For written notes on this lecture, please read
Chapters 7 & 8 of *Algorithms in Bioinformatics: A Practical Introduction*, and
Chapter 17 of *Algorithms on Strings, Trees, and Sequences*.

CS2220 Introduction to Computational Biology

Unit 9: Phylogenetic Trees

Wong Limsoon

Outline

What a phylogeny is

Overview of phylogeny reconstruction

Mitochondrial Eve and other romances

Distance-based phylogeny reconstruction

Robustness of reconstructions

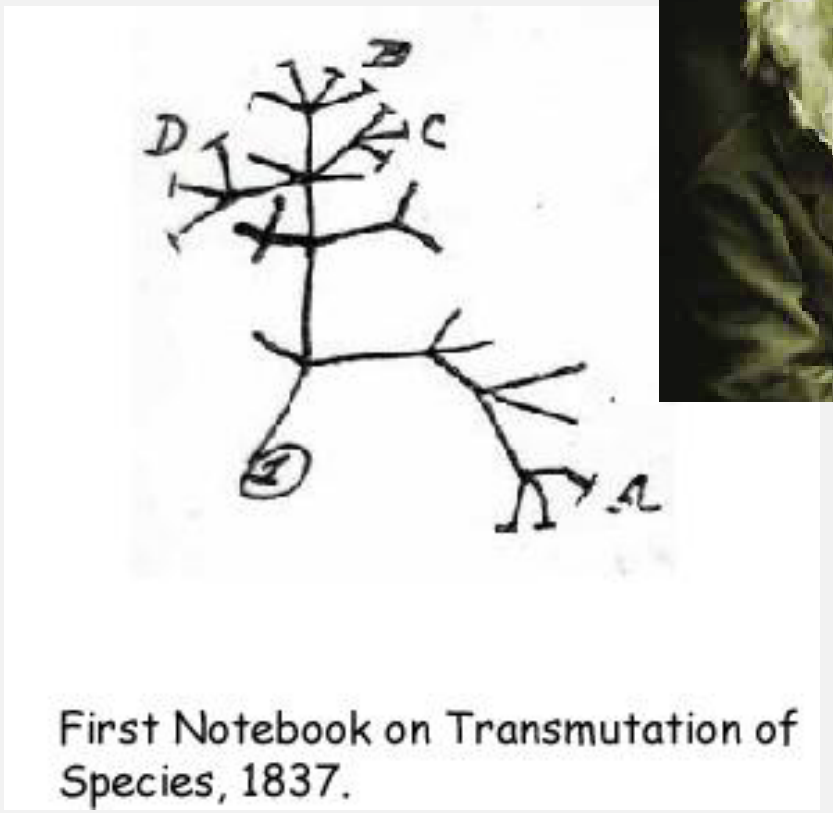
Phylogenetic tree comparison

Phylogeny

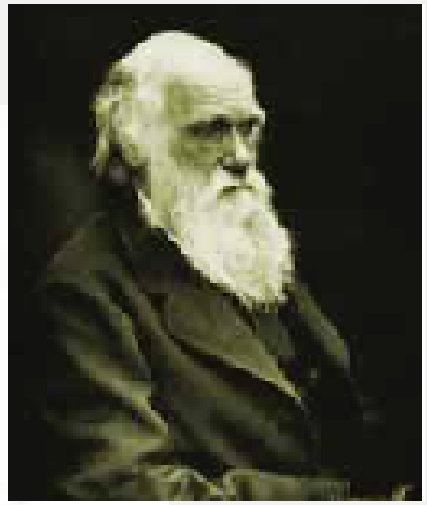
Phylogeny: Reconstruction of evolutionary history of a set of species

Long ago, it was a leaf-labeled tree where the internal nodes referred the hypothetical ancestors and the leaves were labeled by the species

Edges of the tree would represent the evolutionary relationships



First Notebook on Transmutation of Species, 1837.



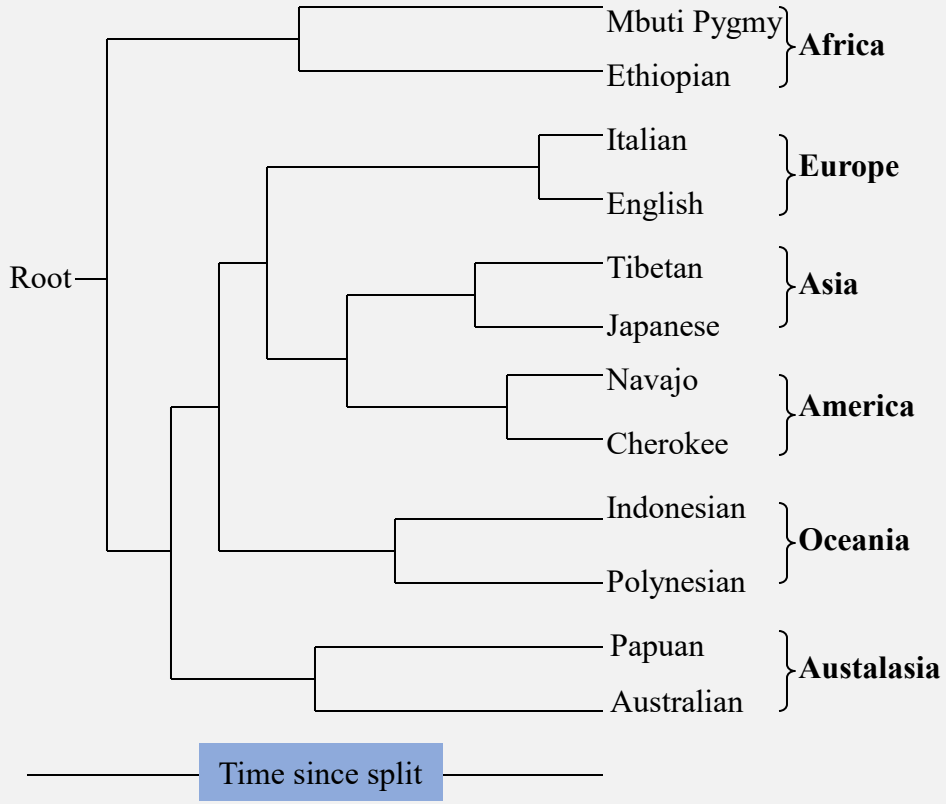
Population tree of a bygone era

Estimate order in which “populations” evolved based on assimilated frequency of many different genes

But …

Is human evolution a succession of population fissions?

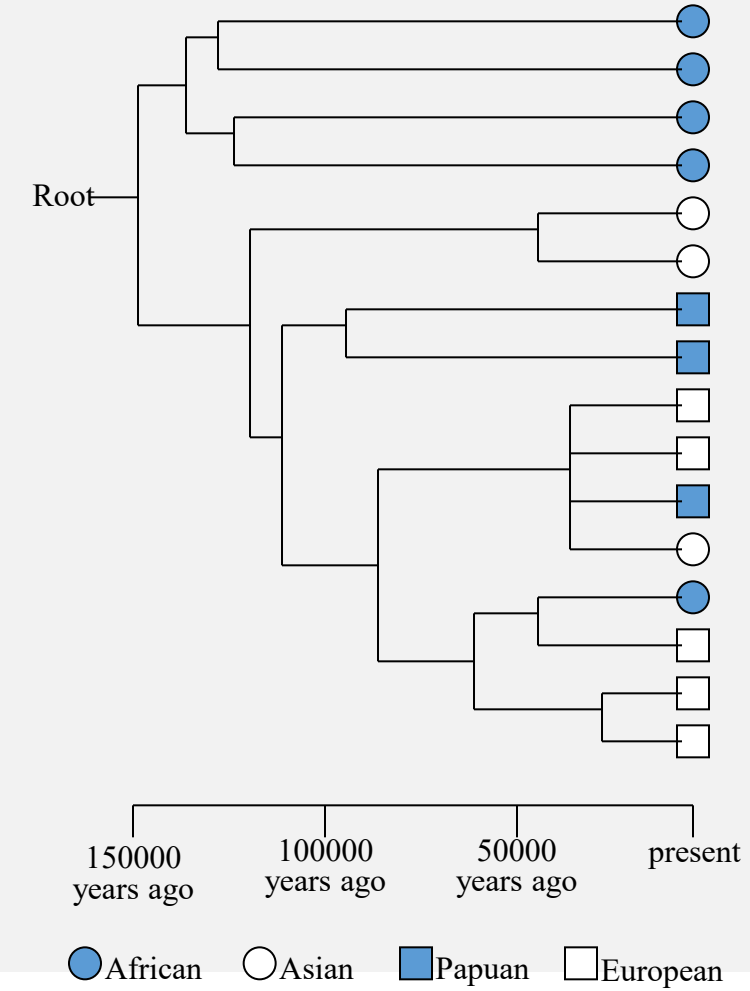
Is there such thing as a proto-Anglo-Italian population which split, never to meet again, and became inhabitants of England and Italy?



Evolution tree used nowadays

Leaves and nodes are individual persons---real people, not hypothetical concept like “proto-population”

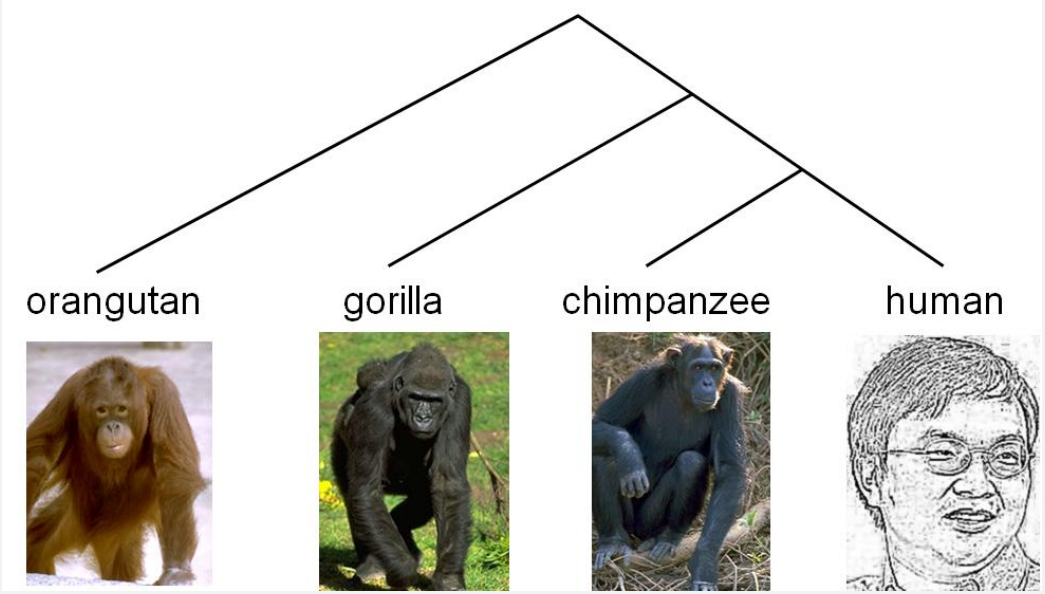
Lines drawn to reflect genetic differences between them in one special gene called mitochondrial DNA



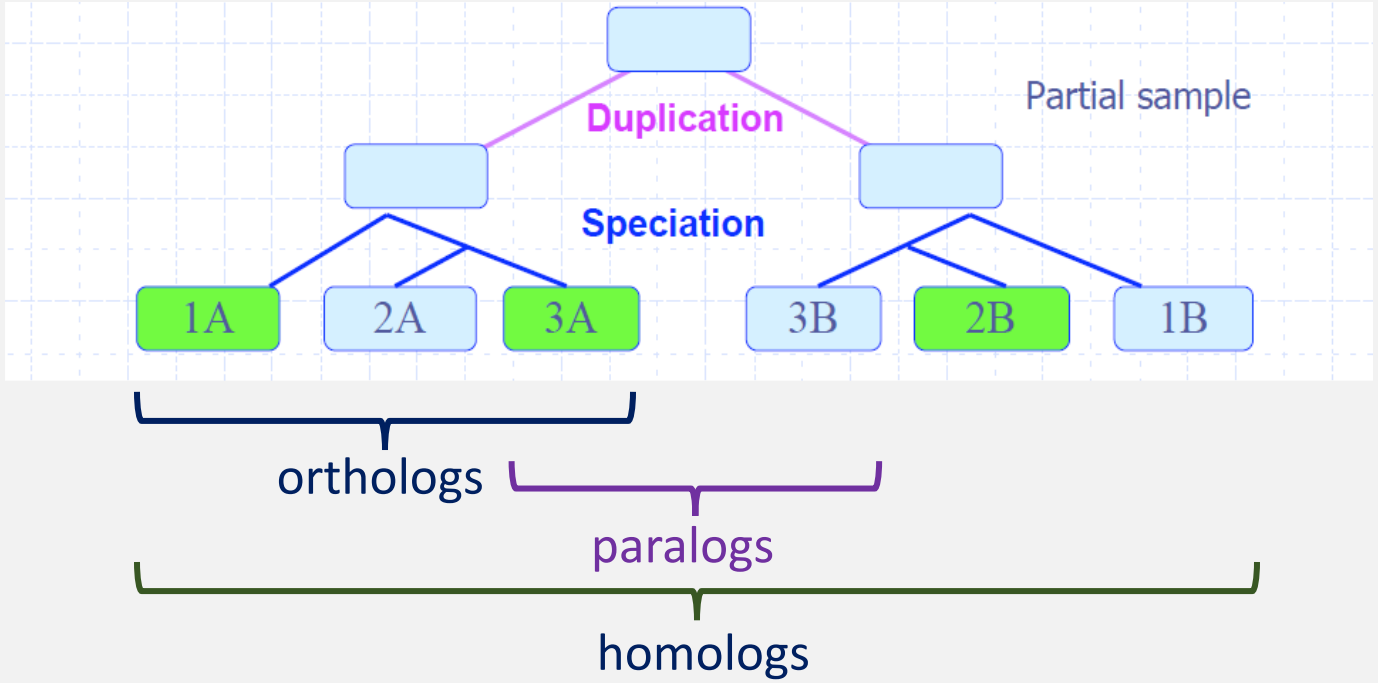
Phylogeny reconstruction

By looking at extent of conserved positions in the “multiple sequence alignment” of different groups of sequences, infer when they last shared an ancestor

