

# Chapter 7

## Severe Acute Respiratory Syndrome (SARS)

### A post-genomic epidemic

#### Phylogenetic analysis

Prof. dr. Antoine van Kampen

Biosystems Data Analysis

Swammerdam Institute for Life Sciences

Bioinformatics Laboratory

Academic Medical Centre

# SARS: the outbreak

- \* February 28, 2003, Hanoi, the Vietnam French hospital called the WHO with a report of an influenza-like infection.
- \* Dr. Carlo Urbani (WHO) came and concluded that this was a new and unusual pathogen.
- \* Next few days Dr. Urbani collected samples, worked through the hospital documenting findings, and organized patient quarantine.
- \* Fever, dry cough, short breath, progressively worsening respiratory failure, death through respiratory failure.

# SARS: the outbreak

- \* Dr. Carlo Urbani was the first to identify *Severe Acute Respiratory Syndrome: SARS*.
- \* In three weeks Dr. Urbani and five other healthcare professionals from the hospital died from the effects of *SARS*.



- \* By March 15, 2003, the WHO issued a global alert, calling *SARS* a worldwide health threat.

# Origin of the SARS epidemic

- \* Earliest cases of what now is called *SARS* occurred in November 2002 in [Guangong](#) (P.R. of China)
- \* [Guangzhou hospital](#) spread 106 new cases
- \* A doctor from this hospital visited Hong Kong, on Feb 21, 2003, and stayed in the 9th floor of the Metropole Hotel
- \* The doctor became ill and died, diagnosed pneumonia
- \* Many of the visitors of the 9th floor of the Metropole Hotel now became disease carriers themselves

# Guangong, Guangzhou, Hong Kong



# Origin of the SARS epidemic

- \* One of the visitors of the 9th floor of the Metropole Hotel was an American business man who went to [Hanoi](#), and was the first patient to bring *SARS* to the Vietnam French hospital of Hanoi.
- \* He infected 80 people before dying
- \* Other visitors of the 9th floor of the Metropole Hotel brought the disease to [Canada](#), [Singapore](#) and [the USA](#).
- \* By end April 2003, the disease was reported in 25 countries over the world, on 4300 cases and 250 deaths.

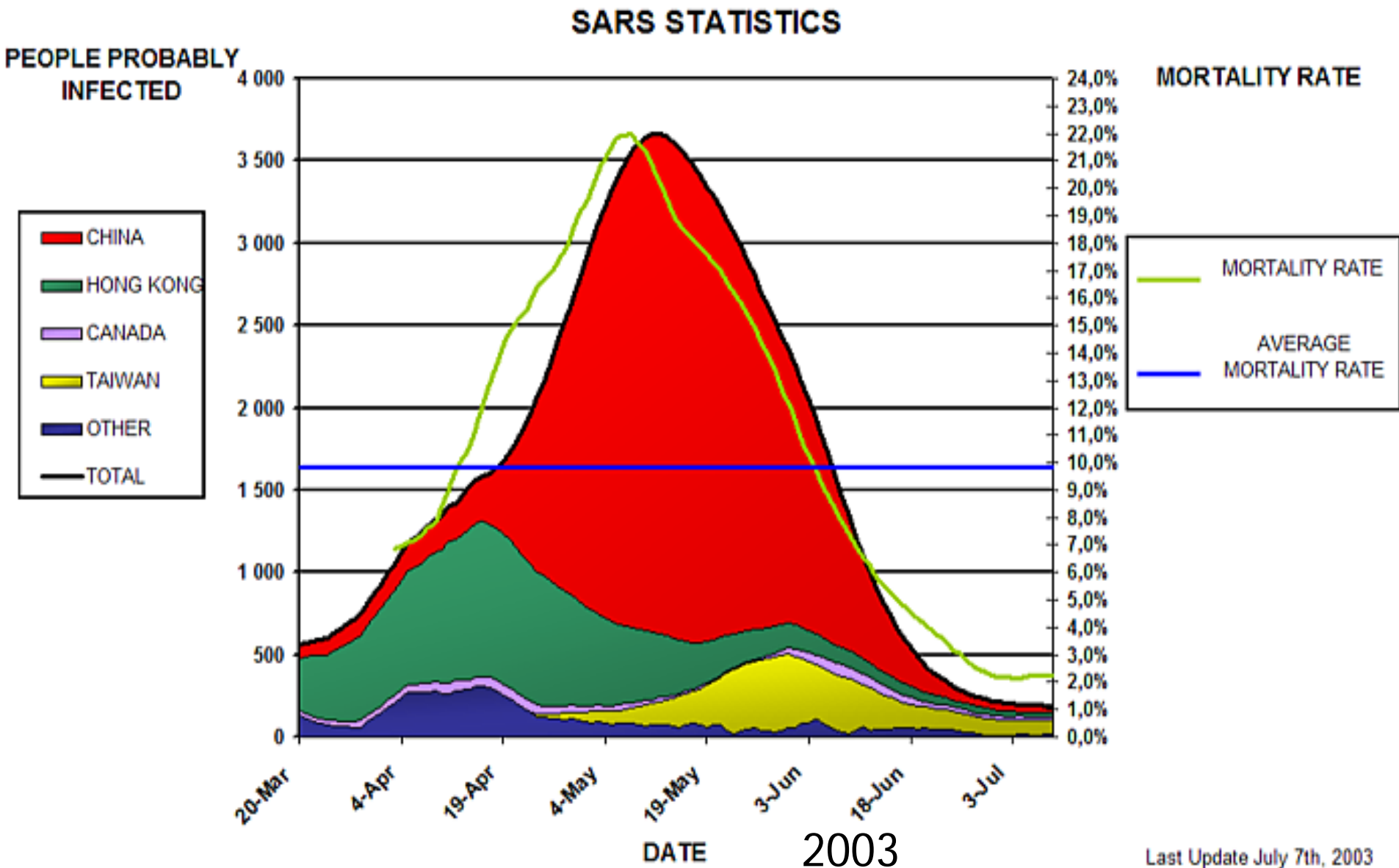
# Hanoi, Vietnam, Singapore



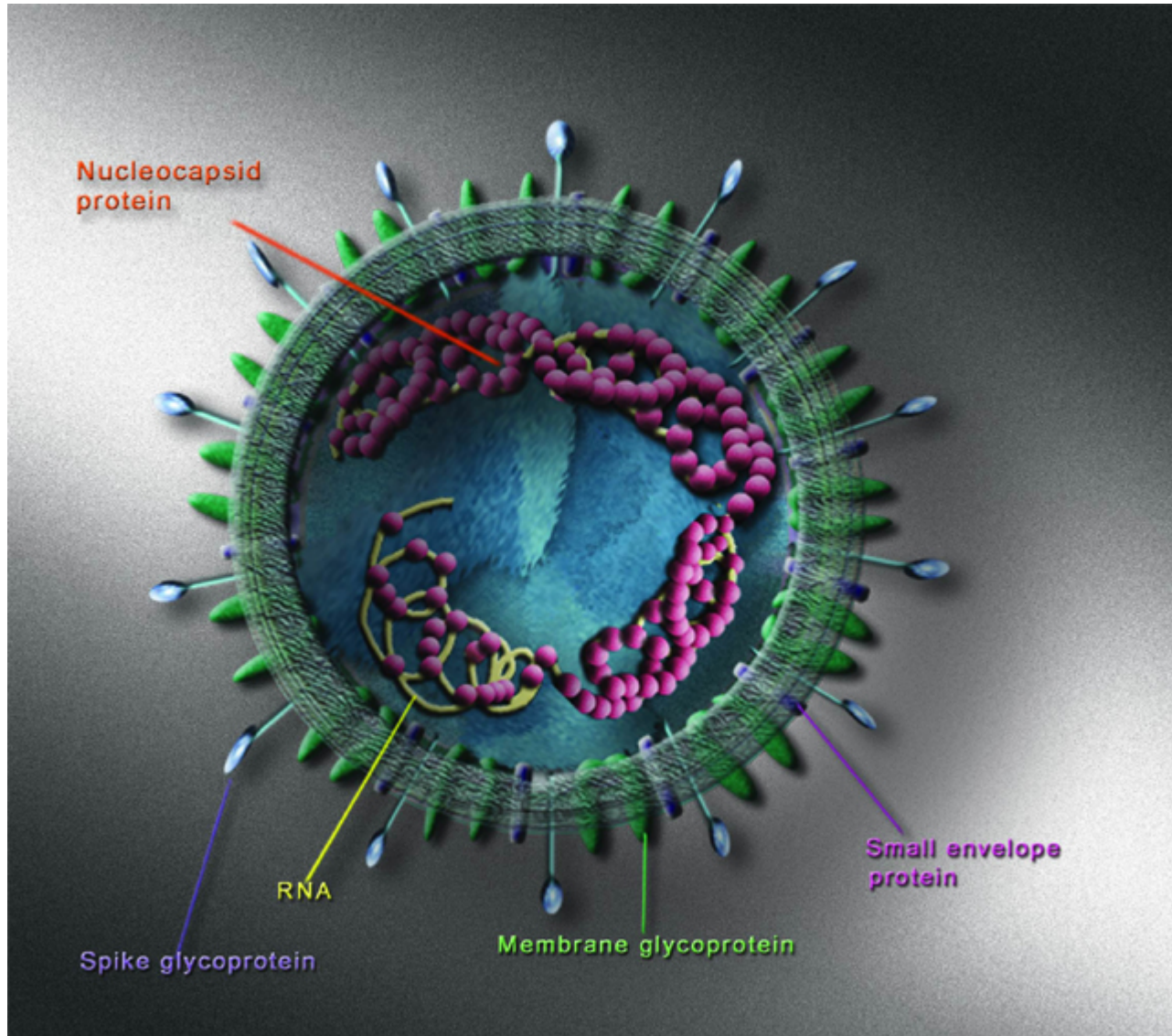
# The SARS corona virus

- \* Early March 2003, the WHO coordinated an international research program.
- \* End March 2003, laboratories in Germany, Canada, United States, and Hong Kong independently identified a novel virus that caused *SARS*.
- \* The *SARS* corona virus (SARS-CoV) is an RNA virus (like HIV).
- \* Corona viruses are common in humans and animals, causing ~25% of all upper respiratory tract infections (e.g. common cold) .

# SARS: the outbreak



# The SARS corona virus



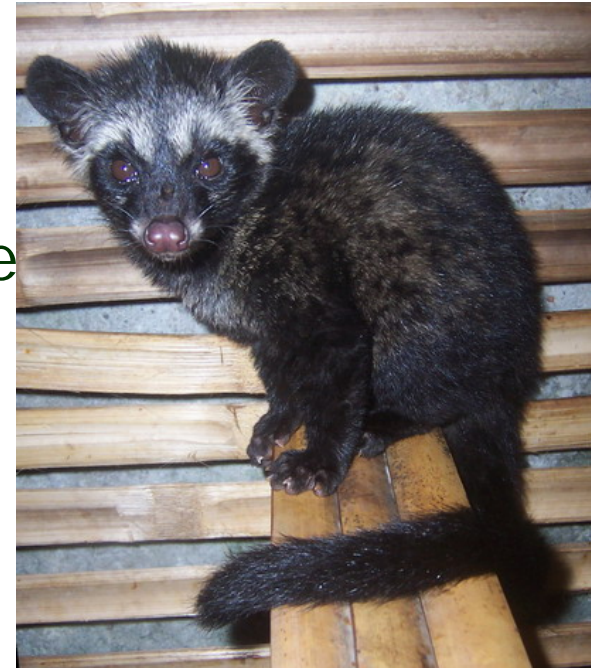
# The SARS corona virus

\* April 2003, a laboratory in Canada announced the entire RNA genome sequence of the SARS CoV virus.

•Phylogenetic analysis of the SARS corona virus showed that the most closely related CoV stem from the *palm civet* which were substantially different from any known human viruses.

•The palm civet is a popular food item in the Guangdong province of China.

•Also different from bird CoVs - so no relation to bird flue.



# SARS genome (about 29.751bp) - 2003

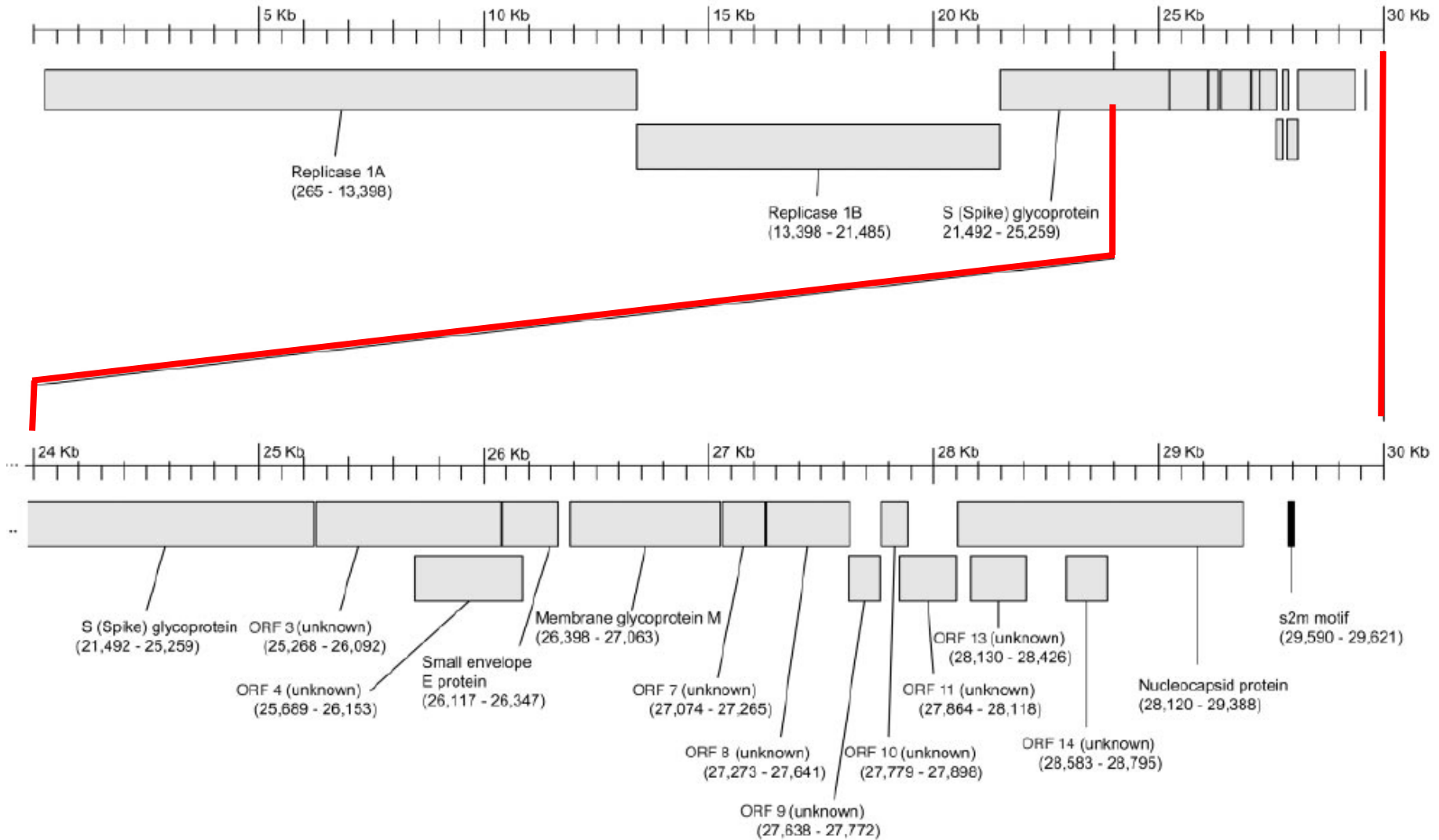


fig. 2. Map of the predicted ORFs and s2m motif in the Tor2 SARS virus genome sequence.

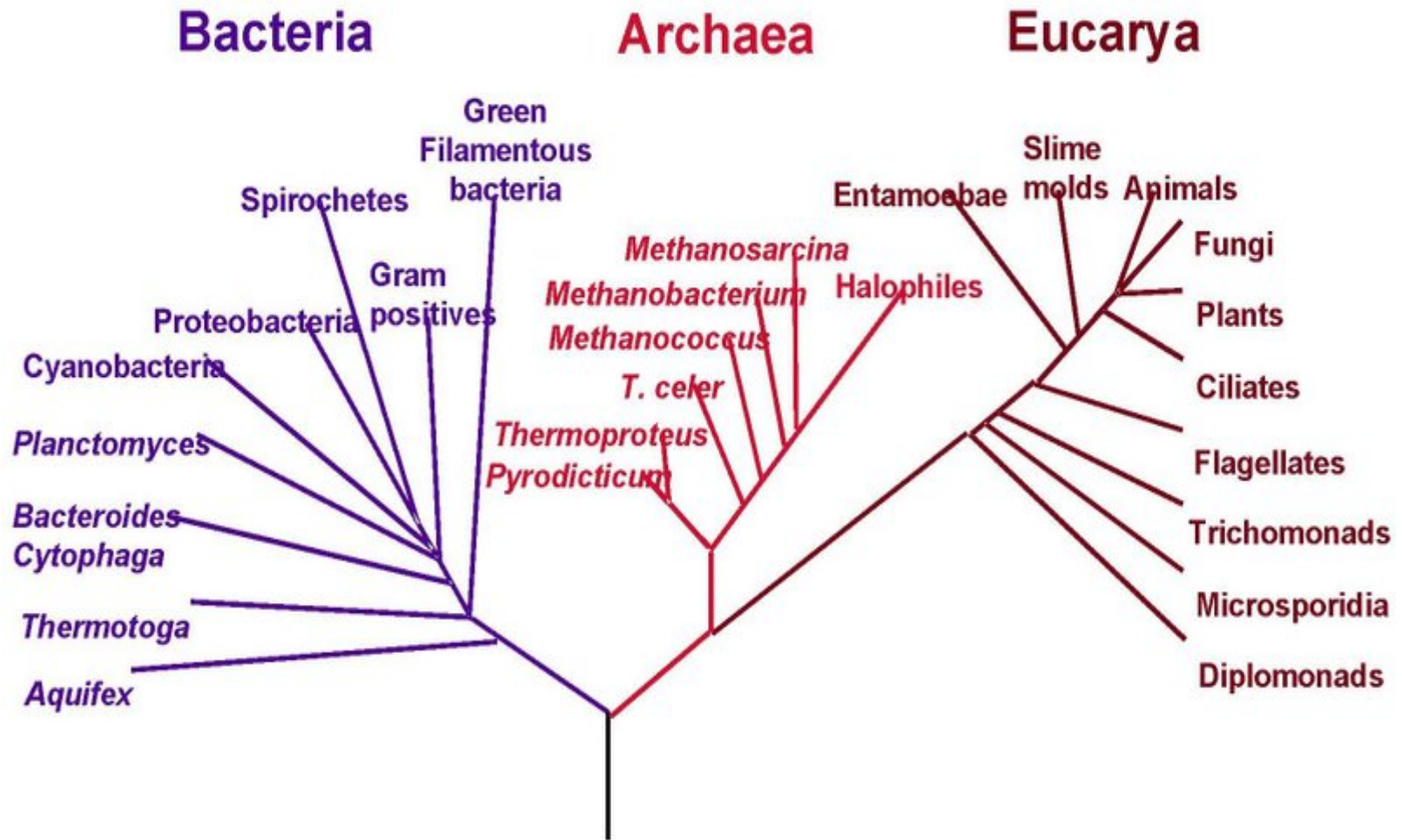
# Phylogenetics

**Phylogenetics** is the study of evolutionary relatedness among various organisms or genes

A **phylogenetic tree** is a tree showing the evolutionary interrelationships among various species or other entities that are believed to have a common ancestor.

