CS4054
Bioinformatics

Spring 2025

Rushda Muneer

Phylogeny

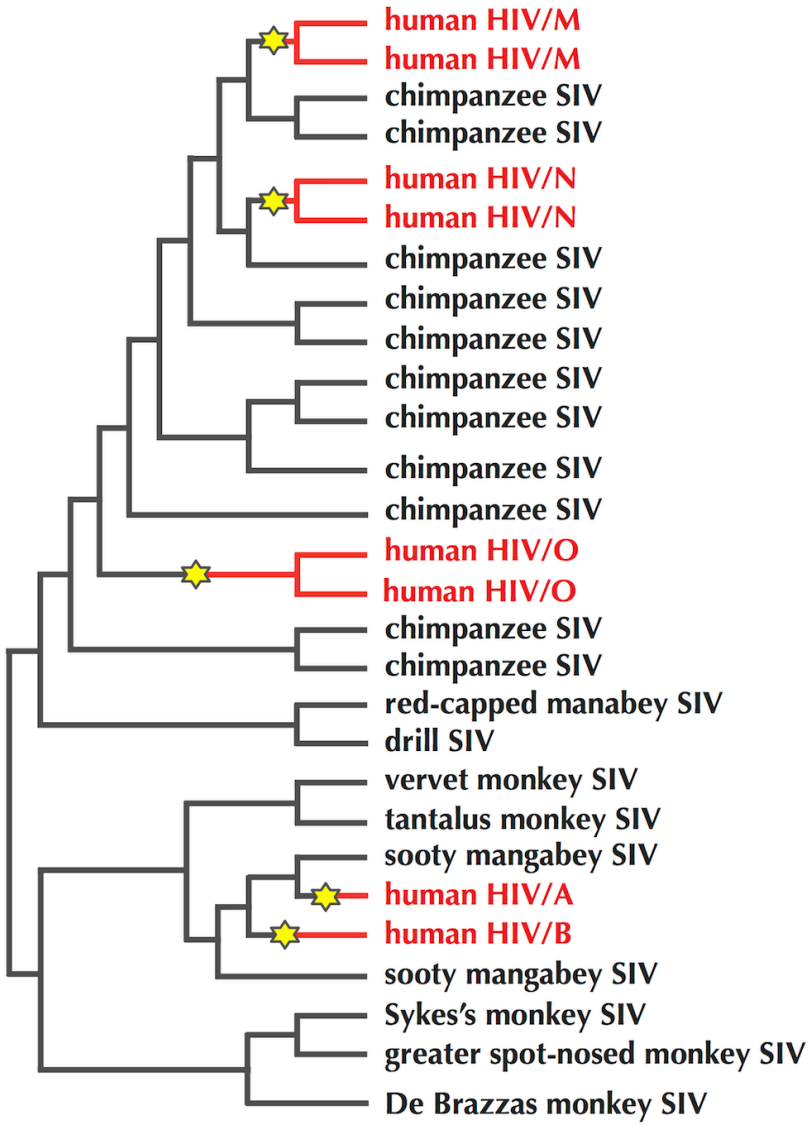
- Phylogeny is the **representation of the evolutionary history** and relationships between groups of organisms.
- The results are represented in a phylogenetic / evolutionary tree that provides a **visual output of relationships based** on shared or divergent physical and genetic characteristics.

The Evolution of SARS

- SARS is caused by a **coronavirus**, named for its **crown-like appearance** ("corona" is Latin for crown).
- These viruses typically infect the **respiratory tracts** of **mammals and birds**.
- **RNA viruses** (like **coronaviruses**, **influenza**, and **HIV**) mutate quickly due to **error-prone RNA replication**.
- This rapid mutation:
  - Explains yearly changes in **flu vaccines**.
  - Accounts for the many **HIV subtypes**.
- Important questions:
  - How did it **jump species**?
  - **When and where** did it start?
  - How did it **spread globally**?

Tracing SARS Through Evolutionary Trees

- Unanswered questions about **SARS** are tied to building **evolutionary trees** (*phylogenies*).
- **Phylogenies** help track how viruses **evolve** and **spread**.
- Example: Scientists used a primate virus **evolutionary tree** to show **HIV** was transmitted to humans **five separate times**.
- Similar methods are used to trace the **origins and spread of SARS-CoV**.



