2 to the different types of offspring and also the contributions of the individuals with the dominant phenotype.
3. Show that the following recursion holds:

$$P'_{11} = (1-\rho)p^2 + \rho\frac{p^2}{1-P_{22}}$$

Exercise 3 Dynamics of phenotype frequencies under assortative mating (continued)

See exercise 2 for details.

1. Derive also the recursion for the other two genotypes, P_{12}, P_{22}.
2. Which conclusions can you draw from this set of recursions for the allelic frequencies.
3. What recursion follow for the special cases of $\rho = 0$ and $\rho = 1$?

Exercise 4 Dynamics of heterozygosity under assortative mating

We are considering the assortative mating case discussed in exercise 2 and 3. The quantity $H = 2P_{12}$ is known as *heterozygosity*. ($2P_{12}$ is the frequency of all heterozygotes, such that $P_{11} + 2P_{12} + P_{22} = 1$.)

1. Derive the recursion for H' (which can be expressed as a function of H and ρ).
2. For the case $\rho = 1$, derive the value of $H(t)$ in generation t given e.g. $p(t = 0) = \frac{1}{2}$.
3. What general conclusion can be derived for $\lim_{t\to\infty} H(t)$ in the case of $\rho < 1$?

Exercise 5 Dynamics of phenotype frequencies with X-linkage

Consider a gene with two alleles A and a, located on the X chromosome (in mammals, females are XX while males are XY). The A allele is dominant (individuals with genotypes AA and Aa have phenotype $[A]$ while individuals with genotype aa have phenotype $[a]$). The frequencies of the A allele in males and females are respectively denoted p and q, initial frequencies are denoted p_0 and q_0.

1. Express the recursion for the allele frequencies p' and q'.
2. Express the allele frequencies in males and females after t generations.
3. Express the frequency of the $[A]$ phenotype in females and in males.
4. Does this dynamics change if the females:males ratio differs from 1:1 in the population?

2 Selection and mutation

Exercise 6 Diploid selection model with multiplicative fitness

We consider an autosomal gene with k alleles with frequencies $p_1, \ldots, p_k$ in a diploid population of individuals. We denote the relative frequencies of A_iA_i homozygotes by P_{ii} and of A_iA_j heterozygotes by P_{ij} and assume full random-mating. We assume that k positive constants $v_1, \ldots, v_k$ exist such that the fitness of the A_iA_j genotype $W_{ij} = v_iv_j$ (*multiplicative* fitness).

1. Express the mean fitness $\overline{W}$ in the population as a function of $\bar{v} = \sum_i p_iv_i$.
2. Express the marginal fitness of allele A_i, and the frequency p_i' of allele A_i in the next generation.
3. What model do you recognize (discuss possible differences with the current model)?

Exercise 7 Selection at a single locus with two alleles and a recessive lethal allele

In a randomly mating population we consider a gene with two alleles a and A where the homozygotes aa and heterozygotes Aa have the same phenotype and fitness, whereas the homozygotes AA die before the reproductive stage. The frequencies of the A and a alleles are denoted p and q ($p + q = 1$).

1. Express the fitnesses W_{11}, W_{12}, and W_{22}. What are the corresponding values for the selection coefficient s and the degree of dominance h?
2. Express the marginal fitnesses of the a and A alleles, as well as the mean fitness of the population $\bar{W}$.
3. Express the frequency of the lethal A allele at the next generation.
4. Give the frequency of the A allele after t generations, as a function of t and the initial frequency of the A allele.

Exercise 8 Diploid case with two alleles mutation and selection, without backward mutation

We consider the two-alleles case in discrete time, in a randomly mating population of diploid individuals where the fitness values of the genotypes aa, Aa and AA are $W_{11} = 1$, $W_{12} = 1 - hs$, and $W_{22} = 1 - s$, respectively (we assume $0 < s < 1$). The frequencies of a and A are denoted q and p. The mutation rate from a to A (probability that an a allele mutates to an A allele in a generation) is denoted μ, the backward mutation (from A to a) is assumed to be zero.

