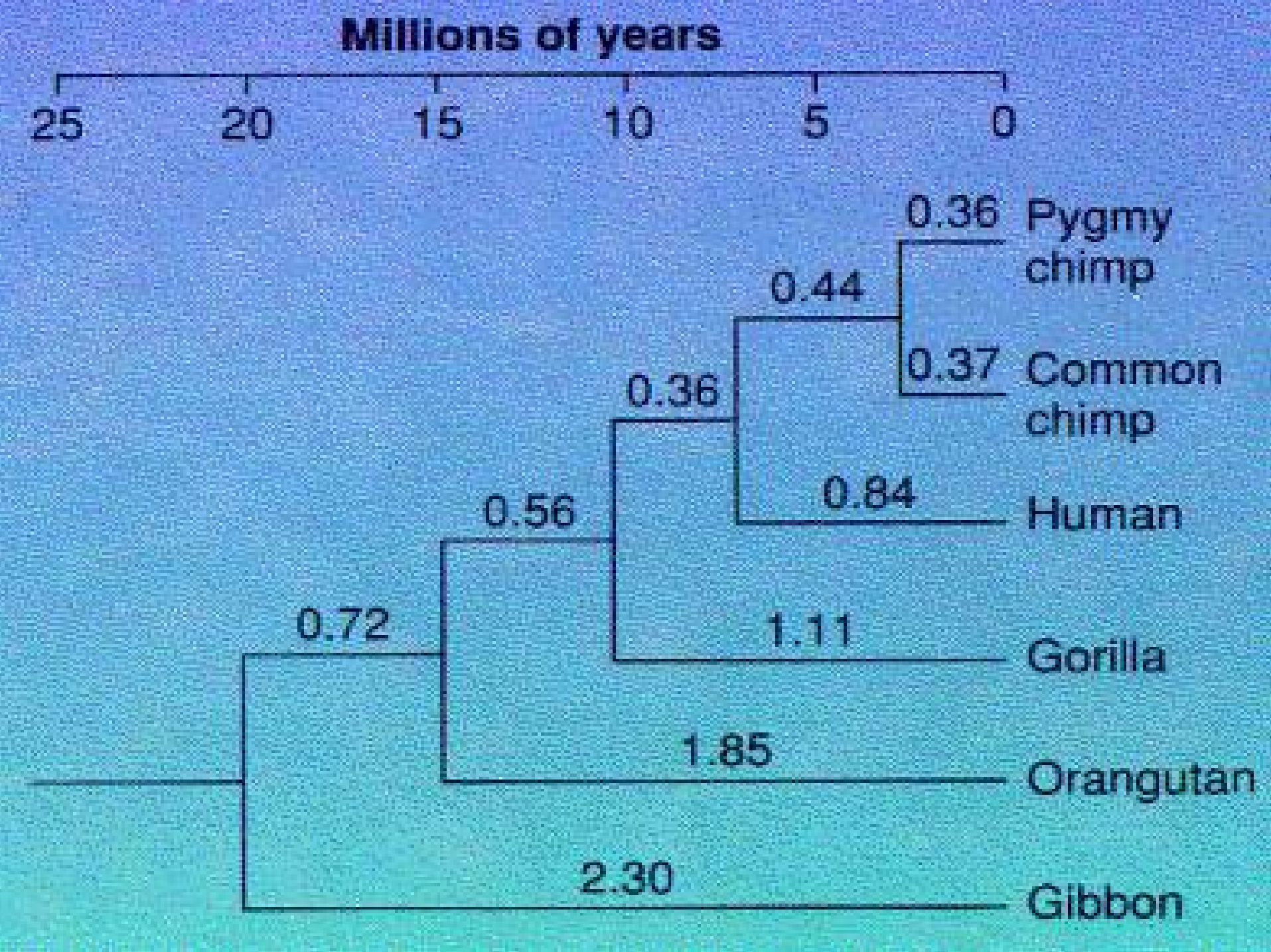
Theory of Evolution

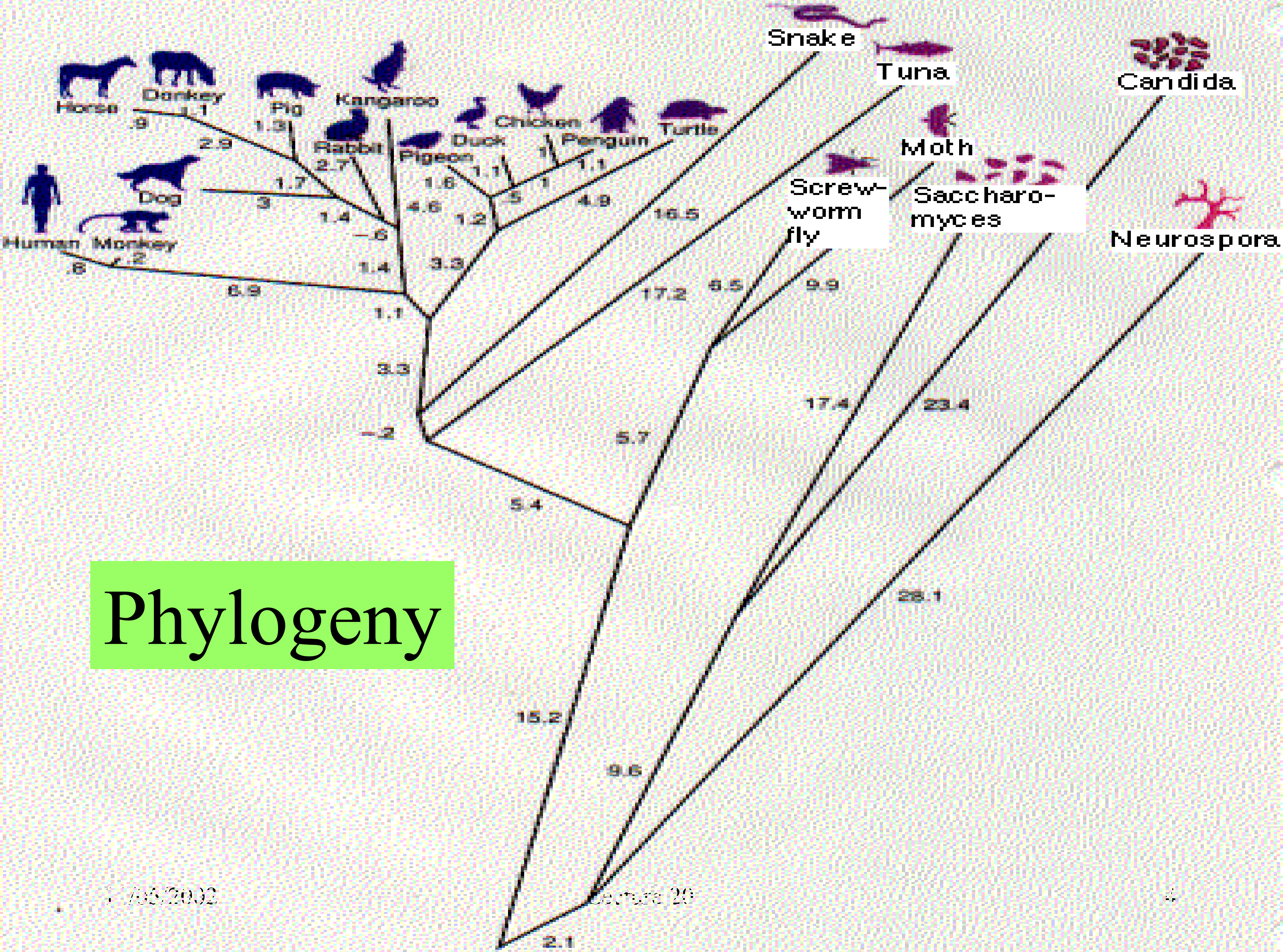
- Charles Darwin
  - **1858-59:** *Origin of Species*
  - 5 year voyage of H.M.S. Beagle (1831-36)
  - Populations have variations.
  - Natural Selection & Survival of the fittest: *nature selects best adapted varieties to survive and to reproduce.*
  - Speciation arises by splitting of one population into subpopulations.
  - Gregor Mendel and his work (1856-63) on inheritance.


