For written notes on this lecture, please read chapter 11 of *The Practical Bioinformatician*, 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

NUS
National University
of Singapore

Evolution

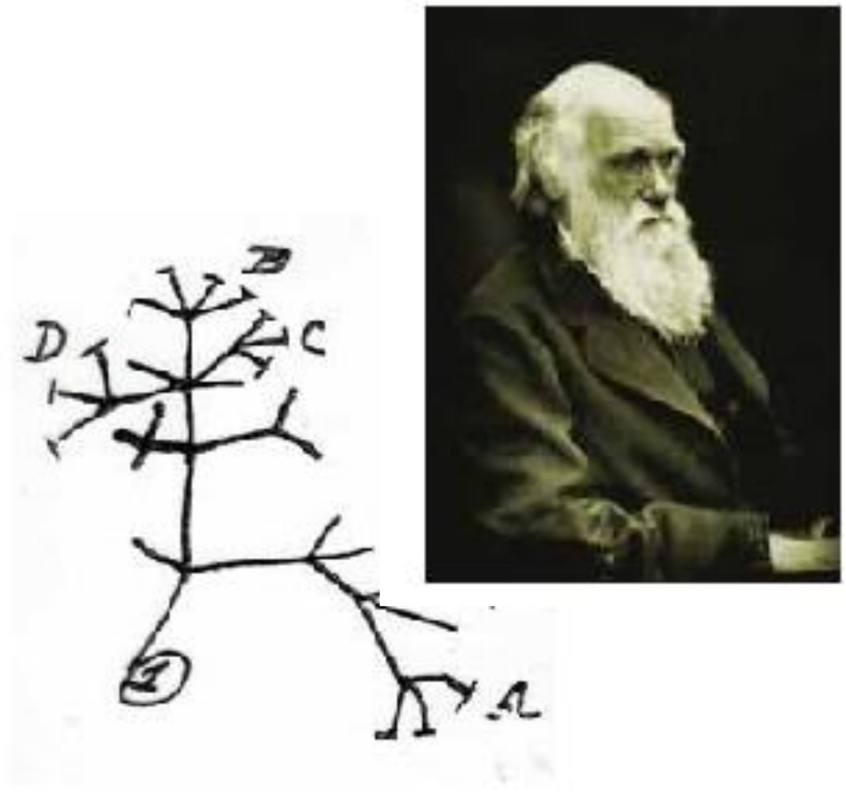
- DNA encodes blue print of life
- Living things pass DNA info to their children
- Due to mutations, DNA is changed a little bit
- After a long time, different species would evolve
- Phylogenetics studies genetic relationship between different species

Phylogeny

- **Phylogeny: Reconstruction of evolutionary history of a set of species**

- **Usually, it is a leaf-labeled tree where the internal nodes refer the hypothetical ancestors and the leaves are labeled by the species**

- **Edges of the tree represent the evolutionary relationships**

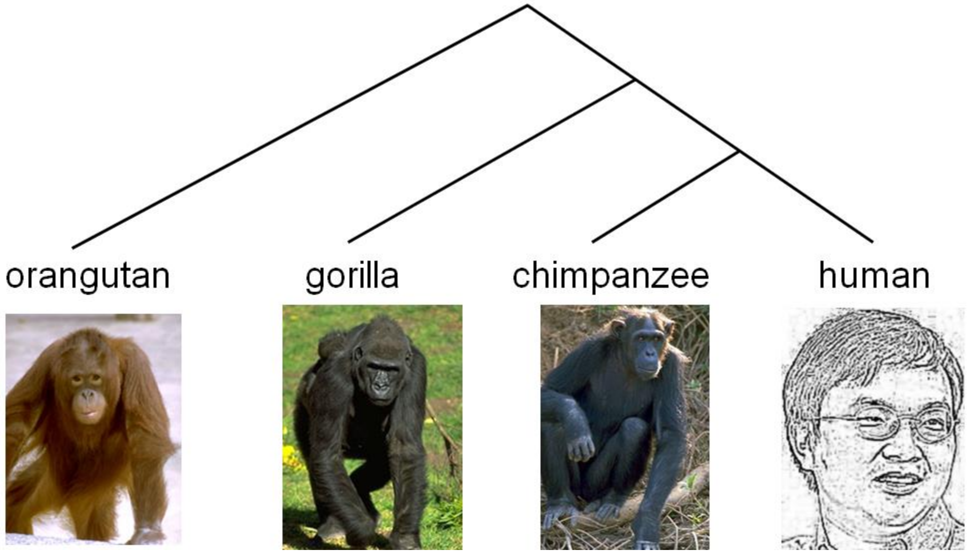


First Notebook on Transmutation of Species, 1837.

Phylogeny: Example

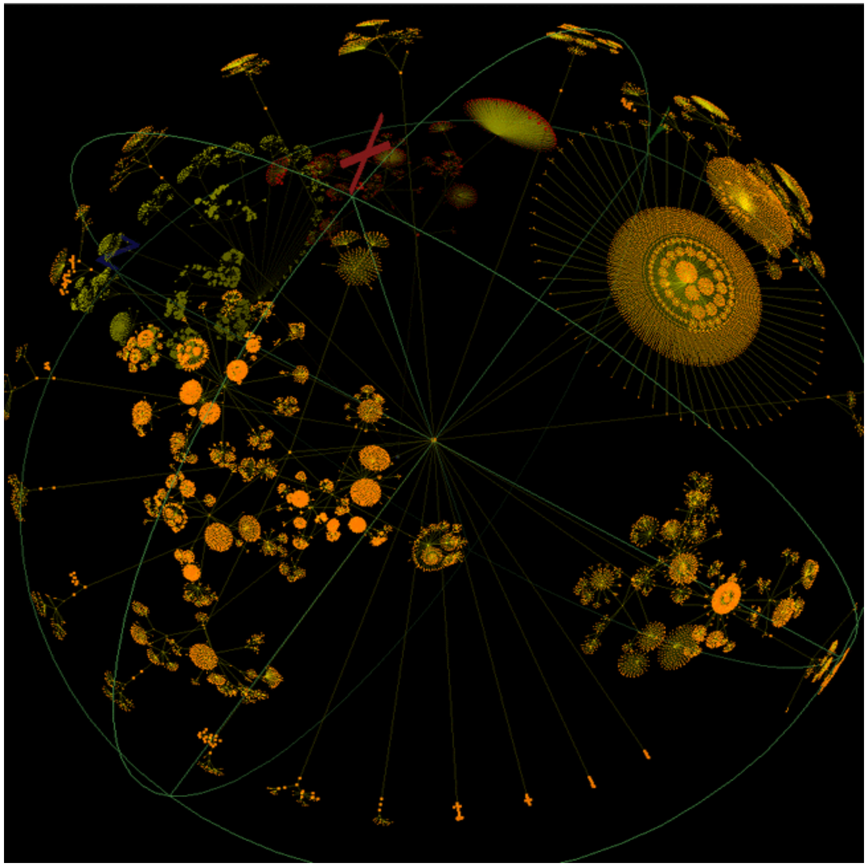
- By looking at extent of conserved positions in the "multiple seq alignment" of different groups of seqs, can infer when they last shared an ancestor

⇒ Construct "family tree" or phylogeny



Applications

- Understanding history of life
- Understanding rapidly mutating viruses (like HIV)
- Predict protein/RNA struct
- Do multiple seq alignment
- Explain and predict gene expression
- Explain and predict ligands
- Design enhanced organisms
- Design drug



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

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**

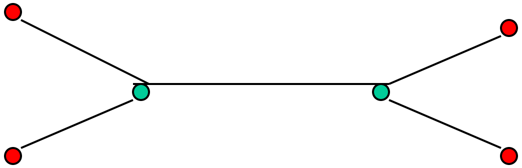
⇒ **Try to use molecules that carry a great deal of evolutionary history, like mitochondrial DNA, and ribosomal RNA**

Phylogeny Reconstruction

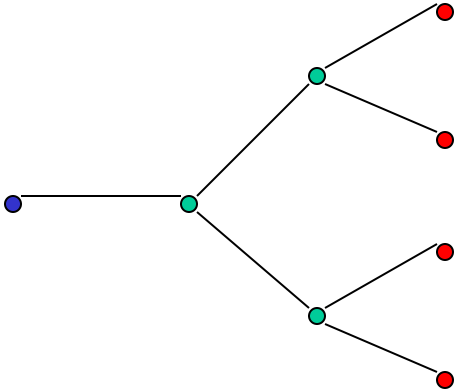
NUS
National University
of Singapore

Rooted and unrooted tree

- **Normally, the reconstructed tree is unrooted since estimating the root is difficult**



- **Rooted tree can be reconstructed by systematic biologists based on using outgroup**
  - Outgroup is a species which is clearly less related with all other species in the phylogeny



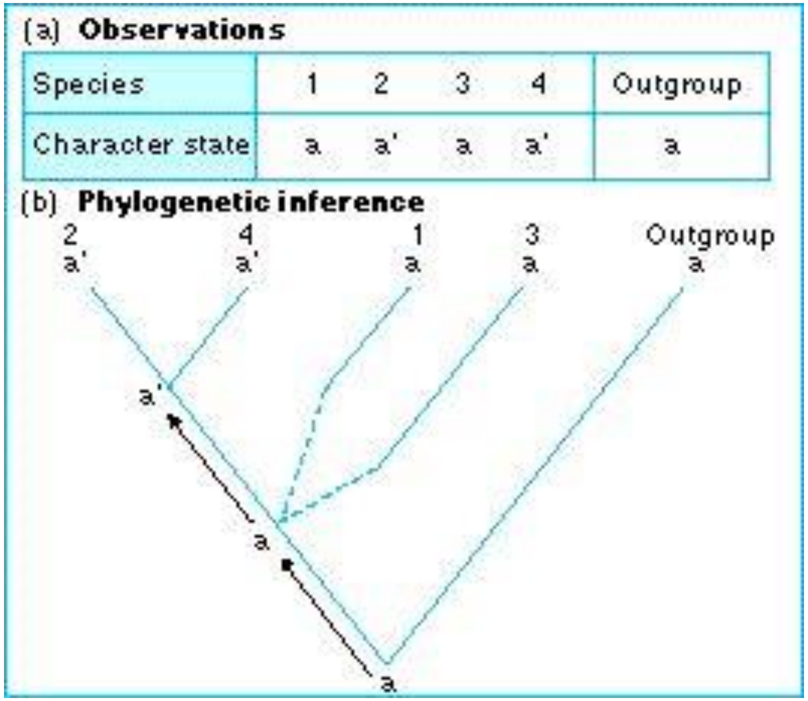
How does outgroup work?

- More similar to outgroup

⇒ More “ancient”

- More diff from outgroup

⇒ More “recent”, because more time to evolve



(a) **Observations**

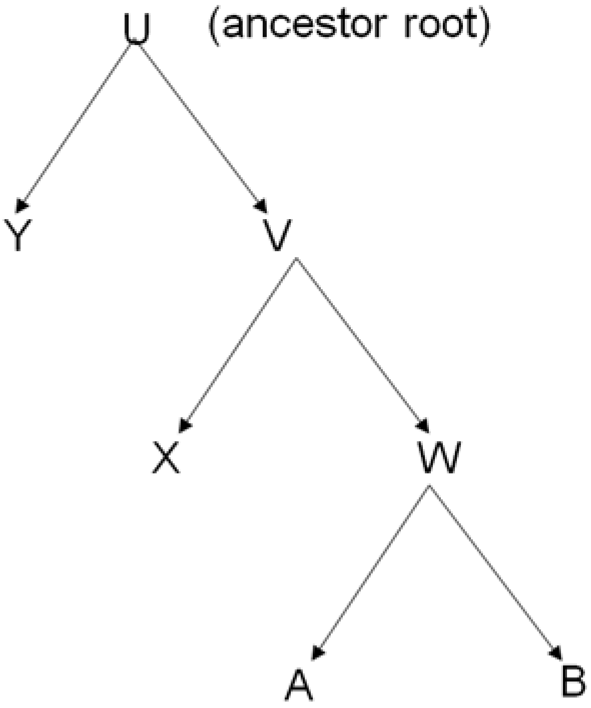
Species	1	2	3	4	Outgroup
Character state	a	a'	a	a'	a

(b) **Phylogenetic inference**

Image credit: Mark Ridley

Exercise

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



- **What is the most likely sequence for U?**
- **Hint: A phylogeny with fewer mutations is more likely than a phylogeny with more mutations**

U = ACCA**G**–TACTT**[C or –]**GATA**A**

If position 1 is "T", then both Y and W has a mutation in this position. If position 1 is "A", then only X has a mutation in this position. By the parsimony assumption, position 1 must be "A".

