# Computational Biology HW 2

Probst Jennifer, 16703423

26. October 2020

## Question 1

Nucleotide substitutions can either be transitions or transversions. Transition are a little more likely, occur between T and C and therefore balance out. Transversions from T or C to A or G have even a lower probability. Therefore we expect the distribution of the 10 species' sequences to stay more or less the same with about 50% T and C.

## Question 2

Now transvertions to A or G are highly likely to happen, so this will result in the stationary distribution of the substitution model over time. The equilibria can vary a lot depending on the model used. This is for example roughly 25% of each nucleotide for the JC69 model, where no particular nucleotide is favoured.

## Question 3

In a stationary distribution the matrix $P(t) = \exp(Q * t)$ would look like

$$\text{Ps} = \begin{pmatrix} 0.22 & 0.26 & 0.33 & 0.19 \\ 0.22 & 0.26 & 0.33 & 0.19 \\ 0.22 & 0.26 & 0.33 & 0.19 \\ 0.22 & 0.26 & 0.33 & 0.19 \end{pmatrix}$$

where the entries of column i are the element pi[i] of the vector of equilibrium frequencies, which is in our case $pi = (0.22, 0.26, 0.33, 0.19)$. I wrote a function that checks after what time t this is reached exactly for each value in P(t), which is at t = 1000 mya. If we define an error measure $e = sum(abs(P(t) - Ps)) < 1e - 10$ which is the sum of the difference of the distribution at time t to the stationary distribution, we result in 844 mya.

## Question 4

We know from the exercise that the overall rate of change from a nucleotide i to any other nucleotide is $-q_{ii}$. This we denote as $\lambda$. Now the exponential distribution f with rate $\lambda$; $f(t) = \lambda * e^{-\lambda * t}$; models the probability for a substitution at time t. We can draw samples from f to get the time when the next substitution happens.

## Question 5

Look in the row of the chose nucleotide i and the entries $q_{ij}$ with $i \neg j$, which denote the rate of change from nucleotide i to j. To sample the nucleotide it is substituted by, we draw from the 3 different nucleotide probabilities $P(i- > j) = q_{ij}/(q_{ij1} + q_{ij2} + q_{ij3})$.