Millions of years
25
20
15
10
5
0
0.36
Pygmy chimp
0.44
0.37
Common chimp
0.36
0.84
Human
0.56
0.72
1.11
Gorilla
1.85
Orangutan
2.30
Gibbon

Dominant View of Evolution

- All existing organisms are derived from a common ancestor and that new species arise by splitting of a population into subpopulations that do not cross-breed.
- Organization: **Directed Rooted Tree**; Existing species: Leaves; Common ancestor species (divergence event): Internal node; Length of an edge: Time.

Phylogeny

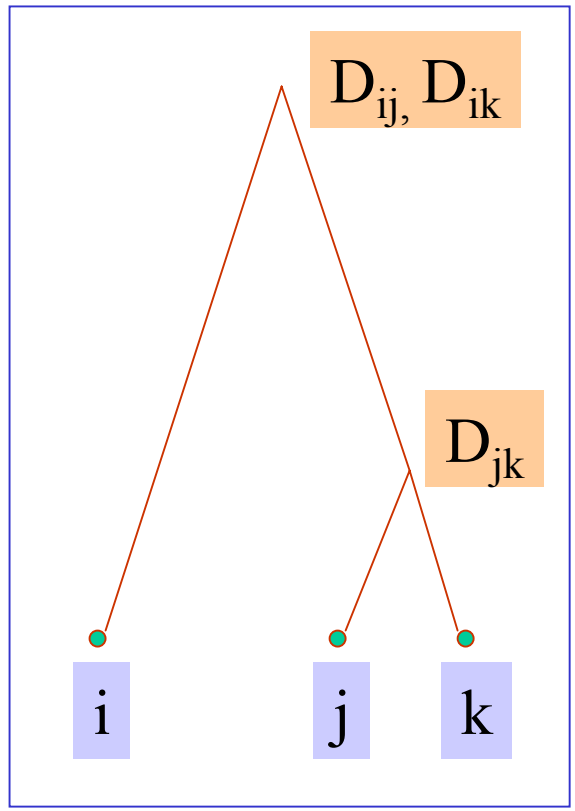

Snake
Tuna
Candida
Moth
Screw-worm fly
Saccharo-myces
Neurospora
Horse
Donkey
Pig
Kangaroo
Rabbit
Pigeon
Duck
Chicken
Penguin
Turtle
Dog
Human
Monkey
16.5
17.2
6.5
9.9
17.4
23.4
28.1
5.7
5.4
15.2
9.6
2.1
6.9
3.3
1.1
4.9
4.6
2.9
1.7
3

Constructing Evolutionary/Phylogenetic Trees

- 2 broad categories:
  - Distance-based methods
    - Ultrametric
    - Additive:
      - UPGMA
      - Transformed Distance
      - Neighbor-Joining
  - Character-based
    - Maximum Parsimony
    - Maximum Likelihood
    - Bayesian Methods

Ultrametric

- An ultrametric tree:
  - decreasing internal node labels
  - distance between two nodes is label of least common ancestor.
- An ultrametric distance matrix:
  - Symmetric matrix such that for every i, j, k, there is tie for maximum of D(i,j), D(j,k), D(i,k)



