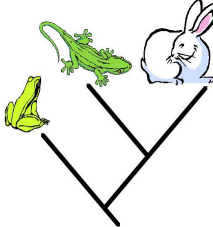
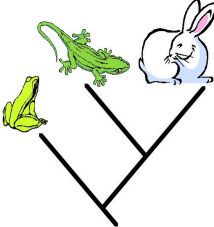



Algorithms in Bioinformatics THREE Phylogenetic Trees

Sami Khuri
Department of Computer Science
San José State University

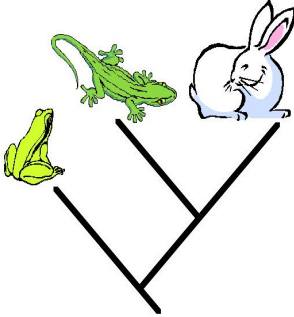
sami.khuri@sjsu.edu



©2018 Sami Khuri




Phylogenetic Trees



- ❖ Distance Methods
- ❖ Character Methods
- ❖ Molecular Clock
- ❖ UPGMA
- ❖ Maximum Parsimony
- ❖ Maximum Likelihood
- ❖ Fitch and Margoliash

©2018 Sami Khuri



Phylogeny Terminology

- **Phylogeny**- the history of descent of a group of organisms from a common ancestor

From Greek:


- **phylon** = tribe, race
- **genesis** = source

- **Taxonomy**- the science of classification of organisms

From Greek:

- **taxis** = to arrange, classify

©2018 Sami Khuri



Phylogeny: Inference Tool

- **Phylogeny** is the inference of evolutionary relationships.
- Traditionally, phylogeny relied on the comparison of morphological features between organisms.
- Today, molecular sequence data are also used for phylogenetic analyses.

©2018 Sami Khuri



Importance of Phylogeny

- How many genes are related to my favorite gene?
- Was the extinct quagga more like a zebra or a horse?
- Was Darwin correct when he stated that humans are the closest to chimps and gorillas?
- How related are whales and dolphins to cows?
- Where and when did HIV originate?
- What is the history of life on earth?

©2018 Sami Khuri



Picture of Last Quagga



Died in Amsterdam zoo in 1883.

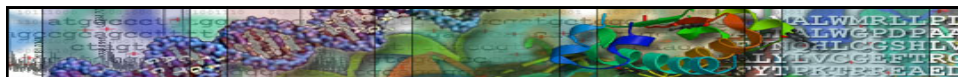
©2018 Sami Khuri



Phylogenetic Analysis

- A **phylogenetic analysis** of a family of related nucleic acid or protein sequences is a determination of how the family might have been derived during evolution.
- Two sequences that are very much alike will be located as neighboring outside branches (leaves) and will be joined by a common branch beneath them.


©2018 Sami Khuri



Phylogenetic Trees

- **Phylogenetic tree**: diagram showing evolutionary paths of species/genes.
- Why do we construct phylogenetic trees?
 - To understand the path (**lineage**) of various species.
 - To understand how various **functions** evolved.
 - To perform **multiple alignment**.


©2018 Sami Khuri



Additional Uses of Phylogenetic Trees

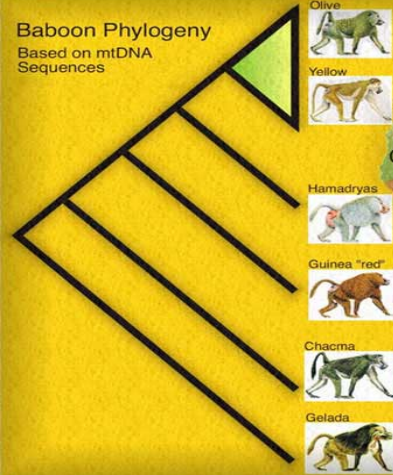
- To study the **evolutionary relationships** of different species and to understand how species relate to one another.
- To **predict** the unknown gene's function according to its phylogenetic relationship to other genes.