**In general:** we are able to draw evolutionary trees because all species/individuals on earth share a common ancestor

# Phylogenetic Tree of Life

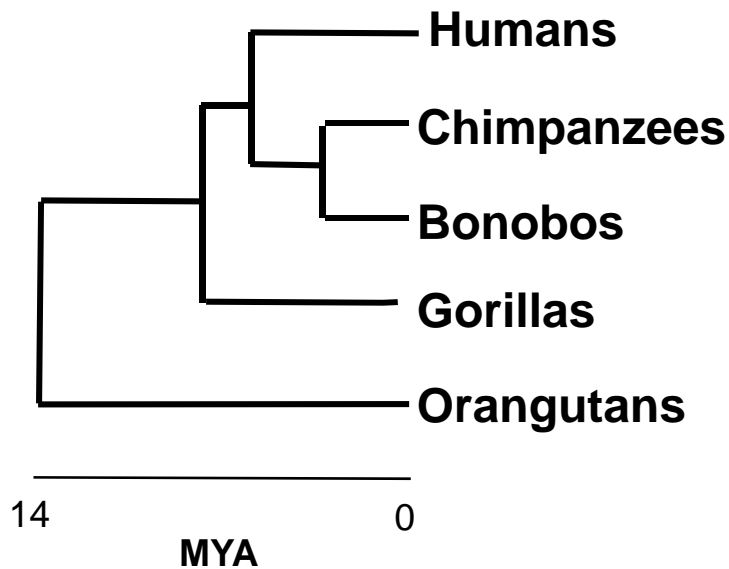


# Phylogenetic analysis of SARS CoV

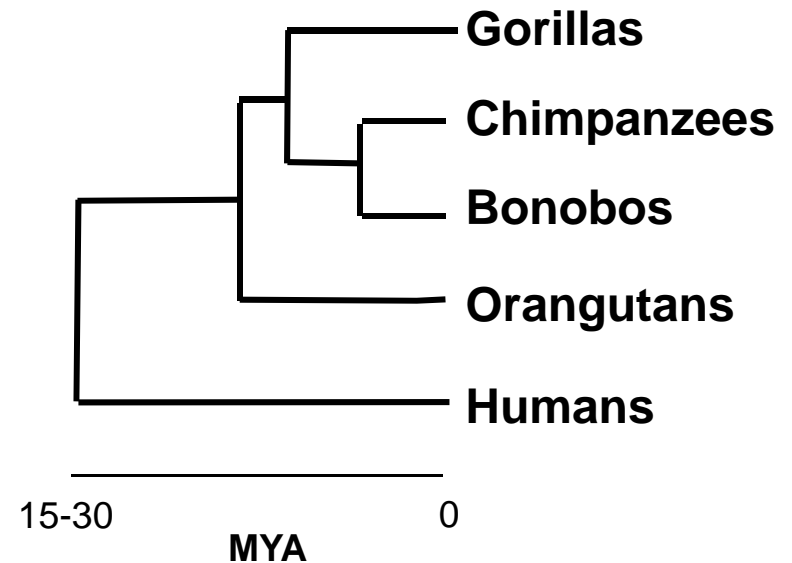
Phylogenetic analysis helps to answer:

- \* What kind of virus caused the original infection?
- \* What is the source of the infection?
- \* When and where did the virus cross the species border?
- \* What are the key mutations that enabled this switch?
- \* What was the trajectory of the spread of the virus?

# Which species are the closest living relatives of modern humans?



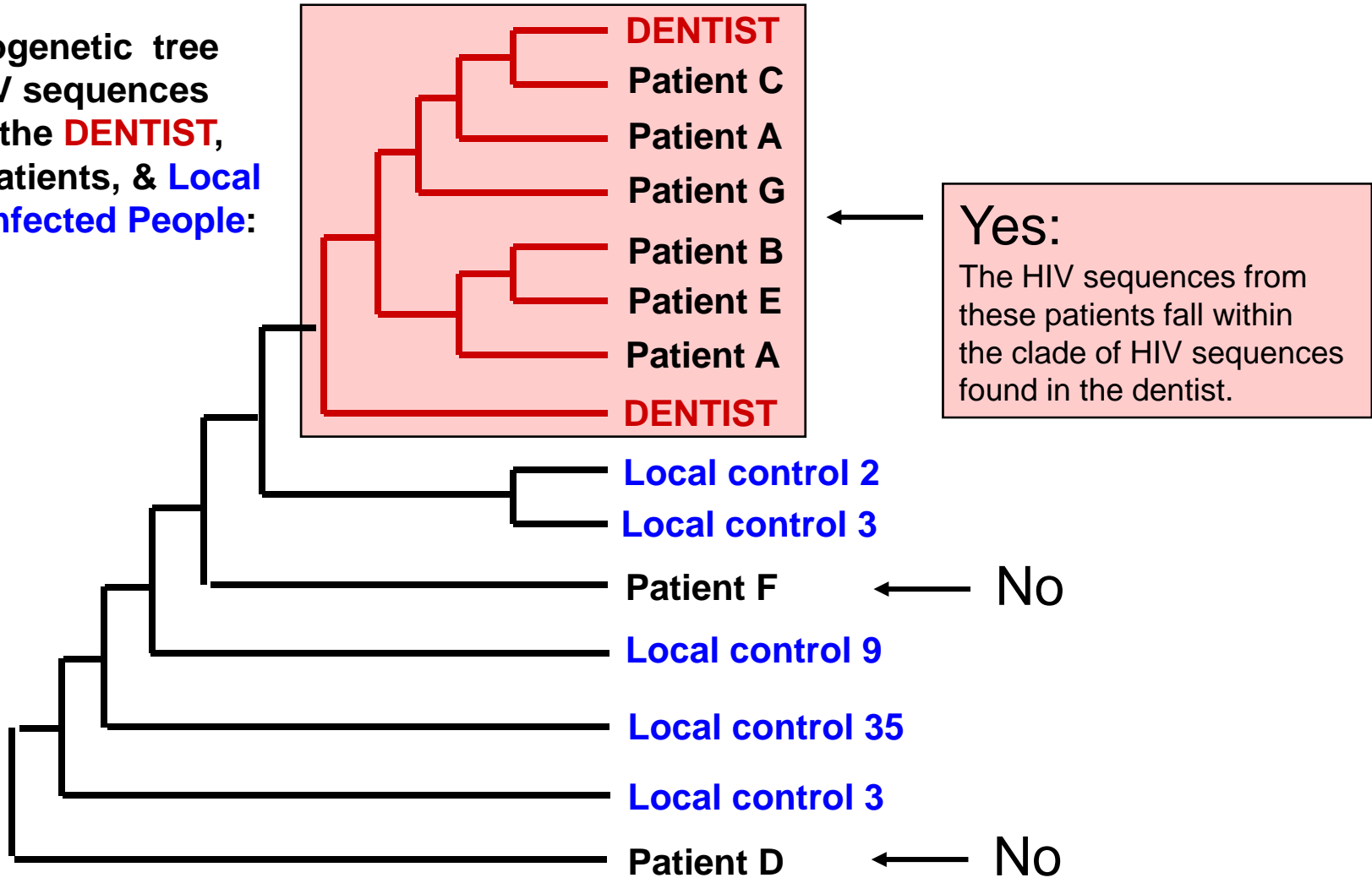
Mitochondrial DNA, most nuclear DNA-encoded genes, and DNA/DNA hybridization all show that bonobos and chimpanzees are related more closely to humans than either are to gorillas.



The pre-molecular view was that the great apes (chimpanzees, gorillas and orangutans) formed a clade separate from humans, and that humans diverged from the apes at least 15-30 MYA.

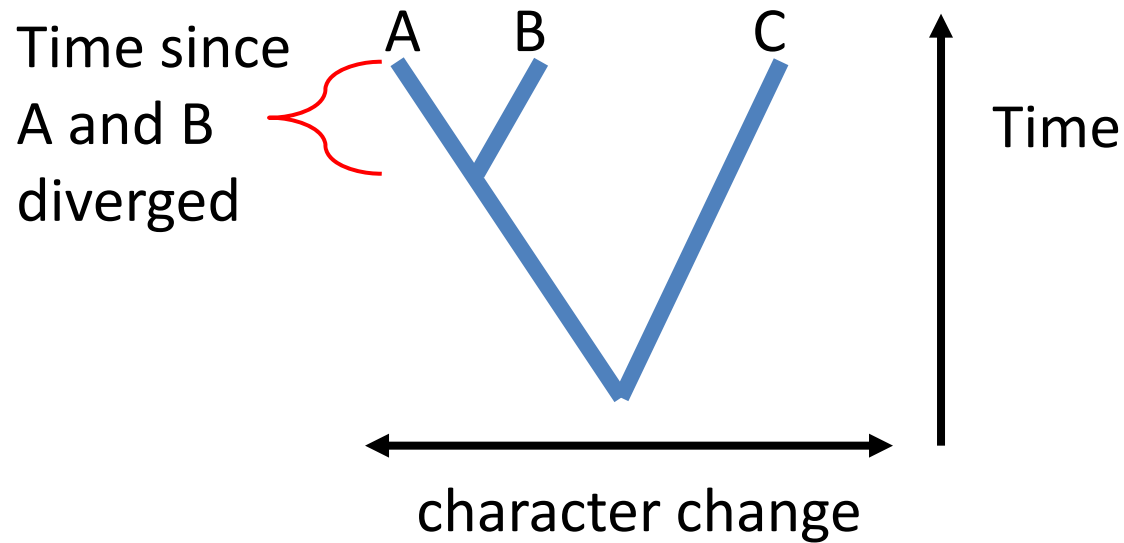
# Did the *Florida Dentist* infect his patients with HIV?

Phylogenetic tree of HIV sequences from the **DENTIST**, his Patients, & **Local HIV-infected People**:



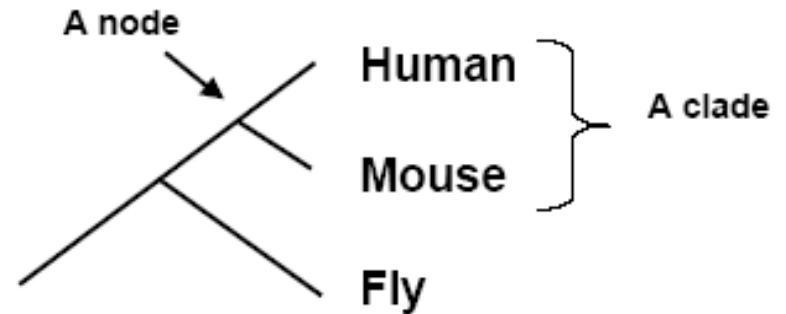
# The structure of phylogenetic trees

A is closer to B than it is to C

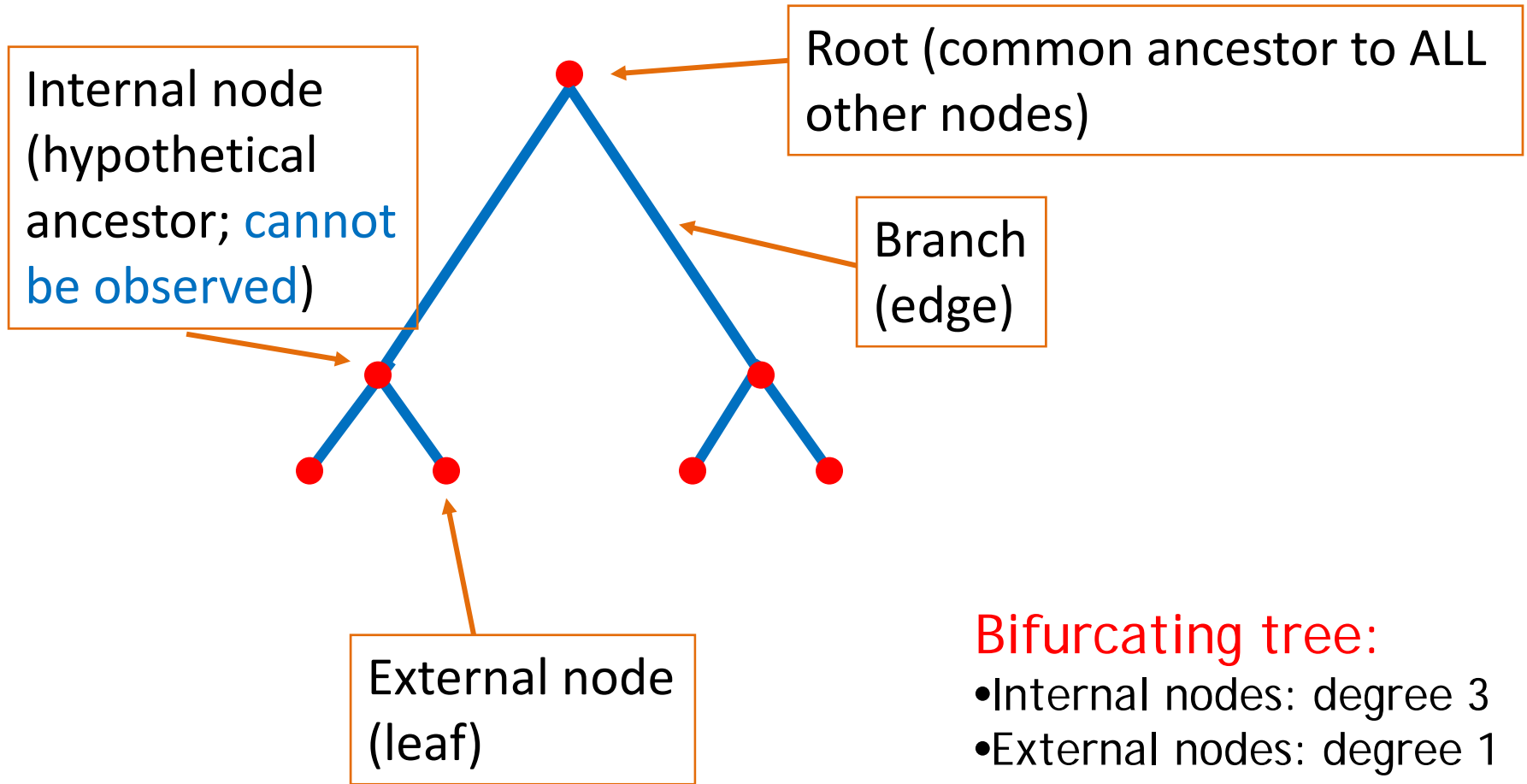


# A Phylogenetic Tree

- **Taxon** -- Any named group of organisms - evolutionary theory not necessarily involved.
- **Clade** -- A monophyletic taxon (i.e., have a **common ancestor**; evolutionary theory utilized)



# The structure of phylogenetic trees



Leafs represent current species or genes

## Bifurcating tree:

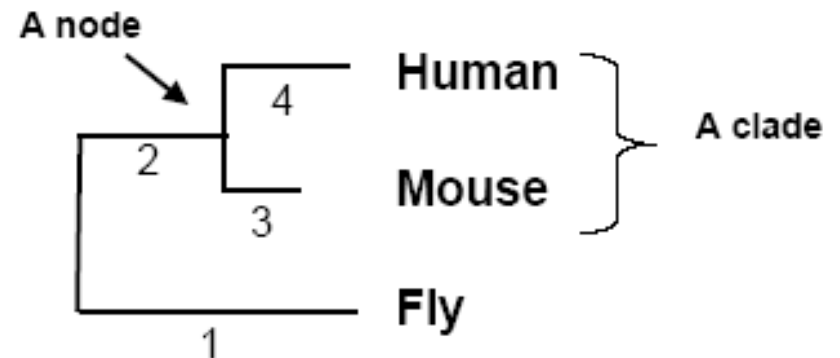
- Internal nodes: degree 3
- External nodes: degree 1
- Root node: degree 2

## Multifurcating trees:

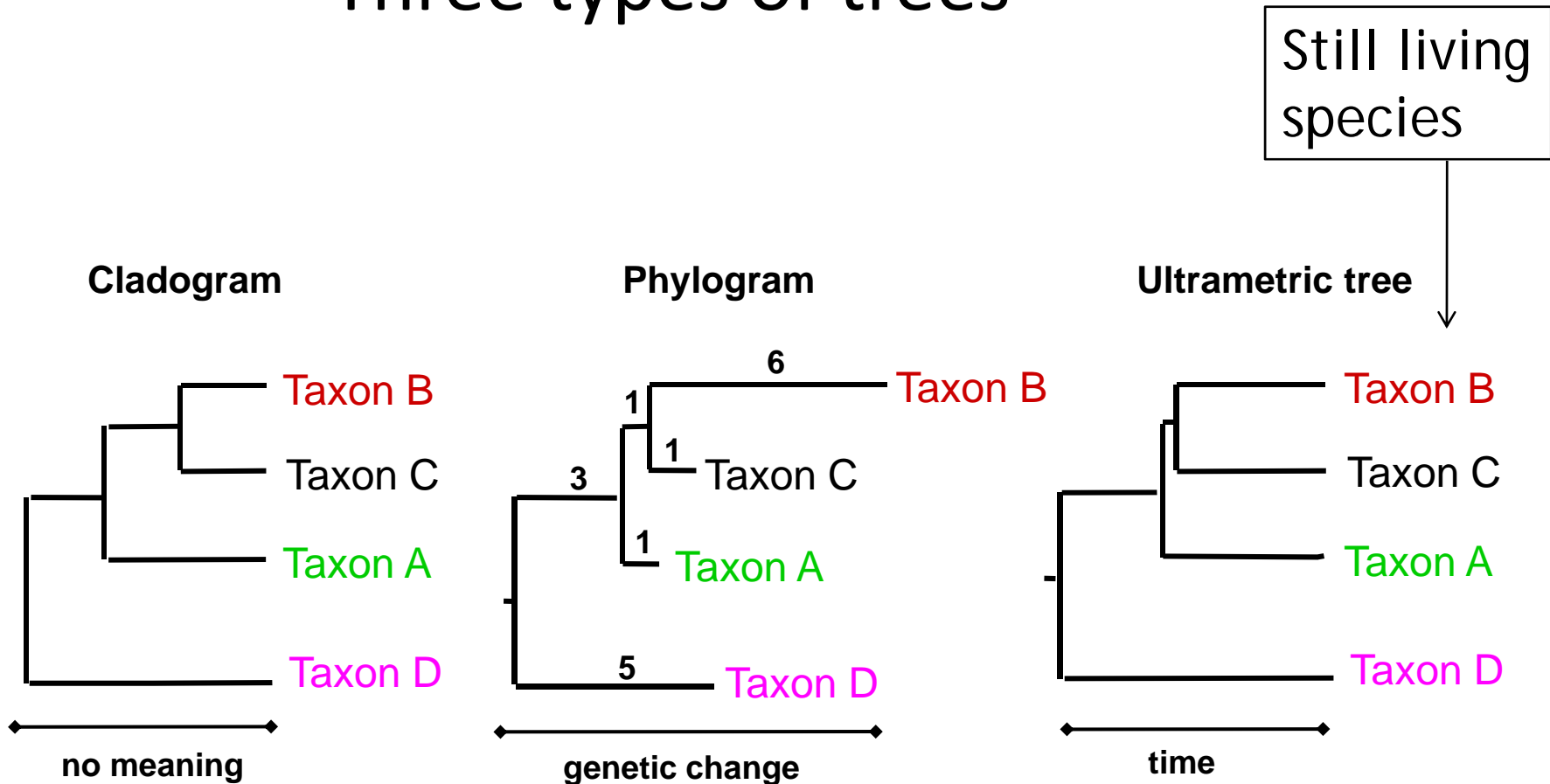
used to represent uncertainty in the order of splitting events

