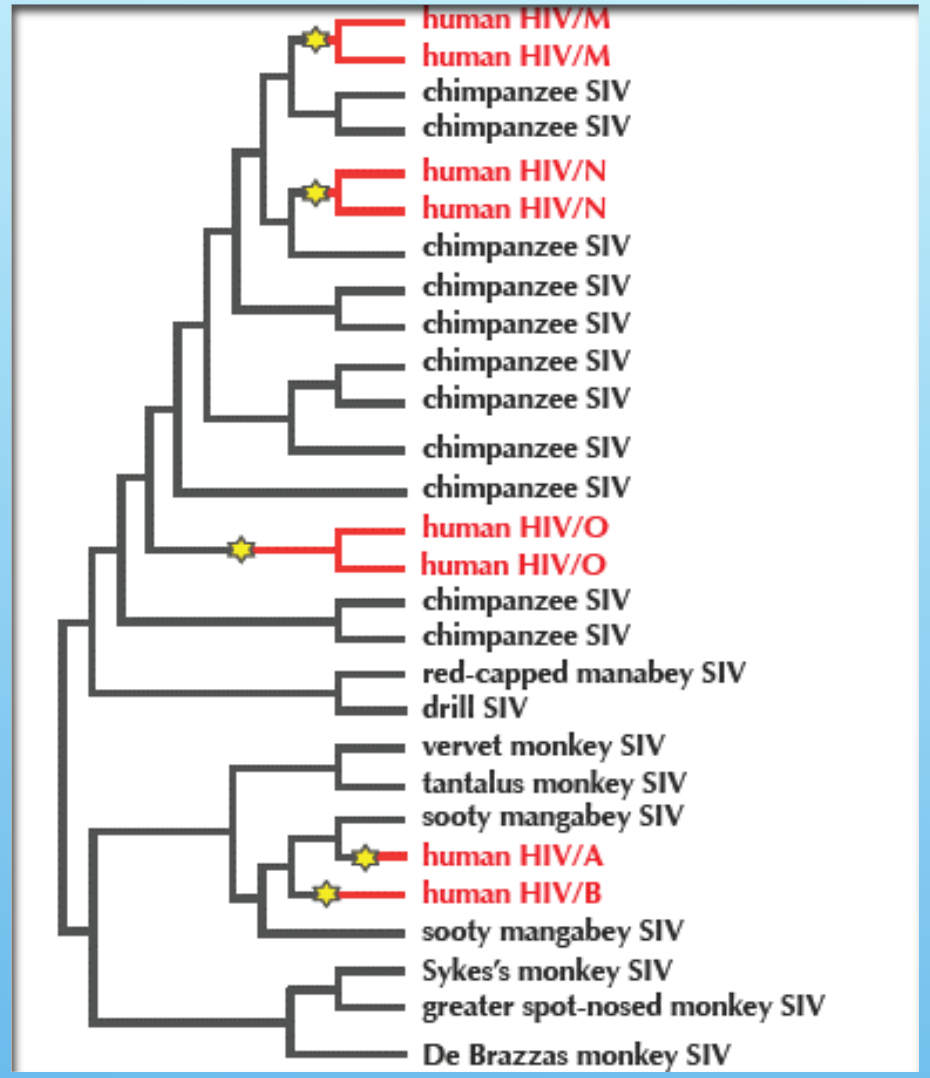
Which Animal Gave Us SARS

Evolutionary tree Reconstruction

Introduction to Bioinformatics
Lecture 3

Agenda

- The Fastest Outbreak
- From Distance Matrix to Evolutionary Tree
- Distance-Based Phylogeny



THE FASTEST OUTBREAK

Trouble at the Metropole Hotel

- February 21, 2003: a Chinese doctor named Jianlun flew to Hong Kong to attend a wedding and checked into 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 China where a deadly, highly contagious respiratory illness had infected hundreds of people. The Chinese government made brief mention of this incident to the WHO but had concluded that a common bacterial infection most likely the cause, it was already too late to stop the outbreak.
- February 23: a man who stayed across the hall from Dr. Jianlun at the Metropole traveled to Hanoi and died after infecting 80 people.
- February 26: a woman checked out of the Metropole, traveled back to Toronto, and died after initiating an outbreak there.
- March 1: a third guest was admitted to a hospital in Singapore, where sixteen additional cases of the illness arose within two weeks.

