Bioinformatics: Phylogeny Construction

Swakkhar Shatabda

Department of Computer Science and Engineering
BRAC University

Book Reference

Bioinformatics Algorithms, An Active Learning Approach , Vol 2, Chapter 7

BRAC UNIVERSITY
Inspiring Excellence

SARS CoV-1

- On February 21, 2003, a Chinese doctor named Liu Jianlun flew to Hong Kong to attend a wedding and checked into Room 911 of the Metropole Hotel. The next day, he became too ill to attend the wedding and was admitted to a hospital. Two weeks later, Dr. Jianlun was dead.
- On his deathbed, Jianlun told doctors that he had recently treated sick patients in Guangdong Province, China
- On February 23, a man who had stayed across the hall from Dr. Jianlun at the Metropole traveled to Hanoi and died after infecting 80 people.
- On February 26, a woman checked out of the Metropole, traveled back to Toronto, and died after initiating an outbreak there.

BRAC UNIVERSITY
Inspiring Excellence

SARS CoV-1

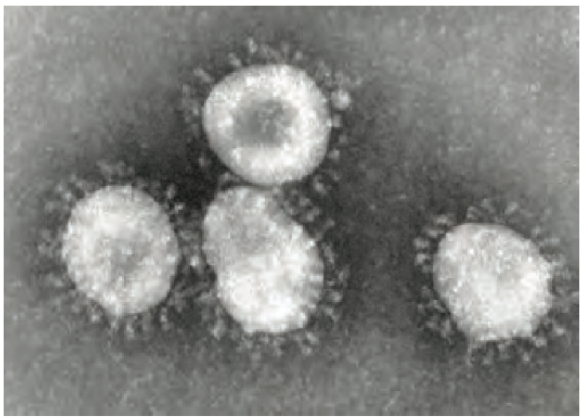
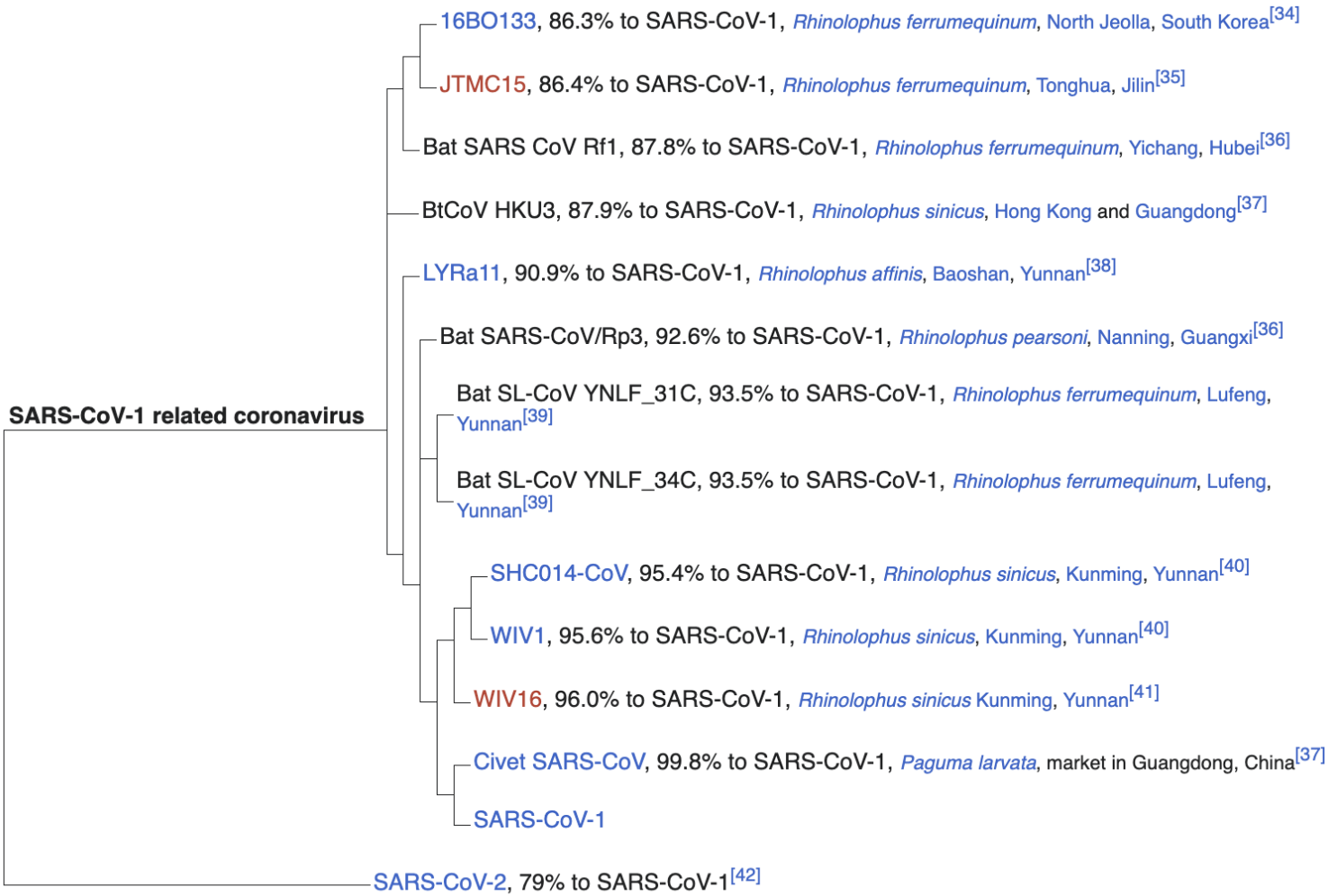