# A phylogenetic tree with branch lengths

- Branch length can be significant.
- In this case it is and mouse is slightly more **similar** to fly than human is to fly (sum of branches  $1+2+3$  is less than sum of  $1+2+4$ )
- But human is more **closely related** (in evolution) to mouse than to fly

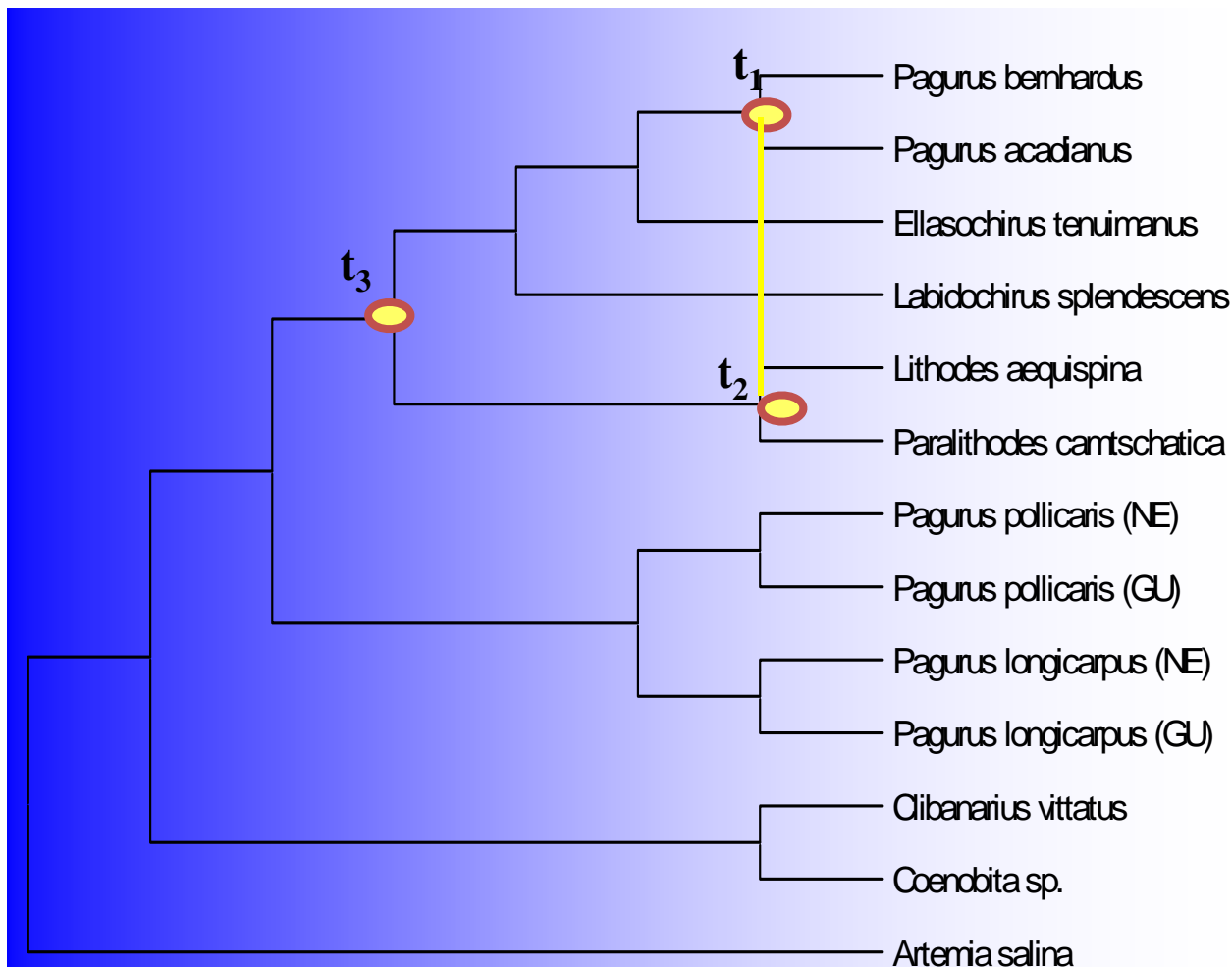


# Three types of trees



All show the same evolutionary relationships, or branching orders, between the taxa.

# Types of trees: cladogram

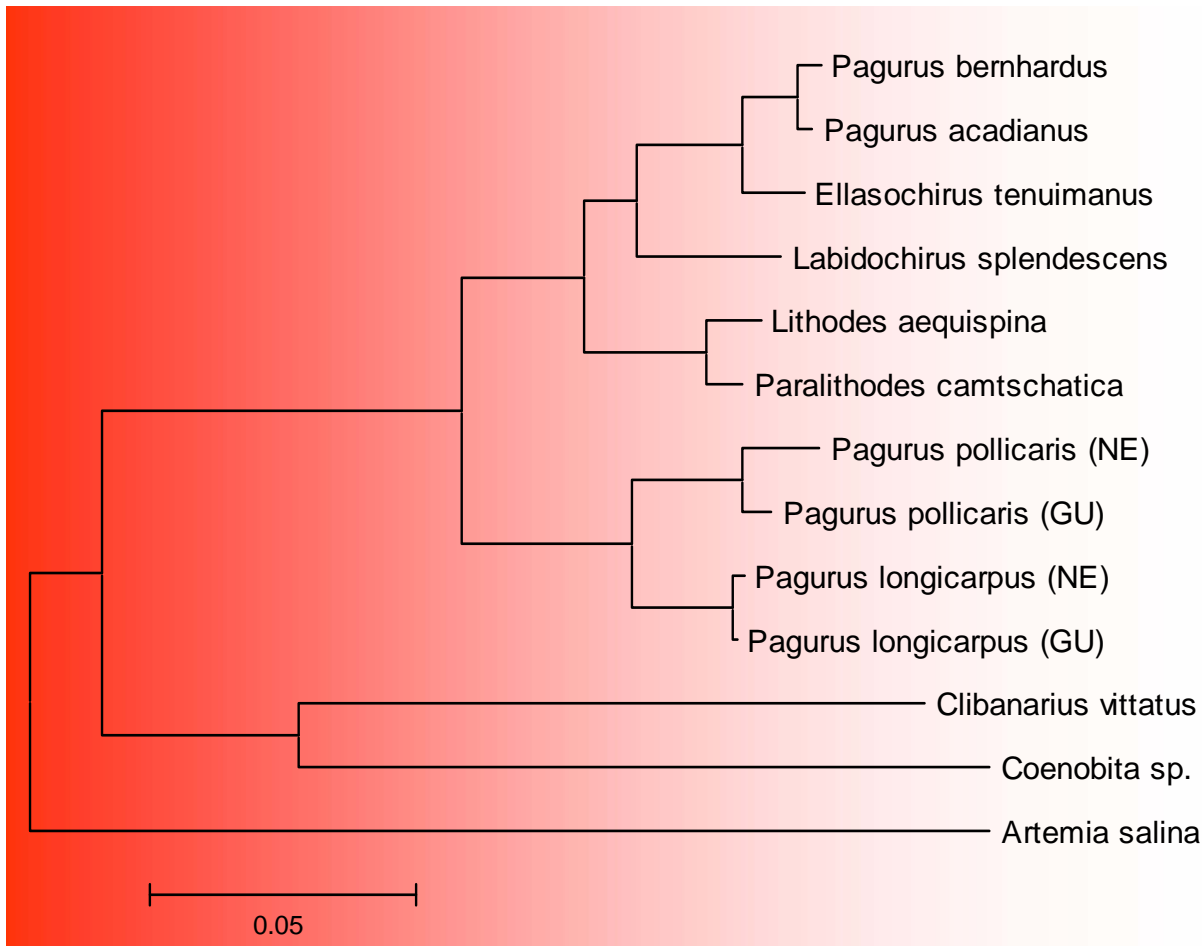


## cladogram

- *Does not* imply that ancestors on the same line necessarily speciated at the same time.
- $t_1$  can be before or after  $t_2$  but not before  $t_3$

(no time scale)

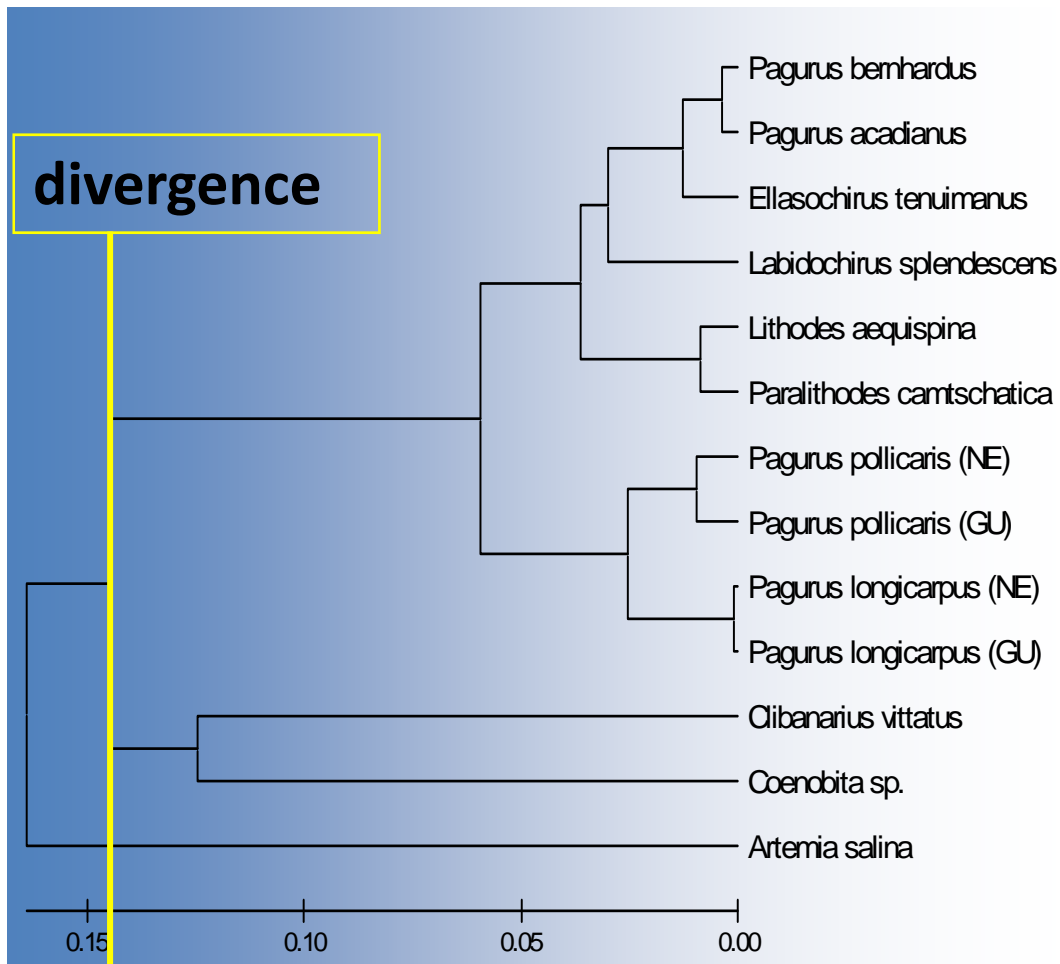
# Types of trees: phylogram



phylogram  
(additive tree:  
branch lengths can  
be summed)

branch lengths =  
amount of change

# Types of trees: ultrametric



**Ultrametric tree  
(linearized tree)**

All tree tips are  
equidistant from  
the root

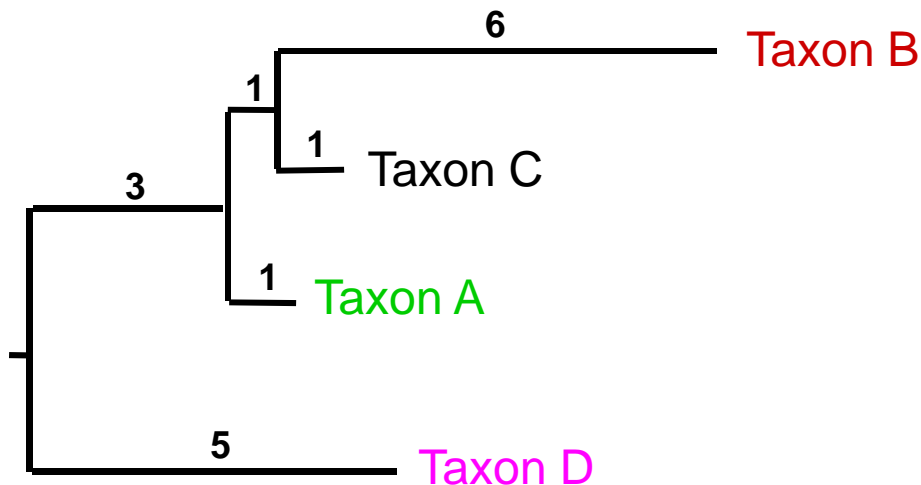
# Similarity vs. Evolutionary Relationship

Similarity and relationship are *not* the same thing, even though evolutionary relationship is inferred from certain types of similarity.

Similar: having likeness or resemblance (an observation)

Related: genetically connected (an historical fact)

Two taxa can be most similar without being most closely-related:



**C is more similar in sequence to A ( $d = 3$ ) than to B ( $d = 7$ ), but C and B are most closely related (that is, C and B shared a common ancestor more recently than either did with A).**

# Rooted and unrooted trees

Trees can either be rooted or unrooted

## Rooted Tree

- ✿ One node identified as root from which ultimately all other roots descent
- ✿ Rooted tree has a direction, which corresponds to evolutionary time

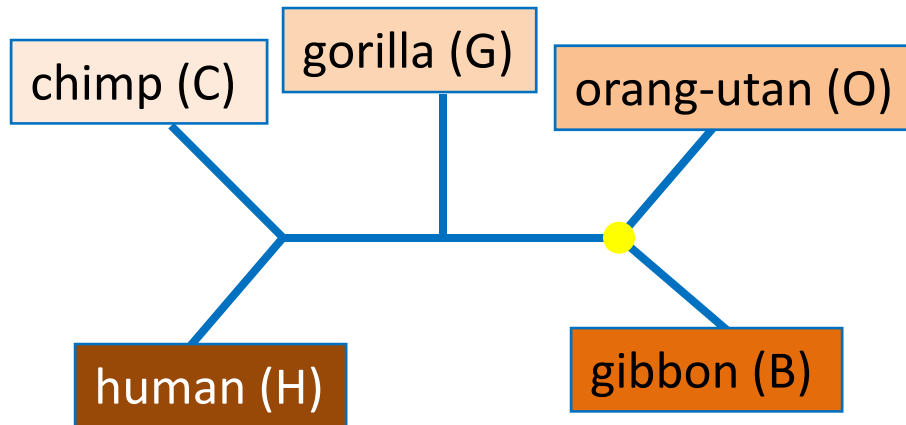
## Unrooted Tree

Do not specify evolutionary relationships in the same way:

- ✿ Do not allow to talk about ancestors and descendants.
- ✿ Adjacent sequences need not to be evolutionary closely related

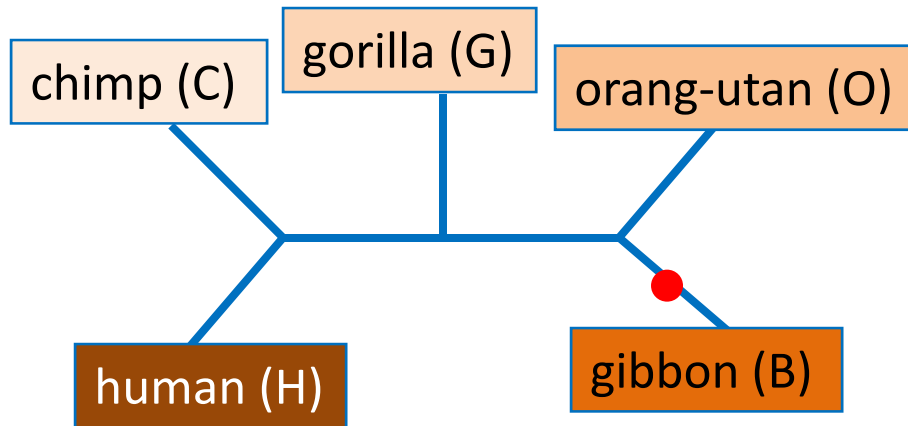
# Rooted and unrooted trees

UNROOTED TREE (no direction)  
Only shows branching pattern

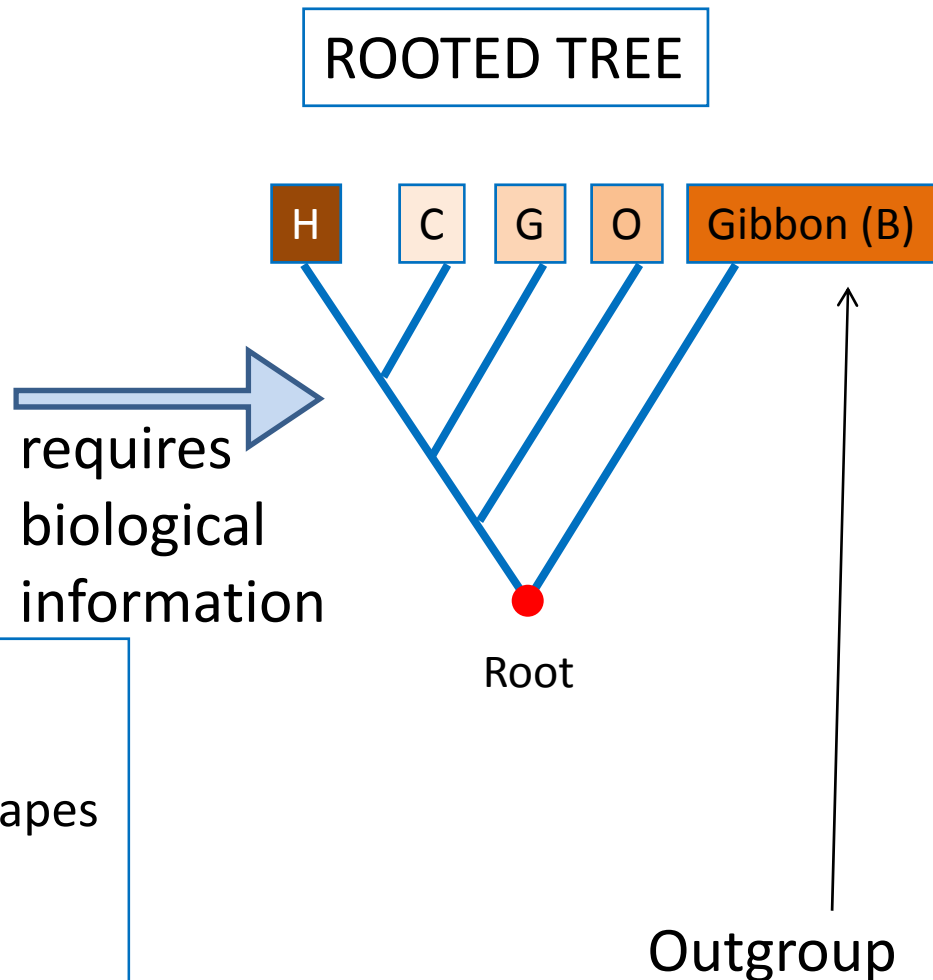


- Orang-utan and Gibbon are neighbors in this tree but
  - 1) not necessarily the closest related
  - 2) Yellow point is not necessarily the common ancestor

