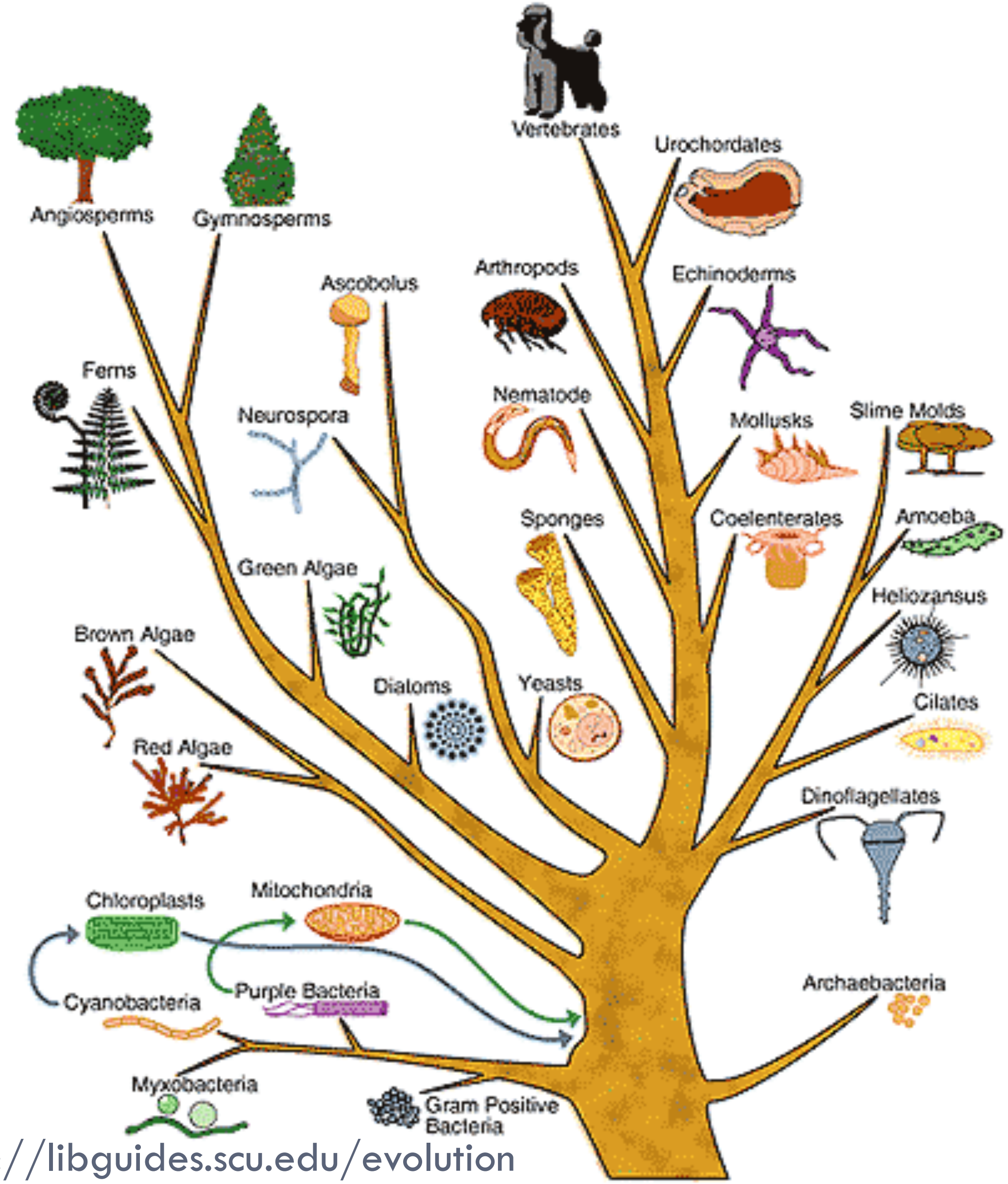
Phylogenetic tree construction



http://libguides.scu.edu/evolution

Outline

- Phylogenetic tree types
- Distance Matrix method
  - UPGMA
  - Neighbor joining
- Character State method
  - Maximum likelihood

Phylogenetic tree?

