Phylogenetic Trees

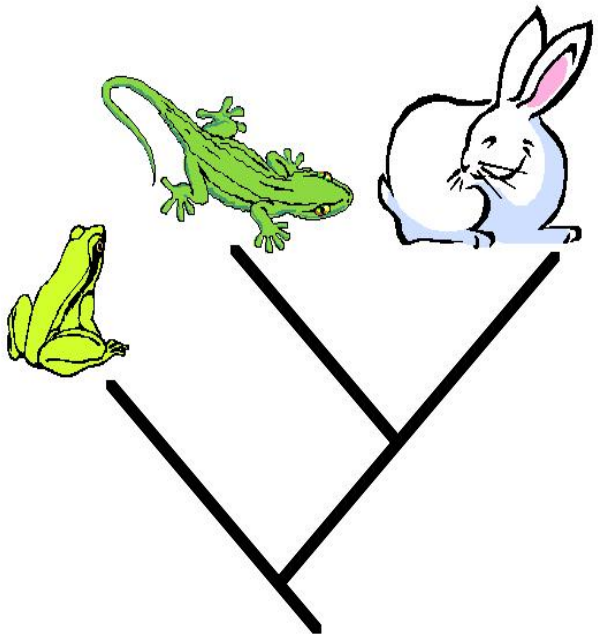
Eight

Wendy Lee
Department of Computer Science
San José State University

Biology/CS/SE 123A

Fall 2014

Phylogenetic Trees



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

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

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.

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?

Picture of Last Quagga

Died in Amsterdam zoo in 1883.

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.

Aim of Phylogenetic Analysis

- The evolutionary relationships among the sequences are depicted by placing the sequences as outer branches on a tree.
- The branching relationships on the inner part of the tree then reflect the degree to which different sequences are related.
- The aim of phylogenetic analysis is to discover all of the branching relationships in the tree and the branch lengths.

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.

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.