FIGURE 7.1 (Left) Coronavirus particles. (Right) A solar eclipse with the sun's corona visible.

- Coronaviruses, influenza viruses, and HIV are all RNA viruses, meaning that they possess RNA instead of DNA.
- RNA replication has a higher error rate than DNA replication, and so RNA viruses are capable of mutating more quickly into divergent strains.
- The rapid mutation of RNA viruses explains why the flu shot changes from year to year and why there are many different subtypes of HIV.

BRAC UNIVERSITY
Inspiring Excellence

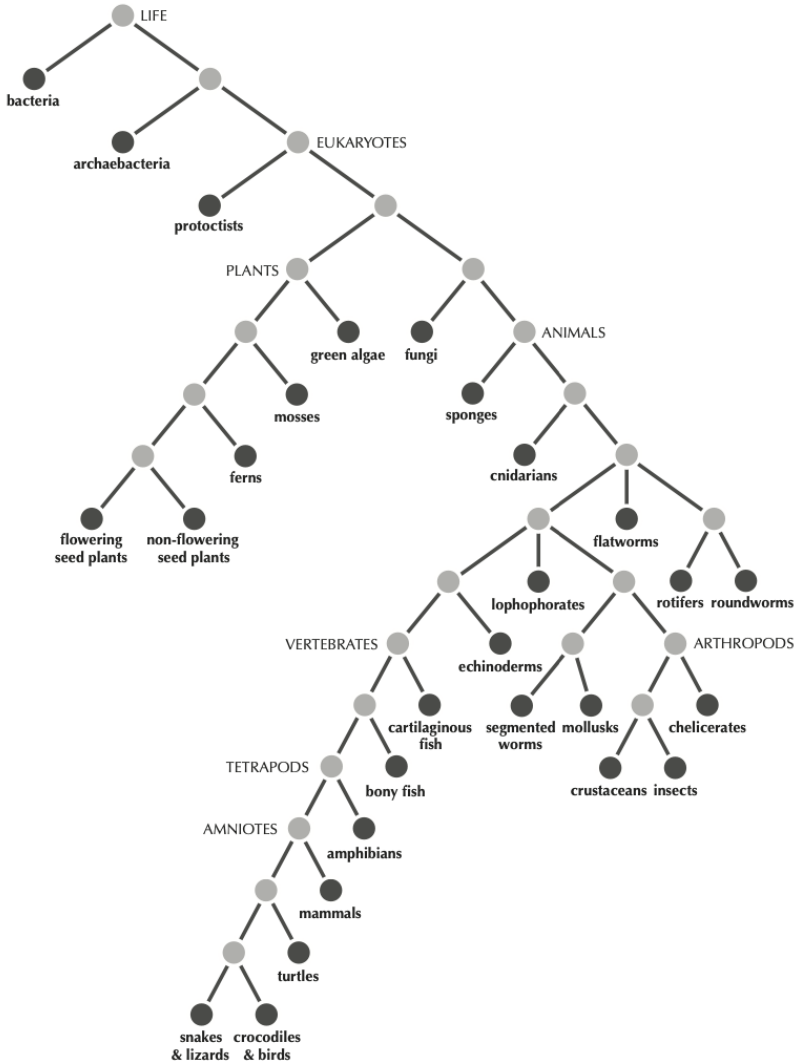
Origin of SARS?