$\Rightarrow$ Construct “family tree” or phylogeny

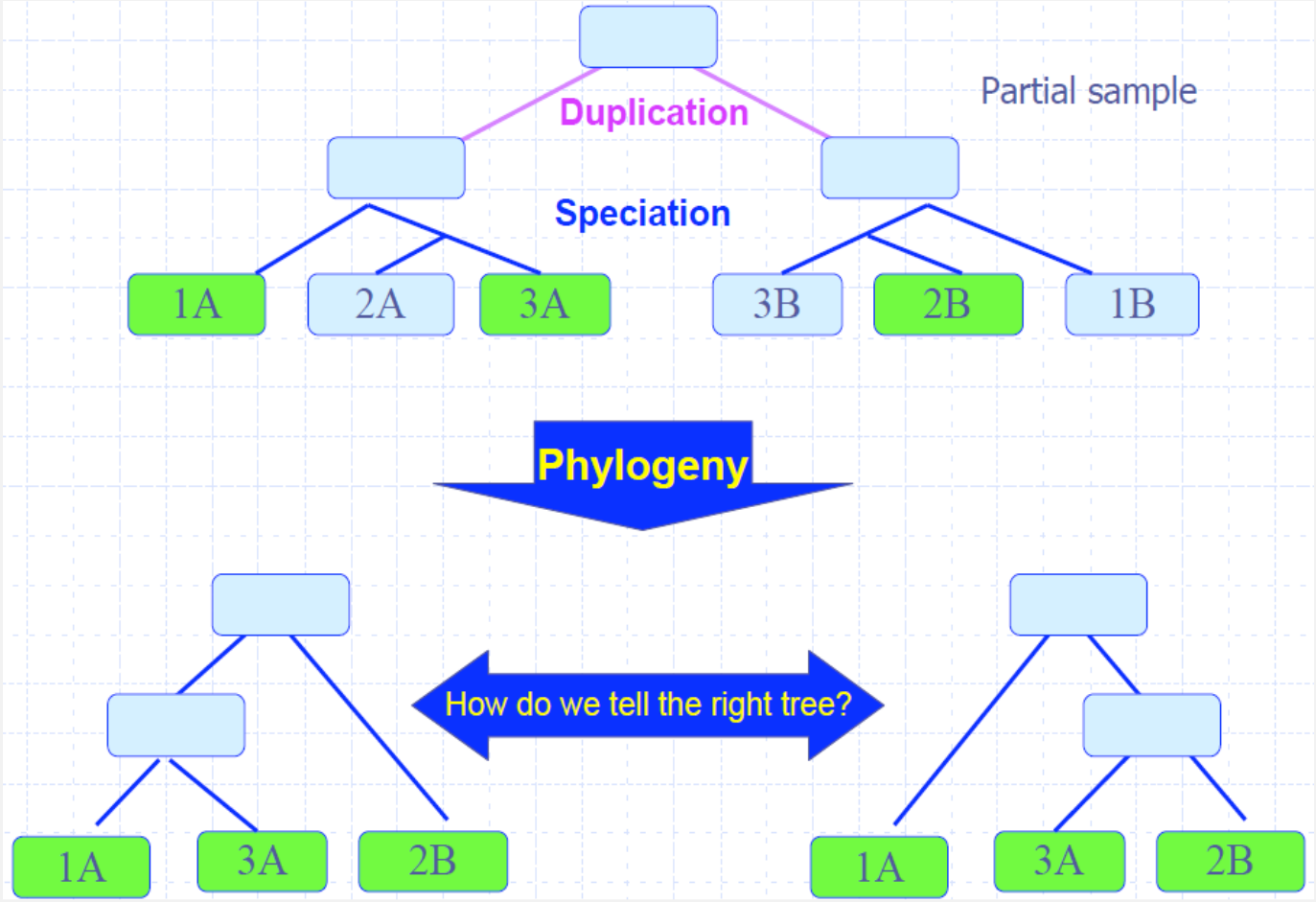


Ortholog, paralog, and homolog



Source: Yechiam Yemini

Phylogeny reconstruction: Tell evolution from homology



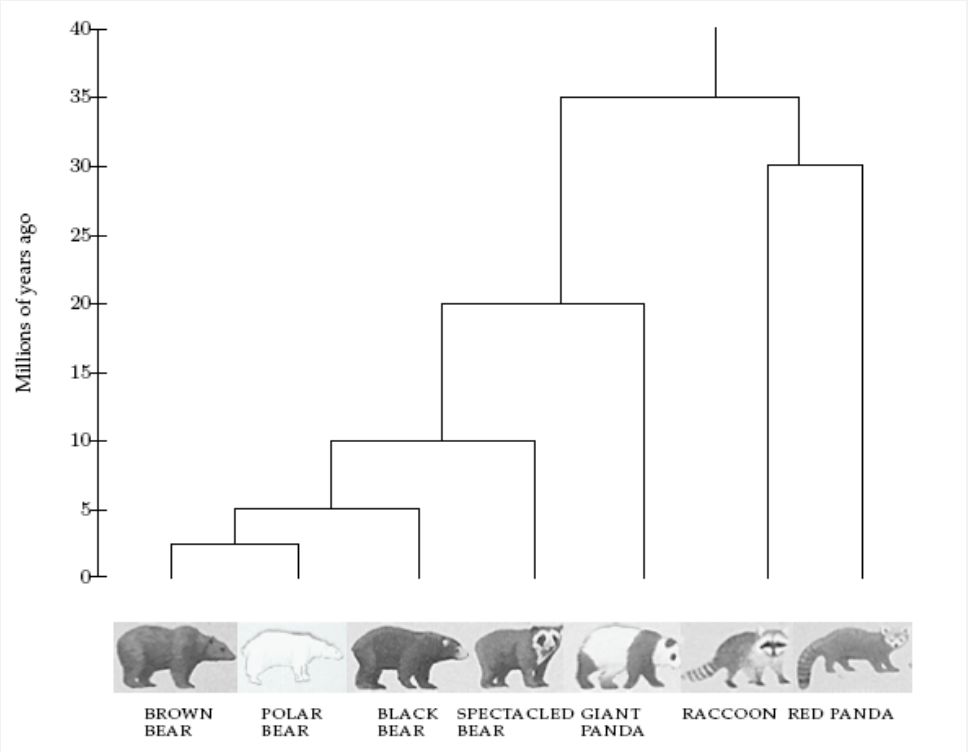
Source: Yechiam Yemini

Example application: Giant panda riddle

For ~100 years scientists couldn't figure out which family giant panda belongs to

They look like bears but have features unusual for bears but typical for raccoons, e.g., they do not hibernate

In 1985, Steven O'Brien et al. solved the giant panda classification problem using DNA sequences and algorithms



Source: Somayyeh Koohi

Example application: Flu vaccine

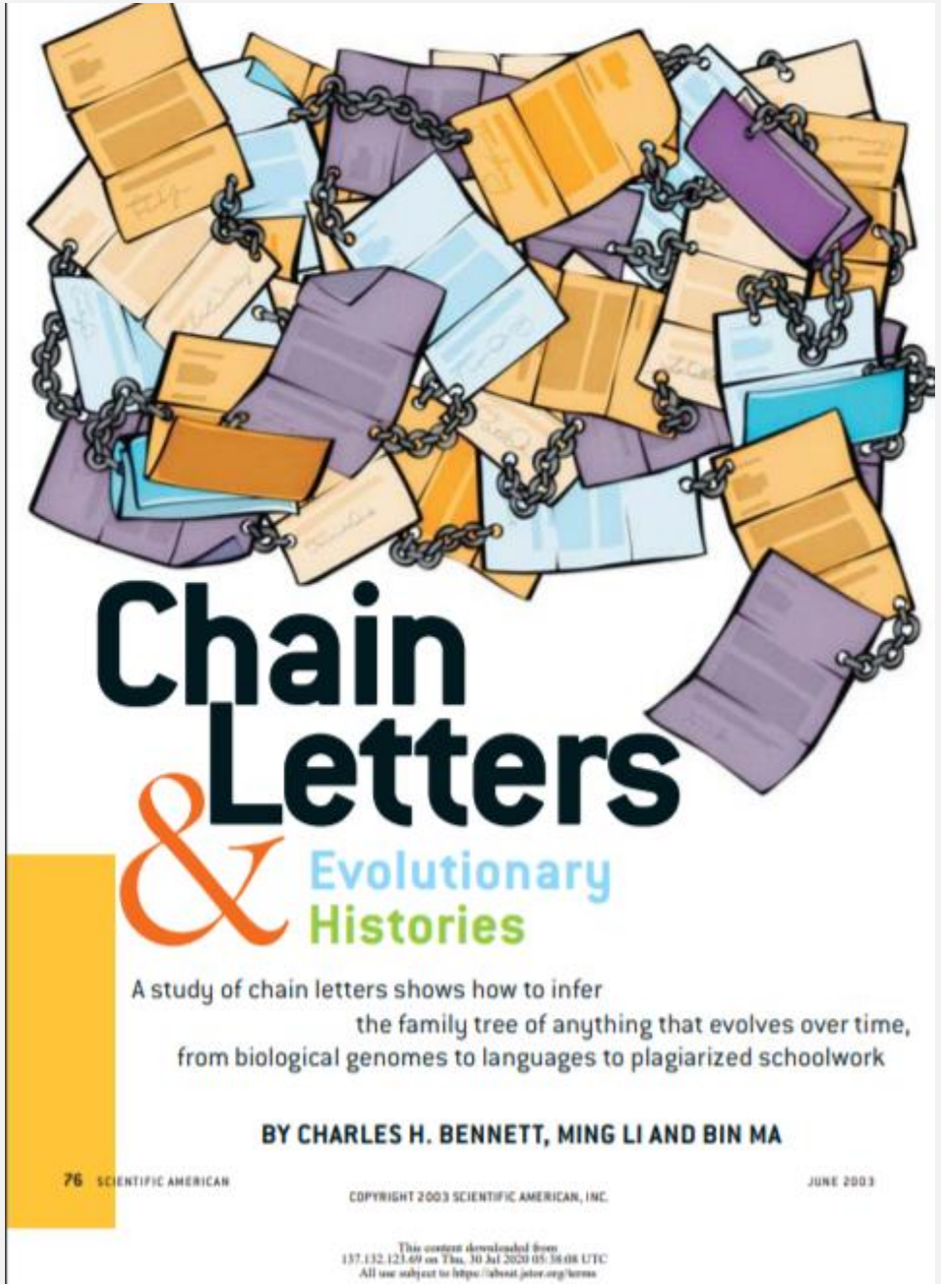
Influenza is a fast-evolving virus

Phylogenetic analyses of human influenza A (subtype H3) virus can be used to make predictions about the evolutionary course of future human influenza strains

The predicted strains of flu virus is included in the vaccine prepared each year to protect against the upcoming influenza season

R. M. Bush et al. Predicting the evolution of human influenza A. *Science*, 286:1921-1925, 1999

Example application: Chain-letter forensic



Caution

Genomes of most organisms have complex origin

Some parts of the genome are passed by vertical descent thru normal reproductive cycle

Some parts may have arisen by horizontal xfer of genetic material thru a virus, symbiosis, etc.

When a particular gene is being subjected to phylogenetic analysis, the evolutionary history of that gene may not coincide with the evolutionary history of another gene

∴ Use molecules that carry a great deal of evolutionary history, like mitochondrial DNA, and ribosomal RNA

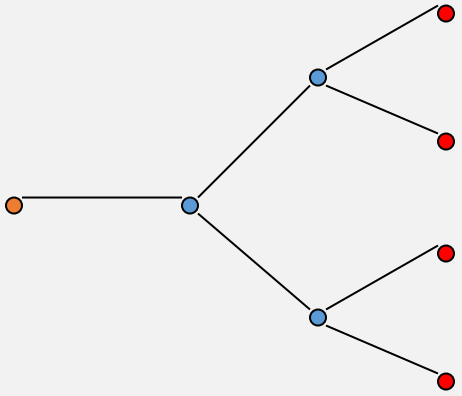
Overview of phylogeny reconstruction

Rooted and unrooted tree

Normally, the reconstructed tree is unrooted since estimating the root (i.e., oldest ancestor) is difficult

Rooted tree can be reconstructed by systematic biologists using an outgroup

Outgroup is a related species which is clearly less related with all other species in the phylogeny



How does outgroup work?

More similar to outgroup

⇒ More “ancient”

More different from outgroup

⇒ More “recent”, because more time to evolve

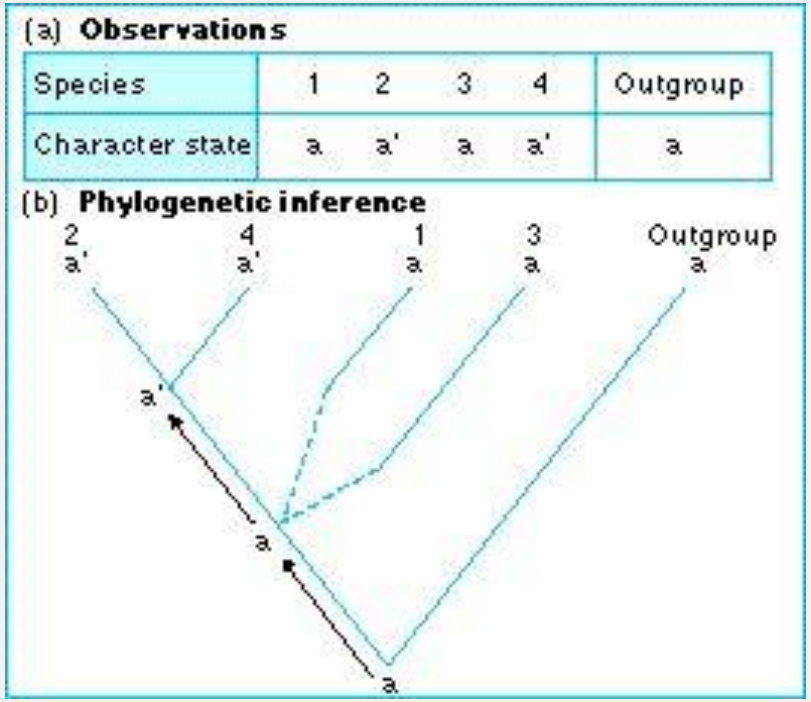


Image credit: Mark Ridley

Exercise

Outgroup is a species which is clearly less related with all other species in the phylogeny

Should you use an outgroup that is completely unrelated to the other sequences in the phylogeny reconstruction?

Distance, character, & branch length

A tree can be based on:

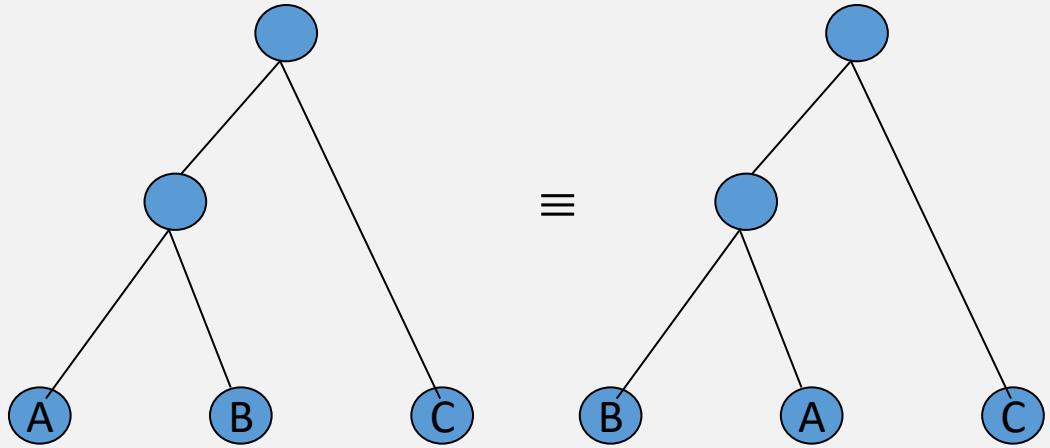
Quantitative measures like the distance or similarity between species, or

Qualitative aspects like common characters

Branches indicate phylogenetic relationship:

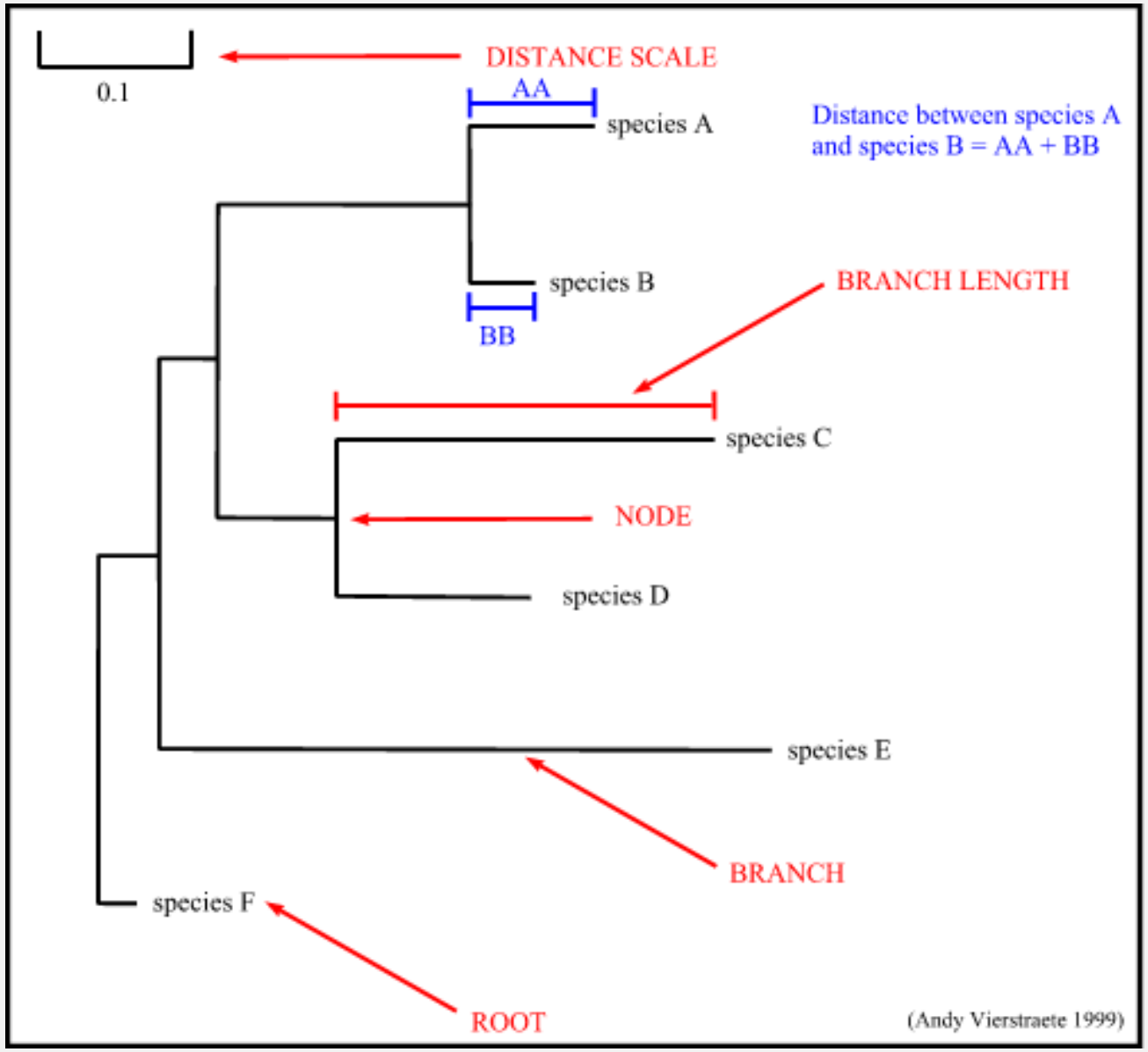
Branches can have length that indicate phylogenetic distance or closeness

