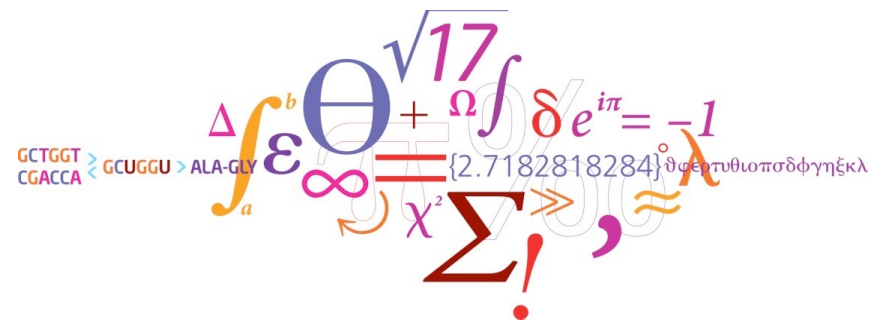


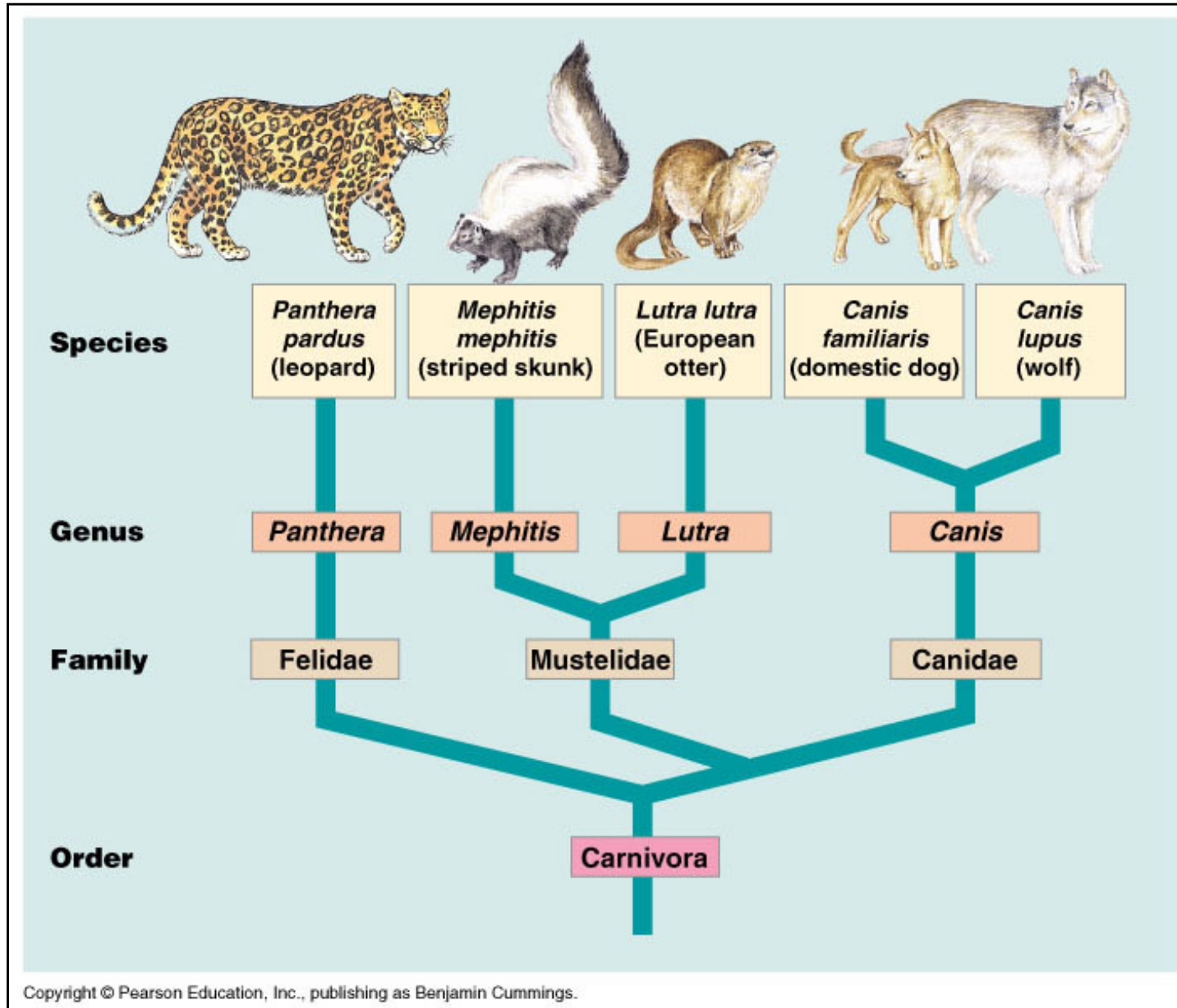
Phylogenetic Reconstruction: Distance Matrix Methods