1. Express the marginal fitnesses of the a and A alleles, as well as the mean fitness in the population.
2. Express the allele frequencies p' and q' at the next generation.
3. Show that the non-trivial equilibrium frequencies for the A allele, $p^*_{1,2} \neq 1$ are solutions of a quadratic equation.
4. Express the values of the equilibrium points and study their stability in the special case when $h = \frac{1}{2}$.

Exercise 9 Limiting cases of the mutation-selection case with two alleles

For the two non-trivial solutions $p^*_{1,2}$ obtained in exercise 8 we will now investigate some limiting cases.

1. If $h = 0$ prove
$$p_1^* = \sqrt{\frac{\mu}{s}}.$$
2. If $h \gg \sqrt{\frac{\mu}{s}}$ show
$$p_1^* \approx \frac{\mu}{hs}.$$
Remember the Taylor expansion $\sqrt{1-\varepsilon} = 1 - \frac{\varepsilon}{2} + o(\varepsilon)$.

3. If $h \gg \sqrt{\frac{\mu}{s}}$ show that if $h > \dfrac{1-\frac{\mu}{s}}{1-\mu}$ the second equilibrium p_2^* is admissible, but unstable and satisfies

$$p_2^* \approx \frac{h}{2h-1} - \frac{\mu}{hs}.$$

The same Taylor expansion could be useful.

4. Finally prove for mutiplicative selection coefficients $W_{11} = 1$, $W_{12} = 1-t$, and $W_{22} = (1-t)^2$ that one obtains exactely

$$p_1^* = \frac{\mu}{t}$$

3 Recombination and drift

Exercise 10 Haldane's Mapping Function

The recombination fraction r between two loci on the same chromosome is the probability of an odd number of recombination events between these loci. Ignoring interference between adjacent recombination events (in biological reality a recombination events reduces the probability of another recombination events in its close proximity), r relates to the genetic map distance d via *Haldane's mapping function*. Thereby d is the average number of recombination events in a given interval.

1. Derive *Haldane's mapping function*

$$r = \frac{1}{2}(1 - \exp[-2d])$$

2. Why is it useful to have a map function which converts recombination distance r into mapping distance d and vice versa? What might be easier to measure in data?

3. When could it be appropriate to use the approximation $r = d$? What does that mean in the biological context?

Tips for the derivation (question 1): We denote the expected proportions of $0, 1, \ldots, k$ recombination events between to loci $p_0, p_1, \ldots, p_k$ and ignoring interference between recombination events we can assume that they are Poisson distributed, with $P[k] = \frac{\lambda^k \exp[-\lambda]}{k!}$. Then r and d can be written as functions of p_i. Show that $d = \lambda$ and then you can use this result to derive the mapping function.

Exercise 11 Recombination with additive fitness in discrete time

We consider a haploid population, where at two loci, $\mathcal{A}$ and $\mathcal{B}$ there segregate two alleles each, denotes by A_1 and A_2 and B_1 and B_2. Fitness is additive as shown in the following scheme:

	B_1	B_2
A_1	$a_1 + b_1$	$a_1 + b_2$
A_2	$a_2 + b_1$	$a_2 + b_2$

1. Write down mean fitness $\overline{\omega}$ and show that it can be written as the sum of single-locus mean fitness.

2. Show that the time derivative of mean fitness does not depend on r or D. Show that mean fitness is non-decreasing (hint: use Jensen's inequality).

3. Show that all equilibria are in linkage equilibrium (LE)

4. Show that linkage equilibrium $D = 0$ is not maintained, if the population is not at equilibrium, *e.g.* give an example.

Exercise 12 Covariance(genotype, fitness) governs dynamics

We consider the discrete time dynamics of a diploid population with two loci $\mathcal{A}$ and $\mathcal{B}$. The dynamics for the four haplotypes $A_1B_1, A_1B_2, A_2B_1, A_2B_2$ (alternative notation ab, aB, Ab, AB), with frequencies $P_\bullet$ and marginal fitness $\omega_\bullet$ indexed by $\bullet = *_{11}, *_{12}, *_{21}, *_{22}$ read

$$P'_{ij} = \frac{\omega_{ij}}{\overline{\omega}} P_{ij} + \eta_{ij} r \underbrace{\frac{\omega_{A_1B_1A_2B_2}}{\overline{\omega}} D}_{\hat{D} \text{ for diploids}} \tag{7}$$

where $\eta_{ij} = 1$ if $i = j$, and otherwise -1.