THE FASTEST OUTBREAK

Severe Acute Respiratory Syndrome (SARS)

- Businesses were closed, sick passengers were removed from airplanes, and Chinese officials threatened to execute infected patients who violated quarantine.
- This mysterious new disease had crossed the Pacific Ocean within a week of entering Hong Kong, while the HIV took two decades to circle the globe.
- International travel may have helped the disease spread rapidly, but international collaboration would eventually contain it.
- In a matter of a few weeks, biologists identified a virus that had caused the epidemic and sequenced its genome.
- In the process, the mysterious new disease earned a name: **Severe Acute Respiratory Syndrome**, or **SARS**.

THE FASTEST OUTBREAK

The evolution of SARS

- The virus causing SARS belongs to a family of viruses called **coronaviruses**
- Coronaviruses infect the respiratory tracts of mammals and birds but typically cause only minor problems, like the common cold**.**
- 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, which explains the flu shot changes from year to year and why there are many different subtypes of HIV.

THE FASTEST OUTBREAK

The evolution of SARS

- Researchers initially hypothesized that, **SARS-Cov** had jumped from animals to humans. They first named birds as the likely suspect because of the similarities between SARS and “bird flu”. But after more research, they believed that SARS did not come from birds because its genome did not resemble avian coronaviruses.
- When and where did it happen? How did SARS spread around the world, and who infected whom?
- Scientists started sequencing coronaviruses from various species to determine which one is the most similar to SARS-CoV to determine how SARS jumped from animals to humans.
- Scientists focused on only one of the genes in SARS-CoV which encodes the **Spike protein**, which identifies and binds to receptor sites on the host’s cell membrane.

FROM DISTANCE MATRIX TO EVOLUTIONARY TREE

Distance Matrix

- After constructing a multiple alignment of genes from n different species, biologists often transform this alignment into an $n \times n$ **distance matrix** D.
- In the distance matrix, $D_{i,j}$ represents the number of differing symbols between the genes representing rows i and j of the alignment
  - It may be representing any other distance functions in order to suit different applications (e.g., edit distance)
- D must satisfy three properties
  - **Symmetric** (for all i and j, $D_{i,j} = D_{j,i}$),
  - **Nonnegative** (for all i and j, $D_{i,j} \geq 0$)
  - satisfy the **triangle inequality** (for all i, j, and k, $D_{i,j} + D_{j,k} \geq D_{i,k}$).

FROM DISTANCE MATRIX TO EVOLUTIONARY TREE

Distance Matrix

- For example:

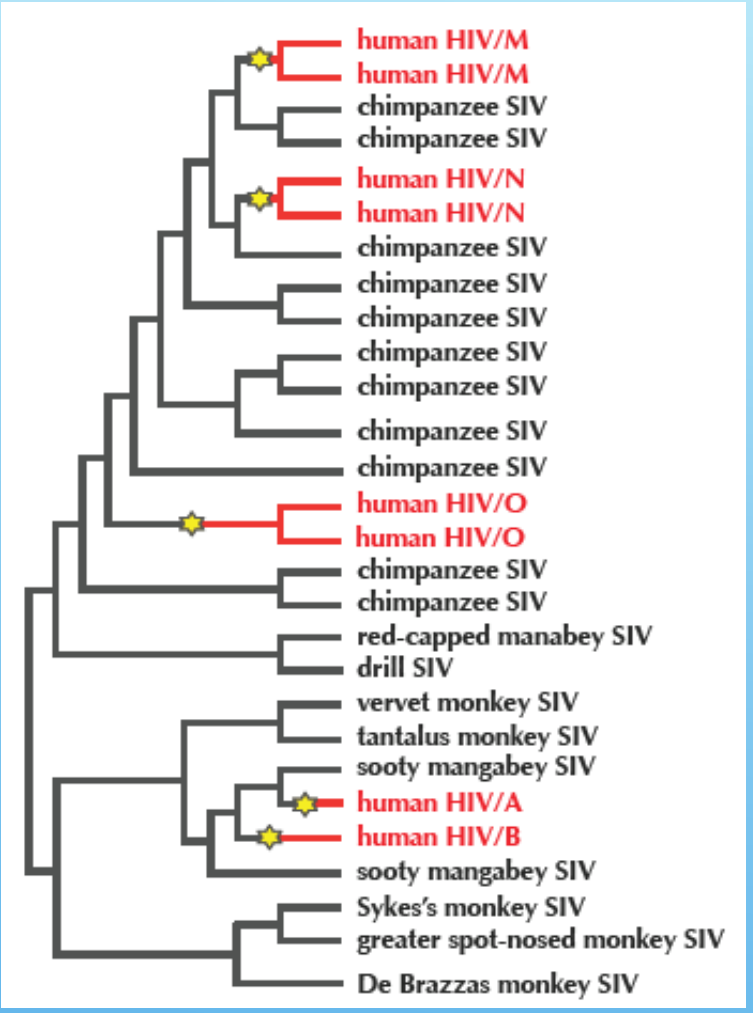
Species	Alignment	Distance Matrix			
		Chimp	**Human**	**Seal**	**Whale**
Chimp	ACGTAGGCCT	0	3	6	4
Human	ATGTAAGACT	3	0	7	5
Seal	TCGAGAGCAC	6	7	0	2
Whale	TCGAAAGCAT	4	5	2	0

- By the end of 2003, bioinformaticians had sequenced many coronaviruses taken from a variety of animals and SARS patients and then computed the associated distance matrix.
- They needed to use this information in order to construct a coronavirus **phylogeny** and understand the origin and spread of the SARS epidemic.

FROM DISTANCE MATRIX TO EVOLUTIONARY TREE

Evolutionary trees as graphs

- This is the HIV **evolutionary tree** (**phylogeny**)
- We want to contract a similar tree for SARS-Cov.
- **An evolutionary tree** (or any tree) is a graph that must satisfy 2 conditions: Connected and Contains no cycles
- Nodes with degree larger than 1 are called **internal nodes** while, those with degree 1 are called a **leaves**.
- Given a leaf j, there is only one node connected to j by an edge called the **parent** of j, denoted **PARENT(j).**
- An edge connecting a leaf to its parent is called a **limb**.
- A **rooted tree** is a tree with one node designated as a special node called the **root.**



FROM DISTANCE MATRIX TO EVOLUTIONARY TREE

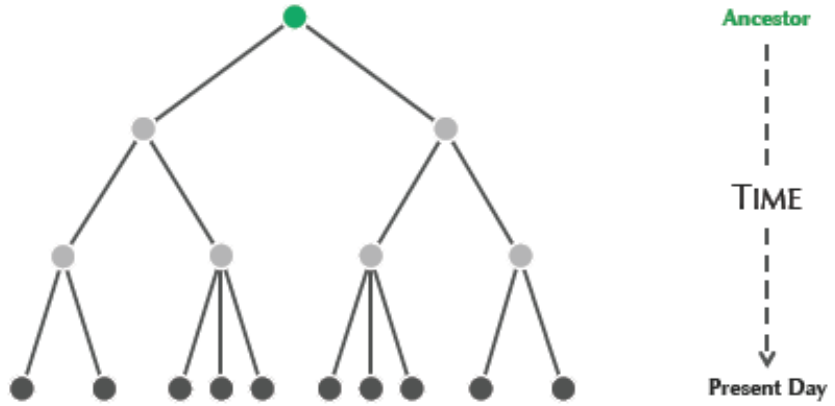


FIGURE 7.5 A rooted tree, with the root (representing an ancestor of all species in the tree) indicated in green at the top of the tree. The presence of the root implies an orientation of edges in the tree away from the root.

We use rooted trees when dealing with node corresponding to the ancestor of all species in the tree; otherwise, we use unrooted trees.