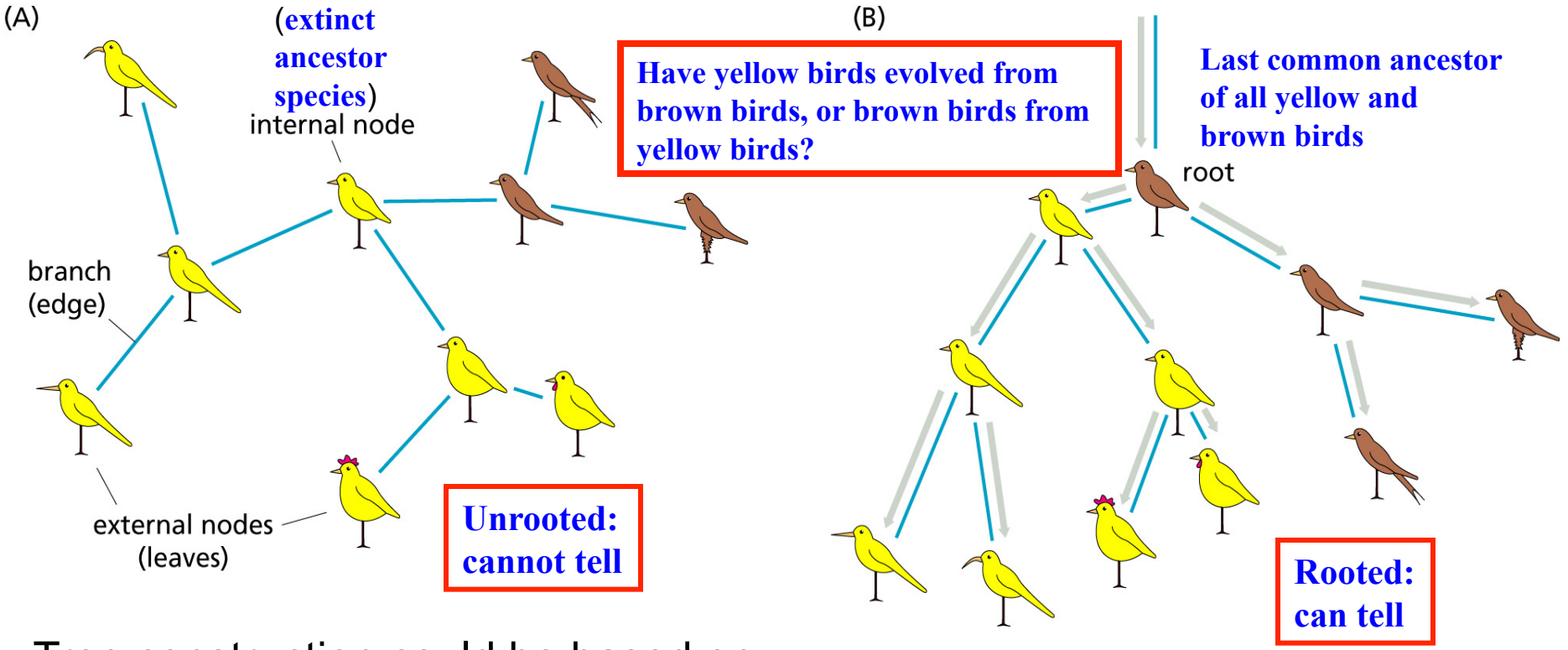
More Terminology

- Leaves represent objects (genes, species) being compared
  - Taxon refers to the leaves when they represent species and broader classifications of organisms.
- Internal nodes are hypothetical ancestral units
- In a rooted tree, the path from root to a node represents an evolutionary path.
- An unrooted tree specifies relationships among objects, but not evolutionary paths.

Rooted and Unrooted Trees

- All objects in a rooted tree have a single common ancestor.
  - In general, rooted trees require more information to construct than unrooted ones.
- Objects are leaves in an unrooted tree and internal nodes are common ancestors.
  - In general, given any two leaves, we cannot tell if they have a common ancestor.

Unrooted and Rooted Trees

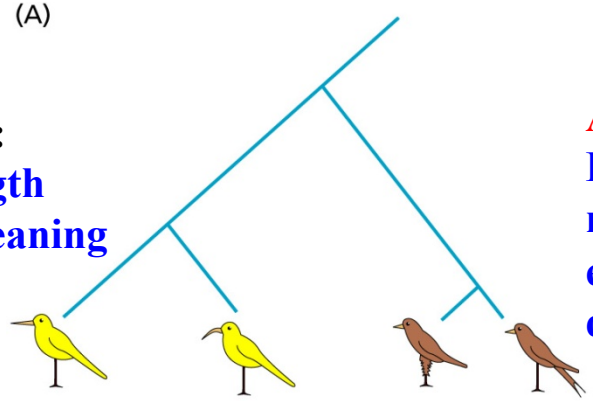


Tree construction could be based on:

- morphological features, or
- sequence data

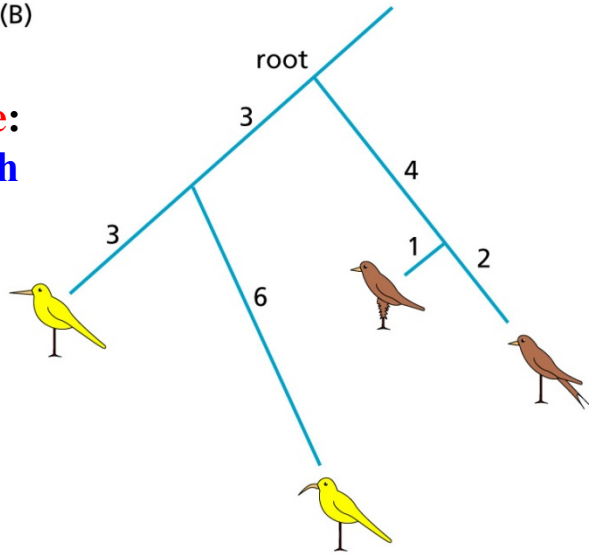
Cladogram:
Branch length carry no meaning

(A)



