Lecture 08: Phylogenetic Analyses

CSCI-478/CSCI-578/BIOL-510 Bioinformatics, Fall 2021

Hua Wang, Ph.D.

Department of Computer Science
Colorado School of Mines

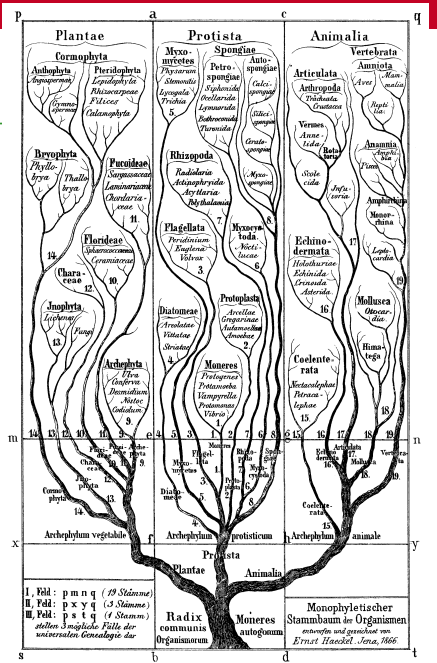
September 23, 2021

What is Molecular Phylogenetics?

What is Molecular Phylogenetics?

Phylogeny

- Term coined by Ernst Haeckel (1866)
  - Phylon ($\phi\upsilon\lambda o\nu$)
    - Tribe
    - Race
  - Genus
    - Birth
    - Origin
- At every node in the tree, a new lineage is born.
- All lineages in a tree are **related** because they descend from the same root.
- Tree topology shows how the lineages are related.



What is Molecular Phylogenetics?

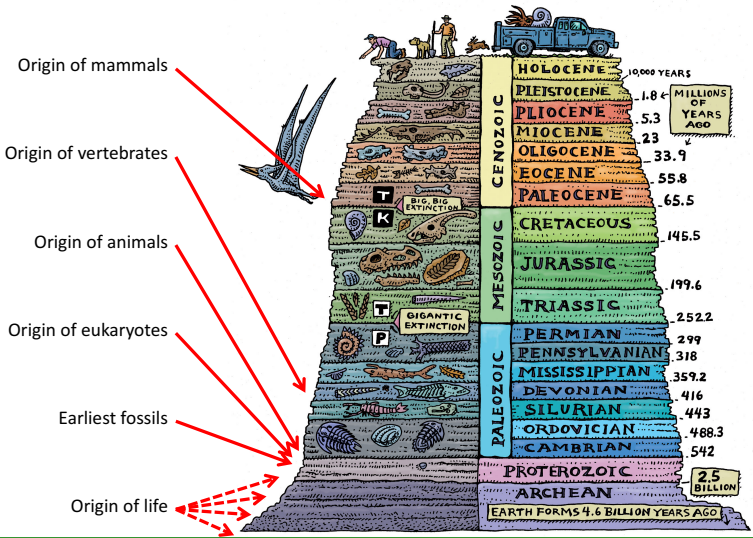
- **Phylogenetics**
  - the study of evolutionary relationships in organisms,
  - one part of the larger field of **systematics**, which also includes taxonomy.
  - The term **taxonomy** connotes the process and methodology for the naming and classification of organisms.

- **The systematics**
  - the branch of biology that deals with classification and nomenclature; **taxonomy**

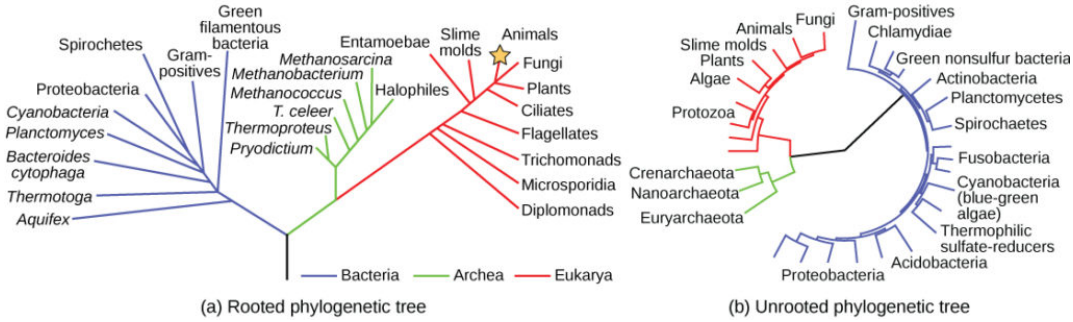
What is phylogeny good for?