- A tree represents graphical relation between organisms, species, or genomic sequence
- In Bioinformatics, it’s based on genomic sequence

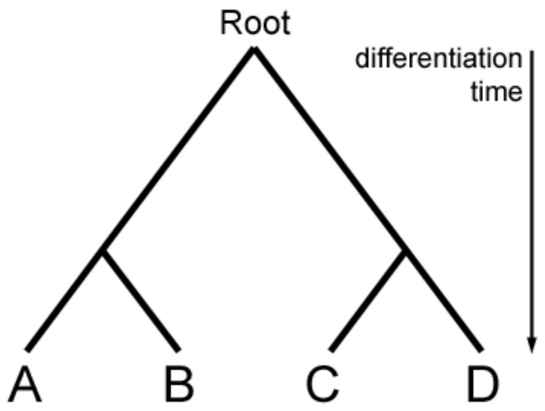
What do they represent?

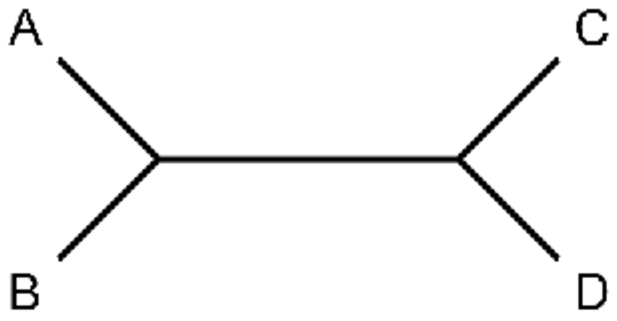
- Root: origin of evolution
- Leaves: current organisms, species, or genomic sequence
- Branches: relationship between organisms, species, or genomic sequence
- Branch length: evolutionary time

(in cladogram, it doesn't represent time)

Rooted / Unrooted trees

- Rooted tree: directed to a unique node
  - (2 * number of leaves) - 1 nodes,
  - (2 * number of leaves) - 2 branches
- Unrooted tree: shows the relatedness of the leaves without assuming ancestry at all
  - (2 * number of leaves) - 2 nodes
  - (2 * number of leaves) - 3 branches

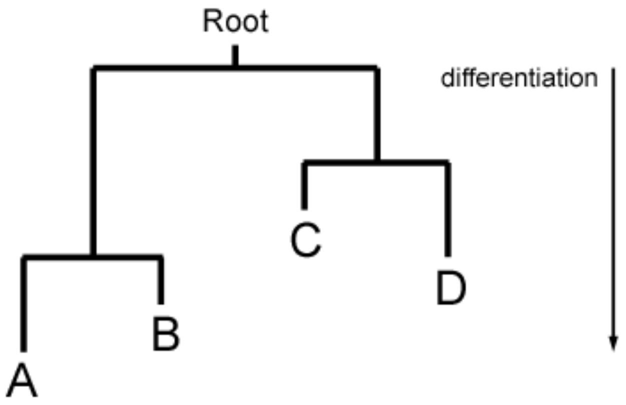


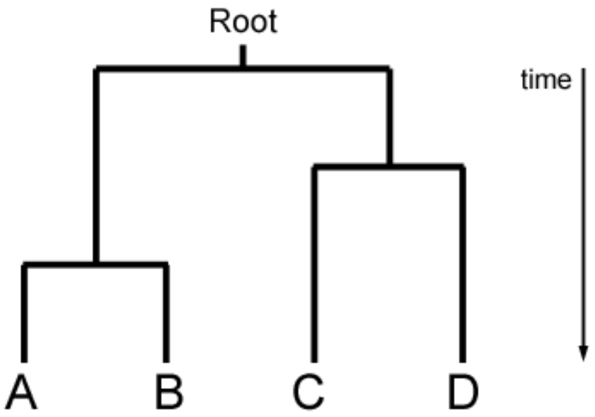


https://www.nescent.org/wg_EvoViz/Tree

More tree types used in bioinformatics (from cohen article)

