Tentamentsskrivning i **TMS106/MSN560: Population genetics, 5p.**

Tid: Onsndagen den 8 mars 2006 kl 8.30-12.30.
Lärare och Jour: Serik Sagitov, tel. 772-5351, mob.tel. 073-6907613, rum MC 1421.
Hjälpmedel: Räknedosa utan manualer och med tömda minnen, egen formelsamlingen fyra A4 sidor, utdelade tabeller

Grading system (CTH):	marks	0-11	12-17	18-23	24-30
	grade	U	3	4	5

Grading system (GU):	marks	0-11	12-20	21-30
	grade	U	G	VG

1. (5 marks) Consider two sequences 1 and 2 stemming from a sequence 3, which is their most recent common ancestor. Use the Jukes-Cantor model for the following computations.

   a. If sequences 1 and 3 differ at 10% of their sites, what is the genetic distance between them? If the sequence 3 evolves into the sequence 2 by changing a completely different 10% of its sequence, what is the genetic distance between these two sequences?

   b. Find the genetic distance between the sequences 1 and 2 which differ at 20% of their sites. Compare it to the sum of the other two distances. Why the discrepancy?

   c. Consider two sequences which differ at 75% of their sites. The Jukes-Cantor genetic distance between these two sequences is infinite. Justify this drastic correction of the genetic distance form the observed 0.75 differences per site up to infinitely many changes per site.

2. (4 marks) Suppose that due to a mutation a single gene copy has appeared with the relative fitness $w_a = 1 + s$, $s > 0$ compared to that $w_A = 1$ of the wild-type allele. Consider the early fate of the mutant allele assuming the population size is large and selection is additive.

   a. Explain how the mutant allele can disappear from the population even if the mutation increases fitness.

   b. Given $s = 0.1$ suggest a number for the probability of fixation of the advantageous allele using an appropriate formula for the Wright-Fisher model with selection and without mutation. What is the expected time to fixation if the diploid population size is 10^5.

3. (6 marks) Two inbred varieties of tobacco are crossed and give a variance in leaf number in the F1 generation of 1.5. The variance in the F2 generation is 6.0.

   a. What are the genotypic and environmental variance components? What assumptions do you make with your calculations?

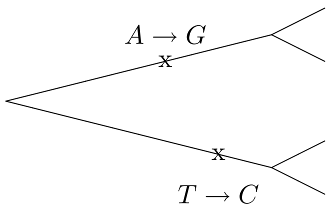
   b. Find the broad sense heritability. What are the corresponding allele frequencies? Explain the fact that heritability can differ for two subpopulations with different allele frequencies.

   c. A recent study of violent behavior in humans was based on father-son comparison. A linear regression model was applied with dependent variable Y = violent behavior of the son, and independent variable X = violent behavior of the father. The study has shown that 80% of variation in Y, is explained by variation in X.

   Many people think this means that people with the at-risk alleles (if they could be identified) are destined to become violent. What is the fallacy of this argument?

4. (4 marks) If linkage equilibrium is maintained with respect to a pair of loci, then natural selection at one locus will not affect the other. Justify this general principle by answering the following question. Suppose the population is in linkage equilibrium. Then if a plague carries off all but the AA individuals, what will happen to the genotype frequencies at the unselected locus B?

5. (6 marks) Four DNA sequences of 100 sites length have evolved from an ancestral sequence as shown in the diagram. The crosses indicate nucleotide changes. We are interested in estimating the compound mutation rate $\theta = 4N\mu$, where N stands for the effective population size and μ is the mutation rate per site per generation.



   a. Describe the resulting 4 by 100 data matrix. How many different alleles do you have in the given sample of size 4? Estimate θ using the corresponding allele frequencies.

b. Compute an estimate of θ based on the number of segregating sites.

c. Compute an estimate of θ from the nucleotide diversity.

d. Are the estimates b) and c) supposed to be the same? In what sense?

6. (5 marks) In a sample of 200 individuals from a population which is expected to be at Hardy-Weinberg equilibrium for a locus with 3 alleles, the number of the 6 possible genotypes found are

Genotype	Number
A_1A_1	76
A_1A_2	54
A_1A_3	33
A_2A_2	18
A_2A_3	16
A_3A_3	3

a. Calculate the gene frequencies of the three alleles, and what numbers of the six genotypes you would have expected from those frequencies.

b. Without doing a formal statistical test, can you see any apparent discrepancies between the numbers in the sample and Hardy-Weinberg proportions? Where?

c. A general question. How realistic are the key assumptions of the Hardy-Weinberg population model?

Partial answers and solutions are also welcome. Good luck!

The answers

Problem 1a. The distances from the leaves to the root $d_{\mathrm{JC}}(1,3) = d_{\mathrm{JC}}(2,3) = 0.107$ are the same. The correction by 0.07 takes account of multiple hits not seen as mismatches.

Problem 1b. The distance between the leaves is $d_{\mathrm{JC}}(1,2) = 0.233$, which is larger than the sum of the other two distances. To explain this discrepancy recall that the Jukes-Cantor distance estimates the total number of changes based on the number of observed differences. For twice as many differences the predicted number of mutations becomes more than twice as large.

Problem 1c. Think of two aligned sequences which are totally unrelated (so that one can claim that the true genetic distance between the sequences is infinite). Then clearly, by chance, 25% of the the aligned nucleotide pairs would match. In this sense a 75% mismatch does correspond to an infinite genetic distance.