BRAC UNIVERSITY
Inspiring Excellence

Tree of Life

- Present-day species have been assigned to the leaves of the tree, or nodes having degree 1.
- Nodes with degree larger than 1 are called internal nodes and represent unknown ancestor species.
- Given a leaf j, there is only one node connected to j by an edge, which we call the **parent** of j.
- An edge connecting a leaf to its parent is called a **limb**.



Rooted Trees

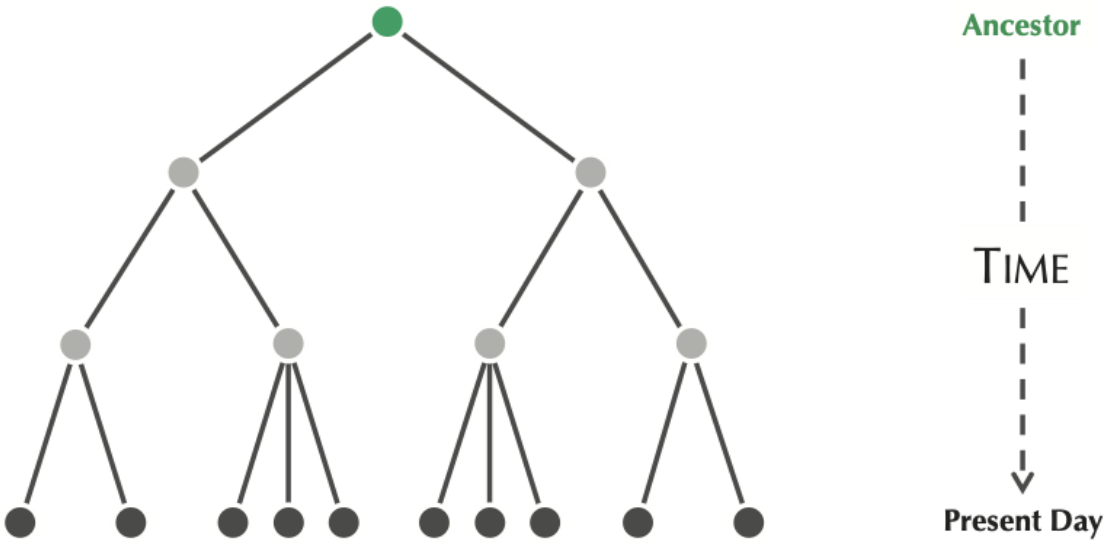


FIGURE 7.5 A rooted tree, with the root (representing an ancestor of all species in the tree) indicated in green at the top of the tree. The presence of the root implies an orientation of edges in the tree away from the root.

- In a rooted tree, the edges in the tree automatically inherit an implicit orientation away from the root, which is placed at the top or left of the tree.
- This edge orientation models time: the ancestor of all species in the tree is found at the root, and evolution proceeds from the root outward through the tree.