- Unrooted tree
- Rooted tree
  - Cladograms: Branch length have no meaning
  - Phylograms: Branch length represent evolutionary change
  - Ultrametric: Branch length represent time, and the length from the root to the leaves are the same





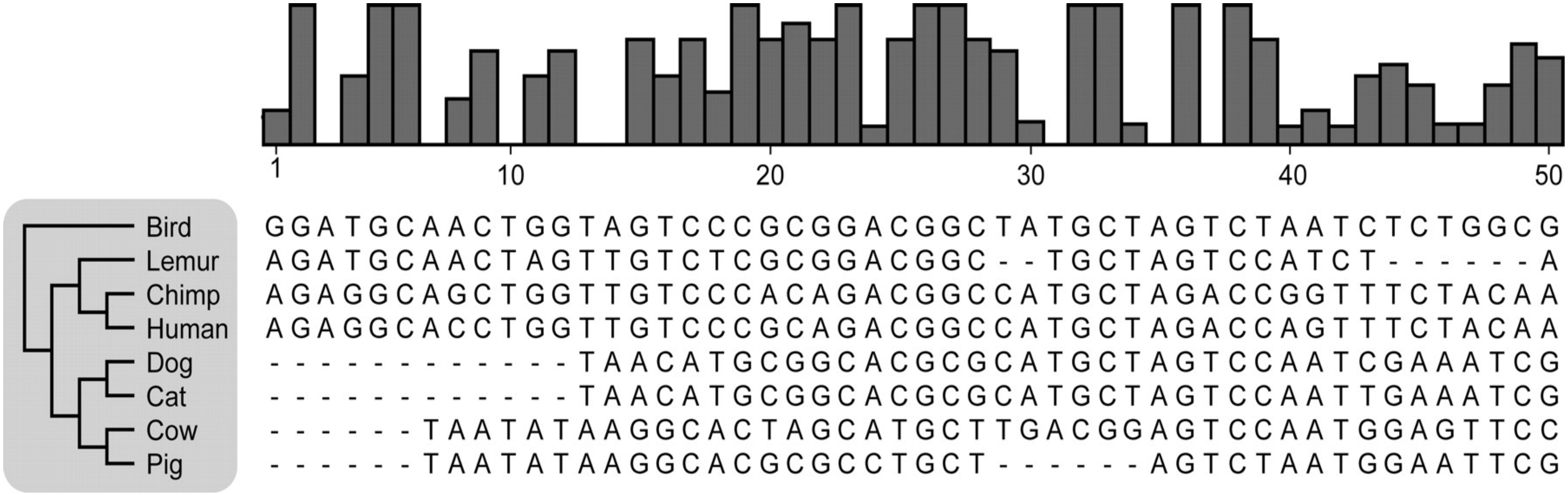
https://www.nescent.org/wg_EvoViz/Tree

How to construct a phylogenetic tree?

- Step1:

  Make a multiple alignment from base alignment or amino acid sequence (by using MUSCLE, BLAST, or other method)

How to construct a phylogenetic tree?



```
Bird   GGATGCAACTGGTAGTCCCGCGGACGGCTATGCTAGTCTAATCTCTGGCG
Lemur  AGATGCAACTAGTTGTCTCGCGGACGGC--TGCTAGTCCATCT------A
Chimp  AGAGGCAGCTGGTTGTCCCACAGACGGCCATGCTAGACCGGTTTCTACAA
Human  AGAGGCACCTGGTTGTCCCGCAGACGGCCATGCTAGACCAGTTTCTACAA
Dog    ------------TAACATGCGGCACGCGCATGCTAGTCCAATCGAAATCG
Cat    ------------TAACATGCGGCACGCGCATGCTAGTCCAATTGAAATCG
Cow    ------TAATATAAGGCACTAGCATGCTTGACGGAGTCCAATGGAGTTCC
Pig    ------TAATATAAGGCACGCGCCTGCT------AGTCTAATGGAATTCG
```