Choosing outgroup

- **Outgroup seq should be closely related to rest of seqs, but there should also be significantly more diff betw outgroup and rest of seqs**
- **Outgroup that is too distant may lead to incorrect tree because of more random & complex nature of diff betw outgroup and rest of seqs**
- **In choosing outgroup, one assumes that the evolutionary history of the gene is same as rest of seqs. If this assumption is incorrect (e.g., horizontal gene xfer has occurred), an incorrect analysis could result**

Methods for phylogeny reconstruction

- **Maximum parsimony**
- **Distance**
  - – Straightforward
  - – Applicable to large number of seqs

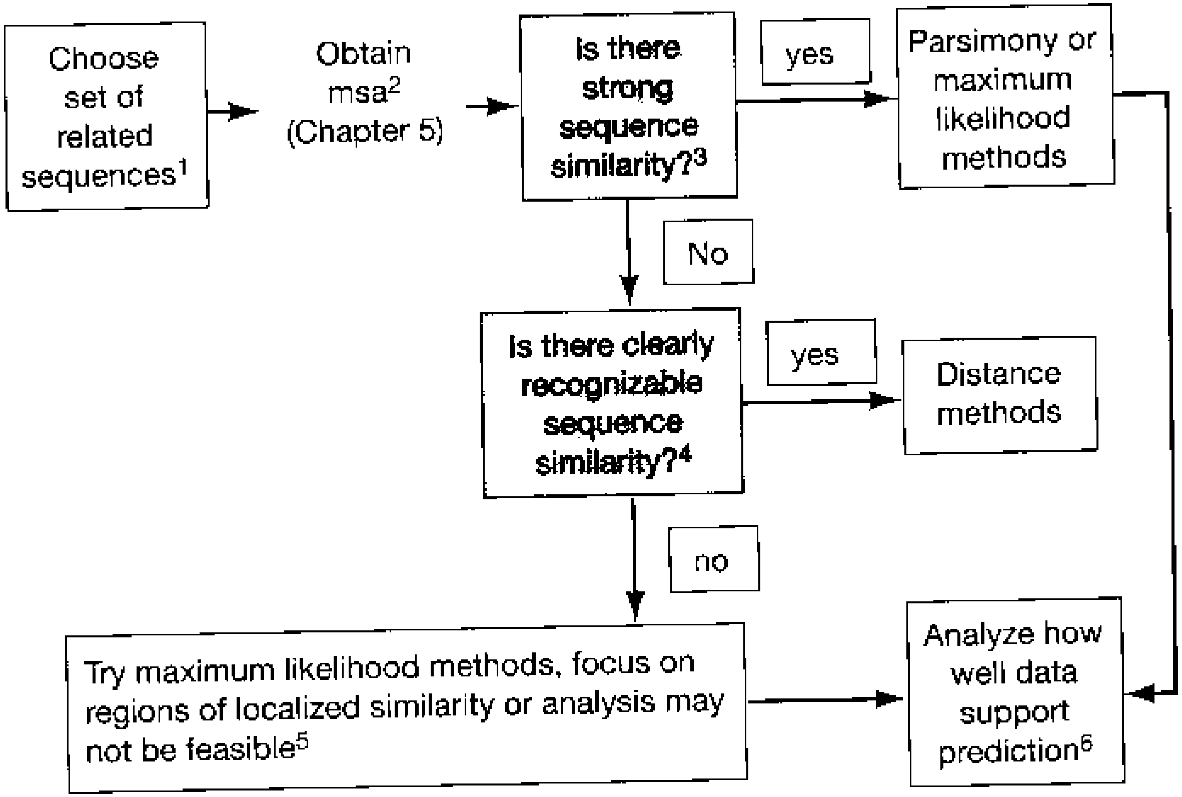
  ⇒Commonly used in mol biol labs

  ⇒ We consider only this one here!
- **Maximum likelihood**
  - – Require more understanding of evolutionary models on which they are based
  - – Involve exponential number of steps

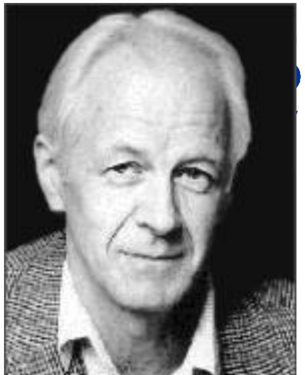
  ⇒Limited to small number of seqs

Exercise: What are the characteristics of max parsimony?

When to use which method?


Choose set of related sequences[1]
Obtain msa[2] (Chapter 5)
Is there strong sequence similarity?[3]
yes
Parsimony or maximum likelihood methods
No
Is there clearly recognizable sequence similarity?[4]
yes
Distance methods
no
Try maximum likelihood methods, focus on regions of localized similarity or analysis may not be feasible[5]
Analyze how well data support prediction[6]

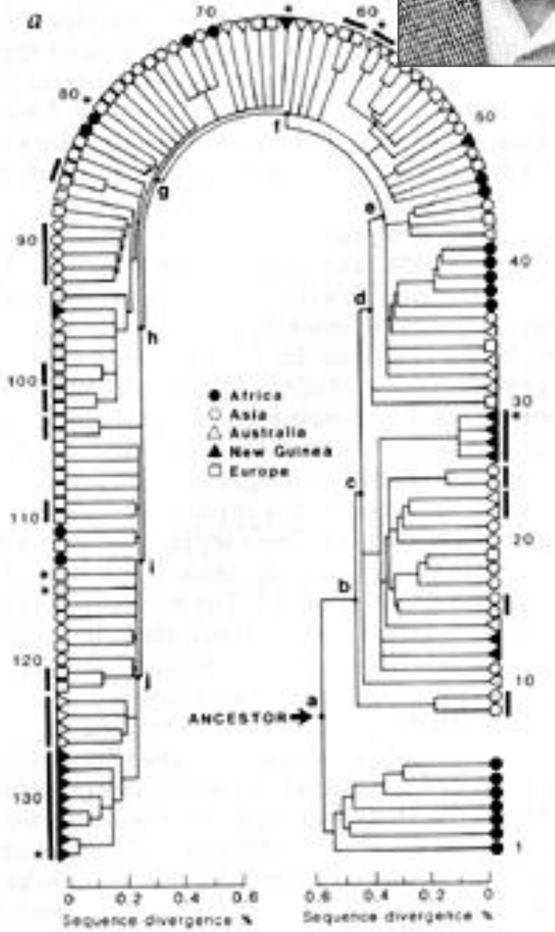
Source: D.W.Mount, Bioinformatics: Sequence and Genome Analysis, Cold Spring Harbor Press, 2004



Allan Wilson

- **“Molecular clock”: Dating by genetic mutations**
  - Deduced in 60s that proto-hominids evolved 5m yrs ago, contrary to the 25m yrs believed by anthropologists
  - In 80s, his findings became more widely accepted

- **Molecular approach to understand evolution**
  - Concluded in 80s that modern man evolved from “African Eve”
  - 20 yrs to convince palaeontologists, but when they did, it married their science with that of genetics



About mitochondrial Eve

- **Human mitochondrial DNA (mtDNA)**
  - Circular double-stranded consisting of ~16k bp
  - Everyone inherits the mtDNA from his/her mother
  - The pointwise mutation substitution rates of mtDNA is ~10 times faster than nuclear DNA
  - Every cell has lots of mtDNAs
  - No recombination

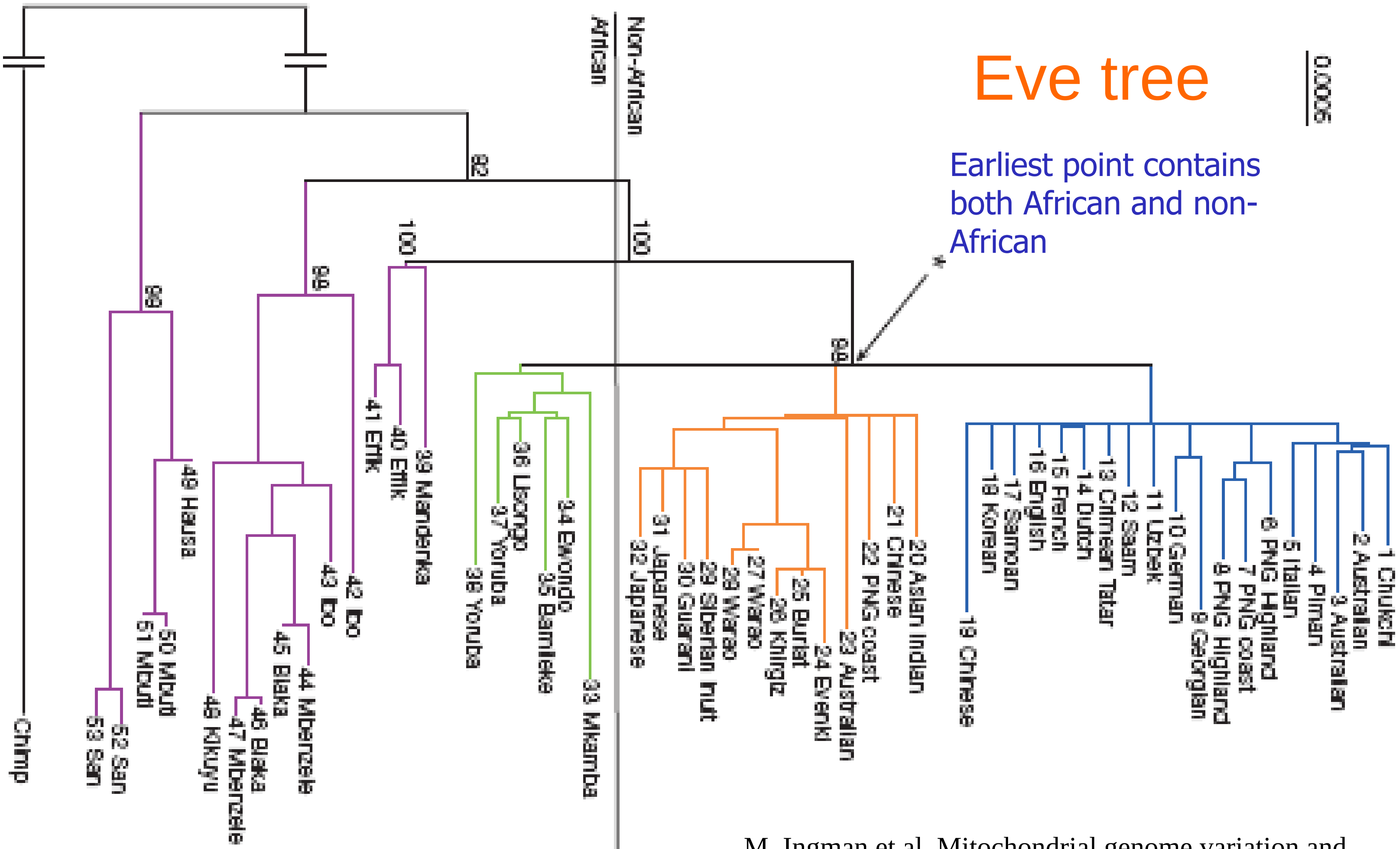
⇒ **We all inherit the mtDNA from the mother of human (Eve)!**

Genetics helps find 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.

Eve tree



Earliest point contains both African and non-African

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

Exercise: What is the outgroup?

Y-chromosome Adam

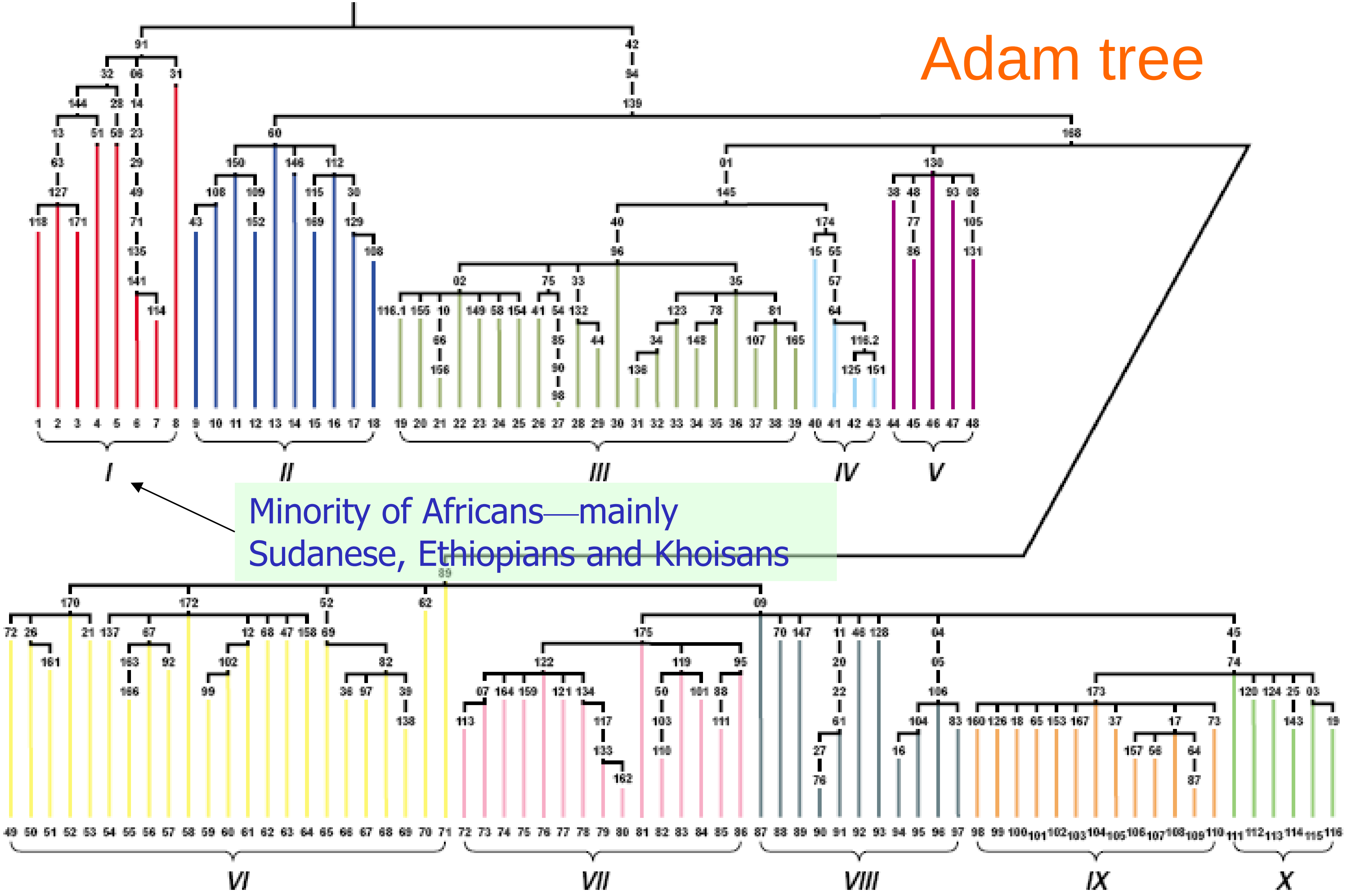
- **Y chromosome is unique to males and it can help to find the father of human**
  - – Mutation rate of Y chromosome not as fast as mtDNA

  ⇒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 ~59,000 years ago

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

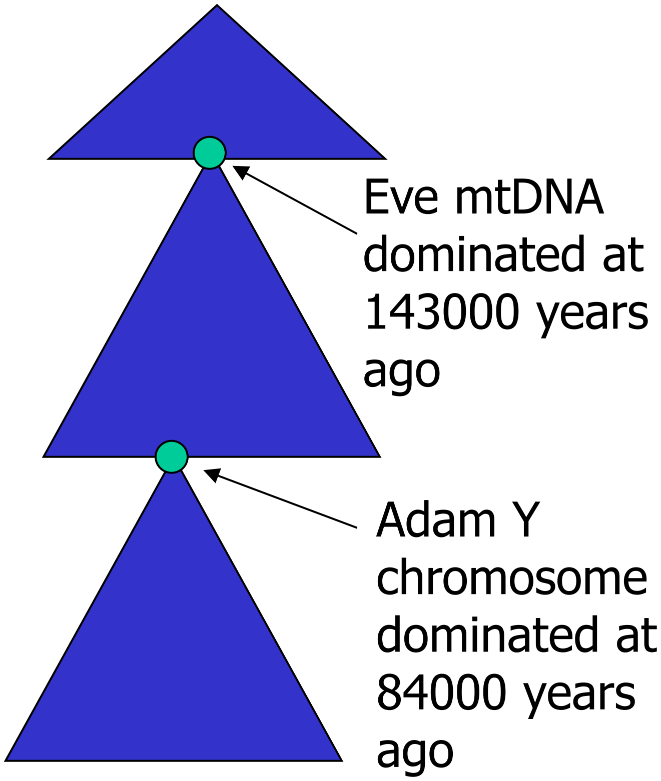
Adam tree



Minority of Africans—mainly Sudanese, Ethiopians and Khoisans

Why Adam &Eve appeared in different time

- **~143,000 years ago,**
  - Among diff human mtDNAs, Eve’s mtDNA had advantages and started to dominate
  - All other versions of mtDNAs eventually disappeared


