Evolutionary distance: How many substitutions have fixed?

We will simulate evolution of sequences w/ a transition/transversion bias κ.

transitions: $A \longleftrightarrow G$, $C \longleftrightarrow T$

transversions: $A \longleftrightarrow T$, $A \longleftrightarrow C$, $C \longleftrightarrow A$, $C \longleftrightarrow G$

$$\Pr(T \to C) \propto \kappa \longrightarrow \frac{\kappa}{2+\kappa}$$

$$\Pr(T \to A) \propto 1 \longrightarrow \frac{1}{2+\kappa}$$

$$\Pr(T \to G) \propto 1 \longrightarrow \frac{1}{2+\kappa}$$

	Generation	
	0	CAT GCA
transversion C1A	1	AAT GCA
transition T3C	2	AAC GCA
transition C5T	3	AAC GTA
~~transition~~ transversion A6C	4	AAC GTC
transition C3T	5	AAT GTC
transition C6T	6	AAT GTT

6 substitutions

4 transitions

2 transversions

But all we observe:

CAT GCA
<u>A</u>AT G<u>TT</u>

Hamming distance = 3
transition substitutions = 1
transversion substitutions = 2

Jukes-Cantor model of substitution

4 nucleotides, each equally likely to mutate to any of the other three. The rate of substitution is μ, so we expect μt change in time t. So rate at which each character mutates to another specific one is $\mu/3$

Let $p_x(t)$ be probability a given site is x at time t.

Choose $p_x(t=0)=1$

$$\frac{dp_x}{dt} = -\mu p_x + \frac{\mu}{3}(1-p_x) = \frac{\mu}{3} - \frac{4\mu}{3}p_x$$

$$\Rightarrow dt = \frac{dp_x}{\frac{\mu}{3} - \frac{4\mu}{3}p_x} = -\frac{3}{\mu} \cdot \frac{dp_x}{4p_x - 1}$$

$$\Rightarrow t = \frac{-3}{\mu}\int \frac{dp_x}{4p_x-1} = \frac{-3}{4\mu} \cdot \ln(4p_x - 1) + C$$

What is constant of integration C?

At $t=0$, $p_x = 1$, so:

$$0 = -\frac{3}{4\mu}\cdot\ln(4p_x - 1) + C = -\frac{3}{4\mu}\cdot\ln 3 + C$$

$$\Rightarrow C = \frac{3}{4\mu}\cdot\ln 3$$

So,

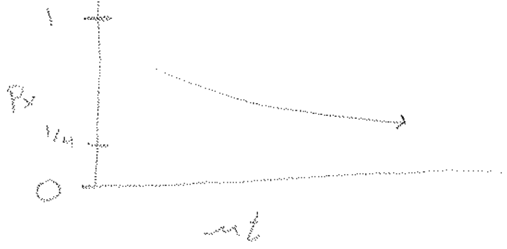
$$t = -\frac{3}{4\mu}\cdot\ln(4p_x-1) + \frac{3}{4\mu}\cdot\ln 3 = \frac{3}{4\mu}\cdot\ln\left(\frac{3}{4p_x-1}\right)$$

$$\Rightarrow \frac{4\mu t}{3} = \ln\left(\frac{3}{4p_x-1}\right) \Rightarrow e^{\frac{4\mu t}{3}} = \frac{3}{4p_x-1}$$

$$\Rightarrow 4p_x - 1 = 3e^{-\frac{4\mu t}{3}} \Rightarrow \boxed{p_x = \frac{3}{4}e^{-\frac{4\mu t}{3}} + \frac{1}{4}}$$

Some values:

p_x	μt
1.0	0
0.91	0.1
0.63	0.5
0.45	1.0
0.30	2.0
0.25	5.0



In our small simulated example, $p_x = 0.5$. Using the Jukes-Cantor model, we would estimate

$$\mu t = \frac{3}{4}\cdot\ln\frac{3}{4\cdot 0.5 - 1} = \frac{3}{4}\ln 3 \approx 0.82$$

In reality, we had 6 substitutions at 6 sites, so $\mu t = 1$.

Why is the Jukes-Cantor estimate not quite right:

1) Sampling statistics (this problem would diminish for larger sequences)

2) The model is not quite right. In our simulations, transitions were more likely than transversions. In general, would you expect Jukes-Cantor to overestimate or underestimate distances of evolution that actually occurred under a more complex model?

More general nucleotide substitution models.

Let $i, j, k, \ldots$ denote possible characters. For instance:

- * nucleotides: A, T, C, G
- * amino acids: A, C, D, E, F, ...
- * codons: AAA, AAC, AAG, ...

Let w_{ij} be the rate of substitution from $i \to j$. For Jukes-Cantor,

$$w_{ij} = \frac{\mu}{3} \quad \forall \; i \neq j$$

Let $w_{ii} = 1 - \sum_{j \neq i} w_{ij}$ (rate that i does not change). In Jukes-Cantor, $w_{ii} = 1 - \mu$

What is $p_i(t)$ given $p_i(t=0)$?