- **Step 2:**

  Check the multiple alignment if it reflects the evolutionary process.

http://genome.cshlp.org/content/17/2/127.full

How to construct a phylogenetic tree? cont

- Step3:

  Choose what method we are going to use and calculate the distance or use the result depending on the method

- Step 4:

  Verify the result statistically.

Distance Matrix methods

- Calculate all the distance between leaves (taxa)
- Based on the distance, construct a tree
- Good for continuous characters
- Not very accurate
- Fastest method
  - UPGMA
  - Neighbor-joining

UPGMA

- Abbreviation of “Unweighted Pair Group Method with Arithmetic Mean”
- Originally developed for numeric taxonomy in 1958 by Sokal and Michener
- Simplest algorithm for tree construction, so it's fast!

How to construct a tree with UPGMA?

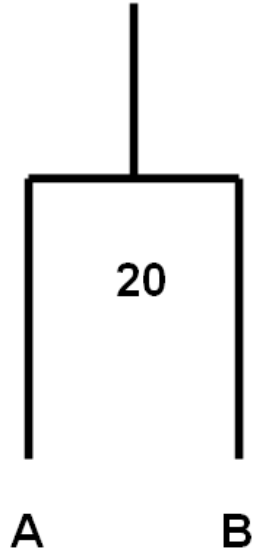
- Prepare a distance matrix
- Repeat step 1 and step 2 until there are only two clusters
- Step 1:

  Cluster a pair of leaves (taxa) by shortest distance
- Step 2:

  Recalculate a new average distance with the new cluster and other taxa, and make a new distance matrix

Example of UPGMA

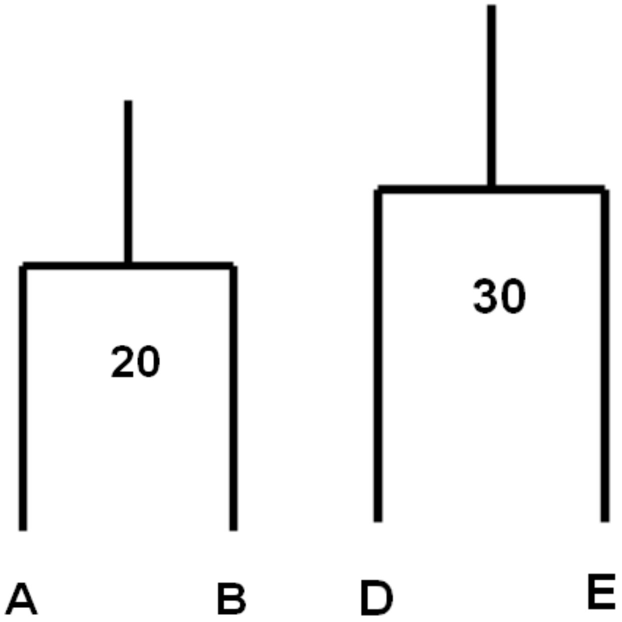
	A	B	C	D	E
A	0				
B	20	0			
C	60	50	0		
D	100	90	40	0	
E	90	80	50	30	0



- New average distance between AB and C is:
  - C to AB = (60 + 50) / 2 = 55
- Distance between D to AB is:
  - D to AB = (100 + 90) / 2 = 95
- Distance between E to AB is:
  - E to AB = (90 + 80) / 2 = 85