- Evolutionary history ("tree of life")
- Population history
- Rates of evolutionary change
- Origins of diseases
- Prediction of sequence function
- Can be applied to organisms, sequences, viruses, languages, *etc.*

What is phylogeny good for? — look back in time



What is phylogeny good for? — understand evolution



What is phylogeny good for? — understand virus transmission

Molecular evidence of HIV-1 transmission in a criminal case

Michael L. Metzker*†, David P. Mindell‡, Xiao-Mei Liu*§, Roger G. Ptak¶, Richard A. Gibbs*, and David M. Hillis**

*Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030; †Department of Ecology and Evolutionary Biology and Museum of Zoology, University of Michigan, Ann Arbor, MI 48109-1079; §School of Dentistry, Biologic and Materials Sciences, University of Michigan, Ann Arbor, MI 48109; and **Section of Integrative Biology and Center for Computational Biology and Bioinformatics, University of Texas, Austin, TX 78712

Edited by Walter M. Fitch, University of California, Irvine, CA, and approved September 4, 2002 (received for review May 2, 2002)

A gastroenterologist was convicted of attempted second-degree murder by injecting his former girlfriend with blood or blood-products obtained from an HIV type 1 (HIV-1)-infected patient under his care. Phylogenetic analyses of HIV-1 sequences were admitted and used as evidence in this case, representing the first use of phylogenetic analyses in a criminal court case in the United States. Phylogenetic analyses of HIV-1 reverse transcriptase and *env* DNA sequences isolated from the victim, the patient, and a local population sample of HIV-1-positive individuals showed the victim's HIV-1 sequences to be most closely related to and nested within a lineage comprised of the patient's HIV-1 sequences. This finding of paraphyly for the patient's sequences was consistent with the direction of transmission from the patient to the victim. Analysis of the victim's viral reverse transcriptase sequences revealed genotypes consistent with known mutations that confer resistance to AZT, similar to those genotypes found in the patient. *A priori* establishment of the patient and victim as a suspected transmission pair provided a clear hypothesis for phylogenetic testing. All phylogenetic models and both genes examined strongly supported the close relationship between the HIV-1 sequences of the patient and the victim. Resampling of blood from the suspected transmission pair and independent sequencing by different laboratories provided precaution against laboratory error.

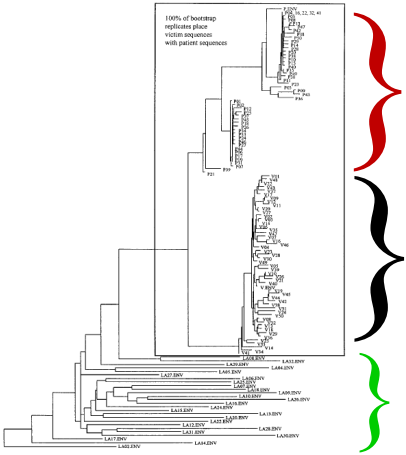
(13). This case was the first time that phylogenetic analysis has been used as evidence in a United States criminal proceeding. Here we present the phylogenetic evidence that constituted part of the prosecution's case that resulted in the conviction of the Louisiana gastroenterologist on the charge of attempted second degree murder.

Materials and Methods

Criminal Investigation. The prosecution's case was based on circumstantial evidence indicating that on August 4, 1994, a Lafayette, LA, gastroenterologist made a mixture of blood or blood-products from two patients under the doctor's care, one infected with HIV-1 and the other with hepatitis C, and infected his former girlfriend by intramuscular injection. Our efforts for the criminal investigation involved only the molecular analysis of HIV-1 sequences, which represented only one part of the prosecution's case against the physician.

Risk factors associated with HIV-1 infection for the victim were determined through the course of the criminal investigation. From 1984 to 1995, the victim reported having sexual contacts with seven men, including the doctor, all of whom were interviewed by local law enforcement agents. The seven men were tested between the