Ultrametric: Assumptions

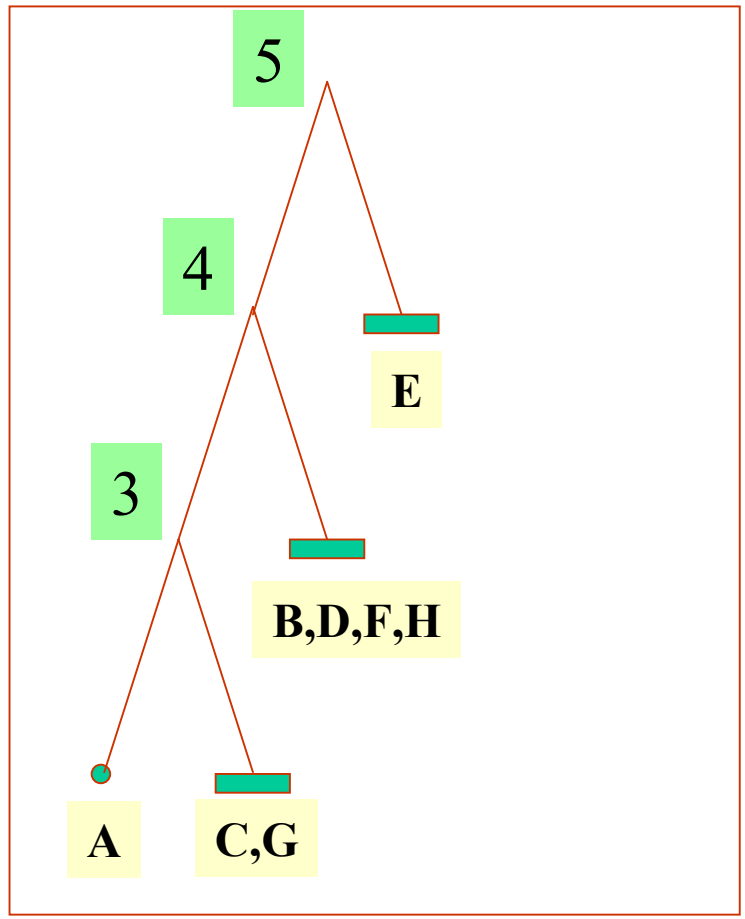
- **Molecular Clock Hypothesis**, Zuckerkandl & Pauling, 1962: **Accepted** point mutations in amino acid sequence of a protein occurs at a **constant** rate.
  - Varies from protein to protein
  - Varies from one part of a protein to another

Ultrametric Data Sources

- Lab-based methods: hybridization
  - Take denatured DNA of the 2 taxa and let them hybridize. Then measure energy to separate.
- Sequence-based methods: distance

