1 Reminder

Recall we defined a matrix $R$ containing the rates of nucleotide base transformations. $R$ suffices that the sum of every row equals 0 (i.e. every row is the rates of change in distribution so the sum remains constant) and defined as:

$$ R = \frac{\partial P(0)}{\partial t} = \lim_{\epsilon \to 0} \frac{P(\epsilon) - P(0)}{\epsilon} = \lim_{\epsilon \to 0} \frac{P(\epsilon) - I}{\epsilon} $$

(1.1)

Note $[P(t)]_{ij}$ is the probability to transfer from $i$ to $j$ given $t$ (another notation: $P(i \xrightarrow{t} j)$) we get:

$$ P(\epsilon) \approx I + \epsilon R $$

(1.2)

$$ P(t) \xrightarrow{taylor\ series} I + tR + \frac{(tR)^2}{2!} + \frac{(tR)^3}{3!} + ... $$

(1.3)

As $t \to \infty$ the probability to undergo substitution increases, resulting in the decrease of probability to stay the same at one position (assuming $R_{ij} \neq 0$ for any $i, j$). We would like to find a model for $R$ satisfying this demands. In other words we would like a matrix $R$ that is:

- Negative on the main diagonal
- Positive elsewhere

2 The Jukes-Cantor Model

The simplest substitution model was proposed in 1969 by Thomas Jukes and Charles Cantor\textsuperscript{1}. The single-parameterized matrix takes the form:

$$ R_{JC} = \begin{bmatrix}
-3\alpha & \alpha & \alpha & \alpha \\
\alpha & -3\alpha & \alpha & \alpha \\
\alpha & \alpha & -3\alpha & \alpha \\
\alpha & \alpha & \alpha & -3\alpha
\end{bmatrix} $$

which means that there is equal rate to change to any other nucleotide. We would like to develop $P_{JC}(t)$ according to $R_{JC}$:

$$ P_{JC}(t) = e^{tR_{JC}} $$

(2.1)

\textsuperscript{1}https://en.wikipedia.org/wiki/Models_of_DNA_evolution#JC69_model(Jukes_and_Cantor, 1969)[2]
In the previous lesson we saw the following differential equation:

\[ P(t + \epsilon) = P(t) \cdot P(\epsilon) \approx P(t) + P(t) \epsilon R \] (2.2)

\[ \frac{\partial P(t)}{\partial t} = \lim_{\epsilon \to 0} \frac{P(t + \epsilon) - P(t)}{\epsilon} = P(t) \cdot \lim_{\epsilon \to 0} \frac{P(\epsilon) - P(0)}{\epsilon} = P(t) \cdot \frac{\partial P(0)}{\partial t} \] (2.3)

So an expression for \( P(t) \) would be one that its derivative is equal to the above equation.

\( R \) and \( P(0) \) are symmetric suggesting we can give \( P(t) \) the following form:

\[
P(t) = \begin{bmatrix}
r & s & s & s \\
s & r & s & s \\
s & s & r & s \\
s & s & s & r \\
\end{bmatrix}
\]

Since every row denotes a probability distribution, we derive -

\[ r = 1 - 3s \] (2.4)

Therefore:

\[
\frac{\partial P(t)}{\partial t} = P(t) \cdot R_{JC} = \begin{bmatrix}
r & s & s & s \\
s & r & s & s \\
s & s & r & s \\
s & s & s & r \\
\end{bmatrix} \cdot \begin{bmatrix}
-3\alpha & \alpha & \alpha & \alpha \\
\alpha & -3\alpha & \alpha & \alpha \\
\alpha & \alpha & -3\alpha & \alpha \\
\alpha & \alpha & \alpha & -3\alpha \\
\end{bmatrix} = \begin{bmatrix}
r' & s' & s' & s' \\
s' & r' & s' & s' \\
s' & s' & r' & s' \\
s' & s' & s' & r' \\
\end{bmatrix}
\]

while

\[ r' = -3\alpha r + 3\alpha s^2 \] (2.5)

\[ s' = -3\alpha s + \alpha r + 2\alpha s^2 (1 - 3s) \alpha - s\alpha = \alpha(1 - 4s) \] (2.6)

To find the expression for \( P(t) \) we calculate the integral of the above expressions and get:

\[
[P_{JC}(t)]_{i,j} = \begin{cases} 
\frac{1}{4}(1 + 3e^{-4\alpha t}) & i = j \\
\frac{1}{4}(1 - e^{-4\alpha t}) & i \neq j 
\end{cases}
\]

In order to verify the solution satisfies the initial equations, we can differentiate it. For the case where \( i \neq j \):

\[
\frac{\partial}{\partial t} \frac{1}{4}(1 - e^{-4\alpha t}) = -\frac{1}{4}e^{-4\alpha t} \cdot (-4\alpha) = \alpha \cdot e^{-4\alpha t} \stackrel{(1)}{=} \alpha(1 - 4s) = s'
\]

(1) - verification: when placing \( (1 - 4s) \) back into the term of \( s \) instead of \( e^{-4\alpha t} \) we get: \( \frac{1}{4}(1 - (1 - 4s)) = s. \)

In a similar way we can verify the solution for \( i = j \) and get \( r' \).