Eve mtDNA dominated at 143000 years ago
Adam Y chromosome dominated at 84000 years ago

- **In parallel, diff versions of Y chromosomes appeared**
  - Took another ~60,000 years before Adam’s Y chromosome started to take over

Distance-Based Phylogeny-Reconstruction Methods

NUS
National University of Singapore

Distance between species

- **In character-based methods, we try to minimize # of mutations**

- **Species which look similar should be evolutionary more related**

⇒ **Define distance betw 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 Betw 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**
- **Looks reasonable, as 2 mismatches can be thought of as 2 mutations**

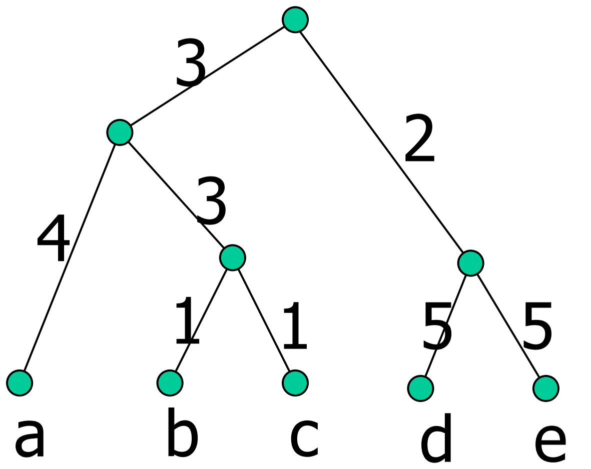
- **However, this fails to capture “multiple” mutations on the same site**
- **In practice, need to apply some corrective distance transformation**

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)
- **Output: Tree of degree 3 that is consistent with M**

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

→



Metric Space

- **A distance metric M which satisfies**
  - Symmetry

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

  - Self identity

$$M_{ii} =0$$

  - Triangular inequality

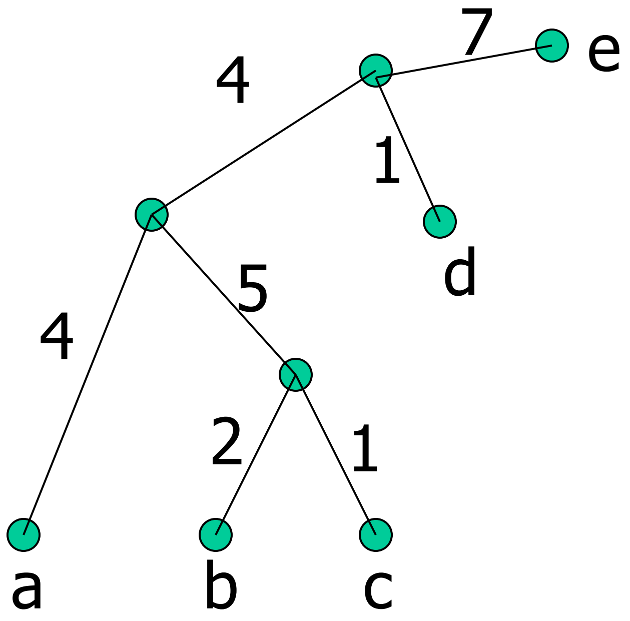
$$M_{ij} + M_{jk} \geq M_{ik}$$

Additive Metric

- **Let S be a set of species**
- **Let M be distance matrix for S**
- **If there is a rooted tree T where**
  - every edge has a positive weight and 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

	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



- **Don’t know the root! We can only build an unrooted phylogeny**

Why Additive Metric?

- **Distance captures actual number 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 number of mutations for each edge,**
  - The distance (the number of mutations) betw two species i and j should be the sum of the edge weights along the path from i to j

⇒ **Additive metric seems reasonable**

Is Hamming distance additive?

- **For any two species i and j, can we define M_{ij} to be Hamming distance betw species i and j?**
  - Example: Assume # of characters m=5
    - **Species i: (A, C, G, C, T)**
    - **Species j: (C, C, A, C, T)**
    - **Hamming distance $h_{ij} = 2$**
  - No! Hamming distance can’t capture “multiple” mutations on the same site. It is not additive metric
- **Poisson correction**
  - Corrected distance $M_{ij} = -\ln(1- h_{ij}/m)$
  - As # of characters increases, M converges to an additive metric

Properties of Additive Metric

- Buneman's 4-point condition

  M is additive if and only if

  for every four species in S,

  we can label them i, j, k, l such that

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

- Based on the 4-point condition, we can check whether a matrix M is additive or not

Proof

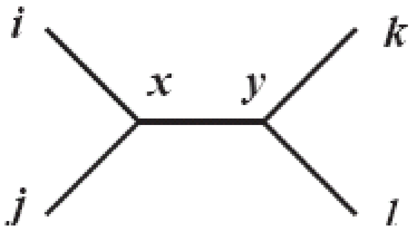


Figure 8.3: Buneman's 4-Point Condition

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

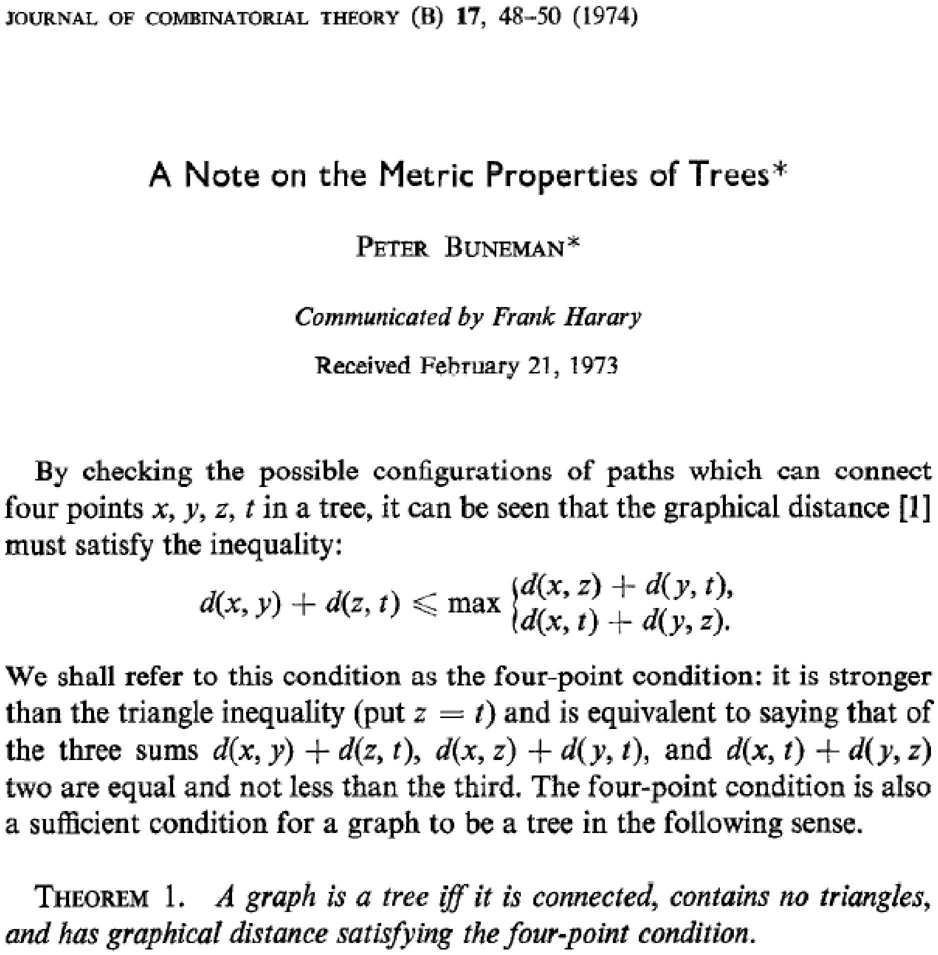
$$
\begin{aligned}
& M_{jk} + M_{il} \\
= & (M_{jx} + M_{xy} + M_{ik}) + (M_{ix} + M_{xy} + M_{yl}) \\
= & M_{ix} + M_{jx} + M_{yk} + M_{yl} + 2M_{xy}
\end{aligned}
$$

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

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

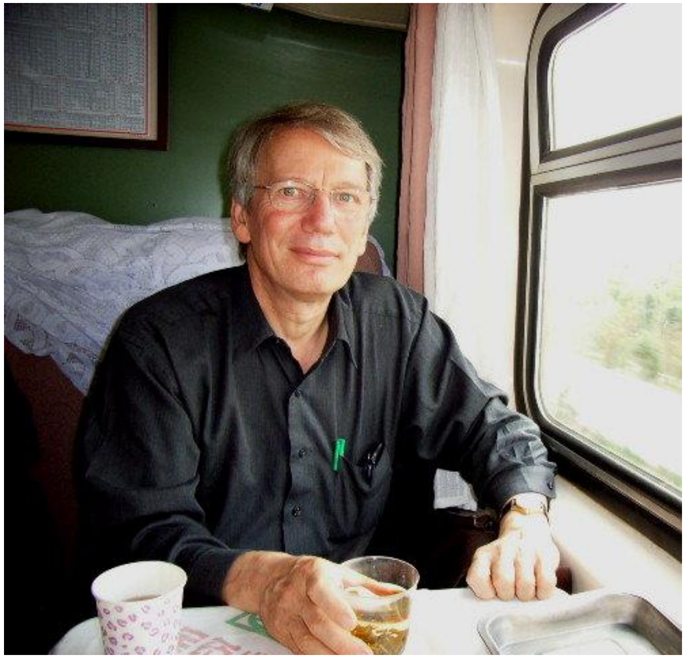
Received February 21, 1973

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:

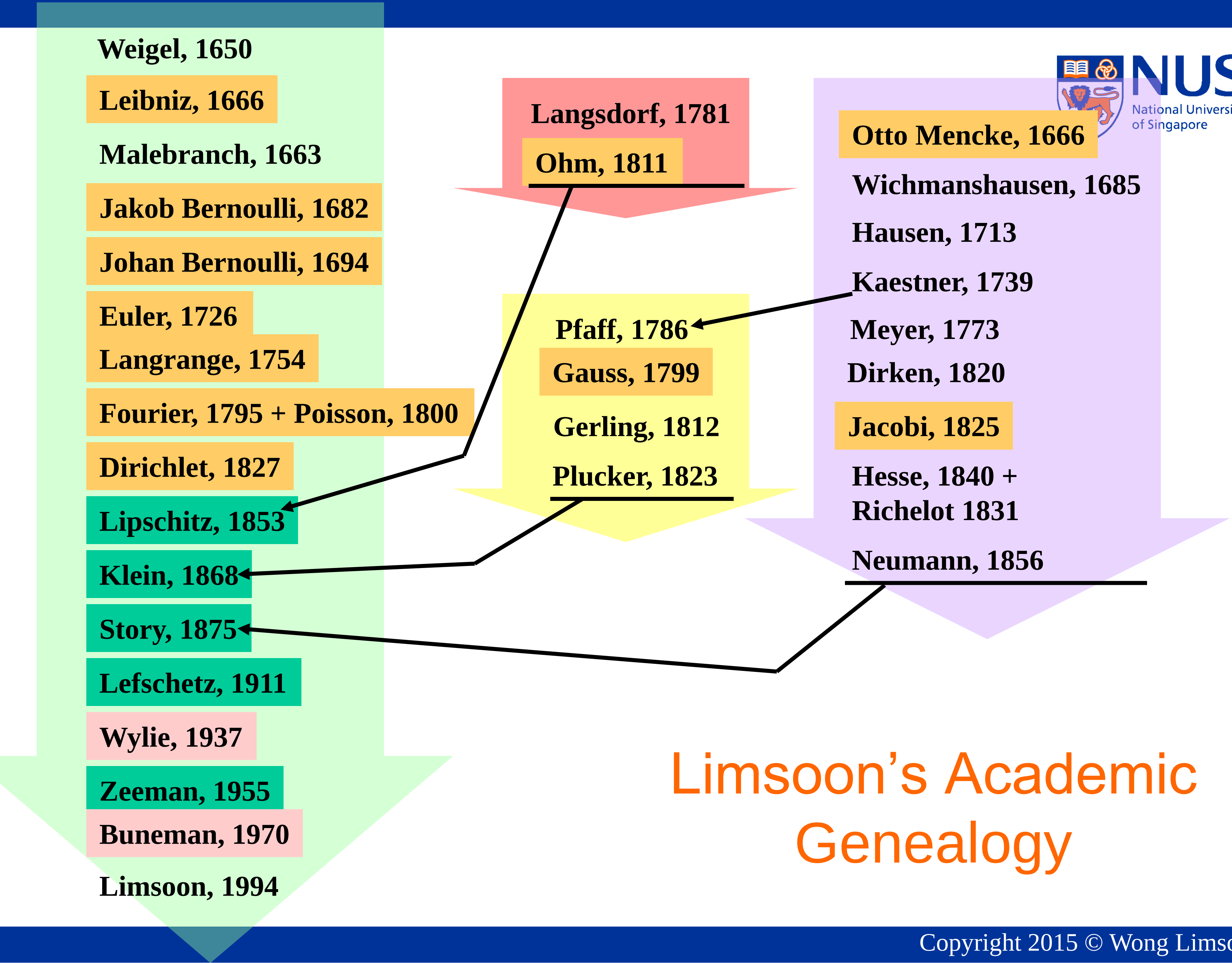
$$d(x, y) + d(z, t) \leqslant \max \left\{ \begin{array}{l} d(x, z) + d(y, t), \\ d(x, t) + d(y, z). \end{array} \right.$$

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



Weigel, 1650

Leibniz, 1666

Malebranch, 1663

Jakob Bernoulli, 1682

Johan Bernoulli, 1694

Euler, 1726

Langrange, 1754

Fourier, 1795 + Poisson, 1800

Dirichlet, 1827

Lipschitz, 1853

Klein, 1868

Story, 1875

Lefschetz, 1911

Wylie, 1937

Zeeman, 1955

Buneman, 1970

Limsoon, 1994

Langsdorf, 1781

Ohm, 1811

Pfaff, 1786

Gauss, 1799

Gerling, 1812

Plucker, 1823

Otto Mencke, 1666

Wichmanshausen, 1685

Hausen, 1713

Kaestner, 1739

Meyer, 1773

Dirken, 1820

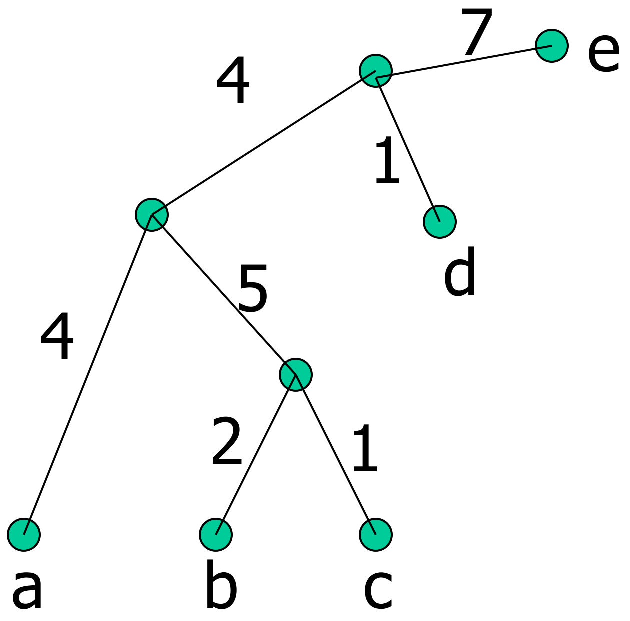
Jacobi, 1825

Hesse, 1840 +
Richelot 1831

Neumann, 1856

Let’s Check!