Problem 2a. In terms of the haploid reproduction model, the relative fitness coefficient $(1+s)$ of the mutant allele gives the average number of offspring surviving to maturity as compared to the offspring of the wild-type allele. Thus if there were no variation in the offspring numbers, the mutant allele would always get fixed. However, the offspring numbers are random and by chance the first mutant may have zero offspring. Or it may produce one offspring which leaves no viable descendants. On the other hand, if the reproduction of the mutant allele does not get extinct during the first few generations, then with a high probability the advantageous mutant allele will take over the population.

Problem 2b. Fixation probability ≈ 0.2. Mean time to fixation 244 generations.

Problem 3a. The F1 generation is the offspring of mating pairs of type $AA \times aa$. Thus all individuals in the F1 generation have the same genotype Aa. In this case the genotypic variance is zero and $\sigma_e^2 = 1.5$. The F2 generation is random bred with $p = q = 0.5$. Its variance $\sigma_p^2 = 6.0$ contains both components - genotypic and environmental. By substraction we find the genotypic variance $\sigma_g^2 = 6.0 - 1.5 = 4.5$. Here we assume that the genotypic and environmental components are independent.

Problem 3b. The broad sense heritability $H^2 = 4.5/6 = 0.75$. It is specific for the allele frequency $p = 0.5$ in the F2 generation. Changing the allele frequency in a population implies a change in the phenotype value distribution, because the genotype frequencies will change.

Problem 3c. Here dependence is not due to genetic heritability but rather due to inheritance of environment. With fathers exposed to a violent environment, their sons would tend to exhibit violent behavior as well.

Problem 4. The genotype frequencies in the HWE and linkage equilibrium case.

Mother\Father	AB	Ab	aB	ab	Gamete total
AB	$p_1^2p_2^2$	$p_1^2p_2q_2$	$p_1q_1p_2^2$	$p_1q_1p_2q_2$	p_1p_2
Ab	$p_1^2p_2q_2$	$p_1^2q_2^2$	$p_1q_1p_2q_2$	$p_1q_1q_2^2$	p_1q_2
aB	$p_1q_1p_2^2$	$p_1q_1p_2q_2$	$q_1^2p_2^2$	$q_1^2p_2q_2$	q_1p_2
ab	$p_1q_1p_2q_2$	$p_1q_1q_2^2$	$q_1^2p_2q_2$	$q_1^2q_2^2$	q_1q_2
Gamete total	p_1p_2	p_1q_2	q_1p_2	q_1q_2	1

The joint one-locus genotype frequencies.

A-locus\\B-locus	BB	Bb	bb	Total
AA	$p_1^2p_2^2$	$2p_1^2p_2q_2$	$p_1^2q_2^2$	p_1^2
Aa	$2p_1q_1p_2^2$	$4p_1q_1p_2q_2$	$2p_1q_1q_2^2$	$2p_1q_1$
aa	$q_1^2p_2^2$	$2q_1^2p_2q_2$	$q_1^2q_2^2$	q_1^2
Total	p_2^2	$2p_2q_2$	q_2^2	1

After the plague: the genotype frequencies

Mother\Father	AB	Ab	Gamete total
AB	p_2^2	p_2q_2	p_2
Ab	p_2q_2	q_2^2	q_2

and the B-locus genotype frequencies

Genotype	BB	Bb	bb	Total
Frequency	p_2^2	$2p_2q_2$	q_2^2	1

Problem 5a. The data matrix looks something like

nucleotide site position	1	...	24	25	26	...	84	85	86	...	100
sequence 1	A	...	G	G	T	...	A	T	C	...	G
sequence 2	A	...	G	G	T	...	A	T	C	...	G
sequence 3	A	...	G	A	T	...	A	C	C	...	G
sequence 4	A	...	G	A	T	...	A	C	C	...	G

There are 2 alleles with sample frequencies $p = q = 0.5$. The HWE heterozygosity is $H = 2pq = 0.5$. Using the IAM formula $H = \frac{\Theta}{1+\Theta}$ we estimate $\hat{\Theta} = 2$. Here Θ corresponds to the mutation rate per gene per generation. Thus $\hat{\theta} = 2/100 = 0.02$.

Problem 5b. Using $\hat{\theta} = S/a_1$ with $S = 2/100$ and $a_1 = 1.83$ we find $\hat{\theta} = 0.011$.

Problem 5c. $\pi = 8/600 = 0.013$.

Problem 5d. The estimates b) and c) are both unbiased estimates of θ under the null hypothesis of neutrality, which claims that the observed genetic variation is due to the random genetic drift and neutral mutations. Another important assumption is the infinitely many sites mutation model. Under these assumptions the two estimates for large samples are expected to be close to each other.

Problem 6a. Allele frequencies are $p_1 = 0.5975, p_2 = 0.2650, p_3 = 0.1375$.

Problem 6b. No striking deviations from the HWE. Some deficit of the A_1A_2 heterozygotes.

Genotype	A_1A_1	A_1A_2	A_1A_3	A_2A_2	A_1A_3	A_2A_2	Total
Observed	76	54	33	18	16	3	200
Expected	71.4	63.4	32.8	14.0	14.6	3.8	200

Problem 6c. No natural population will be exactly in Hardy-Weinberg proportions. But in many instances the population will be close to the HWE. The HWE model assumptions are realistic in the short term since the evolutionary forces like mutation and genetic drift act rather slowly. The random mating assumption is relevant in the cases when the genetic variation does not influence mating decisions.