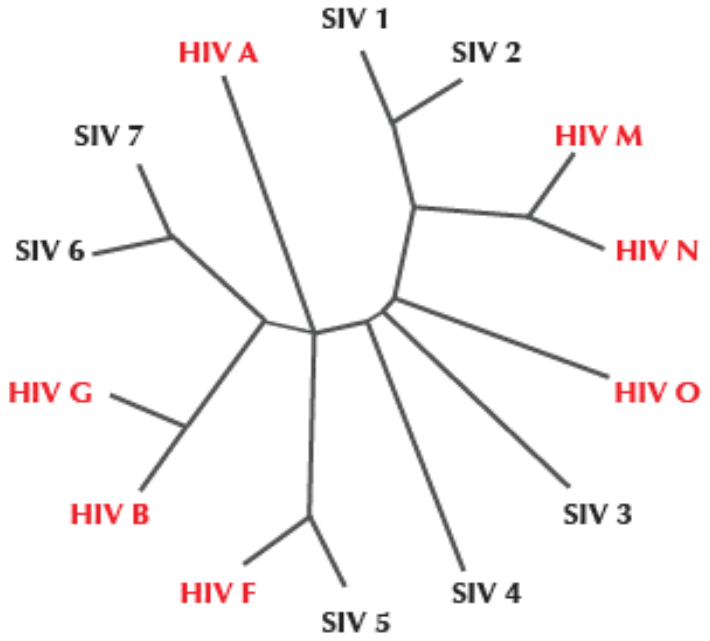
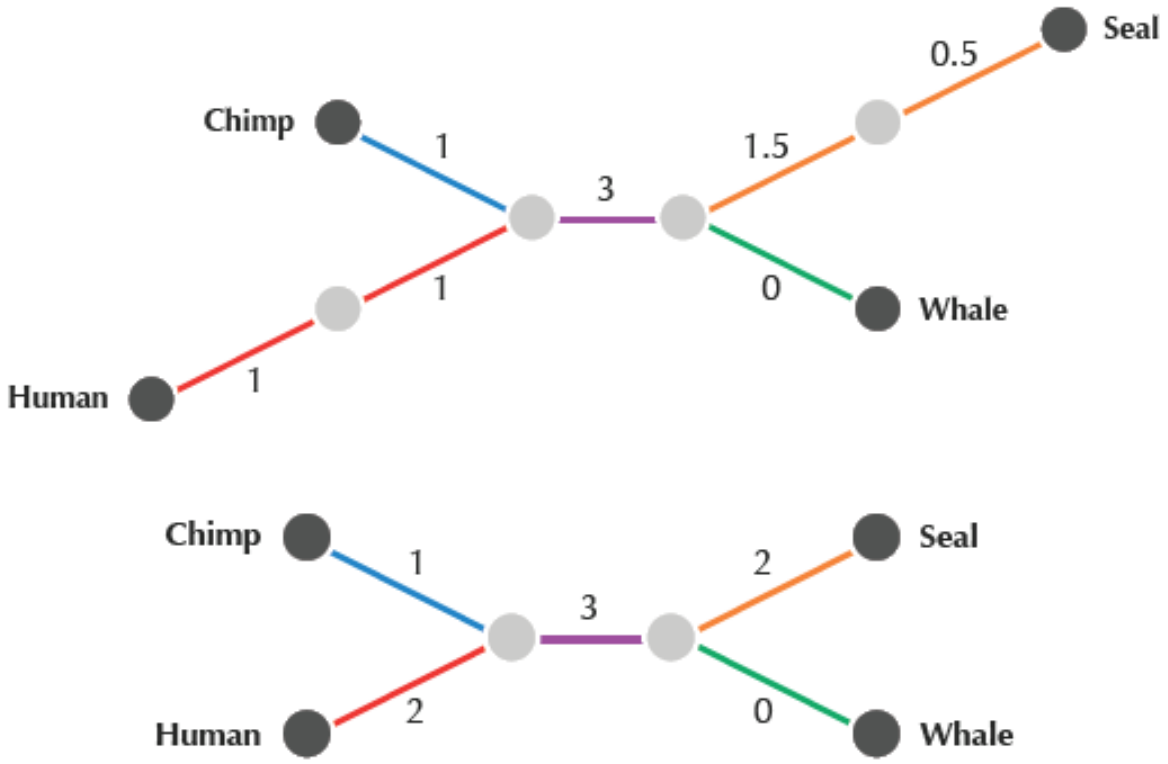