©2018 Sami Khuri




Baboon Phylogeny

Baboon Phylogeny
Based on mtDNA
Sequences



Olive
Yellow
Hamadryas
Guinea "red"
Chacma
Gelada

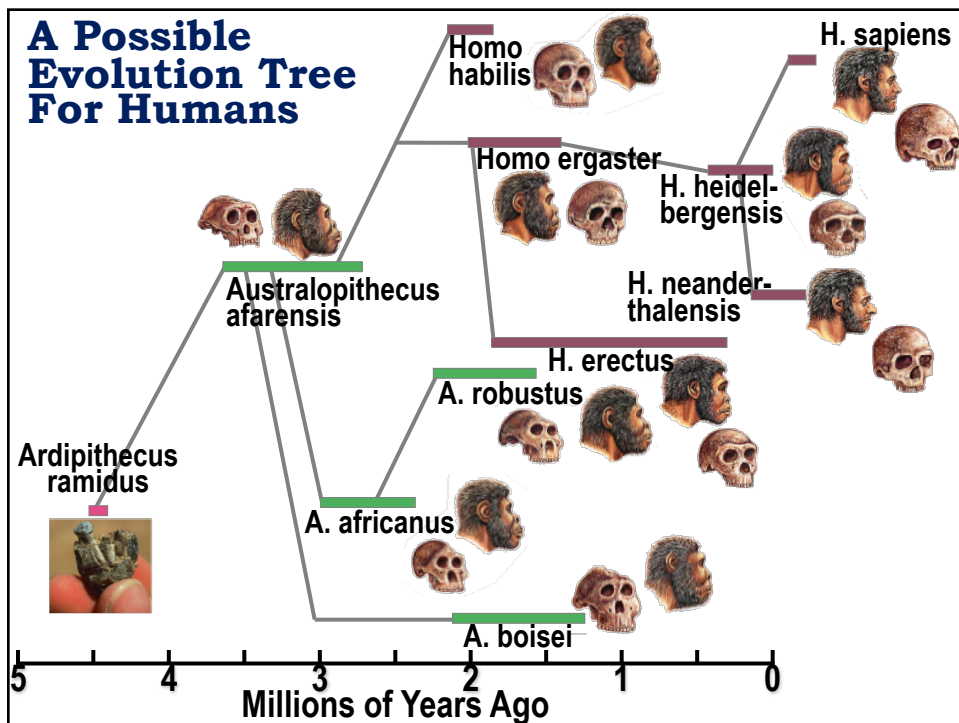
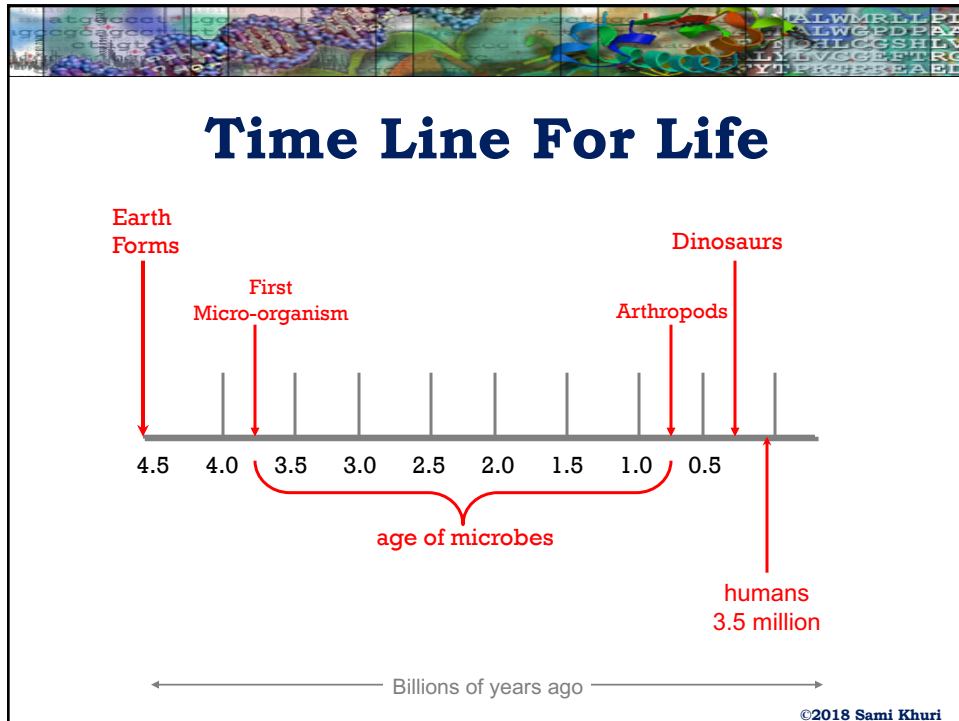
AFRICA



Guinea Hamadryas
Olive
Yellow
Chacma

Baboon Population
Distribution

©2018 Sami Khuri



Unrooted and Rooted Trees

(A) Unrooted: cannot tell

(B) Rooted: can tell

Tree construction could be based on:

- morphological features, or
- sequence data

©2018 Sami Khuri


(A) **Cladogram:**
Branch length carry no meaning

(B) **Additive Tree:**
Branch length measure evolutionary divergence

(C) **Ultrametric Tree:**
Additive tree and constant rate of mutation along branches

(D) **Additive Tree:**
with outgroup


©2018 Sami Khuri



Building of a Phylogenetic Tree

- Sequence Selection:
 - Identify a DNA or protein sequence.
 - Obtain related sequences by performing a database search.
- Perform multiple alignment.
- Build a phylogenetic tree.
- Check the robustness of the tree.