Unrooted Trees

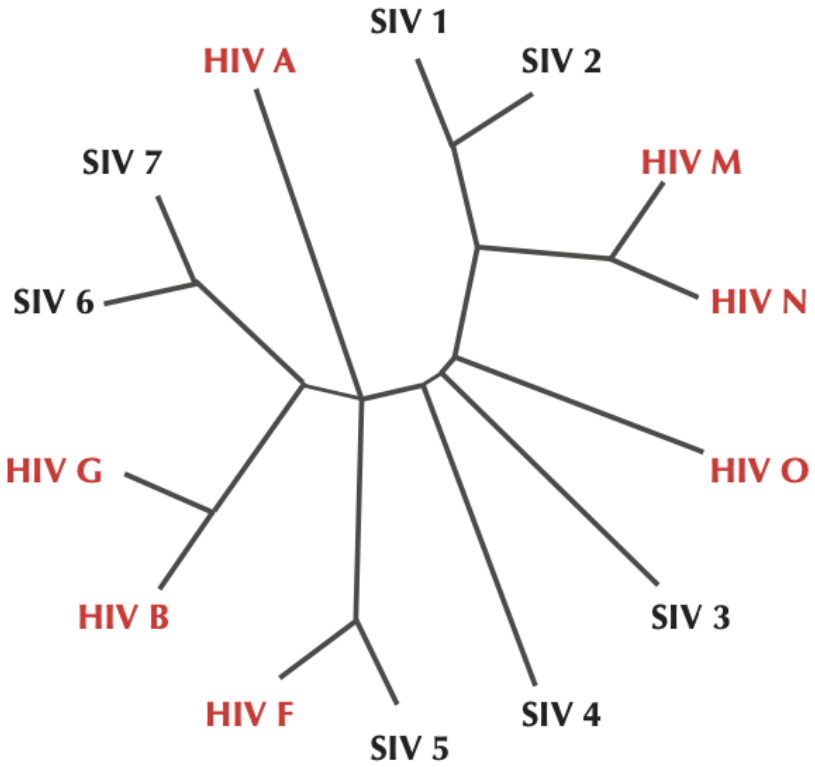


FIGURE 7.6 An unrooted tree of HIV and SIV viruses that suggests additional viral families F and G in addition to the viral families A, B, M, N, and O shown in Figure 7.2.

- Trees without a designated root are called unrooted.

BRAC UNIVERSITY
Inspiring Excellence

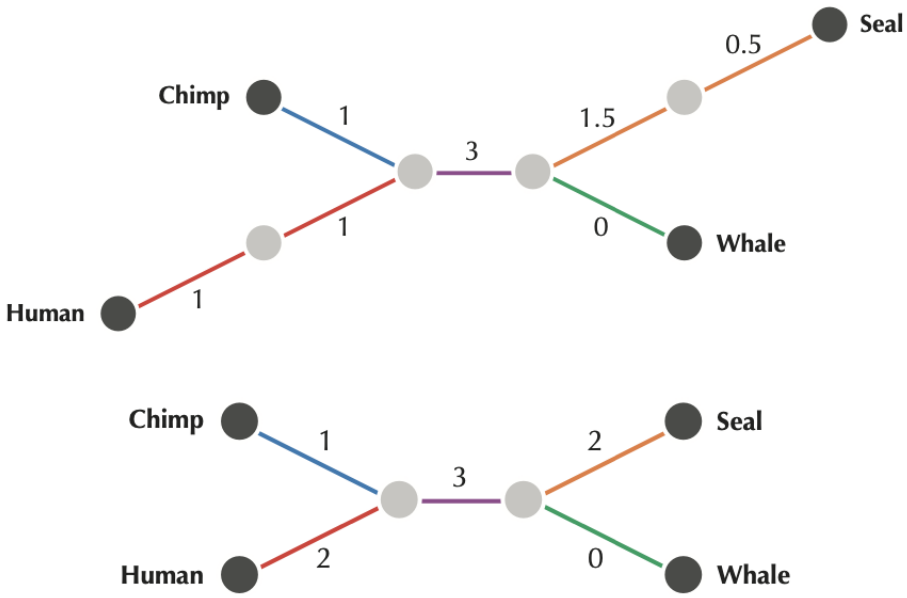
Distance Based Phylogeny

Species	Alignment	Distance Matrix			
		Chimp	**Human**	**Seal**	**Whale**
Chimp	ACGTAGGCCT	0	3	6	4
Human	ATGTAAGACT	3	0	7	5
Seal	TCGAGAGCAC	6	7	0	2
Whale	TCGAAAGCAT	4	5	2	0

Figure 7.3 A multiple alignment of hypothetical DNA sequences from four species, along with the distance matrix produced by counting the number of differing symbols between each pair of rows in this multiple alignment.

- Regardless of which distance function we use, in order to be a distance matrix, D must satisfy three properties.
- It must be symmetric (for all i and j, $D_{i,j} = D_{j,i}$), non- negative (for all i and j, $D_{i,j} \geq 0$) and satisfy the triangle inequality (for all i, j, and k, $D_{i,j} + D_{j,k} \leq D_{i,k}$).

