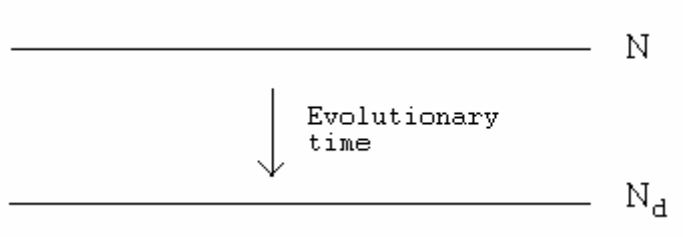
Lecture Overview:

1) Estimating substitution distances – Jukes Cantor Model
2) Constructing correct phylogenetic trees – Felsenstein

Estimating substitution distances:



N = number of nucleotides
N_d = number of observed substitutions (nucleotides that changed)

But what we are really interested in is the number of actual substitutions that occurred (including non-observable substitutions like A -> A, or if comparing two genomes with common ancestor, A -> C that occurred in both genomes being compared)

Jukes-Cantor Model

Assumption: every letter can mutate (change) into another letter with equal probability (not really true for biology because nucleotides have small bias to mutate to their complements (A <-> T, C <-> G), and some organisms favour certain nucleotides (e.g. thermostable bacteria have GC rich genomes because these pairings are more stable than AT).
Nonetheless, this is the classical model and holds up fairly well in practice.

	A	C	G	T
A	*1-3α*	*α*	*α*	*α*
C	*α*	*1-3α*	*α*	*α*
G	*α*	*α*	*1-3α*	*α*
T	*α*	*α*	*α*	*1-3α*

α = probability of nucleotide change

Let's model probability that any observed nucleotide (C in this example) is the same after some time *t:*

$$P_{C_1} = 1 - 3\alpha$$
$$P_{C_2} = (1-3\alpha)P_{C_1} + \alpha(1 - P_{C_1})$$
.
.
.
.
$$P_{C_T} = (1-3\alpha)P_{C_{T-1}} + \alpha(1 - P_{C_{T-1}})$$

without solving directly (messy), use continuous approximation and solve differential equation to derive:

$$P_{C_T} = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t}$$

The probability that a nucleotide changes is 1 – P(nt stays the same):

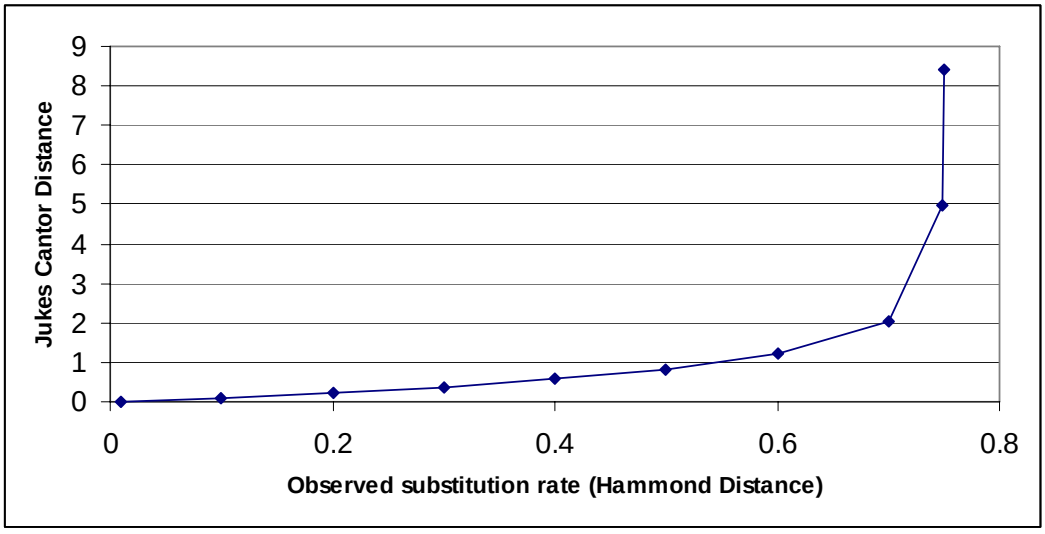
$$\frac{N_d}{N} = 1 - P_{C_t}$$

$$\frac{N_d}{N} = \frac{3}{4}(1 - e^{-4\alpha t})$$

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

αt = Jukes-Cantor distance (alpha is the substitution rate which you cannot get directly because you do not know the time)

This makes sense qualitatively: if observed substitutions are 75% (basically, meaning you have two random sequences), alpha-t approaches infinity, if it small, alpha-t goes to zero:



e.g. mouse-human alpha-t ≈ 0.5

Other models that incorporate nucleotide and substitution biases also exist.