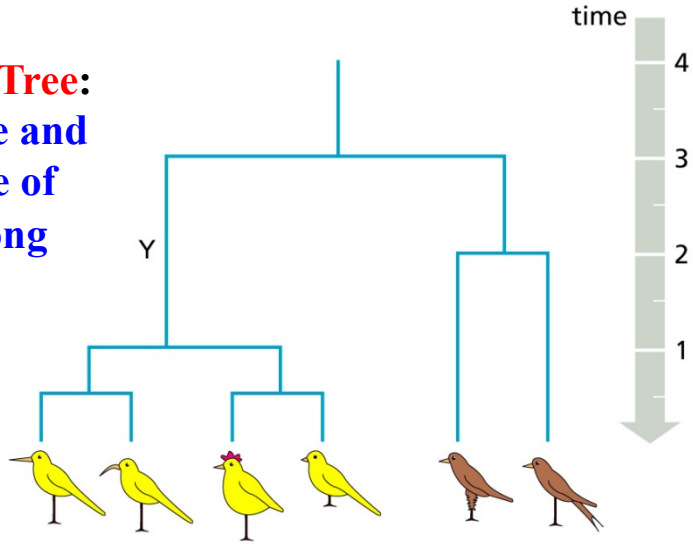
Additive Tree:
Branch length measure evolutionary divergence

(B)



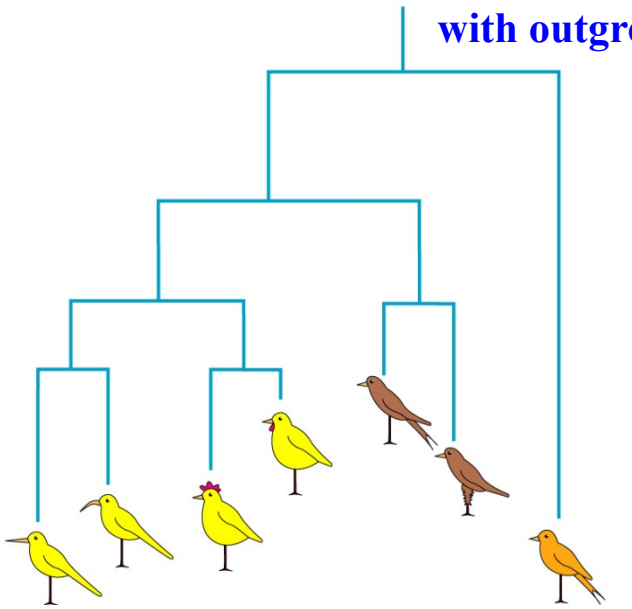
Ultrametric Tree:
Additive tree and constant rate of mutation along branches

(C)