Example of UPGMA cont 1

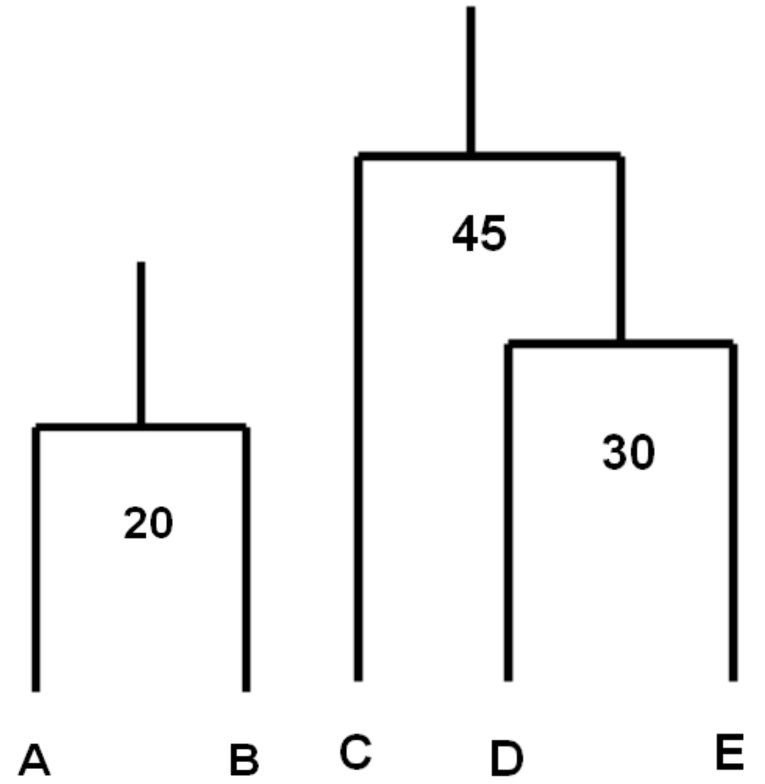
	AB	C	D	E
AB	0			
C	55	0		
D	95	40	0	
E	85	50	30	0



- ❑New average distance between AB and DE is:
  - ❑AB to DE = (95 + 85) / 2 = 90

Example of UPGMA cont 2

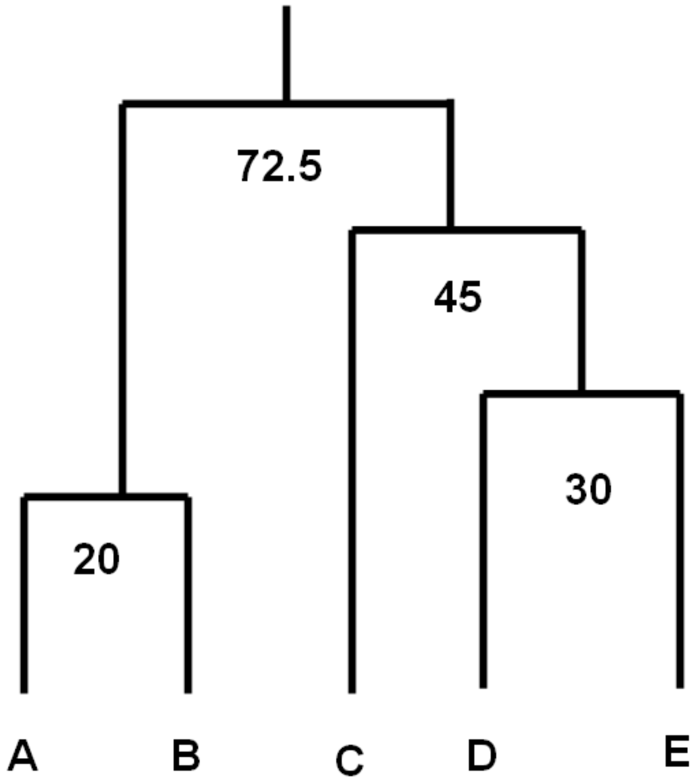
	AB	C	DE
AB	0		
C	55	0	
DE	90	45	0



- New Average distance between CDE and AB is:
  - CDE to AB = (90 + 55) / 2 = 72.5