Building a Distance Matrix from Coronavirus Genomes

- To trace **SARS-CoV's jump to humans**, scientists sequenced coronaviruses from **different species**.
- Whole-genome comparison is difficult due to:
  - **Gene rearrangements**, **insertions**, and **deletions**.
- Focus was placed on the **Spike protein gene**:
  - Crucial for **binding to host cells**.
  - **1,255 amino acids** long in SARS-CoV.
  - Shows **weak similarity** to other coronaviruses but enough for **multiple alignment**.
- These alignments helped compare viruses and build a **distance matrix**.

Transforming Distance Matrices into Evolutionary Trees

- After aligning genes from **n species**, scientists create an **n × n distance matrix (D)**.
- Each entry $\mathbf{D_{ij}}$ often represents:
  - The number of **differences** between aligned gene sequences (e.g., symbol mismatches).

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

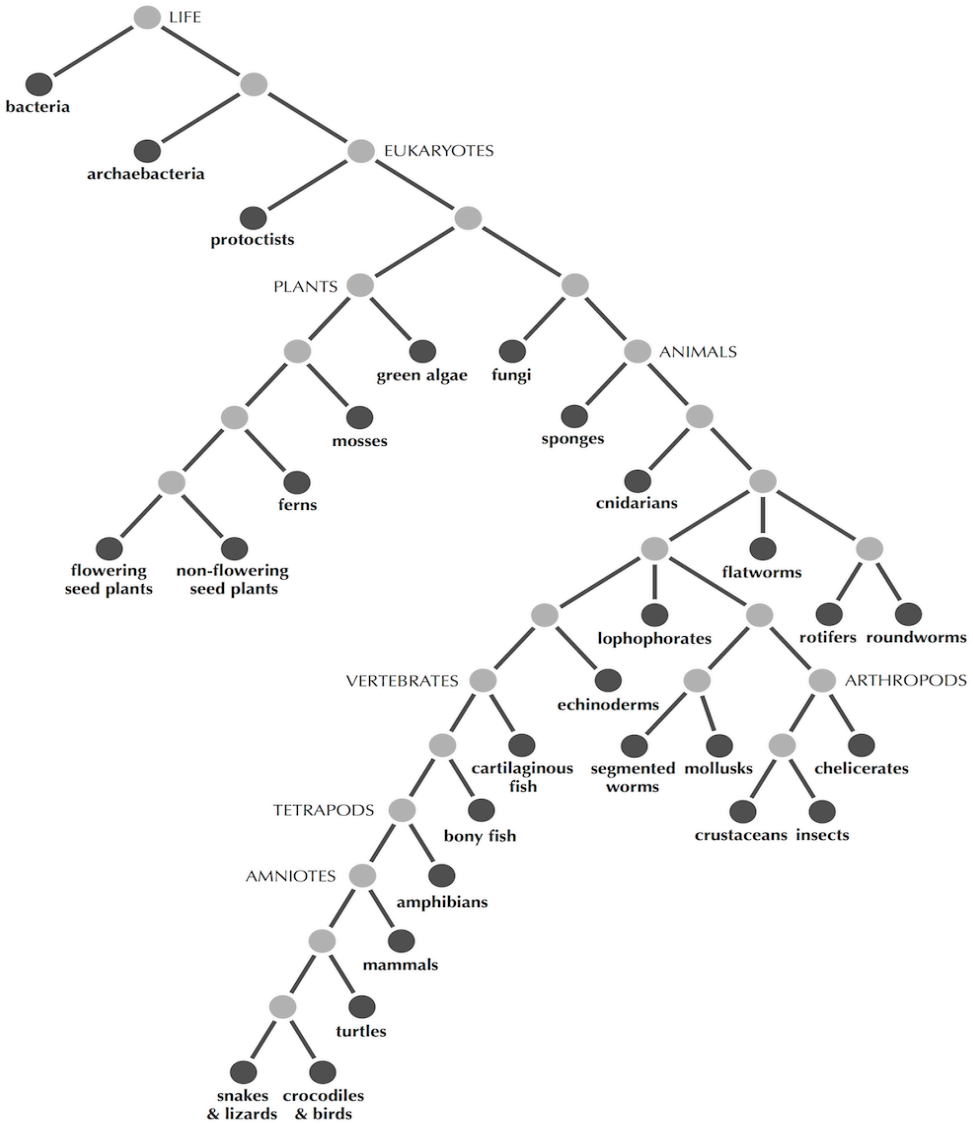
- The choice of distance function depends on the **application and data type**.