We define a measure g_{ij} for the genotype, as the frequency of haplotype A_iB_j within a (diploid) genotype $(A_kB_l|A_mB_n)$. Thus g_{ij} can take only the values $0, 0.5, 1$. Show that we can write the change in haplotype frequency, $\Delta P_{ij} = P'_{ij} - P_{ij}$, in terms of the covariance between the measure of the genotype g_{ij} and fitness, *i.e.*

$$\Delta P_{ij} = \frac{1}{\overline{\omega}} \Big(\underbrace{\mathrm{Cov}(\omega, g_{ij})}_{=P_{ij}(\omega_{ij} - \overline{\omega})} + \eta_{ij} r \hat{D} \Big) \tag{8}$$

Hints:

- Set $g_{ij} = g_{11}$ for the change in P_{11} and show $\mathrm{cov}(\omega, g_{11}) = P_{11}(\omega_{11} - \overline{\omega})$ for this measure of genotype. This then easily generalises to arbitrary i, j.
- $g_{ij}$is the probability to sample the given haplotype A_iB_j from a genotype (that consists of two haplotypes): it can be either 0 (the genotype does not contain the given haplotype), $\frac{1}{2}$ (the haplotype occurs only once in the genotype) or 1 (the haplotype occurs twice in the genotype). For instance, g_{11} is the measure for the haplotype A_1B_1 an we have, (among other possible genotypes) $g_{11}(A_1B_2|A_2B_1) = 0$, $g_{11}(A_1B_1|A_2B_1) = \frac{1}{2}$ and $g_{11}(A_1B_1|A_1B_1) = 1$.
- Remember the definition of Covariance: If there are n possible realizations of pairs of random variables (X, Y), namely (x_i, y_i) for $i = 1, ..., n$, with possibly unequal probabilities p_i , then the covariance of x and Y is

$$\mathrm{Cov}(X, Y) = \sum_i^n p_i \big(x_i - \mathbb{E}(X))(y_i - \mathbb{E}(Y)\big).$$

Exercise 13 Homo- and heterozygosity without replacement

In the lecture we have defined homozygosity F_t and heterozygosity H_t as the probability that two randomly haploid alleles (single chromosomes with diploid individuals) are identical or different by descent, respectively. Sampling occurs with replacement, such that a chosen allele, can be sampled twice with rat $\frac{1}{2N}$.

Let us now consider the case, where sampling occurs without replacement in a population of size $2N$ with k alleles. Derive the formulas for F_t and H_t as expressions of the allele frequencies, as well as the recursion equations. What can be said about the long term evolution of these measures in comparison when sampling occurs with replacement?

4 Genetic drift, neutral theory and the coalescent

Exercise 14 Heterozygosity in a Wright-Fisher model with a finite number of alleles

The lecture notes deal with heterozygosity in the infinite alleles model or infinite sites model, where each mutation creates a new allele. The detailed results were obtained for the two-alleles model.
We consider a Wright-Fisher population of N haploid individuals.

1. Derive the recurrence equation for heterozygosity H_t at time t and give the equation for H at equilibrium under mutation and drift for an arbitrary number of alleles n, where $\forall (i, j) \in \{1, \ldots, n\}, i \neq j, \mu_{ij} = \frac{\mu}{(n-1)}$, so that $\sum_{j \neq i} \mu_{ij} = \mu$.

2. *Think about:* Would this model be relevant at the single nucleotide level? What is the appropriate value for n and what are the model assumptions?

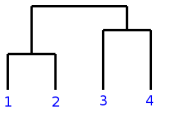
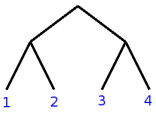
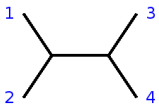
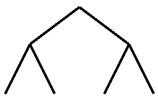
****

Hint: Use the fact that mutation rates are small to approximate the solutions and ignore all terms of order μ^2 in the derivations.