Distance Based Phylogeny



- Not every distance matrix has a tree fitting it
- A distance matrix additive if there exists a tree that fits this matrix and non-additive otherwise.

BRAC UNIVERSITY
Inspiring Excellence

The first algorithm?

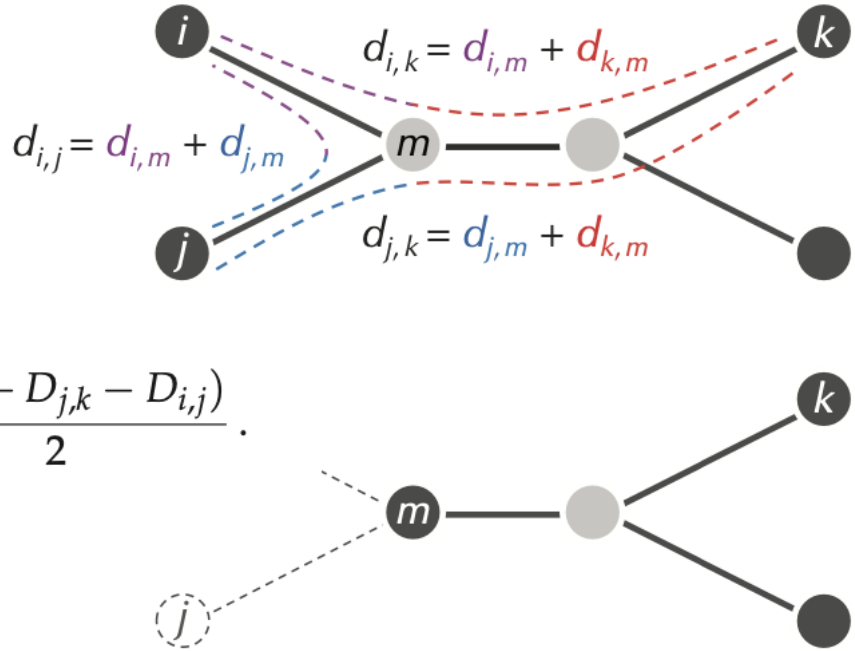


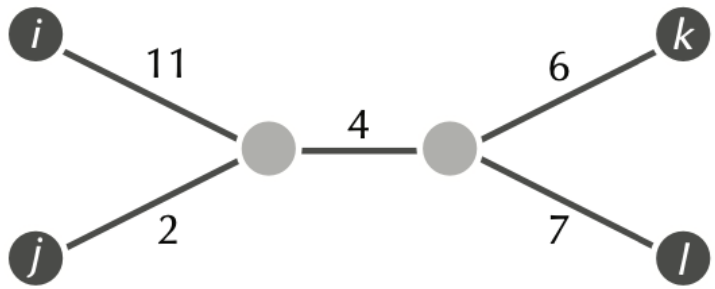
FIGURE 7.8 For neighboring leaves i and j and their parent node m, $d_{k,m} = (d_{i,k} + d_{j,k} - d_{i,j})/2$ for every other leaf k in the tree. (Bottom) Removing leaves i and j from the tree turns m into a leaf (we assume that m has degree 3). The distances from this new leaf to any other leaf k can be recomputed as $d_{k,m} = (D_{i,k} + D_{j,k} - D_{i,j})/2$.

BRAC UNIVERSITY
Inspiring Excellence

A first algorithm?

1. Find a pair of neighboring leaves i and j by selecting the minimum element $D_{i,j}$ in the distance matrix
2. Replace i and j with their parent, and recompute the distances from this parent to all other leaves as described in the last slide
3. Solve the Distance-Based Phylogeny problem for the smaller tree.
4. Add the previously removed leaves i and j back to the tree.

	i	*j*	*k*	*l*
i	0	13	21	22
j	13	0	12	13
k	21	12	0	13
l	22	13	13	0



BRAC UNIVERSITY
Inspiring Excellence

Additive Phylogeny - Idea

- Rather than looking for a pair of neighbors in TREE(D), we will instead reduce the size of the tree by trimming its leaves one at a time.
- Given a leaf j in a tree, we denote the length of the limb connecting j with its parent as LIMBLENGTH(j).