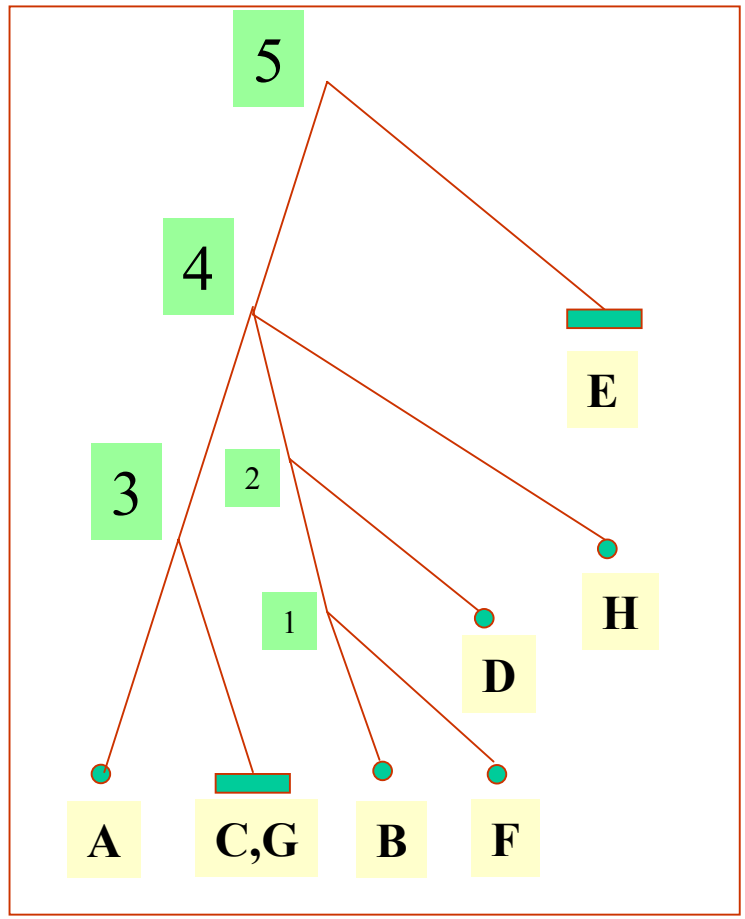
Ultrametric: Example

	A	B	C	D	E	F	G	H
A	0	4	3	4	5	4	3	4
B								
C								
D								
E								
F								
G								
H								



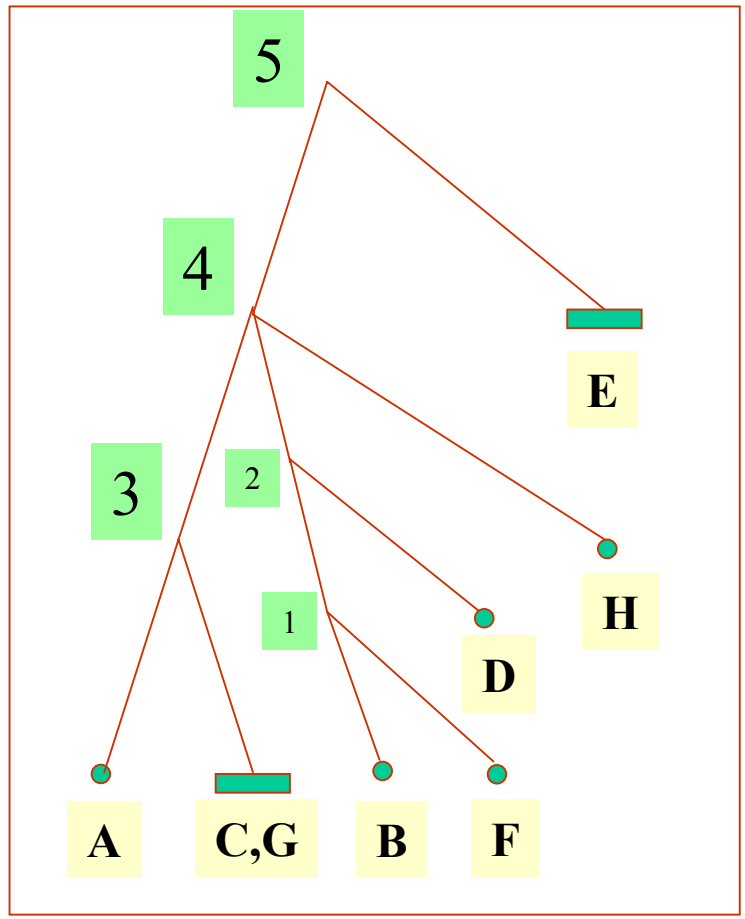
Ultrametric: Example

	A	B	C	D	E	F	G	H
A	0	4	3	4	5	4	3	4
B		0	4	2	5	1	4	4
C								
D								
E								
F								
G								
H								



Ultrametric: Distances Computed

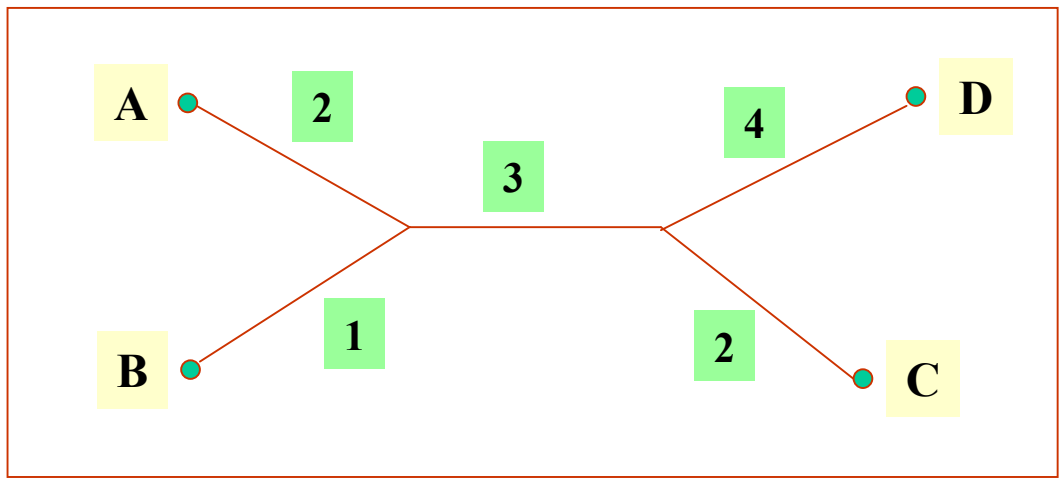
	A	B	C	D	E	F	G	H
A	0	4	3	4	5	4	3	4
B		0	4	2	5	1	4	4
C							2	
D								
E								
F								
G								
H								