Bent Petersen

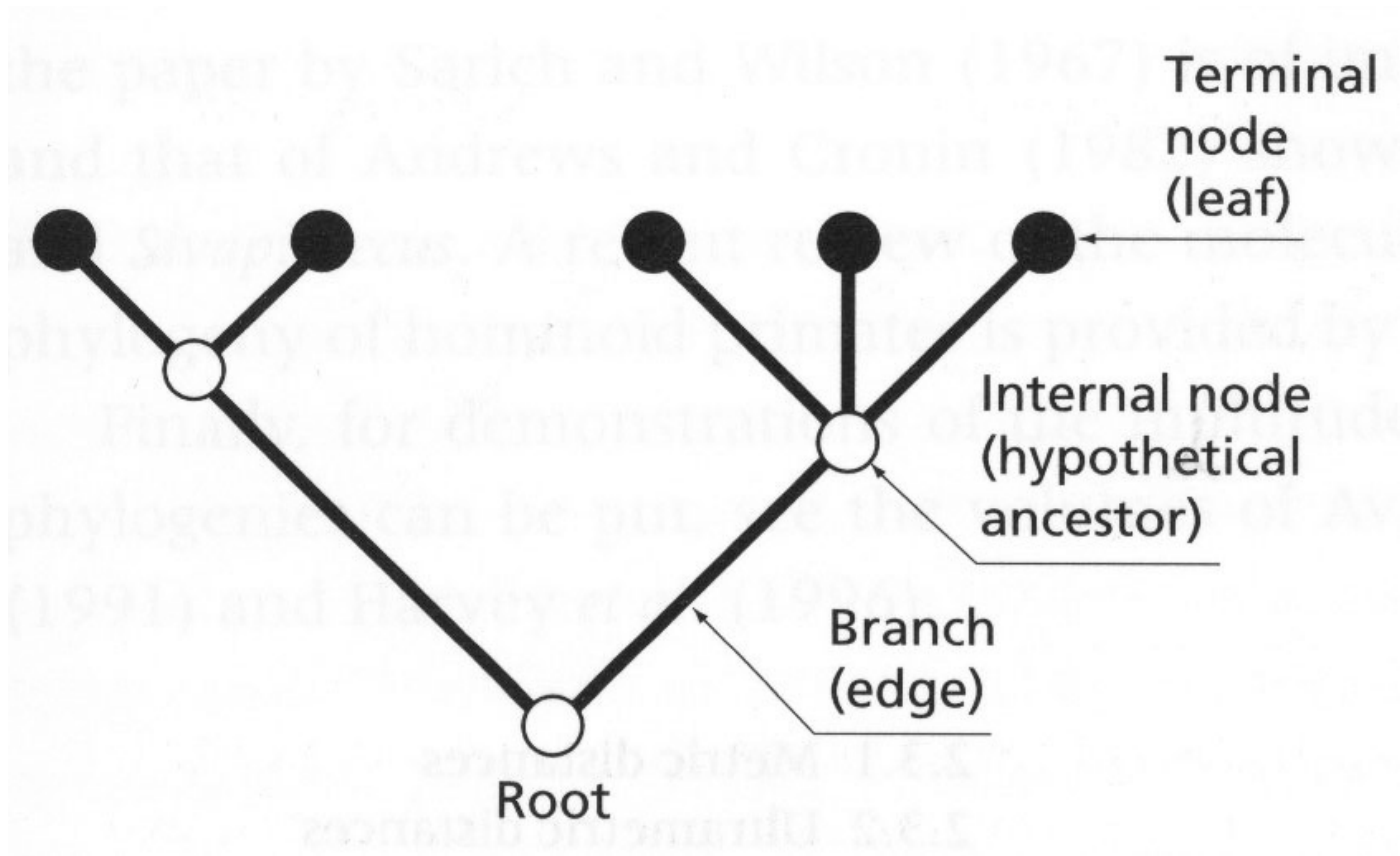
With slides from Anders Gorm Pedersen



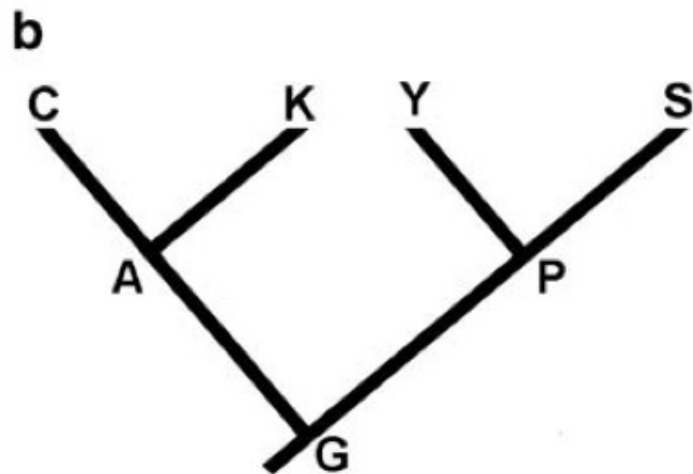
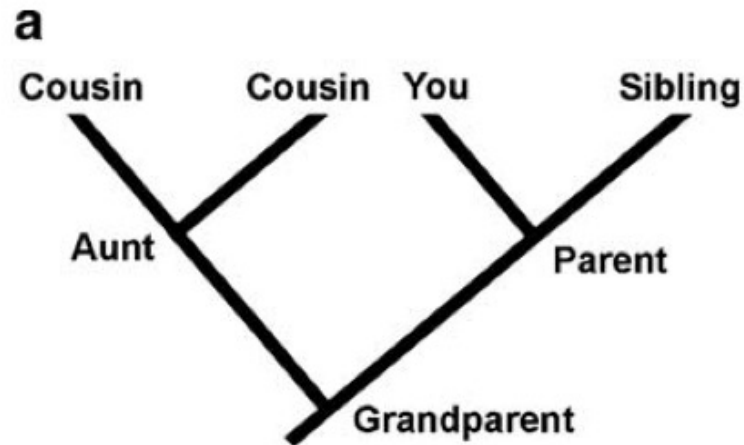
Recap: From lecture #1



Trees: meaning

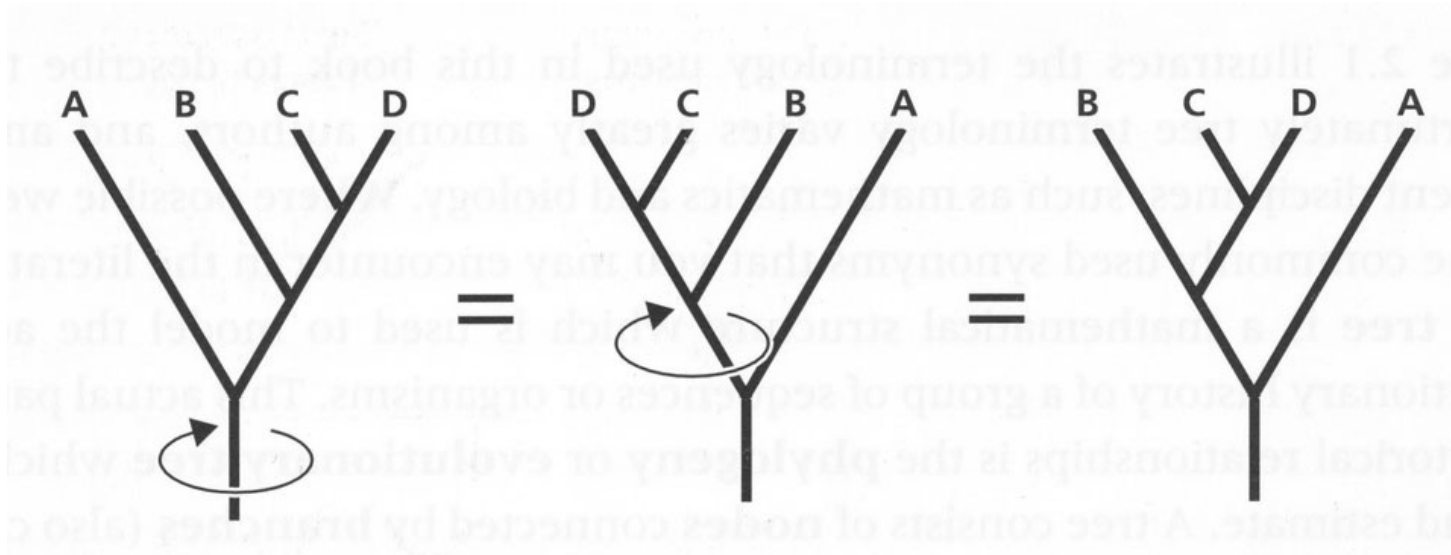


Trees: meaning



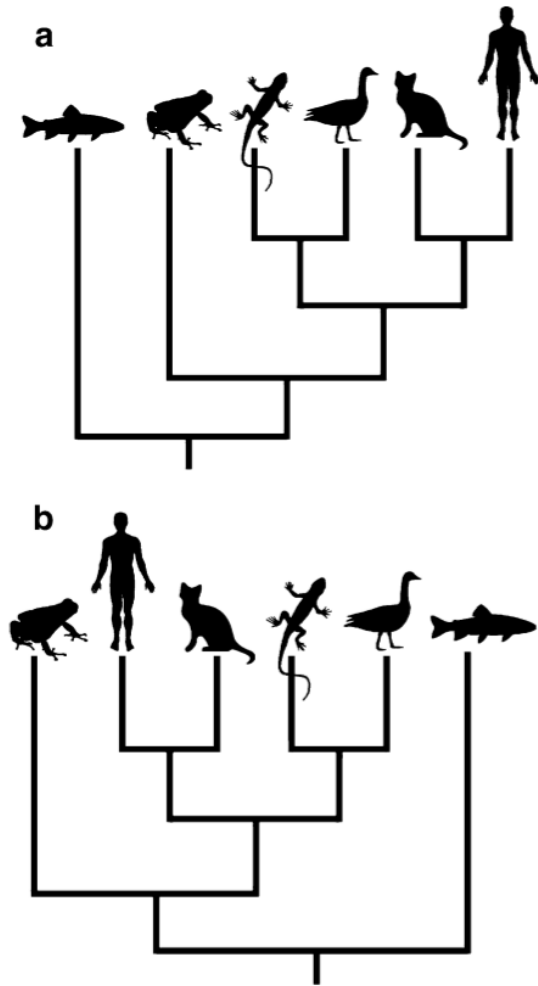
- Phylogenies indicate both relatedness and historical descent
- Y(ou) not descended from S(ister) (or vice versa) - both are contemporary and descended from P(arent)
- S and C are less closely related than S and Y: Their common ancestor is deeper in the tree