Limb Length Theorem

Given an additive matrix D and a leaf j, LIMBLENGTH(j) is equal to the minimum value of $(D_{i,j} + D_{j,k} - D_{i,k})/2$ over all leaves i and k.

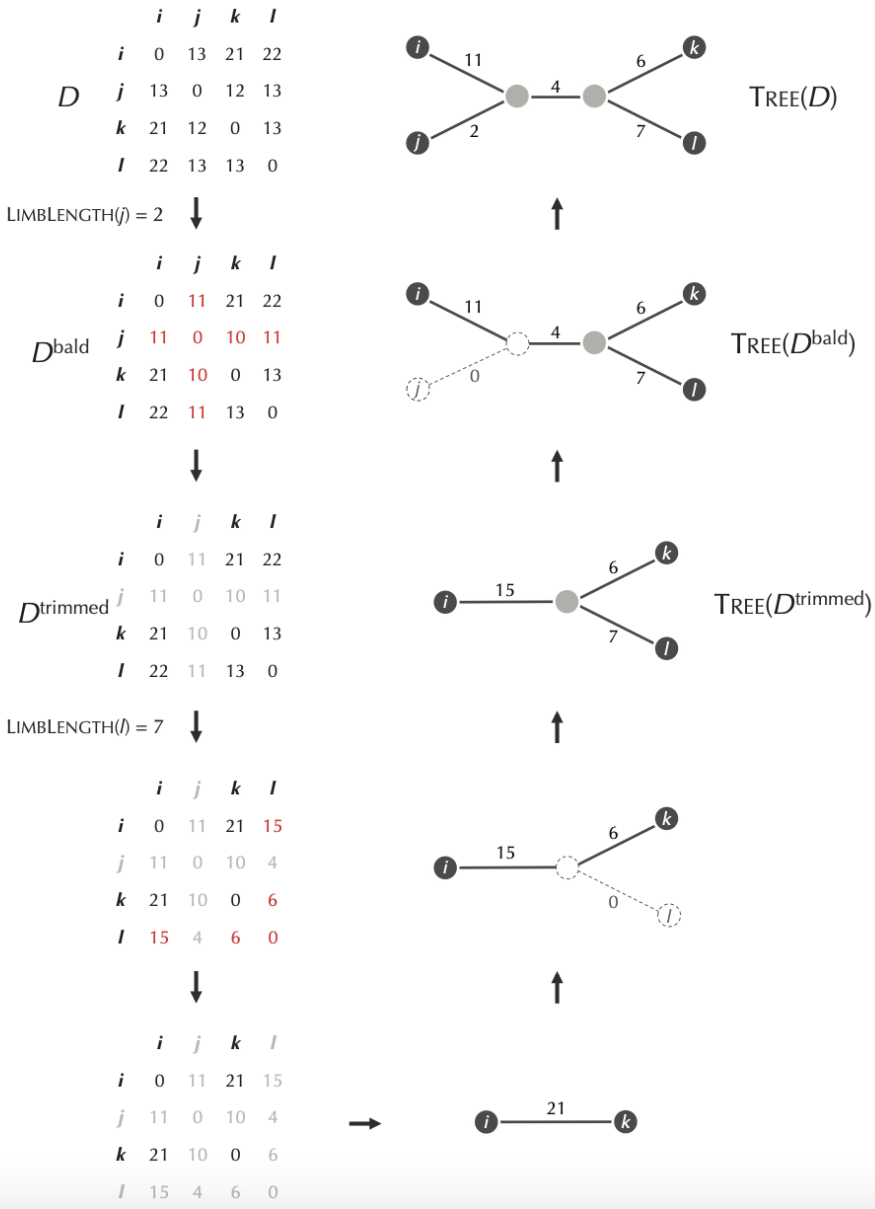
BRAC UNIVERSITY
Inspiring Excellence

Additive Phylogeny - Algorithm

1. Pick an arbitrary leaf j, compute LIMBLENGTH(j), and construct the distance matrix $D_{trimmed}$
2. Solve the Distance-Based Phylogeny Problem for $D_{trimmed}$
3. Identify the point in TREE($D_{trimmed}$) where leaf j should be attached in TREE(D)
4. Add a limb of length LIMBLENGTH(j) growing from this attachment point in TREE($D_{trimmed}$) to form TREE(D).