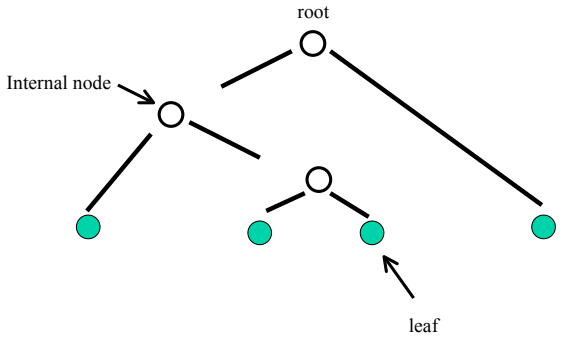
HIV Phylogeny



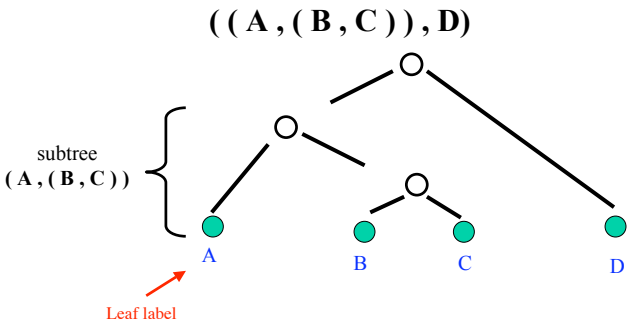
Tree terminology

Tree — a graph, a data structure

In graph theory, a tree is an undirected graph in which any two vertices are connected by exactly one path, or equivalently a connected acyclic undirected graph.



Tree terminology

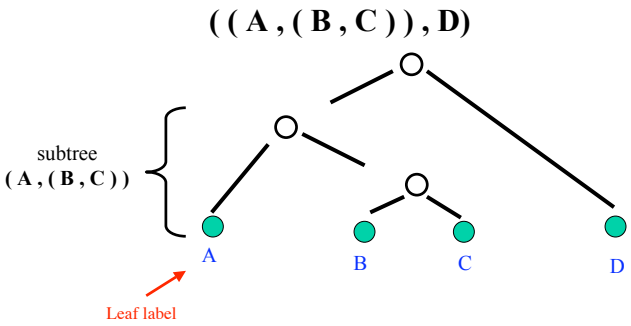


- A **tree** is either
  - a **leaf**

  or
  - (**LeftTree**, **RightTree**)

  where both LeftTree and RightTree are trees.

Tree terminology



- A **tree** is either

  a **leaf**

  or

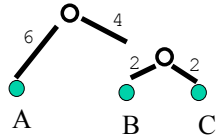
  (**LeftTree**, **RightTree**)

  where both LeftTree and RightTree are trees.

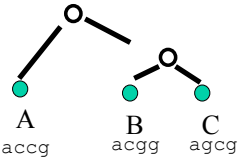
Tree terminology

Trees with labels

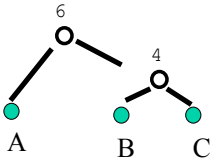
- We can add data to
  - Leaves
  - Branches
  - Nodes



(A:6 , (B : 2 , C : 2):4)



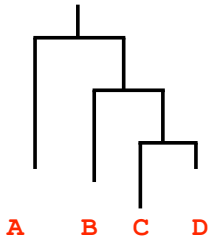
(A[accg] , (B [acgg] , C [agcg]))

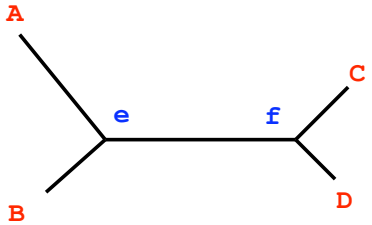


6:(A , 4:(B, C))

Tree terminology

Trees can be unrooted





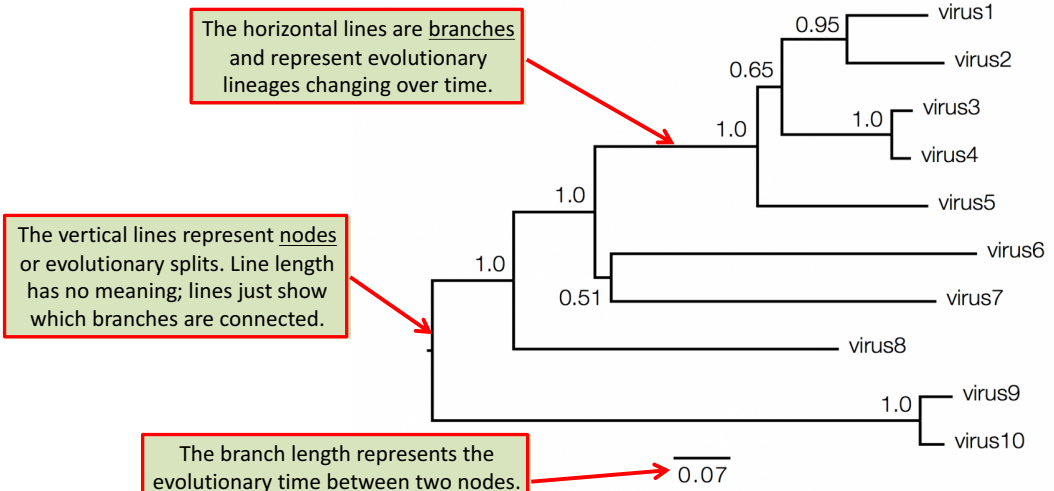
Are there alternative rootings?
Draw them…

Phylogenetic trees

The horizontal lines are branches and represent evolutionary lineages changing over time.