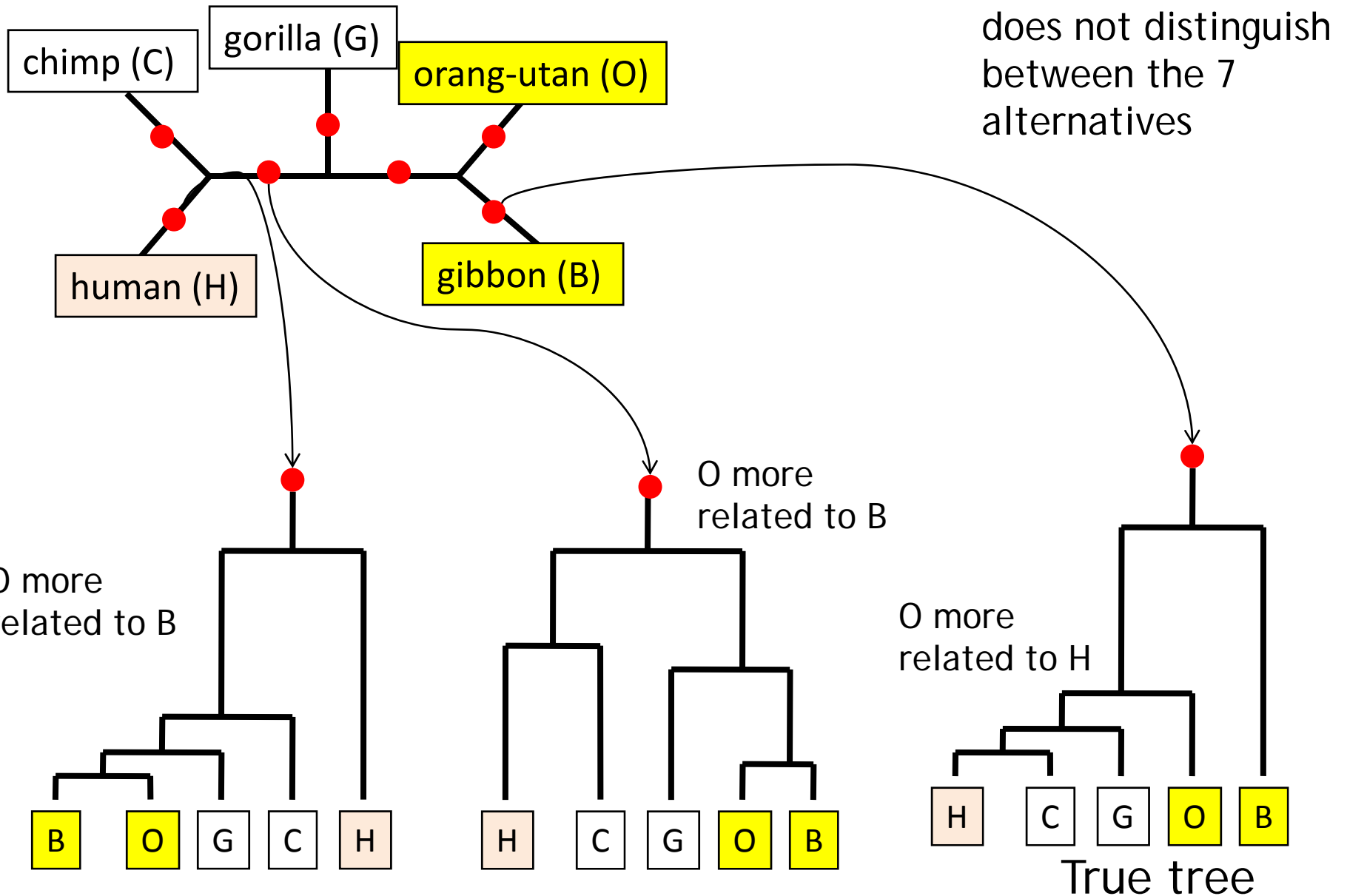
# Rooted and unrooted trees



- Orang-utan and Gibbon are neighbors in this tree but
- Orang-utan is related more closely other apes (including) human than to the Gibbon
- This is because root lies on the branch to Gibbon



# Rooted and unrooted trees



# Number of possible trees

Number of unrooted trees for  $N \geq 3$ ):

$$\frac{(2n - 3)!}{2^{n-2}(n - 2)!}$$

Number of rooted trees for  $N \geq 2$ ):

$$\frac{(2n - 5)!}{2^{n-3}(n - 3)!}$$

Problem: find TRUE tree from all possible trees

Species	Number of rooted trees
2	1
3	3
4	15
5	105
6	945
...	...
15	213,458,046,676,875
...	...
50	$2.76292 \times 10^{76}$

# Inferring trees from data

- ✿ Inferring phylogenetic trees is closely linked to the analysis of DNA as a result of advances in DNA sequencing technologies
- ✿ More recently, it has been based on whole-genome comparisons
- ✿ Unknown tree can be inferred because sequences are always changing
  - ◇ They leave behind a trail of mutations that will be present in the descendants of mutant genes and absent from all other individuals

# The biological basis of evolution

*Mother DNA:* tctgcctc

gatgcctc

tctgcctcggg

gatgcatc

gacgcctc

gctgcctcggg

gatgaatc

gccgcctc

gctaagcctcggg

present species

# Inferring trees from data

- ✿ Inference of trees is only possible because mutation is rare
  - ◇ Therefore, unrelated individuals will have the same sequence very infrequently
- ✿ If a gene or any segment of DNA did not change over time, we would have no record of its past.
- ✿ Within a set of organisms we expect that every gene that they share will lead to the same (or very similar) tree.
  - ◇ Each gene might evolve at different rates but all of the genes will be inherited as a group and will be passed to descendants together, resulting in the same tree.

# Inferring trees from data

## Two groups of methods

- ✿ (1) Rank all possible trees and use some criterion to find the optimal one
  - ◇ Generally, find tree with the smallest number of necessary mutations to explain the data
  - ◇ Given the huge number of trees, these methods may take a long time
  - ◇ May not find the true tree as result of approximations that must be used to speed up the search
  - ◇ Likelihood-based methods are favored for in-depth phylogenetic analysis
- ✿ (2) build the tree from the data (without explicitly stating a scoring function)
  - ◇ Often based on computing pairwise distance between taxa
  - ◇ Very fast
  - ◇ Not as well behaved statistically as other methods

# Molecular phylogenetic tree building methods

Mathematical and/or statistical methods for inferring the divergence order of taxa, as well as the lengths of the branches that connect them. There are many phylogenetic methods available today, each having strengths and weaknesses. [Most can be classified as follows:](#)

		COMPUTATIONAL METHOD	
		Optimality criterion	Clustering algorithm
DATA TYPE	Characters	<b>PARSIMONY</b> <b>MAXIMUM LIKELIHOOD</b>	
	Distances	<b>MINIMUM EVOLUTION</b> <b>LEAST SQUARES</b>	<b>UPGMA</b> <b>NEIGHBOR-JOINING</b>

# Types of data used in phylogenetic inference

**Character-based methods:** Use the aligned characters, such as DNA or protein sequences, directly during tree inference.

Taxa	Characters
Species A	ATGGCTATTCTTATAGTACG
Species B	ATCGCTAGTCCTTATATTACA
Species C	TTCACTAGACCTGTGGTCCA
Species D	TTGACCAGACCTGTGGTCCG
Species E	TTGACCAGTTCTCTAGTTTCG

**Distance-based methods:** Transform the sequence data into pairwise distances (dissimilarities), and then use the matrix during tree building.

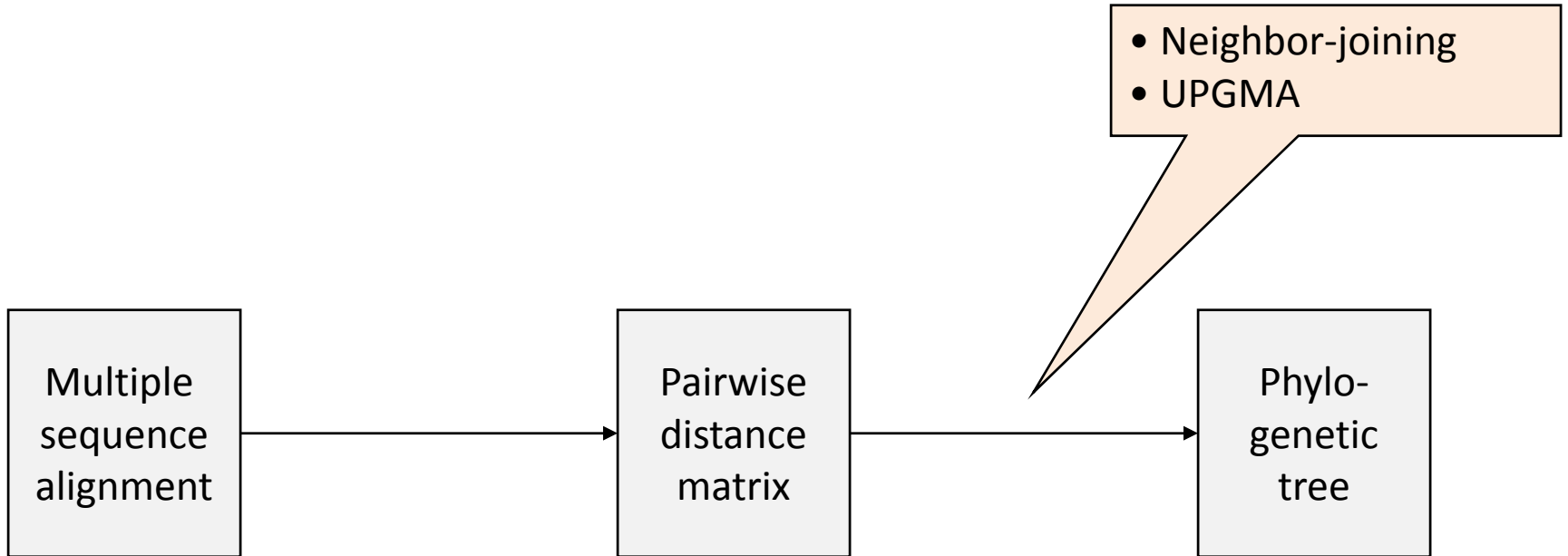
	A	B	C	D	E
Species A	----	0.20	0.50	0.45	0.40
Species B	0.23	----	0.40	0.55	0.50
Species C	0.87	0.59	----	0.15	0.40
Species D	0.73	1.12	0.17	----	0.25
Species E	0.59	0.89	0.61	0.31	----

← Example 1:  
Uncorrected  
“p” distance  
(=observed percent  
sequence difference)



Example 2: Kimura 2-parameter distance  
(estimate of the true number of substitutions between taxa)

# Distance methods



**METRIC DISTANCES** between any two or three taxa (a, b, and c) have the following properties:

**Property 1:**  $d(a, b) \geq 0$

**Non-negativity**

**Property 2:**  $d(a, b) = d(b, a)$

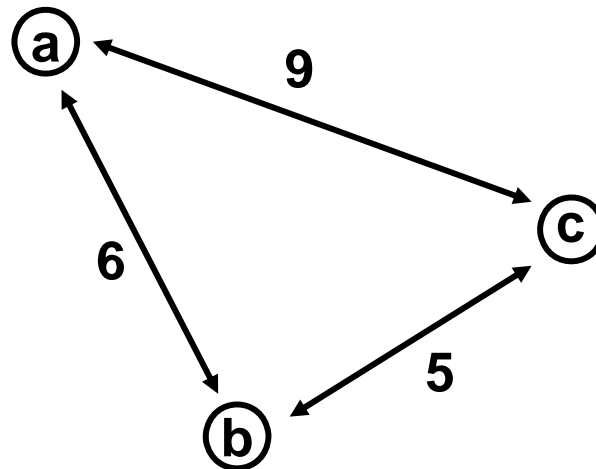
**Symmetry**

**Property 3:**  $d(a, b) = 0$  if and only if  $a = b$

**Distinctness**

**Property 4:**  $d(a, c) \leq d(a, b) + d(b, c)$

**Triangle inequality:**

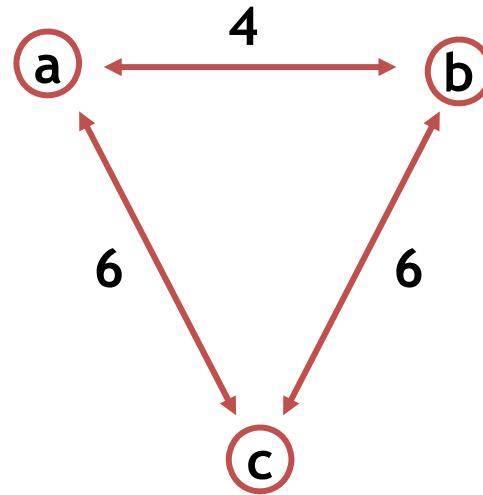


# ULTRAMETRIC DISTANCES

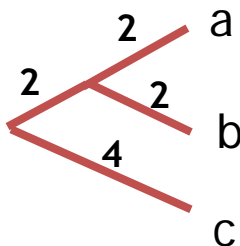
must satisfy the previous four conditions, plus:

**Property 5**  $d(a, b) \leq \text{maximum} [d(a, c), d(b, c)]$

This implies that the two largest distances are equal



**Similarity = Relationship if the distances are ultrametric!**

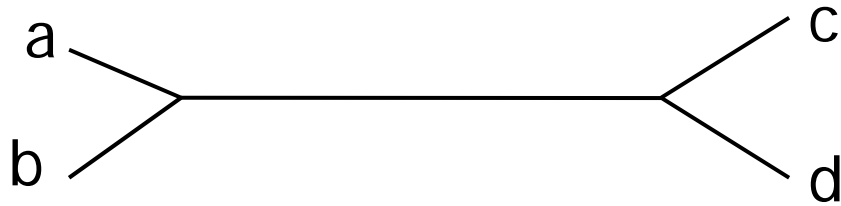


If distances are ultrametric, then the sequences are evolving in a perfectly clock-like manner, thus can be used in [UPGMA trees](#) and for the most precise calculations of divergence dates.

## ADDITIVE DISTANCES:

### Property 6:

$$d(a, b) + d(c, d) \leq \text{maximum} [d(a, c) + d(b, d), d(a, d) + d(b, c)]$$



For distances to fit into an evolutionary tree, they must be either metric or ultrametric, and they must be additive.

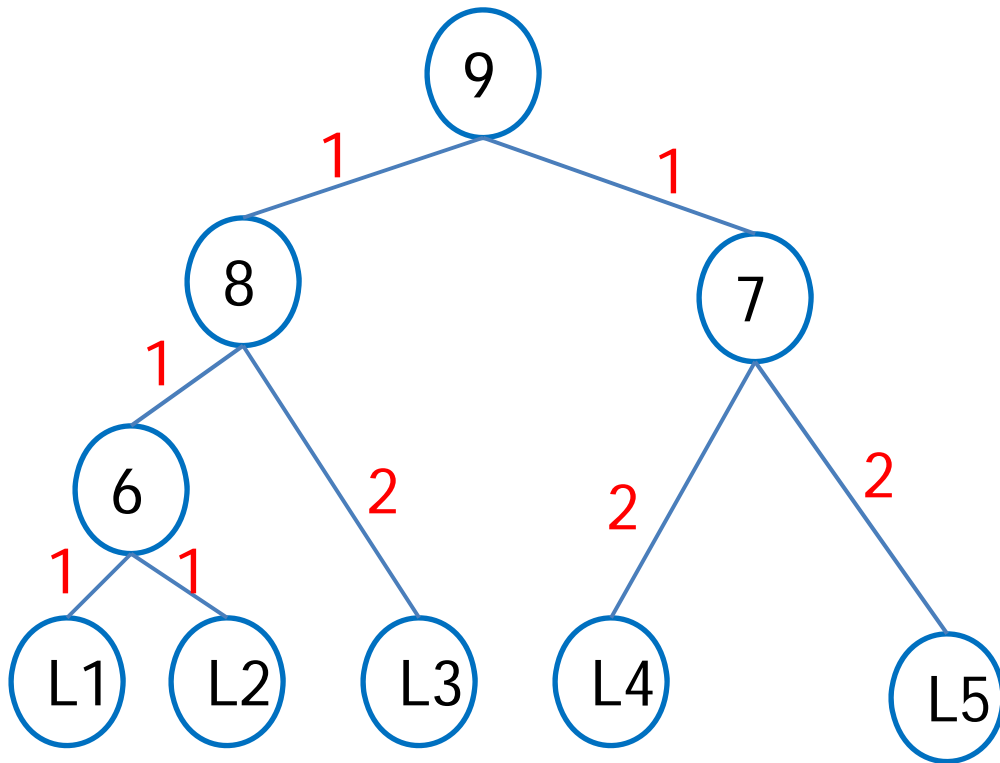
Estimated distances often fall short of these criteria, and thus can fail to produce correct evolutionary trees.