©2018 Sami Khuri




Distance and Character Based Trees

The construction of the tree is:

- **distance-based**: measures the distance between species/genes (eg. mutations, time, distance metric).
 - First calculate the overall distance between all pairs of sequences, then construct a tree based on the distances.
- **character-based**: morphological features (eg. number of legs), DNA/protein sequences.
 - Use the individual substitutions among sequences to determine the most likely ancestral relationships.

The tree is constructed based on the gain or loss of traits.


©2018 Sami Khuri



Methods for Constructing Phylogenetic Trees

- Distance-Based Methods:
 - Unweighted Pair Group Method Using Arithmetic Averages (UPGMA)
 - Fitch Margoliash (FM)
 - Neighbor Joining (NJ)
- Character-Based Methods:
 - Maximum Parsimony (MP)
 - Maximum Likelihood (ML)

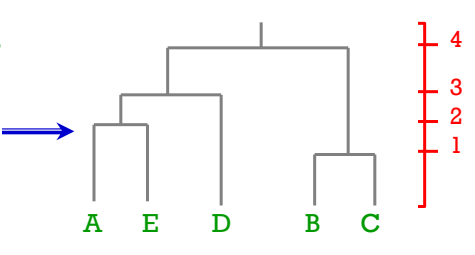
©2018 Sami Khuri



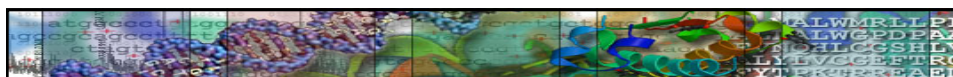
Distance-Based Method

- **Given:** an $n \times n$ matrix M , where $M(i,j)$ is the distance between objects i and j
- **Build** an edge-weighted tree such that the distances between leaves i and j correspond to $M(i,j)$

	A	B	C	D	E
A	0	8	8	5	3
B		0	3	8	8
C			0	8	8
D				0	5
E					0

→


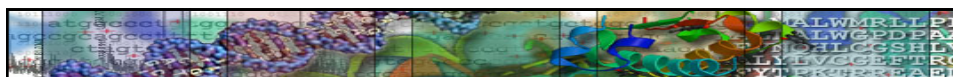
©2018 Sami Khuri



UPGMA

- UPGMA is a sequential clustering algorithm.
 - It works by clustering the sequences, at each stage amalgamating two operational taxonomic units (OTUs) and at the same time creating a new node in the tree.
 - The edge lengths are determined by the difference in the heights of the nodes at the top and bottom of an edge.

©2018 Sami Khuri




The Molecular Clock

- **UPGMA** assumes that:
 - the gene substitution rate is constant, in other words: divergence of sequences is assumed to occur at the same rate at all points in the tree.
 - Known as the **Molecular Clock**.
 - the distance is linear with evolutionary time.



©2018 Sami Khuri




UPGMA Algorithm

- The algorithm iteratively picks two clusters and merges them, thus creating a new node in the tree.
- The average **distance** between two clusters is determined by:

$$d_{ij} = \frac{1}{|C_i| + |C_j|} \sum_{p \in C_i, q \in C_j} d_{pq}, \text{ where } C_i \text{ and } C_j \text{ are clusters.}$$


©2018 Sami Khuri



The UPGMA Algorithm

- **Initialization**
 - Assign each sequence i to its own cluster C_i ,
 - Define one leaf of T for each sequence; place at height zero.
- **Iteration** while more than two clusters, do
 - Determine the two clusters C_i, C_j for which d_{ij} is minimal.
 - Define a new cluster $C_k = C_i \cup C_j$; compute d_{kl} for all l .
 - Define a node k with children i and j ; place it at height $d_{ij}/2$.
 - Replace clusters C_i and C_j with C_k .
- **Termination**
 - Join last two clusters, C_i and C_j ; place the root at height $d_{ij}/2$.