BRAC UNIVERSITY
Inspiring Excellence

Additive Phylogeny - Example



Additive Phylogeny - Pseudocode

```
ADDITIVEPHYLOGENY(D, n)
    if n = 2
        return the tree consisting of a single edge of length D_{1,2}
    limbLength ← LIMB(D, n)
    for j ← 1 to n − 1
        D_{j,n} ← D_{j,n} − limbLength
        D_{n,j} ← D_{j,n}
    (i, n, k) ← three leaves such that D_{i,k} = D_{i,n} + D_{n,k}
    x ← D_{i,n}
    remove row n and column n from D
    T ← ADDITIVEPHYLOGENY(D, n − 1)
    v ← the (potentially new) node in T at distance x from i on the path between i and k
    add leaf n back to T by creating a limb (v, n) of length limbLength
    return T
```

Ultrametric Evolutionary Trees

- Biologists often assume that every internal node in an evolutionary tree corresponds to a species that underwent a **speciation event**, splitting one ancestral species into two descendants.
- If we had a **molecular clock** measuring evolutionary time, then we could assign an age to every node v in a rooted binary tree (denoted AGE(v)), where all of the leaves of the tree have age 0 because they correspond to present-day species.
- A tree, in which the distance from the root to any leaf is the same, is called **ultrametric**
- **UPGMA** (which stands for Unweighted Pair Group Method with Arithmetic Mean) is a simple clustering heuristic that introduces a hypothetical molecular clock for constructing an ultrametric evolutionary tree.
- Given an $n \times n$ matrix D, UPGMA first forms n trivial clusters, each containing a single leaf. The algorithm then finds a pair of "closest" clusters.
- To clarify the notion of closest clusters, UPGMA defines the distance between clusters C1 and C2 as the average pairwise distance between elements of C1 and C2,

$$D_{C_1,C_2} = \frac{\sum_{i \in C_1} \sum_{j \in C_2} D_{i,j}}{|C_1| \cdot |C_2|}$$

BRAC UNIVERSITY
Inspiring Excellence

UPGMA Algorithm

UPGMA(D, n)

$Clusters \leftarrow n$ single-element clusters labeled $1, \ldots, n$

construct a graph T with n isolated nodes labeled by single elements $1, \ldots, n$

for every node v in T

$\text{AGE}(v) \leftarrow 0$

while there is more than one cluster

find the two closest clusters C_i and C_j (break ties arbitrarily)

merge C_i and C_j into a new cluster C_{new} with $|C_i| + |C_j|$ elements

add a new node labeled by cluster C_{new} to T

connect node C_{new} to C_i and C_j by directed edges

$\text{AGE}(C) \leftarrow D_{C_i,C_j}/2$

remove the rows and columns of D corresponding to C_i and C_j

remove C_i and C_j from *Clusters*

add a row/column to D for C_{new} by computing $D(C_{new}, C)$ for each C in *Clusters*

add C_{new} to *Clusters*

root $\leftarrow$ the node in T corresponding to the remaining cluster

for each edge (v, w) in T

length of $(v, w) \leftarrow \text{AGE}(v) - \text{AGE}(w)$

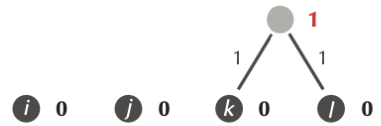
return T

BRAC UNIVERSITY
Inspiring Excellence

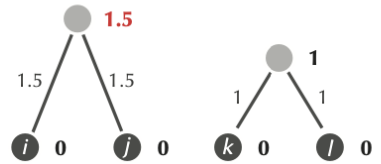
UPGMA Example

	i	j	k	l
i	0	3	4	3
j	3	0	4	5
k	4	4	0	2
l	3	5	2	0

	i	j	k	l
i	0	3	4	3
j	3	0	4	5
k	4	4	0	**2**
l	3	5	**2**	0



	i	j	$\{k, l\}$
i	0	**3**	3.5
j	**3**	0	4.5
$\{k, l\}$	3.5	4.5	0



	$\{i, j\}$	$\{k, l\}$
$\{i, j\}$	0	**4**
$\{k, l\}$	**4**	0