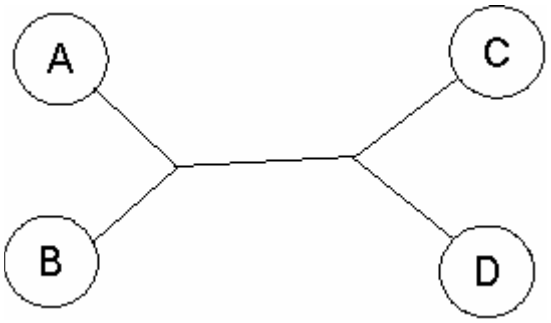
-Hasegawa-Kishino-Yano (HKY) model (does not assume equal base frequencies)

-Kimura (1980), takes into account transitions, and transversions (a transition refers to a change between two bases that have similar structure; A and G are similarly structured (collectively called purines), and C and T are similar (collectively called pyrimidines, a transversion refers to a change between a purine and pyrimidine, e.g. C -> A)

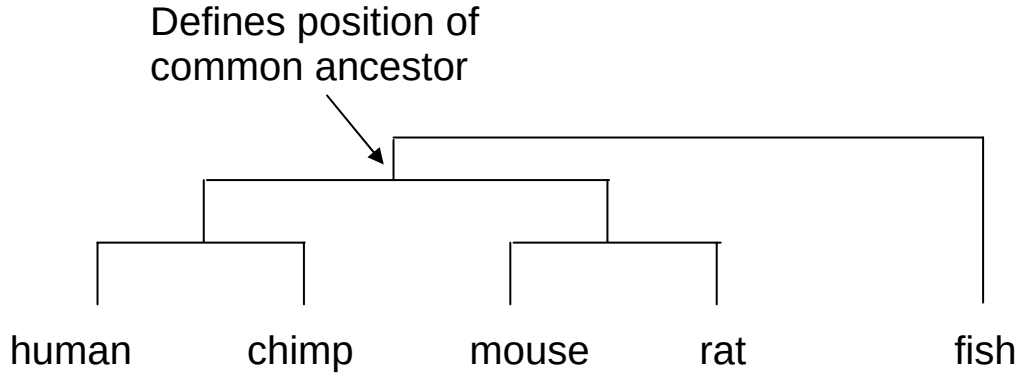
Phylogenetic Trees:

Depiction of substitution distances that accurately depict common ancestors (i.e. when two different genomes were identical).

e.g.



Demonstrates that A, B, C, and D shared a common ancestor, but we don't know when (unrooted tree). To find out where along the horizontal line common ancestor exist, we need an outgroup (an organism that diverged prior to existence of common ancestor). In other words, the outgroup defines the distance to common ancestry between the two highest nodes on a hierarchical tree (the two ancestors adjoining the root of the tree). For example if A, B, C, and D were human, chimp, mouse, and rat as shown below.



fish defines the distance from the mouse-rat common ancestor and human-chimp common ancestor to the common ancestor of both lineages.

Building Trees

There are two types of approaches to building trees: (i) Distance based methods, which work from pairwise distances between the sequences (e.g. UPGMA, Neighbour Joining), (ii) character-based methods, which work directly from multiply aligned sequences (e.g. parsimony and likelihood approaches; these were not covered in class).

Distance Methods

The proofs for both distance methods described below are based on the assumption that the we are dealing with an additive tree (i.e. for any two leaves the distance between them is the sum of edges along a path that connects them).

