

MS2a, Week 3, Model Solution

Rune Lyngsø

November 2, 2011

A Phylogeny Reconstruction

- a. Consider a binary character – for convenience denote the two possible states 0 and 1 – evolving according to the rate matrix

$$Q = \begin{bmatrix} -\alpha & \alpha \\ \alpha & -\alpha \end{bmatrix}$$

Determine $P(t) = e^{Qt}$.

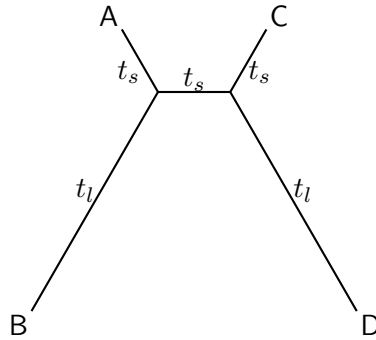
This is very similar to the problem on the problem sheet for week 2, except that we only have two possible characters. We can proceed in the same manner as then, or we could observe that the characteristic polynomial is $(-\alpha - \lambda)^2 - \alpha^2 = \lambda^2 + 2\alpha\lambda$. Hence Q has eigenvalues 0 and -2α . This immediately tells us that Q has two independent eigenvectors and thus can be diagonalised, which again means that

$$P(t) = e^{Qt} = B \begin{bmatrix} 1 & 0 \\ 0 & e^{-2\alpha t} \end{bmatrix} B^{-1}$$

for some properly chosen matrix B . Without even finding B we know that $P(t)_{00} = a + be^{-2\alpha t}$ for some $a, b \in \mathbf{R}$. We further know that $P(0)_{00} = 1 \Rightarrow a + b = 1$ and $P'(0)_{00} = -\alpha \Rightarrow b = 1/2$. We can conclude that $a = b = 1/2$. Identical reasoning can be used for $P(t)_{11}$, and $P(t)_{01} = 1 - P(t)_{00}$ and $P(t)_{10} = 1 - P(t)_{11}$ completes the picture to give

$$P(t) = \begin{bmatrix} \frac{1}{2} + \frac{1}{2}e^{-2\alpha t} & \frac{1}{2} - \frac{1}{2}e^{-2\alpha t} \\ \frac{1}{2} - \frac{1}{2}e^{-2\alpha t} & \frac{1}{2} + \frac{1}{2}e^{-2\alpha t} \end{bmatrix}$$

- b. Assume that we have a sequence of this binary character evolving on the following tree



with observed sequences A, B, C, and D. Let t_s be chosen such that $P(t)_{01} = 1/25$ and t_l be chosen such that $P(t)_{01} = 1/4$. What are the values of αt_s and αt_l meeting this requirement?

$$P(t_s)_{01} = \frac{1}{2} - \frac{1}{2}e^{-2\alpha t_s} = \frac{1}{25} \Rightarrow \alpha t_s = \ln \frac{5}{\sqrt{23}}$$

$$P(t_l)_{01} = \frac{1}{2} - \frac{1}{2}e^{-2\alpha t_l} = \frac{1}{4} \Rightarrow \alpha t_l = \ln \sqrt{2}$$

- c. What are the probabilities of observing each of the 16 possible combinations of the binary character at the four sequences, *i.e.* the probability of observing a 0 in all four sequences, a 0 in sequences A, B, and C and a 1 in sequence D, *etc.*?

Due to the symmetry of the evolutionary model, the patterns come in pairs with identical probability, *e.g.* $p_{0100} = p_{1011}$, so we only need to calculate probabilities for, say, patterns with 0 observed at A. With just two internal nodes it is as easy enumerating the four possible combinations of characters in these nodes as anything else. This gives

	00	00	00	00	p
0000:	$\frac{124416}{500000}$	$\frac{72}{500000}$	$\frac{72}{500000}$	$\frac{24}{500000}$	$\frac{124584}{500000}$
0001:	$\frac{41472}{500000}$	$\frac{216}{500000}$	$\frac{24}{500000}$	$\frac{72}{500000}$	$\frac{41784}{500000}$
0010:	$\frac{5184}{500000}$	$\frac{1728}{500000}$	$\frac{3}{500000}$	$\frac{576}{500000}$	$\frac{7491}{500000}$
0011:	$\frac{1728}{500000}$	$\frac{5184}{500000}$	$\frac{1}{500000}$	$\frac{1728}{500000}$	$\frac{8641}{500000}$
0100:	$\frac{41472}{500000}$	$\frac{24}{500000}$	$\frac{216}{500000}$	$\frac{72}{500000}$	$\frac{41784}{500000}$
0101:	$\frac{13824}{500000}$	$\frac{72}{500000}$	$\frac{72}{500000}$	$\frac{216}{500000}$	$\frac{14184}{500000}$
0110:	$\frac{1728}{500000}$	$\frac{576}{500000}$	$\frac{9}{500000}$	$\frac{1728}{500000}$	$\frac{4041}{500000}$
0111:	$\frac{576}{500000}$	$\frac{1728}{500000}$	$\frac{3}{500000}$	$\frac{5184}{500000}$	$\frac{7491}{500000}$

- d. For sequence length $n \rightarrow \infty$ what tree would you expect to be preferred by the parsimony method, *i.e.* the tree requiring the fewest character changes when summed over all sequence positions?