Trees: representations



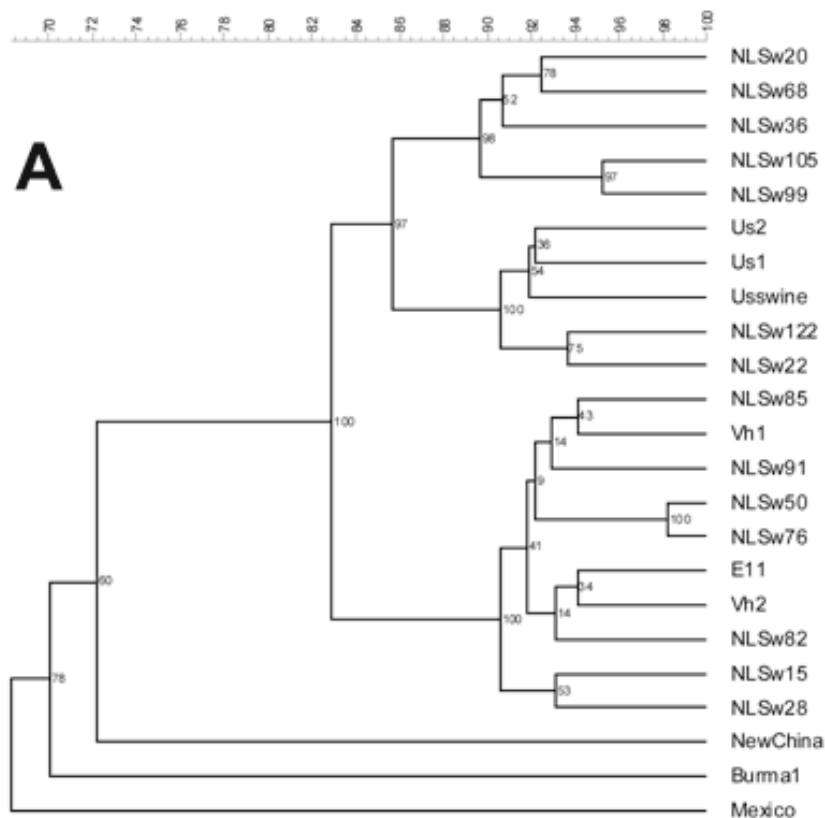
Three different representations of the same tree

Trees: Order of leafs is meaningless



- Order of terminal nodes contain no information about relatedness
- Frogs and humans are equally closely related to fishes

Trees: rooted vs. unrooted



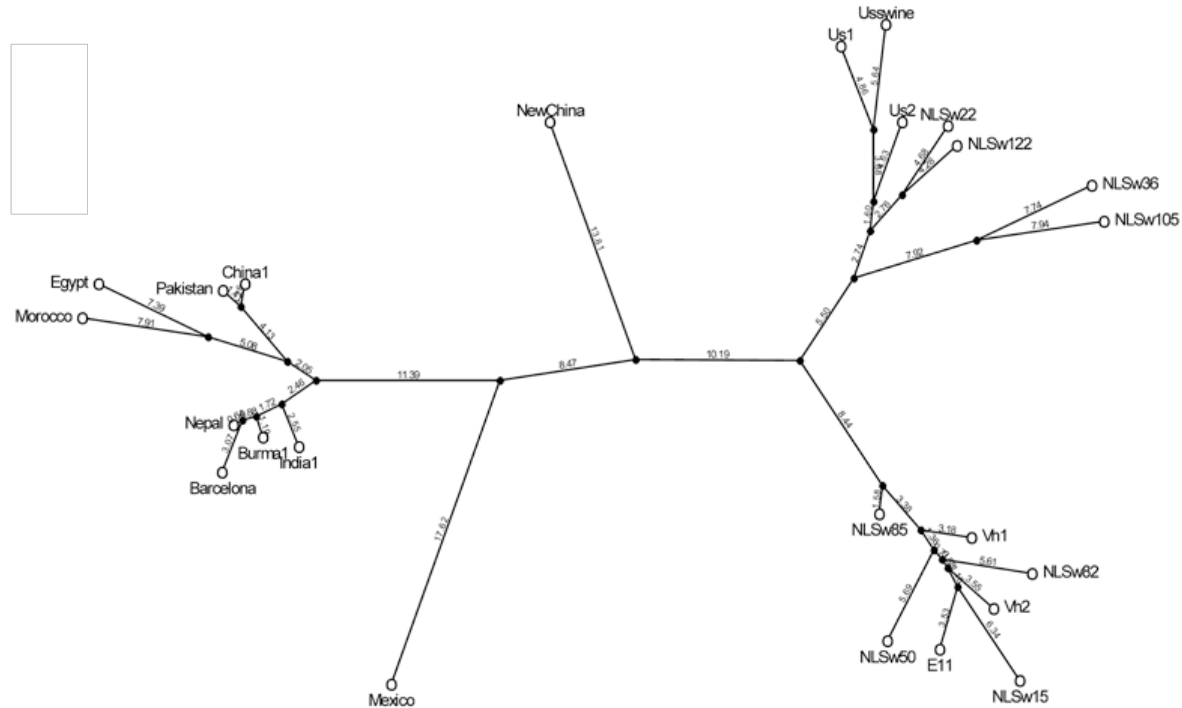
A rooted tree has a single node (the root) that represents a point in time that is earlier than any other node in the tree.

A rooted tree has directionality (nodes can be ordered in terms of “earlier” or “later”).

In the rooted tree, distance between two nodes is represented along the time-axis only (the second axis just helps spread out the leafs)

Early  Late

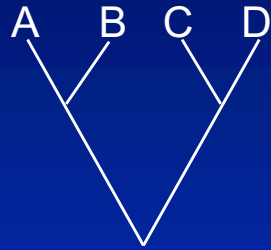
Trees: rooted vs. unrooted



- In unrooted trees there is no directionality: we do not know if a node is earlier or later than another node
- Distance along branches directly represents node distance

Notice: All distance based trees are unrooted, unless an “outgroup” can be specified

Trees: representation in computer files



`((A , B) , (C , D));`

Newick format:

- Leafs: represented by taxon name
- Internal nodes: represented by pair of matching parentheses
- Descendants of internal node given as comma-delimited list.
- Tree string terminated by semicolon

Trees: representation in computer files



Newick format:

- Leafs: represented by taxon name
- Internal nodes: represented by pair of matching parentheses
- Descendants of internal node given as comma-delimited list.
- Tree string terminated by semicolon

Trees: representation in computer files



Newick format:

- Leafs: represented by taxon name
- Internal nodes: represented by pair of matching parentheses
- Descendants of internal node given as comma-delimited list.
- Tree string terminated by semicolon

Trees: representation in computer files



Newick format:

- Leafs: represented by taxon name
- Internal nodes: represented by pair of matching parentheses
- Descendants of internal node given as comma-delimited list.
- Tree string terminated by semicolon

Trees: representation in computer files



Newick format:

- Leafs: represented by taxon name
- Internal nodes: represented by pair of matching parentheses
- Descendants of internal node given as comma-delimited list.
- Tree string terminated by semicolon

Trees: representation in computer files



Newick format:

- Leafs: represented by taxon name
- Internal nodes: represented by pair of matching parentheses
- Descendants of internal node given as comma-delimited list.
- Tree string terminated by semicolon

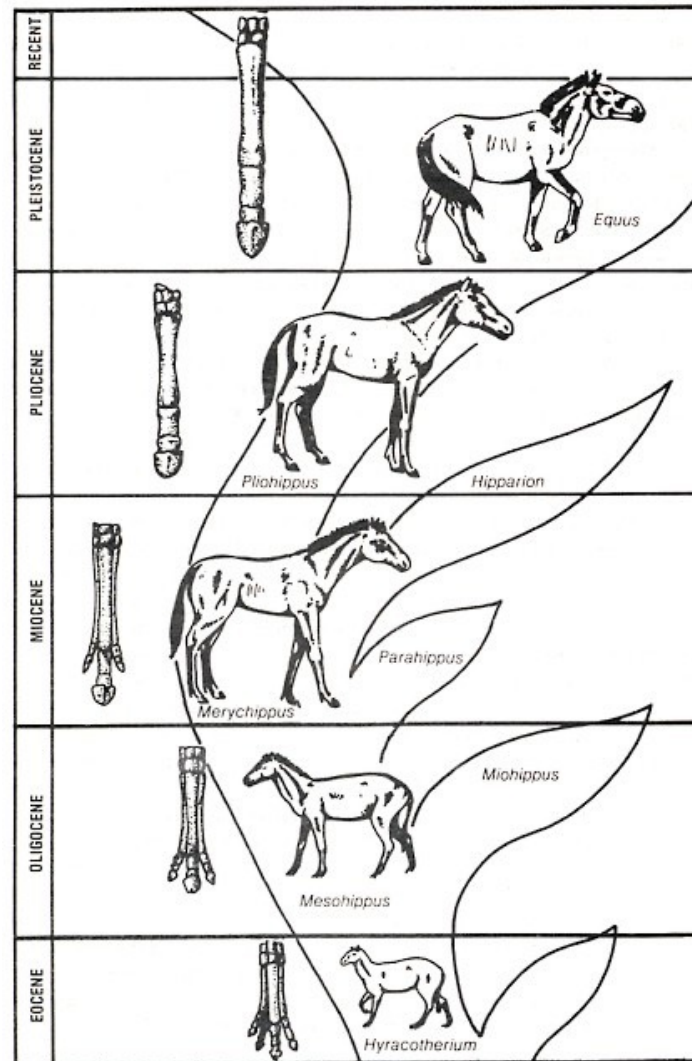
Trees: representation in computer files



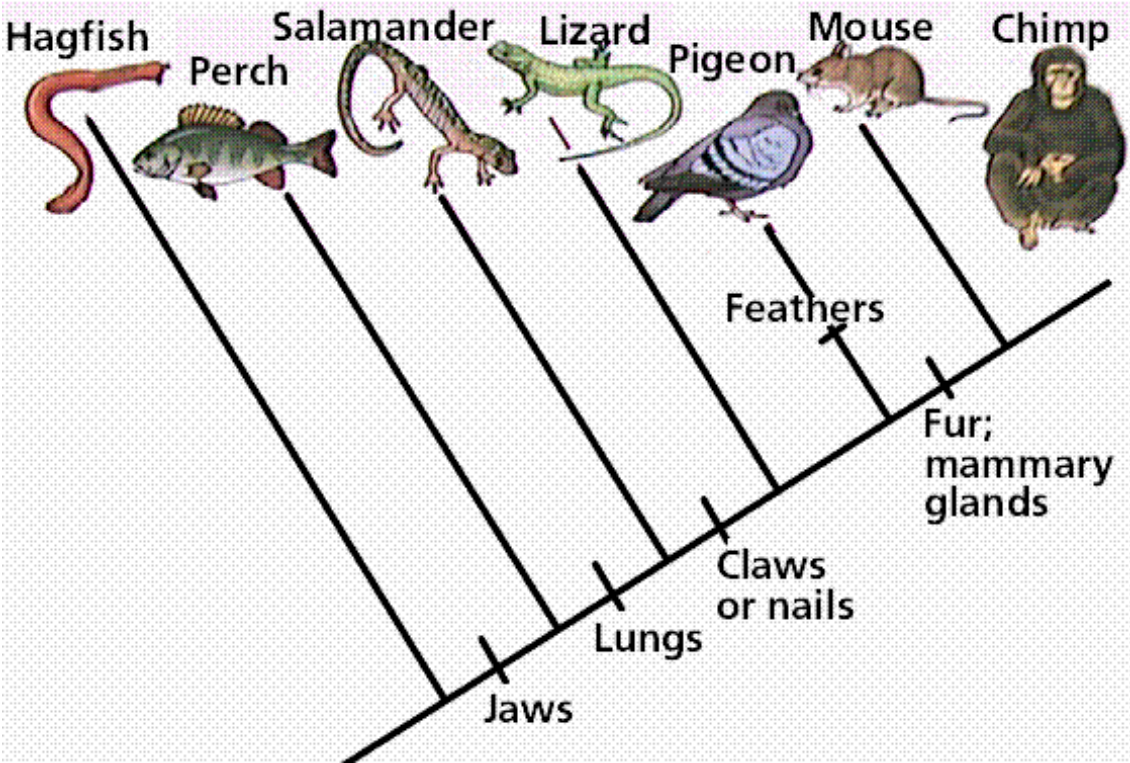
Newick format:

- Leafs: represented by taxon name
- Internal nodes: represented by pair of matching parentheses
- Descendants of internal node given as comma-delimited list.
- Tree string terminated by semicolon

Reconstructing a tree using non-contemporaneous data



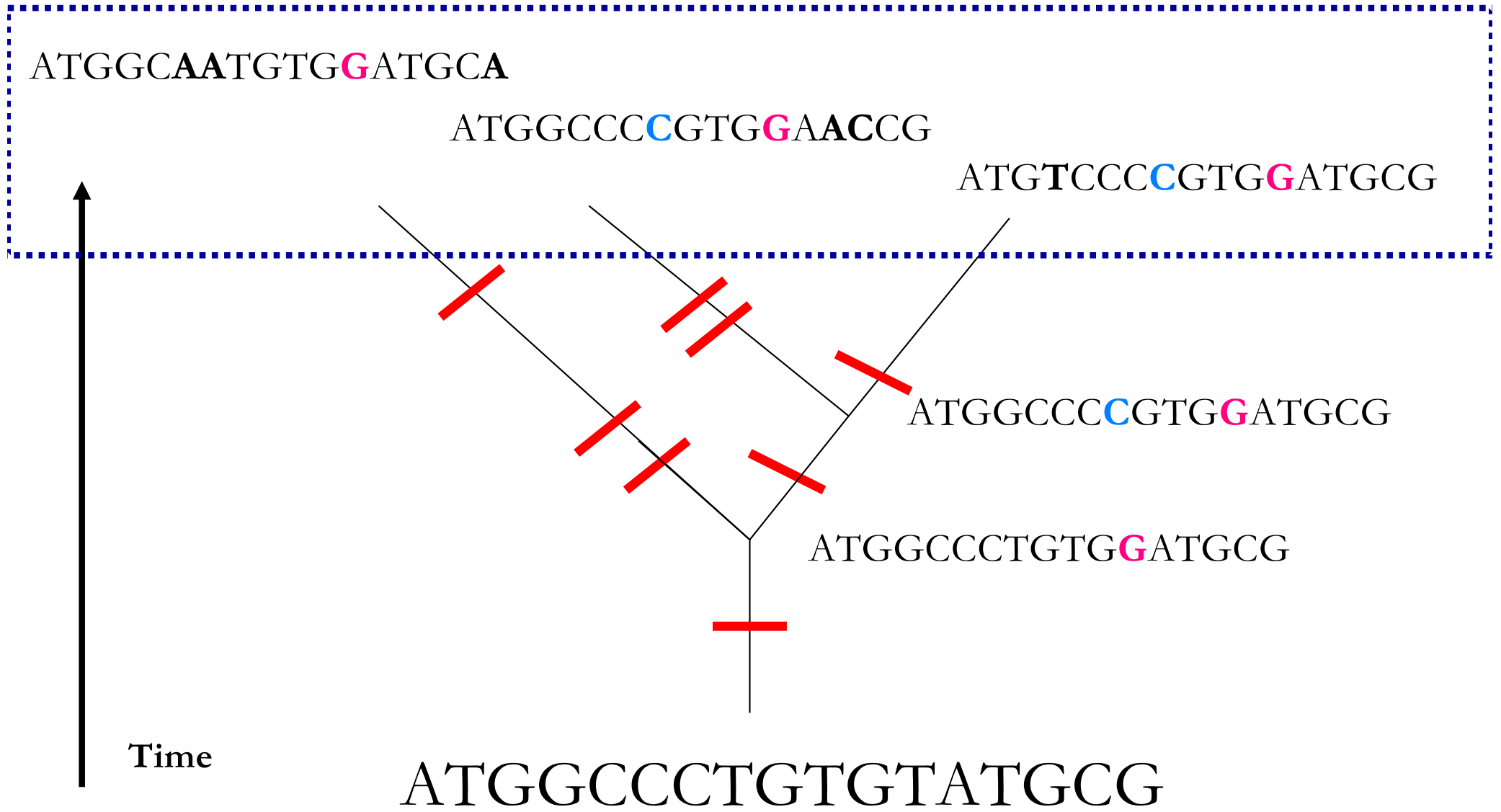
Reconstructing a tree using present-day data