Rotations of internal branches keep the phylogenetic relations intact



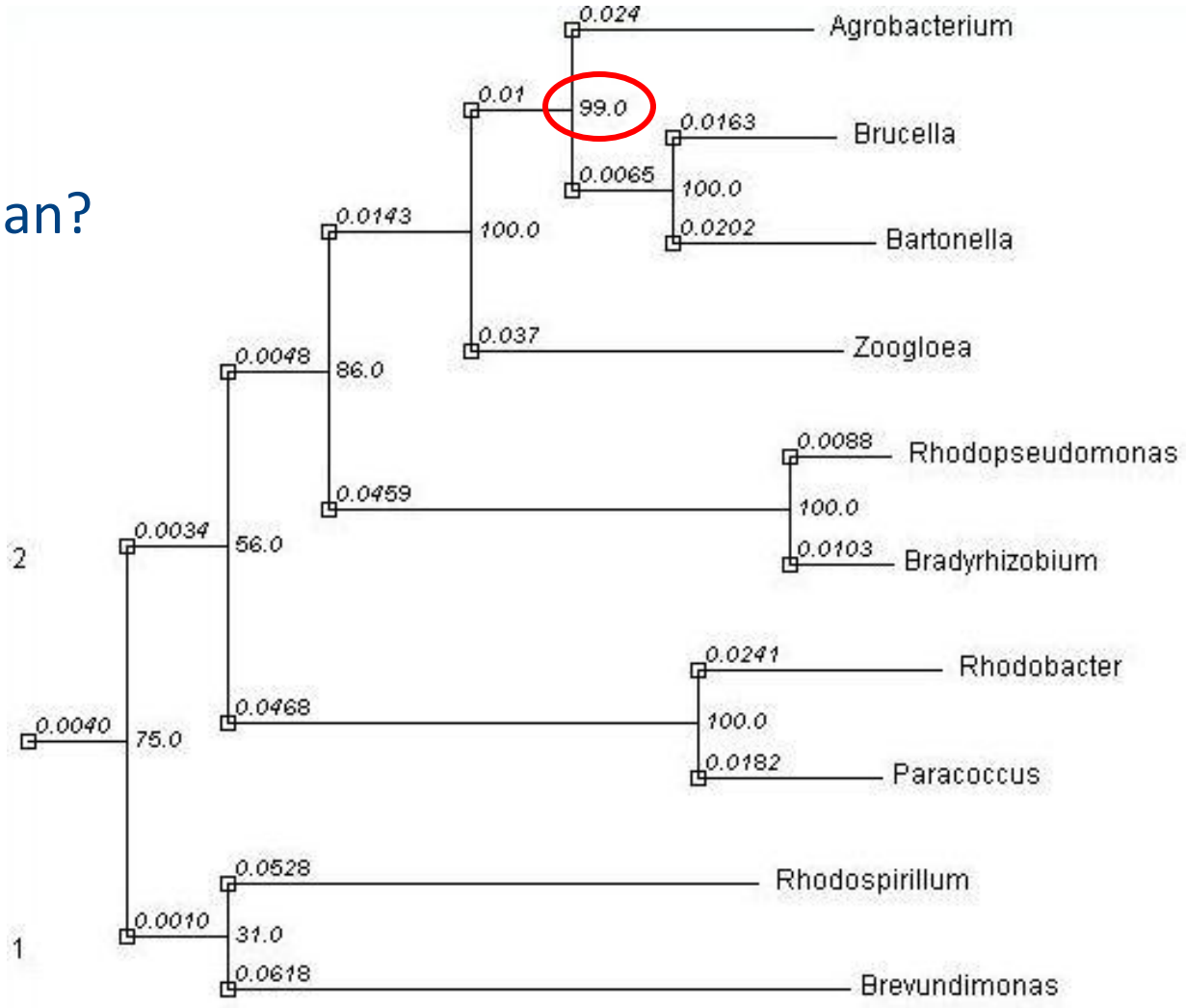
Source: Somayyeh Koohi

Tree terminologies



Exercise

What do numbers on the nodes mean?



Parsimony principle

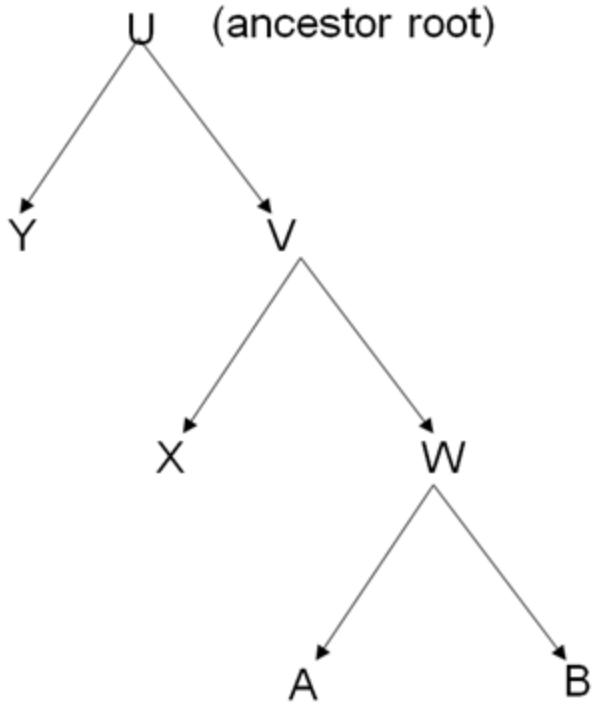
Given multiple possible evolutionary trees

Choose the evolutionary tree that requires the fewest changes (mutations, substitutions, insertions, deletions, etc.) to explain the observed data

i.e., **minimize the total number of evolutionary events** needed to account for the differences among the sequences

Exercise

```
X    ACCTG-TACTTCGATAA
Y    ACCAG-TACTT-GATAA
A    ACCAGGTACTTCGATAT
B    ACCAGGTACTTCGATTT
        1 2     3    4
```



What is the most likely sequence for U?

Brute-force phylogeny reconstruction

How?

Enumerate all trees

Compute evolutionary likelihood

Select best tree

Complexity?

More practical approaches

Maximum parsimony

Minimize # of mutations

Distance

Minimize "evolutionary" distance

OK for large # of seqs

Commonly used; we study this one

Maximum likelihood

Maximize likelihood of mutations

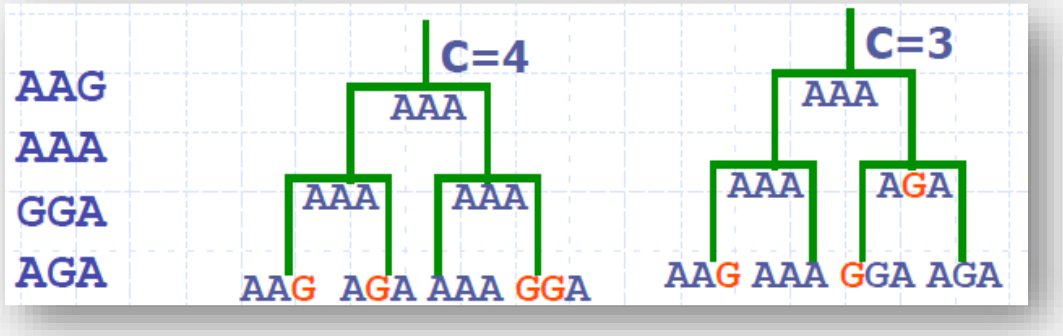
Require more understanding of evolutionary models

Involve exponential # of steps

Limited to small number of seqs

Maximum parsimony

Find tree with minimal character-state changes to explain data



Source: Yechiam Yemini

Input

- A **multiple sequence alignment** of DNA, RNA, or protein sequences.
- Each column (site) in the alignment is assumed to evolve independently.

Scoring a tree

For each site in the alignment:

- The algorithm computes the **minimum number of changes** required to produce the observed nucleotides or amino acids across the taxa, given a particular tree topology.
- This is often done using **Fitch's algorithm** (for unordered characters) or **Sankoff's algorithm** (for weighted costs).

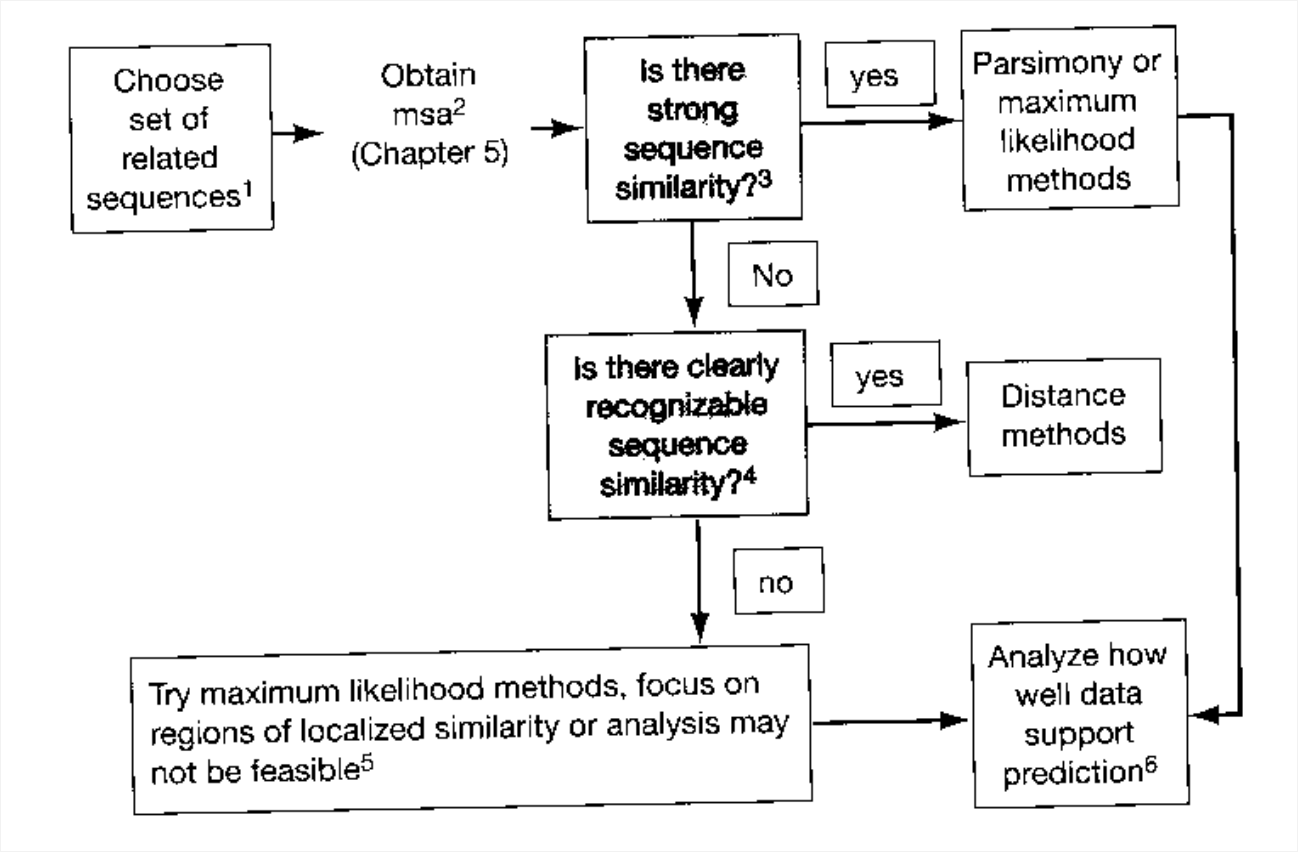
The **total parsimony score** of a tree is the sum of these minimum changes across all sites.

Exercise

What are the characteristic of maximum parsimony?

Is maximum parsimony more likely to over or under estimate the evolutionary change that has occurred?

When to use which method?



Source: D.W.Mount, Bioinformatics: Sequence and Genome Analysis, Cold Spring Harbor Press, 2004

Mitochondrial Eve and other romances

The pioneers

While Zuckerkandl and Pauling conceived molecular phylogeny

Wilson made it a science through:

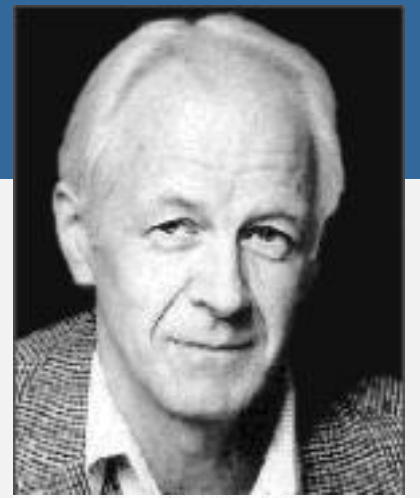
Rigorous quantitative comparison of molecular data

Development of empirical distance methods

Pioneering applications to real biological questions

Stage	Key figures	Contribution
Conceptual proposal	Zuckerkandl & Pauling (1962–65)	Proposed molecular sequence as evolutionary documents
Algorithmic foundation	Fitch, Margoliash (1967–71)	Developed parsimony and distance methods
Experimental realization	**Allan Wilson** (1960s–80s)	Empirically reconstructed phylogenies using molecular (protein → DNA) data

Allan Wilson



"Molecular clock": Dating by genetic mutations

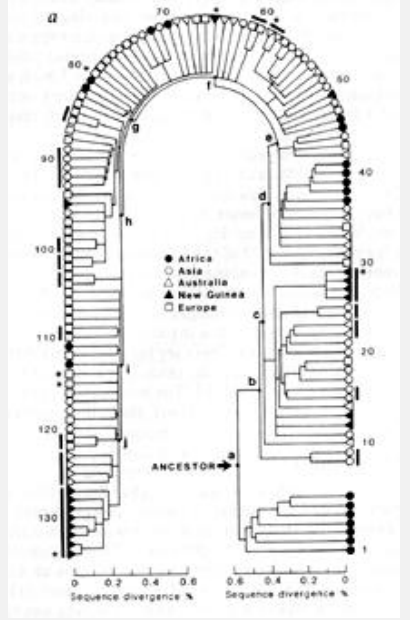
Deduced in 1960s that proto-hominids evolved 5m yrs ago, contrary to the 25m yrs believed by anthropologists

In 1980s, his findings became more widely accepted

Molecular approach to understand evolution

Concluded in 1980s that modern man evolved from "African Eve"

20 yrs to convince palaeontologists, but when they did, it married their science with that of genetics



Human mitochondrial DNA (mtDNA)

Circular double-stranded consisting of ~16k bp,
Enough to look at the 500bp control region

Everyone inherits the mtDNA from his/her mother,
No recombination

Pointwise mutation substitution rates of mtDNA is ~10x faster than nuclear DNA,
Accumulate about 1 mutation every 10,000 years *in the control region*

Every cell has lots of mtDNAs

Genetics finds origin of human

Statistical analysis of mtDNAs from placental tissue of 147 women of different races & regions

Wilson's group and others construct phylogenetic tree assuming constant molecular clock

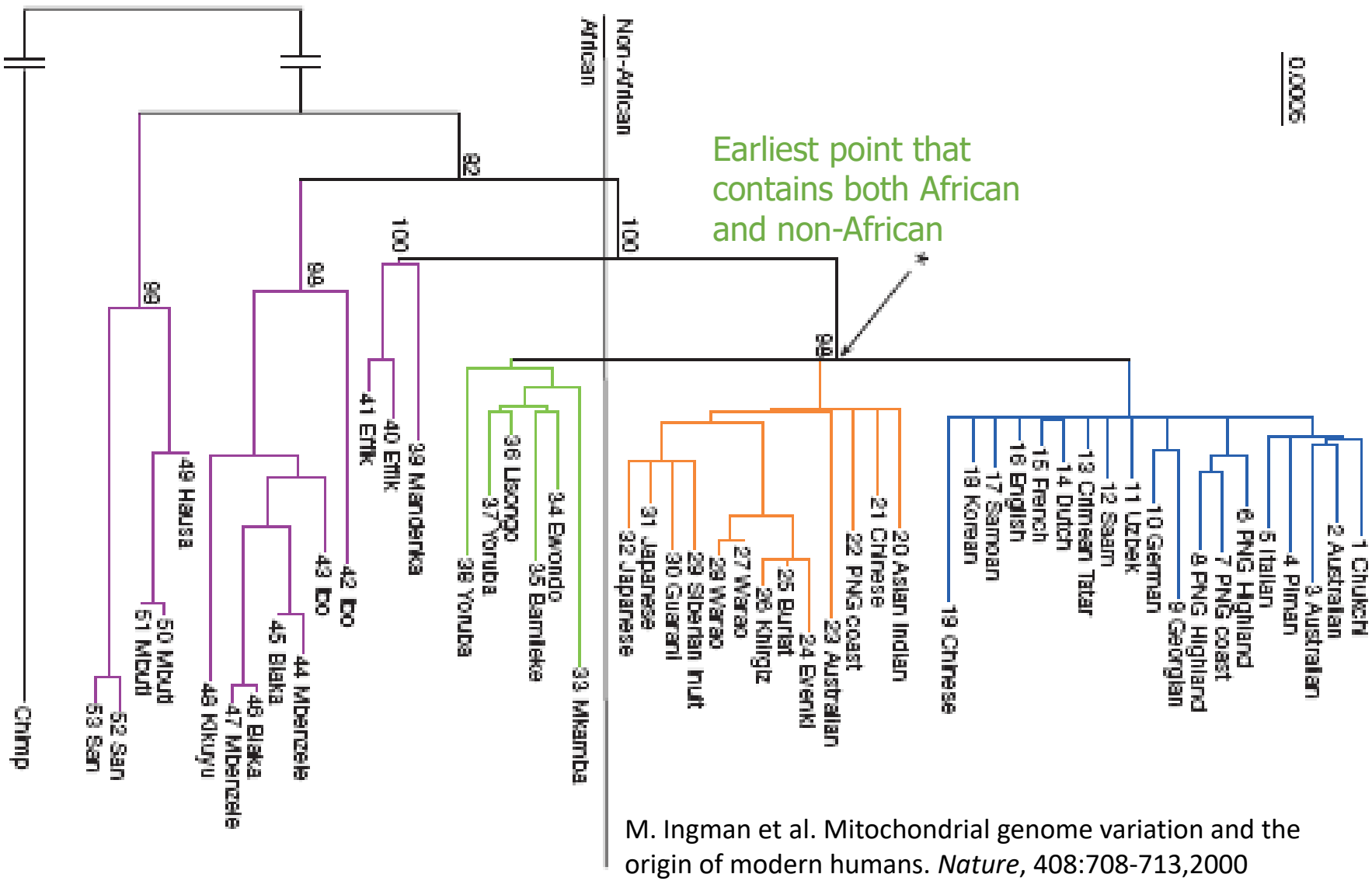
The tree implies that the common ancestor of modern human appear ~143,000 years ago

- L. Vigilant et al. African populations and the evolution of human mitochondrial DNA. *Science*, 253:1503-1507, 1991.
- R. L. Cann et al. Mitochondrial DNA and human evolution. *Nature*, 325:31-36, 1987.