To see how \( P_{JC} \) behaves, we can calculate its limits for different values of \( t \) -

\[
\lim_{t \to 0} P_{JC}(t) = \begin{bmatrix} 
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
\end{bmatrix} = I
\]

\[
\lim_{t \to \infty} P_{JC}(t) = \begin{bmatrix} 
\frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} \\
\frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} \\
\frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} \\
\frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} \\
\end{bmatrix}
\]
Visualization of the two limits as a function of $t$:

![Graph of $P_{JC}(t)$](image)

Figure 1: Graph of $P_{JC}(t)$

The upper curve describes the case where $i = j$ and the lower when $i \neq j$. The Constant function is $f(x) = \frac{1}{4}$. Note that when $t \to \infty$, $[P_{JC}(t)]_{i,j} = \frac{1}{4}$ for all $i$ and $j$, meaning that no matter what base we started from, the probability to transfer into any nucleotide is equal. This suggests that when waiting enough time, we practically forget where we come from (random walk). Hence, we can derive the Jukes-Cantor model does not take into consideration the different types of substitutions (transition and transversion) which occur on different probabilities (transitions are slightly likelier to happen). To account for this issue, a new model was proposed a few years later.

### 3 The Kimura Model

In 1980 Motoo Kimura\(^ 2\) proposed a rate matrix considering different rates for:

- Transitions: from purine to purine or from pyrimidine to pyrimidine ($A \leftrightarrow G$ or $C \leftrightarrow T$)
- Transversions: from purine to pyrimidine or vice versa

Denote $\alpha$ as the transition rate and $\beta$ as the transversion rate, the two-parameterized model is of the form:

$$R_k = \begin{bmatrix}
-2\beta - \alpha & \beta & \alpha & \beta \\
\beta & -2\beta - \alpha & \beta & \alpha \\
\alpha & \beta & -2\beta - \alpha & \beta \\
\beta & \alpha & \beta & -2\beta - \alpha
\end{bmatrix}$$

Using the same procedure as before, we can find a general time-dependent form for $P$.

Similar to Jukes-Cantor matrix, when $t \to \infty$, all $P_k(t)$ values converge to $\frac{1}{4}$.

---

4 Efficiency

Since in both cases $R$ is squared, real and symmetric, it can be diagonalized. By the spectral theorem we can write $R = UDU^T$, where $U$ is unitary and $D$ is diagonal. Now, the power of $R$ (and of $tR$ as a result) takes a simpler form. For instance:

$$(tR)^2 = UD^2U^T$$

and generally $(tR)^n = U D^n U^T$. Since $D$ is diagonal, this means we only have to power the eigenvalues down the diagonal when calculating the approximation of $P(t)$.

5 Ergodicity

When receiving a matrix $R$ containing real substitution rates concerning two sequences, unlike the theoretical models we’ve seen so far, it is not guaranteed that $P(t)$ will converge. We would like to phrase a restriction on such matrices in order to force convergence. It appears that ergodicity is a sufficient restriction.

**Definition:** $P(t)$ is ergodic if $\forall_{a,b} \forall_{t > 0} \ [P(t)]_{a,b} > 0$.

So for two nucleotides $a, b$ and any given time $t > 0$, the probability to pass from one state to another is strictly larger than 0.

**Theorem:** If $R$ defines an ergodic Markov process then -

$$P(t) \xrightarrow{t \to \infty} \begin{bmatrix} \pi & \pi \\ \pi & \pi \\ \pi & \pi \end{bmatrix}$$

$\pi$ is the stationary distribution of $P(t)$, meaning it remains unchanged in the Markov chain as time progresses (a fixed point). Given $P(t)$ the following holds:

$$\pi [P(t)] = \bar{\pi}$$

so $\bar{\pi}$ is an eigenvector of $[P(t)]^T$ with eigenvalue 1. How can we prove there exists such vector for $[P(t)]$ as well? Since every matrix and its transpose has the same characteristic polynomial, they both have the same eigenvalues, suggesting that $[P(t)]$ indeed has an eigenvector with eigenvalue 1.

6 Probabilistic Models of Evolution

Given a tree, we would like to be able to calculate its likelihood when considering it as a hidden Markov model. Moreover we would like to calculate the likelihood of a certain sequence (also given a tree $T$). The problems we will account for during this subject:

- Branch lengths given $T$
- Tree topology
- Sequence of a certain ancestor

Say we have a rooted tree $T$ and the branch lengths $\{t_{ij}\}$ are given as well ($i$ and $j$ are neighbors in the tree).

**Notations and restrictions:**

**Observed:**

- Let $x_1, x_2, ..., x_n$ be the sequences in the tree leaves.
- Every node in $T$ is a sequence of length 1.
Hidden:

- Let $x_{n+1}, \ldots, x_{2n-2}$ be the sequences in the internal nodes of $T$.
- Let $x_r = x_{2n-1}$ be the sequence of the root.

Now a formal form of our questions according to the notations will be:

- **Likelihood** - $P(x_1, \ldots, x_n)$: the probability to observe the leaves given the topology of the tree.
- **Posterior** - $P(x_i = a| x_1, \ldots, x_n)$ for $i \in \{n+1, \ldots, 2n-1\}$: the probability of a sequence for an ancestor given the leaves (note that $a$ is a single nucleotide since we require length of 1 for each sequence in $T$).
- **A-posterior of transition** - $P(x_i = a, x_j = b|x_1, \ldots, x_n)$: the probability of a mutation in a certain branch ($i$ is the father of $j$ or vice versa).
- **MPE** - $P(x_{n+1}, \ldots, x_{2n-1}| x_1, \ldots, x_n)$: given an observation over the leaves, what is the most probable explanation for the other nodes in $T$. 