Jukes-Cantor model

Simplest case: characters are nucleotides, each nucleotide is equally likely to mutate to any other.

Let's say our initial probability distribution over nucleotides at a site is:

$$\vec{p}_0 = \begin{pmatrix} p_A \\ p_C \\ p_G \\ p_T \end{pmatrix}$$

Let w_{ij} be the probability that a single substitution changes nucleotide from i to j:

$$w_{ij} = \begin{cases} 1/3 & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

we can write a matrix $\underline{W} = [w_{ij}] = \begin{pmatrix} 0 & 1/3 & 1/3 & 1/3 \\ 1/3 & 0 & 1/3 & 1/3 \\ 1/3 & 1/3 & 0 & 1/3 \\ 1/3 & 1/3 & 1/3 & 0 \end{pmatrix}$

After exactly one substitution, the new probability distribution over nucleotides is

$$\underline{W}\,\vec{p}_0$$

So what about after some amount of time? We don't know exactly how many substitutions have occurred, but assuming a molecular clock, the average will be μt where μ is substitution rate and t is time.

If substitutions are independent, then $\Pr(m|\mu t) = e^{-\mu t}\frac{(\mu t)^m}{m!}$

So:

$$\begin{aligned}
\vec{p}(t) &= \Pr(m=0|\mu t)\cdot\vec{p}_0 + \Pr(m=1|\mu t)\,\underline{W}\,\vec{p}_0 + \Pr(m=2|\mu t)\cdot\underline{W}^2\vec{p}_0 + \dots \\
&= \sum_{m=0}^{\infty} \Pr(m|\mu t)\,\underline{W}^m\,\vec{p}_0 \\
&= \sum_{m=0}^{\infty} e^{-\mu t}\frac{(\mu t)^m}{m!}\underline{W}^m\,\vec{p}_0 \\
&= e^{-\mu t}\left(\sum_{m=0}^{\infty}\frac{(\mu t\underline{W})^m}{m!}\right)\vec{p}_0 \\
&= e^{-\mu t}\,e^{\mu t\underline{W}}\,\vec{p}_0 \\
&= e^{\mu t(\underline{W}-\underline{I})}\,\vec{p}_0 = e^{-t\underline{P}}\,\vec{p}_0
\end{aligned}$$

where $\underline{P} = \underline{W} - \underline{I}$ is transition matrix or substitution matrix.

So for Jukes-Cantor model:

$$P_{ij} = \begin{cases} 1/3 & \text{if } i \neq j \\ -1 & \text{if } i = j \end{cases}$$

and

$$\vec{p}(t) = e^{+\mu t \underline{P}} \vec{p}_0$$

For this very simple $\underline{P}$, it turns out that there is an analytic expression for the matrix exponential:

$$\left[e^{\mu t \underline{P}}\right]_{ij} = \begin{cases} \frac{1}{4} + \frac{3}{4} e^{-\mu t} & \text{if } i = j \\ \frac{1}{4} - \frac{1}{4} e^{-\mu t} & \text{if } i \neq j \end{cases}$$

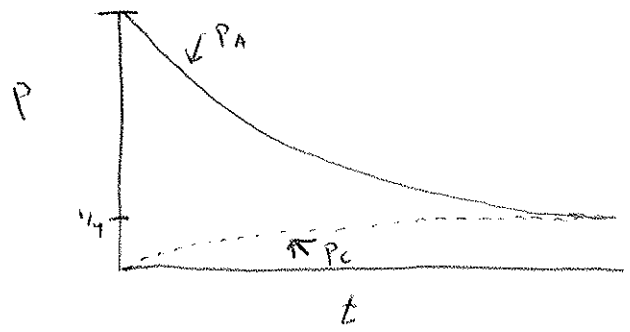
So let's say that

$$\vec{p}_0 = \begin{pmatrix} P_A = 1 \\ P_C = 0 \\ P_G = 0 \\ P_T = 0 \end{pmatrix}$$

Then at time t,

$$P_A(t) = \frac{1}{4} + \frac{3}{4} e^{-\mu t}$$

$$P_C(t) = P_G(t) = P_T(t) = \frac{1}{4} - \frac{1}{4} e^{-\mu t}$$



$e^a e^b = e^{a+b}$

Using Jukes-Cantor for maximum likelihood estimate:

AA —— μt —— AC

$$\Pr(AA \mid AC, \mu t) = \Pr(A \mid A, \mu t) \cdot \Pr(A \mid C, \mu t)$$

$$= \left(\tfrac{1}{4} + \tfrac{3}{4} e^{-\mu t}\right)\left(\tfrac{1}{4} - \tfrac{1}{4} e^{-\mu t}\right)$$

$$= \tfrac{1}{16}\left(1 + 3e^{-\mu t}\right)\left(1 - e^{-\mu t}\right)$$

$$= \tfrac{1}{16}\left[1 - e^{-\mu t} + 3e^{-\mu t} - 3e^{-2\mu t}\right]$$

$$= \tfrac{1}{16}\left(1 + 2e^{-\mu t} - 3e^{-2\mu t}\right)$$

$$2\Pr(AA \mid \mu t) = 0 = 2\mu t e^{-\mu t} + 6\mu t e^{-2\mu t} = 2\mu t\left(3e^{-2\mu t} + e^{-\mu t}\right)$$

$$e^{\mu t} = 3$$

Not worth the algebra, doesn't clarify the point

μt	$\Pr(AA \mid AC, \mu t)$
0	0
0.1	0.028
0.5	0.077
0.824	0.083
1.0	0.082
2.0	0.07
4.0	0.063

For not much data, likelihood is not sharply peaked. Would be more peaked with more data.

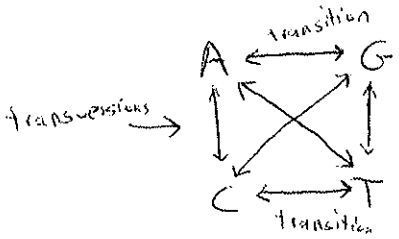
Even for best value of μt, likelihood $\ll 1$

More general substitution models

In practice, we typically use models that are more complex than Jukes Cantor and do the calculations using digital computing machines rather than pen and paper.

One example of such a model: HKY85. Captures 2 things not in Jukes-Cantor:

1)



] purines

] pyrimidines

For most polymerases, transitions are more common than transversions. Why?

2) Empirically, most genes have unequal frequencies of the four nucleotides:

$$\phi_A \neq \phi_C \neq \phi_G \neq \phi_T$$

HKY85 substitution matrix

	A	C	G	T
A	*	ϕ_C	$\times\phi_G$	ϕ_T
C	ϕ_A	*	ϕ_G	$\times\phi_T$
G	$\times\phi_A$	ϕ_C	*	ϕ_T
T	ϕ_A	$\times\phi_C$	ϕ_G	*

ϕ_x is equilibrium frequency of nucleotide x

$\times$ is transition-transversion ratio. Is $\times$ typically $>$ or <1?