The vertical lines represent nodes or evolutionary splits. Line length has no meaning; lines just show which branches are connected.

The branch length represents the evolutionary time between two nodes. Unit: substitutions per sequence site.



http://epidemic.bio.ed.ac.uk/how_to_read_a_phylogeny

Phylogenetic trees

Task definition — constructing phylogenetic trees

Given:

- data characterizing a set of species/genes

Do:

- infer a phylogenetic tree that accurately characterizes the evolutionary lineages among species/genes

Phylogenetic trees

Task definition — constructing phylogenetic trees

Given:

- data characterizing a set of species/genes

Do:

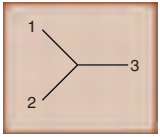
- infer a phylogenetic tree that accurately characterizes the evolutionary lineages among species/genes

A brute-forth algorithm

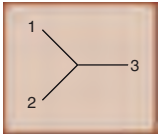
- Do an exhaustive search to examine **all** possible trees and selects the one with **the most optimal features**, such as the shortest overall sum of the branch lengths.
- But, is this computationally possible?

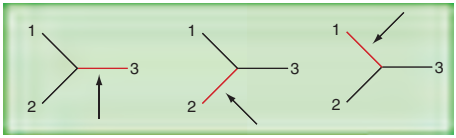
Phylogenetic trees — how many phelogenetic tree?

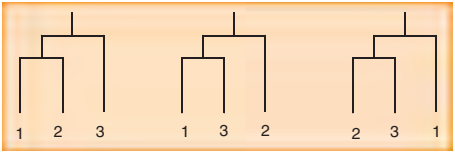
Phylogenetic trees — how many phelogenetic tree?



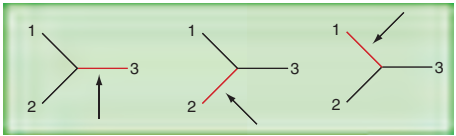
Phylogenetic trees — how many phelogenetic tree?

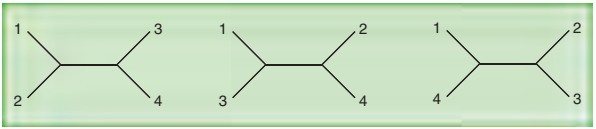




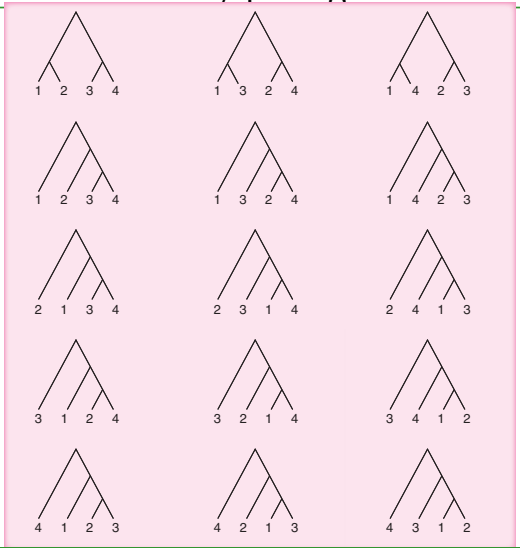


Phylogenetic trees — how many phelogenetic tree?





Phylogenetic trees — how many phelogenetic tree?



Phylogenetic trees — how many phelogenetic tree?

Given a set of n sequences, we can construct a rooted tree with

- $2n - 1$ nodes and $2n - 2$ edges.

Phylogenetic trees — how many phelogenetic tree?

Given a set of n sequences, we can construct a rooted tree with

- $2n - 1$ nodes and $2n - 2$ edges.

Given a set of n sequences, we can construct an unrooted tree with

- $2n - 2$ nodes and $2n - 3$ edges,
- which can produced $2n - 3$ difference rooted trees by placing the root on any one edge.

Phylogenetic trees — how many phelogenetic tree?

Given a set of n sequences, we can construct a rooted tree with

- $2n-1$ nodes and $2n-2$ edges.

Given a set of n sequences, we can construct an unrooted tree with

- $2n-2$ nodes and $2n-3$ edges,
- which can produced $2n-3$ difference rooted trees by placing the root on any one edge.

Given a set of n sequences, we can construct

- $\prod_{i=3}^{n}(2i-5)$ possible unrooted trees, and
- $(2n-3)\prod_{i=3}^{n}(2i-5)$ possible rooted trees.

Phylogenetic trees — how many phelogenetic tree?