Key Properties of a Distance Matrix

- For a matrix **D** to qualify as a **distance matrix**, it must satisfy:
- **Symmetry**
  - For all *i* and *j*, $D_{i,j} = D_{j,I}$
- **Non-negativity**
  - For all *i* and *j*, $D_{i,j} \geq 0$
- **Triangle Inequality**
  - For all *i*, *j*, and *k*, $D_{i,j} + D_{j,k} \geq D_{i,k}$
- These properties ensure that **D** behaves like a true measure of distance across species or genomes.

Evolutionary trees as graphs

- Phylogeny of all life can be represented as a graph
- A connected graph without cycles that models an evolutionary tree of life on Earth. Present-day species are shown as darker nodes

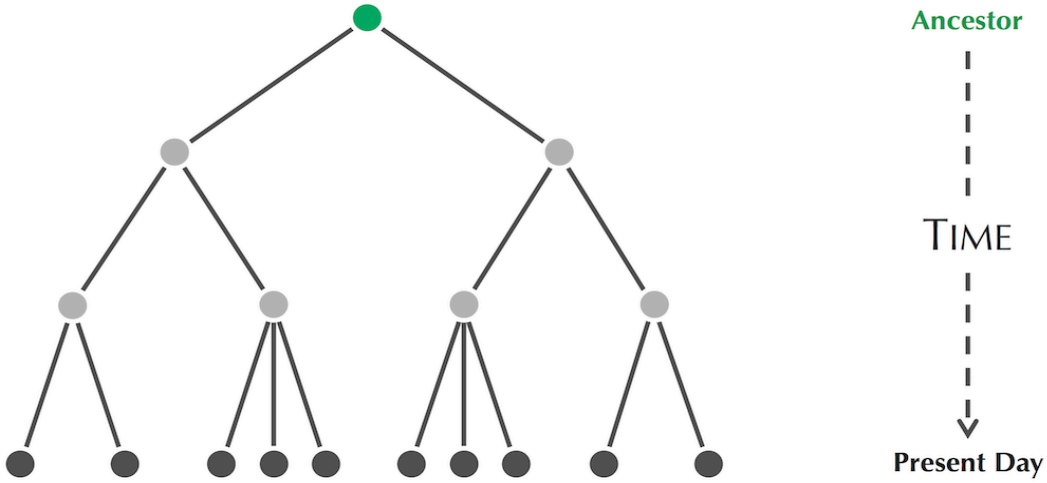


Trees in Phylogenetic Modeling

- **Phylogenetic trees** are a type of **graph** with two main properties:
  - **Connected** – any node can be reached from any other node.
  - **Acyclic** – the graph contains **no cycles**.
- Therefore, a **tree** is a **connected, cycle-free graph**.
- **Node types**:
  - **Leaves**: nodes with **degree 1.**
  - **Internal nodes**: nodes with **degree > 1**.
- **Leaves, Parents, and Limbs in a Tree**
  - For any **leaf** j, there is exactly **one connected node**.
  - This node is called the **parent** of j, denoted as **Parent(j)**.
  - The **edge** between a leaf and its parent is called a **limb**.

Rooted vs. Unrooted Trees

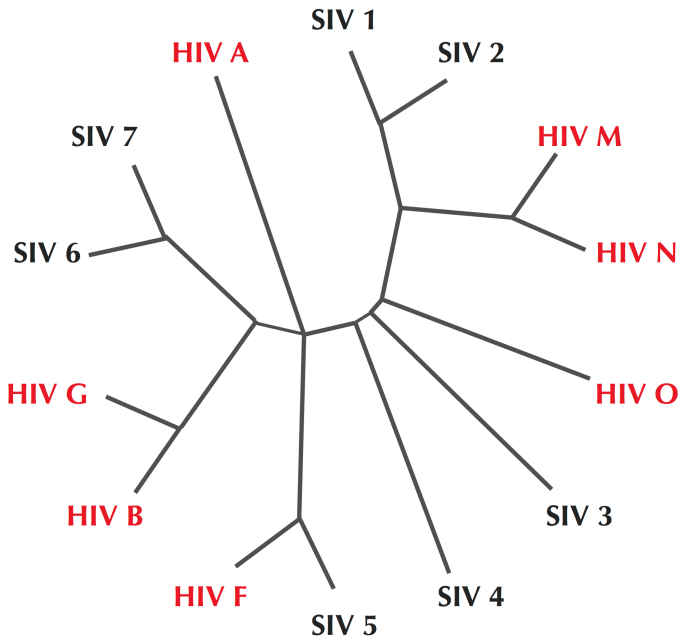
- A **rooted tree** has a special node called the **root**.
  - **Edges** are implicitly **oriented away** from the root.
  - The **root** represents the **common ancestor** of all species.
  - Orientation reflects the **flow of evolutionary time**.
- In contrast, **unrooted trees** do **not specify a root** and lack directional flow.