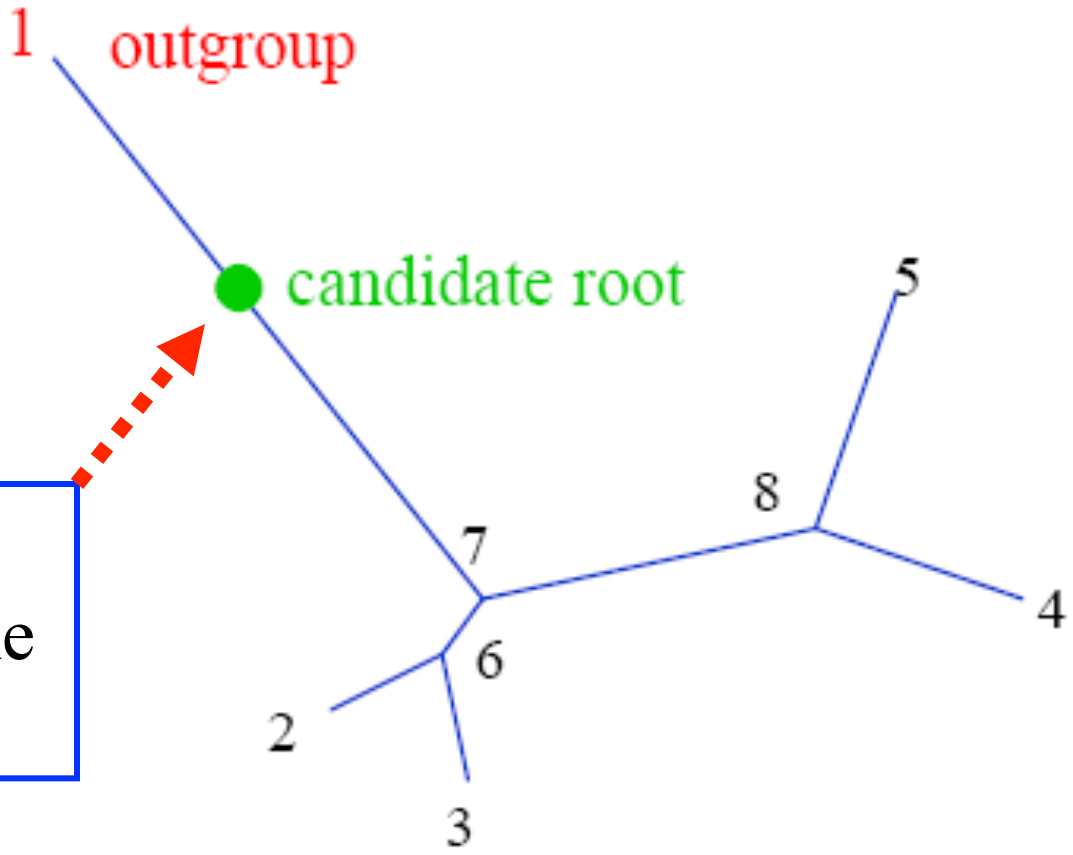


(D)

Additive Tree:
with outgroup



Rooting a Tree



The best place for having the root of the phylogenetic tree

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.

Distance and Character Based Trees

The construction of the tree is:

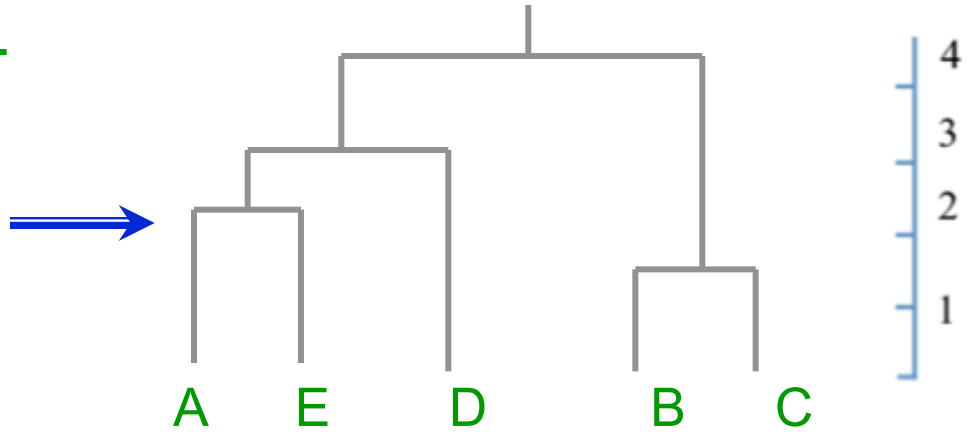
- distance-based: measures the distance between species/ genes (eg. mutations, time, distance metric).
  - First calculate the overall distance between all pairs of sequences, then construct a tree based on the distances.
- character-based: morphological features (eg. number of legs), DNA/protein sequences.
  - 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.

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



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.

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.

Rates of Evolutionary Change

- Different rates throughout genomic DNA base-pair sequence, based mainly on coding.
- ORFs: codon position 3 changes faster than positions 1 and 2.
- Introns change faster than exons.
- Intergenic DNA (especially repeats) changes faster than intragenic (ORF) DNA.
- DNA overall: transition mutations more frequent than transversion mutations.

