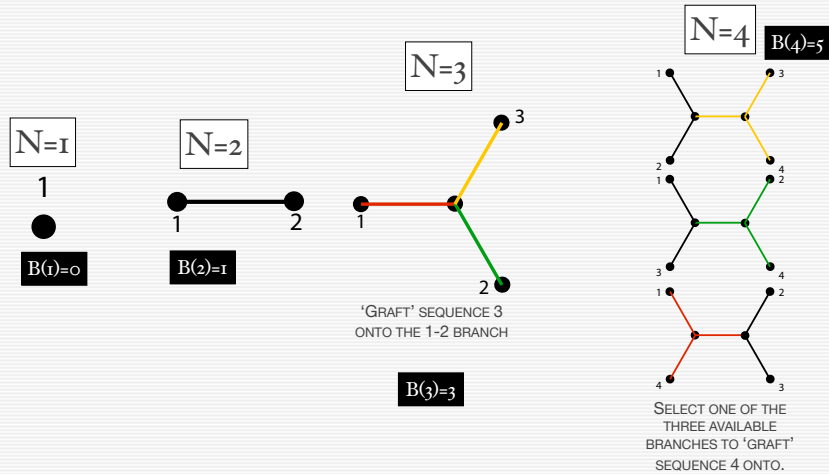
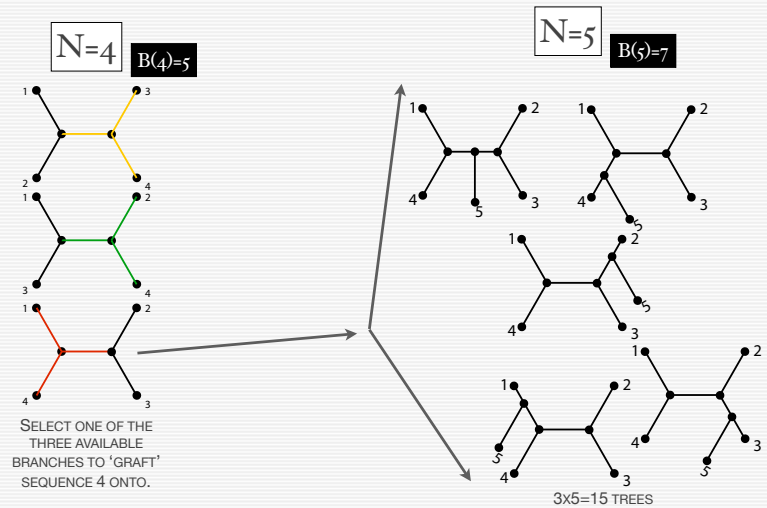


COUNTING BRANCHES & TREES

- How many branches $B(N)$ does an **unrooted tree** on N leaves have and how many different **unrooted labeled trees**, $T_u(N)$, with N leaves are there?



COUNTING BRANCHES & TREES



THERE ARE COMBINATORIALLY MANY TREES

N	$T_u(N)$
3	1
4	3
5	15
6	105
7	945
8	10395
9	135135
10	2027025

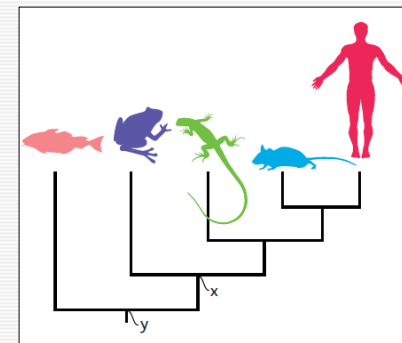
N	$T_u(N)$
20	2.22E+20
50	2.84E+74
100	1.70E+182

