Review from last lecture:

All we observe is differences between sequences:

ATG GAA ...
ATG GTA ...

We want to infer evolutionary distance t. We do this in terms of 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: $\vec{\pi} = \vec{\pi}\,\underline{W}$

Likelihood:

$Pr(data \mid model)$

Simple example: sequence 1 AC —— μt —— AA sequence 2

In this case, the *model* is just the value of μt plus any free parameters is substitution model. For instance, for HKY, the substitution model would have 4 parameters: κ, π_A, π_C, π_G.

Here we will just use Jukes-Cantor, so the model is just μt.

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

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

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

μt	$Pr(AC \mid AA, \mu t)$
0	0
0.1	0.028
0.5	0.077
0.824 ← maximum likelihood estimate	0.083
1.0	0.082
2.0	0.070
4.0	0.063

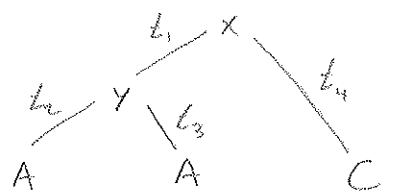
For not much data (2 nts), likelihood is not sharply peaked. Would be more peaked with more data.

Even for best model, likelihood $\ll 1$. Why?

More than 2 sequences:

Alignment of 4 nt sequences:

A
A
C



Here is one tree topology T.

Model is now

(substitution model, T, t_1, t_2, t_3, t_4)

Given this topology T:

$$Pr(data|model) = \sum_x \sum_y Pr(A, A, C, x, y | t_1, t_2, t_3, t_4)$$

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

(This sum will get very large for large trees. Felsenstein pruning algorithm, or dynamic programming)

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

For Jukes-Cantor, what is $Pr(x)$?

$$Pr(x) = \pi_x = 1/4$$

For Jukes-Cantor, what is $Pr(C|x,t_4)$?

$$Pr(C|x,t_4) = \begin{cases} p = \frac{3}{4} e^{-\frac{4}{3}\mu t} + \frac{1}{4} & \text{if } C = x \\ 1-p = \frac{3}{4} - \frac{3}{4} e^{-\frac{4}{3}\mu t} & \text{if } C \neq x \end{cases}$$

To find maximum likelihood, must optimize parameters (branch lengths, model parameters) for each topology plus search over tree topologies.

Model comparison:

Should we use a more complex model? For instance Kimura 2-parameter vs. Jukes-Cantor?

Likelihood ratio test (nested models)

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

Bayes Theorem:

$$\Pr(\text{model}|\text{data}) = \frac{\Pr(\text{data}|\text{model}) \cdot \Pr(\text{model})}{\Pr(\text{data})}$$

$$= \frac{\Pr(\text{data}|\text{model}) \cdot \Pr(\text{model})}{\sum_{\text{model}'} \Pr(\text{data}|\text{model}') \cdot \Pr(\text{model}')}$$

Advantage: Tells us what we really want, $\Pr(\text{model}|\text{data})$. No overfitting problem given appropriate priors.

Disadvantages: 1) Hard to compute (less of a problem now)

2) What is prior, $\Pr(\text{model})$?

How to compute?

Markov Chain Monte Carlo (MCMC)

1) start w/ model m_1

2) pick a new model m_2. If using Metropolis method, choose m_2 such that proposal rate $m_1 \to m_2 = m_2 \to m_1$. Typically local steps (m_2 similar to m_1)

3) Compute $R = \frac{\Pr(m_2|\text{data})}{\Pr(m_1|\text{data})} = \frac{\Pr(\text{data}|m_2) \cdot \Pr(m_2)}{\Pr(\text{data}|m_1) \cdot \Pr(m_1)}$

4) If $R \geq 1$, move to model m_2. Otherwise move to m_2 w/ probability R.

5) Repeat step (2)

If MCMC is iterated repeatedly, the fraction f_i of samples when the chain is at model m_i is $f_i = \Pr(m_i|\text{data})$.

To see this, consider models m_i and m_j with $\Pr(m_i|\text{data}) \geq \Pr(m_j|\text{data})$

At equilibrium,

$$f_i \cdot \Pr(m_i \to m_j) = f_j \cdot \Pr(m_j \to m_i)$$

Now $\Pr(m_i \to m_j) = R = \frac{\Pr(m_j|\text{data})}{\Pr(m_i|\text{data})}$

$$\Pr(m_j \to m_i) = 1$$

So: $f_i \cdot \frac{\Pr(m_j|\text{data})}{\Pr(m_i|\text{data})} = f_j \Rightarrow \frac{f_i}{f_j} = \frac{\Pr(m_i|\text{data})}{\Pr(m_j|\text{data})}$

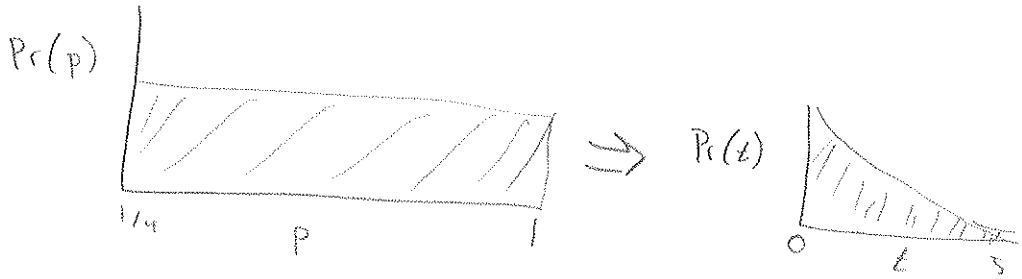
In MCMC, how do we know when we have run chain "long enough"?

About the prior!

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

Choosing a "flat" prior over parameters can be dangerous...

Flat prior on p



Flat prior on t