Number of OTUs	Number of rooted trees	Number of unrooted trees
2	1	1
3	3	1
4	15	3
5	105	15
6	945	105
7	10,395	945
8	135,135	10,395
9	2,027,025	135,135
10	34,489,707	2,027,025
15	213,458,046,676,875	8×10^{12}
20	8×10^{21}	2×10^{20}
50	2.8×10^{76}	3×10^{74}

Approaches to phylogenetic trees

Three general types of methods:

- **Distance**: find tree that accounts for estimated evolutionary distances
- **Parsimony**: find tree that requires minimum number of changes to explain the data
- **Maximum likelihood**: find tree that maximizes the likelihood of the data

Approaches to phylogenetic trees

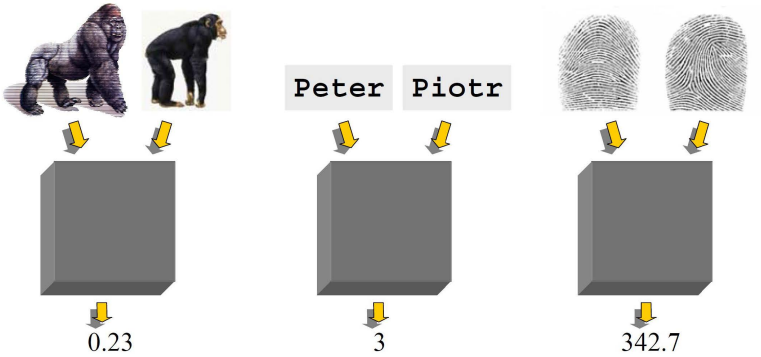
Three general types of methods:

- **Distance**: find tree that accounts for estimated evolutionary distances
- **Parsimony**: find tree that requires minimum number of changes to explain the data
- **Maximum likelihood**: find tree that maximizes the likelihood of the data

Distance-Based Methods

Distance

Let O_1 and O_2 be two objects from the universe of possible objects. The distance dissimilarity between O_1 and O_2 is a real number denoted by $D(O_1, O_2)$.



Distance-Based Methods

Properties should a distance measure

- Symmetry: $D(A, B) = D(B, A)$
  - *Otherwise you could claim: Alex looks like Bob, but Bob looks nothing like Alex.*
- Constancy of Self-Similarity: $D(A, A) = 0$
  - *Otherwise you could claim: Alex looks more like Bob, than Bob does.*
- Positivity Separation: $D(A, B) = 0$ If $A = B$
  - *Otherwise there are objects in your world that are different, but you cannot tell apart.*
- Triangular Inequality: $D(A, B) \leq D(A, C) + D(B, C)$
  - *Otherwise you could claim: Alex is very like Bob, and Alex is very like Carl, but Bob is very unlike Carl.*

Distance-Based Methods

Where do we get distances for input sequences?

- Commonly obtained from sequence alignments. In alignment of sequence i with sequence j:

$$dist(i,j) = \frac{\#mismatches}{\#mismatches + \#matches}$$

- To correct for multiple substitutions at a single position:

$$dist_{\text{Jukes-Cantor}} = -\frac{3}{4} \ln \left(1 - \frac{4}{3} dist(i,j)\right)$$

The UPGMA method

The Unweighted Pair Group Method using Arithmetic Averages

The Unweighted Pair Group Method using Arithmetic Averages (UPGMA) will reconstruct the tree T that is consistent with the data.

- Basic idea:
  - Iteratively pick two taxa/clusters and merge them
  - Create new node in tree for merged cluster
- Distance d_{ij} between clusters C_i and C_j of taxa is defined as:

$$d_{ij} = \frac{1}{|C_i||C_j|} \sum_{p \in C_i, q \in C_j} d_{pq}$$

The UPGMA method

The UPGMA algorithm

- Assign each taxon to its own cluster
- Define one leaf for each taxon; place it at height 0
- While more than two clusters:
  - Determine two clusters i and j with smallest d_{ij}
  - Define a new cluster $C_k = C_i \cup C_j$
  - Define a node k with children i and j, place it at height $d_{ij}/2$
  - Replace clusters i and j with k
  - Compute distance between k and other clusters
- Join last two clusters i and j by root at height $d_{ij}/2$

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

0.0

UPGMA:

Unweighted **P**air-**G**roup **M**ethod with **A**rithmetic mean

Unweighted – all pairwise distances contribute equally.

Pair-Group – groups are combined in pairs (dichotomies only).

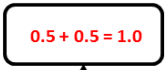
Arithmetic mean – pairwise distances to each group (clade) are mean distances to all members of that group.