# Inferring trees from distance data

- ✿ Set of  $n$  taxa  $\{t_1, \dots, t_n\}$
- ✿ matrix  $D$  of pairwise genetic distances + JC-correction for multiple substitutions
- ✿ **Additive** distances: distance over path from  $i \rightarrow j$  is:  
 $d(i, j)$

# Additivity and distance matrices

- ✿ If branches within a tree each have a specified length then the distance between any two nodes can easily be computed: **length of path**



	L1	L2	L3	L4	L5
L1	0	2	4	6	6
L2		0	4	6	6
L3			0	6	6
L4				0	4
L5					0

Distance matrix is '**additive**', i.e., distance can be represented by a tree

# Additivity and distance matrices

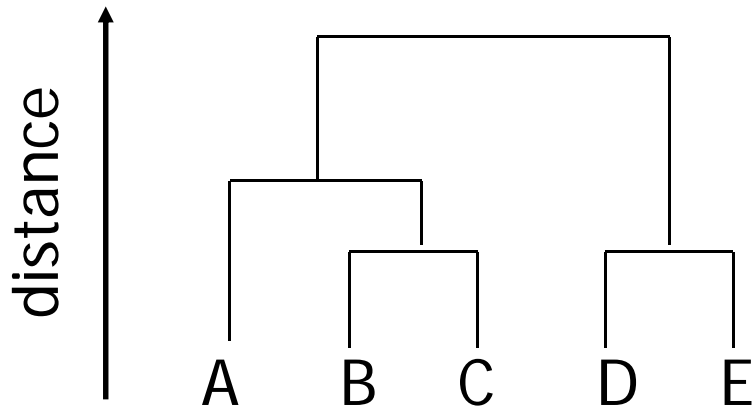
- ✿ Biological interpretation: additivity is the actual number of substitutions separating two taxa from their last **common** ancestor
- ✿ **Jukes-Cantor:**
  - ◇ Attempt to make distance matrix more additive
  - ◇ This results in easier inference of tree

# UPGMA

- ✿ Unweighted Pair Group Method with Arithmetic mean.
- ✿ Also known as agglomerative hierarchical clustering.
- ✿ Basic idea: iteratively connect the two most closely related sequences.
- ✿ **Assumes that all taxa have constant evolutionary rates**
  - ◇ **Molecular clock assumption**

# Hierarchical Clustering

Group elements in a tree-like structure



The more similar objects are, the shorter the path



# Hierarchical Clustering

**Algorithm** (agglomerative clustering)

- Start: all objects in a separate cluster
- Clustering: combine the 2 clusters with the shortest distance

Distance between clusters is:

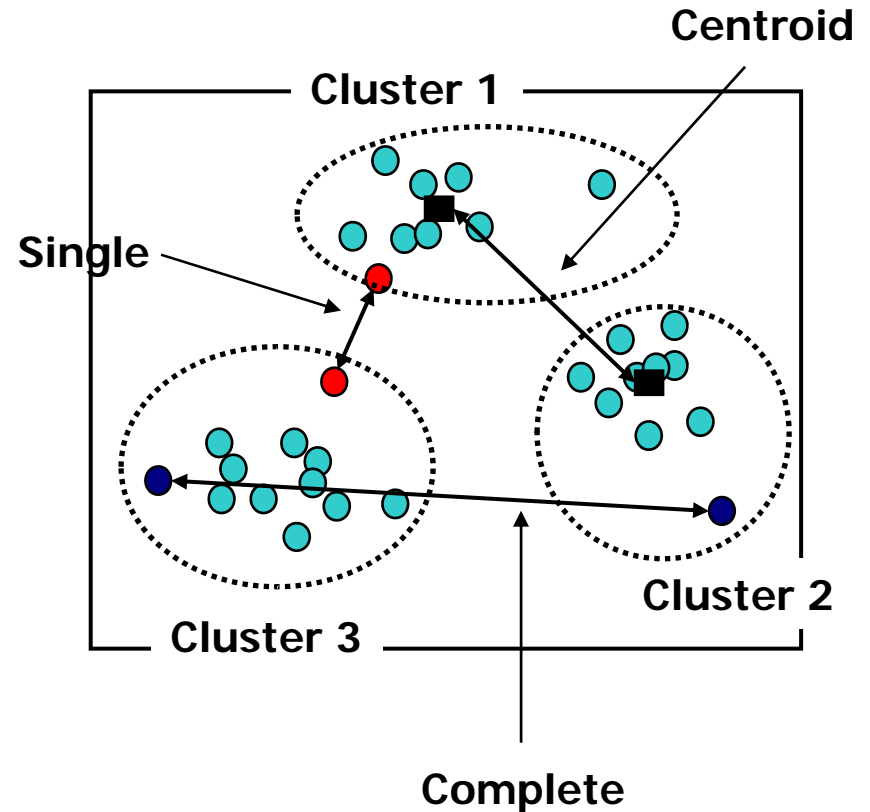
single, complete, centroid, average, median, Ward, ...

- Repeat till only 1 cluster is left

# Combining Clusters

Distance between two clusters is:

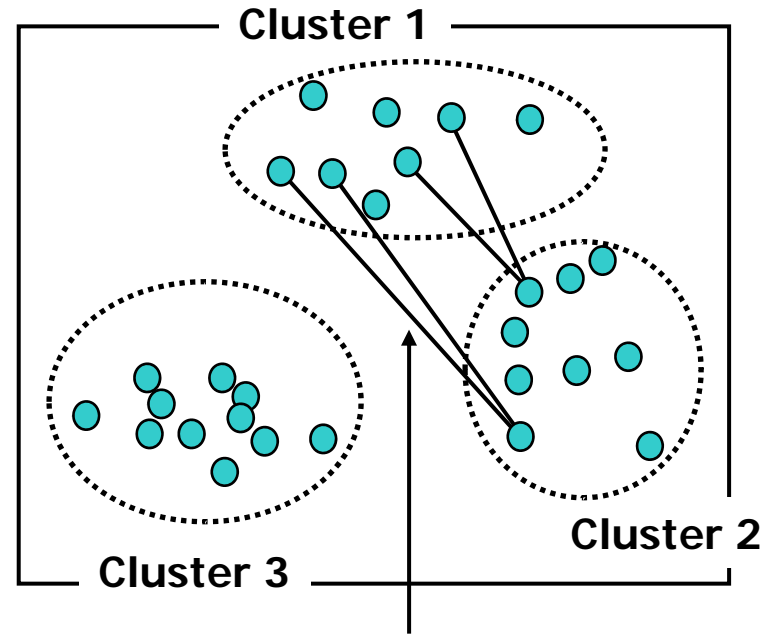
- **Single**  
distance between two closest cluster members
- **Complete**  
distance between two most distant cluster members
- **Centroid**  
distance between means of each cluster



# Combining Clusters

Distance between two clusters is:

- **Average**  
average distance between all members of the two clusters
- **Median**  
median distance between all members of the two clusters
- **Ward**  
average distance between all members of the two clusters with adjustment for covariance



**Mean/median/adjusted mean  
of all pairwise distances**

# Hierarchical Clustering: Example

Distance matrix:

array \ gene	A	B	C	D	E		array	A	B	C	D
gene1	5	3	3	-1	0	→	B	8			
gene2	3	1	2	-3	-4		C	7	3		
gene3	2	0	-1	-3	-3		D	22	14	15	
gene4	1	-1	0	-4	-4		E	22	14	15	2

Manhattan distance:

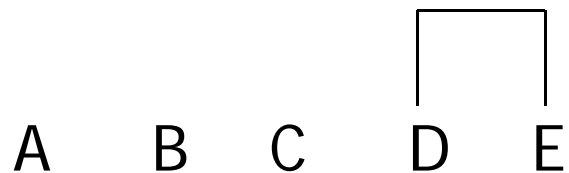
$$\begin{aligned} \text{dist}(A,B) &= |5-3| + |3-1| + |2-0| + |1+1| \\ &= 8 \end{aligned}$$

# Hierarchical Clustering: Example

Distance matrix:

array	A	B	C	D
B	8			
C	7	3		
D	22	14	15	
E	22	14	15	2

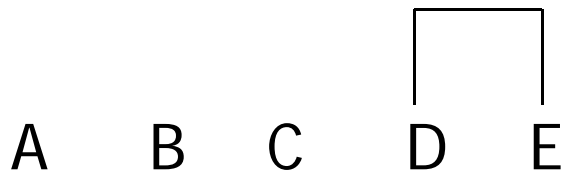
↑  
minimum



tree (or dendrogram)

# Hierarchical Clustering: Example

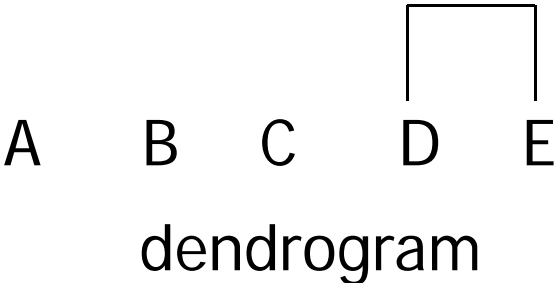
array	A	B	C	D
B	8			
C	7	3		
D	22	14	15	
E	22	14	15	2



dendrogram

# Hierarchical Clustering: Example

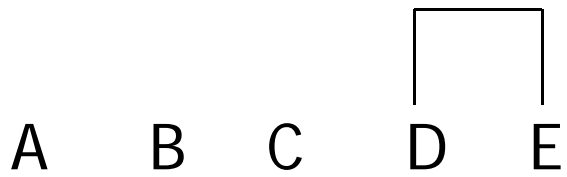
array	A	B	C	D		array	A	B	C
B	8					B			
C	7	3			→	C			
D	22	14	15			DE			
E	22	14	15	2					



# Hierarchical Clustering: Example

array	A	B	C	D	array	A	B	C
B	8				B	8		
C	7	3			C	7	3	
D	22	14	15		DE	22		
E	22	14	15	2				

$$\text{dist}(A, DE) = \min(\text{dist}(A, D), \text{dist}(A, E)) = 22$$

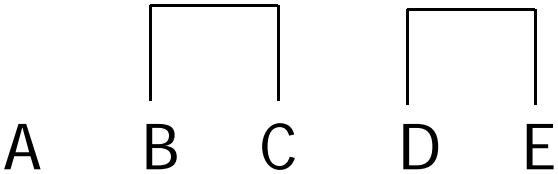


dendrogram

Single linkage: minimum distance

# Hierarchical Clustering: Example

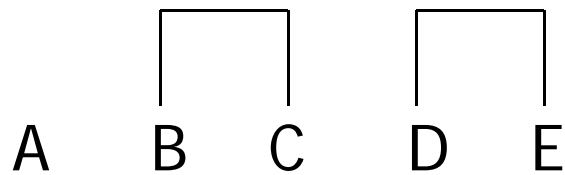
array	A	B	C	D	array	A	B	C
B	8				B	8		
C	7	3			C	7	3	
D	22	14	15		DE	22	14	15
E	22	14	15	2				



dendrogram

# Hierarchical Clustering: Example

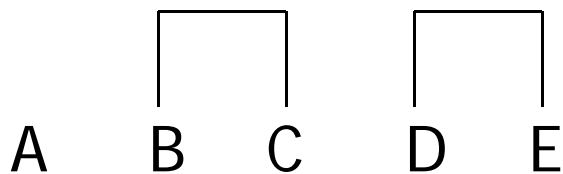
array	A	B	C
B	8		
C	7	3	
DE	22	14	15



dendrogram

# Hierarchical Clustering: Example

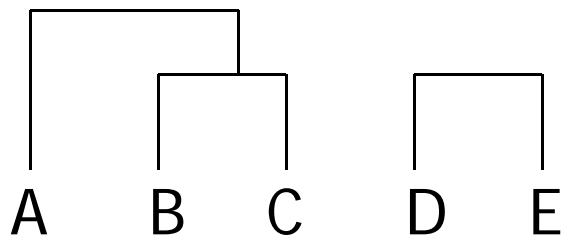
array	A	B	C		array	A	BC
B	8			→	BC	7	
C	7	3			DE	22	14
DE	22	14	15				



dendrogram

# Hierarchical Clustering: Example

array	A	B	C		array	A	BC
B	8			→	BC	7	
C	7	3			DE	22	14
DE	22	14	15				



dendrogram

# Hierarchical Clustering: Example

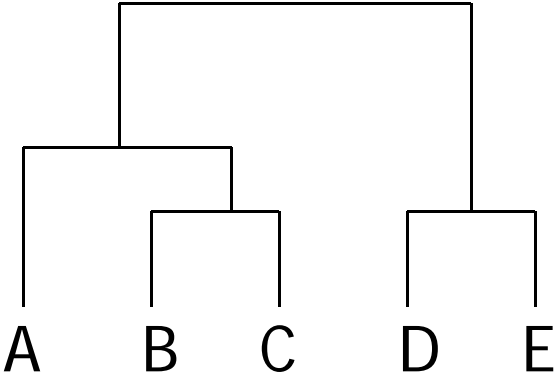
array	A	B	C
B	8		
C	7	3	
DE	22	14	15



array	A	BC
BC	7	
DE	22	14



array	ABC
DE	14



dendrogram



# Inconsistency

- ✿ UPGMA is inconsistent for data that isn't ultrametric (clock-like).
- ✿ Next we'll look at a method that is consistent for any additive data.

# Neighbor Joining (NJ) method

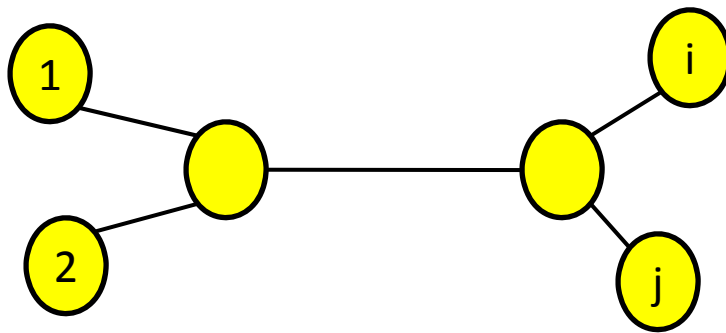
- ✿ NJ is similar to UPGMA
- ✿ NJ does not assume molecular clock
- ✿ Produces **unrooted trees**
- ✿ It corrects for unequal evolutionary rates between sequences by using a conversion step
- ✿ NJ tree construction process is opposite to that used in UPGMA
  - ◇ Start with a completely unresolved **star tree**
  - ◇ Subsequently, **decompose tree** based on corrected distances
  - ◇ Taxa with shortest corrected distances are joint first

# Neighbor Joining (NJ) method

Use property of additive distances to find neighbors in the underlying tree.

If 1-2 and i-j are two pairs of neighbors then:

$$d(1,2) + d(i,j) < d(i,1) + d(2,j)$$



# Neighbor Joining (NJ) method

This leads to a criterion for detecting neighbors when there are an arbitrary number of external nodes in the tree

First define total distance from taxon  $t_i$  to all other taxa:

$$R_i = \sum_j d(t_i, t_j)$$

Define 'closeness' of neighbors (minimizes the distance between external nodes and total distance in the tree):

$$M(i, j) = (n-2)d(i, j) - R_i - R_j$$

For two nodes that are neighbors:

$$M(i, j) < M(i, k) \text{ for all } k \text{ not equal to } j$$

# Neighbor Joining (NJ) method

This gives crucial piece of NJ algorithm:

From the distance matrix (containing all pair-wise distances) compute a new table of values for  $M(i,j)$ .

Then join the pair of taxa with the smallest  $M(i,j)$  (this is called the 4-point condition)

Subsequently,

- \* NJ merges neighbors in a new node
- \* Distance between  $V$  and other nodes is calculated using the 3-point formula

# Neighbor Joining (NJ) method

Three-point formula:

$$L_x + L_y = d_{AB}$$

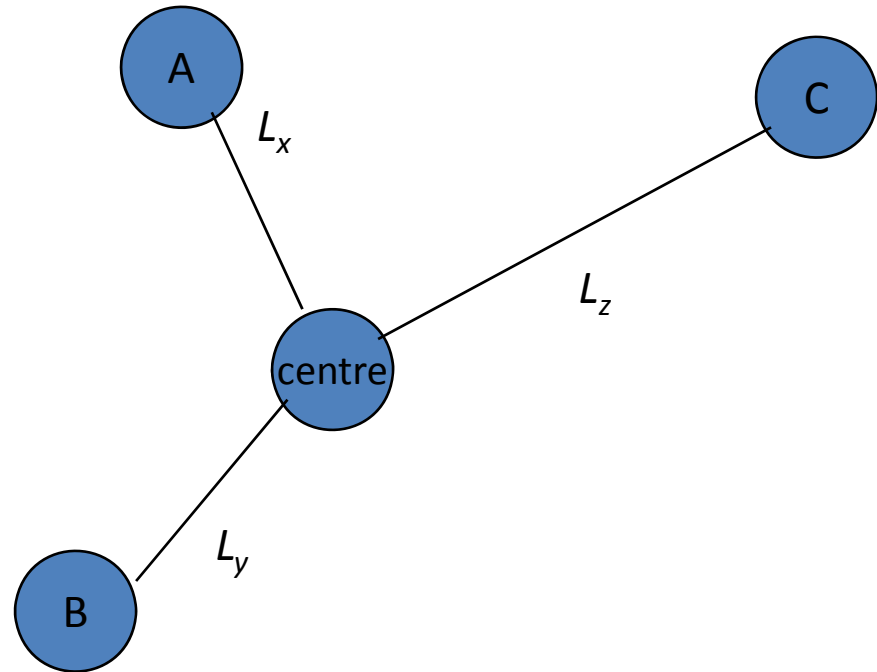
$$L_x + L_z = d_{AC}$$

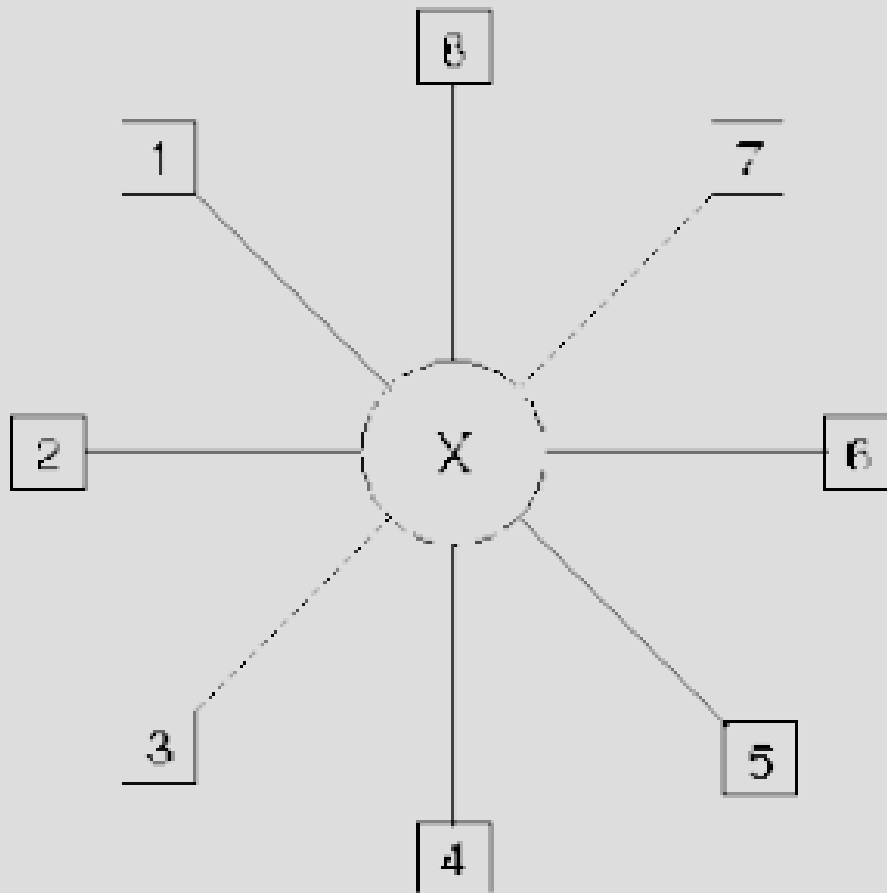
$$L_y + L_z = d_{BC}$$

$$L_x = (d_{AB} + d_{AC} - d_{BC}) / 2$$

$$L_y = (d_{AB} + d_{BC} - d_{AC}) / 2$$

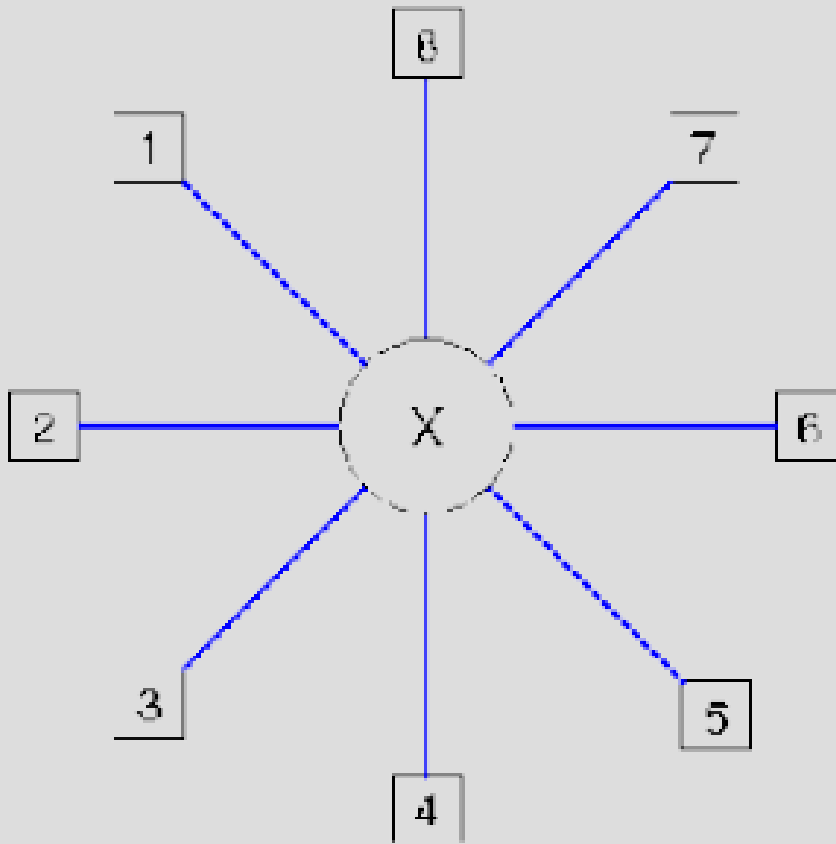
$$L_z = (d_{AC} + d_{BC} - d_{AB}) / 2$$