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.

Unrooted HIV tree

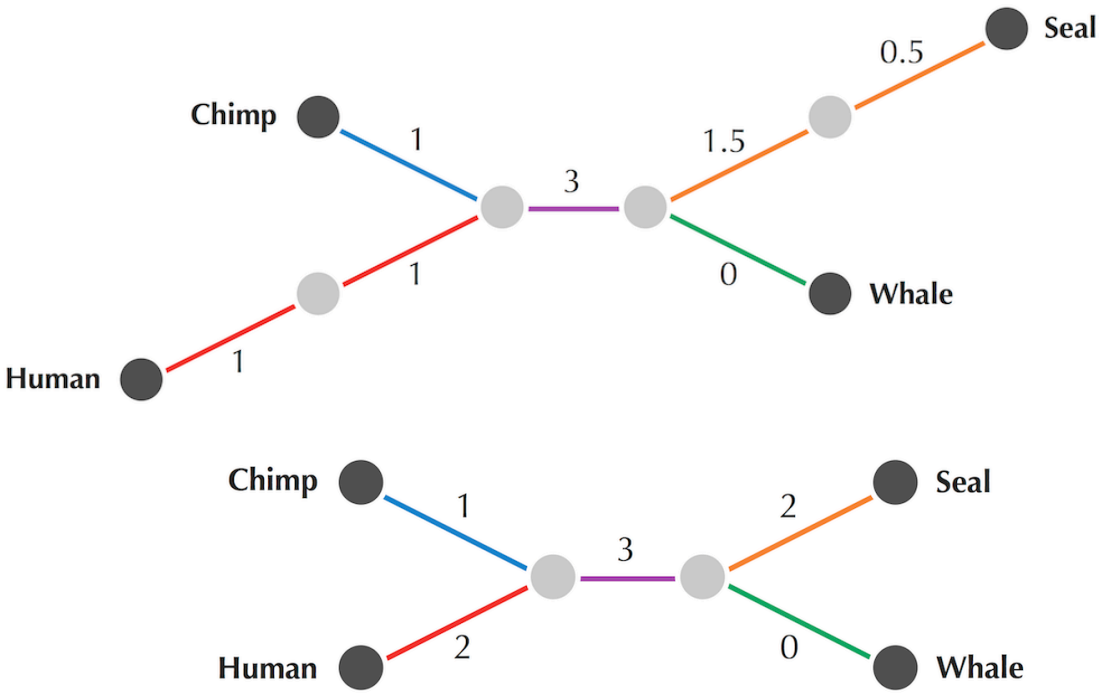
- Shows HIV viruses from **different datasets**.



Distance-Based Phylogeny Construction

- Focus on deriving an **unrooted tree** from a **distance matrix**.
  - **Leaves** represent the **species** in the matrix.
  - **Internal nodes** represent **ancestral species**.
- Each **edge** in the tree has a **non-negative length**, reflecting the **evolutionary distance** between the connected organisms.