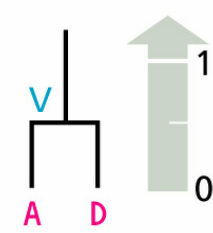
©2018 Sami Khuri




UPGMA: Example (1st Iteration)

Sequences A and D are the closest and are combined to create a new cluster V of height $\frac{1}{2}$ in T.

d_{ij}	A	B	C	D	E	F
A	-	6	8	1	2	6
B		-	8	6	6	4
C			-	8	8	8
D				-	2	6
E					-	6



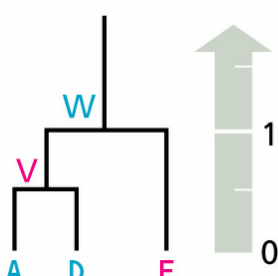
Understanding Bioinformatics by M. Zvelebil and J. Baum
©2018 Sami Khuri



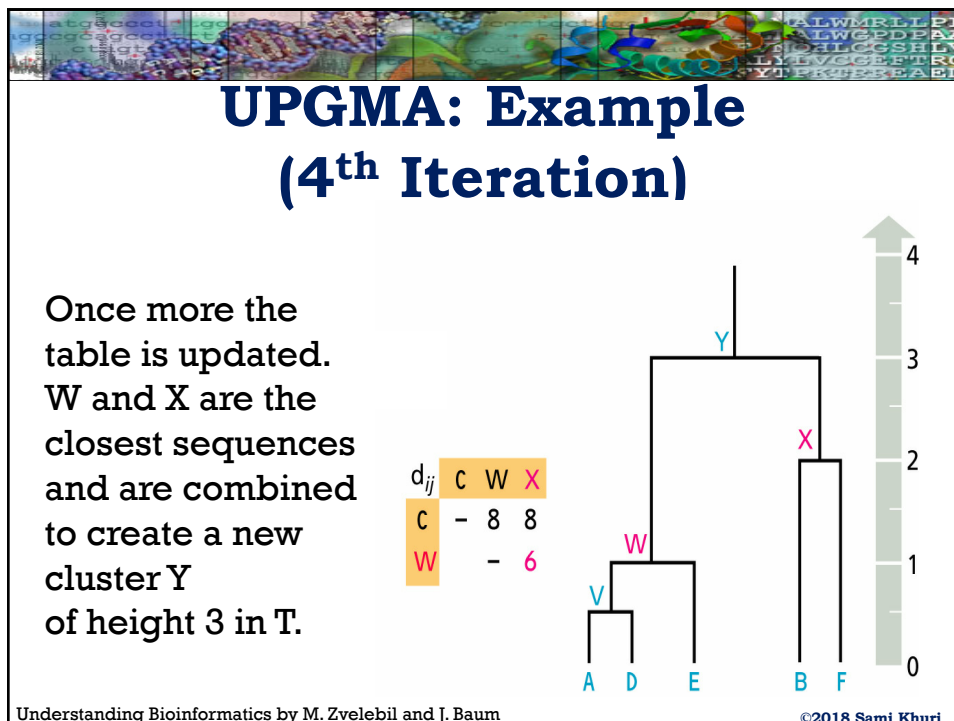
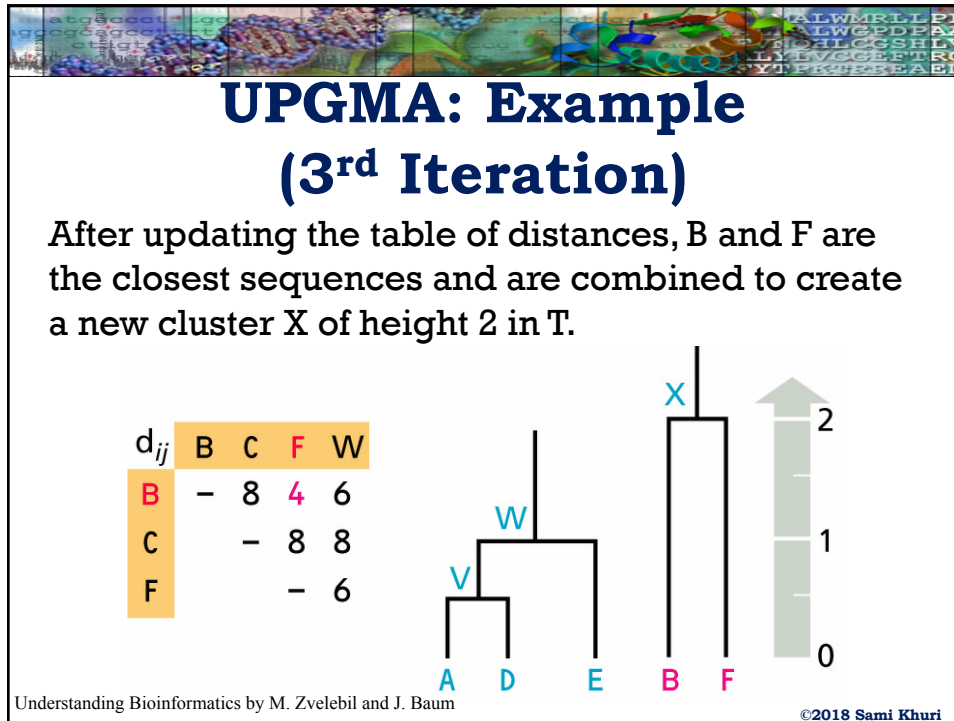
UPGMA: Example (2nd Iteration)

