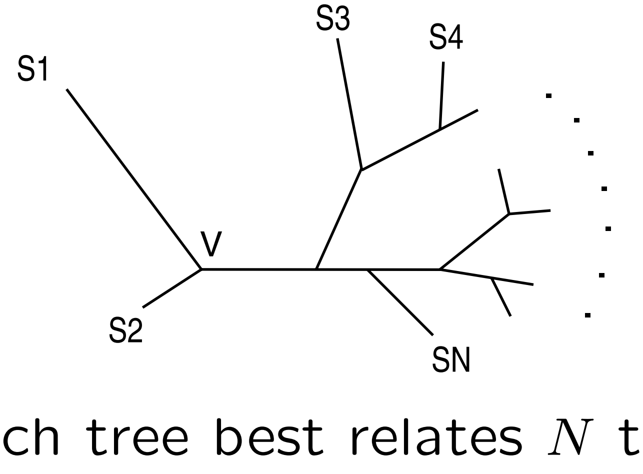
Models of Molecular Evolution: An Introduction

Elizabeth S. Allman
University of Southern Maine

MAT 500 Computational Methods in Genomics
University of Maine, Orono
November 11, 2003

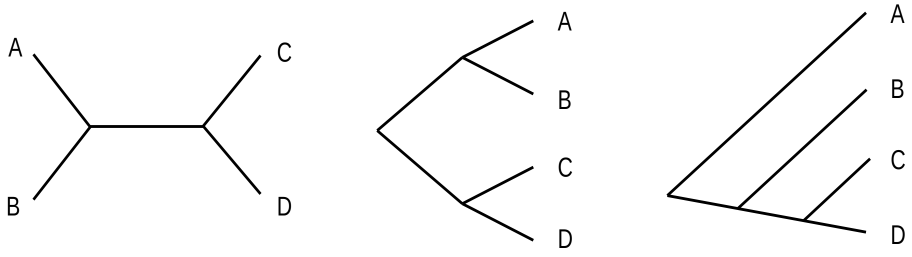


Which tree best relates N taxa?

Problem:

Given aligned biological sequences, presumed to have arisen from a common ancestral sequence, infer their evolutionary history.

```
A:  AATCGCTGCTCGACC...
B:  AAATGCTACTGGACC...
C:  AAACGTTACTGGAGC...
D:  AATCGTGGCTCGATC...
```



Do we care about root location? edge lengths? description of mutation process along edges? sequences at internal nodes?

In addition to the intrinsic interest of wanting to understand evolutionary history, there are *many* less obvious applications:

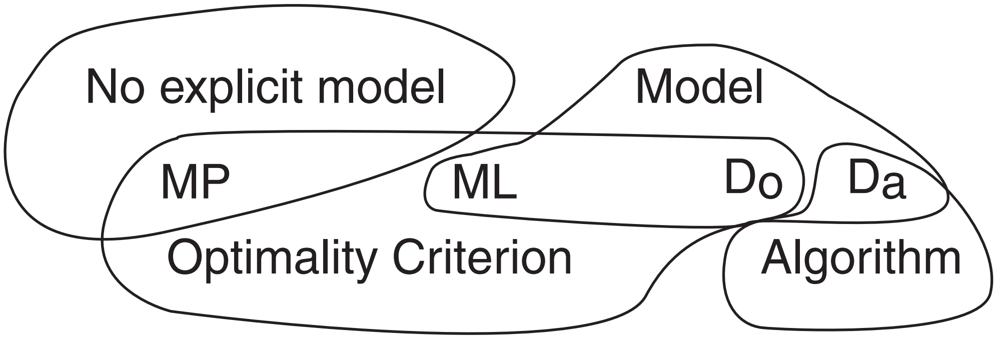
- Epidemiology – Florida dentist AIDS cluster
- Ecology – co-evolution of species and parasites; assessing diversity
- Conservation – Whales
- History – Dead Sea scrolls

Classical applications:

- inferring evolutionary relationships between primates

Major approaches in current use:

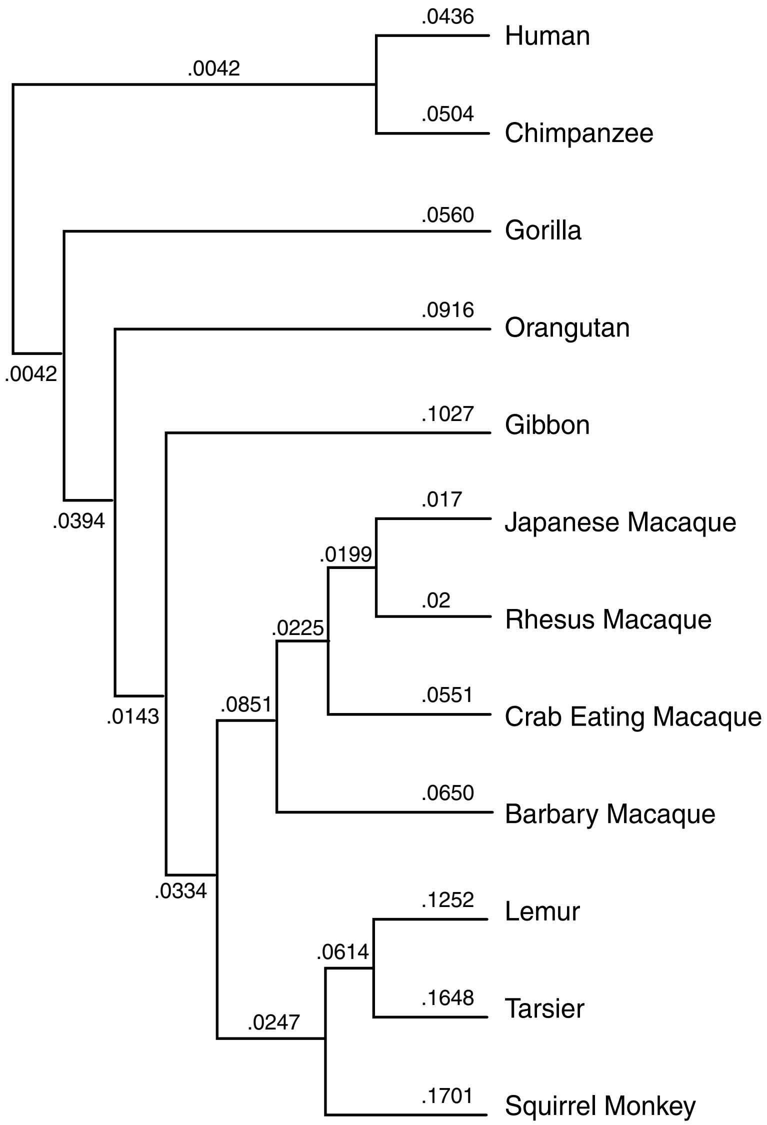
- •Maximum Parsimony (MP)
- •Distance Methods (D$_o$, D$_a$)
- •Maximum Likelihood (ML)



Distance methods begin with distance matrix, pairwise distances between species

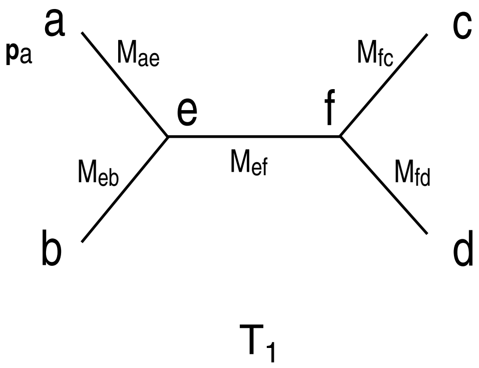
```
Gor Orangu  Human  Chimp Gibbon  CEMac Lemur  BMacaq JMacaq SqMonk  RhMac ...
  0  .1890  .1100  .1130  .2150  .3150  .3470  .2850  .2740  .3290  .2710 ...
         0  .1790  .1920  .2110  .3170  .3440  .2790  .2890  .3390  .2920 ...
                0  .0940  .2050  .2920  .3720  .3040  .2680  .3290  .2710 ...
                       0  .2140  .3240  .3720  .2920  .2850  .3480  .2980 ...
                              0  .3080  .3540  .2860  .2930  .3220  .2800 ...
                                     0  .3610  .1350  .0880  .3540  .0990 ...
                                            0  .3430  .3360  .3440  .3470 ...
                                                   0  .1370  .3570  .1300 ...
                                                                    .
                                                                    .
                                                                    .
```

Neighbor Joining leads to....


.0436
Human
.0042
.0504
Chimpanzee
.0560
Gorilla
.0916
Orangutan
.0042
.1027
Gibbon
.017
Japanese Macaque
.0394
.0199
.02
Rhesus Macaque
.0225
.0551
Crab Eating Macaque
.0143
.0851
.0650
Barbary Macaque
.0334
.1252
Lemur
.0614
.1648
Tarsier
.0247
.1701
Squirrel Monkey

How to find distances between two DNA sequences?