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

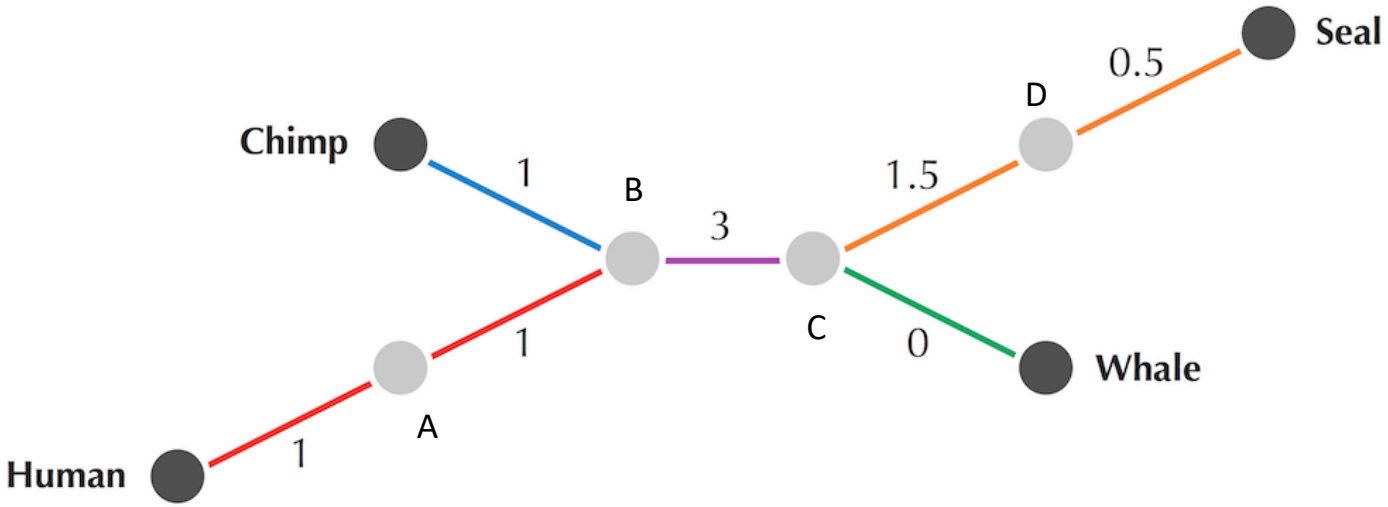


Distance-Based Phylogeny Problem

- **Goal**: Reconstruct an **unrooted evolutionary tree** from a **distance matrix**.
- A tree **T** fits a distance matrix **D** if $d_{i,j}(T) = D_{i,j}$ for all pairs of leaves i and j.
- **Problem Overview**:
  - **Input**: A **distance matrix**.
  - **Output**: A **tree** that fits the given distance matrix.

Non-Branching Paths

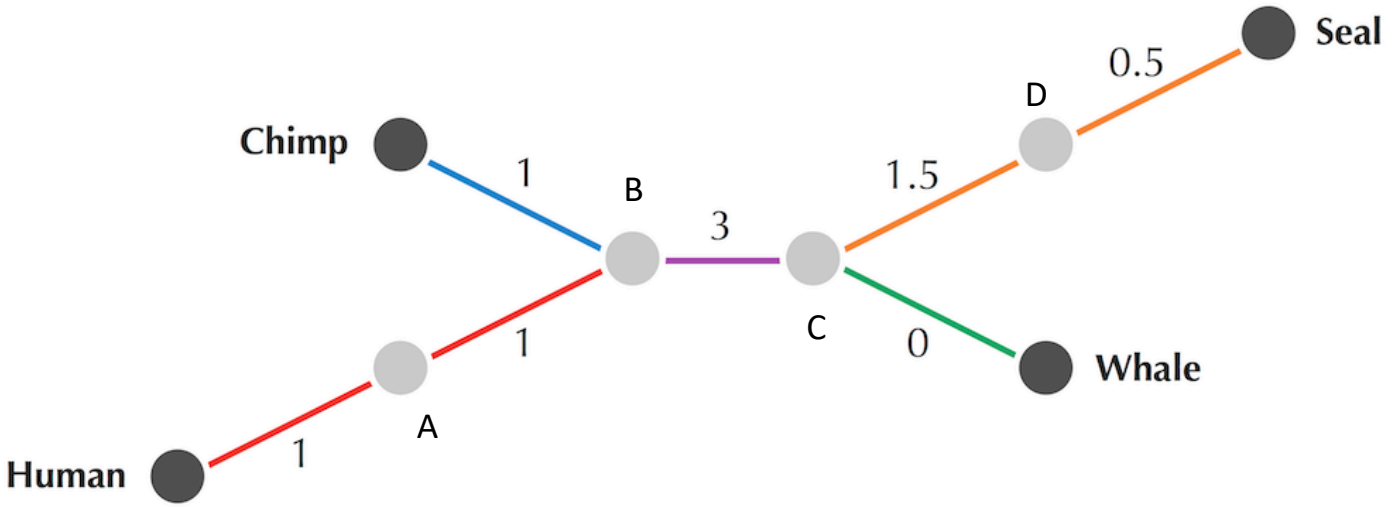
- A path in a tree is **non-branching** if every node, except the starting and ending nodes has a degree 2.



Is the path from Human to Seal non-branching?
Which paths are non-branching?

Maximal Non-Branching Path

- A non-branching path is **maximal** if it isn't part of a longer non-branching path.



Identify the maximal non-branching paths

Transformation to a Simple Tree

- Replace every **maximal non-branching path** with a single edge whose length equals the path's total length.
- After this transformation, the tree becomes a **simple tree**, with **no nodes of degree 2**.