	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?

Ultrametric

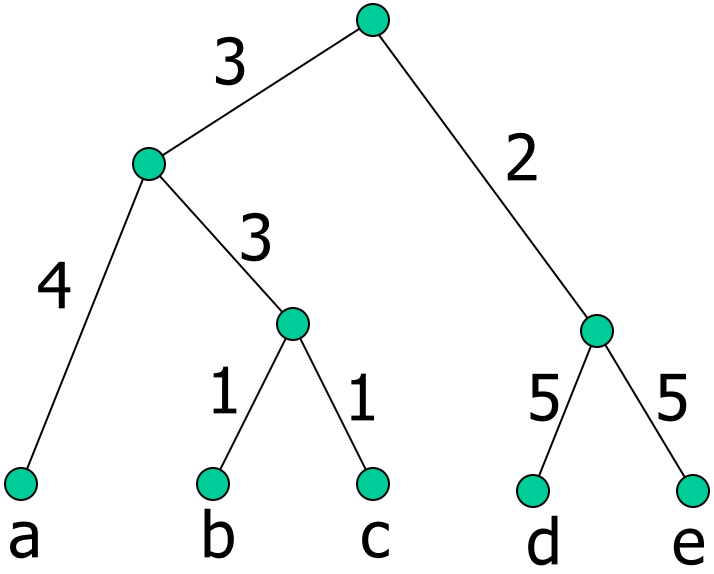
- **Assume M is additive. That is, there exists a tree T such that**
  - the distance between any two species i and j equals the sum of the edge weights along the path from i to j

- **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!**

Properties of Ultrametric

- **Ultrametric is an additive metric**

⇒ It satisfies 4-point condition

- **Additional property: Buneman's 3-point condition**

  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 or not**

Proof

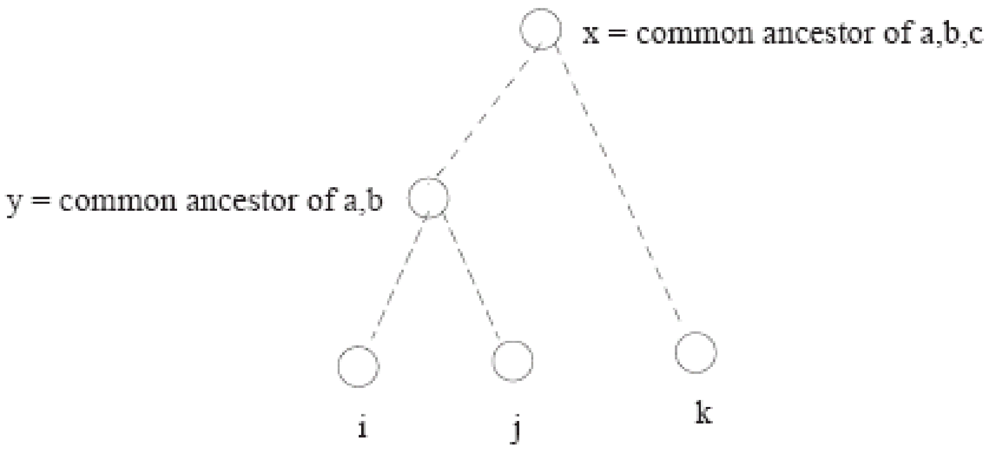


Figure 8.4: Ultrametric Tree

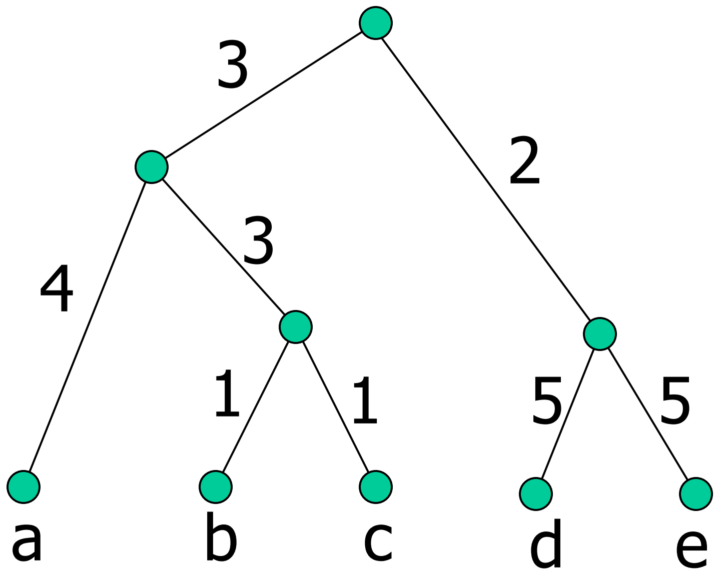
From the above formulas,and by Property 3 of an Ultrametric tree. There is

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

proven!
(⇐)Exercise. ■

Let’s Check!

	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 ($M_{ik}= M_{jk} \geq M_{ij}$) satisified?

Constant molecular clock

- **Constant molecular clock is an assumption in biology**
  - It states that # 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. Thus, 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:**

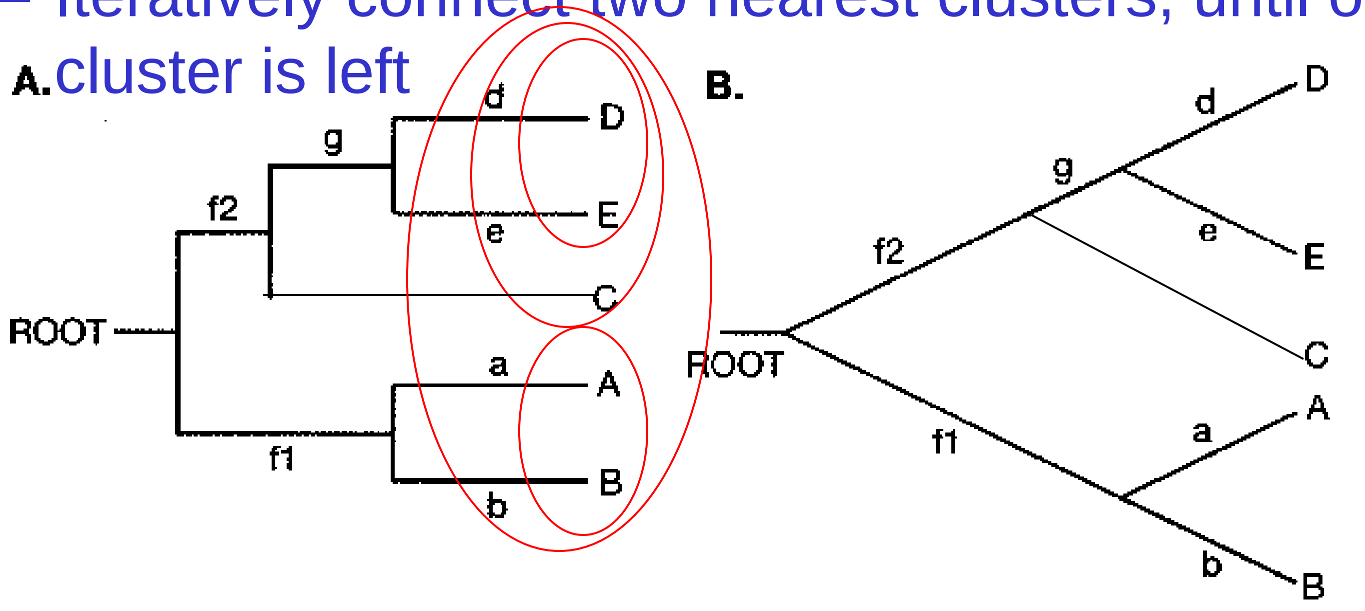
  An ultrametric matrix M for a set of species S

- **Problem:**

  Reconstruct the phylogenetic tree T for S

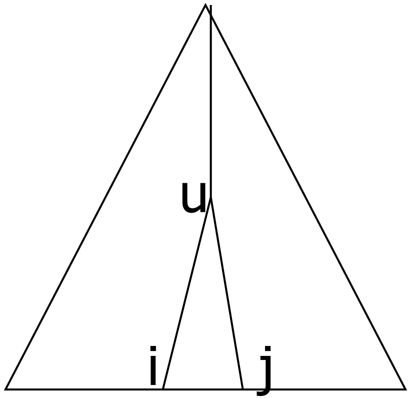
Unweighted Pair Group Method With Arithmetic Mean (UPGMA)

- **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



Definition - 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) = $M_{ij}/2$

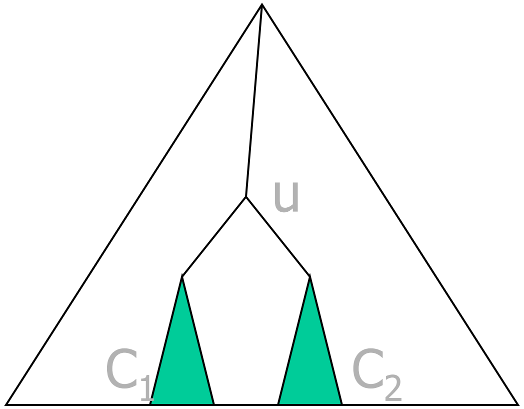


Distance betw two clusters

- **For any two clusters C_1 and C_2 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 clusters C_1,C_2, D,**

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

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

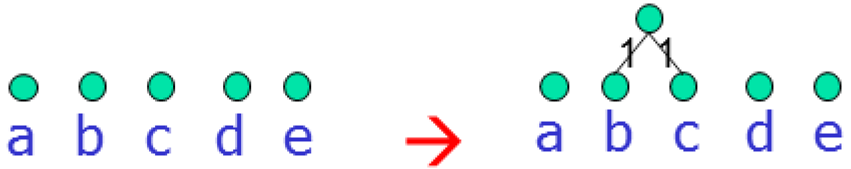
$$\begin{aligned} dist(C_1 \cup C_2, D) &= \frac{\sum_{i \in C_1 \cup C_2, j \in D} M_{ij}}{|C_1 \cup C_2| \cdot |D|} \\ &= \frac{\sum_{i \in C_1, j \in D} M_{ij} + \sum_{i \in C_2, j \in D} M_{ij}}{|C_1 \cup C_2| \cdot |D|} \\ &= \frac{|C_1||D| \, dist(C_1, D) + |C_2||D| \, dist(C_2, D)}{|C_1 \cup C_2|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