Example of UPGMA cont 3

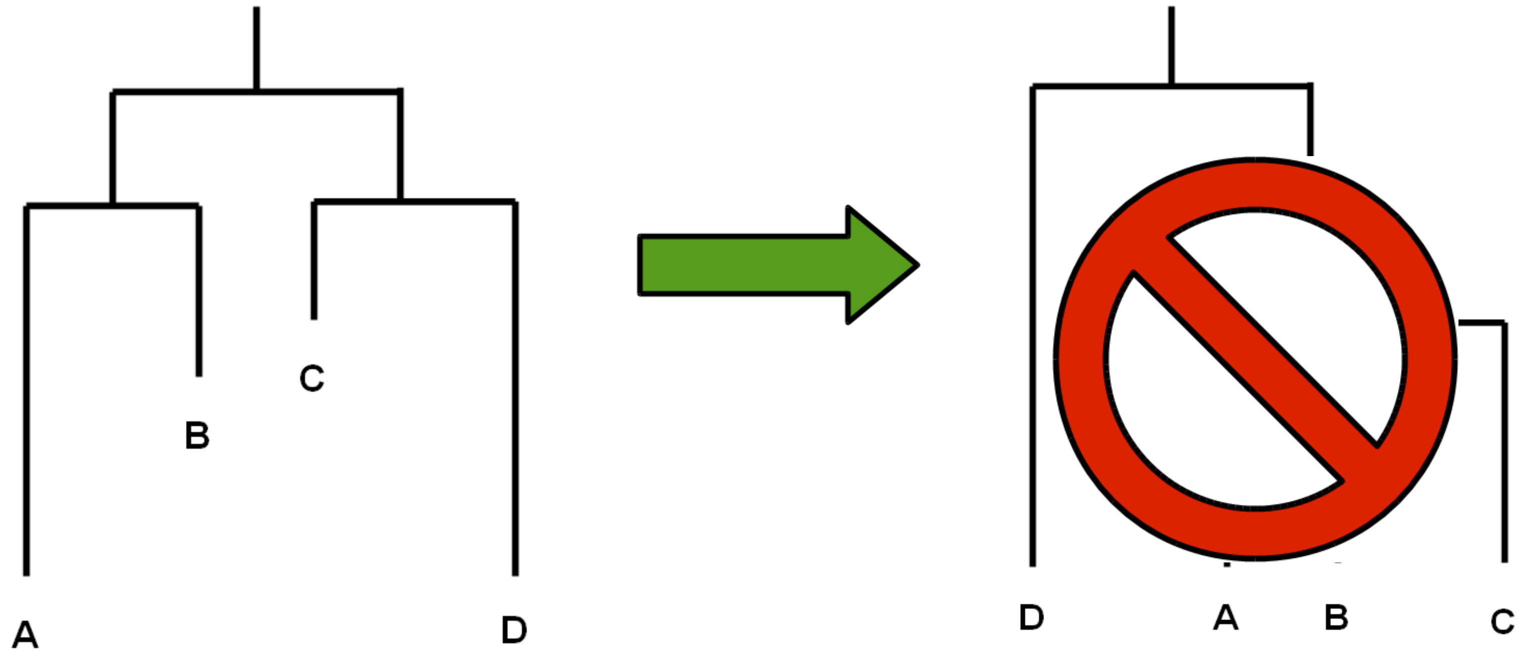
	AB	CDE
AB	0	
CDE	72.5	0



- There are only two clusters. so this completes the calculation!