The pattern 0011 supports the tree grouping sequences A and B, the pattern 0101 supports the tree grouping sequences A and C together, while the pattern 0110 supports the tree grouping sequences A and D together. The remaining patterns (with 0 observed at A) do not distinguish between the three possible topologies. The 0101 pattern will occur the most for sufficiently long sequences (almost twice as many occurrences as the pattern 0011 and more than three times as often as the pattern 0110) so we would expect the parsimony method to choose the tree grouping sequences A and C together.

- e. Write an expression in terms of the probabilities of the 16 possible character combinations (*i.e.* your variables will be $p_{0000}, \dots, p_{1111}$) that should be maximised to find the phylogeny the maximum likelihood method will converge to for $n \rightarrow \infty$?

The likelihood is just the probability of the data given the parameters (topology and rates, here represented by the probabilities of the 16 possible character combinations). For $n \rightarrow \infty$ we know that the fraction of columns observed with a particular character combination will converge to the probability of observing that combination, which we computed above. So the likelihood expression we should maximise is

$$\left(p_{0000}^{124584} p_{0001}^{41784} p_{0010}^{7491} p_{0011}^{8641} p_{0100}^{41784} p_{0101}^{14184} p_{0110}^{4041} p_{0111}^{7491} p_{1000}^{7491} p_{1001}^{4041} p_{1010}^{14184} p_{1011}^{41784} p_{1100}^{8641} p_{1101}^{7491} p_{1110}^{41784} p_{1111}^{124584} \right)^{n/500000}$$

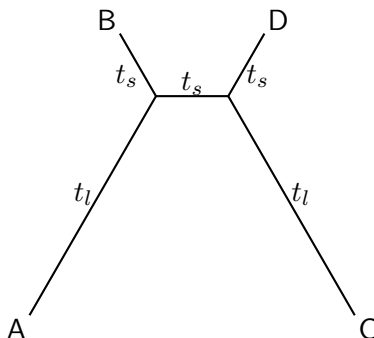
For the MLE we can ignore n as this will not change the location of the maximum (for non-negative likelihoods).

Without analytically solving for the MLE phylogeny, which phylogeny do you expect it to be?

We know that the maximum likelihood method converges to the true tree under this model, so the true phylogeny is the expected outcome. To further corroborate this claim, we can observe that the likelihood will be maximised for phylogenies where the probability of observing each character combination equals the observed fraction of this combination. This evidently holds for the phylogeny we have simulated

the data from. We can now write an equation system for this identity, and possibly using the assistance of *e.g.* Maple establish that the true phylogeny is the only solution to this equation system for probabilities of observed changes restricted to be between 0 and 1/2.

- f. Assume now that half the positions of our sequence evolve on the tree above, and half the positions evolve on the following tree



that is the tree where A and C sit at the end of long branches instead of B and D. What is now the probability of observation of the 16 possible patterns of character states at the four sequences?

Again it suffices to consider combinations with a 0 observed at sequence A. Apart from that, we just need to take the average of two probabilities from the model where only one tree was assumed, *e.g.* we get the probability of the pattern 0010 by taking the average of the probabilities of observing the patterns 0010 and 0001 under the one tree model. For some patterns, the two patterns averaged over are the same. The results are summarised by the following table:

0000	0001	0010	0011	0100	0101	0110	0111
0000	0001/0010	0010/0001	0011	0100/0111	0101	0110	0111/0100
$\frac{249,168}{1,000,000}$	$\frac{49,275}{1,000,000}$	$\frac{49,275}{1,000,000}$	$\frac{17,282}{1,000,000}$	$\frac{49,275}{1,000,000}$	$\frac{28,368}{1,000,000}$	$\frac{8,082}{1,000,000}$	$\frac{49,275}{1,000,000}$

What is the tree expected to be preferred by the parsimony method?

The probabilities of the three patterns distinguishing between the three possible topologies is unchanged from the previous case, so the parsimony method will still postulate the wrong topology of grouping sequence A with sequence C.

- g. If you were told that the correct topology is in fact not the topology the maximum likelihood method will converge to, which of the alternate topologies would be your guess for the one the maximum likelihood method converges to instead?

If we know the right topology is not the MLE, then the only two remaining possibilities are the one that group sequence A with sequence C and the one that group sequence A with sequence D. Sequence A and Sequence D are not evolutionary close in either tree, so the best guess would be that the MLE groups sequence A and sequence C. This is indeed what happens. The equations are beyond Maple's capabilities, but the `Dnaml` program from the PHYLIP phylogeny inference package consistently infers a tree topology grouping sequences A and C.