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

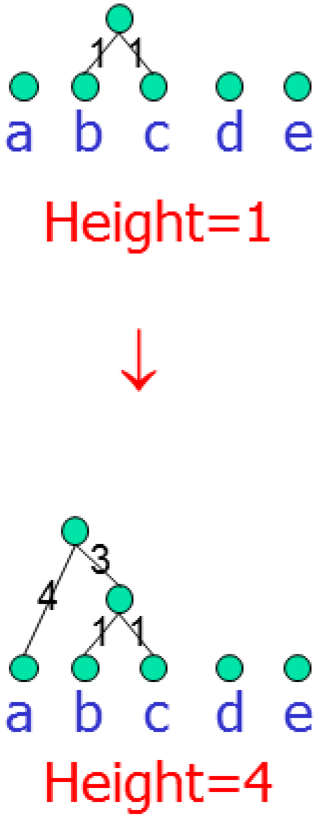


Height=1

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

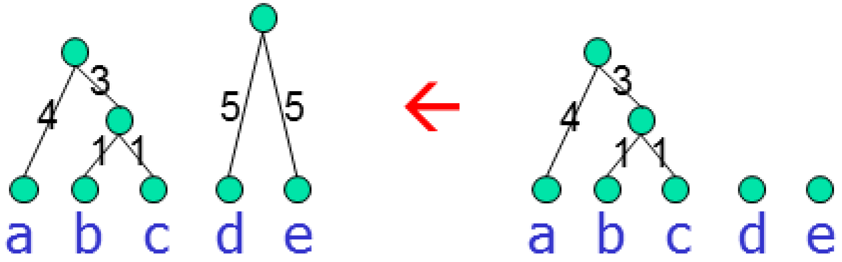


M	a,b,c	d	e
a,b,c	0	14	14
d	14	0	10
e	14	10	0

Example

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

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

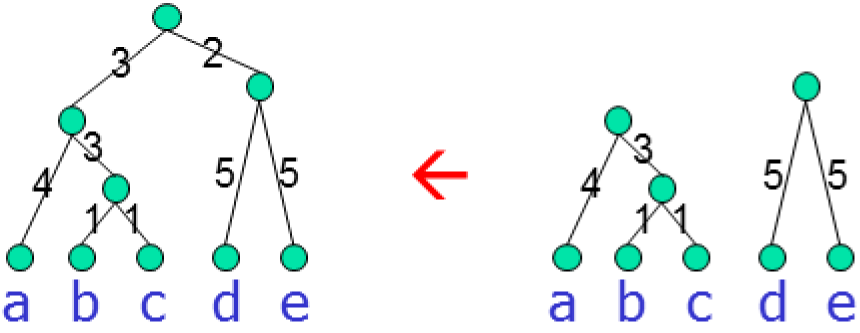


Height=5

Height=4

Example

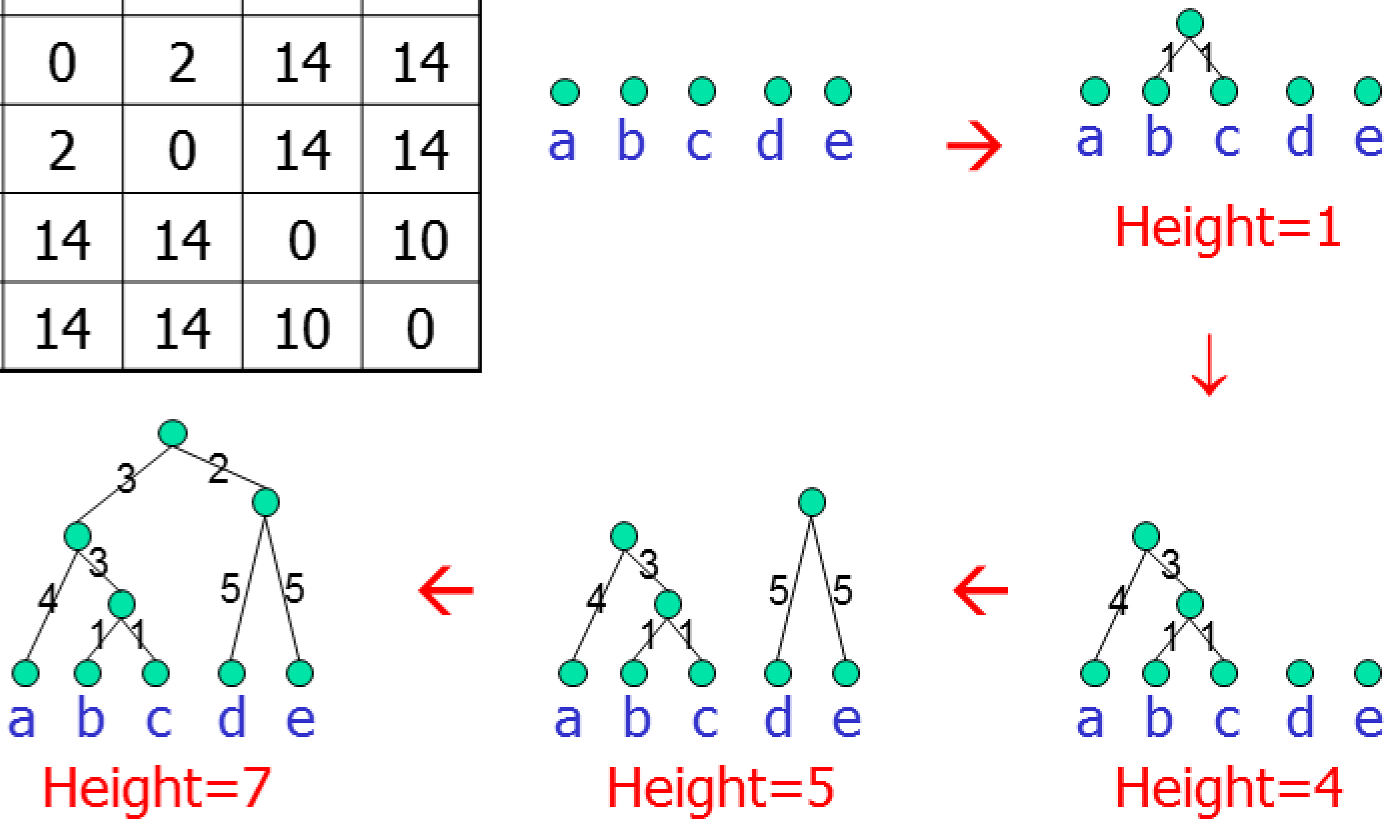
M	a,b,c	d,e
a,b,c	0	14
d,e	14	0



Height=7 Height=5

Example

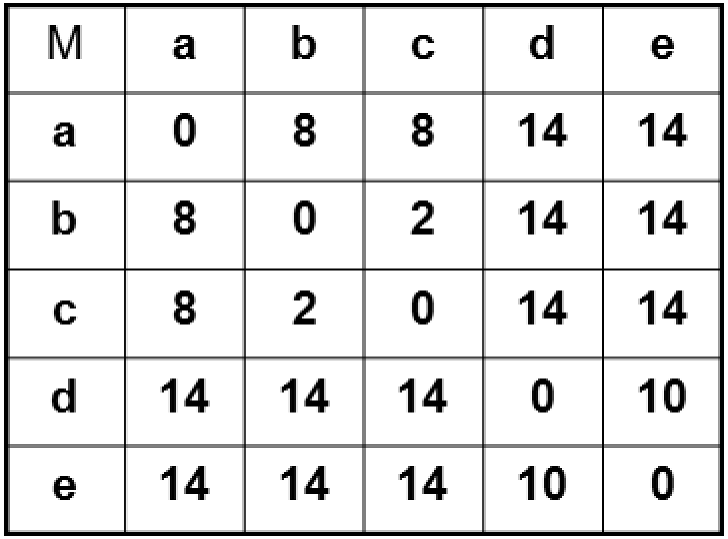
M	a	b	c	d	e
a	0	8	8	14	14
b	8	0	2	14	14
c	8	2	0	14	14
d	14	14	14	0	10
e	14	14	14	10	0



The algo runs in $O(n^3)$ time

- Given n x n ultrametric distance matrix M
- $O(n)$ Initialize set Z to consist of n singleton clusters {1}, {2}, …, {n}
- $O(n^2)$ For all {i}, {j} $\in$ Z, initialize dist({i}, {j}) = M_{ij}
- $O(n^3)$ Repeat n-1 times
  - $O(n^2)$ Determine cluster A, B $\in$ Z where dist(A, B) is min
  - $O(1)$ Define a new cluster C = A $\cup$ B
  - $O(1)$ Z := Z – {A, B} $\cup$ {C}
  - $O(1)$ Define new node c and let c be parent of A and B. Also, define height(c) = dist(A, B)/2
  - $O(n)$ For all D $\in$ Z – {C}, define dist(D, C) = dist(C, D) = (|A| dist(A, D) + |B| dist(B, D)) / (|A| + |B|)

Achieving quadratic complexity



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

- **Use a vector $\min_P[i]$ to record the column id j such that M[i,j] is min of row i**
- **When searching for clusters to merge, look for $x = \text{argmin}_i \min_P[i]$**
- **Then merge cluster x with cluster $\min_P[x]$**

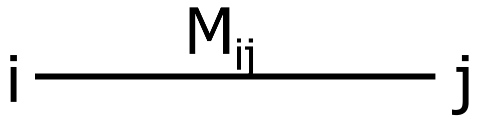
Now algo runs in $O(n^2)$ time

```
              • Given n x n ultrametric distance matrix M
$O(n)$        • Initialize set Z to consist of n singleton clusters {1}, {2}, …, {n}
$O(n^2)$      • For all {i}, {j} $\in$ Z, initialize dist({i}, {j}) = $M_{ij}$
$O(n^2)$      • For each P $\in$ Z, set $min_P = argmin_{Q \in Z-\{P\}}$ dist(P,Q)
$O(n^2)$ {    • Repeat n-1 times
         $O(n)$ – Determine cluster A $\in$ **Z** such that dist(A, $min_A$) is minimized
         $O(1)$ – Let B = $min_A$; Define a new cluster C = A $\cup$ B
         $O(1)$ – Z := Z – {A, B} $\cup$ {C}
         $O(1)$ – Define new node c and let c be parent of A and B. Also,
                  define height(c) = dist(A, B)/2
         $O(n)$ – For all D $\in$ Z – {C}, define dist(D, C) = dist(C, D) =
                  (|A| dist(A, D) + |B| dist(B, D)) / (|A| + |B|)
         $O(n)$ – Let $min_C = argmin_{Q \in Z}$ dist(C,Q)
         $O(n)$ – For each cluster D $\in$ **Z**–{C}, if dist(D, C)$\leq$dist(D, $min_D$), set
                  $min_D$ = C
```

Additive-tree reconstruction

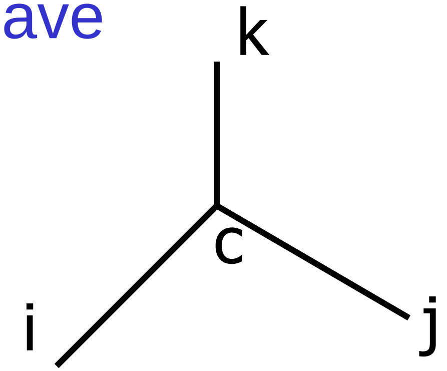
- Suppose M is an additive metric. We show an algorithm which reconstructs the additive tree in O(n^2) time

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



Additive tree for 3 species

- **For any three species i, j, k, we can find their center c as follows. (* Call this 3-star method *)**
  - 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} = dic + dcj**
  - 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!**

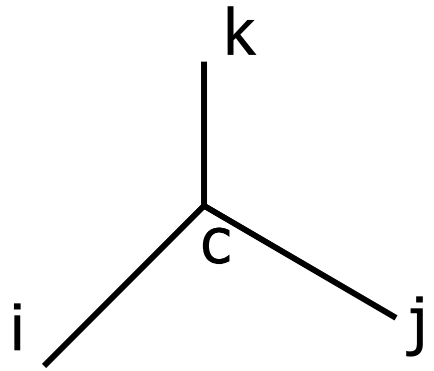
Additive tree for 4 species (I)

- **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**

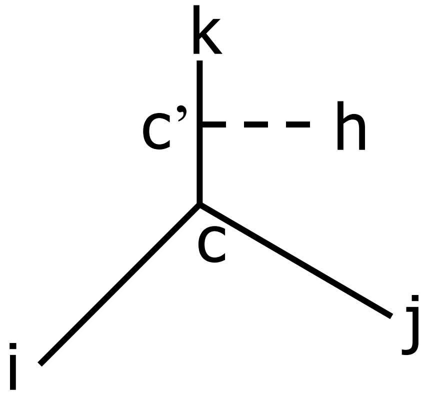
- **To include h into the tree, we need to introduce one more internal node c'**

- **c' splits either (i, c), (j, c) or (k, c)**



Additive tree for 4 species (II)

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

- **Note:**
  - The time required is O(k-1)
  - Also, the tree is unique!

Time complexity

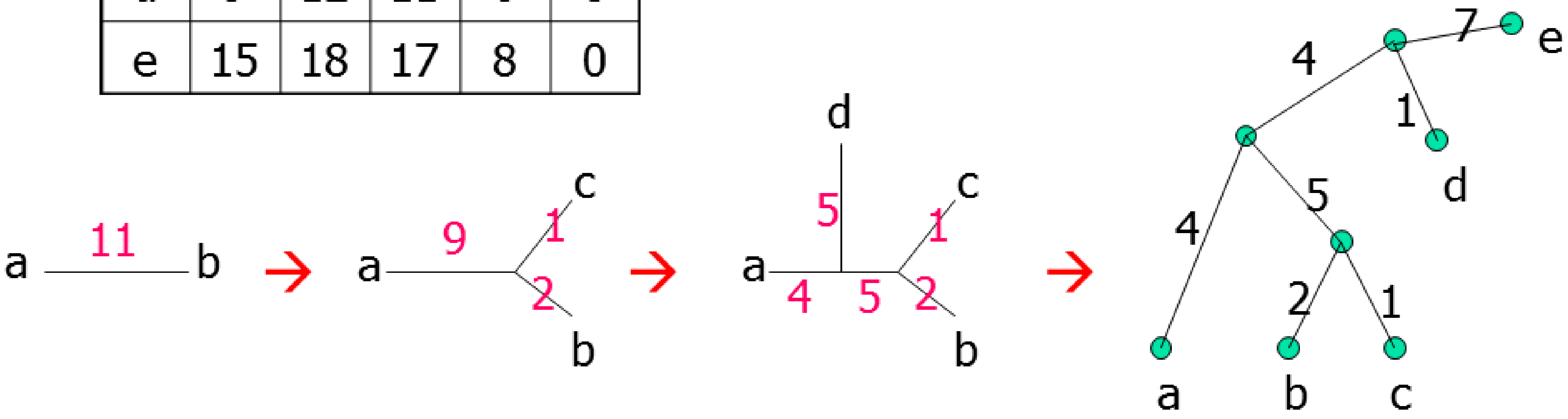
- In summary, to recover an additive tree with n species, the time is

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

- Note: 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



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

- Definition: For any tree T for the set of species S, let D be its corresponding distance matrix. We 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 which minimizes SSQ(T). Such tree is known as Least Squares Tree

- This problem is NP-hard

Can tree reconstruction methods infer the correct tree? (I)

- **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? (II)

- **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**

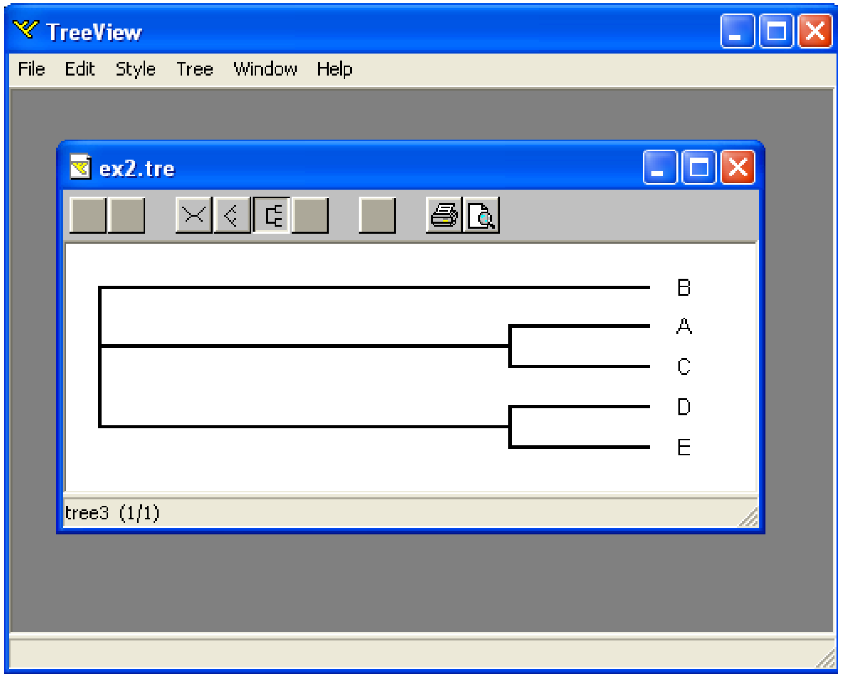
T. Zhu et al. An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. *Nature*, 391: 594-597, 1998

S/w for constructing phylogenetic tree

- **Felsenstein's PHYLIP**
  - Large # of methods, including ML, MP and NJ
  - 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

S/w for visualizing phylogenetic tree

- **Treeview**
  - A simple program for displaying phylogenies.
  - http://taxonomy.zoology.gla.ac.uk/rod/treeview.html



Phylogenetic Tree Comparison

NUS
National University
of Singapore

Why tree comparison?