Exercise: Eve tree

What is the outgroup?

Why is Eve considered African?



M. Ingman et al. Mitochondrial genome variation and the origin of modern humans. *Nature*, 408:708-713,2000

Y-chromosome Adam

Y chromosome is unique to males and it can help to find the father of humans

Mutation rate of Y chromosome not as fast as mtDNA

Need more samples to study Y-chromosome evolution

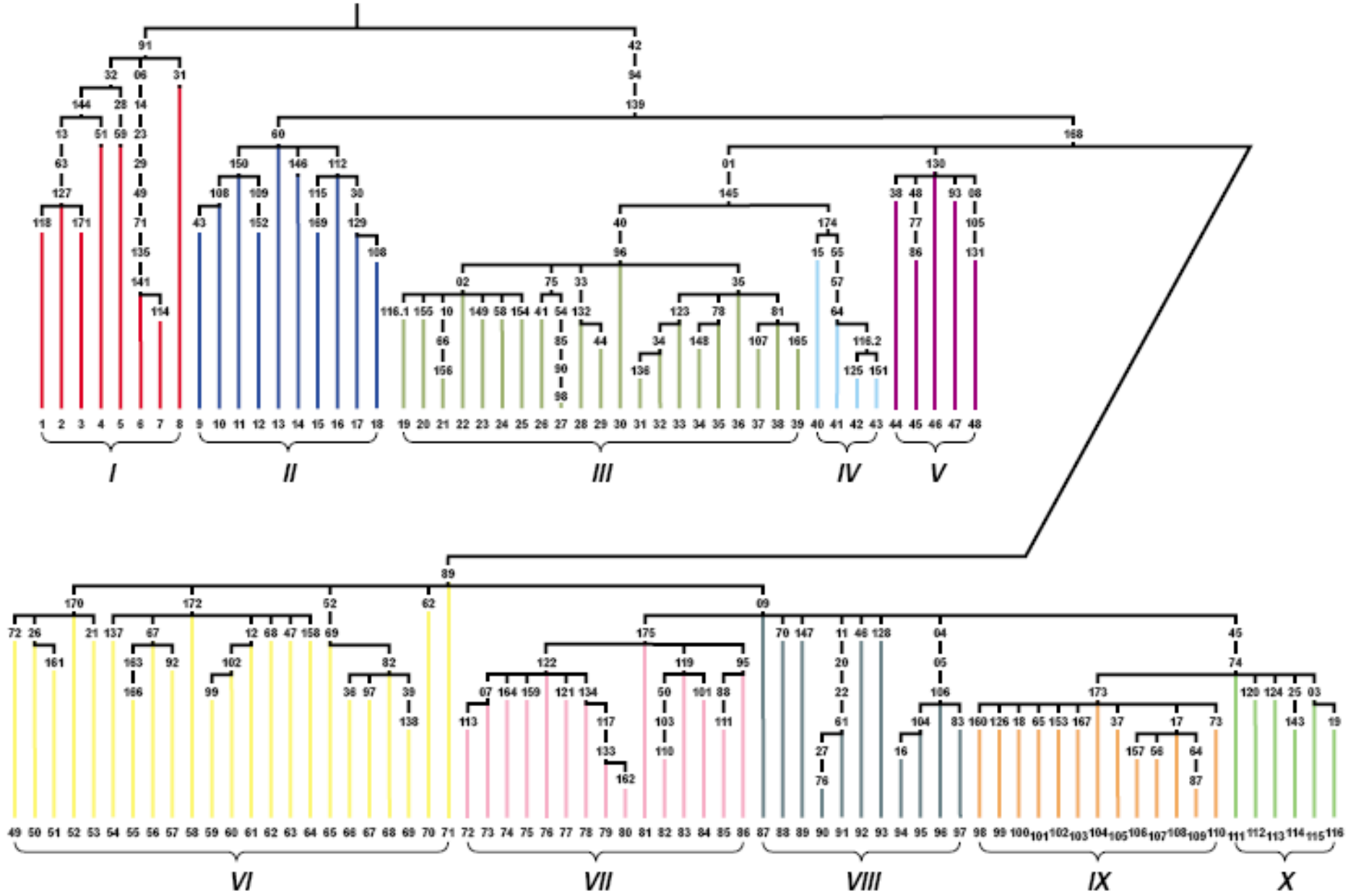
Y chromosome of 1,062 males from 22 different geographic areas were analyzed

167 haplotypes identified

Common ancestor of the 167 haplotypes estimated to appear ~60,000 years ago

Underhill et al. Y chromosome sequence variation and the history of human populations. *Nature Genetic*, 26:358-361, 2000

Adam tree



Exercise

Eve appeared ~143,000 years ago

Adam appeared ~60,000 years ago

Can Adam & Eve appear in different time? How?

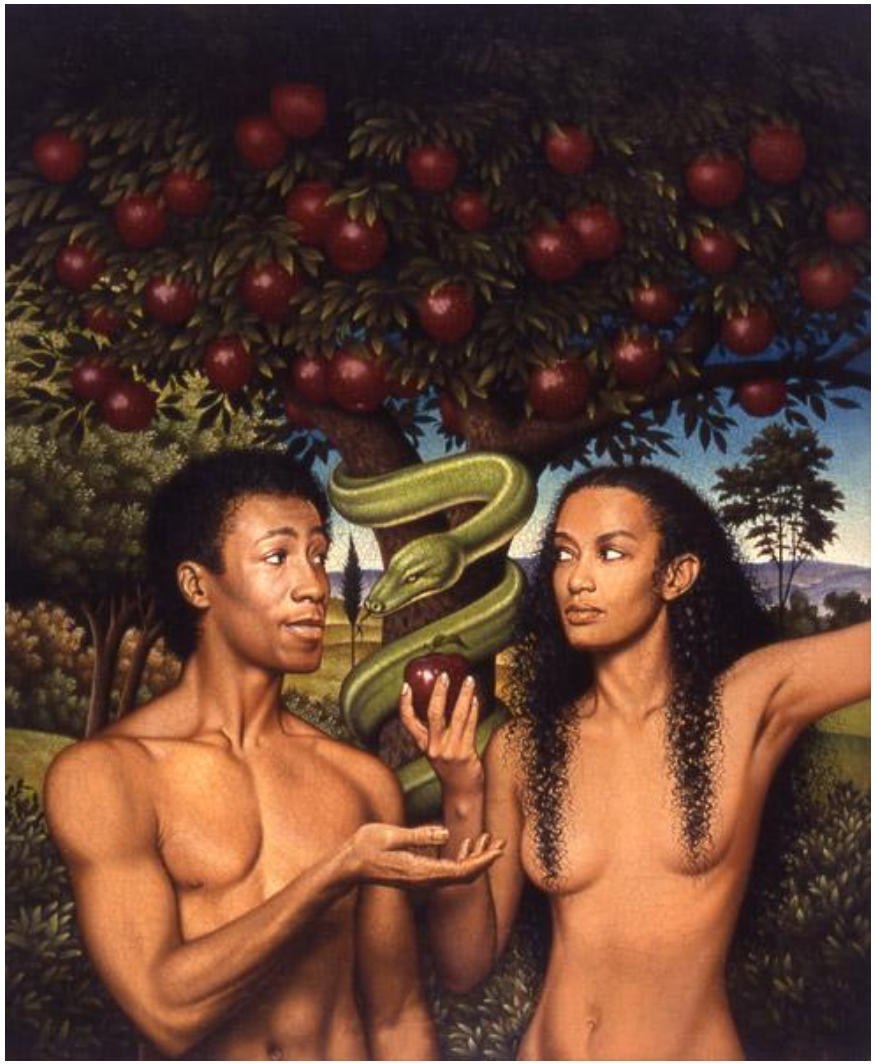
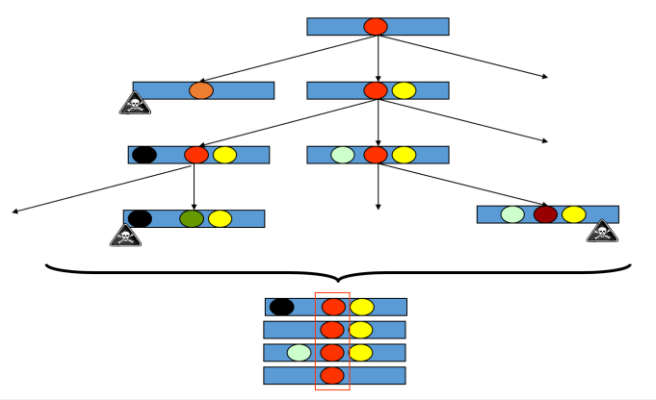
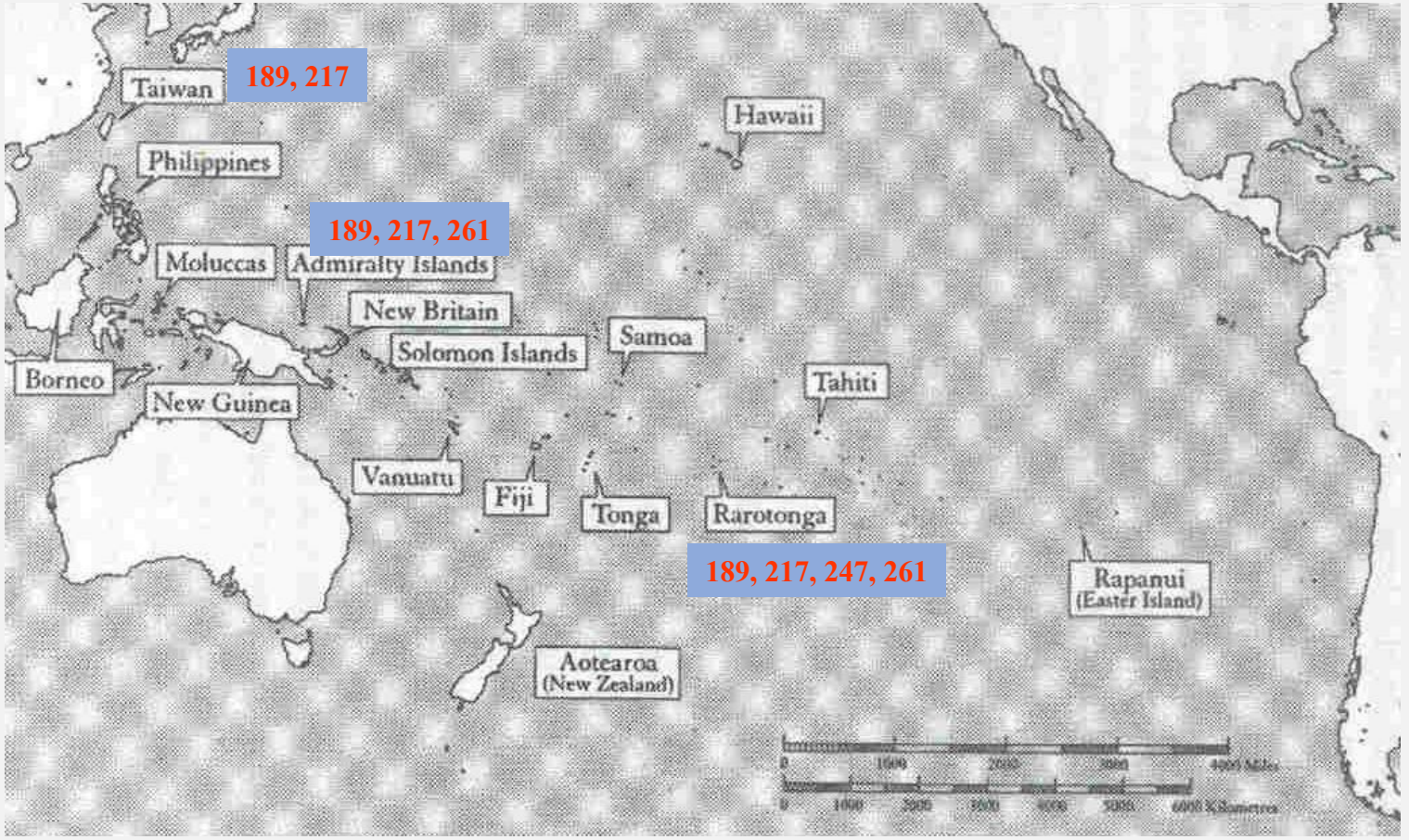


Image credit: Newsweek

Origin of Polynesians

Do they come from Asia or America?





Origin of Polynesians, cont'd

Common mitochondrial control sequences from Rarotonga have variants at positions 189, 217, 247, 261. Less common ones have 189, 217, 261

Sequences from Taiwan natives have variants 189, 217

Sequences from regions in between have variants 189, 217, 261

More 189, 217 closer to Taiwan

More 189, 217, 261 closer to Rarotonga

247 not found in America

$\Rightarrow$ Polynesians came from Taiwan!

Taiwan sequences sometimes have extra mutations not found in other parts

$\Rightarrow$These are mutations that happened since Polynesians left Taiwan!

Neanderthal vs Cro Magnon

Are Europeans descended purely from Cro Magnons? Pure Neanderthals? Or mixed?

Neanderthal

Cro Magnon

Analysis in Brian Sykes' popular book

Based on palaeontology, Neanderthal & Cro Magnon last shared an ancestor 250,000 yrs ago

Mitochondrial control regions accumulate 1 mutation per 10,000 yrs

If Europeans have mixed maternal ancestry, the mitochondrial control regions between 2 Europeans should have ~25 diff w/ high probability

The average number of differences between Welsh is ~3, at most 8

When compared w/ other Europeans, 14 differences at most

$\Rightarrow$ Maternal ancestor quite likely either 100% Neanderthal or 100% Cro Magnon

Mitochondrial control sequences from Neanderthal have 26 differences from Europeans

$\Rightarrow$ Maternal ancestor 100% Cro Magnon

Exercise

Brian Sykes' framing suggested a binary pure Cro-Magnon vs Neanderthal view

In reality,

Modern Europeans have mixed ancestry, not all lineages come from Cro-Magnons

No Neanderthal mtDNA survived in modern Europeans

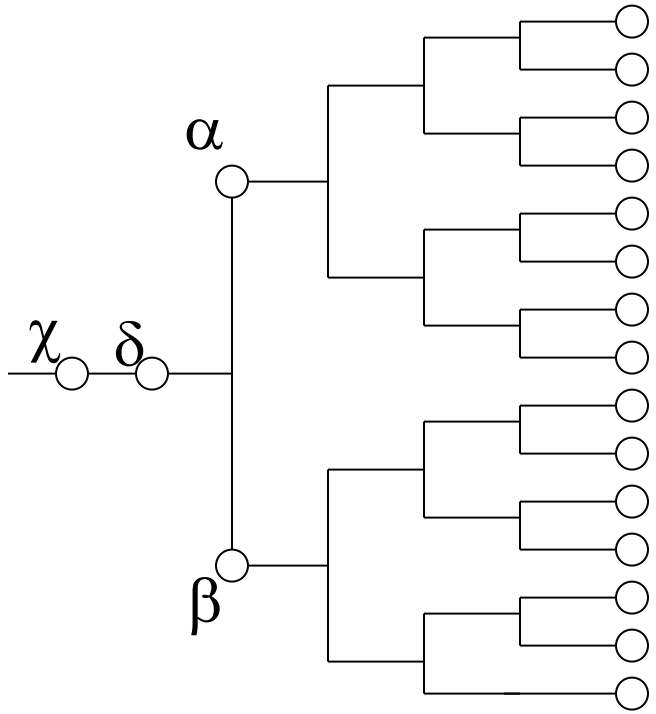
~2% of a modern European nuclear DNA is Neanderthal

Can you explain this?