BRANCHES $B(N) = 2N - 3, N \geq 1$

TREES $(2N - 5)!!, N \geq 3$

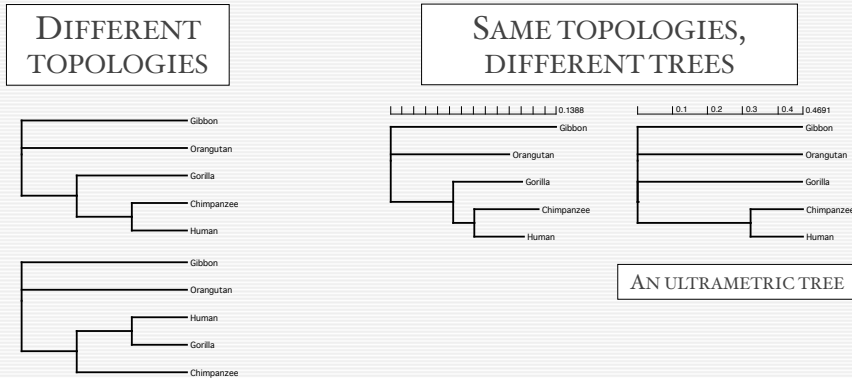
HIERARCHICAL CLUSTERING

- We have a collection of aligned nucleotide sequences from different species, and wish to construct their evolutionary hierarchy/history – a phylogeny.



TOPOLOGY VS TREE

- Topology defines the structure of the tree (unweighted edges)
- Topology combined with branch lengths constitutes a phylogenetic tree



HIERARCHICAL CLUSTERING

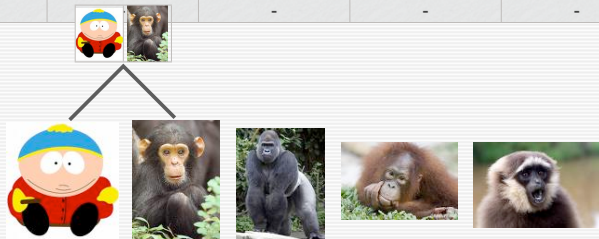
- distance matrix on 5 nucleotide sequences.
- here we consider number of mismatches
- *p*-distances (number of mismatches normalized by the length of the sequence)
- model-based distances (later)

	Human	Chimpanzee	Gorilla	Orangutan	Gibbon
Human	-	8	10	15	18
Chimpanzee	-	-	12	17	19
Gorilla	-	-	-	16	19
Orangutan	-	-	-	-	20
Gibbon	-	-	-	-	-

CLUSTERING PROCEDURE

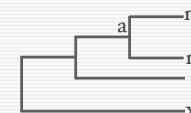
- At each step, we select the two closest sequences and **join** them to form a clade.
- We then replace the two just joined sequences with their ancestor
- This reduces the size of the data matrix by one
- We need to compute the distances from the new ancestor to the remaining sequences

	Human	Chimpanzee	Gorilla	Orangutan	Gibbon
Human	-	8	10	15	18
Chimpanzee	-	-	12	17	19
Gorilla	-	-	-	16	19
Orangutan	-	-	-	-	20
Gibbon	-	-	-	-	-



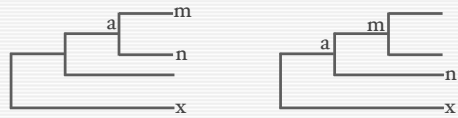
UPDATING DISTANCES

- **Single linkage:** $d(x,a) = \min(d(x,m), d(x,n))$
- **Complete linkage:** $d(x,a) = \max(d(x,m), d(x,n))$

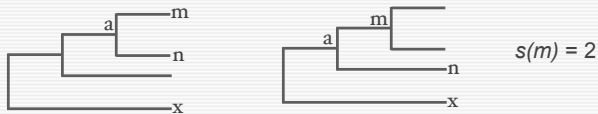


UPDATING DISTANCES

- **UPGMA** (Unweighted Pair Group Method with Arithmetic Mean): $d(x,a) = 1/2 [d(x,m) + d(x,n)]$



- **WPGMA** (Weighted Pair Group Method with Arithmetic Mean): $d(x,a) = [s(m) d(x,m) + s(n) d(x,n)] / [s(m) + s(n)]$, where $s(n)$ measures the number of actual sequences represented by node n .



