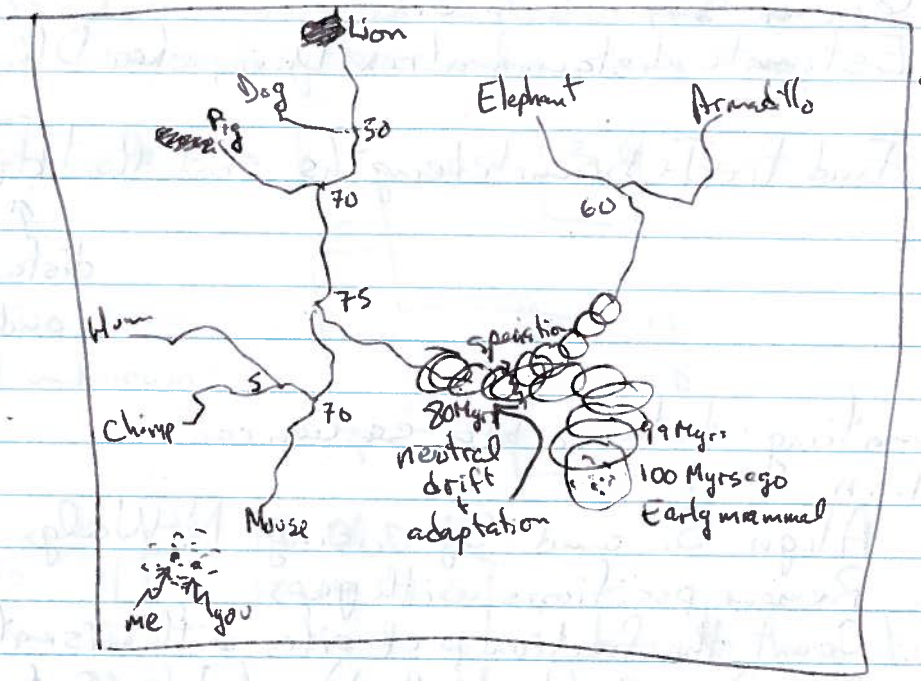
Sequence Evolution + Phylogenetics



$\{A, C, G, T\}^*$ = Space of all DNA sequences

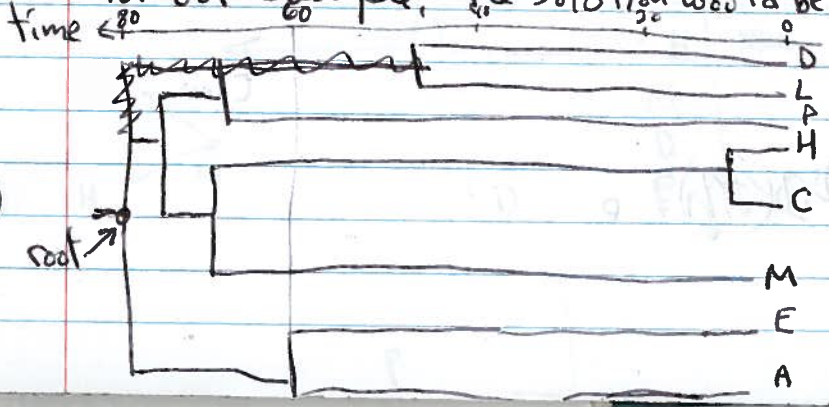
Phylogenetic Inference Problem (Biological version)

Given: DNA sequences from multiple species

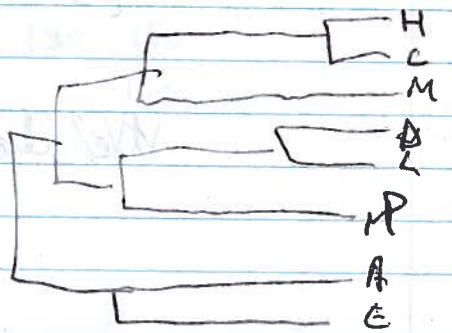
Find: Evolutionary tree that describes their evolution

→ Topology, Dating of speciation events

For our example, the solution would be:



Note: this is the same topology as



Distance-based Phylo. Inference methods

Idea: Given seq. $S_1 \ldots S_n$

(A) Estimate distance matrix $D_{n \times n}$, where $D(i,j)$ = distance btw S_i, S_j

(B) Find tree + Branch lengths such that $d_T(i,j) \approx D(i,j)$

($d_T(i,j)$: distance btw nodes i and j in tree T)

(A) Estimating distance btw sequences

For $i,j = 1 \ldots n$

1. Align S_i and S_j using N-W alg.
2. Remove positions with gaps
3. Count the fraction p of sites with mismatches
4. $D(i,j) = -\frac{3}{4}\left(\log\left(1-\frac{4}{3}p\right)\right)$ (Jukes-Cantor model)