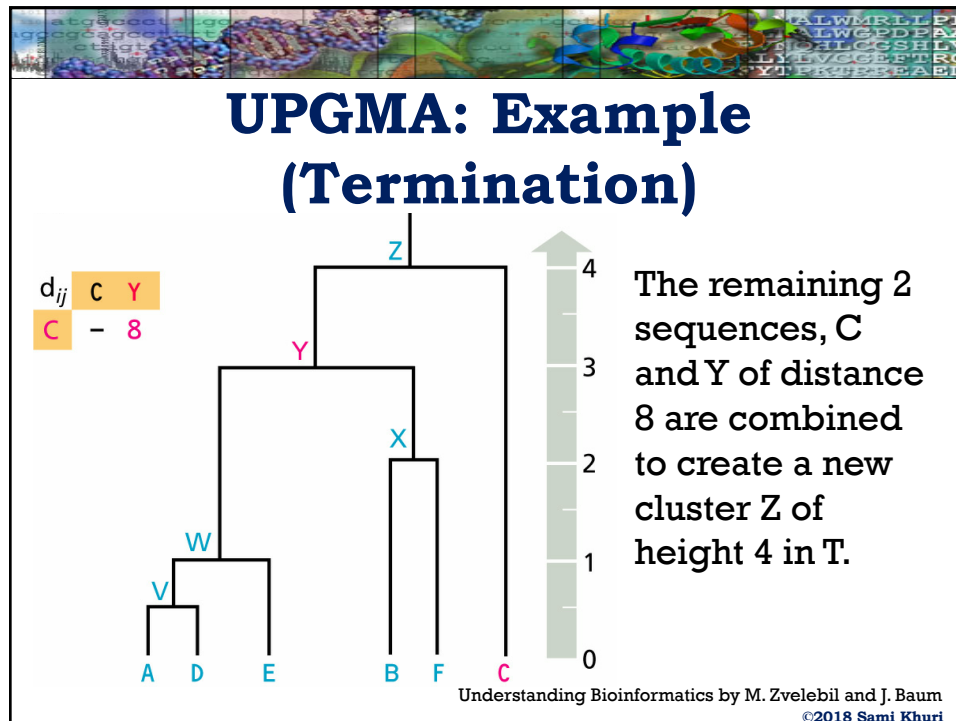
The table of distances is updated to reflect the average distances from V to the other sequences. V and E are the closest and are combined to create a new cluster W of height 1 in T.

d_{ij}	B	C	E	F	V
B	-	8	6	4	6
C		-	8	8	8
E			-	6	2
F				-	6



Understanding Bioinformatics by M. Zvelebil and J. Baum
©2018 Sami Khuri





Limitations of Distance-Based Phylogenetic Trees

The **distance-based phylogenetic tree** is derived from the pairwise distance of aligned sequences and not from the original sequence data. The distance information may not contain all the sequence information.

©2018 Sami Khuri



Observable Features

- Sometimes we do not have a distance metric between the species we are interested in.
- What we have instead, are **observable features**.
- We then use the **observable features** to build the tree. These trees are called **Character-Based trees**.

©2018 Sami Khuri



Character-Based Trees

- The building of the tree is based on **morphological features** and not on distances.
- Examples of **morphological features**:
 - has feathers
 - has a backbone
 - has a certain amino acid at a certain position in the sequence
 - whether or not a certain protein regulates another protein.

©2018 Sami Khuri