Exercise

Clan mother is the most recent maternal ancestor common to all members of the clan

Which of α, β, χ, δ is the clan mother? Why?



How many clans in Europe?

Cluster sequences according to mutations

Each cluster thus represents a major clan

European sequences cluster into 7 major clans

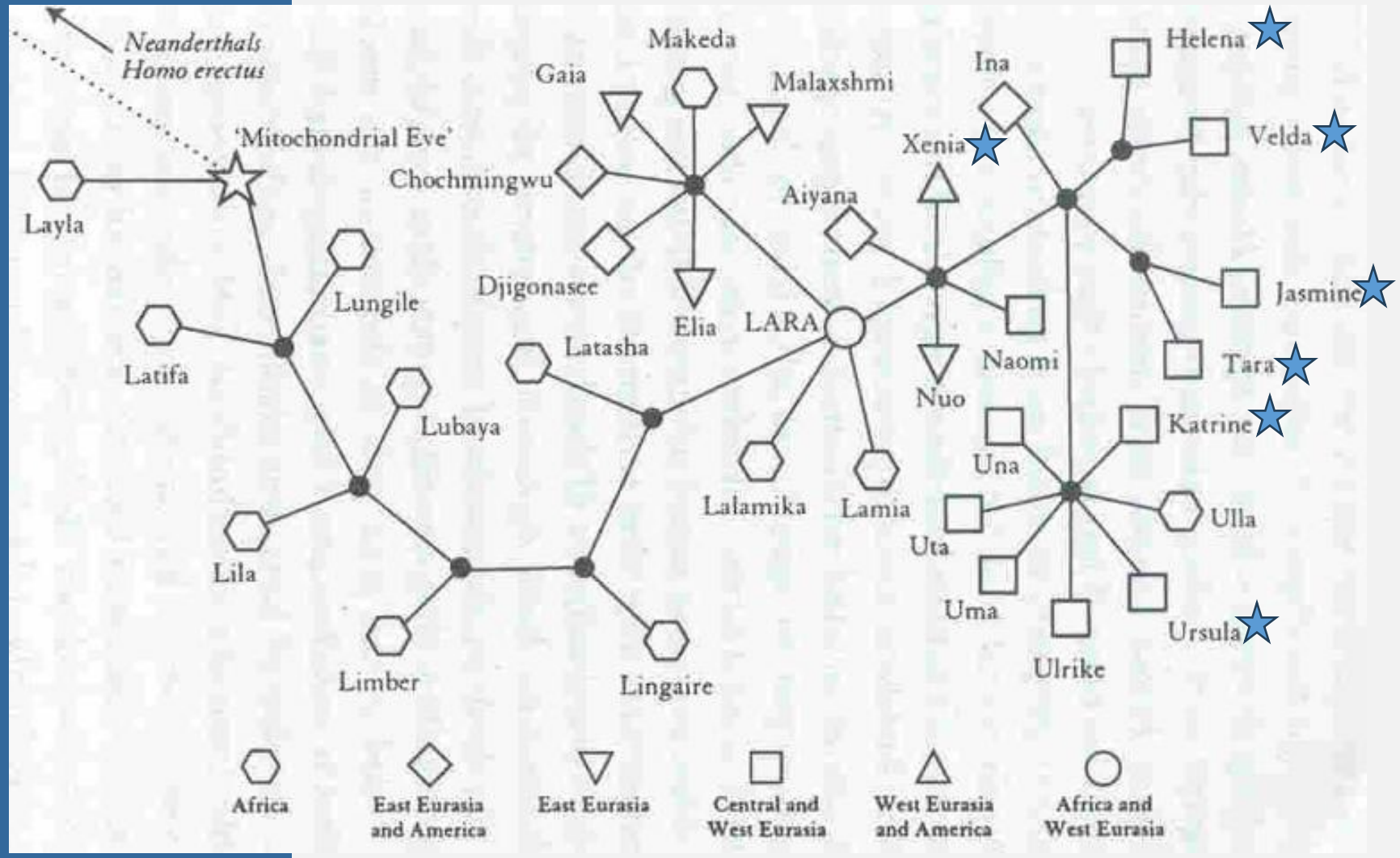
The 7 clusters age between 45,000 and 10,000 years (length of time for all mutations in a cluster to arise from a single founder sequence)

The founder sequence carried by just 1 woman in each case---the clan mother

Note that the clan mother did not need to be alone. There could be other women, it was just that their descendants eventually died out

7 daughters of Eve and other world clans according to Brian Sykes

Take this with a grain of salt!



Nobel Prize 2022

Discoveries concerning the genomes of extinct hominins and human evolution

Neaderthal-Eurasian inbreeding, via nuclear DNA sequence analysis

Europeans with Neanderthal heritage are at greater risks to develop more severe COVID-19 disease

Svante Pääbo

ForMemRS

Pääbo in 2016

Distance-based phylogeny-reconstruction

Distance between species

Try to minimize # of mutations

Species which look similar should be evolutionary more related

Define distance between two species to be # of mutations needed to change one species to another

Try to construct a phylogeny based on distance info among species

Finding distance between two species

Consider two species with these DNA fragments:

Species i: (A, C, G, C, T)

Species j: (C, C, A, C, T)

2 mismatches, so can estimate distance to be 2

Reasonable, as 2 mismatches can be thought of as 2 mutations

However, this fails to capture “multiple” mutations on the same site

In practice, need to apply some corrective distance transformation

Distance-based methods: Specification

Input: Distance matrix M satisfying constraints

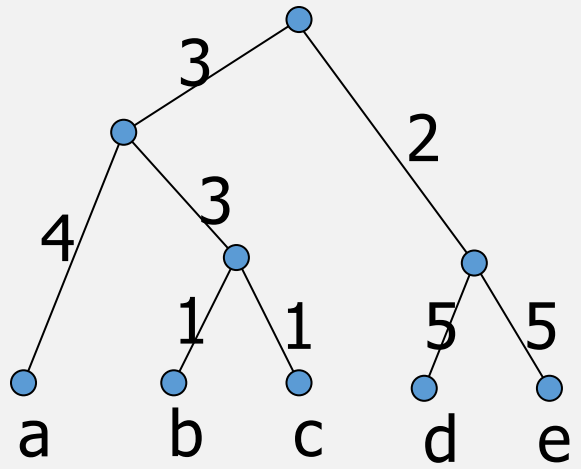
M should satisfy metric space properties

M is an additive metric

M is ultrametric (optional)

M	a	b	c	d	e
a	0	8	8	14	14
b	8	0	2	14	14
c	8	2	0	14	14
d	14	14	14	0	10
e	14	14	14	10	0

Output: Tree of degree 3 consistent with M



Metric space

A distance metric M which satisfies

Symmetry, $M_{ij} = M_{ji} \geq 0$

Self identity, $M_{ii} = 0$

Triangular inequality, $M_{ij} + M_{jk} \geq M_{ik}$

Additive metric

Let S be a set of species and M be distance matrix for S

If there is a tree T where:

Every edge has a positive weight

Every leaf is labeled by a distinct species in S; and

For every i, j $\in$ S, M_{ij} = the sum of the edge weights along the path from i to j

Then M is called an additive metric

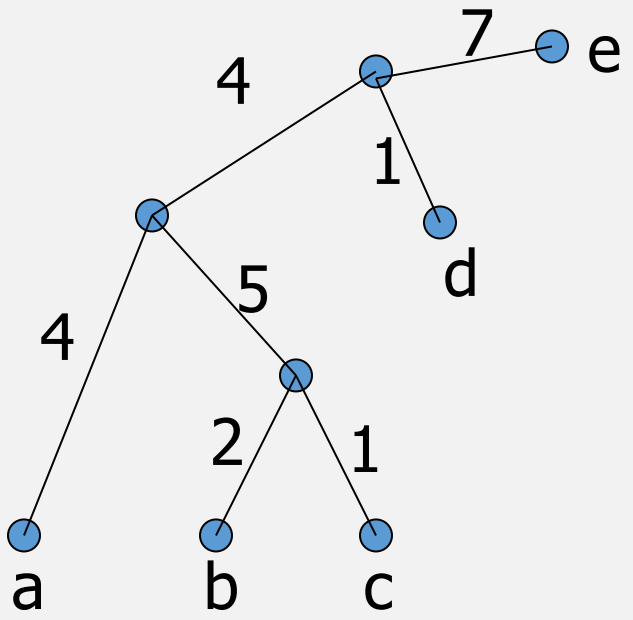
The corresponding tree T is called additive tree

Additive metric example

Don’t know the root!

We can only build an unrooted phylogeny

	a	b	c	d	e
a	0	11	10	9	15
b	11	0	3	12	18
c	10	3	0	11	17
d	9	12	11	0	8
e	15	18	17	8	0



Why additive metric?

Distance captures actual # of mutations between a pair of species

If (1) the correct tree for a set of species is known and (2) we get the exact # of mutations for each edge, the distance (the # of mutations) between two species i and j should be the sum of the edge weights along the path from i to j

Additive metric seems reasonable

Buneman's 4-point condition

M is additive if and only if

for every four species in S,

we uniquely partition them into { i, j }, { k, l } such that

$$M_{ik} + M_{jl} = M_{il} + M_{jk} \geq M_{ij} + M_{kl}$$

i.e., of the 3 sums, two are equal and large than the 3rd

Based on the 4-point condition, we can check whether a matrix M is additive

Half of the proof

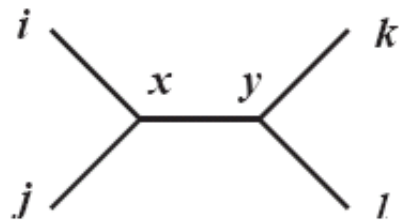


Figure 8.3: Buneman's 4-Point Condition

$$
\begin{aligned}
& M_{ik} + M_{jl} \\
&= (M_{ix} + M_{xy} + M_{yk}) + (M_{jx} + M_{xy} + M_{yl}) \\
&= M_{ix} + M_{jx} + M_{yk} + M_{yl} + 2M_{xy}
\end{aligned}
$$

$$
\begin{aligned}
& M_{jk} + M_{il} \\
&= (M_{jx} + M_{xy} + M_{ik}) + (M_{ix} + M_{xy} + M_{yl}) \\
&= M_{ix} + M_{jx} + M_{yk} + M_{yl} + 2M_{xy}
\end{aligned}
$$

$$
\begin{aligned}
& M_{ij} + M_{kl} \\
&= M_{ix} + M_{xj} + M_{ky} + M_{yl}
\end{aligned}
$$

So it can be easily verified that: $M_{ik} + M_{jl} = M_{il} + M_{jk} \geq M_{ij} + M_{kl}$.
($\Leftarrow$) Will not present here. ■

Peter Buneman

JOURNAL OF COMBINATORIAL THEORY (B) 17, 48–50 (1974)

A Note on the Metric Properties of Trees*

PETER BUNEMAN*

Communicated by Frank Harary

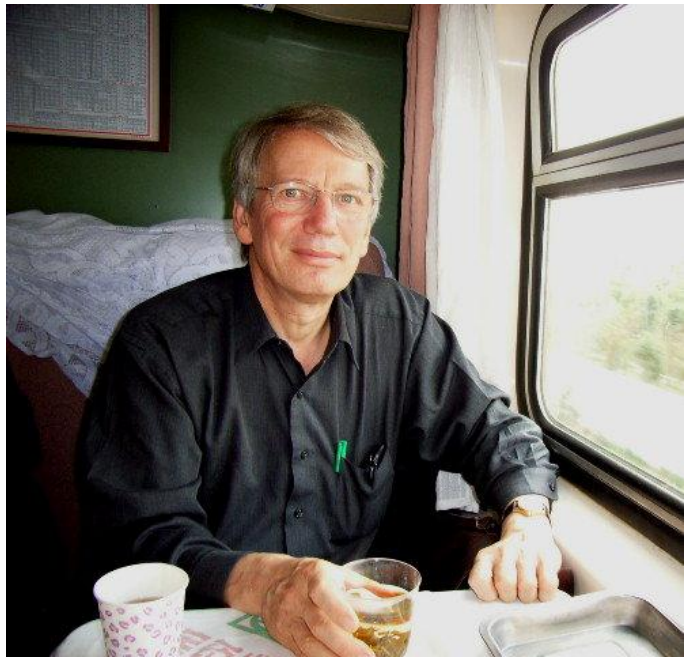
Received February 21, 1973

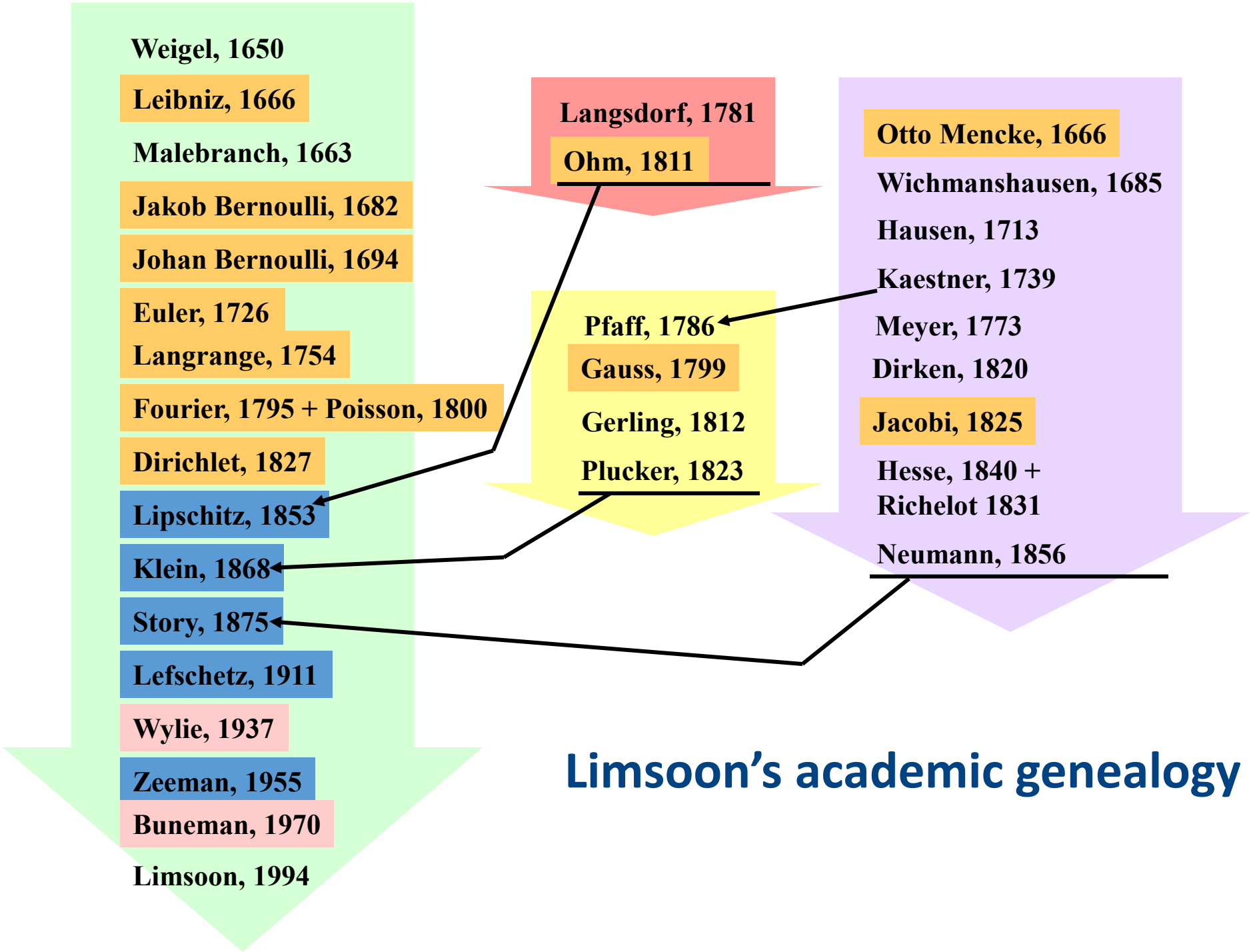
By checking the possible configurations of paths which can connect four points x, y, z, t in a tree, it can be seen that the graphical distance [1] must satisfy the inequality:

$$d(x, y) + d(z, t) \leqslant \max \begin{cases} d(x, z) + d(y, t), \\ d(x, t) + d(y, z). \end{cases}$$

We shall refer to this condition as the four-point condition: it is stronger than the triangle inequality (put $z = t$) and is equivalent to saying that of the three sums $d(x, y) + d(z, t)$, $d(x, z) + d(y, t)$, and $d(x, t) + d(y, z)$ two are equal and not less than the third. The four-point condition is also a sufficient condition for a graph to be a tree in the following sense.

