Review from last lecture:

All we can observe is differences between final sequence

ATG GAA ...
ATG GCA ...

We want to estimate evolutionary distance $t$. We do this in terms of $p$ (or $\vec{p}$), the probability of a difference given $t$:

Jukes-Cantor: $p = \frac{3}{4} e^{-\frac{4}{3}\mu t} + \frac{1}{4}$

General: $\vec{p} = \vec{p}_0 e^{\mu t \underline{W}}$

Equilibrium frequencies: stationary state of substitution process

Example: HKY model

$$\begin{pmatrix} \pi_A & \pi_C & \pi_G & \pi_T \end{pmatrix} \begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{cccc} A & T & C & G \\ \multicolumn{4}{c}{\left(\begin{array}{cccc} 1-\pi_T-\pi_C-\kappa\pi_G & \pi_T & \pi_C & \kappa\pi_G \\ \pi_A & 1-\pi_A-\kappa\pi_C-\pi_G & \kappa\pi_C & \pi_G \\ \pi_A & \kappa\pi_T & 1-\pi_A-\kappa\pi_T-\pi_G & \pi_G \\ \kappa\pi_A & \pi_T & \pi_C & 1-\kappa\pi_A-\pi_T-\pi_C \end{array}\right)} \end{array} = \begin{pmatrix} \pi_A(1-\pi_T-\pi_C-\kappa\pi_G)+\pi_A\pi_T+\pi_A\pi_C \\ -\pi_A\kappa\pi_G \\ \pi_C\pi_A+\pi_C(1-\pi_A-\kappa\pi_C-\pi_G) \\ +\kappa\pi_A\pi_C+\pi_A\pi_G \\ \vdots \\ \vdots \end{pmatrix}^T = \begin{pmatrix} \pi_A \\ \pi_C \\ \pi_G \\ \pi_T \end{pmatrix}^T$$

So $\vec{\pi} = (\pi_A \;\; \pi_C \;\; \pi_G \;\; \pi_T)$ is stationary state or equilibrium frequency of HKY.

Likehihood:

$P_r(\text{data} \mid \text{model})$

For instance, data: $\overset{\text{sequence 1}}{AC} \overset{\mu t}{\overline{\qquad\qquad\qquad}} \overset{\text{sequence 2}}{AA}$

In this case, the model is just the value of $\mu t$ plus the substitution model. If we used HKY model, the substitution model would have 4 free parameters: $\kappa$, $\pi_A$, $\pi_C$, $\pi_G$. So the model would be specified by $\mu t$, $\kappa$, $\pi_A$, $\pi_C$, and $\pi_G$.

Here we will use Jukes-Cantor for simplicity, so model is just specified by $\mu t$.

Likelihood:

$$\Pr(AC \mid AA, \mu t) = \Pr(A \mid A, \mu t) \cdot \Pr(C \mid A, \mu t) \leftarrow \text{what does this line assume?}$$

$$= p \cdot \left(\frac{1-p}{3}\right)$$

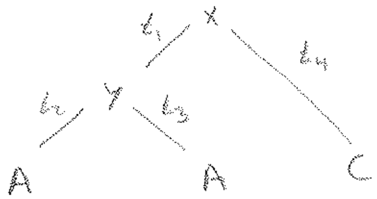
$$= \left(\frac{3}{4} e^{-\frac{4}{3}\mu t} + \frac{1}{4}\right)\left(\frac{1-\left[\frac{3}{4} e^{-\frac{4}{3}\mu t} + \frac{1}{4}\right]}{3}\right)$$

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

For not much data (just two nucleotides) the likelihood is not sharply peaked. It would be more peaked with more data.

Also note that even for the best model, the likelihood typically $\ll 1$. Why?

What about more than two sequences:



$$\Pr(\text{data} \mid \text{model}) = \Pr(A, A, C, x, y \mid t_1, t_2, t_3, t_4)$$

$$= \Pr(x) \cdot \Pr(y \mid x, t_1) \cdot \Pr(C \mid x, t_4) \cdot \Pr(A \mid y, t_2) \cdot \Pr(A \mid y, t_3)$$

If we care about tree topology, we sum over internal nodes:

$$\Pr(A, A, C \mid t_1, t_2, t_3, t_4) = \sum_x \sum_y \Pr(x) \cdot \Pr(y \mid x, t_1) \cdot \Pr(C \mid x, t_4) \cdot \Pr(A \mid y, t_2) \cdot \Pr(A \mid y, t_3)$$

(this sum will get very large for larger trees)

(Felsenstein pruning algorithm, or dynamic programming)

$$= \sum_x \Pr(x) \cdot \Pr(C \mid x, t_4) \cdot \sum_y \Pr(y \mid x, t_1) \cdot \Pr(A \mid y, t_2) \cdot \Pr(A \mid y, t_1)$$

To find the maximum likelihood tree, the branch lengths can usually be optimized using gradient. However, there is no general way to maximize over tree topologies.

Model comparison:

More complex models (additional free parameters in model framework) always have higher likelihoods.

Akaike Information Criterion (AIC): $AIC = -2 \cdot \ln L + 2 \cdot \text{parameters}$