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:

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

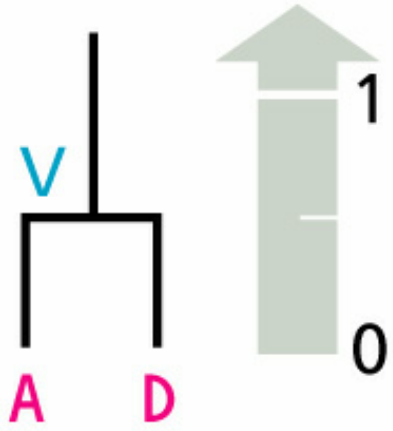
The UPGMA Algorithm

- **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 $d_{ij}/2$.

UPGMA: Example (1st Iteration)

Sequences A and D are the closest and are combined to create a new cluster V of height ½ 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

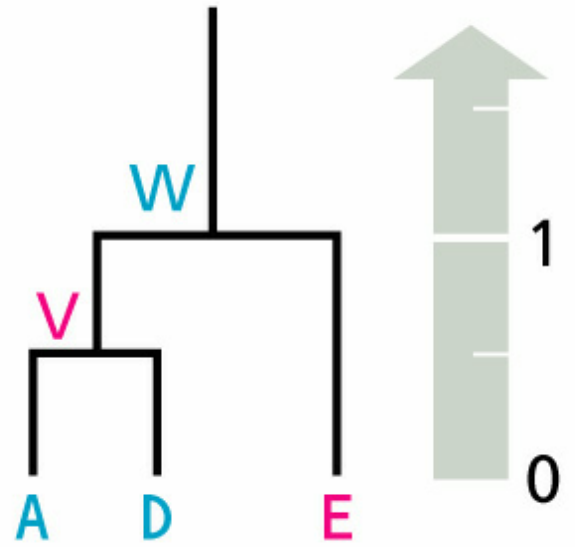


Understanding Bioinformatics by M. Zvelebil and J. Baum

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

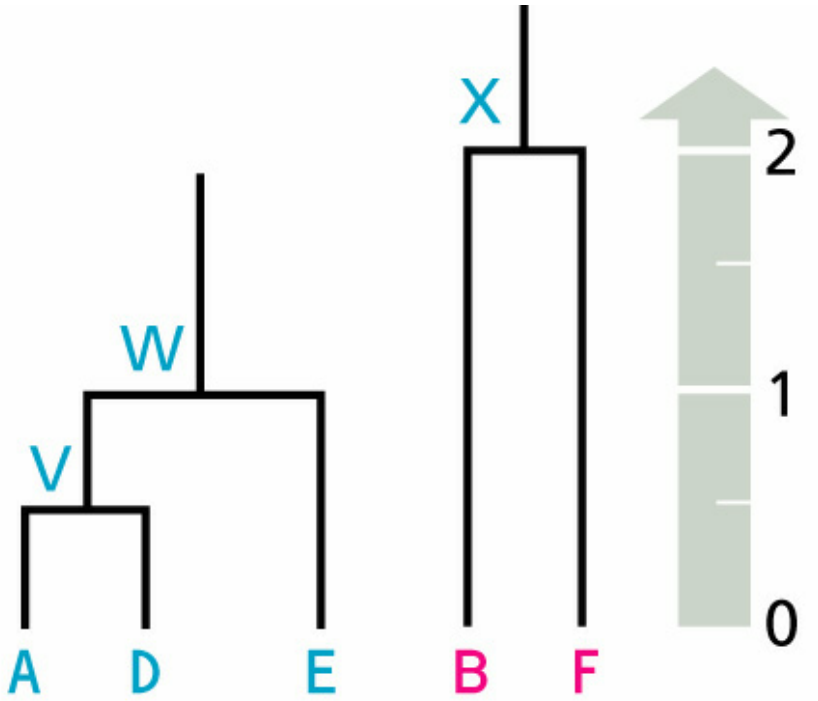


Understanding Bioinformatics by M. Zvelebil and J. Baum

UPGMA: Example (3rd Iteration)

After updating the table of distances, B and F are the closest sequences and are combined to create a new cluster X of height 2 in T.