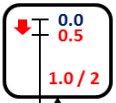
(Ultrametric – assumes molecular clock)

Dr Richard Edwards • University of Southampton • r.edwards@southampton.ac.uk

The U

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	





1. Find the shortest pairwise distance.
2. Join two sequences/groups with shortest distance.
3. Depth of new branch = ½ shortest distance.
4. Tip-to-tip path length = shortest distance.

The U

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

(19 + 18) / 2 = 18.5

(31 + 32) / 2 = 31.5

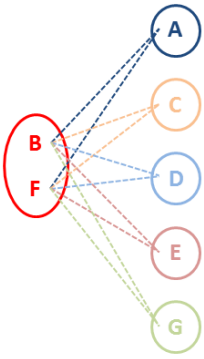
	A	BF	C	D	E	G
A						
BF	18.50					
C	27.00	31.50				
D	8.00	17.50	26.00			
E	33.00	35.50	41.00	31.00		
G	13.00	12.50	29.00	14.00	28.00	

(18 + 17) / 2 = 17.5

(36 + 35) / 2 = 35.5

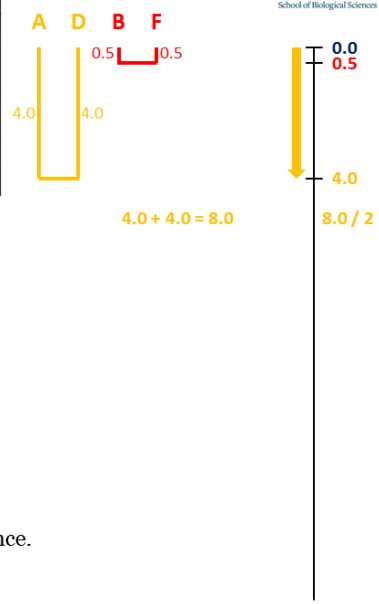
(13 + 12) / 2 = 12.5

5. Calculate mean pairwise distances with other sequences in new matrix.



The U

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	



	A	BF	C	D	E	G
A						
BF	18.50					
C	27.00	31.50				
D	8.00	17.50	26.00			
E	33.00	35.50	41.00	31.00		
G	13.00	12.50	29.00	14.00	28.00	

6. Repeat cycle with new shortest distance.

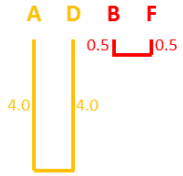
The U

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

	A	BF	C	D	E	G
A						
BF	18.50					
C	27.00	31.50				
D	8.00	17.50	26.00			
E	33.00	35.50	41.00	31.00		
G	13.00	12.50	29.00	14.00	28.00	

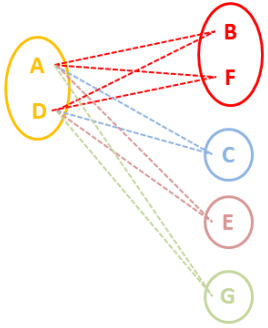
The U

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	



(19 + 18 + 18 + 17) / 4 = 18.0

	AD	BF	C	E	G
AD					
BF	18.00				
C	26.50	31.50			
E	32.00	35.50	41.00		
G	13.50	12.50	29.00	28.00	



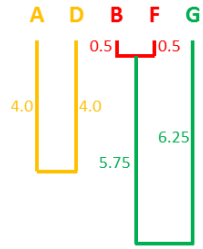
(27 + 26) / 2 = 26.5

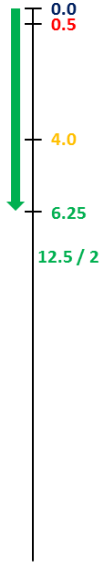
(33 + 31) / 2 = 32.0

(13 + 14) / 2 = 13.5

The U

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	





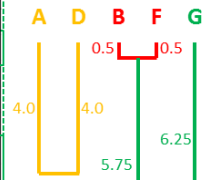
	AD	BF	C	E	G
AD					
BF	18.00				
C	26.50	31.50			
E	32.00	35.50	41.00		
G	13.50	12.50	29.00	28.00	

0.5 + 5.75 + 6.25 = 12.5

12.5 / 2

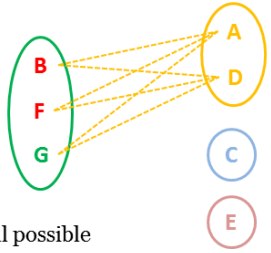
The U

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	



(19 + 18 + 13 + 18 + 17 + 14) / 6 = 16.5

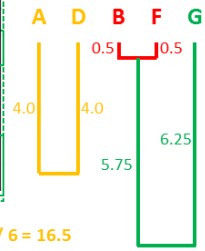
	AD	BFG	C	E
AD				
BFG	16.50			
C	26.50	30.67		
E	32.00	33.00	41.00	



New distances are mean values for all possible pairwise distances **between** groups.