Simplest distance/model of molecular evolution: Jukes Cantor model



Quick review of elementary Probability:

$$\begin{aligned}\mathcal{P}_G &= \text{probability that base } G \text{ occurs at root } a.\\ \mathcal{P}_{A|G} &= \text{probability that a base } G \text{ mutates to}\\ &\quad\ \text{become an } A.\end{aligned}$$

Conditional probability

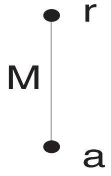
$$\mathcal{P}_{A|G} = \frac{\mathcal{P}(G,A)}{\mathcal{P}_G} = \frac{\mathcal{P}(G \text{ at } a \text{ and } A \text{ at } e)}{\mathcal{P}_G}$$

Explicit Models:

Model base substitutions at a single site
Assume

•i.i.d. – each site is an independent trial of the same probabilistic process

•Markov – probabilities of each substitution along an edge depend only on immediate ancestor base



Model parameters: the tree T

root distribution $\mathbf{p}_r = \begin{pmatrix} \mathcal{P}_A \\ \mathcal{P}_G \\ \mathcal{P}_C \\ \mathcal{P}_T \end{pmatrix}$

Markov matrix $M =$

$$\begin{pmatrix} \mathcal{P}(A|A) & \mathcal{P}(A|G) & \mathcal{P}(A|C) & \mathcal{P}(A|T) \\ \mathcal{P}(G|A) & \mathcal{P}(G|G) & \mathcal{P}(G|C) & \mathcal{P}(G|T) \\ \mathcal{P}(C|A) & \mathcal{P}(C|G) & \mathcal{P}(C|C) & \mathcal{P}(C|T) \\ \mathcal{P}(T|A) & \mathcal{P}(T|G) & \mathcal{P}(T|C) & \mathcal{P}(T|T) \end{pmatrix}$$

Jukes-Cantor Model
Additional assumptions:

• All bases occur with equal probability in the root distribution $\mathbf{p}_r = (.25, .25, .25, .25)$

• All possible base substitutions are equally likely, $A \leftrightarrow G$, $A \leftrightarrow C$, $A \leftrightarrow T$, $G \leftrightarrow T$, etc.
Markov matrix $M_{JC} =$

$$\begin{pmatrix} 1-\alpha & \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & 1-\alpha & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & 1-\alpha & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} & 1-\alpha \end{pmatrix}$$

Parameter α is a probability, but also may be interpreted as a rate.

Rate at which observable base substitutions occur over one time step and is measured in units (substitutions per site)/(time step)

Powers M_{JC}^t of the Jukes Cantor Markov matrix give the conditional probabilities after t time steps. Markov matrix $M_{JC}^t =$

$$\begin{pmatrix} \frac{1}{4}+\frac{3}{4}(1-\frac{4}{3}\alpha)^t & \frac{1}{4}-\frac{1}{4}(1-\frac{4}{3}\alpha)^t & \frac{1}{4}-\frac{1}{4}(1-\frac{4}{3}\alpha)^t & \dots \\ \frac{1}{4}-\frac{1}{4}(1-\frac{4}{3}\alpha)^t & \frac{1}{4}+\frac{3}{4}(1-\frac{4}{3}\alpha)^t & \vdots & \\ \frac{1}{4}-\frac{1}{4}(1-\frac{4}{3}\alpha)^t & \frac{1}{4}-\frac{1}{4}(1-\frac{4}{3}\alpha)^t & \vdots & \\ \frac{1}{4}-\frac{1}{4}(1-\frac{4}{3}\alpha)^t & \frac{1}{4}-\frac{1}{4}(1-\frac{4}{3}\alpha)^t & & \end{pmatrix}$$

(Think: the Jukes-Cantor matrix modeling the evolutionary process (probability of base substitutions) at a single site between two sequences over t time steps.)

(The product of two JC matrices is again JC.)

Let $p(t)$ = the fraction of sites that differ between two sequences S_0 and S_1.

Then $p(t)$ can be estimated from data (proportion of sites that differ). Using modeling principles,

$$p(t) = \frac{3}{4} - \frac{3}{4}(1-\frac{4}{3}\alpha)^t$$

Jukes-Cantor distance:

$$p = \frac{3}{4} - \frac{3}{4}(1 - \frac{4}{3}\alpha)^t$$

Solving for t, gives

$$t = \frac{\ln(1 - \frac{4}{3}p)}{\ln(1 - \frac{4}{3}\alpha)}$$

Unrealistic to find either t or α. However,

$$\begin{aligned} \alpha t &= \text{(mutation rate)(no. of time steps)} \\ &= \text{(no. of subst per site/time step)} \\ &\qquad\qquad \text{(no. of time steps)} \\ &= \text{expected no. substitutions per site} \\ &\qquad\qquad \text{during the elapsed time} \end{aligned}$$

Approximating $\ln(1 - \frac{4}{3}\alpha) \approx -\frac{4}{3}\alpha$, then

$$\alpha t \approx -\frac{3}{4}\ln(1 - \frac{4}{3}p) \equiv d_{JC},$$

where p is the fraction of the sites that disagree in S_0 and S_1.

Eg. Consider the two aligned sequences

S_0 : ACTTGTCGGATGATCAGCGGTCCATGCACCTGACAACGGT

S_1 : ACATGTTGCTTGACGACAGGTCCATGCGCCTGAGAACGGC

Compute the JC distance between them.

$$d_{JC} \equiv -\frac{3}{4} \ln \left(1 - \frac{4}{3} p\right)$$