B Recombination

- h. Can we find a tree for the data set

Pan	TTATCC
Gorilla	TTGTTC
Pongo	CCACCC
Hylobates	CCGTTC

such that only one substitution is required in each position? If yes, provide such a tree. If no, why not?

We cannot. The first two sites require the tree to group Pan and Gorilla together, against Pongo and Hylobates, to be able to explain them with just one substitution. However, positions 3 and 5 require that we group Pan and Pongo together against Gorilla and Hylobates to explain them with just one substitution.

- i. Compute the minimum number of substitutions required for the above data set for each of the three possible unrooted tree topologies, e.g. by using Fitch's algorithm (or just eye-balling it if you feel confident about doing this).

Independent of topology, site 4 will always require 1 substitution and site 6 will not require any substitutions. For the remaining four sites, the cost depends on whether the topology has the right grouping into pairs. Hence, we get the following minimum number of substitutions:

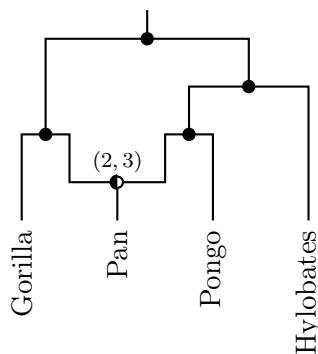
	Site	1	2	3	4	5	6	Total
Pan								
	Pongo							
		1	1	2	1	2	0	7
Gorilla								
	Hylobates							
Pan								
	Gorilla	2	2	1	1	1	0	7
Pongo								
	Hylobates							
Pan								
	Pongo	2	2	2	1	2	0	9
Hylobates								
	Gorilla							

- j. Assume that apart from substitutions, you are also allowed events arbitrarily changing the tree topology between consecutive sites (this is a simplification of recombination events – recombination events only allow certain changes to tree topology). What is the minimum number of events you need to explain the above data set.

If we start with the topology grouping Pan with Gorilla, but between sites 2 and 3 swap to the topology grouping Pan with Pongo, we get a total of 5 substitutions and one change of topology for a total of 6 events.

Give an ancestral recombination graph explaining the data set with this number of events.

There are many possible ancestral recombination graphs possible, but if we assume that Pan is the recombinant species with closest relative Gorilla for the first two positions and closest relative Pongo for the remaining positions, we get the following



where the partly filled node represents the recombination. At a recombination node we choose the branch in the direction of the filled half of the node for positions less than or equal to the first position in the pair

indicating the recombination point, otherwise we choose the branch in the direction of the hollow half of the node.

- k. How many recombination nodes are there in the ancestral recombination graph (ARG) you constructed in j?

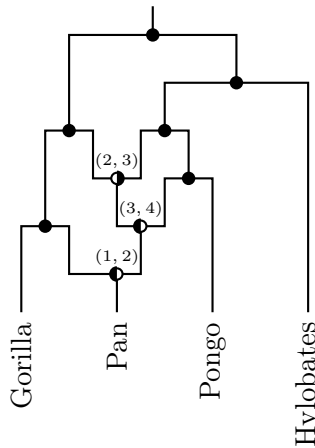
There is one recombination node in the ARG above.

Can you construct a data set by permuting the columns in the above data set that requires more recombination nodes for any ARG explaining it? If yes, give an example.

One example would be the data set obtained by swapping columns 2 and 3:

Pan	TATTCC
Gorilla	TGTTTC
Pongo	CACCCC
Hylobates	CGCTTC

Now the first site requires that Pan and Gorilla are grouped together, the second site that Pan and Pongo are grouped together, the third site that Pan and Gorilla are grouped together, and the fifth site that Pan and Pongo are grouped together. Hence, we will need recombinations between sites 1 and 2, between sites 2 and 3, and between sites 3 and 5. An ARG yielding this relationship would be



How many different marginal trees does the ARG you constructed in j have (a marginal tree is the tree relating the species at a particular position)?

For positions 1 and 2 we get the tree grouping Pan and Gorilla, while for positions 3 to 6 we get the tree grouping Pan and Pongo. This gives two marginal trees.

Can you construct a data set by permuting the columns in the above data set that has more different marginal trees in any ARG explaining it? If yes, give an example.

All informative sites in the data set groups Pan with either Gorilla or Pongo. Hence we do not need other marginal trees to explain the data. We can always construct an ARG with any given sequence of marginal trees: first add recombinations nodes above each species to split its sequence into single positions, then connect the lineages corresponding to the same position according to the marginal tree desired for that position, and finally connect the lineages for each position in arbitrary order. It follows that any permutation of the columns will allow an ARG with just the two marginal trees of the ARG in j .