

Chapter 4 SELF-TEST Molecular evolution

This test covers material in Chapters 3 and 4.

1. The rate of substitutions in a certain region of DNA of length 1000 bases is estimated as 10^{-9} per base per year. If two species diverged approximately 10 million years ago, the fraction of sites that differ between them should be approximately

- A. 1%
- B. 2%
- C. 20%
- D. 75%

B is correct. The probability of a base change in 10 million years is $10^{-9} \times 10^7 = 1\%$. Each species can change by this amount; therefore the fraction of sites that differ should be 2%. Double changes at one site can be neglected when the probability of change is small.

2. For the same sequences as above, which of these statements is true?

- A. The probability that the two sequences will not differ at all is greater than 50%.
- B. The probability that the two sequences will not differ at all is less than 1 in a million.
- C. If there were considerable variation in the substitution rate between sites, the fraction of sites that differ would be greater than if all sites changed at the average rate.
- D. The two sequences should be almost completely randomised with respect to each other after 100 million years.

The probability that the two sequences will not differ at all is $(0.98)^{1000} = 1.68 \times 10^{-9}$.

This is less than 1 in a million, therefore B is correct.

C is wrong because variation in the substitution rate means that there are many changes in some places and none in others, *i.e.* the fraction of sites that differ would be less than if all change at the same rate.

D is wrong – the difference after 10 million years is 2%, so after 100 million years it will be somewhat less than 20% (because of double mutations in the same place). This is a long way from being completely randomized.

3. It is expected that synonymous sites change more rapidly than non-synonymous sites because

- A. The fraction of transversions is larger at synonymous sites.
- B. Stabilizing selection reduces the rate of substitutions at non-synonymous sites.
- C. The mutation rate at synonymous sites is higher.
- D. Natural selection favours new variants arising at synonymous sites.

A – doesn't make sense

B – correct

C – Wrong – the mutation process does not 'know' about whether a site is synonymous or non-synonymous

D – Wrong – natural selection on protein sequences cannot affect synonymous sites, and if there were any weak selection acting on the DNA (*i.e.* codon usage), it would be stabilizing selection, not selection for new variants.

4. A population of fixed size, N , is evolving according to the coalescent theory, with no selection acting. Let T be the time since the last common ancestor of two individuals that are chosen randomly from the population. Which of the following is true?

- A. The mean value of T is dependent on the product of the mutation rate, u , and the population size, N .
- B. T has a Normal distribution, with a mean value equal to N generations.
- C. T will always be very close to N generations when N is very large.
- D. T depends on the particular positions of the branches in the tree, and therefore fluctuates greatly from one population to another.

A – wrong – T does not depend on the mutation rate

B – wrong – T has an exponential distribution

C – wrong – the exponential distribution is not closely peaked around the mean value, even for large N .

D – correct

5. It is believed that all present-day copies of human mitochondrial DNA descended from a single person, 'Eve', living in Africa around 200,000 years ago. Which of the following is true?

- A. Differences in mitochondrial DNA sequences between different African populations are larger than differences between non-African populations.
- B. There are no fossil human remains prior to 200,000 years ago.
- C. Mitochondrial sequences in present day African populations have changed *less* since the time of Eve than have the sequences in non-African populations.
- D. Mitochondrial sequences in present day African populations have changed *more* since the time of Eve than have the sequences in non-African populations.

A – correct

B – wrong – there are fossil human remains from before then, but presumably these people died out and were replaced by our ancestors.

C and D are both wrong – we are assuming that all populations have changed at the same rate. The reason the non-African populations are more similar to one another is that they have a common ancestor at a time that is more recent than Eve.

6. A new mutant allele has just arisen in a population. Which statement is true?

- A. If the mutant is neutral with respect to the original allele, there is a 50% probability that the mutant allele will replace the original allele.
- B. It is very likely to disappear in a few generations due to random drift.
- C. It will only become fixed in the population if there is a strong selective advantage.
- D. If the mutant allele reaches a frequency of 50%, it will almost always go on to fixation.

A – wrong – the probability of fixation of a neutral allele is only $1/N$.

B – correct

C – wrong – advantageous, neutral and deleterious mutations can all sometimes become fixed.

D – wrong – if a neutral mutation is at 50% frequency it will have 50% chance of extinction and 50% chance of fixation.

7. If two sequences evolve according to the Jukes-Cantor model, and they are observed to differ at 20% of sites, which of the following is true?