FIGURE 7.6 An unrooted tree of HIV and SIV viruses that suggests additional viral families F and G in addition to the viral families A, B, M, N, and O shown in Figure 7.2.

FROM DISTANCE MATRIX TO EVOLUTIONARY TREE

Distance-based phylogeny construction

- First, we focus on constructing **unrooted tree** form the distance matrix.
  - Leaves correspond to the species represented by the matrix
  - Internal nodes correspond to unknown ancestral species.
  - Each edge has a non-negative length (wight) representing the distance between the organisms that the edge connects (reflecting the evolutionary distance between species) and is denoted by $d_{i,j}$
  - the length of a path in a tree as the sum of the lengths of its edges.
  - So, the evolutionary distance between two present-day species corresponding to leaves i and j in a tree T is equal to the length of the unique path connecting i and j, denoted $d_{i,j}(T)$



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 7.7 Two unrooted trees fitting the distance matrix in Figure 7.3. Each of the five maximal non-branching paths in the tree on the top is shown using a different color. Replacing each maximal non-branching path in this tree with a single edge (of length equal to the total length of edges) results in the simple tree shown on the bottom.

FROM DISTANCE MATRIX TO EVOLUTIONARY TREE

Distance-based phylogeny construction

Distances Between Leaves Problem:
Compute the distances between leaves in a weighted tree.

Input: A weighted tree with n leaves.
Output: An $n \times n$ matrix $(d_{i,j})$, where $d_{i,j}$ is the length of the path between leaves i and j.

The Distance Between Leaves Problem is straightforward to solve, but we would like to solve the reverse problem, in which we must construct an unrooted tree that models a given distance matrix. We say that a weighted unrooted tree T fits a distance matrix D if $d_{i,j}(T) = D_{i,j}$ for every pair of leaves i and j.

Distance-Based Phylogeny Problem:
Reconstruct an evolutionary tree fitting a distance matrix.

Input: A distance matrix.
Output: A tree fitting this distance matrix.

FROM DISTANCE MATRIX TO EVOLUTIONARY TREE

Distance-based phylogeny construction

- A distance matrix is called **additive** if there exists a **tree that fits this matrix** and **non-additive** otherwise.
- A tree is called **simple tree** if it satisfies the property that **there are no nodes of degree 2.**
- if a matrix is additive, then there exists **a unique simple tree fitting this matrix.**
- In the Distance-Based Phylogeny Problem, we will therefore use the terminology
- TREE(D) to denote the simple tree fitting the additive distance matrix D.
- Our question, then, is how to construct TREE(D) from D.

EXERCISE BREAK: Prove the following statements:

- Every tree with at least two nodes has at least two leaves.
- Every tree with n nodes has $n - 1$ edges.

EXERCISE BREAK: Prove that there exists exactly one path connecting every pair of nodes in a tree. Hint: what would happen if there were two different paths connecting a pair of nodes? What would happen if there were no paths connecting a pair of nodes?

STOP and Think: Does the Distance-Based Phylogeny Problem always have a solution?

EXERCISE BREAK: Prove that every simple tree with n leaves has at most $n - 2$ internal nodes.

TOWARD AN ALGORITHM FOR DISTANCE-BASED PHYLOGENY CONSTRUCTION

A quest for neighboring leaves

- We first assume, that the two closest species with respect to the distance matrix D correspond to neighbors in TREE(D) .
  - i.e. the minimum value $D_{i,j}$ should correspond to leaves i and j having the same parent.
  - we are referring to a minimum off-diagonal element, i.e., a value $D_{i,j}$ such that $i \neq j$.