d_{ij}	B	C	F	W
B	–	8	4	6
C		–	8	8
F			–	6

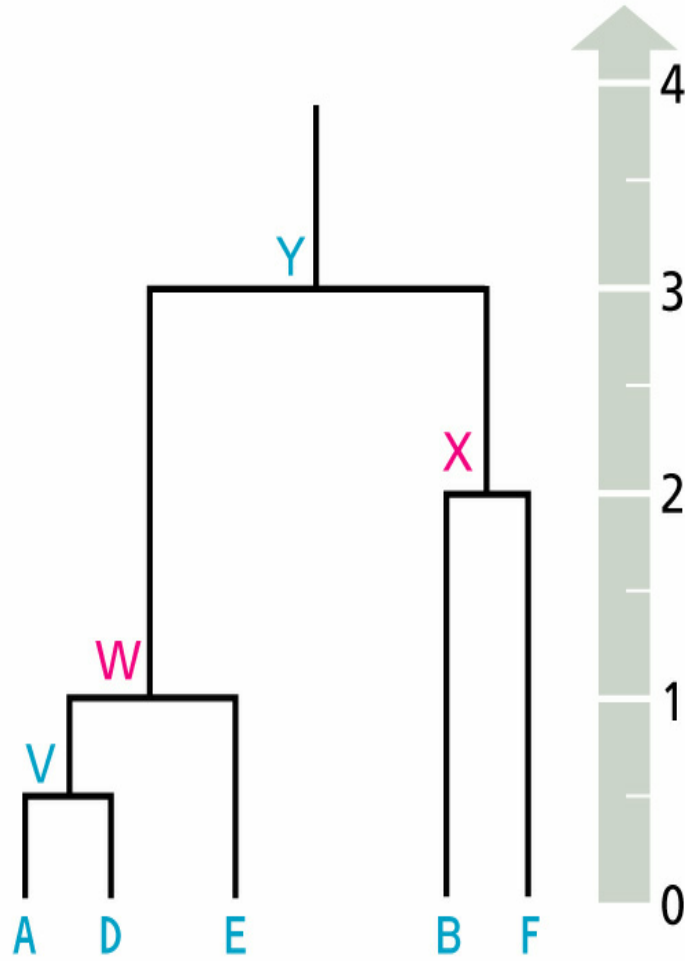


Understanding Bioinformatics by M. Zvelebil and J. Baum

UPGMA: Example (4th Iteration)

Once more the table is updated. W and X are the closest sequences and are combined to create a new cluster Y of height 3 in T.