EXAMPLE CONTINUED

- Use UPGMA. Joining human and chimp...

	Human	Chimpanzee	Gorilla	Orangutan	Gibbon
Human	-	8	10	15	18
Chimpanzee	-	-	12	17	19
Gorilla	-	-	-	16	19
Orangutan	-	-	-	-	20
Gibbon	-	-	-	-	-

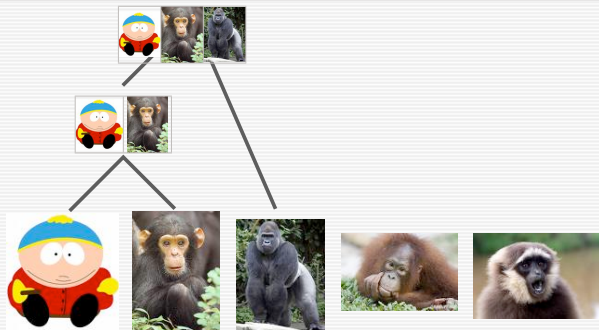


	Human-Chimpanzee	Gorilla	Orangutan	Gibbon
Human-Chimpanzee	-	(10+12)/2=11	(15+17)/2=16	(18+19)/2=18.5
Gorilla	-	-	16	19
Orangutan	-	-	-	20
Gibbon	-	-	-	-

	Human-Chimpanzee	Gorilla	Orangutan	Gibbon
Human-Chimpanzee	-	11	16	18.5
Gorilla	-	-	16	19
Orangutan	-	-	-	20
Gibbon	-	-	-	-



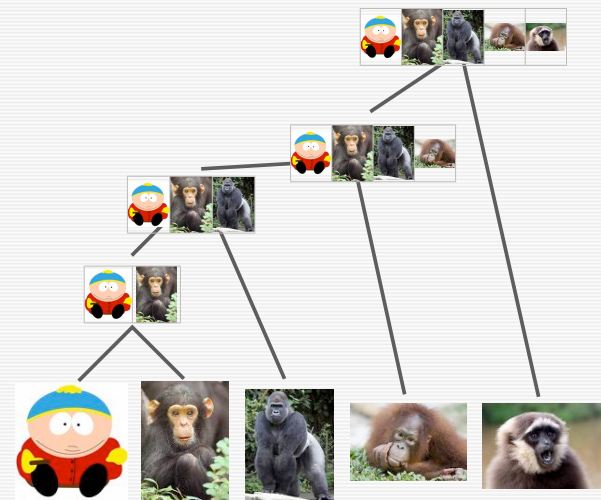
	Human-Chimpanzee-Gorilla	Orangutan	Gibbon
Human-Chimpanzee-Gorilla	-	(16+16)/2 = 16	(18.5+19)/2 = 18.75
Orangutan	-	-	20
Gibbon	-	-	-



	Orangutan	Gibbon
Human-Chimpanzee-Gorilla	16	18.75
Orangutan	-	20
Gibbon	-	-

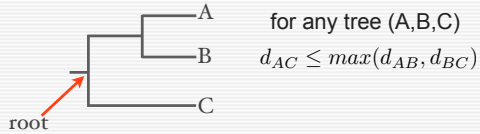


	Gibbon
Hum-Chimp-Gor-Orang	(18.75+20)/2=19.375
Gibbon	-



CLUSTERING APPROACHES

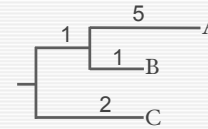
- trees are additive (reconstruct distances by adding edges)
- trees are ultrametric (i.e. all leaf nodes are equidistant from the root)



- assumes a strict molecular clock



FITTING A DISTANCE MATRIX TO A TREE



- this tree is not ultrametric
 - but it can be additive

	A	B	C
A			
B	6		
C	8	4	

can we define an algorithm that allows us to reconstruct a correct additive tree given a distance matrix?