Downside of UPGMA

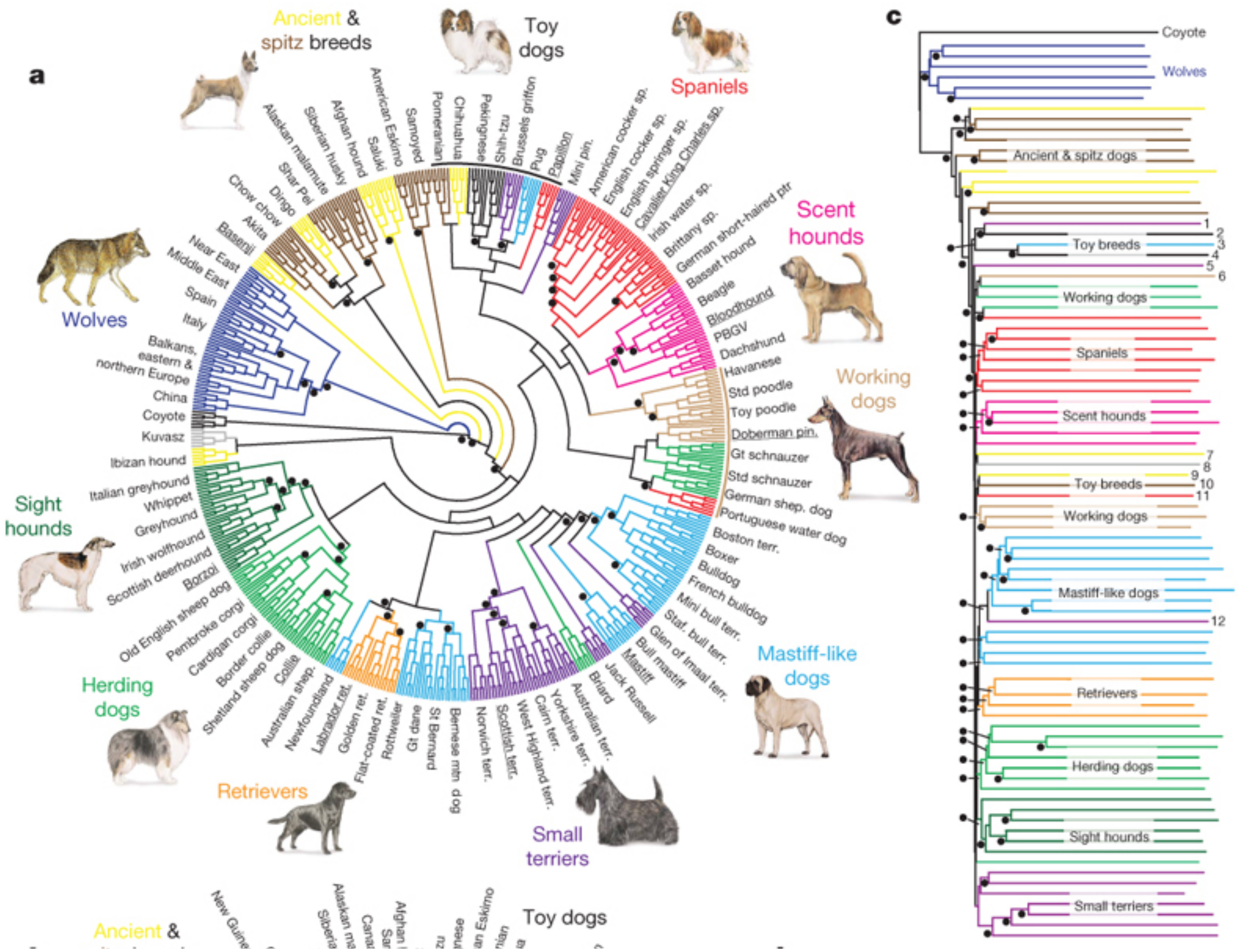
- Assume molecular clock (assuming the evolutionary rate is approximately constant)
- Clustering works only if the data is ultrametric
- Doesn't work the following case:



Neighbor-joining method

- Developed in 1987 by Saitou and Nei
- Works in a similar fashion to UPGMA
- Still fast – works great for large dataset
- Doesn't require the data to be ultrametric
- Great for largely varying evolutionary rates

Example Neighbor-Joining Tree for Dogs



BM vonHoldt *et al. Nature* **000**, 1-5 (2010) doi:10.1038/nature08837