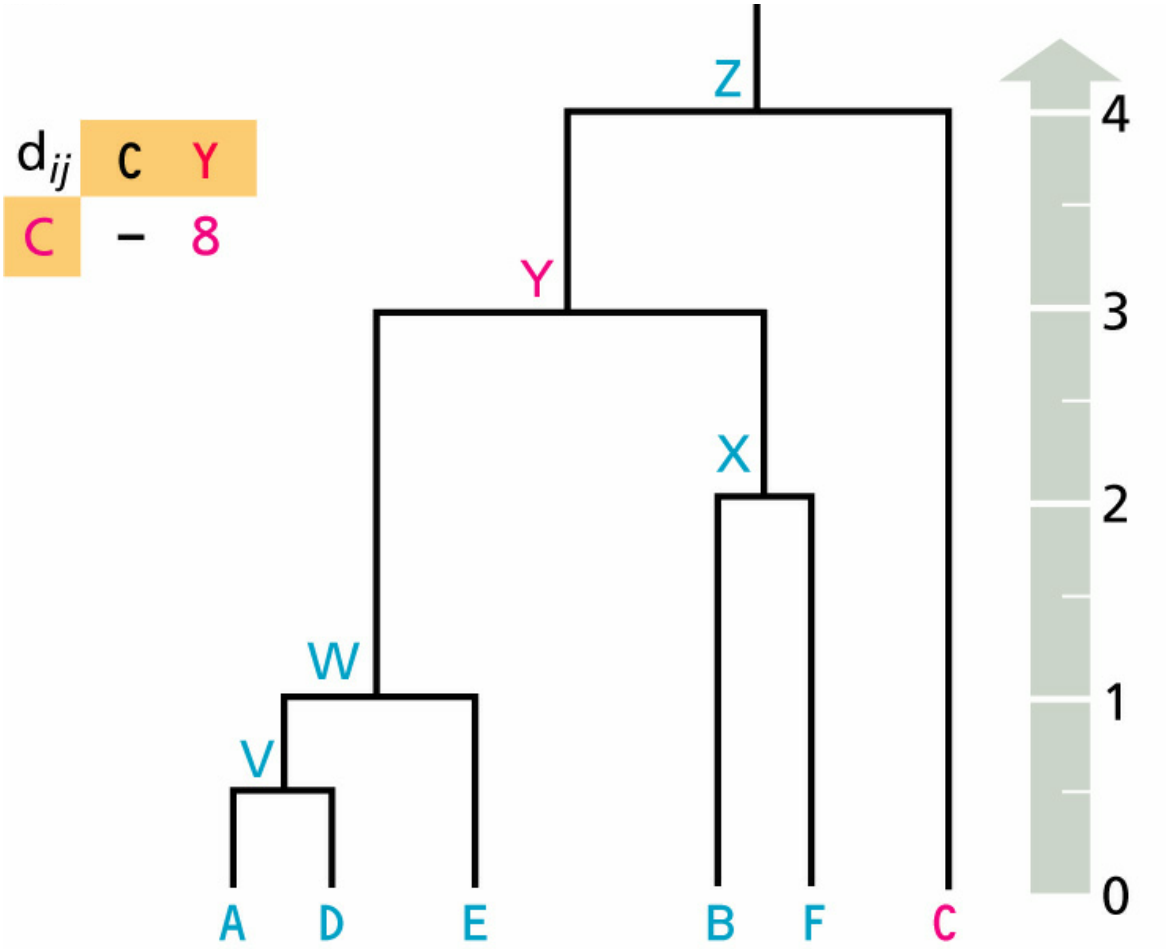
d_{ij}	C	W	X
C	–	8	8
W		–	6



Understanding Bioinformatics by M. Zvelebil and J. Baum

UPGMA: Example (Termination)

d_{ij}	C	Y
C	–	8

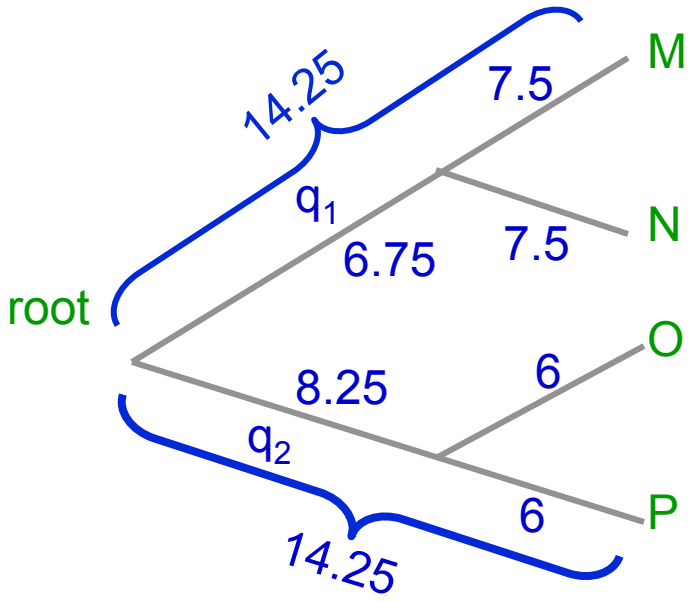


The remaining 2 sequences, C and Y of distance 8 are combined to create a new cluster Z of height 4 in T.

Understanding Bioinformatics by M. Zvelebil and J. Baum

UPGMA Tree: Second Example

	M	N	O	P
M	-	15	26	28
N	-	-	29	31
O	-	-	-	12
P	-	-	-	-



UPGMA assumes a uniform rate of mutation in the tree branches. At any given time, the two sequences should have the same number of changes separating them from the common ancestor.

Bioinformatics by David Mount