- A. The Jukes-Cantor distance is 0.18.
- B. The Jukes-Cantor distance is 0.20.
- C. The Jukes-Cantor distance is 0.23.
- D. The Jukes-Cantor distance cannot be calculated without further information.

The distance is always slightly more than the observed fraction of differences. Therefore C must be true without doing any maths. However, the formula is

$$d = -\frac{3}{4} \ln\left(1 - \frac{4 \times 0.2}{3}\right) = 0.23$$

8. Which of the following statements is correct?

- A. A reversible rate matrix is a symmetrical matrix.
- B. A reversible rate matrix is only used for calculations with rooted trees.
- C. A reversible rate matrix can be used to describe DNA but not protein evolution.
- D. A reversible rate matrix assumes frequencies of different nucleotides are constant in time.

A – wrong – if the matrix were symmetrical, all the base frequencies would be equal.

B – wrong – most methods use unrooted trees and reversible matrices are fine

C – wrong – didn't you read that long section about PAM matrices at all?

D – correct.

9. Which of the following statements is correct?

- A. The PAM250 log-odds matrix applies to sequences that are more distant from one another than the PAM100 matrix.
- B. The BLOSUM85 log-odds matrix applies to sequences that are more distant from one another than the BLOSUM62 matrix.

- C. The BLOSUM85 log-odds matrix applies to sequences that are more distant from one another than the PAM250 matrix.
- D. All three of the above.

A – correct – larger PAM number means more distant sequences

B – wrong – larger BLOSUM number means more similar sequences

C – wrong – you cannot easily compare BLOSUM and PAM numbers but you should have an idea that BLOSUM85 represents quite similar sequences, and PAM250 fairly divergent ones.

D - wrong

10. Which of the following statements concerning the BLOSUM62 matrix below is correct?

- A. Alanine is aligned with arginine more often than expected by chance.
- B. Alanine never changes to cysteine.
- C. Tryptophan evolves the slowest.
- D. The off-diagonal elements are proportional to the rates of substitution from one amino acid to another.

Table 2 - The log odds matrix for BLOSUM 62																				
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	4	0	-2	-1	-2	0	-2	-1	-1	-1	-2	-1	-1	-1	-1	1	0	0	-3	-2
C		9	-3	-4	-2	-3	-3	-1	-3	-1	-1	-3	-3	-3	-3	-1	-1	-1	-2	-2
D			6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	-1	-3	-4	-3
E				5	-3	-2	0	-3	1	-3	-2	0	-1	2	0	0	-1	-2	-3	-2
F					6	-3	-1	0	-3	0	0	-3	-4	-3	-3	-2	-2	-1	1	3
G						6	-2	-4	-2	-4	-3	0	-2	-2	-2	0	-2	-3	-2	-3
H							8	-3	-1	-3	-2	1	-2	0	0	-1	-2	-3	-2	3
I								4	-3	2	1	-3	-3	-3	-3	-2	-1	3	-3	-1
K									5	-2	-1	0	-1	1	2	0	-1	-2	-3	-2
L										4	2	-3	-3	-2	-2	-2	-1	1	-2	-1
M											5	-2	-2	0	-1	-1	-1	1	-1	-1
N												6	-2	0	0	1	0	-3	-4	-2
P													7	-1	-2	-1	-1	-2	-4	-3
Q														5	1	0	-1	-2	-2	-1
R															5	-1	-1	-3	-3	-2
S																4	1	-2	-3	-2
T																	5	0	-2	-2
V																		4	-3	-1
W																			11	2
Y																				7

A – wrong – A to R is -1, *i.e.* less often than expected by chance

B – wrong – A to C is 0, but this means that it happens at the same rate as expected by chance, not that it never happens.

C – correct – because 11 is the largest number on the diagonal

D – wrong – these elements are log odds scores, not rates.

11. The scores in BLOSUM62 are measured in half bits. The score of D against E is 2.

Therefore:

- A. D is aligned with E twice as often as expected by chance.
- B. D is aligned with E four times as often as expected by chance.
- C. D is aligned with E $\ln 2$ times as often as expected by chance.
- D. D is aligned with E e^2 times as often as expected by chance.

If the scores are in half bits, the relative probability for a score S is $2^{S/2}$. If S = 2, we have relative probability = 2^1 , *i.e.* A is correct.