Let $\Pr(m|\mu t)$ be the probability of m mutations in time t.

$$p_i(t) = \Pr(m=0|\mu t) \cdot p_i(t=0) + \Pr(m=1|\mu t) \cdot \sum_j w_{ji} \cdot p_j(t=0)$$
$$+ \Pr(m=2|\mu t) \cdot \sum_j p_j(t=0) \sum_k w_{jk} \cdot w_{ki} \ldots$$

Define $\vec{p}(t) = [p_A(t), p_C(t), p_G(t), p_T(t)]^T$

Define $\underline{W} = [w_{ij}]$

$$\vec{p}(t) = \vec{p}(0) * \Pr(m=0|\mu t) + \Pr(m=1|\mu t)\, \underline{W} \cdot \vec{p}(0) + \Pr(m=2|\mu t) \cdot \underline{W}^2 \vec{p}(0) + \ldots$$
$$= \left(\sum_{m=0}^{\infty} \Pr(m|\mu t)\, \underline{W}^m \right) \vec{p}(0)$$

Assume mutations are Poisson, $\Pr(m|\mu t) = e^{-\mu t} \cdot \frac{(\mu t)^m}{m!}$

So:
$$\vec{p}(t) = e^{-\mu t} \cdot \left(\sum_{m=0}^{\infty} \frac{(\mu t)^m}{m!} \cdot \underline{W}^m \right) \vec{p}(0)$$
$$= e^{-\mu t} \cdot \left(\sum_{m=0}^{\infty} \frac{(\mu t \underline{W})^m}{m!} \right) \cdot \vec{p}(0)$$
$$= e^{-\mu t} \cdot e^{\mu t \underline{W}} \; \vec{p}(0) \qquad \text{Noting } e^x = \sum_{m=0}^{\infty} \frac{x^m}{m!}$$
$$= e^{\mu t (\underline{W} - \underline{I})} \cdot \vec{p}(0)$$

The matrix $Q = \underline{W} - \underline{I}$ is called the substitution matrix.

For Jukes-Cantor, $\underline{Q} = \begin{pmatrix} -\mu & \mu/3 & \mu/3 & \mu/3 \\ \mu/3 & -\mu & \mu/3 & \mu/3 \\ \mu/3 & \mu/3 & -\mu & \mu/3 \\ \mu/3 & \mu/3 & \mu/3 & -\mu \end{pmatrix}$

$\underline{W}$ is a stochastic matrix.

Assuming it is irreducible and acyclic, it has a unique stationary state ("equilibrium frequency") satisfying:

$$\vec{\pi} = \underline{W}\,\vec{\pi} \qquad \left(\text{or} \quad 0 = \underline{Q}\,\vec{\pi}\right) \text{ where } \sum_i \pi_i = 1$$

For Jukes Cantor, $\vec{\pi} = (1/4, 1/4, 1/4, 1/4)$

You can verify that

$$0 = \underline{Q} \cdot \vec{\pi}$$

$$0 = \frac{1}{4} \cdot -\mu + \frac{\mu}{3} \cdot \frac{1}{4} + \frac{\mu}{3} \cdot \frac{1}{4} + \frac{\mu}{3} \cdot \frac{1}{4} = 0$$

Commonly used matrices:

<u>nucleotides</u>: Felsenstein-84 or HKY, GTR

<u>proteins</u>: PAM, BLOSUM, WAG, JTT

<u>codons</u>: Goldman-Yang, Muse-Gaut

When matrices are estimated from protein sequences, do they actually have the property that the matrix $\underline{W}_{62}$ estimated from proteins w/ 62% identity can be written as $(\underline{W}_{90})^{\alpha} = \underline{W}_{62}$?

BLOSUM versus PAM,

Why would this property not be satisfied?

Most substitution models are chosen to satisfy reversibility

$$\pi_j W_{ji} = \pi_i \cdot W_{ij}$$

This implies that there exist $\underline{S}$ (exchangeability matrix)

$$\underline{W} = \underline{S} \cdot \mathrm{diag}(\vec{\pi}) \quad \text{where } \underline{S} \text{ is symmetric}$$

This reversibility is not a strict necessity for most phylogenetic methods, but it entails some algorithmic advantages

Complex (state-of-the-art) substitution model:

Goldman-Yang codon substitution model (61 states, non-stop codons)

$$W_{ij} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ differ by more than one nucleotide} \\ \mu \cdot \pi_j & \text{if differ by synonymous transversion} \\ \mu \cdot \kappa \cdot \pi_j & \text{if differ by synonymous transition} \\ \mu \cdot \omega \cdot \pi_j & \text{if differ by nonsynonymous transversion} \\ \mu \cdot \kappa \cdot \omega \cdot \pi_j & \text{if differ by nonsynonymous transition} \end{cases}$$

How do we get "parameters": κ, ω, $\{\pi_i\}$
(also μ if date-stamped sequences)