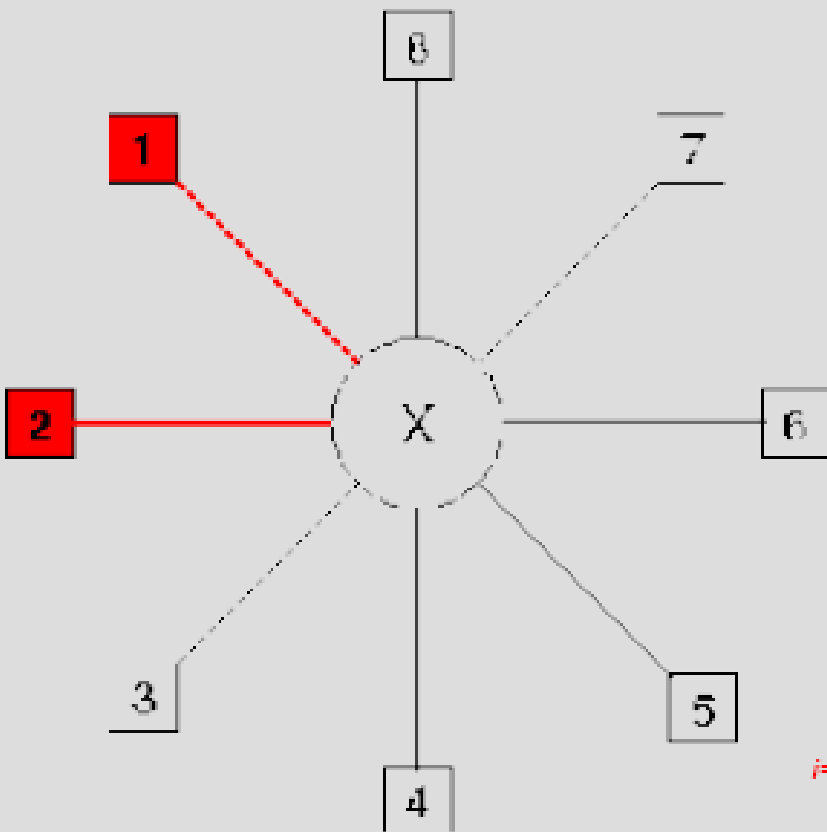
Dij=

0							
7	0						
8	5	0					
11	8	5	0				
13	10	7	8	0			
16	13	10	11	5	0		
13	10	7	8	6	9	0	
17	14	11	12	10	13	8	0



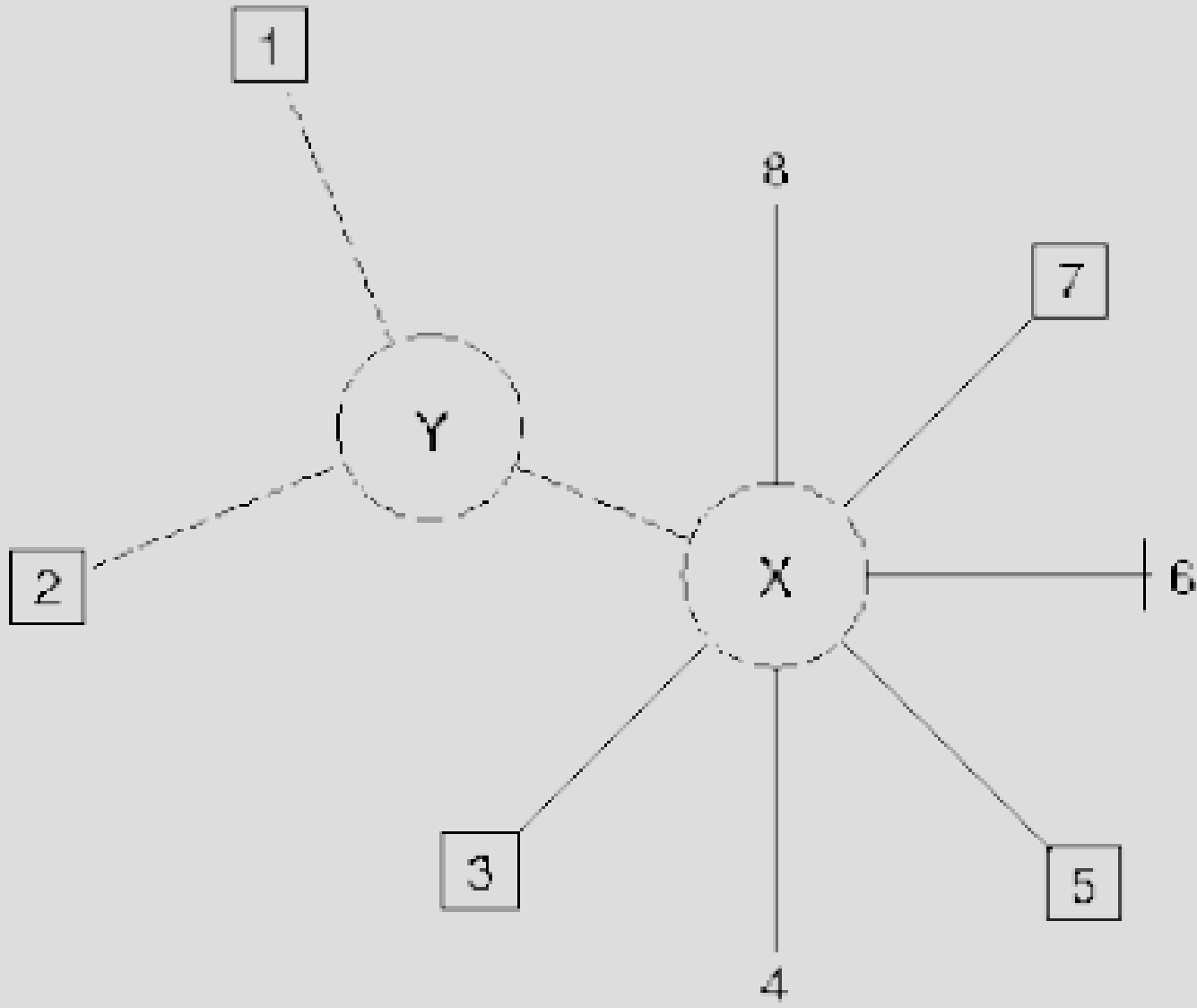
$S_{ij} =$

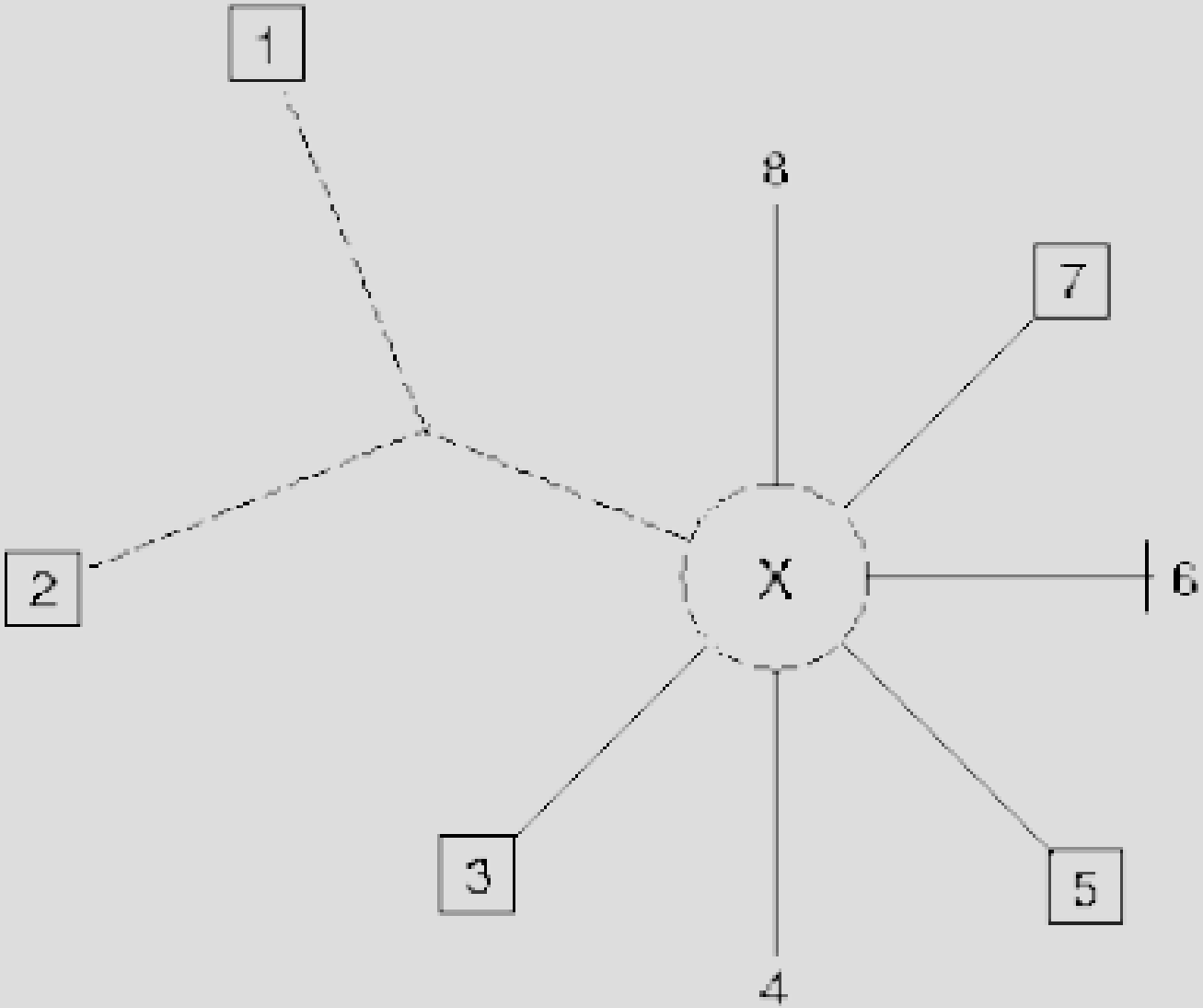
	0							
36.67	0							
38.33	38.33	0						
39.00	39.00	38.67	0					
40.33	40.33	40.00	39.67	0				
40.33	40.33	40.00	39.67	37.00	0			
40.17	40.17	39.83	39.50	38.83	38.83	0		
40.17	40.17	39.83	39.50	38.83	38.83	37.57	0	

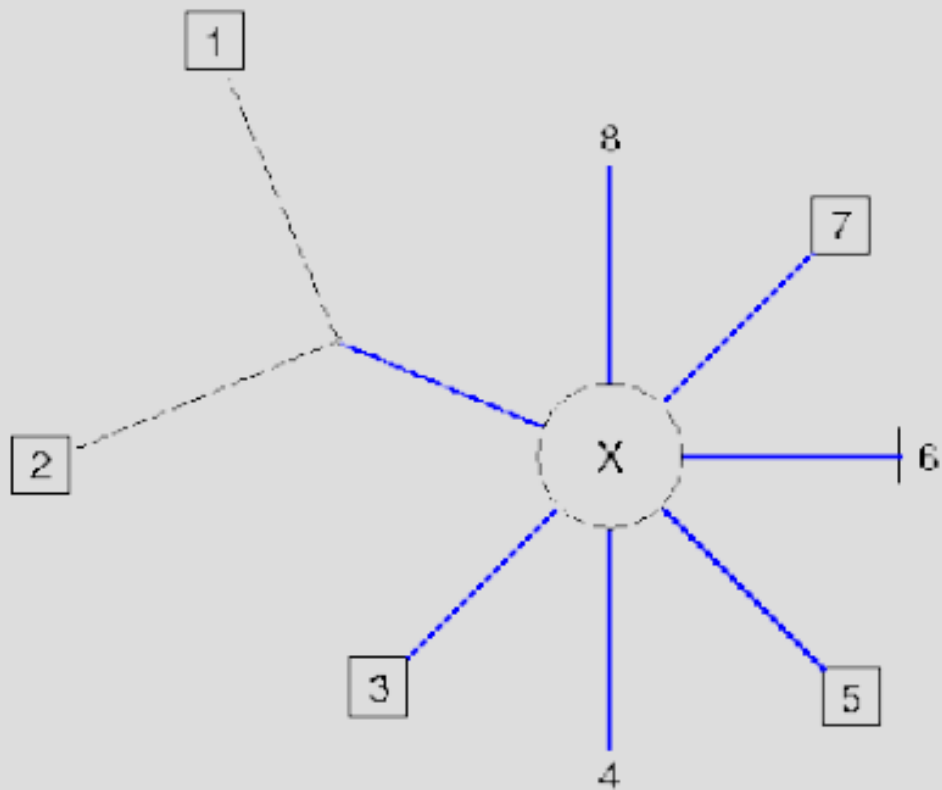


$S_{ij} =$

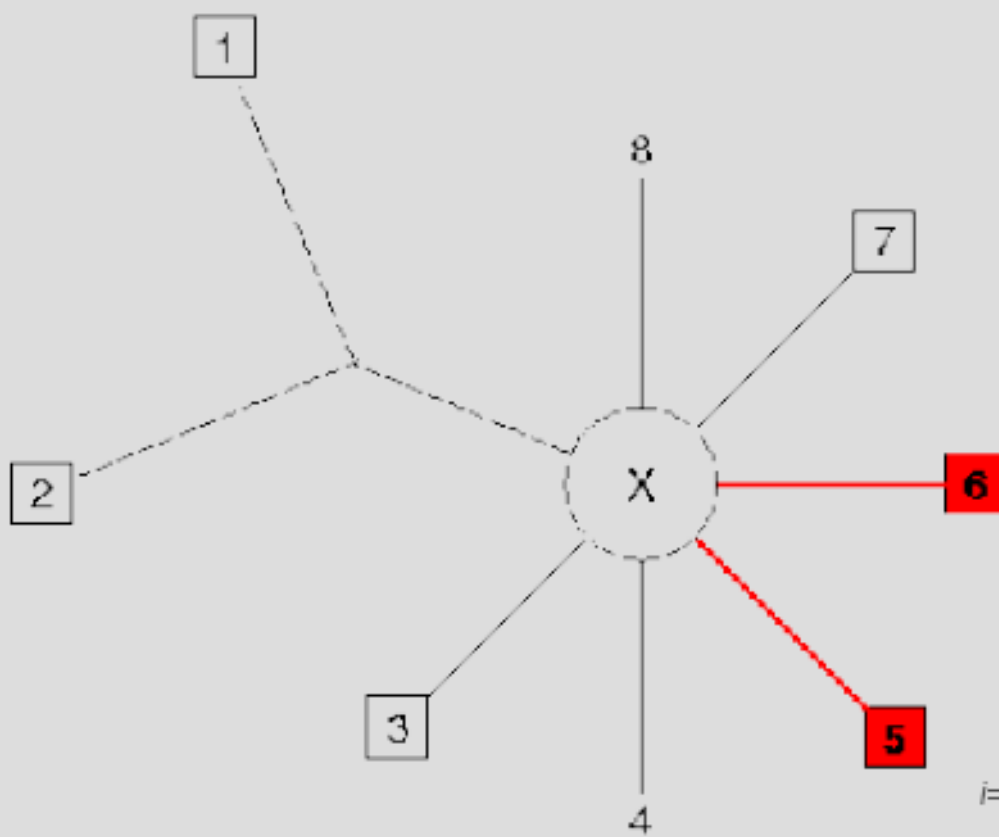
$i=1$	2	3	4	5	6	7	8	$j=1$
0								0
<b>36.67</b>	0							<b>2</b>
38.33	38.33	0						3
39.00	39.00	38.67	0					4
40.33	40.33	40.00	39.67	0				5
40.33	40.33	40.00	39.67	37.00	0			6
40.17	40.17	39.83	39.50	38.83	38.83	0		7
40.17	40.17	39.83	39.50	38.83	38.83	37.57	0	8