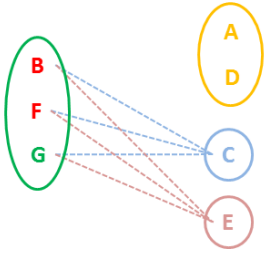
The U

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	



(19 + 18 + 13 + 18 + 17 + 14) / 6 = 16.5

	AD	BFG	C	E
AD				
BFG	16.50			
C	26.50	30.67		
E	32.00	33.00	41.00	



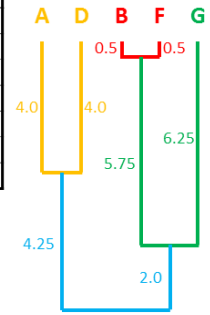
(31 + 32 + 29) / 3 = 30.67

(36 + 35 + 28) / 3 = 33.0

The U

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

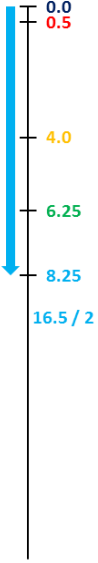
	AD	BFG	C	E
AD				
BFG	16.50			
C	26.50	30.67		
E	32.00	33.00	41.00	



0.5 + 5.75 + 2.0 = 16.5

4.0 + 4.25 +

6.25 + 2.0 = 16.5



16.5 / 2

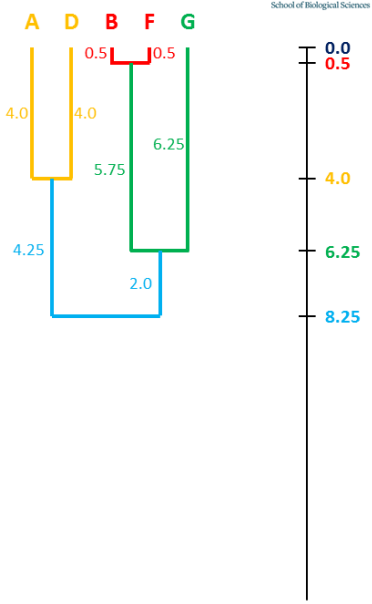
The U

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

(27 + 31 + 26 + 32 + 29) / 5 = 29.00

	ADBFG	C	E
ADBFG			
C	29.00		
E	32.60	41.00	

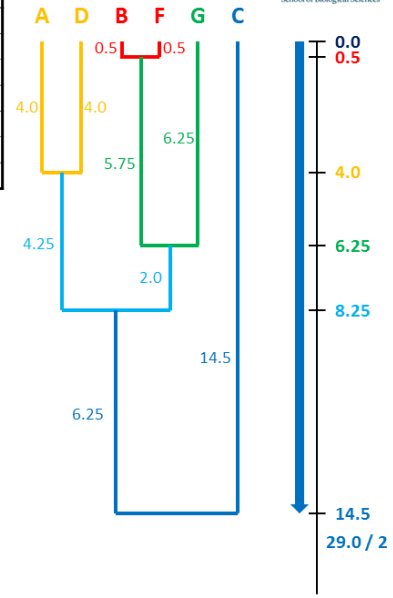
(33 + 36 + 31 + 35 + 28) / 5 = 32.60



The U

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

	ADBFG	C	E
ADBFG			
C	**29.00**		
E	32.60	41.00	



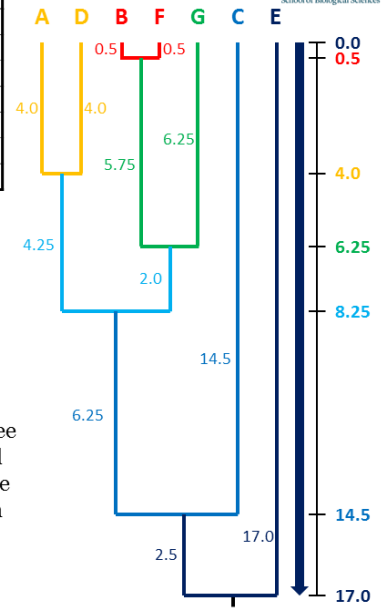
The U

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

(33 + 36 + 41 +31 + 35 + 28) / 6 = 34.00

	ADBFGC	E
ADBFGC		
E	34.00	

UPGMA assumes a molecular clock. The tree is rooted with the final joining of clades. All tip-to-tip distances via the root will have the same total distance, equal to the final mean distance.