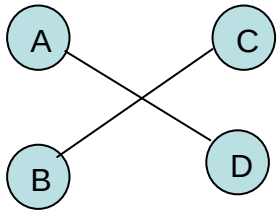
UPGMA

A straightforward method to draw a tree is a greedy clustering approach called the Unweighted Pair Group Method using Arithmetic Averages (UPGMA). A tree is drawn by recursively joining two closest nodes, grouping them into one cluster, and treating them as single node to which new distances are calculated. This is done until only two nodes remain.

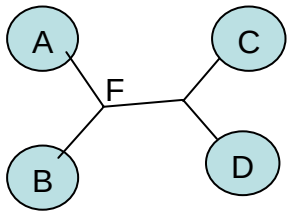
Suppose you are given four sequences separated by the following distances:

	A	B	C	D
A				
B	**4**			
C	**5**	**8**		
D	**7**	**9**	**7**	

To construct the tree we begin with a star tree, where the initial branch point is arbitrary.



A and B are the closest so let's group them under a new node called F and calculate the distances to the new node:



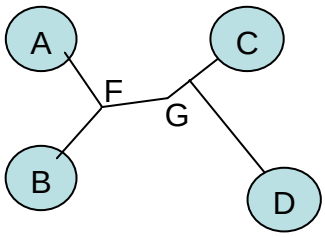
The distance from C to F can be calculated as:

$$d(C,F) = \frac{d(C,B) + d(C,A) - d(A,B)}{2}$$

You divide by 2 because you covered the path from C to F twice. The distance matrix then becomes:

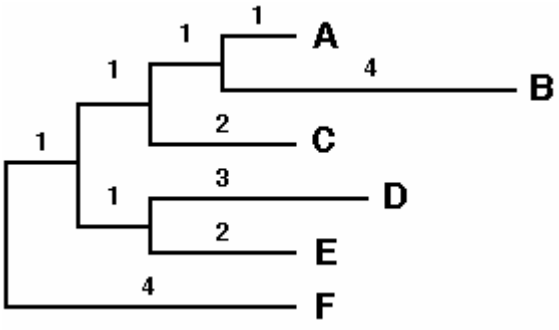
	F	C	D
F			
C	**5**		
D	**6**	**7**	

F and C are now closest, repeating the procedure by grouping C and F into G produces the following tree:



N.B. This is still an unrooted tree as we do not know when C and D diverged.

UPGMA produces the correct tree when the mutation rates along all branches are equal, however, this is not always the case. Consider the following phylogenetic tree (known to be correct):



(http://www.icp.ucl.ac.be/~opperd/private/neighbor.html)

In this case B diverged more quickly than A. Thus, UPGMA would join A and C producing an incorrect tree topology. Neighbour Joining (NJ) was developed to take non-uniform divergence rates into account. The idea is to produce a modified distance matrix that takes into account the net divergence rate of the entire tree, so that nodes close together but far apart from all other nodes are not necessarily joined directly.

Saitou and Nei (1987) derived a model for calculating the modified matrix. If d is the original distance matrix, the modified matrix

$$D_{ij} = d_{ij} - (R_i + R_j)$$

where,

$$R_i = \frac{1}{|N| - 2} \sum_{m \in N} d_{im}.$$

N is the set of leaves in the tree. The number of leaves is $|N|$. The tree topology is constructed as described above (UPGMA) using the modified distance matrix D. Atteson (1999) proved this model to be correct.

Suppose you have the following distance matrix d (corresponding to the above tree):

	A	B	C	D	E
B	5				
C	4	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8

,

R_i, the net divergence matrix for each sequence is calculated as the sum of each node to all other nodes.

$R(A) = (5+4+7+6+8)/4 = 30/4$,
$R(B) = 42/4$,
$R(C) = 32/4$,
$R(D) = 38/4$,
$R(E) = 34/5$,
$R(F) = 44/4$.