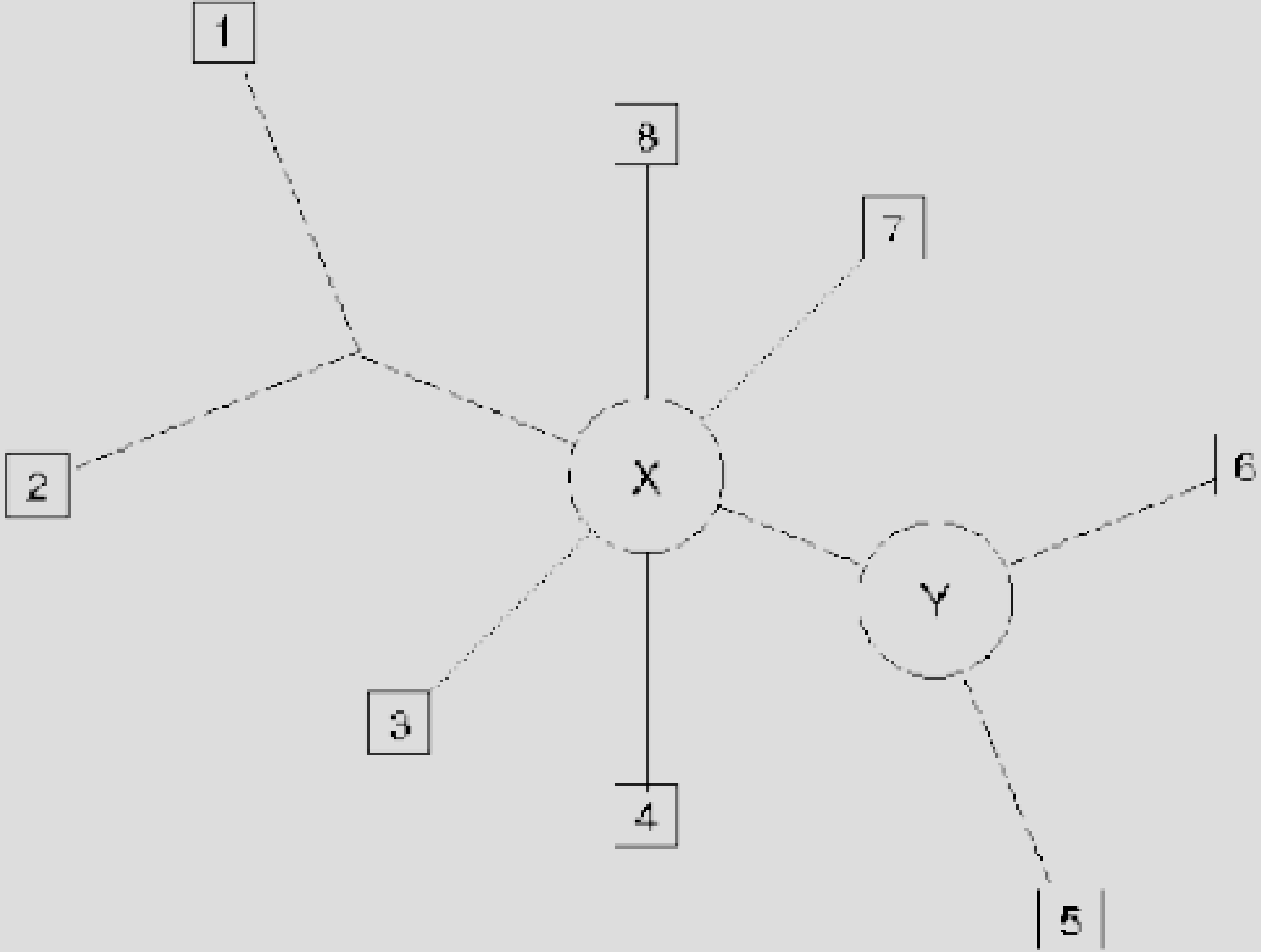


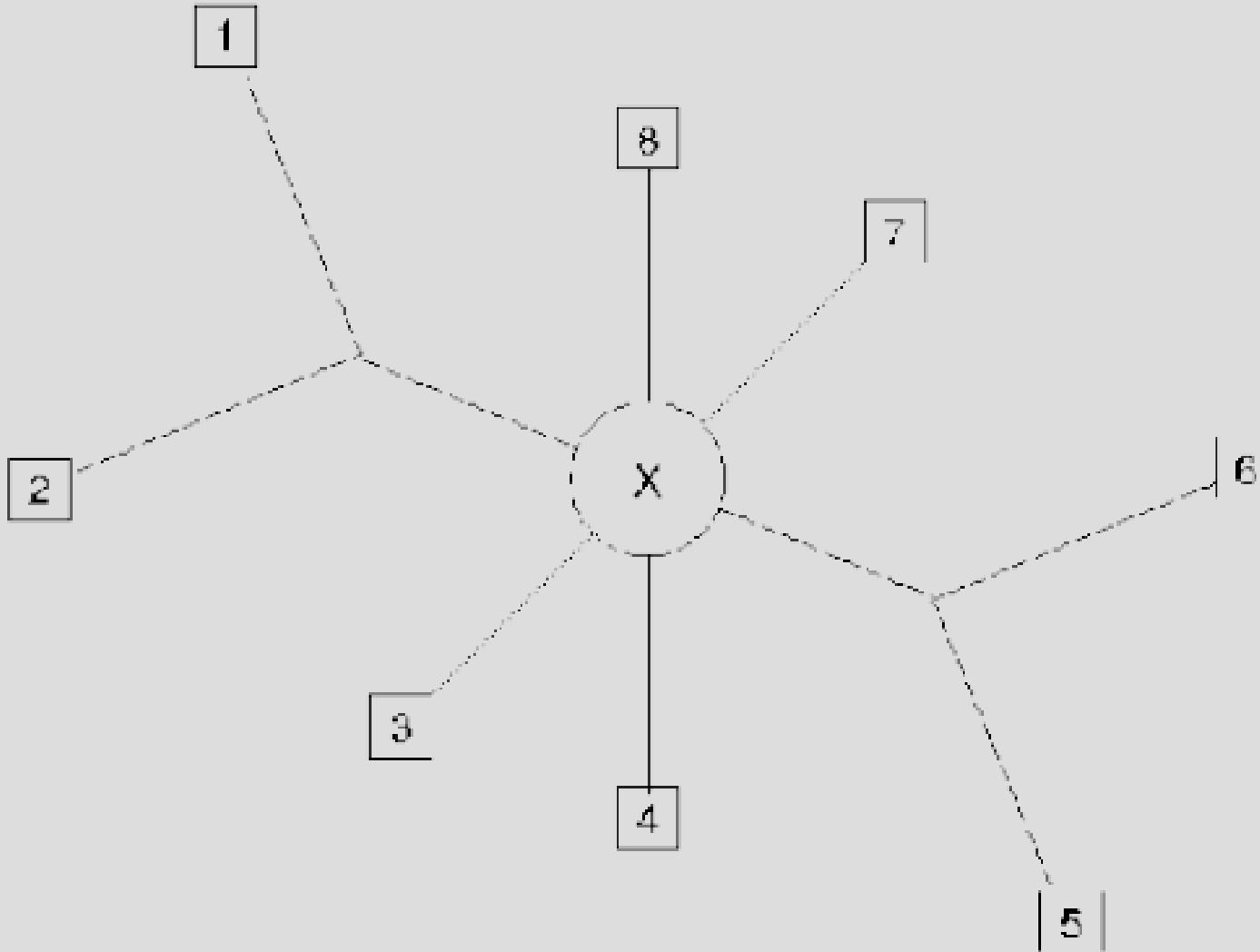


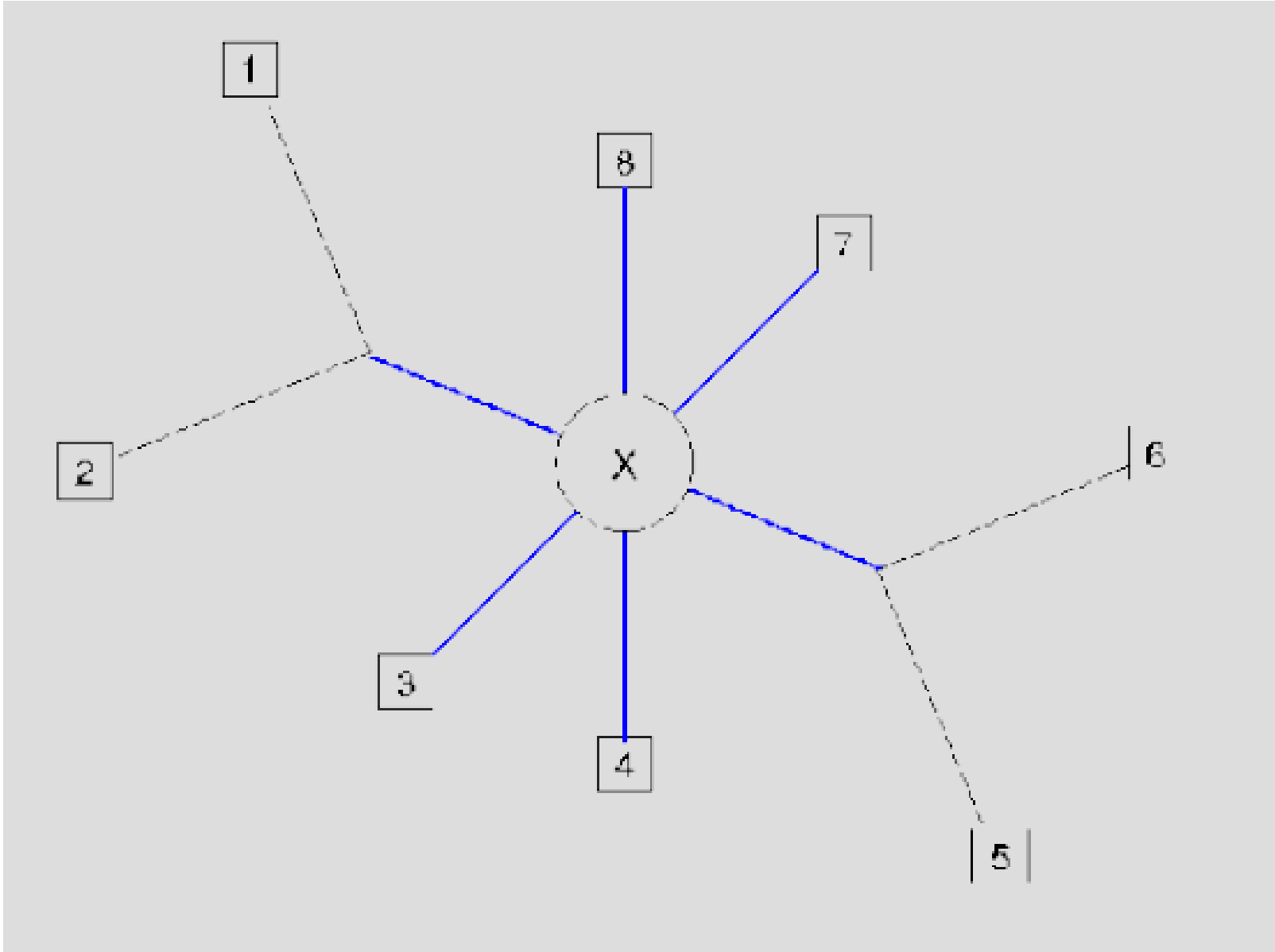
	0						
	31.50	0					
	32.30	32.30	0				
S <sub>ij</sub> =	33.90	33.90	33.70	0			
	33.90	33.90	33.70	31.30	0		
	33.70	33.70	33.50	33.10	33.10	0	
	33.70	33.70	33.50	33.10	33.10	31.90	0

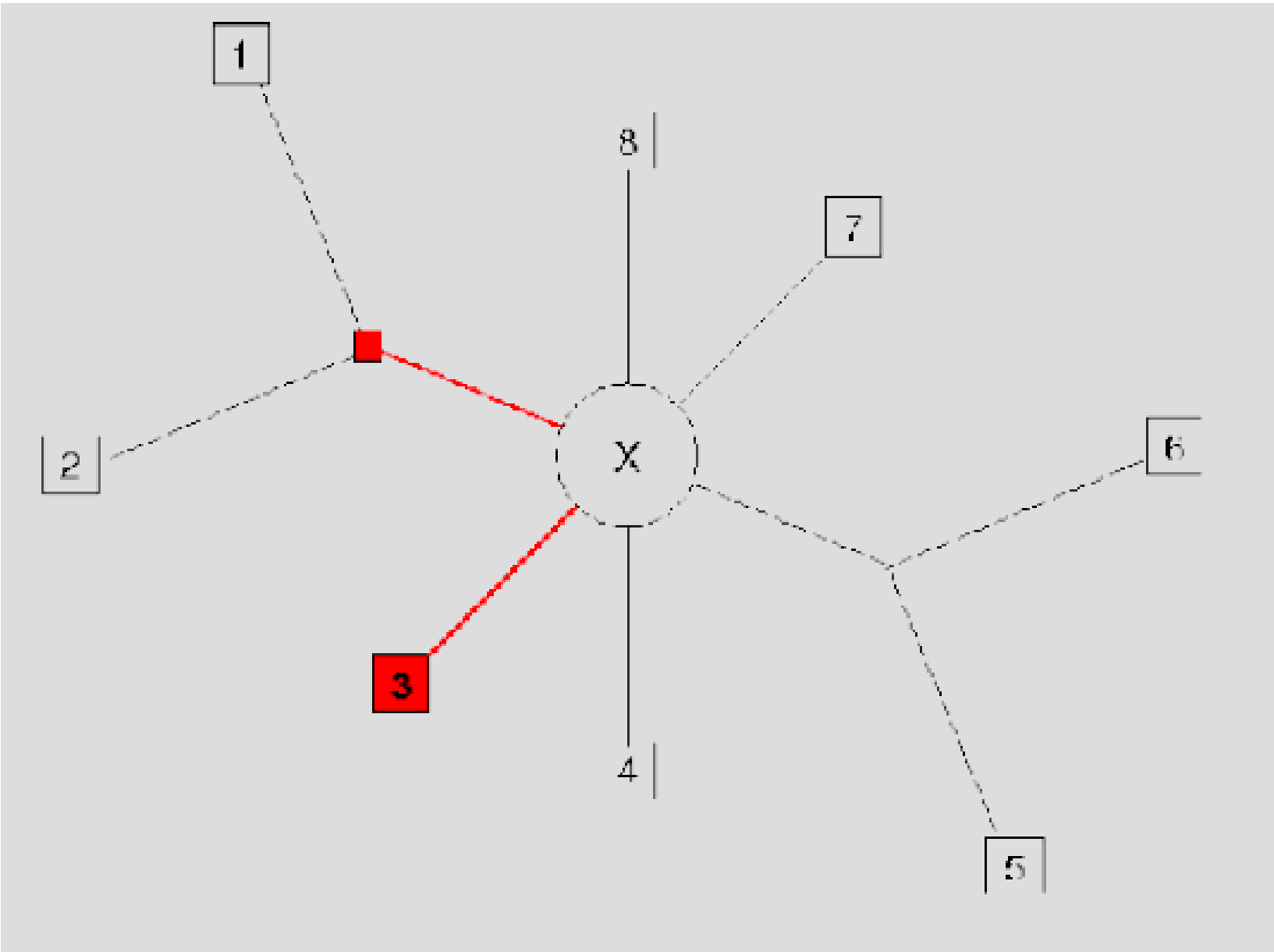


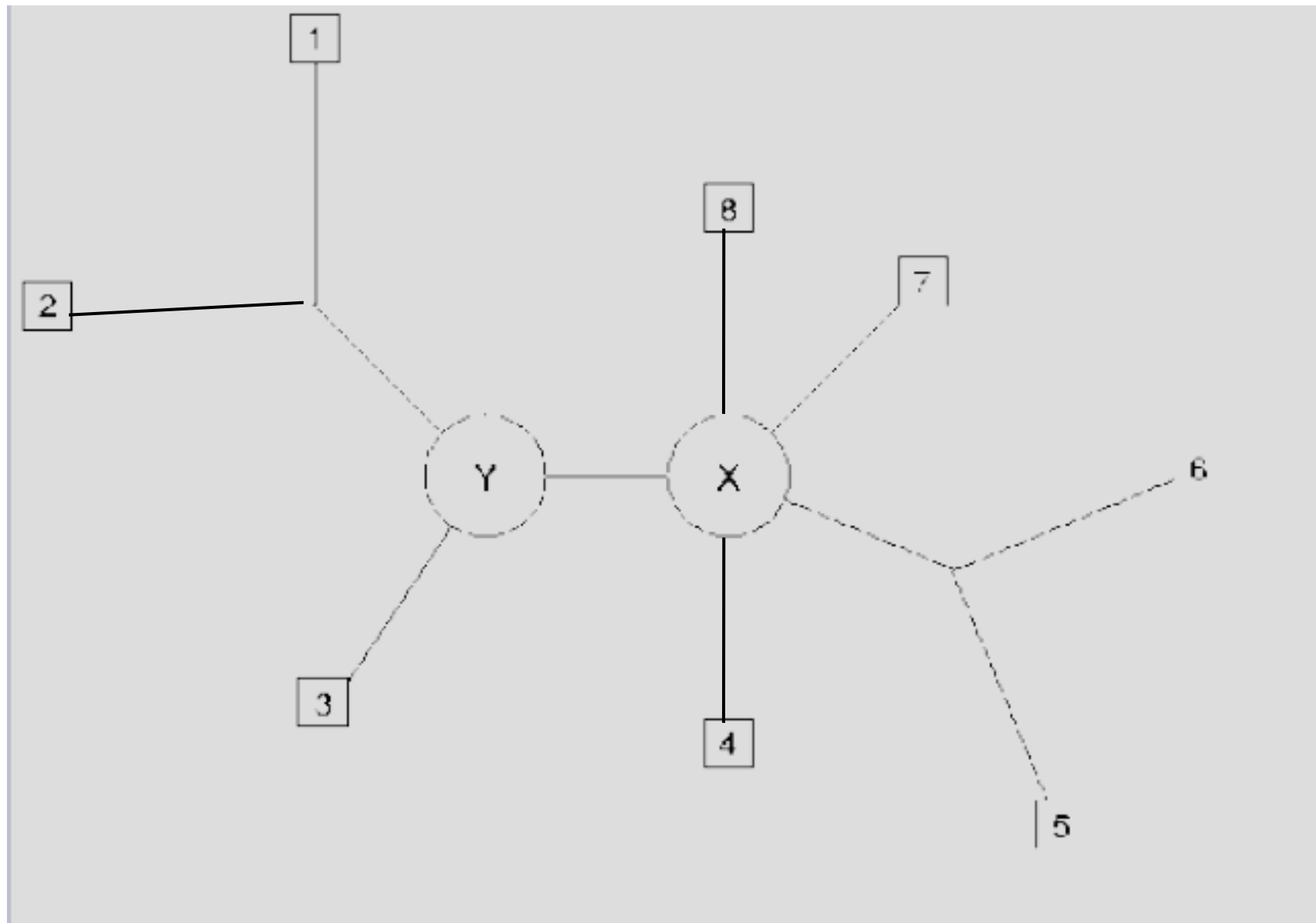
	<i>i=1_2</i>	3	4	5	6	7	8	
	0							<i>j=1_2</i>
	31.50	0						3
	32.30	32.30	0					4
<i>S<sub>ij</sub></i> =	33.90	33.90	33.70	0				5
	33.90	33.90	33.70	31.30	0			6
	33.70	33.70	33.50	33.10	33.10	0		7
	33.70	33.70	33.50	33.10	33.10	31.90	0	8