THEOREM 1. *A graph is a tree iff it is connected, contains no triangles, and has graphical distance satisfying the four-point condition.*

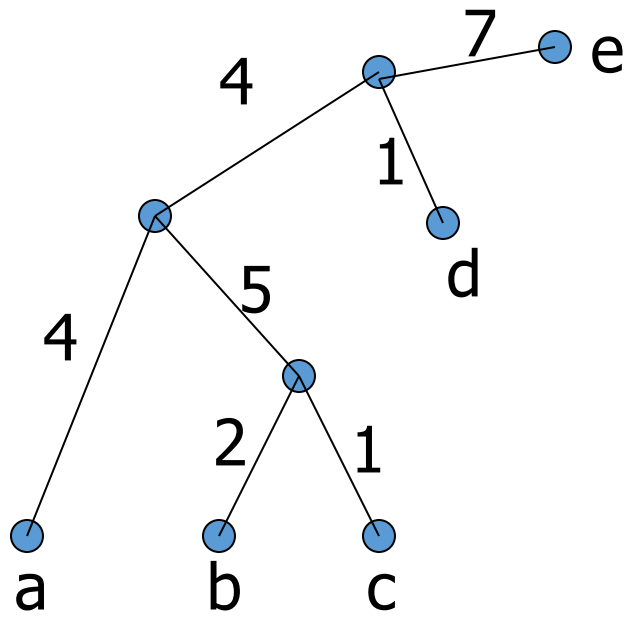




Limsoon's academic genealogy

Exercise

	a	b	c	d	e
a	0	11	10	9	15
b	11	0	3	12	18
c	10	3	0	11	17
d	9	12	11	0	8
e	15	18	17	8	0



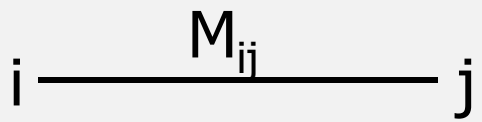
Pick any 4 species

Is 4-point condition $M_{ik} + M_{jl} = M_{il} + M_{jk} \geq M_{ij} + M_{kl}$ satisfied?

Additive-tree reconstruction

Suppose M is an additive metric. We show an algorithm which reconstructs the additive tree in $O(n^2)$ time

For any two species i and j, the additive tree is just an edge with weight M_{ij}



Additive tree for 3 species: 3-star method

For any three species i, j, k, we can find their center c as follows

Let d_{xy} be the length of the path from x to y

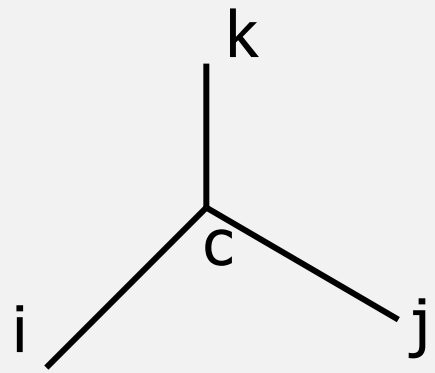
Constraints on c: $M_{ik} = d_{ic} + d_{ck}$, $M_{jk} = d_{jc} + d_{ck}$, $M_{ij} = d_{ic} + d_{cj}$

By solving the three equations, we have

$d_{ic} = (M_{ij} + M_{ik} - M_{jk})/2$

$d_{jc} = (M_{ij} + M_{jk} - M_{ik})/2$

$d_{kc} = (M_{ik} + M_{jk} - M_{ij})/2$



Note: The resulting tree is unique!

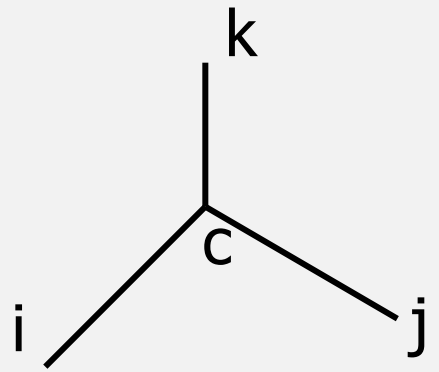
Additive tree for 4 species

Given four species h, i, j, k, we want to recover the additive tree

For species i, j, k, we get the additive tree using the 3-star method

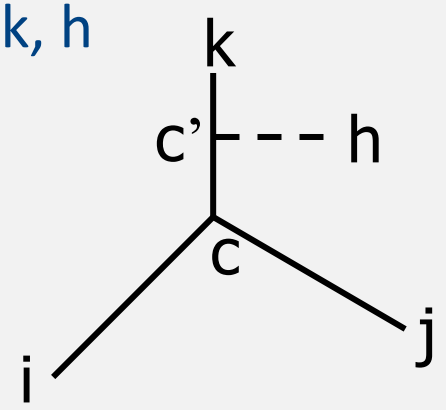
To include h into the tree, we need to introduce one more internal node c’

c’ splits either (i, c), (j, c) or (k, c)



Additive tree for 4 species, cont'd

To check whether c' splits (k, c), we apply 3-star method for species i, k, h



If $d_{kc'} < d_{kc}$, then c' splits (k, c)

Otherwise, use the same approach to check whether c' splits (i, c) or (j, c)

Note: c' can only split exactly one edge. Thus, the additive tree for 4 species is unique

Additive tree for k species

Inductively, assume we know how to recover the additive tree for k-1 species

To recover the additive tree for k species,

Build the additive tree T' for the first k-1 species. Then, insert the last species to T'

The last species should split one of the edges in T'

For every edge in T', we check (using 3-star method) whether the last species splits it

The time required is O(k-1) for each step

Also, each insertion is unique!

Time complexity

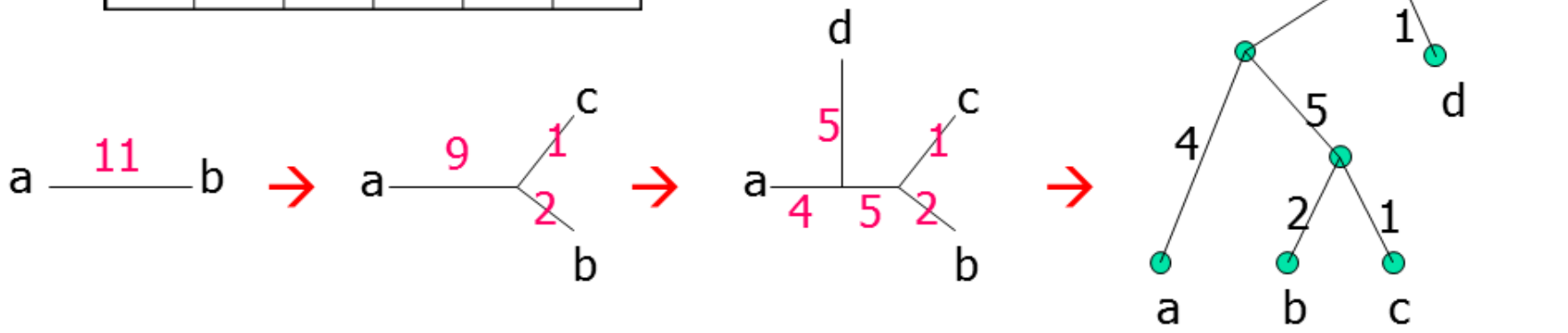
In summary, to recover an additive tree with n species, the time is

$$O(1 + 2 + \ldots + n) = O(n^2)$$

The resulting additive tree for M is unique

Example

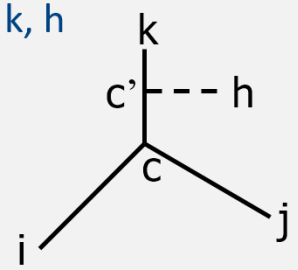
M	a	b	c	d	e
a	0	11	10	9	15
b	11	0	3	12	18
c	10	3	0	11	17
d	9	12	11	0	8
e	15	18	17	8	0



Exercise

Additive tree for 4 species, cont'd

To check whether c' splits (k, c), we apply 3-star method for species i, k, h



If $d_{kc'} < d_{kc}$, then c' splits (k, c)

Otherwise, use the same approach to check whether c' splits (i, c) or (j, c)

Note: c' can only split exactly one edge. Thus, the additive tree for 4 species is unique

Why c' can only split exactly one edge?

Hint: Buneman's 4-point condition

Ultrametric

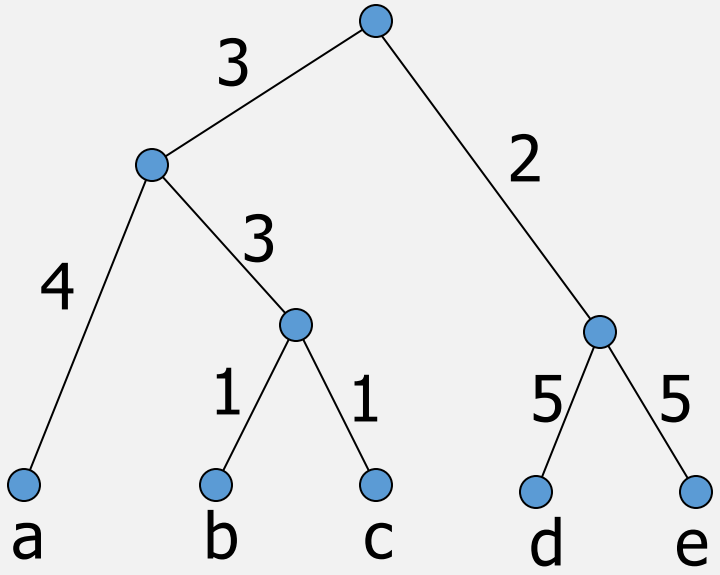
Assume M is additive. i.e., there is a tree T such that the distance between any two species i and j equals the sum of the edge weights along the path from i to j

If we can further identify a root such that the path length from the root of T to every leaf is identical, then M is called an ultrametric

A tree T that satisfies ultrametric is an ultrametric tree

Ultrametric example

	a	b	c	d	e
a	0	8	8	14	14
b	8	0	2	14	14
c	8	2	0	14	14
d	14	14	14	0	10
e	14	14	14	10	0



Every path from root to leaf has the same length

Buneman's 3-point condition

Ultrametric is an additive metric $\Rightarrow$ It satisfies 4-point condition

It has an additional property:

M is ultrametric if and only if

for every three species in S,

we can label them i, j, k such that

$$M_{ik} = M_{jk} \geq M_{ij}$$

Based on the 3-point condition, we can check whether a matrix M is ultrametric

Half of the proof

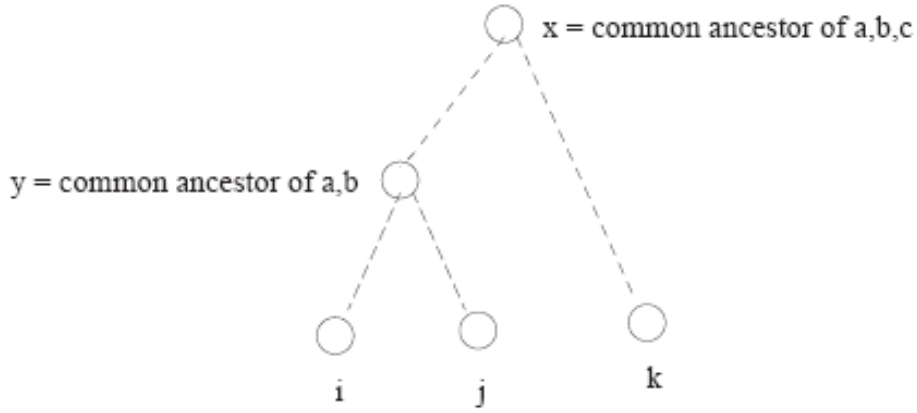


Figure 8.4: Ultrametric Tree

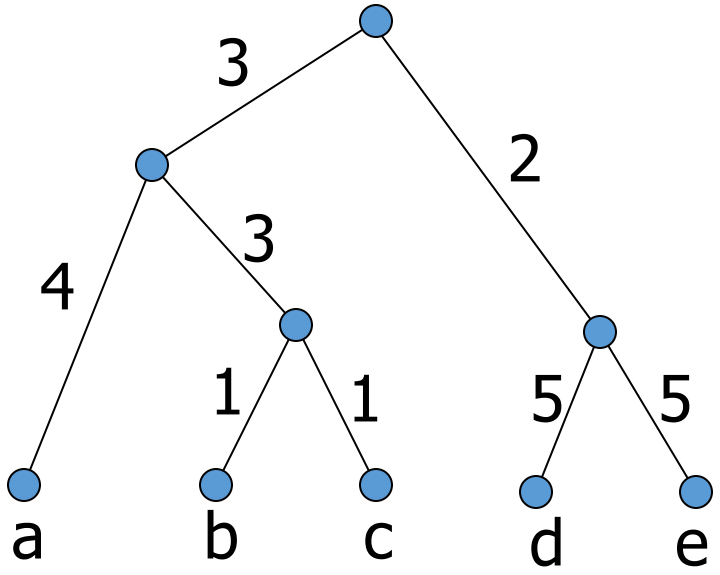
From the above formulas,and by Property 3 of an Ultrametric tree. There is

$$M_{ik} = M_{jk} = 2 * (M_{iy} + M_{yx}) > 2M_{iy} = M_{iy} + M_{jy} = M_{ij}$$

proven!
($\Leftarrow$)Exercise. ■

Exercise

	a	b	c	d	e
a	0	8	8	14	14
b	8	0	2	14	14
c	8	2	0	14	14
d	14	14	14	0	10
e	14	14	14	10	0



Pick any 3 species

Is 3-point condition $M_{ik} = M_{jk} \geq M_{ij}$ satisfied?

Constant molecular clock

Constant molecular clock is an assumption in biology

The # of accepted mutations in any time interval is proportional to the length of that interval

All species evolved at equal rate from a common ancestor

Ultrametric tree states that distance from root to all species are the same

Its correctness is based the constant molecular clock assumption

Some computational problems

Let M be a distance matrix for a set of species S

If M is ultrametric, can we reconstruct the corresponding ultrametric tree T in polynomial time?

If M is additive, can we have a polynomial time algorithm to recover the corresponding additive tree T?

If M is not exactly additive, can we find the nearest additive tree T?

Ultrametric-tree reconstruction

Input: An ultrametric matrix M for a set of species S

Problem: Reconstruct the phylogenetic tree T for S

Unweighted Pair Group Method With Arithmetic Mean (UPGMA)

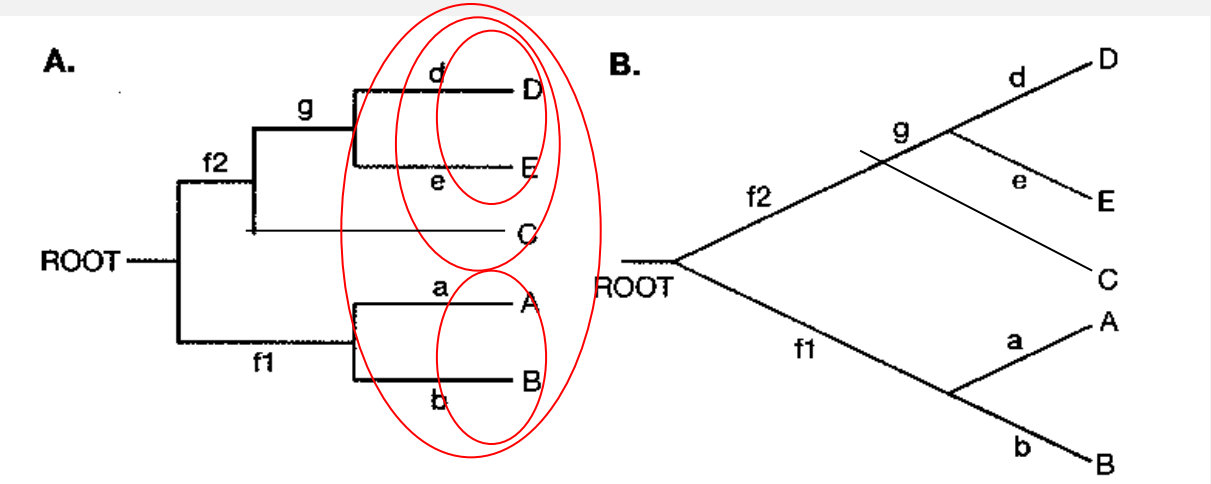
Consider ultrametric tree T

If a subset of species S forms a subtree of T, we call it a cluster

Idea:

Every species forms a cluster

Iteratively connect two nearest clusters, until one cluster is left



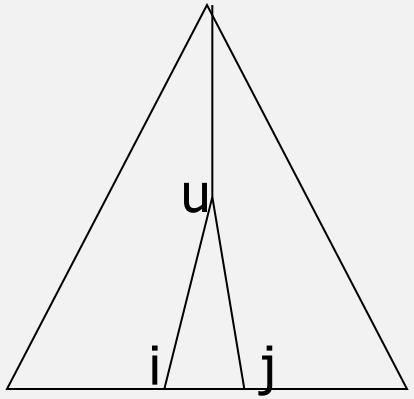
Height

For a node u, define height(u) be path length from u to any of its descendent leaf

Since T is ultrametric, every path should have the same length

Let i and j be descendent leaves of u in two different subtrees

To ensure that distance from the root to both i and j are the same, height(u) = $M_{ij}/2$



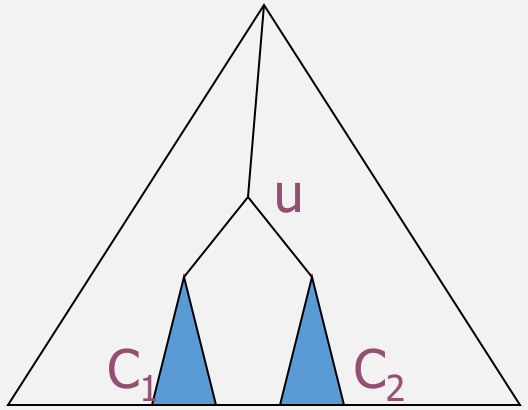
Distance between two clusters

For any two clusters C1 and C2 of T, define

$$dist(C_1, C_2) = \frac{\sum_{i \in C_1, j \in C_2} M_{ij}}{|C_1| \cdot |C_2|}$$

Note that dist(C_1, C_2) = M_{ij} for all i ∈ C_1 and j ∈ C_2 Why?