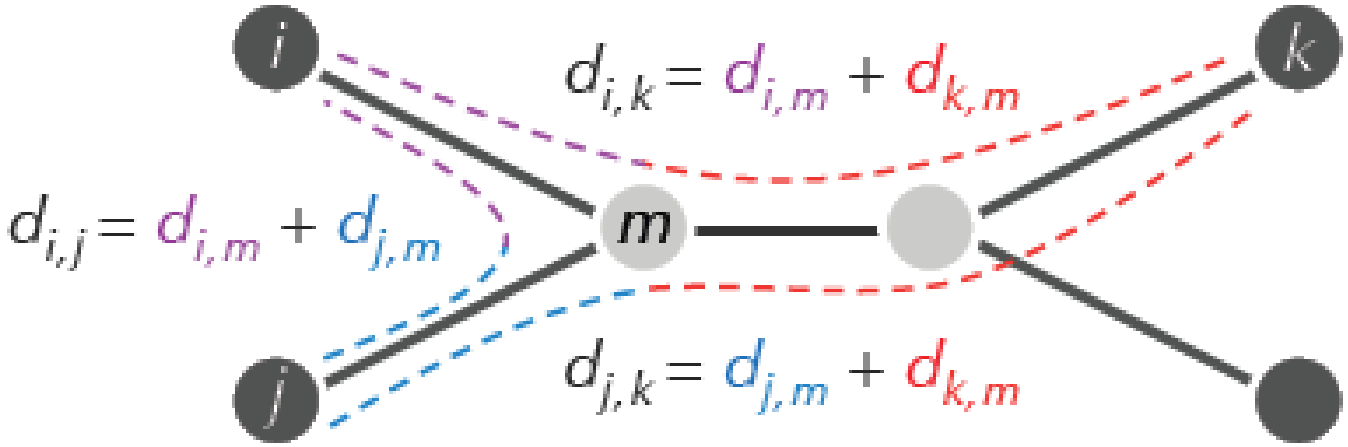
- **Theorem.** Every simple tree with at least three nodes has a pair of neighboring leaves.
  - Proof: look at p.12 in the book

TOWARD AN ALGORITHM FOR DISTANCE-BASED PHYLOGENY CONSTRUCTION

A quest for neighboring leaves

- for every neighboring leaves *i* and *j* sharing a parent node *m*, the following equality holds for every other leaf *k* in the tree:

$$d_{k,m} = \frac{(d_{i,m} + d_{k,m}) + (d_{j,m} + d_{k,m}) - (d_{i,m} + d_{j,m})}{2} = \frac{d_{i,k} + d_{j,k} - d_{i,j}}{2}$$



TOWARD AN ALGORITHM FOR DISTANCE-BASED PHYLOGENY CONSTRUCTION

A quest for neighboring leaves

- Since *i, j* and *k* are leaves, we can compute the distance $d_{k,m}$ between the nodes *k* and *m* in terms of the additive matrix *D* as follows:

$$d_{k,m} = \frac{(D_{i,k} + D_{j,k} - D_{i,j})}{2}$$

- If the degree of *m* is 3, then removing leaves *i* and *j* from the tree turns *m* into a leaf and thus reduces the total number of leaves.
- This is equivalent to removing rows and columns *i* and *j* from *D*, then adding a new row and column corresponding to their parent *m.*
- the distances from *m* to other leaves in the matrix are computed according to the above formula.

TOWARD AN ALGORITHM FOR DISTANCE-BASED PHYLOGENY CONSTRUCTION

A quest for neighboring leaves

A recursive algorithm for the Distance-Based Phylogeny Problem:

- find a pair of neighboring leaves i and j by selecting the minimum element $D_{i,j}$ in the distance matrix;
- replace i and j with their parent, and recompute the distances from this parent to all other leaves as described above;
- solve the Distance-Based Phylogeny problem for the smaller tree;
- add the previously removed leaves i and j back to the tree.

EXERCISE BREAK: Apply this recursive approach to the distance matrix shown in Figure 7.9 (left). (Solve this exercise by hand.)

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