

# Stochastic Modelling of Molecular Evolution

Jotun Hein & Rune Lyngsø

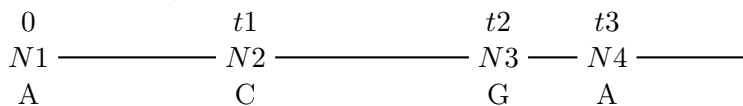
## From rates to probabilities

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 intervals. 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 having changed 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  $\alpha$ , the rate of change from one particular nucleotide to another, and has the following form:

		To			
		A	C	G	T
From	A	$-3\alpha$	$\alpha$	$\alpha$	$\alpha$
	C	$\alpha$	$-3\alpha$	$\alpha$	$\alpha$
	G	$\alpha$	$\alpha$	$-3\alpha$	$\alpha$
	T	$\alpha$	$\alpha$	$\alpha$	$-3\alpha$

The trajectory of a process determined by  $Q$  starting at time  $t$  in state  $N1$  (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  $Ni$  is  $P\{t_i - t_{i-1} > T\} = e^{-3\alpha T}$ , *i.e.*

exponentially distributed with intensity  $3\alpha$ . 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  $\alpha t$  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.

1. Calculate  $Q^2$ ,  $Q^k$  and then  $P(t)$ .
2. What is the probability of observing (A, A, T) in an alignment column?
3. 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?

### 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
```

5. Write (plot if you can) likelihood function of this alignment as a function of  $\alpha t$ , assuming that different positions evolve independently.
6. What is the  $\alpha t$  that makes the alignment most likely.