Let u be lowest common ancestor of i and j, dist(C_1, C_2) = 2 height(u)



Observation

For any non-intersecting clusters C_1,C_2, D,

$$dist(C_1, D) = \frac{\sum_{i \in C_1, j \in D} M_{ij}}{|C_1| \cdot |D|}$$

$$dist(C_2, D) = \frac{\sum_{i \in C_2, j \in D} M_{ij}}{|C_2| \cdot |D|}$$

$$\begin{aligned} dist(C_1 \cup C_2, D) &= \frac{\sum_{i \in C_1 \cup C_2, j \in D} M_{ij}}{|C_1 \cup C_2| \cdot |D|} \\ &= \frac{\sum_{i \in C_1, j \in D} M_{ij} + \sum_{i \in C_2, j \in D} M_{ij}}{|C_1 \cup C_2| \cdot |D|} \\ &= \frac{|C_1||D| dist(C_1, D) + |C_2||D| dist(C_2, D)}{|C_1 \cup C_2| \cdot |D|} \\ &= \frac{|C_1| dist(C_1, D) + |C_2| dist(C_2, D)}{|C_1 \cup C_2|} \end{aligned}$$

Algorithm

Given n x n ultrametric distance matrix M

Initialize set Z to consist of n initial singleton clusters {1}, {2}, …, {n}

For all {i}, {j} $\in$ Z, initialize dist({i}, {j}) = M_{ij}

Repeat n-1 times

Determine cluster A, B $\in$ Z where dist(A, B) is min

Define a new cluster C = A $\cup$ B

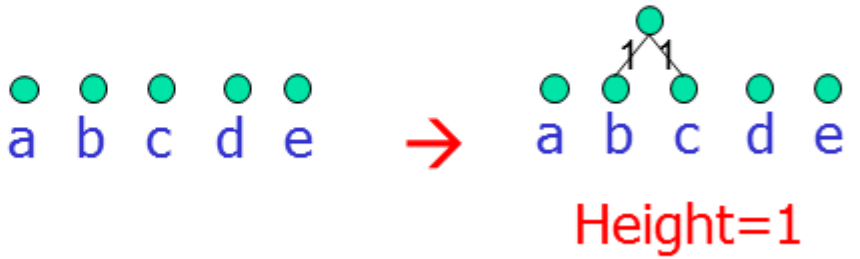
Z := Z – {A, B} $\cup$ {C}

Define new node c and let c be parent of A and B. Also, define height(c) = dist(A, B)/2

For all D $\in$ Z – {C}, define dist(D, C) = dist(C, D) = (|A| dist(A, D) + |B| dist(B, D)) / (|A| + |B|)

Example

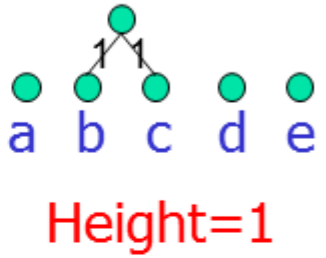
M	a	b	c	d	e
a	0	8	8	14	14
b	8	0	2	14	14
c	8	2	0	14	14
d	14	14	14	0	10
e	14	14	14	10	0



M	a	b,c	d	e
a	0	8	14	14
b,c	8	0	14	14
d	14	14	0	10
e	14	14	10	0

Example

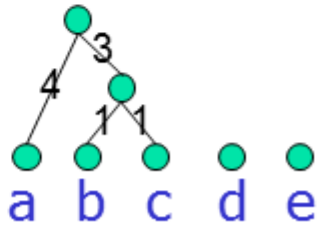
M	a	b,c	d	e
a	0	8	14	14
b,c	8	0	14	14
d	14	14	0	10
e	14	14	10	0



Height=1

↓

M	a,b,c	d	e
a,b,c	0	14	14
d	14	0	10
e	14	10	0

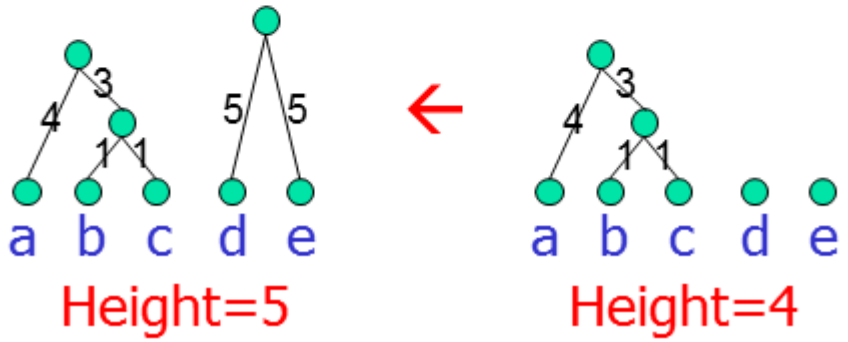


Height=4

Example

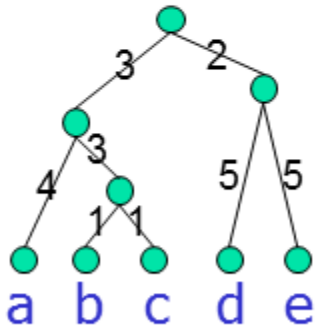
M	a,b,c	d,e
a,b,c	0	14
d,e	14	0

M	a,b,c	d	e
a,b,c	0	14	14
d	14	0	10
e	14	10	0

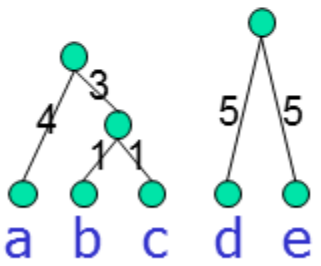


Example

M	a,b,c	d,e
a,b,c	0	14
d,e	14	0



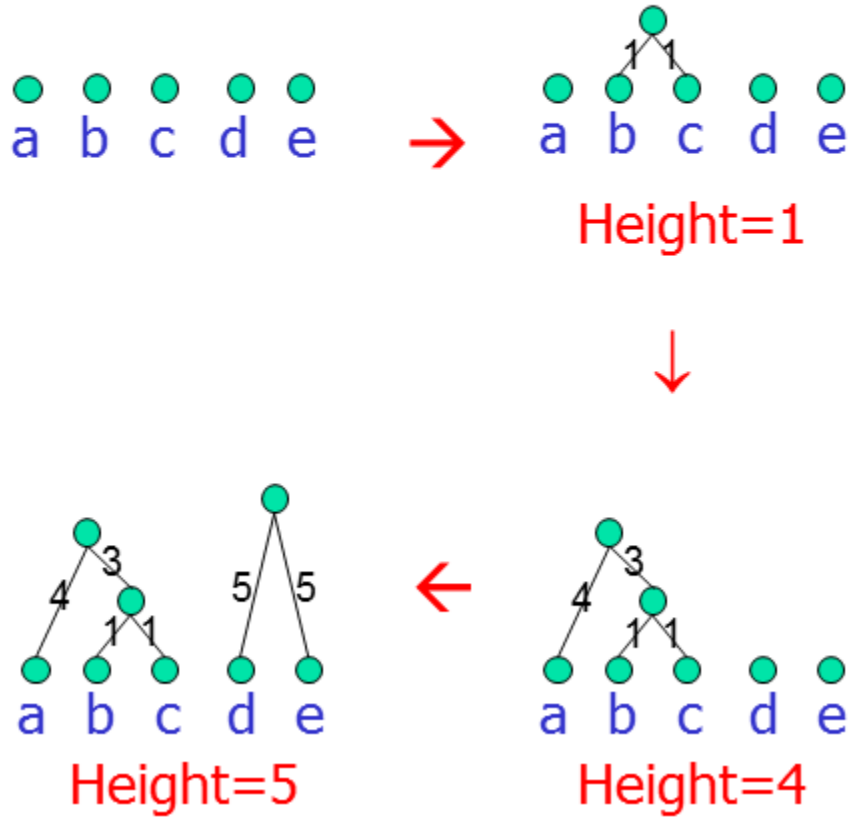
Height=7



Height=5

Example

M	a	b	c	d	e
a	0	8	8	14	14
b	8	0	2	14	14
c	8	2	0	14	14
d	14	14	14	0	10
e	14	14	14	10	0



Exercise

What is the time complexity of UPGMA?

Can it be modified to run in quadratic time?

Reconstruct nearly additive tree

If M is not an additive metric, we can find the nearly additive tree using the following methods

Least Squares Method

Fitch-Margoliash method

Neighbor-Joining Method

L∞-metric

I will show you just the formulation of one of these…

Least-squares method

Input: a metric M for a set of species S

For any tree T for the set of species S, let D be its corresponding distance matrix
Define

$$SSQ(T) = \sum_{i=1}^{n} \sum_{j \neq i} \frac{(D_{ij} - M_{ij})^2}{D_{ij}^2}$$

Aim: Find a tree T minimizing SSQ(T). Such tree is known as Least Squares Tree

This problem is NP-hard

Can tree reconstruction methods infer the correct tree?

Experimentally, bacteriophage T7 was propagated and split sequentially in the presence of a mutagen, where each lineage was tracked

Five different phylogenetic methods were used independently, and each one chose the correct tree, out of 135,135 possible phylogenetic trees

D. M. Hillis et al. Experimental phylogenetics: generation of a known phylogeny. *Science,* 255(5044):589-592, 1992

Can tree reconstruction methods infer the correct tree?

In 1998, researchers used 111 modern HIV-1 (AIDS virus) sequences in a phylogenetic analysis to predict the nucleotide sequence of the viral ancestor of which they were all descendants

The predicted ancestor sequence closely matched, with high statistical probability, an actual ancestral HIV sequence found in an HIV-1 seropositive African plasma sample collected and archived in the Belgian Congo in 1959

T. Zhu et al. An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. *Nature,* 391: 594-597, 1998

A popular tool for reconstructing phylogenetic trees

Felsenstein's PHYLIP

Large # of methods, including maximum likelihood, maximum parsimony and neighbor joining

Command-line mode only

It is the most widely used program suite

Source code is available

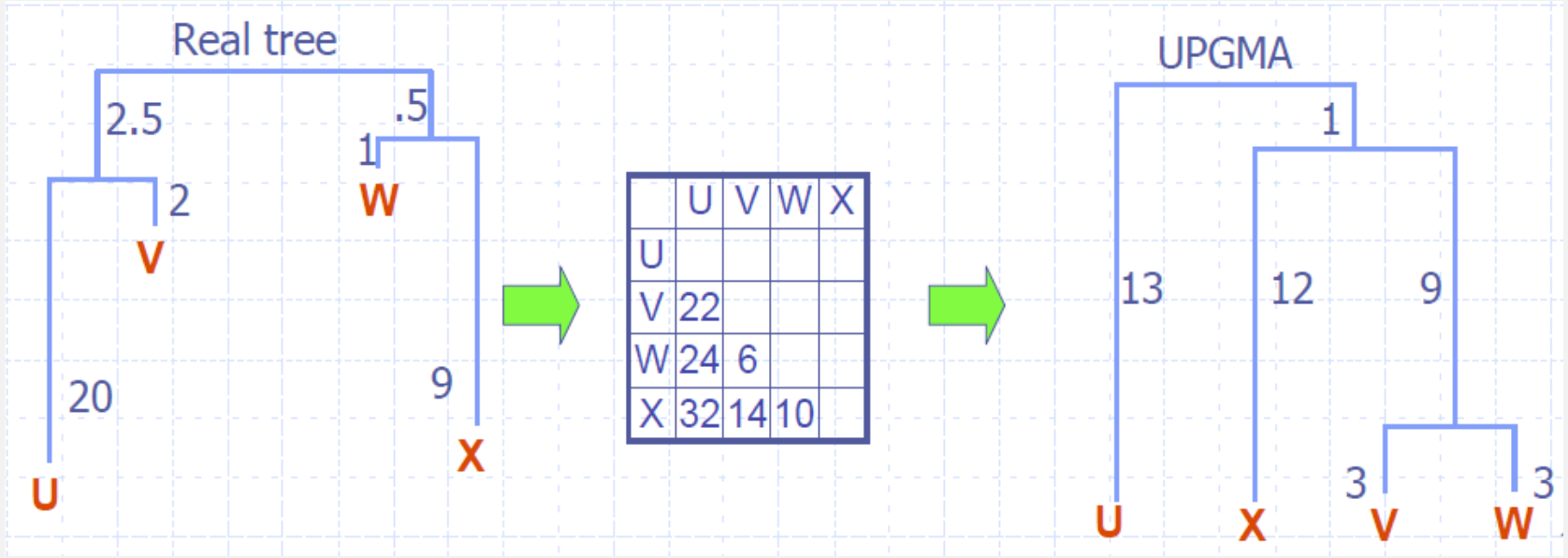
Free of charge

http://evolution.genetics.washington.edu/phylip.html

Robustness of reconstructions

Cautionary note

UPGMA's simple-minded clustering may lead to substantial errors

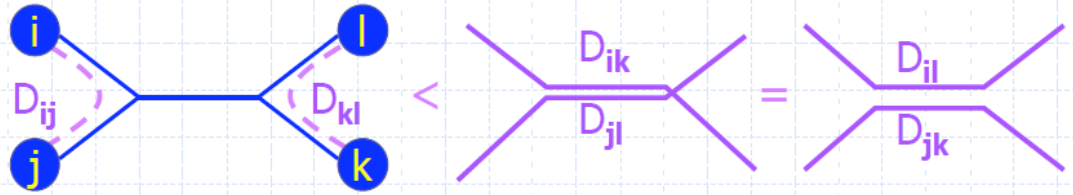


Source: Yechiam Yemini

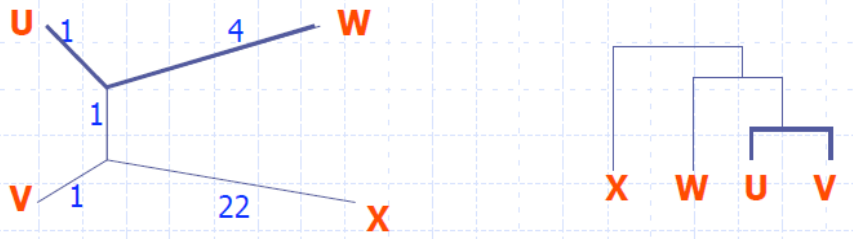
What causes the problem?