Data: molecular phylogeny

- DNA sequences
 - genomic DNA
 - mitochondrial DNA
 - chloroplast DNA
 - Protein sequences
- Restriction site polymorphisms
 - DNA/DNA hybridization
 - Immunological cross-reaction

Recap: Example from Lecture #1



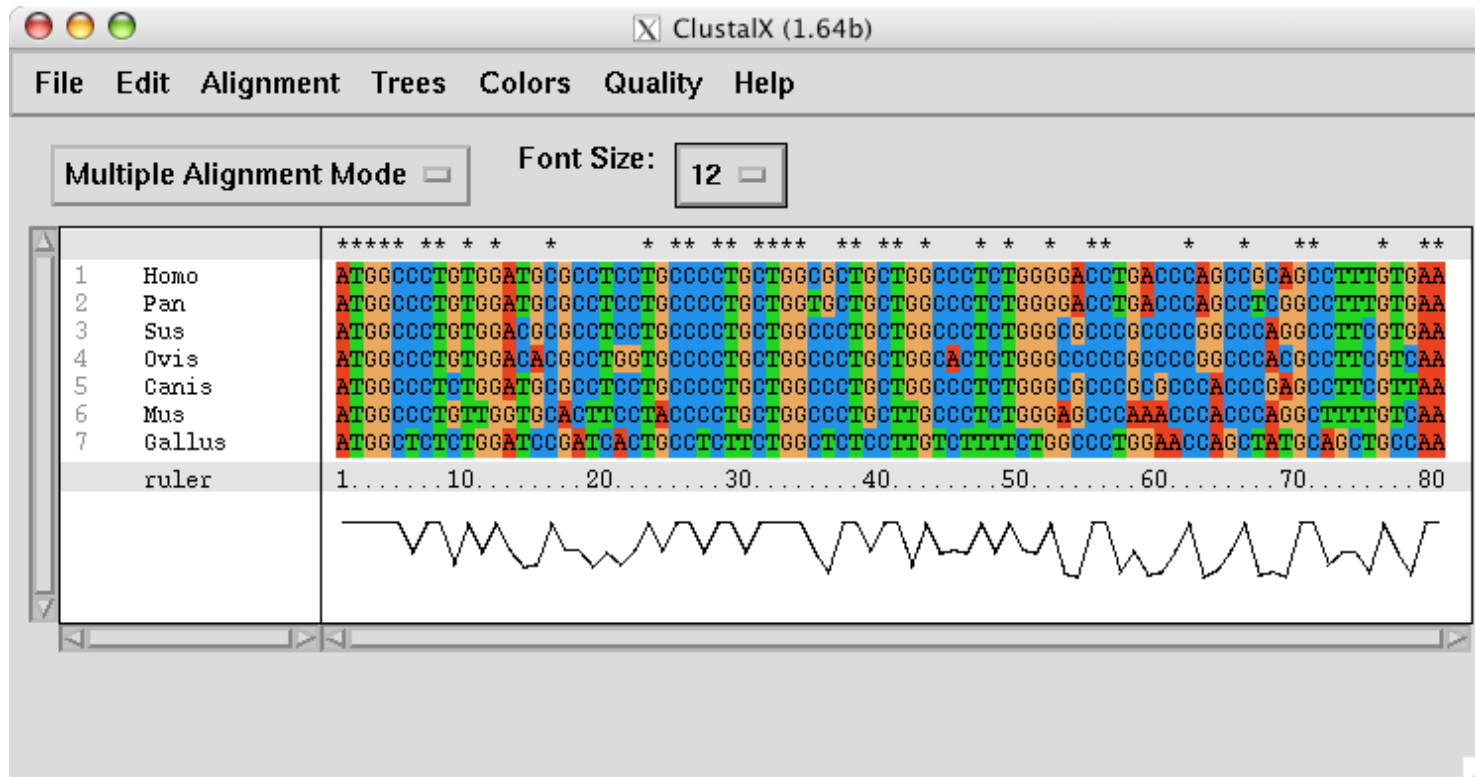
Recap: Example from Lecture #1

- Species1: ATGGC**AA**TGTG**G**ATG**CA**
 - Species2: ATGGCCC**C**GTG**G**A**AC**CG
 - Species3: ATG**T**CCC**C**GTG**G**ATGCG
- Diagram illustrating the alignment of three DNA sequences (Species1, Species2, Species3) with gaps. The sequences are aligned as follows:
- | | | | | | |
|-----------|--|---|---|---|---|
| Species1: | ATGGC AA TGTG G ATG CA |] | 6 |] | 5 |
| Species2: | ATGGCCC C GTG G A AC CG |] | 3 | | |
| Species3: | ATG T CCC C GTG G ATGCG |] | | | |

