Week 8

In this problem, you will study a simple DNA mutation model. One goal of such models is to study the *phylogenetic distance*, i.e. the overall amount of genetic difference between the genomes of two species (e.g. between modern humans and one of their prehistoric ancestor, or a distant relative such as chimpanzees). This distance helps biologists better understand evolutionary relationships of species.

Suppose we have sequenced a strand of DNA from a chromosome of a human and have for comparison a strand of DNA that corresponds to an ancestor of our species. Let N be the unknown number of generations that have passed between the ancestral strand and the contemporary strand. For any position in the sequence, the nucleotide must be either A, G, C, or T. By comparing the ancestral and contemporary strands, we can measure the fraction of DNA sites that are different between the two strands. Let's call this value β. Our task is to construct a mathematical model that somehow connects the measured value of β with the unknown number of generations N.

Let α be the probability of a mutation at one site in one generation.

Let $p_A(n)$, $p_G(n)$, $p_C(n)$, and $p_T(n)$ be the probabilities of each given nucleotide at a particular site in the n-th generation. Mutations from one generation to the next change these probabilities, and we must quantify these changes. Let us consider a nucleotide A at a site in the n-th generation The assumption of the Cantor-Jukes model (1969) is that **all possible changes are equally likely**, i.e. if the current nucleotide at the considered site is A:

- the mutations $A \to G$, $A \to T$ and $A \to C$ occur all with the same rate
- the total rate of mutation at this site is α.

Similar assumptions hold if the current nucleotide is G, C or T.

The goal of this exercise is to estimate the **phylogenetic distance** $d = \alpha N$ for this particular model, knowing only the β parameter, since neither α nor N are easily known when comparing two distant species.

1) Using the above assumptions, write $p_A(n+1)$ as a function of $p_A(n)$, $p_G(n)$, $p_C(n)$, $p_T(n)$ and α.

2) Build the matrix $\boldsymbol{M}$ s.t.

$$\boldsymbol{v}_{n+1} = \begin{pmatrix} p_A(n+1) \\ p_G(n+1) \\ p_C(n+1) \\ p_T(n+1) \end{pmatrix} = \boldsymbol{M}\boldsymbol{v}_n = \boldsymbol{M} \begin{pmatrix} p_A(n) \\ p_G(n) \\ p_C(n) \\ p_T(n) \end{pmatrix},$$

and express $\boldsymbol{v}_N$ as a function of $\boldsymbol{M}$ and $\boldsymbol{v}_0$.

3) What are the eigenvalues and associated eigenvectors of $\boldsymbol{M}$? Hint: the eigenvalues of a circulant matrix $\boldsymbol{C}$ s.t.

$$\boldsymbol{C} = \begin{pmatrix} c_0 & c_{n-1} & c_{n-2} & ... & c_1 \\ c_1 & c_0 & c_{n-1} & ... & c_0 \\ \vdots & & \ddots & & \vdots \\ c_{n-1} & c_{n-2} & c_{n-3} & ... & c_0 \end{pmatrix},$$

are

$$\lambda_k = c_0 + c_{n-1}\omega_k + c_{n-2}\omega_k^2 + ... + c_1\omega_k^{n-1},$$

with $k = 0, 1, ..., n-1$ and $\omega_k = e^{\frac{2ik\pi}{n}}$.

4) Knowing that the change rate over N generation is β and making the same assumptions regarding probability of mutations over N generations, i.e. if considering a site holding a A:

- the mutations $A \to G$, $A \to T$ and $A \to C$ occur all with the same rate
- the total rate of mutation over N generations at this site is β.

Similar assumptions hold if the current site nucleotide is G, C or T.

Build a matrix $\boldsymbol{P}$ depending only on β s.t. $\boldsymbol{v}_N = \boldsymbol{P}\boldsymbol{v}_0$

5) Let us consider the vector $\boldsymbol{u} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$. Write $\boldsymbol{u} = a_1\boldsymbol{v}_1 + a_2\boldsymbol{v}_1$ where $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are eigenvectors of M, $\boldsymbol{v}_1$ being associated with eigenvalue 1 and $\boldsymbol{v}_2$ with another eigenvalue $\lambda \neq 1$, and $a_1, a_2 \in \mathbb{R}$.

Compute $\boldsymbol{M}^N\boldsymbol{u}$.

6) Using the fact that $\boldsymbol{P}\boldsymbol{u} = \boldsymbol{M}^N\boldsymbol{u}$, compute N as a function of β and α.

Finally compute the Cantor-Jukes distance $d = \alpha N$ as a function of β.

Hint: You can assume that α is small (i.e. that only a few number of mutations occur during one generation) and that $\ln(1-x) \approx -x$, for small values of x.

7) Using

$$\boldsymbol{v}(t) = \begin{pmatrix} p_A(t) \\ p_G(t) \\ p_C(t) \\ p_T(t) \end{pmatrix},$$

and the approximation

$$\frac{dp_A(t)}{dt} \approx p_A(n+1) - p_A(n),$$

express the linear system from question 1. as a first order differential equation,

$$\frac{\boldsymbol{v}(t)}{dt} = \boldsymbol{Q}\boldsymbol{v}(t)$$

- What is the relationship between $\boldsymbol{M}$ and the matrix $\boldsymbol{Q}$ involved in this differential equation ?
- What is the relationship between the eigenvalues and eigenvectors of $\boldsymbol{M}$ and those of $\boldsymbol{Q}$?
- What happens to $\boldsymbol{v}(t)$ when the initial probabilities are uniform ?