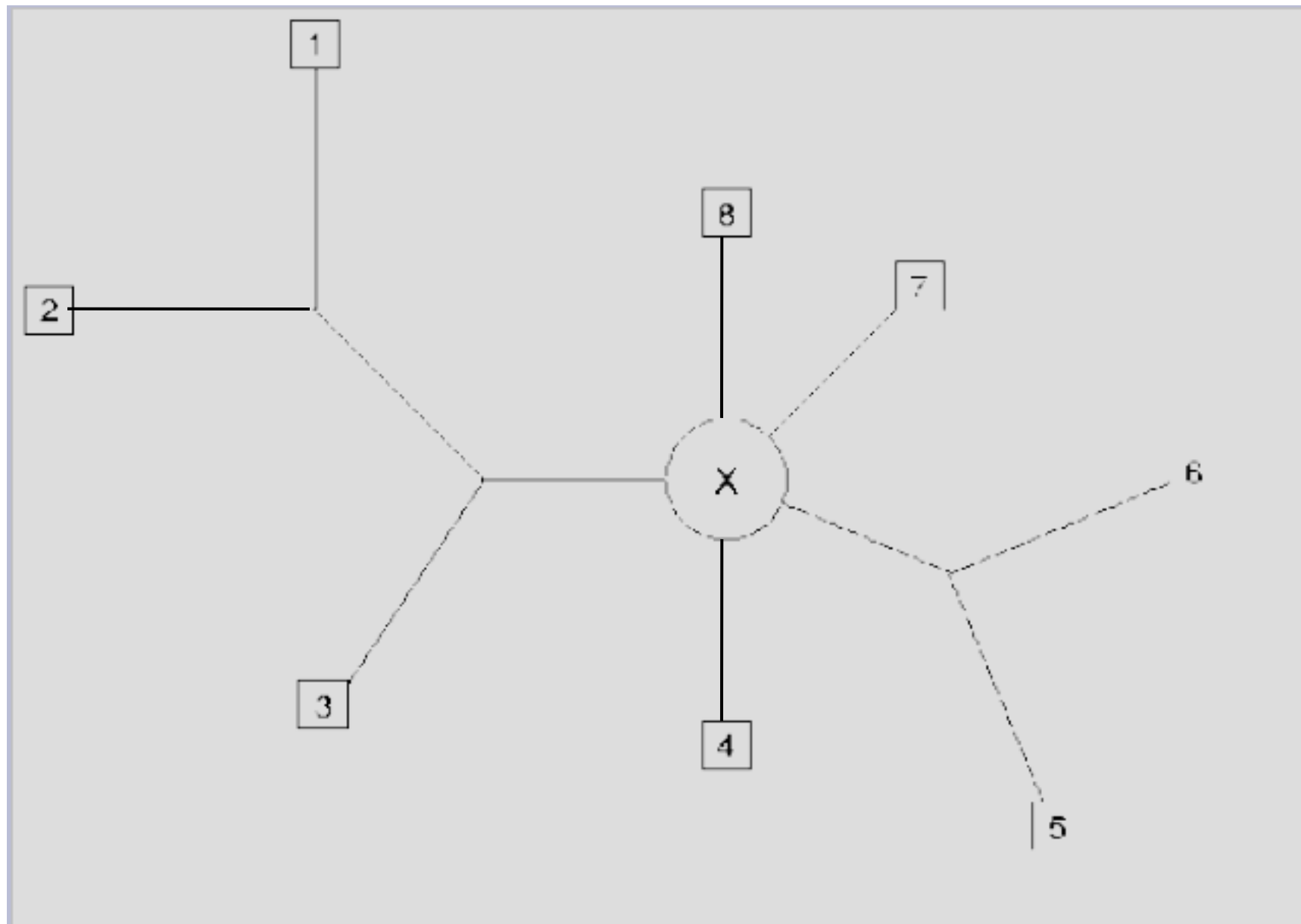


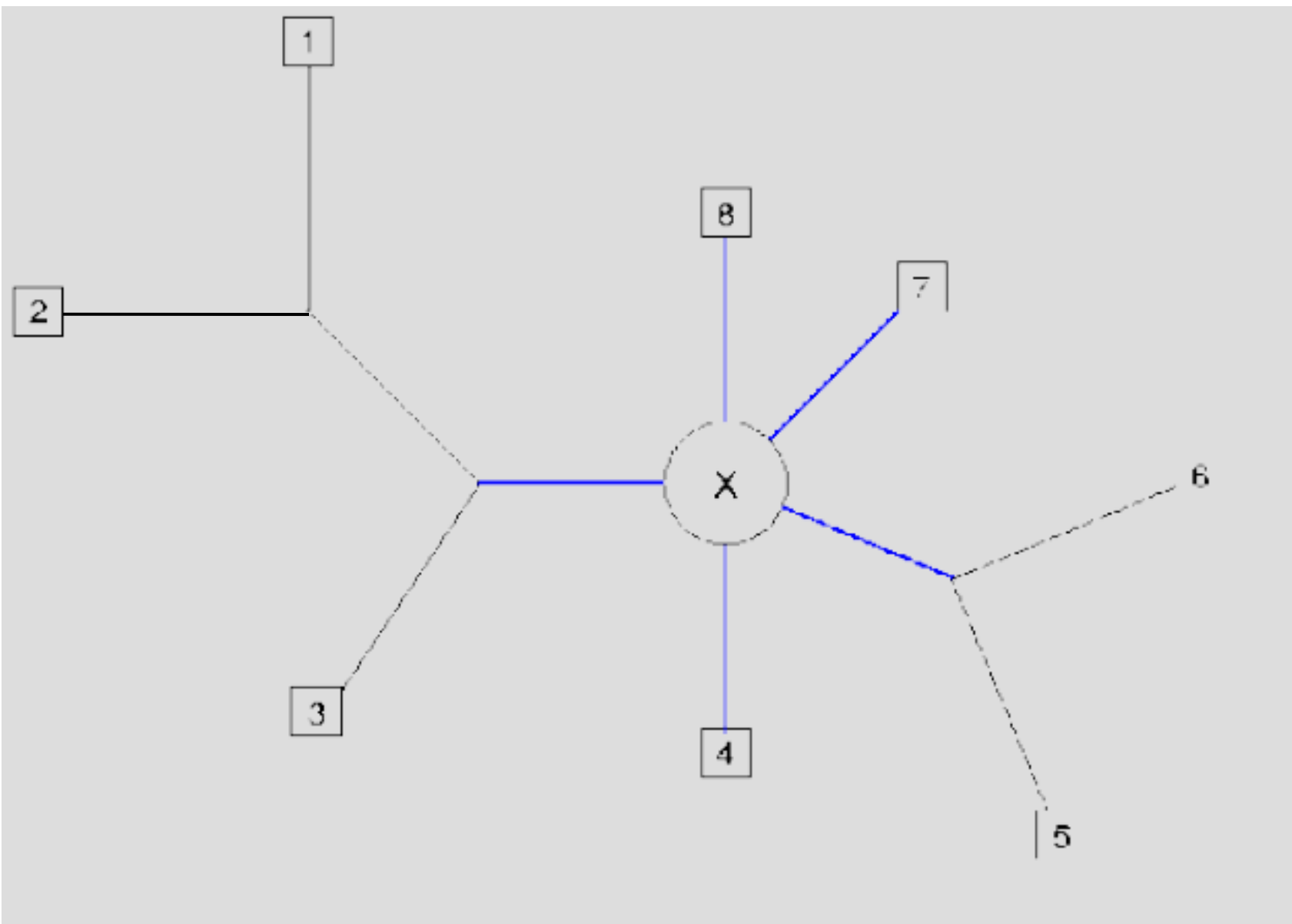












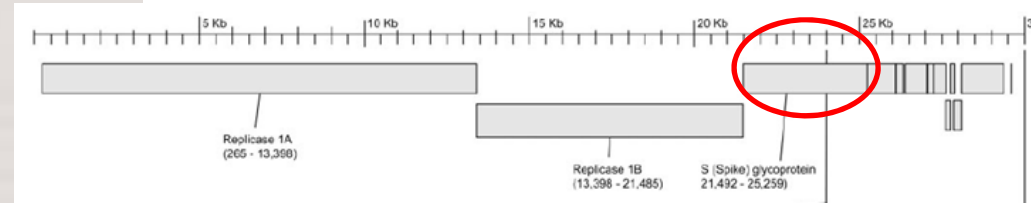
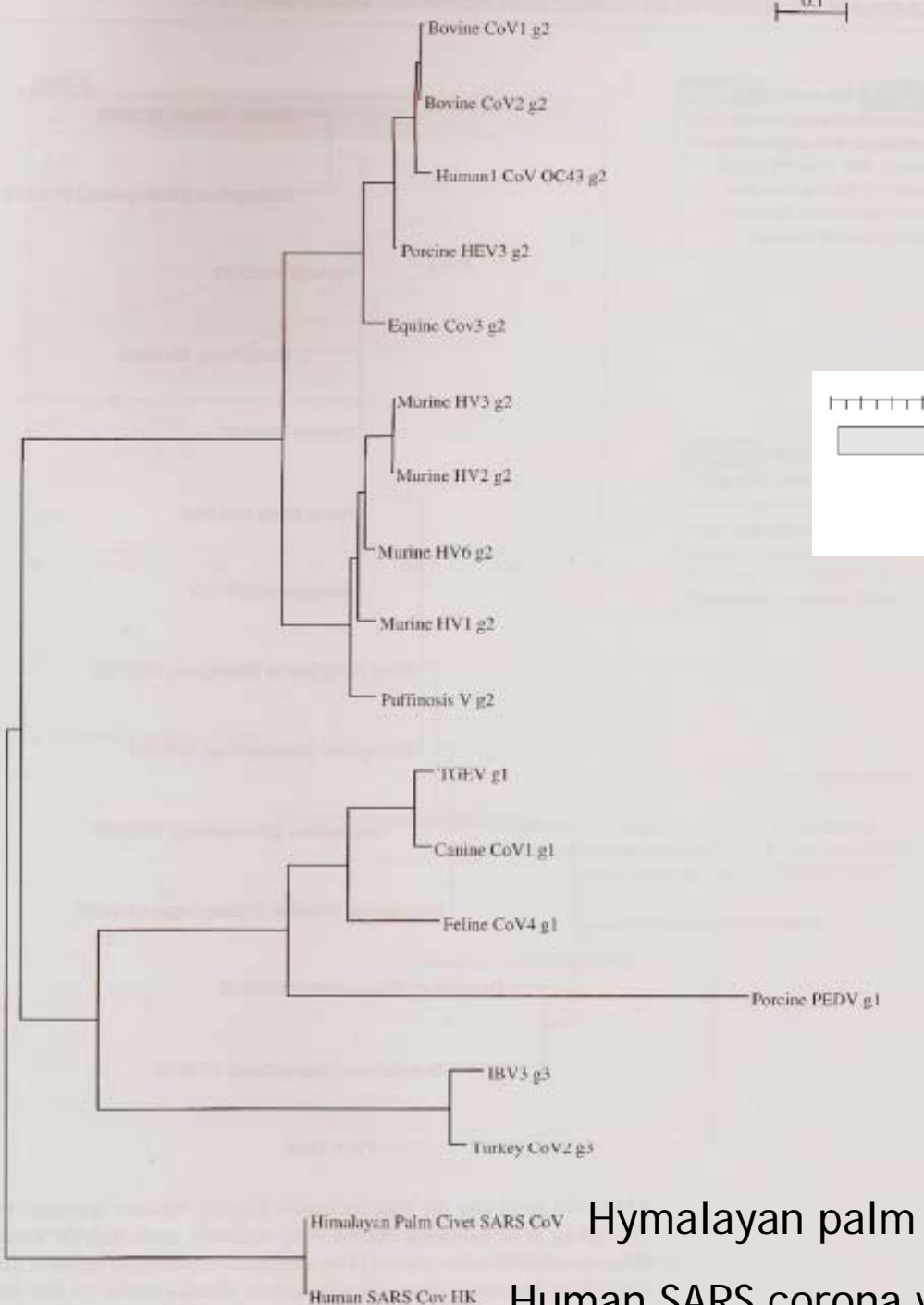
# Case study: phylogenetic analysis of the SARS epidemic

Genome of SARS-CoV: 6 genes

- \* Identify host: Himalayan Palm Civet
- \* The epidemic tree
- \* The date of origin
- \* Area of Origin

# Identifying the host

NJ tree of the 'spike' protein for animal coronaviruses



Himalayan palm civet corona virus

Human SARS corona virus

# Epidemic tree

- Use of 13 genomes
- Data and location annotated in GenBank
- NJ-tree for 'spike' gene
- Jukes-Cantor correction
- Palm civet used as outgroup

Early cases

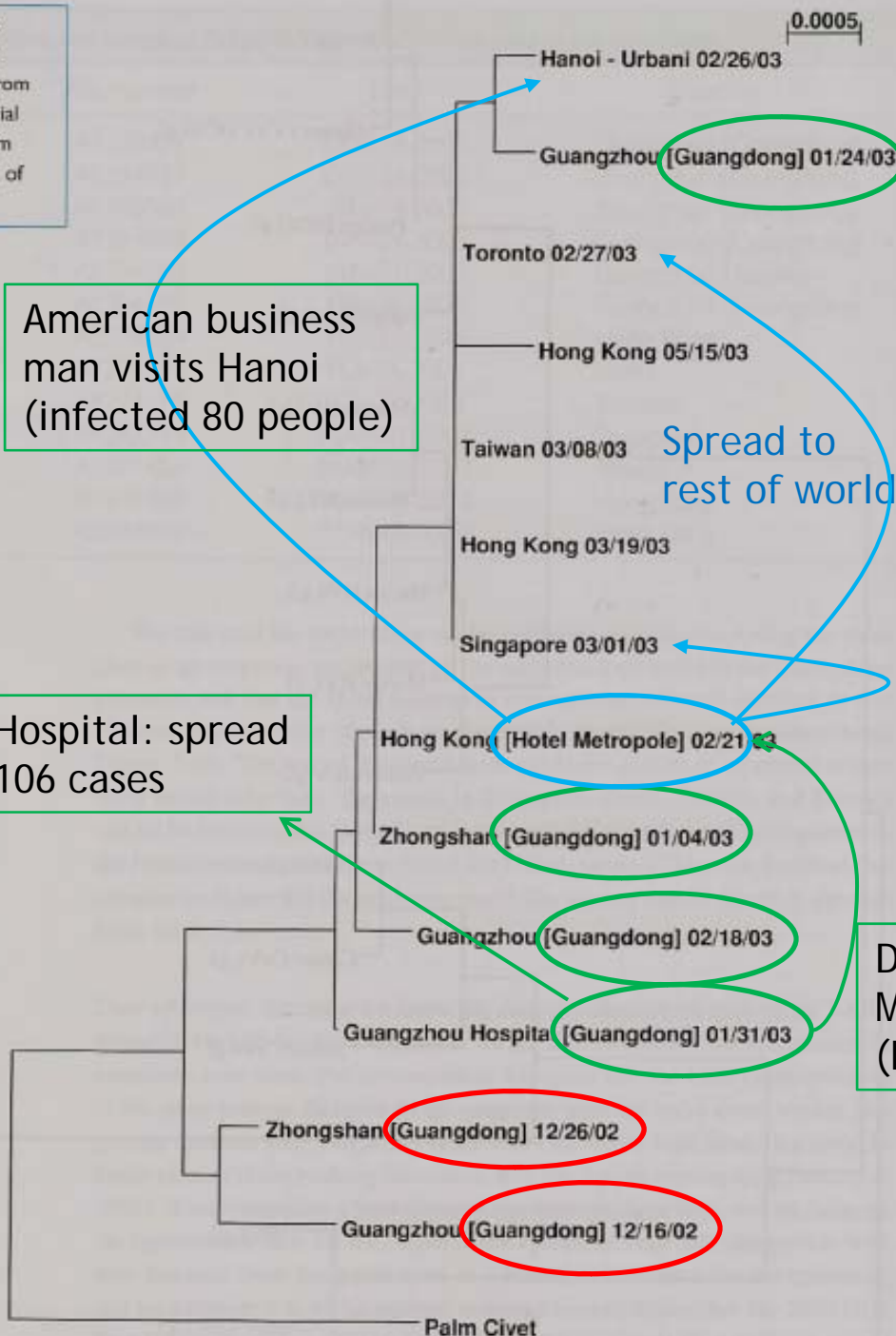
Dec-2002

Jan-2003

All in Guangdong province

Very similar to one of the Guangdong sequences (almost no genetic distance)

Hotel Metropole



American business man visits Hanoi (infected 80 people)

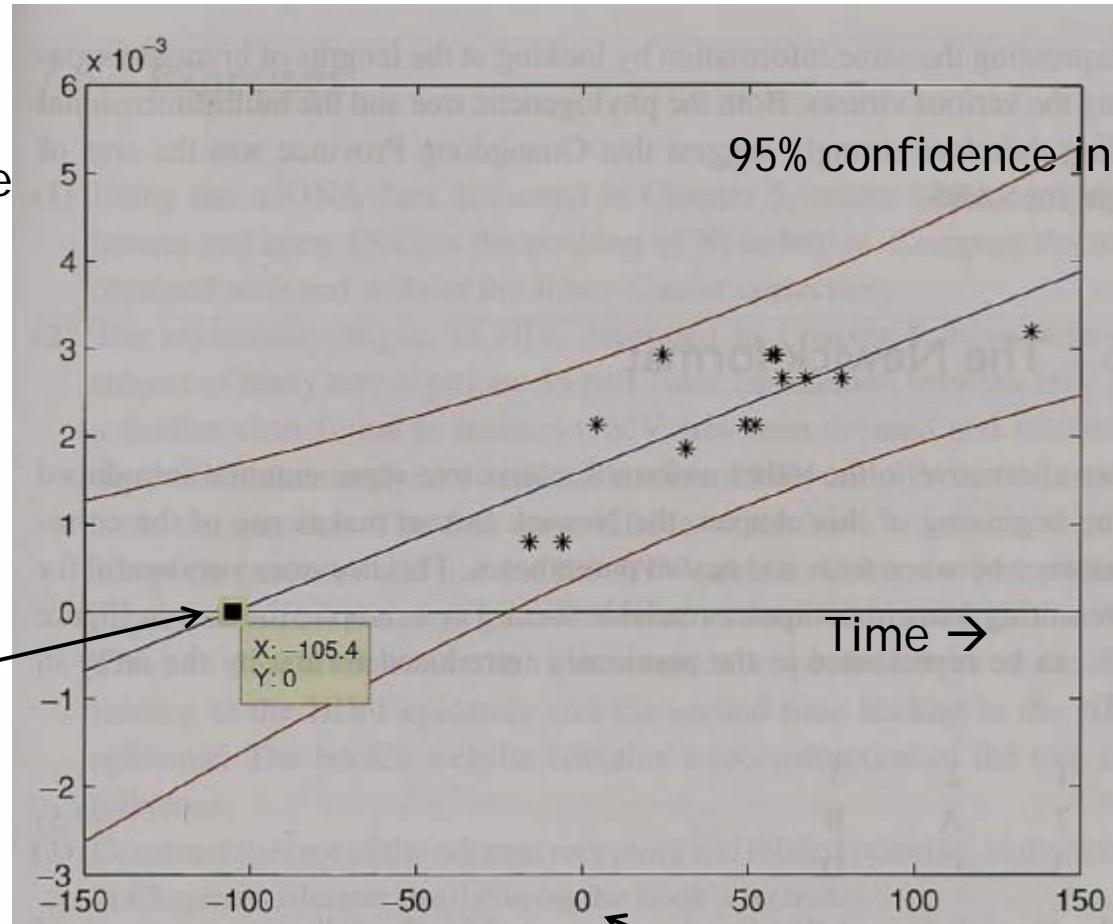
Hospital: spread 106 cases

Doctor visits Metropole hotel (Hong Kong)

# Date of origin

Use ORF of SPIKE protein

Genetic distance



Date of origin  
Extrapolation:  
Sept 16, 2002  
(106 days before  
jan 1)

Jan 1, 2003

*The genetic distance relative to the palm civet increases +/- linearly with time*

# Area of origin

multidimensional scaling