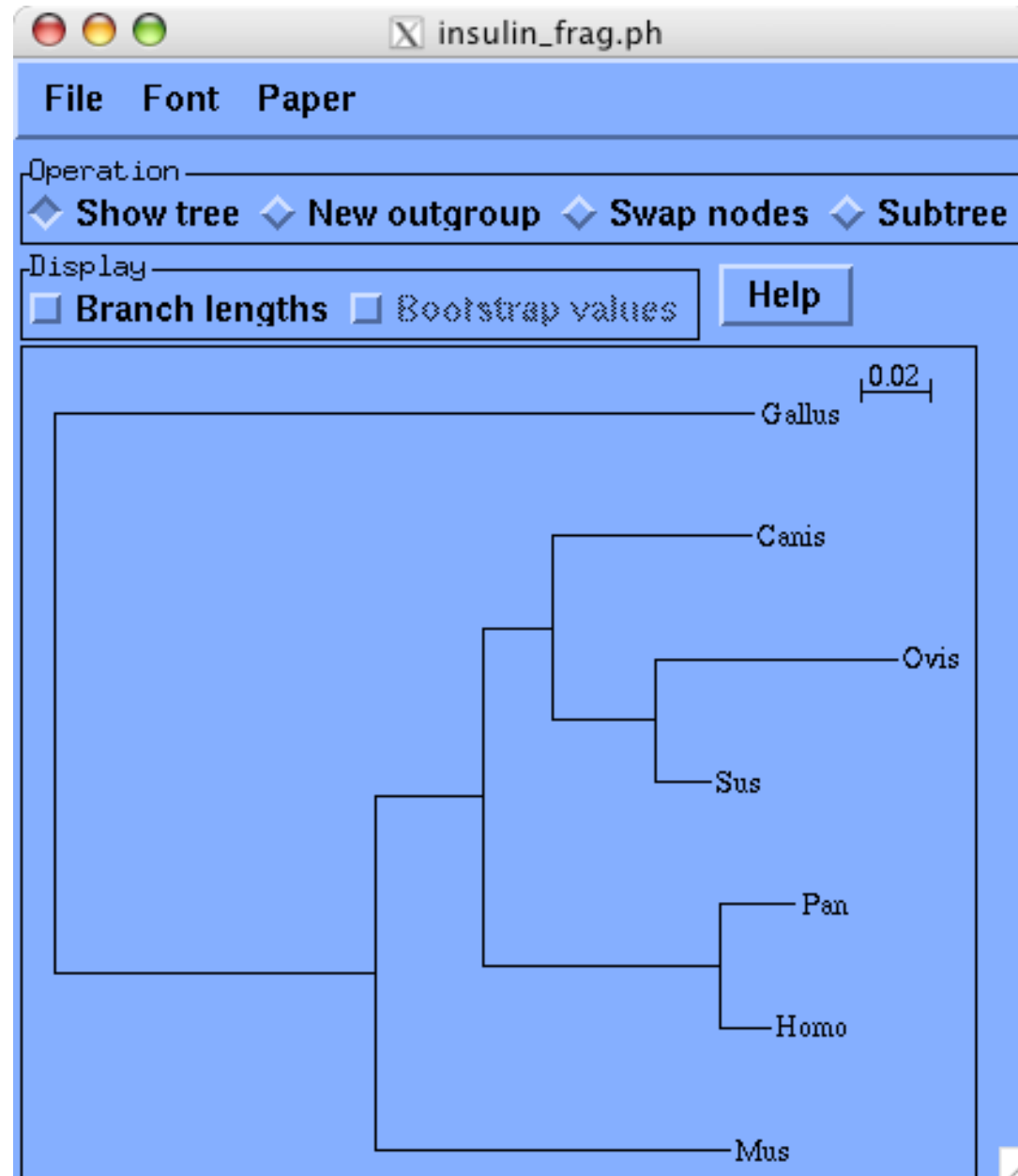
Recap: Example from Lecture #1

Insulin from 7 different species

Homo: ATGGCCCTGTGGATGCGCCTCCTGCCCTGCTGGCGCTGCTGGCCCTCTGGGGACCTGACCCAGCCGACGCTTTGTGAA
 Pan: ATGGCCCTGTGGATGCGCCTCCTGCCCTGCTGGTGCTGCTGGCCCTCTGGGGACCTGACCCAGCCTCGGCCCTTTGTGAA
 Sus: ATGGCCCTGTGGACGCGCCTCCTGCCCTGCTGGCCCTGCTGGCCCTCTGGGGCCCCGCCCCGCCCCAGGCCTTCGTGAA
 Ovis: ATGGCCCTGTGGACACGCCTGGTGCCCTGCTGGCCCTGCTGGCACTCTGGGCCCCGCCCCGCCCCACGCCTTCGTCAA
 Canis: ATGGCCCTCTGGATGCGCCTCCTGCCCTGCTGGCCCTGCTGGCCCTCTGGGGCCCCGCCCCACCCGAGCCTTCGTAA
 Mus: ATGGCCCTGTTGGTGCACCTTCTACCCCTGCTGGCCCTGCTTGCCCTCTGGGAGCCCAAACCCACCCAGGCTTTTGTCAA
 Gallus: ATGGCTCTCTGGATCCGATCACTGCCTCTTCTGGCTCTCCTTGTCTTTTCTGGCCCTGGAACCAGCTATGCAGCTGCCAA



Recap: Example from Lecture #1



Morphology vs. molecular data



African white-backed vulture
(old world vulture)



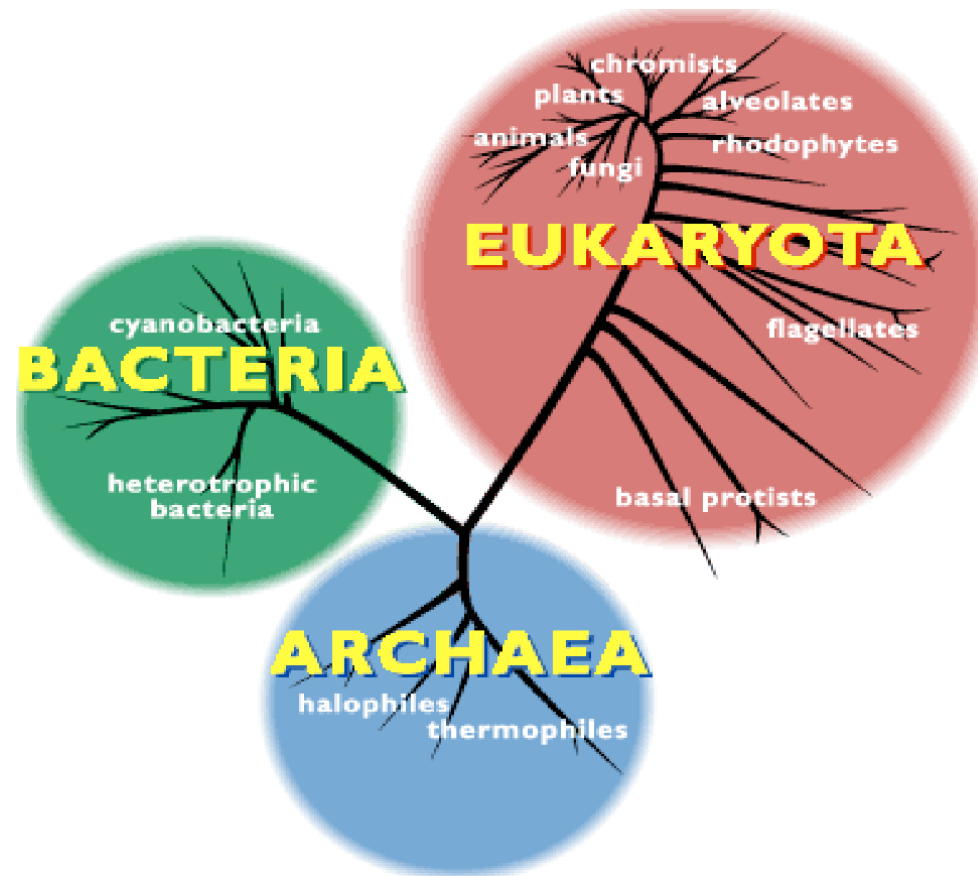
Andean condor
(new world vulture)

New and old world vultures seem to be closely related based on morphology.

Molecular data indicates that old world vultures are related to birds of prey (falcons, hawks, etc.) while new world vultures are more closely related to storks

Similar features presumably the result of convergent evolution

Molecular data: single-celled organisms



Molecular data useful for analyzing single-celled organisms (which have only few prominent morphological features).