D is:

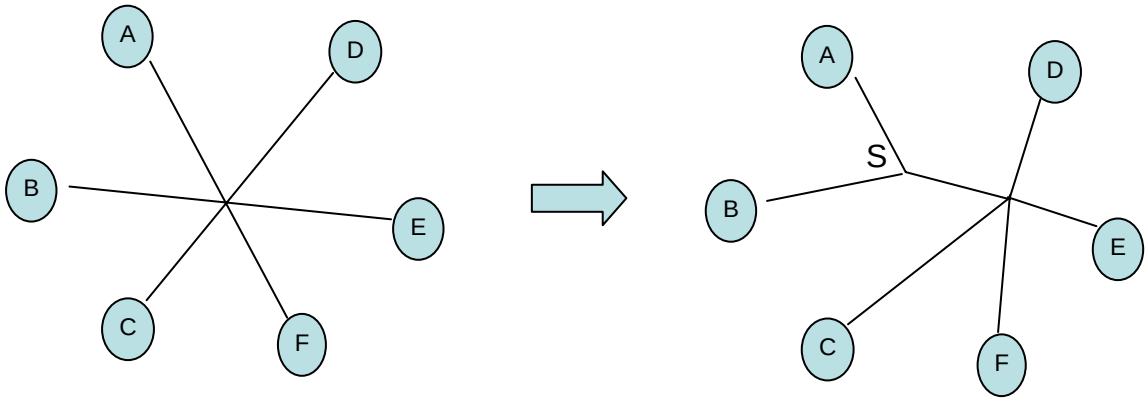
	A	B	C	D	E
B	-13				
C	-11.5	-11.5			
D	-10	-10	-10.5		
E	-10	-10	-10.5	-13	
F	-10.5	-10.5	-11	-11.5	-11.5

It is now apparent that D and E, and A and B are closest and these are initially joined as neighbours.

Sample calculation for D_{ij}:

$$D_{ij} = d_{ij} - (R_i + R_j)$$
$$D_{AB} = d_{AB} - (R_A + R_B)$$
$$D_{AB} = 5 - (\frac{30}{4} + \frac{42}{4})$$
$$D_{AB} = -13$$

We can now use the greedy algorithm to construct the tree. We would start by joining A and B (AB->S), since this is one of the minima of D (we could have also joined D and E, the other minimum), and recompute the distances to S (nodes joining AB; see figure below).



Subbing in S for A and B, we can calculate the distances from the remaining nodes to S,

```
d(CS) = d(AC) + d(BC) - d(AB) / 2 = 3
d(DS) = d(AD) + d(BD) - d(AB) / 2 = 6
d(ES) = d(AE) + d(BE) - d(AB) / 2 = 5
d(FS) = d(AF) + d(BF) - d(AB) / 2 = 7
```

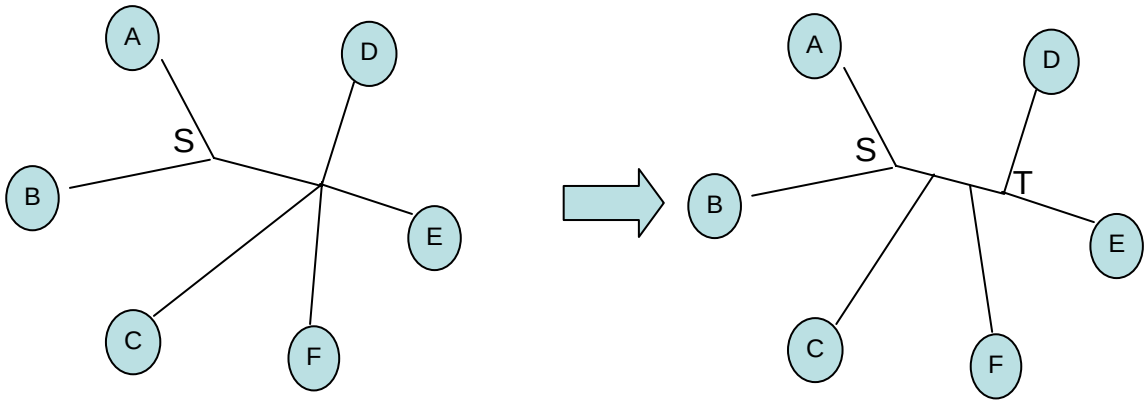
d becomes:

	S	C	D	F
S				
C	3			
D	6	7		
E	5	6	5	
F	7	8	9	8

Where *D* is:

	S	C	D	E	F
S					
C	-19.5				
D	-18	-18.5			
E	-17.5	-18	-20.5		
F	-19.5	-20	-19	-20	

We would next join D and E -> T.



We would continue to join nodes, recalculate *d* and *D* until only two nodes remain.