The U

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	**1.00**	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

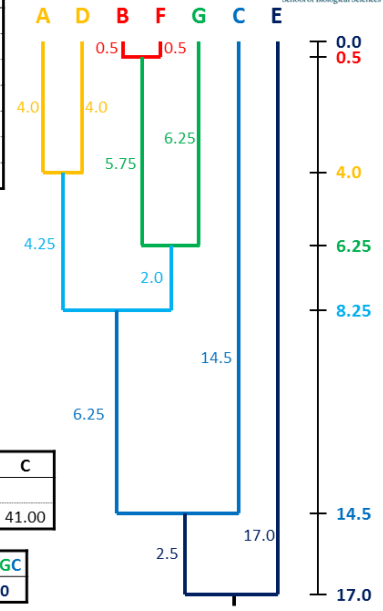
	A	BF	C	D	E
BF	18.50				
C	27.00	31.50			
D	**8.00**	17.50	26.00		
E	33.00	35.50	41.00	31.00	
G	13.00	12.50	29.00	14.00	28.00

	AD	BF	C	E
BF	18.00			
C	26.50	31.50		
E	32.00	35.50	41.00	
G	13.50	**12.50**	29.00	28.00

	AD	BFG	C
BFG	**16.50**		
C	26.50	30.67	
E	32.00	33.00	41.00

	ADBFG	C
C	**29.00**	
E	32.60	41.00

	ADBFGC
E	**34.00**



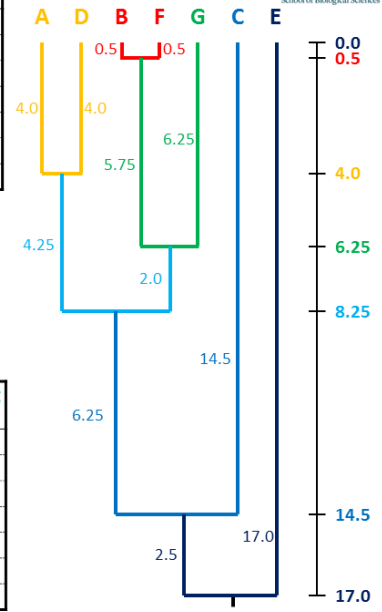
The U

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

The source data for this worked example is a selection of Cytochrome C distances from Table 3 of one of the seminal phylogenetic papers: Fitch WM & Margoliash E (1967). Construction of phylogenetic trees. *Science* **155:**279-84.

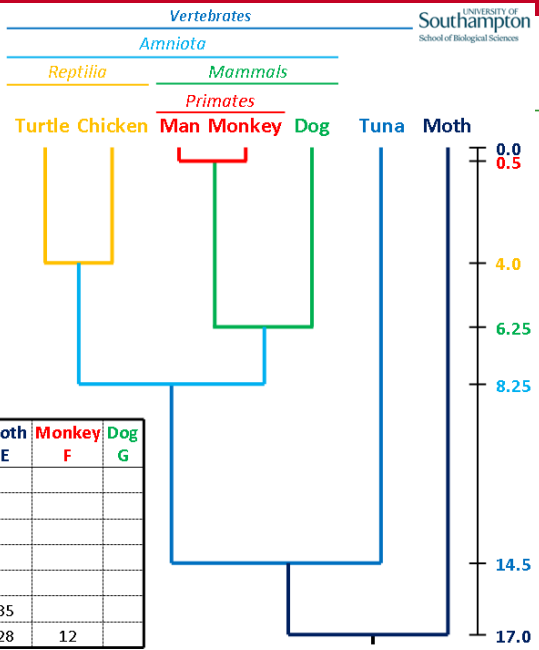
http://www.ncbi.nlm.nih.gov/pubmed/5334057

	Turtle A	Man B	Tuna C	Chicken D	Moth E	Monkey F	Dog G
Turtle							
Man	19						
Tuna	27	31					
Chicken	8	18	26				
Moth	33	36	41	31			
Monkey	18	1	32	17	35		
Dog	13	13	29	14	28	12	



The U

The UPGMA tree based on this Cytochrome C data supports the known evolutionary relationships of these organisms.



	Turtle A	Man B	Tuna C	Chicken D	Moth E	Monkey F	Dog G
Turtle							
Man	19						
Tuna	27	31					
Chicken	8	18	26				
Moth	33	36	41	31			
Monkey	18	1	32	17	35		
Dog	13	13	29	14	28	12	

Copyright statement

This slide set is designed only for teaching CSCI-478/CSCI-578/BIOL-510 Bioinformatics at the Department of Computer Science of Colorado School of Mines in Fall 2021.

Some contents in this slide set are obtained from Internet and maybe copyright sensitive. Copyright and all rights therein are retained by the respective authors or by other copyright holders. Distributing or reposting the whole or part of this slide set not for academic purpose is HIGHLY prohibited, unless the explicit permissions from all copyright holders are granted.