- **We learn a number of 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

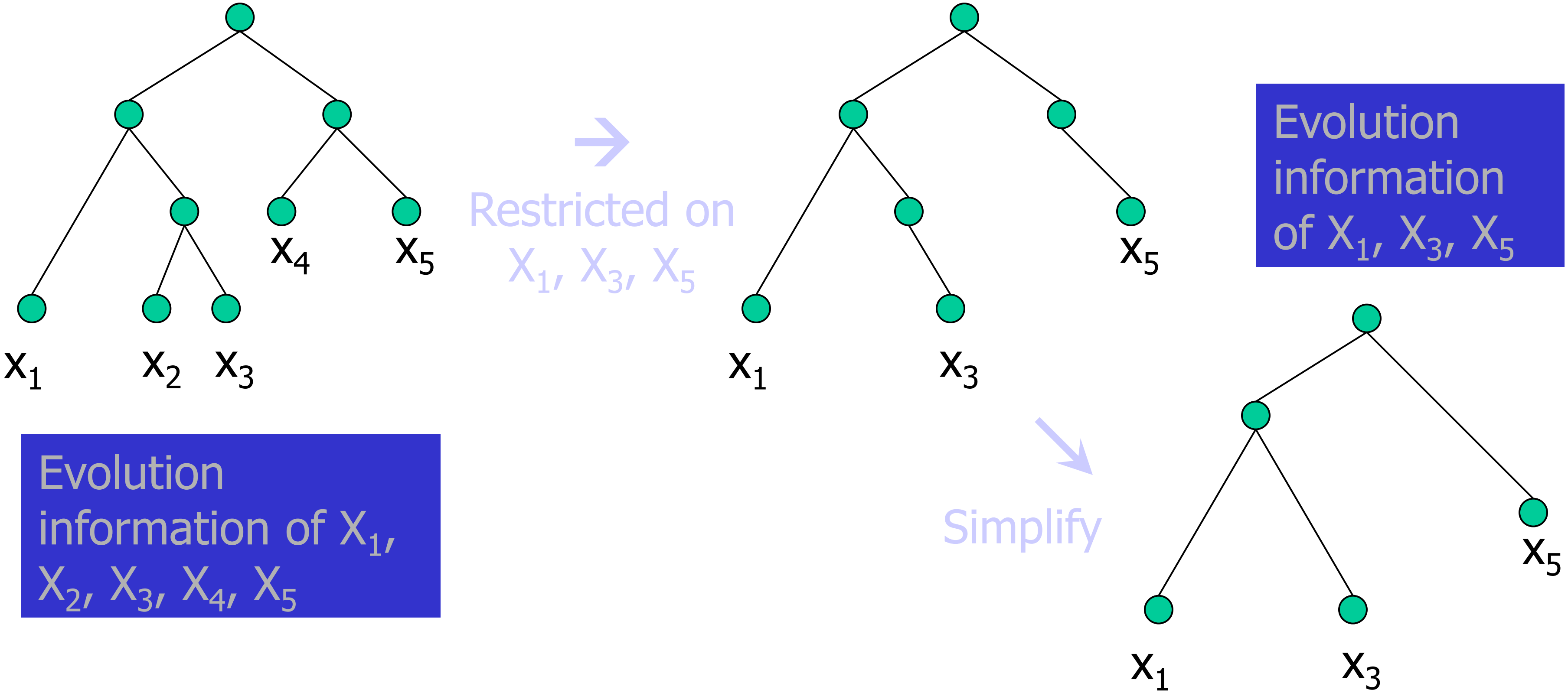
- **Similarity measurement**
  - Find common structure among given trees
    - **Maximum Agreement Subtree**

- **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**



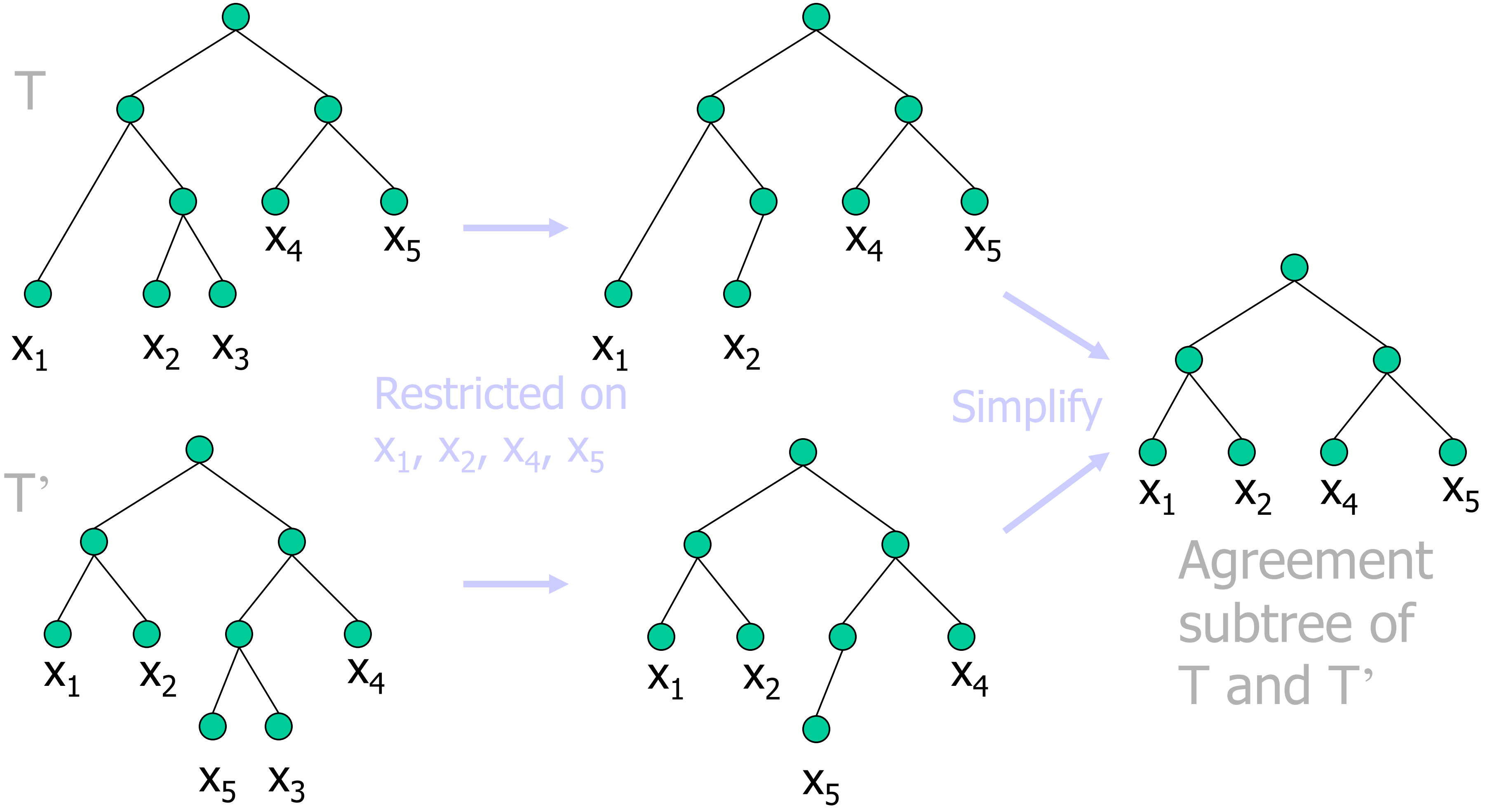
Evolution information of X_1, X_2, X_3, X_4, X_5

Restricted on X_1, X_3, X_5

Evolution information of X_1, X_3, X_5

Simplify

Agreement subtree



Restricted on x_1, x_2, x_4, x_5

Simplify

Agreement subtree of T and T'

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**
  - Since it is agreed by both trees, the 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

Notations

- 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,

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

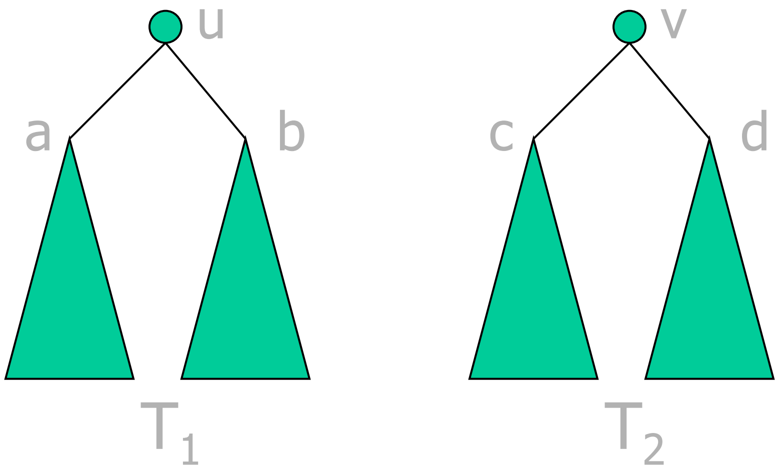
- For any node u in T_1 and v in T_2,

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

Recurrence (I)

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

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

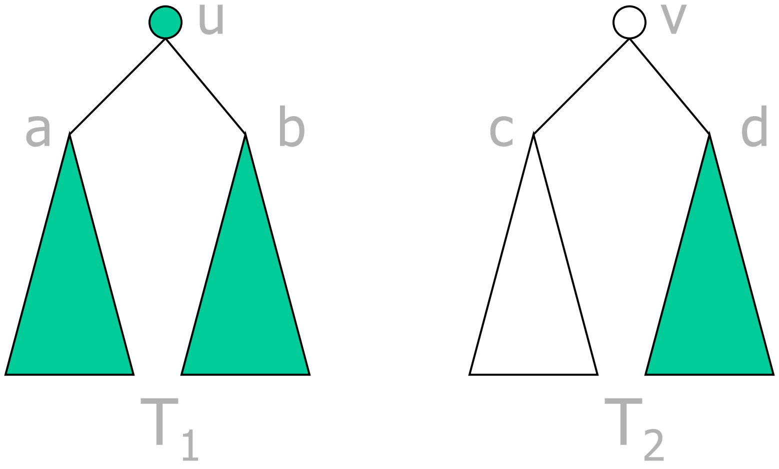


Recurrence (II)

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

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

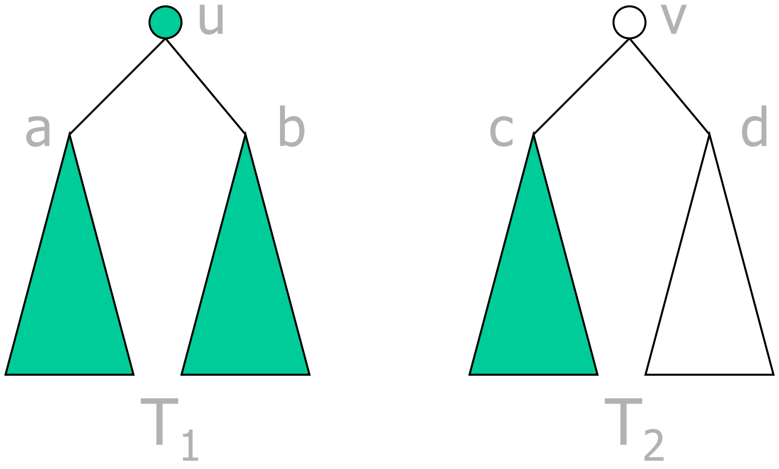
All the species in "agreement" are in right subtree of v



Recurrence (III)

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

All the species in “agreement” are in left subtree of v

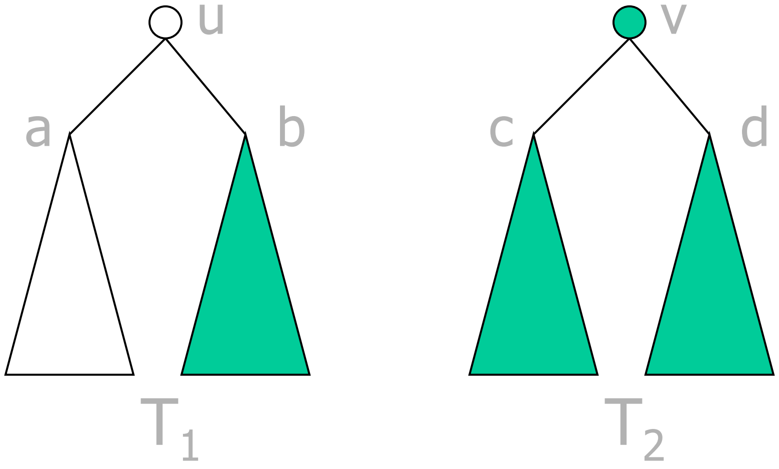


Recurrence (IV)

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

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

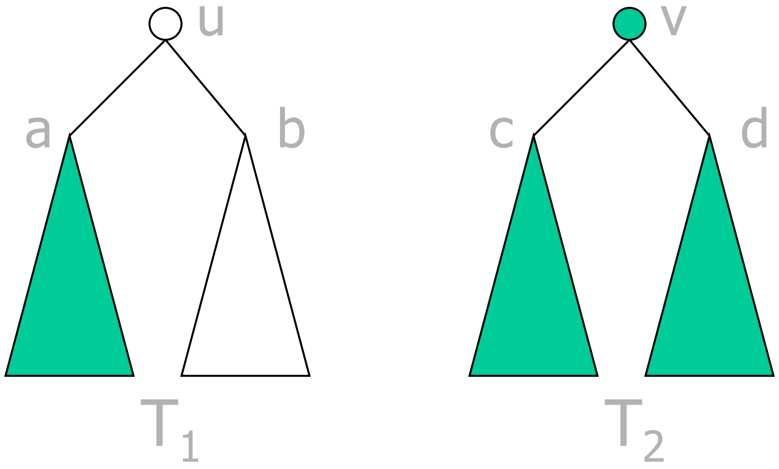
All the species in "agreement" are in right subtree of u