Example

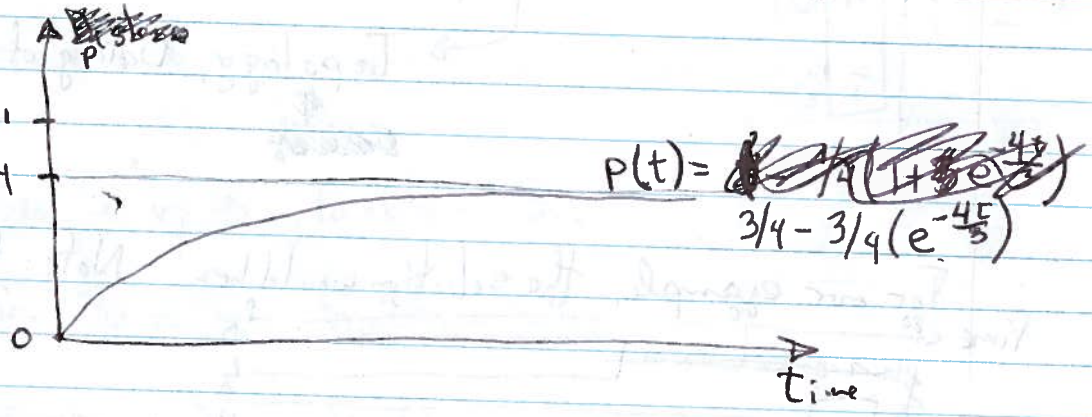
S_i: A C - - T G A C T G

S_j: A G T T T A - C A G

$\Rightarrow p = 3/7$

$D(i,j) = 0.63$

Note: If the two seqs. are very long



$p(t) = 3/4 - 3/4(e^{-\frac{4t}{3}})$

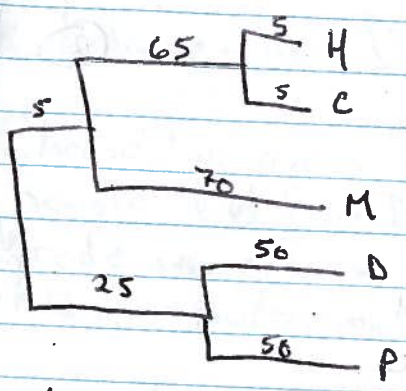
We choose $D(i,j) = p$

Under Kimura 2-parameter model with rates of transition v. transversion

Notes: More advanced distance measures consider separately transitions and transversions

Example:

True tree (unknown)



$D =$

	H	C	M	D	P
H	-	10	140	150	150
C		-	140	150	150
M			-	150	150
D				-	100
P					

How to find tree?

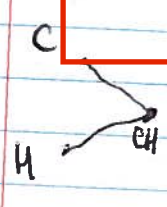
Incorrect
See appendix for corrected version and example

UPGMA

Repeat

1. Choose two closest nodes according to $D_{u,v}$
2. Merge two nodes into a new node w
3. Remove row/col u, v from matrix
4. Insert new row/col for w : $D(w,i) = D(u,i) - \frac{1}{2} D(u,v)$

	CH	M	D	P
CH	~~135~~	135	145	145
M		-	150	150
D				100
P				



M

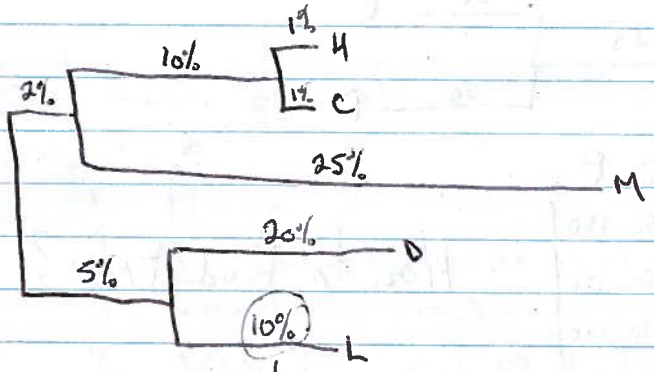
D

P

Running time: ?

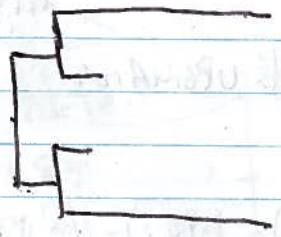
Problem: Only works if ① Distances are estimated perfectly accurately
② Mutation rate is constant

More realistic tree:



Mutation rate varies
⇒ Distances btw seq cannot be calculated in Mys but instead in Expected substitution per site

This means that 10% positions would have gotten mutated. the expected # of sub/s. per site is 0.1



⇒ UPGMA will produce the wrong tree