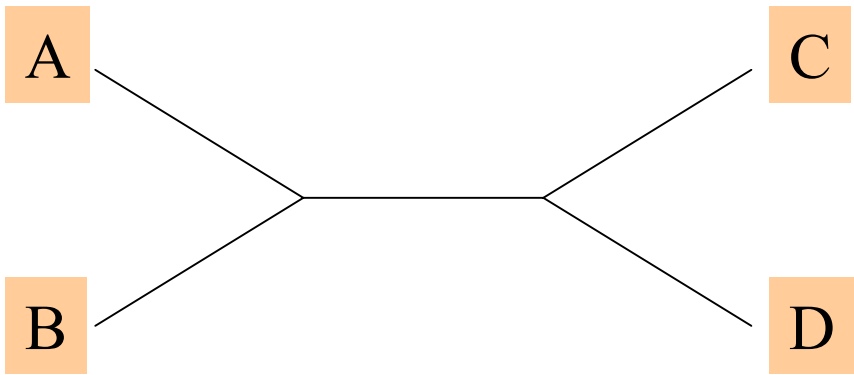
Additive-Distance Trees

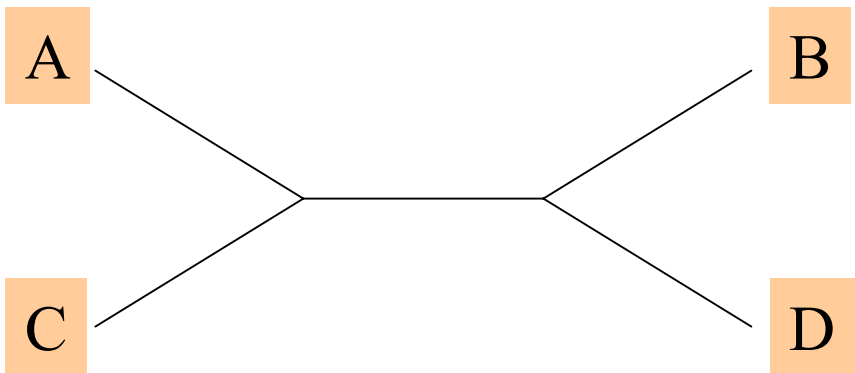
Additive distance trees are edge-weighted trees, with distance between leaf nodes are exactly equal to length of path between nodes.

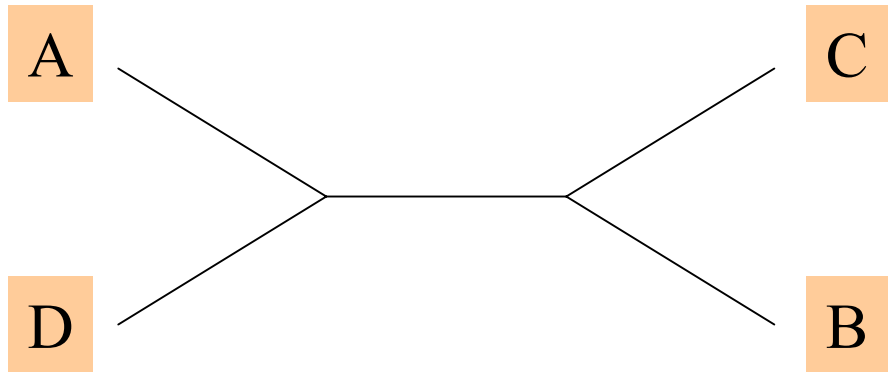
	A	B	C	D
A	0	3	7	9
B		0	6	8
C			0	6
D				0