Distance Matrix Methods

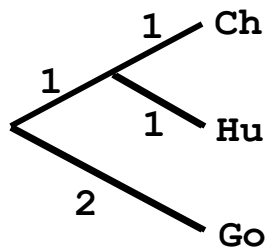
```

                ↓ ↓ ↓ ↓
Gorilla      :  ACGTCGTA
Human        :  ACGTTCCT
Chimpanzee   :  ACGTTTCG
                ↑   ↑
  
```

1. Construct multiple alignment of sequences

	Go	Hu	Ch
Go	-	4	4
Hu		-	2
Ch			-

2. Construct table listing all pairwise differences (distance matrix)

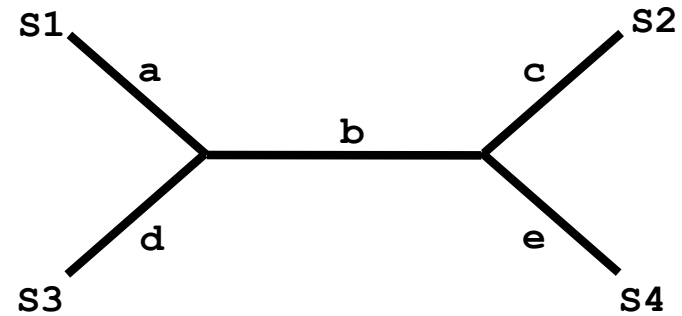


3. Construct tree from pairwise distances

Finding Optimal Branch Lengths

	S ₁	S ₂	S ₃	S ₄
S ₁	-	D ₁₂	D ₁₃	D ₁₄
S ₂		-	D ₂₃	D ₂₄
S ₃			-	D ₃₄
S ₄				-

Observed distance



Distance along tree

Goal:

$$D_{12} \approx d_{12} = a + b + c$$

$$D_{13} \approx d_{13} = a + d$$

$$D_{14} \approx d_{14} = a + b + e$$

$$D_{23} \approx d_{23} = d + b + c$$

$$D_{24} \approx d_{24} = c + e$$

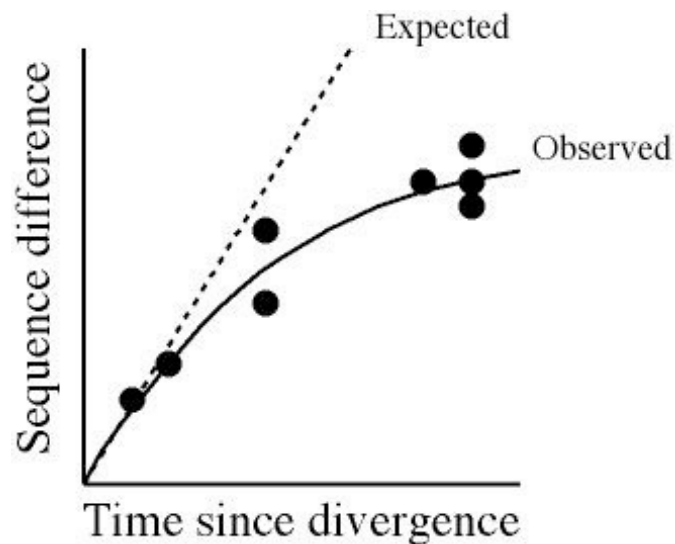
$$D_{34} \approx d_{34} = d + b + e$$

Handout exercise

Superimposed Substitutions

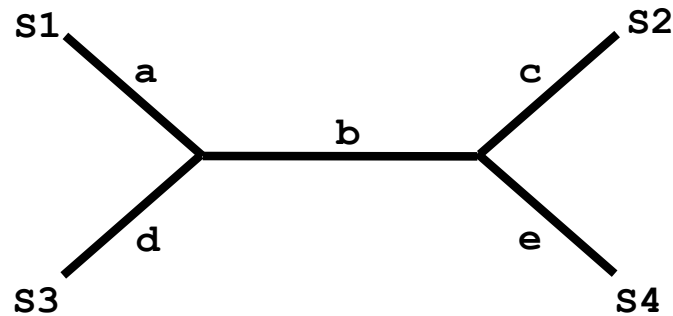
ACGGTGC
↓ ↓
C T
↓ ↓
GCGGTGA

- Actual number of evolutionary events: 5
- Observed number of differences: 2



- Distance is (almost) always underestimated
- Real distance can be estimated from observed distance using models of how evolution occurs

Optimal Branch Lengths: Least Squares



Distance along tree

- Fit between given tree and observed distances can be expressed as “sum of squared differences”:

$$Q = \sum_{j>i} (D_{ij} - d_{ij})^2$$

- Find branch lengths that minimize Q - this is the optimal set of branch lengths for this tree.

Goal:

$$D_{12} \approx d_{12} = a + b + c$$

$$D_{13} \approx d_{13} = a + d$$

$$D_{14} \approx d_{14} = a + b + e$$

$$D_{23} \approx d_{23} = d + b + c$$

$$D_{24} \approx d_{24} = c + e$$

$$D_{34} \approx d_{34} = d + b + e$$

Least Squares Optimality Criterion

- Search through all (or many) tree topologies
- For each investigated tree, find best branch lengths using least squares criterion
- Among all investigated trees, the best tree is the one with the smallest sum of squared errors.

Exhaustive search impossible for large data sets

No. taxa	No. trees
3	1
4	3
5	15
6	105
7	945
8	10,395
9	135,135
10	2,027,025
11	34,459,425
12	654,729,075
13	13,749,310,575
14	316,234,143,225
15	7,905,853,580,625

Heuristic search

1. Construct initial tree; determine sum of squares
2. Construct set of “neighboring trees” by making small rearrangements of initial tree; determine sum of squares for each neighbor
3. If any of the neighboring trees are better than the initial tree, then select it/them and use as starting point for new round of rearrangements. (Possibly several neighbors are equally good)
4. Repeat steps 2+3 until you have found a tree that is better than all of its neighbors.
5. This tree is a “local optimum” (not necessarily a global optimum!)

Clustering Algorithms

- Starting point: Distance matrix
- Cluster least different pair of sequences:
- Repeat until all nodes are linked
- Results in only one tree, there is no measure of tree-goodness.

Neighbor Joining Algorithm

- For each tip compute $u_i = \sum_j D_{ij} / (n-2)$
 (this is essentially the average distance to all other tips, except the denominator is n-2 instead of n)
- Find the pair of tips, i and j, where $D_{ij} - u_i - u_j$ is smallest
- Connect the tips i and j, forming a new ancestral node. The branch lengths from the ancestral node to i and j are:

$$v_i = 0.5 D_{ij} + 0.5 (u_i - u_j)$$

$$v_j = 0.5 D_{ij} + 0.5 (u_j - u_i)$$

- Update the distance matrix: Compute distance between new node and each remaining tip as follows:

$$D_{ij,k} = (D_{ik} + D_{jk} - D_{ij}) / 2$$

- Replace tips i and j by the new node which is now treated as a tip
- Repeat until only two nodes remain.

