

Week 1 Rates (Q) and transition probabilities, $P(t)$.

A. From Q to P

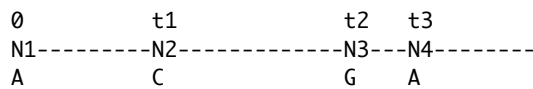
Describing evolution in terms of rates that describes what happens in a very short time interval is easy, but what is needed is a description of what happens during longer time interval as you only observe sequences at the beginning and end of branches on a phylogeny.

Nucleotide evolution will be modelled by 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.

The corresponding rate matrix (Jukes-Cantor) will be (α)

		To				
		A	C	G	T	
F	A !	-3 α	α	α	α	
	!					
R	C !	α	-3 α	α	α	= Q
	!					
O	G !	α	α	-3 α	α	
	!					
M	T !	α	α	α	-3 α	

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



$P\{t_i - t_{i-1} > T\} = e^{-3\alpha t}$, ie exponentially distributed with intensity 3α . The expected number of events (substitutions) in a time interval of length t is $3\alpha t$.

The probability that a nucleotide I has changed into j after time t is the (i,j) entry in $P(t)$, ie $P_{i,j}(t)$. $P(t) = e^{tQ} = I + tQ + \frac{t^2 Q^2}{2} + \dots + \frac{t^k Q^k}{k!} + \dots$ (4 * 4 matrix where $I=Q^0$ is identity matrix). Note that if αt is small then $I+tQ$ is a good approximation to $P(t)$, which corresponds to all substitutions are observable because there aren't more than 1 event in the evolutionary trajectory.

1. Calculate Q^2 , Q^k and then $P(t)$.

(Ie the probability that that the nucleotide is in a new state after time t is $3 * P_{i,j}(t)$!!)

2. Is this process time reversible?

3. What is the probability of observing (A,A,T) in an alignment column?

4. If there on average is 10^{-8} substitutions/(position*year), How many events would you expect in 1000 base pair long sequences that had a common ancestor 5 mill. Years ago? How many differences would one expect to observe? What if they had a common ancestor 50 mill year ago. How would the approximation $I + tQ$ be in the two cases?

From P to Q :

Assume we have observed two sequences nicely aligned:

```

ACGTTGACCTCAAATTTGCTCT
ACGGTGACGTCACAAATGCACT
    
```

5. Write (plot if you can) likelihood function of this alignment as a function of αt , assuming that different positions evolve independently.

6. What is the αt that makes the alignment most likely.