Evolutionary distance: how many substitutions have fixed?

We will simulate evolution with unequal rates of transitions and transversions.

Transitions: $A \leftrightarrow G$, $C \leftrightarrow T$

Transversions: $A \leftrightarrow T$, $A \leftrightarrow C$, $C \leftrightarrow A$, $C \leftrightarrow G$

Define χ as elevation in the rate of transitions

We'll use $\chi = 4$.

For example, if wildtype nucleotide is T.

$$\Pr(T \to C) \propto \chi = \frac{\chi}{2+\chi}$$

$$\Pr(T \to A) \propto 1 = \frac{1}{2+\chi}$$

$$\Pr(T \to G) \propto 1 = \frac{1}{2+\chi}$$

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

In real evolution, 6 substitutions, 4 transitions, 2 transversions

In reality, all we get to observe is this:

CAT GCA
AAT GTT

Hamming distance = 3
one transitions
two transversions

Jukes-Cantor model: 4 letters, all mutations equally likely.

Rate of substitution is μ, so we expect μt substitutions in time t.

Each character mutates to each other at rate $\mu/3$.

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

Choose x such that $p_x(t=0) = 1$

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

$$= -\frac{4}{3}\mu p_x + \frac{\mu}{3}$$

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

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

What is the constant of integration C?

Now at $t=0$, $p_x = 1$,

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

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

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

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

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

We observe p_x, want to know t:

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

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

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

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

Some values:

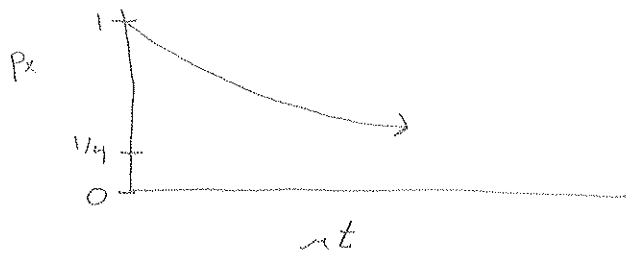
p_x	μt
1.0	0
0.91	0.1
0.63	0.5
0.45	1.0
0.30	2.0
0.25	5.0

In our example, $p_x = 0.5$.

This gives

$\mu t = \frac{3}{4} \cdot \ln \frac{3}{2-1} = \frac{3}{4} \ln 3 \approx 0.82$

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



What will be the effect on Jukes Cantor estimates if all substitution rates are not equal?

More general substitution model. We will derive differently, not using differential equations

Let $i, j, k \in \{A, T, C, G\}$ denote possible character states

For amino acids $i, j, k \in \{A, C, D, E, F, \ldots\}$

For codons, $i, j, k \in \{AAA, AAC, AAG, AAT, ACA, ACC, \ldots\}$

Let w_{ij} be the rate that character i changes to character j. In Jukes-Cantor, $w_{ij} = \frac{\mu}{3}$ for all $i \neq j$. But more generally, this might not be true.

Let $w_{ii} = 1 - \sum_{k \neq i} w_{ik}$ be the rate that i does not change.

In Jukes-Cantor, $w_{ii} = 1 - \mu$

What is $p_i(t)$ — the probability site is i at time t — given $p_i(t=0)$?

Let $Pr(m | \mu t)$ be the probability of m mutations in time t given mutation rate μ,

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

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

Define $\underline{W} = [w_{ij}]^T$ ← matrix

$$\vec{p}(t) = \vec{p}(0) \cdot Pr(m=0|\mu t) + Pr(m=1|\mu t) \cdot \vec{p}(0) \cdot \underline{W} + Pr(m=2|\mu t) \cdot \vec{p}(0) \cdot \underline{W}^2 \ldots$$

$$= \vec{p}(0) \sum_{m=0}^{\infty} Pr(m|\mu t) \cdot \underline{W}^m$$

Assume mutations are Poisson.

$$Pr(m|\mu t) = e^{-\mu t} \cdot \frac{(\mu t)^m}{m!}$$

So:

$$\vec{p}(t) = \vec{p}(0) \cdot e^{-\mu t} \cdot \sum_{m=0}^{\infty} \frac{(\mu t)^m}{m!} \underline{W}^m = \vec{p}(0) \cdot e^{-\mu t} \cdot \sum_{m=0}^{\infty} \frac{(\mu t \underline{W})^m}{m!}$$

$$= \vec{p}(0) \cdot e^{-\mu t} \cdot e^{\mu t \underline{W}} = \vec{p}(0) \cdot e^{\mu t (\underline{W} - \underline{I})}$$

The matrix $\underline{W}$ is called the transition matrix.

For Jukes-Cantor, $\underline{W} = \begin{pmatrix} 1-m & m/3 & m/3 & m/3 \\ m/3 & 1-m & m/3 & m/3 \\ m/3 & m/3 & 1-m & m/3 \\ m/3 & m/3 & m/3 & 1-m \end{pmatrix}$

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

It has a unique stationary state $\vec{\pi}$ satisfying the eigenvector equation

$$\vec{\pi} = \vec{\pi}\,\underline{W}$$

For Jukes-Cantor, $\vec{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$

$$\pi_A = (1-m)\pi_A + \frac{m}{3}\pi_C + \frac{m}{3}\pi_G + \frac{m}{3}\pi_T$$

$$\pi_C = \frac{m}{3}\pi_A + (1-m)\pi_C + \frac{m}{3}\pi_G + \frac{m}{3}\pi_T$$

etc.

This equation is satisfied by $\pi_A = \pi_C = \pi_G = \pi_T$

If we make $\vec{\pi}$ a probability vector ($\sum \pi_i = 1$),

then $\frac{1}{4} = \pi_A = \pi_C = \pi_G = \pi_T$

More general matrices:

nucleotides
- Kimura 2-parameter
- Felsenstein 84 or Hky
- GTR

protein
- PAM
- WAG
- JTT

codon
- Goldman-Yang 1994
- Muse-Gaut 1994

In general, most substitution models are chosen in such a way as to satisfy a property called "reversibility." Specifically, the stochastic matrix $\underline{W}$ is reversible if

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

where $\vec{\pi}$ is the stationary state. $\underline{W} = \underline{S} \cdot \mathrm{diag}(\vec{\pi})$ where $\underline{S}$ is symmetric.

What would evolution be non-reversible?

Systematic changes in nucleotide frequencies with time.

Cycles:

A → C
↑ ↓
T ← G

Example of a complex substitution model:

Goldman - Yang 1994:

The states are codons (usually taken as the 61 non-stop):

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

Recall that all of these models allow us to compute $P_i(t)$ ← probability of site i at time t.

Next lecture we will discuss how this enables us to compute likelihoods: $\Pr(\text{data} \mid \text{model})$

data → observed sequence

model → substitution matrix ($\underline{W}$)
evolutionary distance(s) (t or t_i)
phylogenetic tree