Exercise 15 Topologies of coalescent trees

We use the following definitions:

- A *coalescent tree* is a *labelled* and *time-ordered* history.
- A *rooted tree* is a *labelled* but *not time-ordered* history.
- An *unrooted tree* is a *labelled*, *not time-ordered* history without root.
- A *rooted topology* is an *unlabelled*, *not time-ordered* history.

	coalescent tree	rooted tree	unrooted tree	rooted topology
				
rooted	yes	yes	no	yes
time-ordered	yes	no	no	no
labelled	yes	yes	yes	no

Two coalescent trees are called topologically equivalent if they share the same rooted topology

1. Draw all possible topologically different coalescent trees for a sample of size $n = 5$.
2. Using a forward branching process, give the probability of these given topologies.
3. Propose a recursion to compute the number of topologies for a sample of size n. *Hint:* Define a left and right subtree by "splitting" the tree at the root and relate the number of possible topologies of the entire tree to the number of possible topologies of these two subtrees.
4. How many possible topologies exist for $n = 8$?

Exercise 16 Molecular clock

Based on geological data, the split between two species of fruit flies, *Drosophila differens* and *Drosophila sylvestris*, occupying two different Hawaiian islands, has been estimated to have occurred 2.5 million years ago. In the wild, for fruit flies ten generations per year seem realistic. Based on experiments, the per nucleotide and per generation mutation rate in Drosophila has been estimated to $\mu = 3.5 \cdot 10^{-9}$.

- Give a 'naive' expection, based on an infinite-sites mutation model, of how many differences between the two species can be expected per 1000 nucleotides?
- Do you think it is reasonable to use this substitution rate as the basis of a molecular clock? Check how realistic this 'naive' expectation is: Use the Poisson distribution to give a probability for multiple mutational hits at a single site. For simplicity, assume that mutation rates between all for nucletides, A,C,G,T are equal.
- How do the probabilities for multiple hits change your expectation for the number of segregating polymorphic sites between the two species? You can ignore more than 3 hits at the same site.

Exercise 17 Variance of the number of segregating sites

In the lecture, we have learned that the expected value for S_n, the number of segregating sites in a sample of size n is

$$\mathbb{E}(S_n) = \theta \sum_{i=1}^{n-1} \frac{1}{k}$$

Compute the variance $\mathbb{V}(S_n)$. As for the expected value in the lecture notes assume continuous time for your calculations.

Hints:

- You can express $\mathbb{E}(S_n^2)$ as $\int_0^\infty \mathbb{E}(S_n^2|L_n = t) f_{L_n}(t)\, dt$, where $f_{L_n}(t)$ is the density of the distribution of the total length of a coalescent tree with n tips (similar of course for $\mathbb{E}(S_n)$).
- It is often efficient to partition the coalescent tree into $n-1$ independent parts with $k = 2, \ldots, n$ lineages. These parts of the tree can then be treated independently.

Exercise 18 Coalescent in the Moran model

In the lecture, the coalescent has been derived based on the Wright-Fisher model. However, the coalescent is by no means restricted to the precise assumptions of the Wright-Fisher model. For example, it can also be derived under the Moran model. In the continuous-time version of the model, birth-death events occur in the population at rate N such that (on average) N events occur during a time unit of length one, which defines a generation in the Moran model. At each birth-death event, one random individual is chosen to reproduce and one individual is chosen to die (with replacement).
Consider a sample of n individuals taken from a population of N individuals in an haploid species.

1. Calculate the coalescence rate $p_{c,1}^{(n)}$ for a common ancestor event occurs the n lineages of the sample.
2. Compute the expected time to the most recent common ancestor for a sample of size n in the Moran model. Also calculate the variance of this time (both on the per-generation time scale).
3. Relate the Moran model and the Wright-Fisher model with a coalescent rate of $\frac{n(n-1)}{2N}$ per generation for N haploid individuals. What is the effective population size of a Moran model with N haploid individuals? Assuming that N time steps in the Moran model relate to one generation in the Wright-Fisher model, we want to obtain a relationship between the coalescence effective population size $N_e^{(c)}$ for the Moran model and the Wright-Fisher model.