Unrooted Trees on 4 Taxa

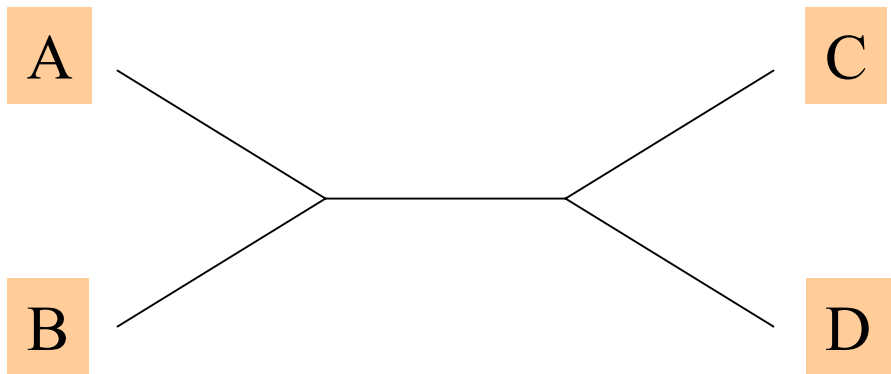






Four-Point Condition

- If the true tree is as shown below, then
  1. $d_{AB} + d_{CD} < d_{AC} + d_{BD}$, and
  2. $d_{AB} + d_{CD} < d_{AD} + d_{BC}$

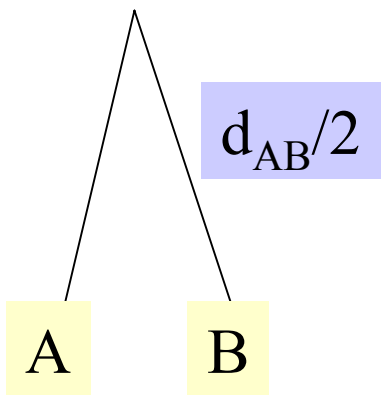


Unweighted pair-group method with arithmetic means (UPGMA)

	A	B	C
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{CD}

	AB	C
C	$d_{(AB)C}$	
D	$d_{(AB)D}$	d_{CD}

$$d_{(AB)C} = (d_{AC} + d_{BC}) /2$$



Transformed Distance Method

- UPGMA makes errors when rate constancy among lineages does not hold.
- Remedy: introduce an outgroup & make corrections

$$D_{ij}' = \frac{D_{ij} - D_{iO} - D_{jO}}{2} + \left(\sum_{k=1}^{n} D_{kO} \Big/ n \right)$$

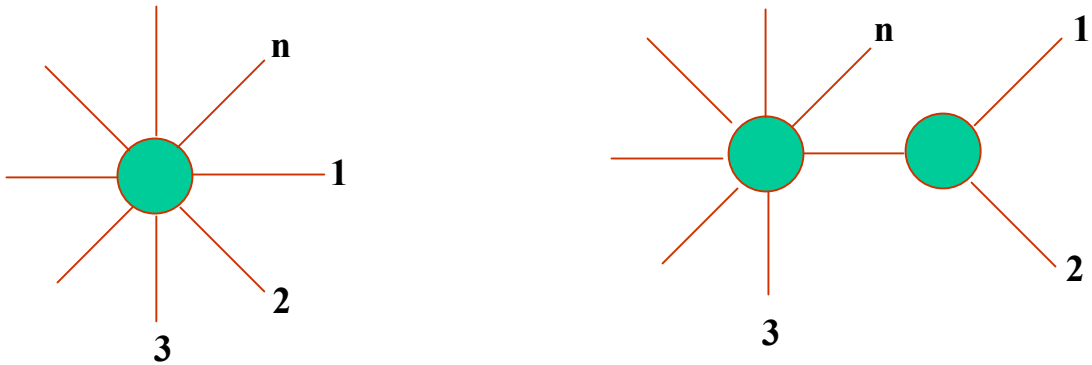
- Now apply UPGMA

Saitou & Nei: Neighbor-Joining Method

- Start with a star topology.
- Find the pair to separate such that the total length of the tree is minimized. The pair is then replaced by its arithmetic mean, and the process is repeated.