NJ visualized

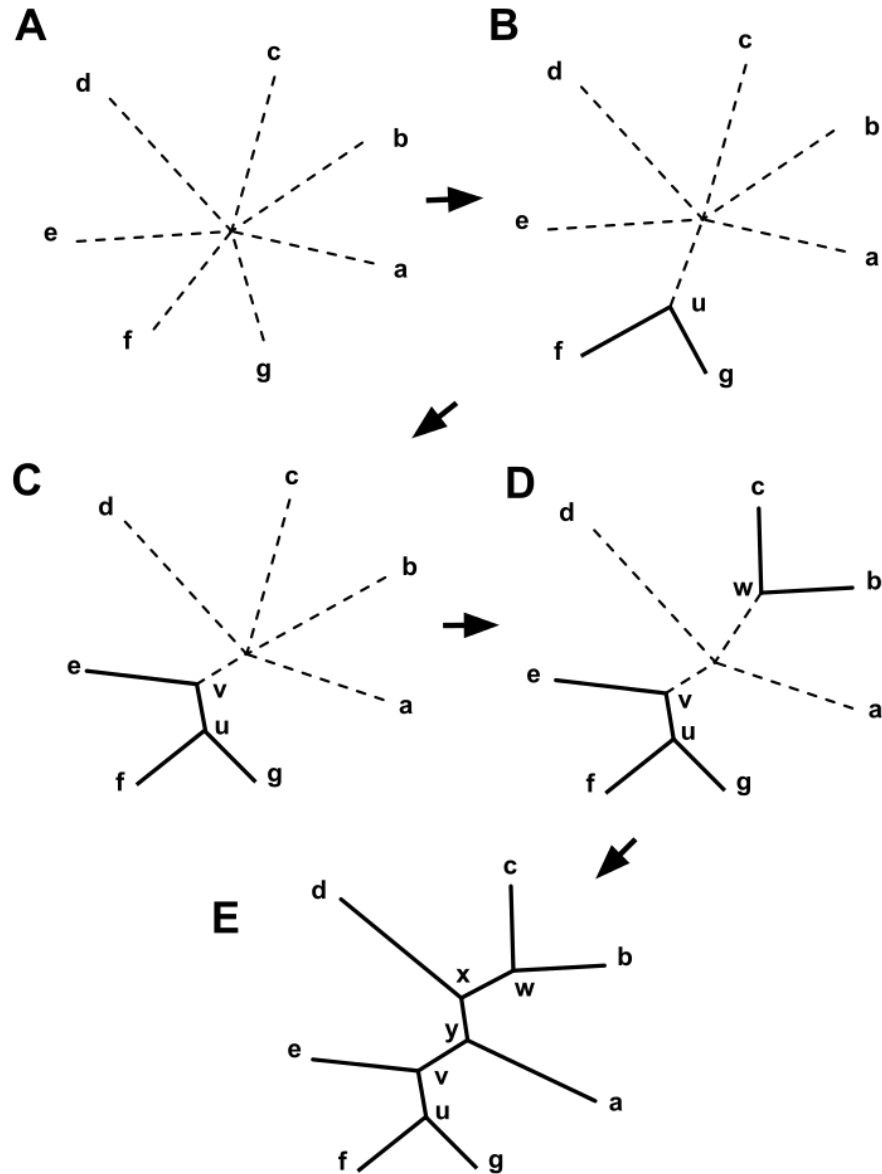


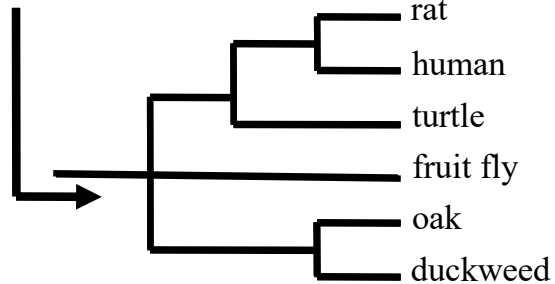
Image source: Wikipedia

Bootstrapping

Original alignment

```

0123456789
rat      GAGGCTTATC
human    GTGGCTTATC
turtle   GTGCCCTATG
fruitfly CTCGCCTTTG
oak      ATCGCTCTTG
duckweed ATCCCTCCGG
  
```

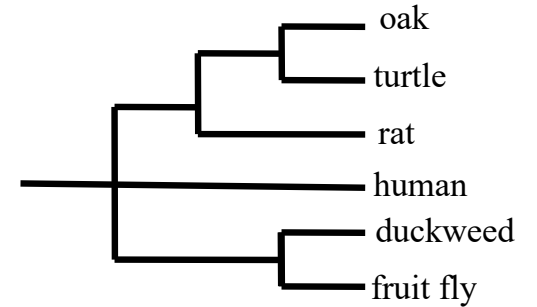


Inferred tree

Bootstrapped alignment 1

```

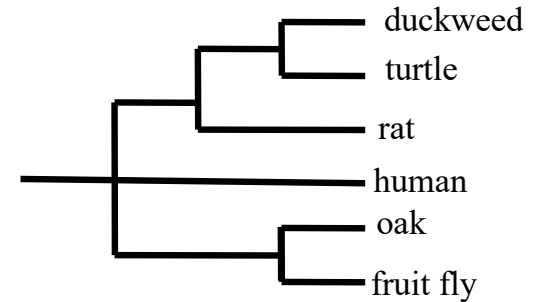
001122234556667
rat      GGAAGGGGCTTTTTA
human    GGTTGGGGCTTTTTA
turtle   GGTTGGGGCCCTTTA
fruitfly CCTTCCCGCCCTTTT
oak      AATTCCCGCTTCCCT
duckweed AATTCCCCCTTCCCC
  
```



Bootstrapped alignment 2

```

445556777888899
rat      CCTTTTAAATTTTCC
human    CCTTTTAAATTTTCC
turtle   CCCCCTAAATTTTGG
fruitfly CCCCCTTTTTTTTGG
oak      CCTTTCTTTTTTTTGG
duckweed CCTTTCCCCGGGGGG
  
```



.

.

.

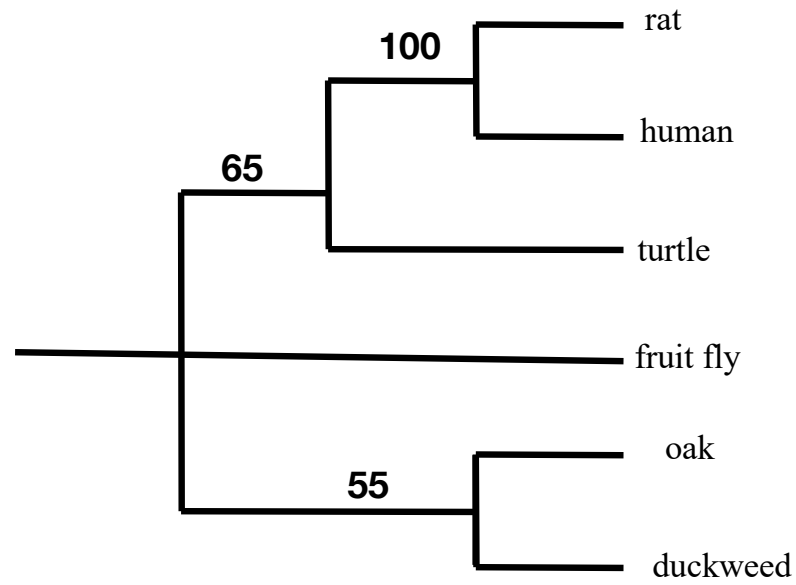
.

.

.

↓ Many more replicates
 (between 100 - 1000)

Bootstrapping values



Key Takeaways

- **Phylogenetic Reconstruction:** Distance matrix methods allow us to explore evolutionary relationships by measuring genetic differences.
- **Tree Structures:** Rooted and unrooted trees provide different perspectives on relatedness and evolutionary time.
- **Importance of Molecular Data:** Molecular sequences are crucial, especially for organisms with few distinguishing physical features.
- **Tree Building Challenges:** Methods like least squares optimization and heuristic searches help manage large datasets and complex tree structures.
- **Confidence in Results:** Bootstrapping and Neighbor Joining provide robust ways to test and refine our phylogenetic trees.