Recurrence (V)

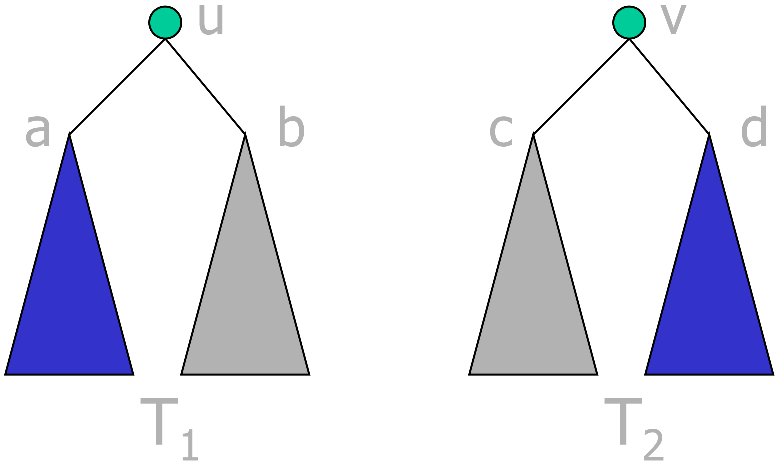
$$MAST(T_1^u, T_2^v) = \max \begin{cases} MAST(T_1^a, T_2^c) + MAST(T_1^b, T_2^d) \\ MAST(T_1^a, T_2^d) + MAST(T_1^b, T_2^c) \\ MAST(T_1^a, T_2^v) \quad \leftarrow \\ MAST(T_1^b, T_2^v) \\ MAST(T_1^u, T_2^c) \\ MAST(T_1^u, T_2^d) \end{cases}$$

All the species in "agreement" are in left subtree of u



Recurrence (VI)

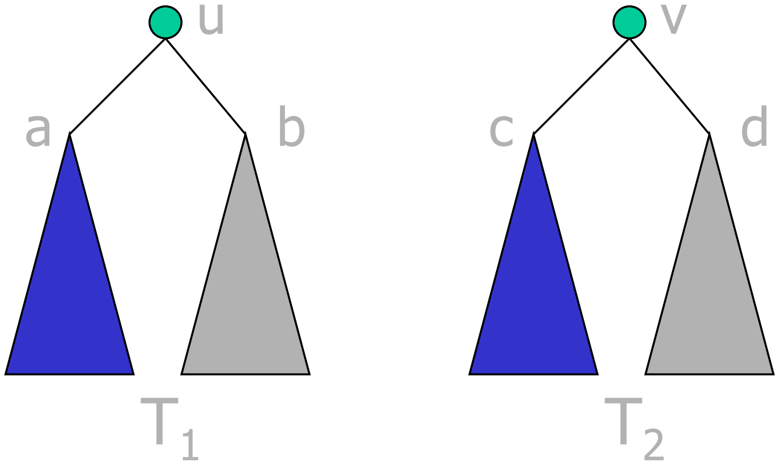
$$MAST(T_1^u, T_2^v) = \max \begin{cases} MAST(T_1^a, T_2^c) + MAST(T_1^b, T_2^d) \\ MAST(T_1^a, T_2^d) + MAST(T_1^b, T_2^c) \quad \leftarrow \\ MAST(T_1^a, T_2^v) \\ MAST(T_1^b, T_2^v) \\ MAST(T_1^u, T_2^c) \\ MAST(T_1^u, T_2^d) \end{cases}$$



Exercise: What does this case correspond to?

Recurrence (VII)

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



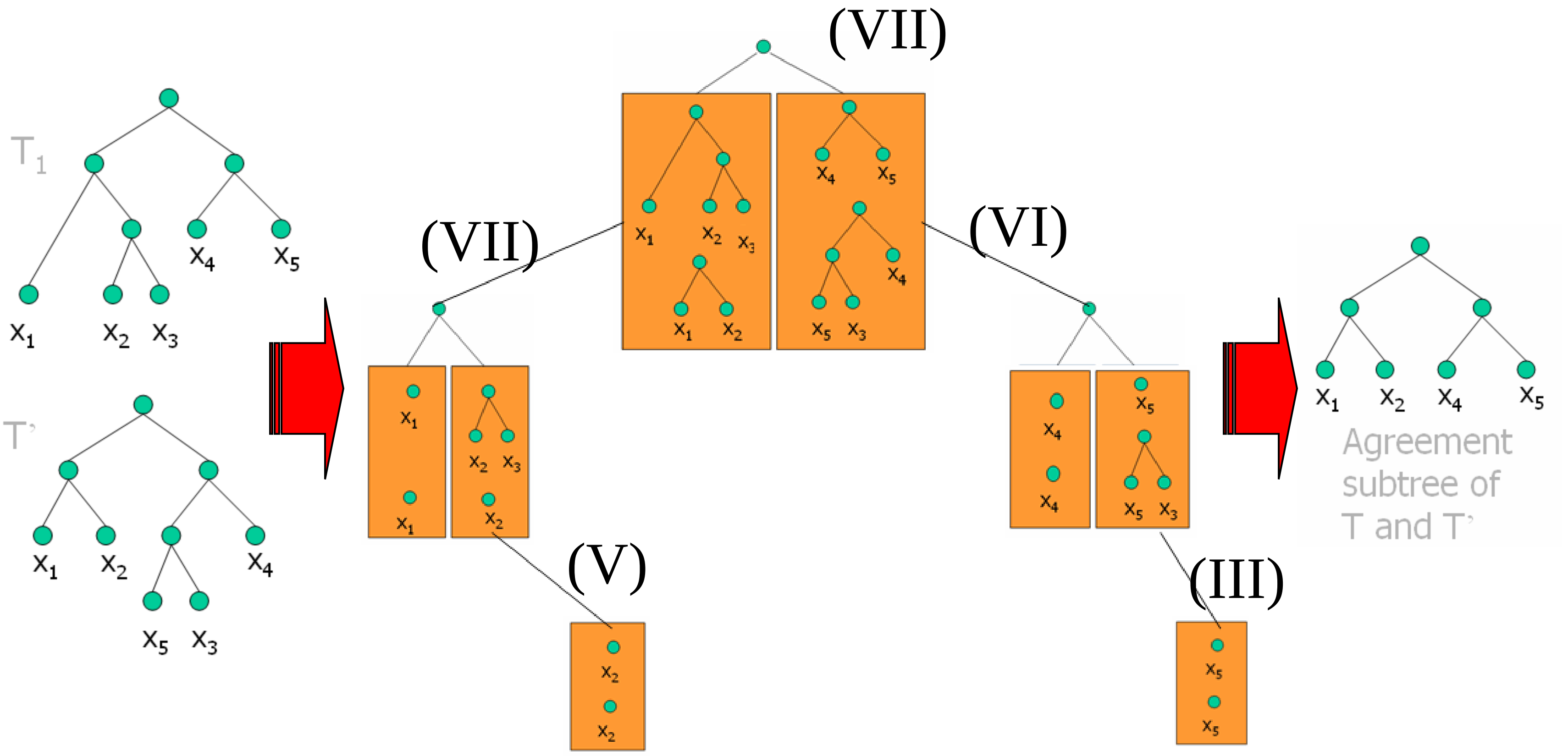
Exercise: What does this case correspond to?

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

provided you have a dynamic programming version of MAST

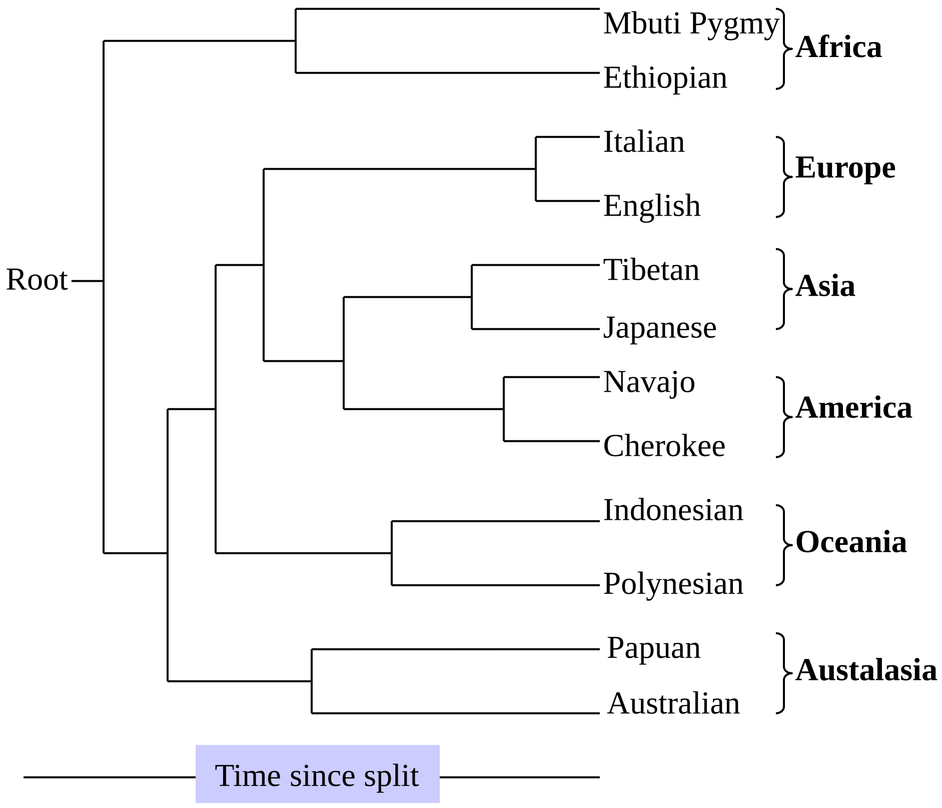
MAST Example



The 7 Daughters of Eve

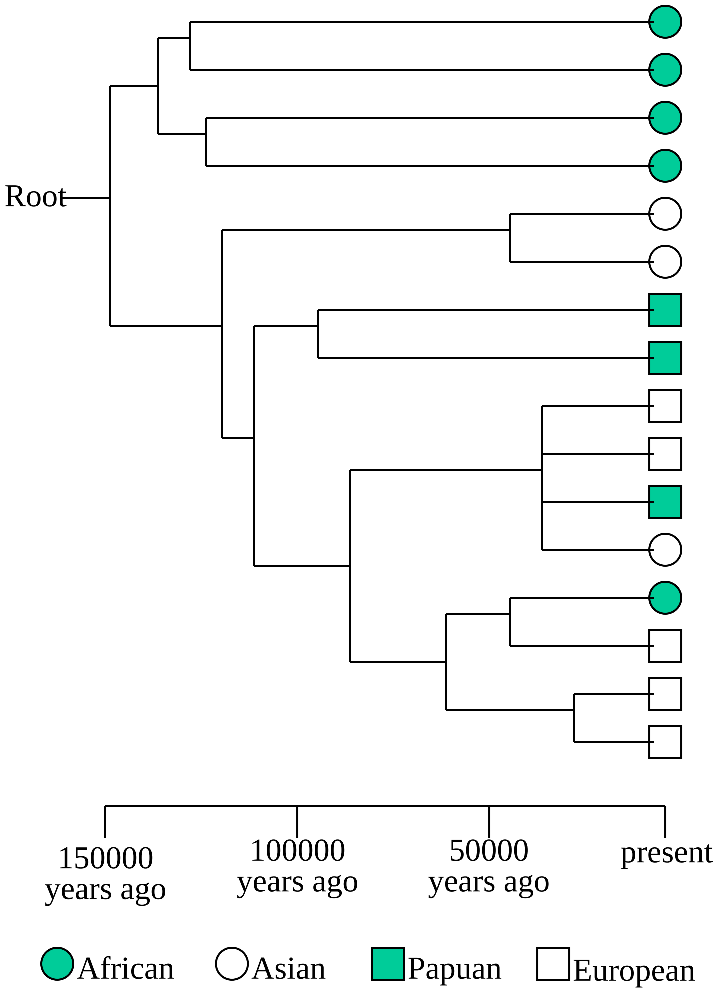
NUS
National University
of Singapore

Population tree



- **Estimate order in which "populations" evolved**
- **Based on assimilated freq 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



- **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**

Why mitochondrial DNA

- **Present in abundance in bone fossils**
- **Inherited only from mother**
- **Sufficient to look at the 500bp control region**
- **Accumulate more neutral mutations than nuclear DNA**
- **Accumulate mutations at the “right” rate, about 1 every 10,000 years**
- **No recombination, not shuffled at each generation**

Mutation rates

- All pet golden hamsters in the world descend from a single female caught in 1930 in Syria
- Golden hamsters “manage” ~4 generations a year :-)
- So >250 hamster generations since 1930
- Mitochondrial control regions of 35 (independent) golden hamsters were sequenced and compared
- No mutation was found

