While BLAST is not the only tool available for sequence homology searches, it is still the most commonly used. Throughout this homework, you will work on alignments without indels, just to keep things simpler.

1. **Simulating Alignments**

   (a) Choose a stationary distribution $\boldsymbol{\pi} = (\pi_{\mathrm{A}}, \pi_{\mathrm{C}}, \pi_{\mathrm{G}}, \pi_{\mathrm{T}})$ on the nucleotides, and a set of five (positive) rates $\gamma_{\mathrm{AC}}, \gamma_{\mathrm{AG}}, \gamma_{\mathrm{AT}}, \gamma_{\mathrm{CG}}, \gamma_{\mathrm{CT}}$). Arrange them in a matrix

$$\boldsymbol{A} = \begin{pmatrix} -- & \gamma_{\mathrm{AC}}\pi_{\mathrm{C}} & \gamma_{\mathrm{AG}}\pi_{\mathrm{G}} & \gamma_{\mathrm{AT}}\pi_{\mathrm{T}} \\ \gamma_{\mathrm{AC}}\pi_{\mathrm{A}} & -- & \gamma_{\mathrm{CG}}\pi_{\mathrm{G}} & \gamma_{\mathrm{CT}}\pi_{\mathrm{T}} \\ \gamma_{\mathrm{AG}}\pi_{\mathrm{A}} & \gamma_{\mathrm{CG}}\pi_{\mathrm{C}} & -- & 1\pi_{\mathrm{T}} \\ \gamma_{\mathrm{AT}}\pi_{\mathrm{A}} & \gamma_{\mathrm{CT}}\pi_{\mathrm{C}} & 1\pi_{\mathrm{T}} & -- \end{pmatrix},$$

   where the diagonal entries are set such that the rows sum to 0. Rescale the matrix such that

$$\boldsymbol{A}_s = -\frac{\boldsymbol{A}}{\sum_{\mathrm{N}\in\{\mathrm{A,C,G,T}\}} \pi_{\mathrm{N}} a_{\mathrm{NN}}},$$

   where a_{NN} is a diagonal entry in $\boldsymbol{A}$. (This operation guarantees that time is measured in expected number of mutations per site.) Compute the matrix exponential

$$P(0.01) = e^{\boldsymbol{A}_s 0.01}$$

   (`expm()` in R; `scipy.linalg.expm` in python). (Sorry I cannot justify this more in class, but I will post supplementary–not examined–CTMC notes that explain why it works.)

   (b) Simulate three aligned sequences (no gaps) of lengths 50, 100, and 1000 using $P^n(0.01)$ for your choice of n.

   (c) Simulate another three sequences with the same lengths but use $P^{2n}(0.01)$.

   Turn in all simulated sequences as fasta files with your solution. Give the sequences self-explanatory names.

2. **Alignment Scoring.**

   (a) Using an appropriate blastn scoring scheme from the data simulated in 1 (b) (click the question mark beside Match/Mismatch Scores at the NCBI blastn website to learn how to choose appropriately for your simulated data), write code to estimate the λ, C and A needed to compute the E value for the highest scoring segment in each of your three sequences. Your method will use simulation, but there will be one simulation, the results of which are applicable to *all* three simulated sequences. (This is what allows such a method to scale to the scope of database searches, where you don't have time to do a custom simulation for every query sequence.)

   (b) Now score the three sequences simulated in 1 (c), but use the same parameters you estimated in Part (a) of this problem. Are you more or less convinced that the sequences are homologous?

   (c) How would the E values of Part (a) change quantitatively if the simulated alignment of length 100 was the best local alignment found when blasting a query sequence of length 1000 against a database of the current size of GenBank? Suppose the database sequence producing the best match was of length 10000.