$$S_{12} = \frac{D_{12}}{2} + \frac{1}{2(n-2)} \sum_{k=3}^{n} (D_{1k} + D_{2k}) + \frac{1}{(n-2)} \sum_{3 \le i \le j \le n} D_{ij}$$

Neighbor-Joining



$$S_{12} = \frac{D_{12}}{2} + \frac{1}{2(n-2)} \sum_{k=3}^{n} (D_{1k} + D_{2k}) + \frac{1}{(n-2)} \sum_{3 \le i \le j \le n} D_{ij}$$

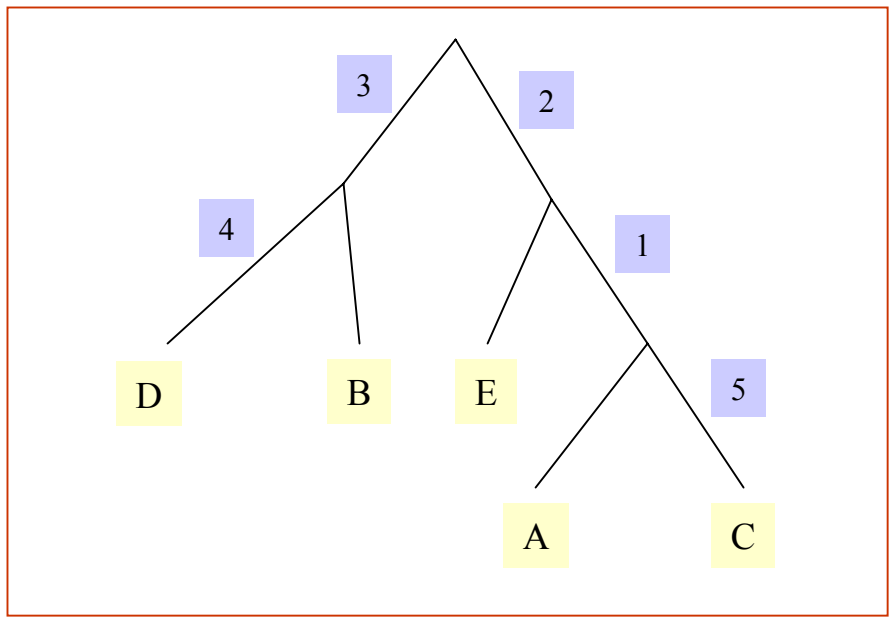
Constructing Evolutionary/Phylogenetic Trees

- 2 broad categories:
  - Distance-based methods
    - Ultrametric
    - Additive:
      - UPGMA
      - Transformed Distance
      - Neighbor-Joining
  - Character-based
    - Maximum Parsimony
    - Maximum Likelihood
    - Bayesian Methods

Character-based Methods

- Input: characters, morphological features, sequences, etc.
- Output: phylogenetic tree that provides the history of what features changed. [Perfect Phylogeny Problem]
- one leaf/object, 1 edge per character, path ⇔changed traits

	1	2	3	4	5
A	1	1	0	0	0
B	0	0	1	0	0
C	1	1	0	0	1
D	0	0	1	1	0
E	0	1	0	0	0



Example

- Perfect phylogeny does not always exist.

	1	2	3	4	5
A	1	1	0	0	0
B	0	0	1	0	1
C	1	1	0	0	1
D	0	0	1	1	0
E	0	1	0	0	1

Maximum Parsimony

- Minimize the total number of mutations implied by the evolutionary history