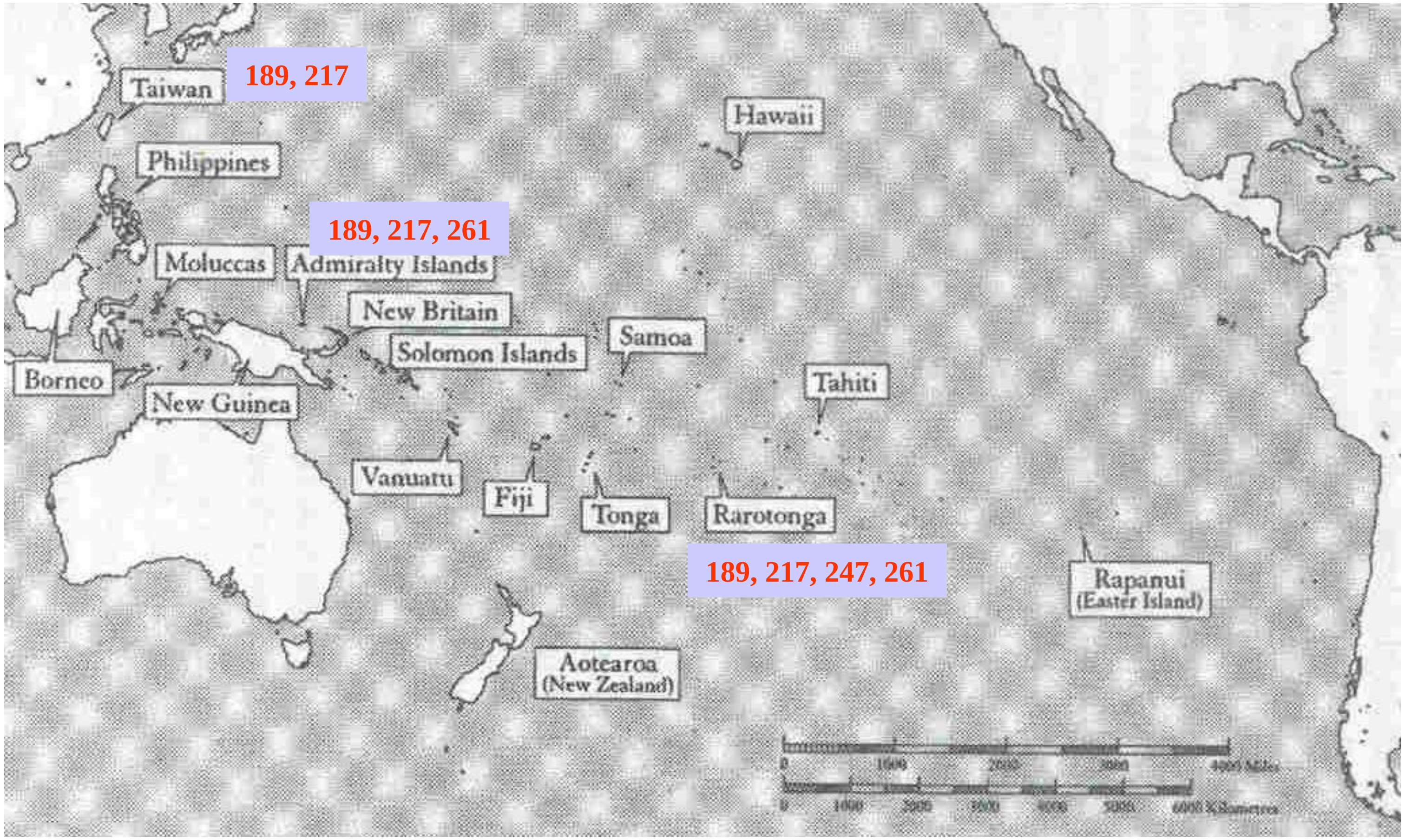
⇒ Mitochondrial control region mutates at the “right” rate

Contamination

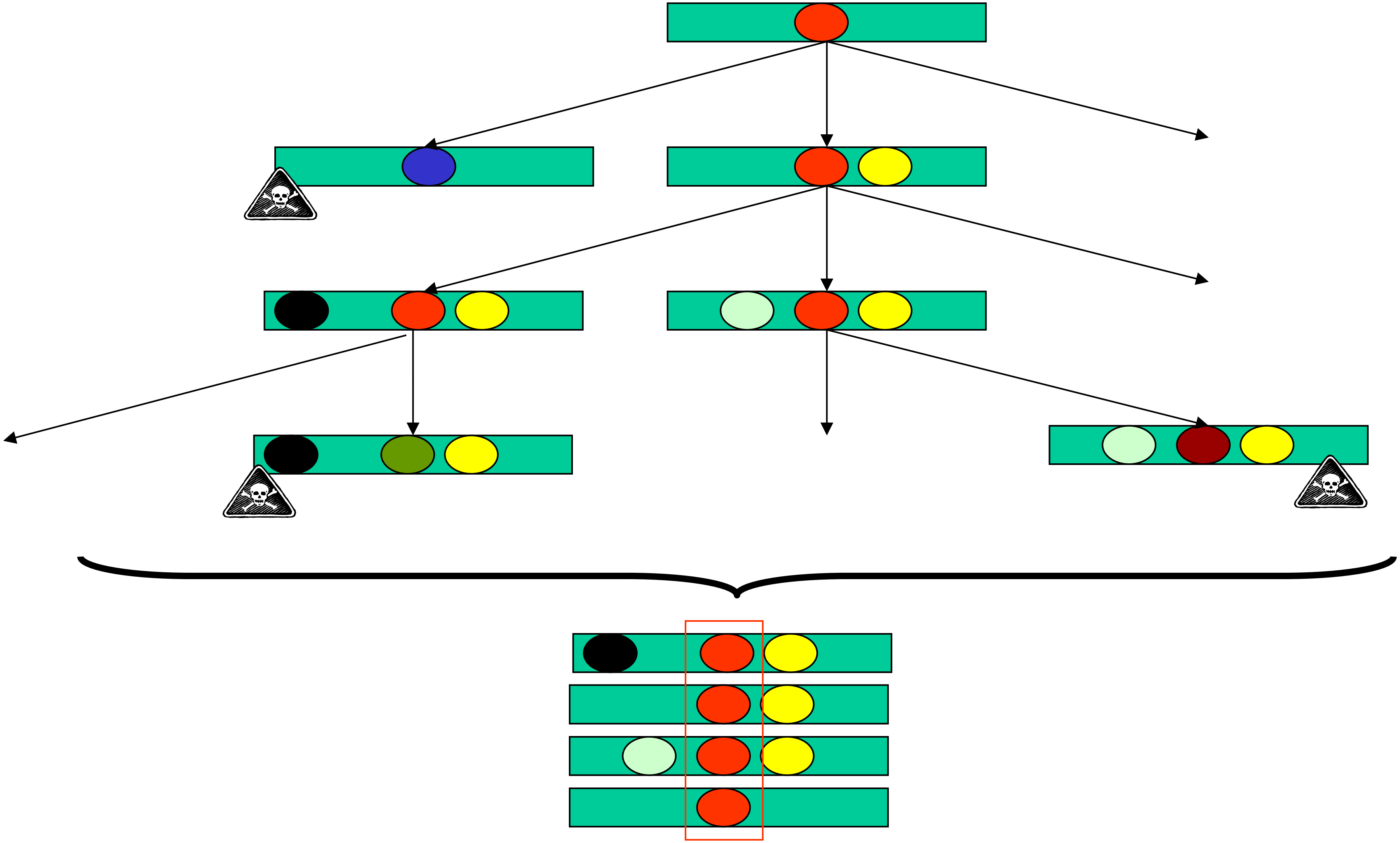
- Need to know if DNA extracted from old bones really from those bones, and not contaminated with modern human DNA
- Apply same procedure to old bones from animals, check if you see modern human DNA
- If none, then procedure is OK

Origin of Polynesians

- **Do they come from Asia or America?**



In the course of evolution…



Origin of Polynesians

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

- Seq from Taiwan natives have variants 189, 217

- Seq from regions in betw have variants 189, 217, 261.

- More 189, 217 closer to Taiwan. More 189, 217, 261 closer to Rarotonga
- 247 not found in America

⇒ Polynesians came from Taiwan!

- Taiwan seq sometimes have extra mutations not found in other parts

⇒ 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

Neanderthal vs Cro Magnon

- Based on palaeontology, Neanderthal & Cro Magnon last shared an ancestor 250000 yrs ago
- Mitochondrial control regions accumulate 1 mutation per 10000 yrs

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

- The number of diff betw Welsh is ~3, & at most 8.
- When compared w/ other Europeans, 14 diff at most

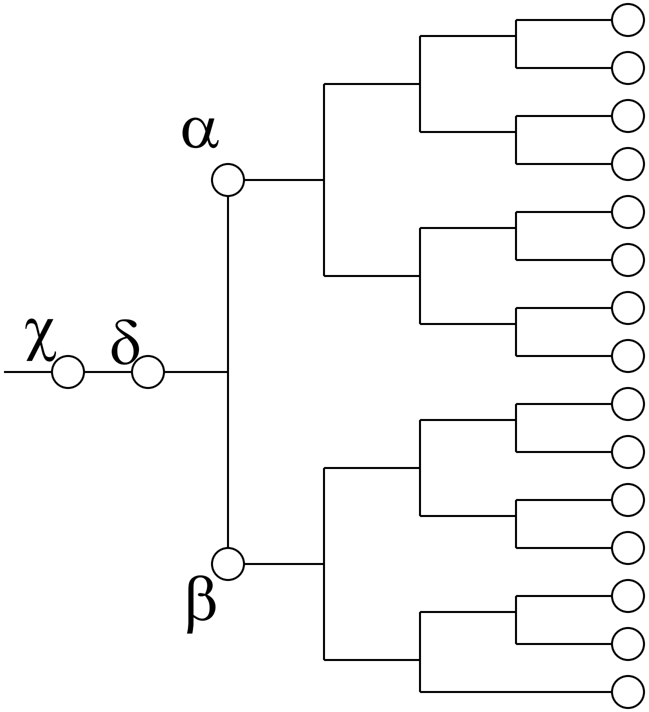
⇒ Ancestor either 100% Neanderthal or 100% Cro Magnon

- Mitochondrial control seq from Neanderthal have 26 diff from Europeans

⇒ Ancestor must be 100% Cro Magnon

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

Clan Mother



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

- **Clan mother is the most recent maternal ancestor common to all members of the clan**
- **A woman with only sons cant be clan mother---her mitochondrial DNA cant be passed on**
- **A woman can’t be clan mother if she has only 1 daughter---she is not most recent maternal ancestor**

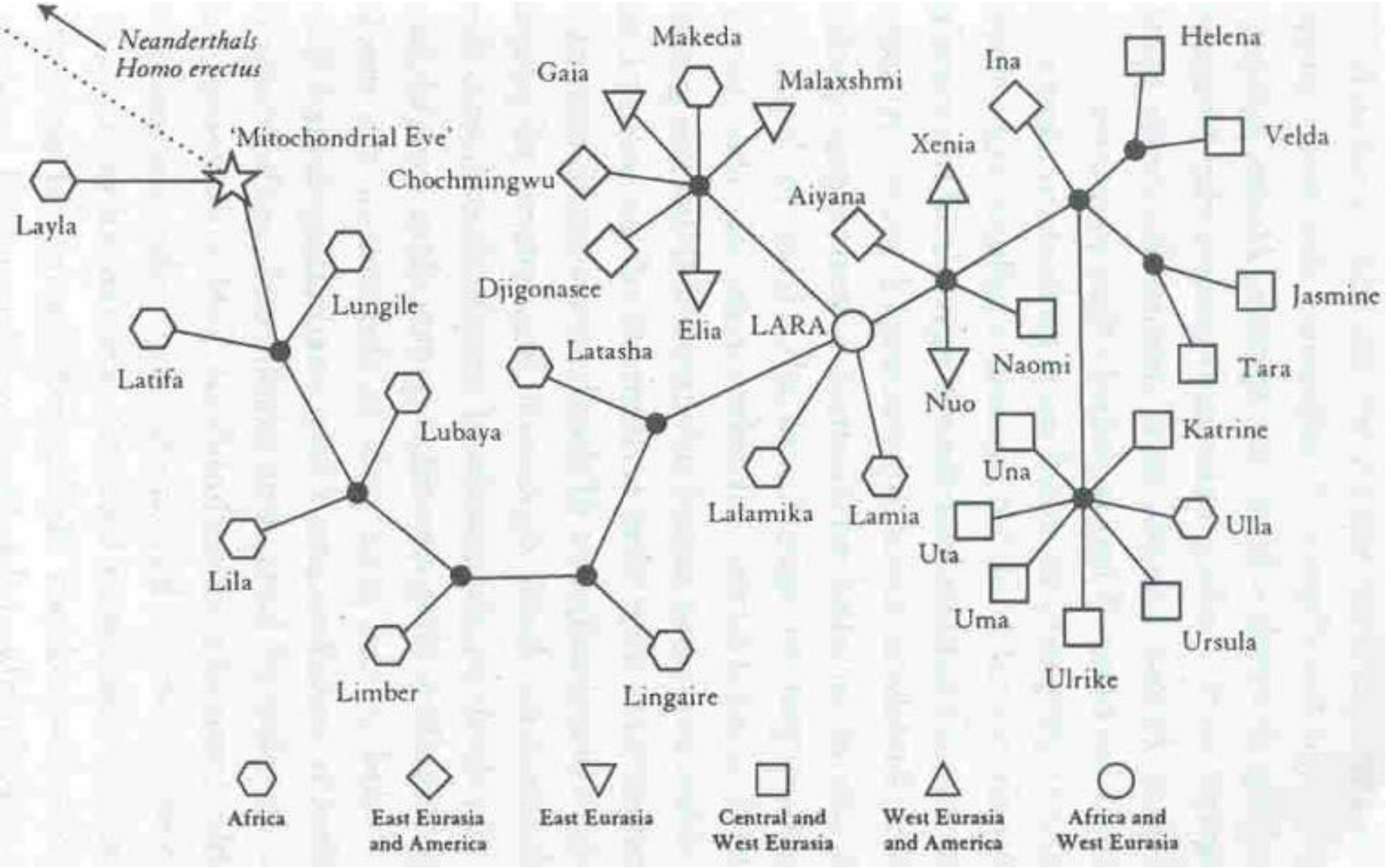
How many clans in Europe?

- Cluster seq according to mutations
- Each cluster thus represents a major clan
- European seq cluster into 7 major clans
- The 7 clusters age betw 45,000 and 10,000 years (length of time taken for all mutations in a cluster to arise from a single founder seq)

- The founder seq 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

Exercise: How about clan father?

World Clans


Neanderthals
Homo erectus
'Mitochondrial Eve'
Layla
Lungile
Latifa
Lubaya
Lila
Limber
Lingaire
Makeda
Gaia
Malaxshmi
Chochmingwu
Djigonasee
Elia
LARA
Latasha
Lalamika
Lamia
Xenia
Aiyana
Nuo
Naomi
Ina
Helena
Velda
Jasmine
Tara
Katrine
Una
Uta
Uma
Ulla
Ulrike
Ursula
Africa
East Eurasia and America
East Eurasia
Central and West Eurasia
West Eurasia and America
Africa and West Eurasia

Any Question?

NUS
National University
of Singapore

Acknowledgements

- **A lot of the slides from this lecture are given to me by Ken Sung**

References

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

- S.-W. Meng. Analysis of Phylogeny: A Case Study on Saururacea, *The Practical Bioinformatician*, chapter 11, pages 245—268, WSPC, 2004

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

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