

MS2a, Week 1

Rune Lyngsø

October 13, 2011

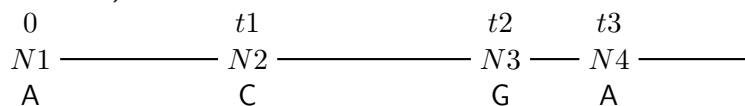
A From rates to probabilities

Describing evolution in terms of rates that describes what happens in a very short time interval is easy, but what is needed is a description of what happens during longer time interval. We only observe sequences at the leaves of a phylogeny, and what happens at the internal branches is hidden. Several, or even numerous events, can have taken place, while still only being observable as the net result of one nucleotide changing into another. It is thus necessary to be able to go from descriptions of instantaneous events to the accumulated result over a time interval.

We model nucleotide evolution as a continuous time discrete Markov process. Let us assume that all nucleotides evolve at the same rate and that one jumps to the alternative nucleotides with the same probability, known as the Jukes-Cantor model. More realistic models exist, but though the mathematics become more involved, conceptually they are equivalent to the simple model we assume here. The rate matrix of our model is parameterised by α , the rate of change from one particular nucleotide to another, and has the following form:

		To			
		A	C	G	T
	A	-3α	α	α	α
	C	α	-3α	α	α
	G	α	α	-3α	α
	T	α	α	α	-3α

The trajectory of a process determined by Q starting at time t in state N_1 (for instance A) could look like this



The probability that we have to wait more than time T from the change to $N_{(i-1)}$ until it again changes to N_i is $P\{t_i - t_{i-1} > T\} = e^{-3\alpha T}$, i.e. exponentially

distributed with intensity 3α . The expected number of events (substitutions) in a time interval of length t is $3\alpha t$. Let $P(t)$ denote the matrix of probabilities of change over time t , i.e. the (i, j) entry in $P(t)$, $P_{i,j}(t)$, is the probability that nucleotide i has changed into nucleotide j after time t . We know that $P(t) = e^{tQ} = I + tQ + t^2Q^2/2 + \dots + t^kQ^k/k! + \dots$ (where $I = Q^0$ is identity matrix). Note that if t is small then $I + tQ$ is a good approximation to $P(t)$, which corresponds to all substitutions being observable because there is at most one event in the evolutionary trajectory. For this simple model of substitution, one can realise that $Q^i = (-4)^{i-1}$ for $i \geq 1$. With a bit of rearrangements of the exponential expansion, this yields

$$P_{i,j}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-4\alpha t} & \text{if } i = j \\ \frac{1}{4} - \frac{1}{4}e^{-4\alpha t} & \text{if } i \neq j \end{cases}$$

by converting the exponential expansion on matrix powers to an exponential expansion on a real number.

- What is the probability of observing (A, A, T) in an alignment column
- If there on average is 10^{-8} substitutions/(position \times year), how many events would you expect in 1000 base pair long sequences that had a common ancestor 5 million years ago?

How many differences would you expect to observe between the two sequences (assuming they are only affected by the substitution process, and not for example insertions and deletions)?

What would your answers be if they had a common ancestor 50 million years ago? How would the approximation $I + Qt$ be in the two cases?

- Kimura's 2 parameter substitution model is a slightly more realistic model than Jukes-Cantor, allowing different substitution rates between transitions (when an A changes to a G or vice versa, or when a C changes to a T or vice versa) and transversions (all other changes). The rate matrix can be written as

		To			
		A	G	C	T
From	A	$-\alpha - 2$	α	1	1
	G	α	$-\alpha - 2$	1	1
	C	1	1	$-\alpha - 2$	α
	T	1	1	α	$-\alpha - 2$

where the four nucleotides have been sorted so we first have the purines and then the pyrimidines. It may look like this only has a single parameter,

α , but as we cannot separate rates and time we have chosen a unit of time corresponding to one transversion to simplify the matrix – if transversion rate was β , we can divide all matrix entries by β if we just use βt as time instead of t and let α denote the relative difference between transition and transversion rate.

This matrix is slightly harder to exponentiate by finding regularities in the powers of Q , compared to the Jukes–Cantor model. However, it can still be done. Calculate Q^2 , Q^3 , Q^k (Hint: try to write the entries depending on α as a power of $(\alpha + 1)$ and term not depending on α) and then $P(t)$.

B From probabilities to rates

Assume we have observed two sequences that have evolved from a common ancestor under just the Jukes–Cantor model of nucleotide substitution, i.e. we know they are correctly aligned as

```

A C G T T G A C C T C A A A T T T G C T C T
A C G G T G A C G T C A C A A A T G C A C T

```

- d. Write (plot if you can) likelihood function of this alignment as a function of αt , assuming that different positions evolve independently.
- e. What is the αt that makes the alignment most likely?