Closest Pair vs. Evolutionary-Neighbors

- Additivity: $D_{ij}+D_{kl} < D_{ik}+D_{jl} = D_{il}+D_{jk}$



- UPGMA overcomes non-additivity by averaging distances
- But, the closest pair may not be evolutionary neighbors
- The evolutionary tree distances may diverge greatly; averaging distorts neighborhood



Source: Yechiam Yemini

Bootstrapping

A statistical technique to increase robustness

Scenario:
Given sample S and result R(S) computed from S

Bootstrapping:
Resample S, to get S';
Compute R(S');
Evaluate match of R(S) with the values R(S')

Credit: yechiam Yemini

Bootstrapping in phylogenetic tree

S = columns of sequences of size n; R(S)=tree

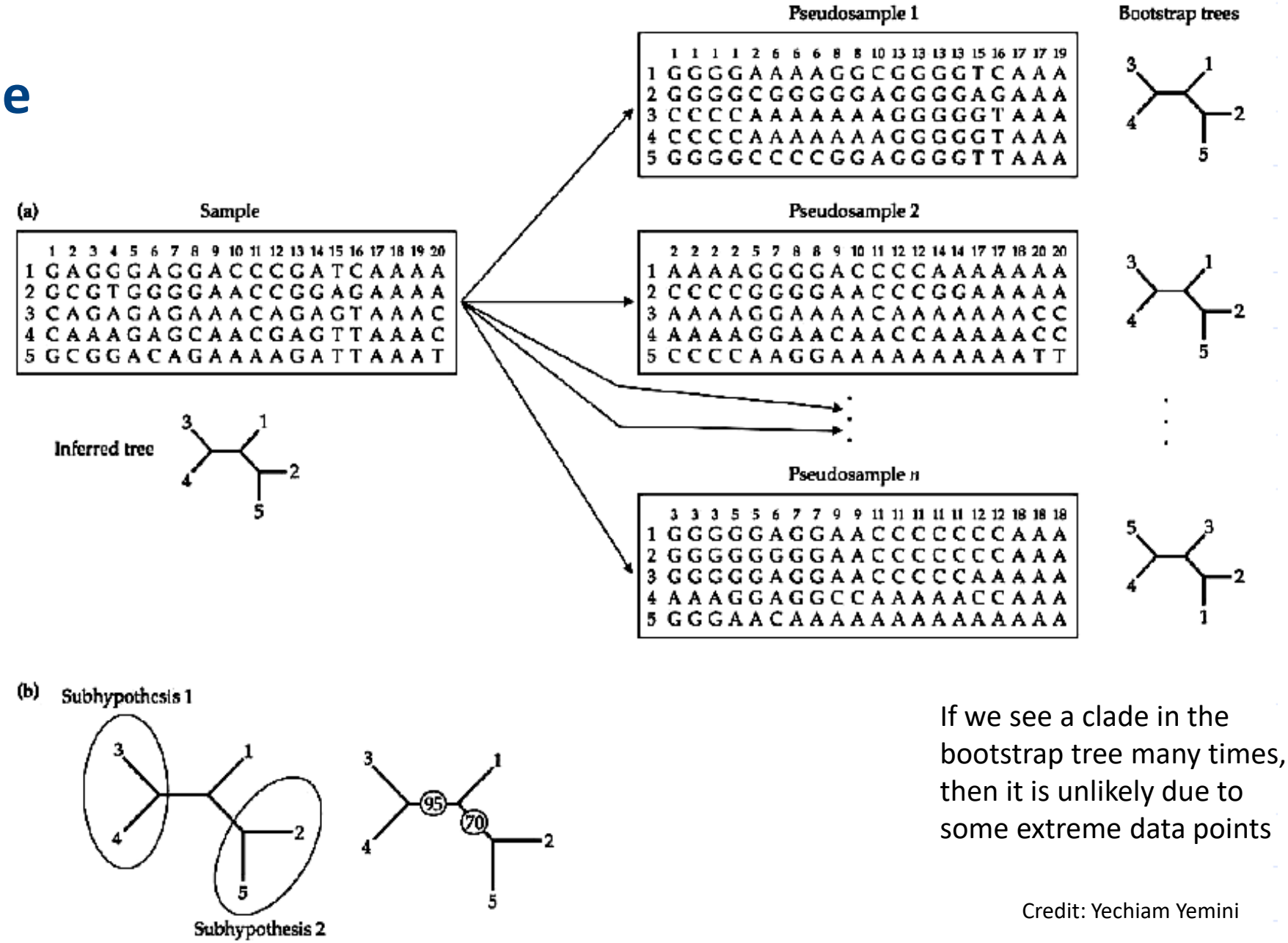
S’ = Sample n random columns of S with possible repetitions

Compute phylogenetic tree R(S’)

Use { R(S’) } to compute likelihood of branches of R(S)

Credit: Yechiam Yemini

Example



If we see a clade in the bootstrap tree many times, then it is unlikely due to some extreme data points

Credit: Yechiam Yemini

Phylogenetic tree comparison

Why tree comparison?

There are several methods to reconstruct phylogeny for the same set of species

Different phylogenies are resulted using
Different data (different segments of genomes)
Different model (Cavender-Farris-Neyman model, Jukes-Cantor Model)
Different reconstruction algorithms

Tree comparison helps us to gain information from multiple trees

Two types of comparisons

Similarity measurement: Find common structure among given trees

Maximum Agreement Subtree

Dissimilarity measurement: Determine differences among given trees

Robinson-Foulds distance

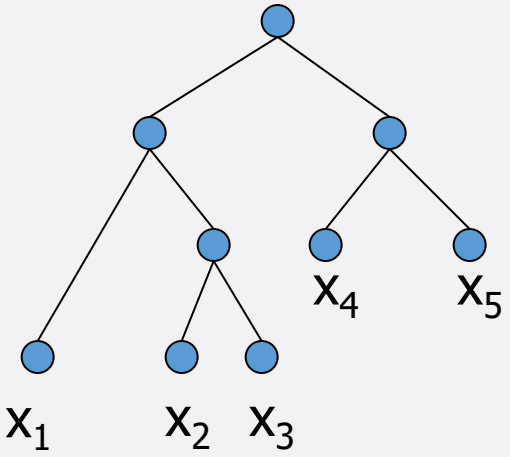
Nearest-neighbor interchange

Subtree transfer distance

In this lecture, we discuss the first method

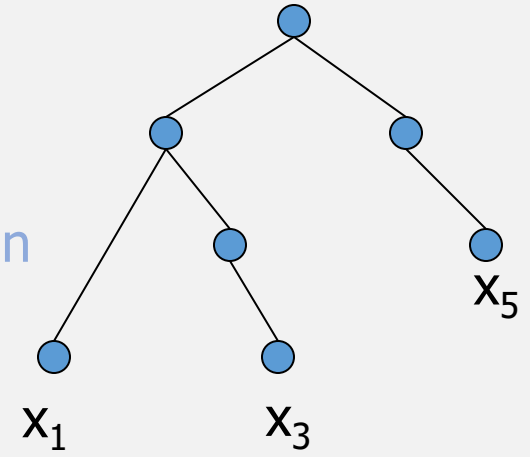
Restricted subtree

Consider tree T



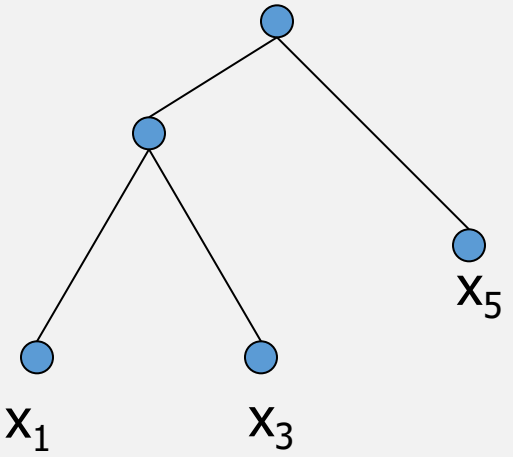
Evolution information of X_1, X_2, X_3, X_4, X_5

→ Restricted on X_1, X_3, X_5

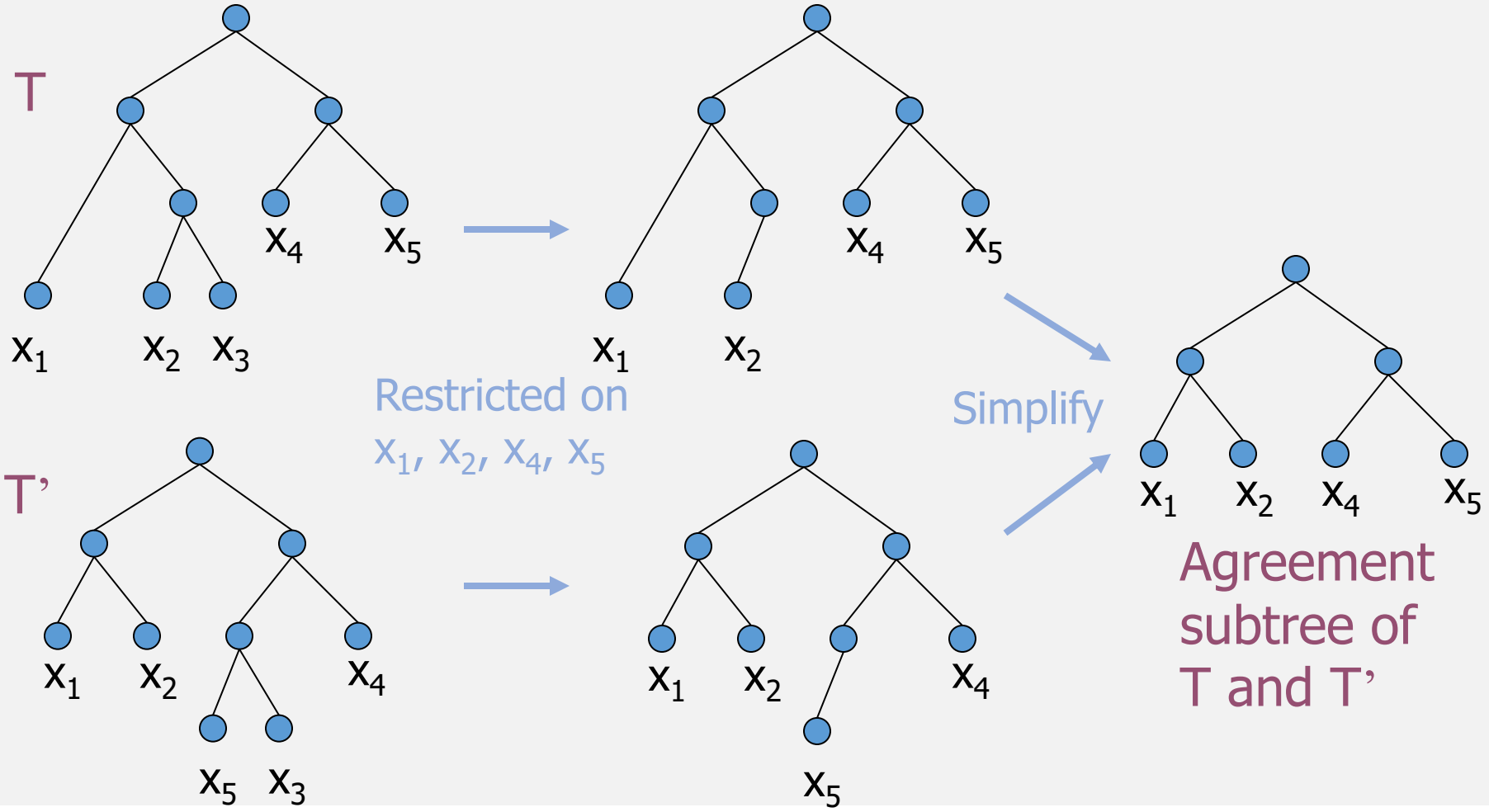


Evolution information of X_1, X_3, X_5

↘ Simplify



Agreement subtree



Maximum Agreement SubTree (MAST)

Given two trees T_1 and T_2

Agreement subtree of T_1 and T_2 is the common info agreed by both trees
Agreed by both trees $\Rightarrow$ evolution of the agreement subtree is more reliable

Maximum agreement subtree problem
Find the agreement subtree with largest possible number of leaves
Such agreement subtree is called the maximum agreement subtree

MAST for rooted trees

MAST of two degree-d rooted trees T_1 and T_2 with n leaves can be computed in

$$O(\sqrt{d}\, n \log(\tfrac{n}{d}))\ \text{time}$$

But the algo for the above is complicated

So here we show you a $O(n^2)$-time algorithm which computes the maximum agreement subtree of two binary trees with n leaves

MAST by dynamic programming

For any two binary rooted trees T_1 and T_2, let MAST(T_1, T_2) be number of leaves in the maximum agreement subtree

For a tree T and a node u, T^u is the subtree of T rooted at u

Base cases

For any leaf x in T_1 and y in T_2,

$$MAST(x, y) = \max \begin{cases} 1 \text{ if } x = y \\ 0 \text{ otherwise} \end{cases}$$

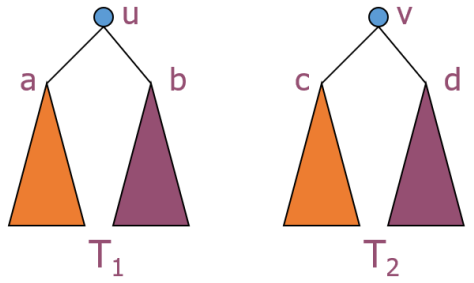
For any node u in T_1 and v in T_2,

$$MAST(T_1^u, \Lambda) = 0, MAST(\Lambda, T_2^v) = 0$$

Recurrence

$$MAST(T_1^u, T_2^v) =$$

$$\max \begin{cases} MAST(T_1^a, T_2^c) + MAST(T_1^b, T_2^d) \\ MAST(T_1^a, T_2^d) + MAST(T_1^b, T_2^c) \\ MAST(T_1^a, T_2^v) \\ MAST(T_1^b, T_2^v) \\ MAST(T_1^u, T_2^c) \\ MAST(T_1^u, T_2^d) \end{cases}$$



Time complexity

Suppose T_1 and T_2 are rooted phylogenies for n species

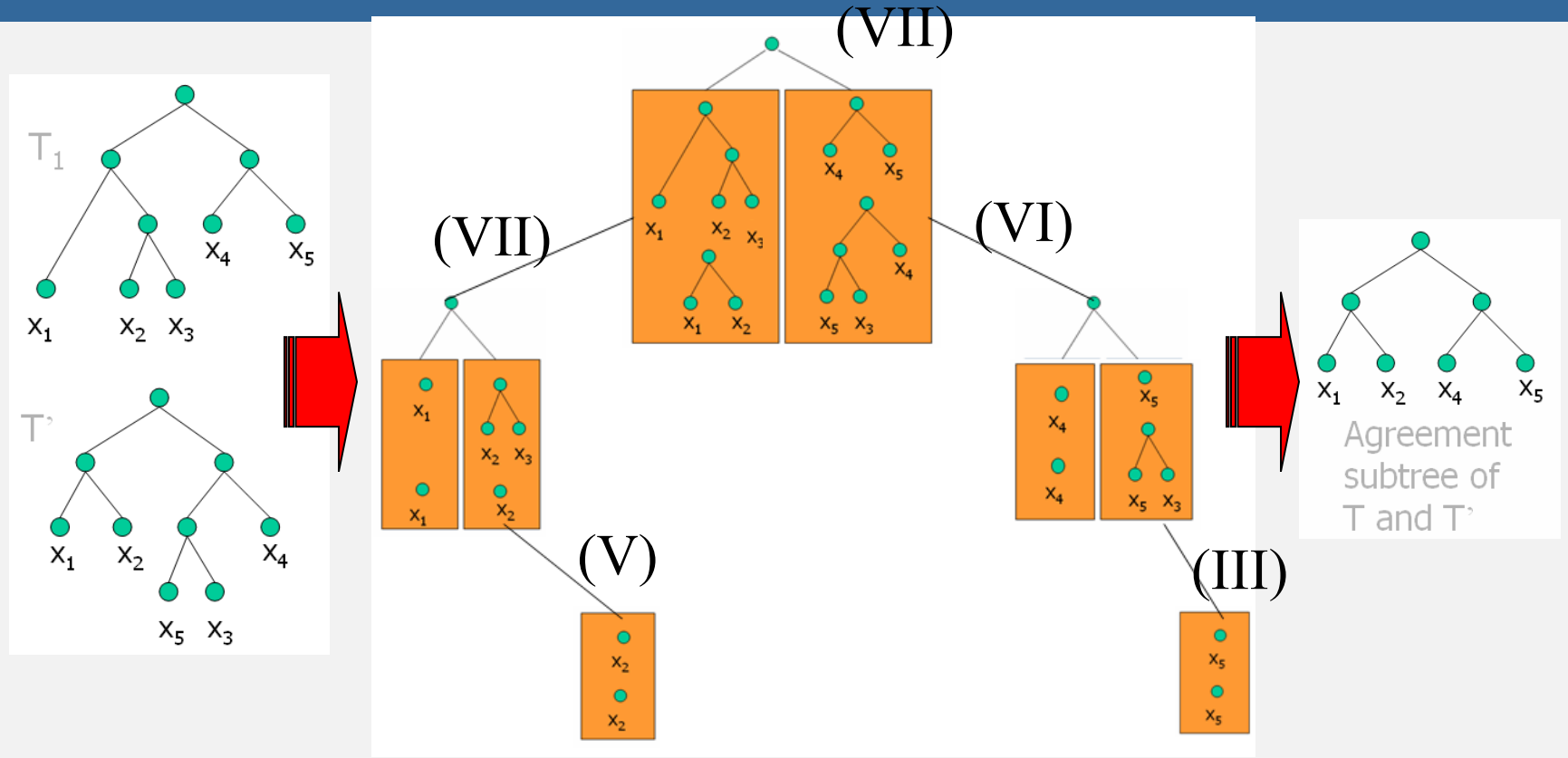
We have to compute MAST(T_1^u, T_2^v) for every u in T_1 and v in T_2

Thus, we need to fill in n^2 entries

Each entry can be computed in O(1) time using dynamic programming

In total, the time complexity is $O(n^2)$

MAST example



Acknowledgements

A lot of the slides from this lecture were given to me by Ken Sung

Many slides are also based on the lecture slides of Yechiam Yemini and Somayyeh Koohi

Good to read

B. Sykes. *The seven daughters of Eve*, Gorgi Books, 2002

J. Kim, T. Warnow. Tutorial on Phylogenetic Tree Estimation, ISMB 1999

W. K. Sung. *Algorithms in Bioinformatics: A Practical Introduction*, chapters 7 & 8, 2009

D. Gusfield. *Algorithms on Strings, Trees, and Sequences,* chapter 17, 1997

http://www.geneticorigins.org/mito/media2.html