UZH BIO390, Nov 2019

Sequence evolution & phylogenetics

Christophe Dessimoz

http://lab.dessimoz.org 🐦 @cdessimoz

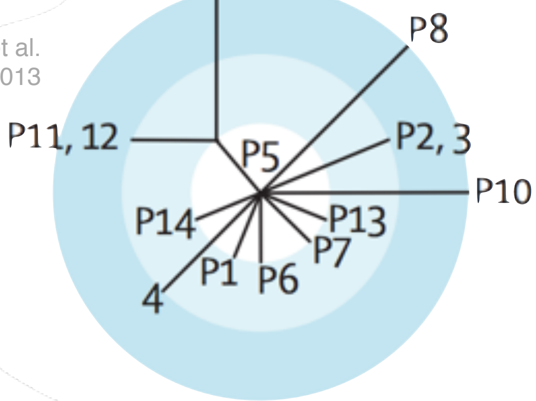
UCL

Unil
UNIL | Université de Lausanne

SIB

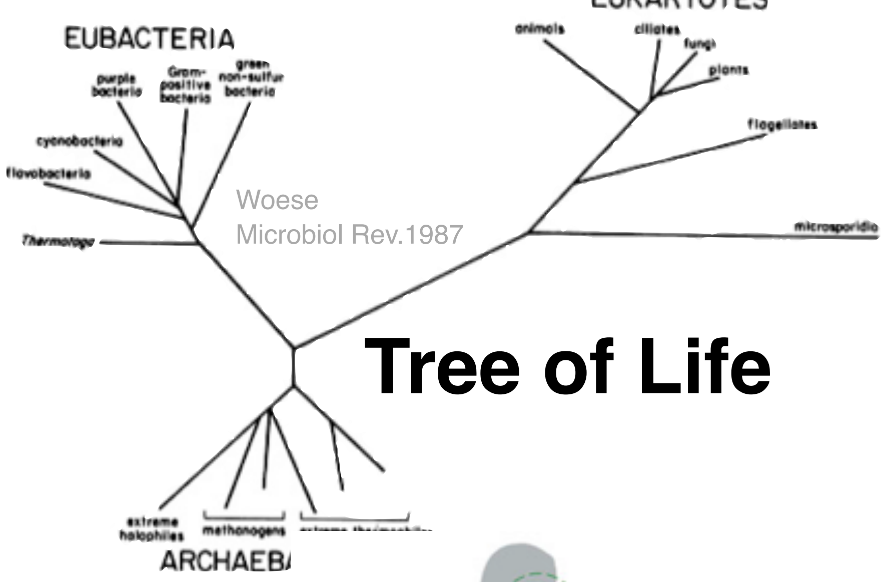
Today’s lecture

- Pairwise distance estimation
- Tree thinking & terminology
- Tree inference
  - Methods
  - Confidence
  - Rooting



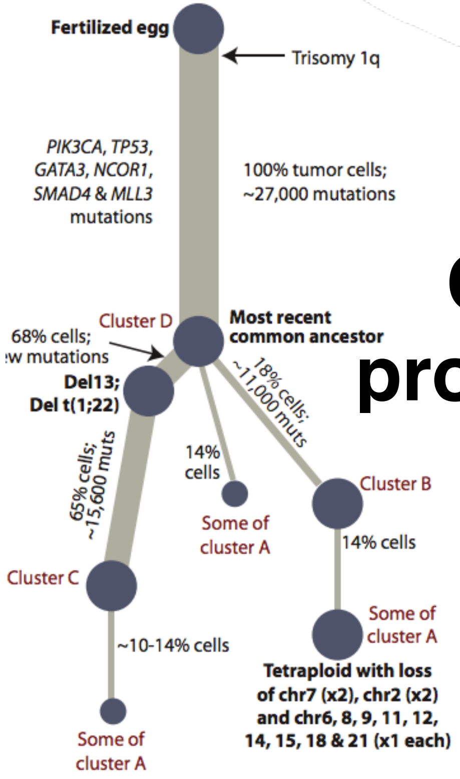
Harris et al.
Lancet Infect Dis 2013

Epidemiology



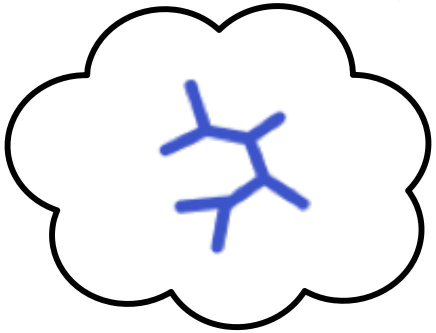
Woese
Microbiol Rev.1987

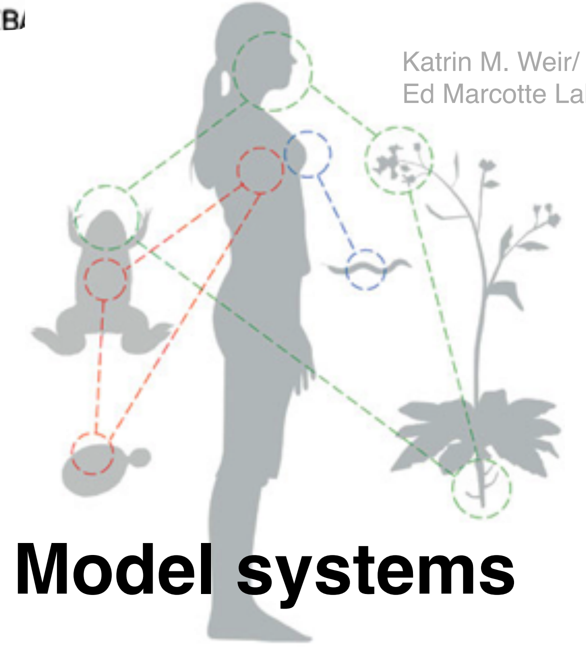
Tree of Life



Nik-Zainal et al. Cell 2012

Cancer progression

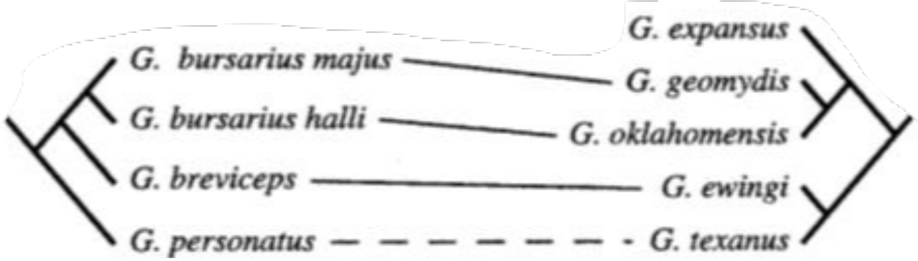




Katrin M. Weir/
Ed Marcotte Lab

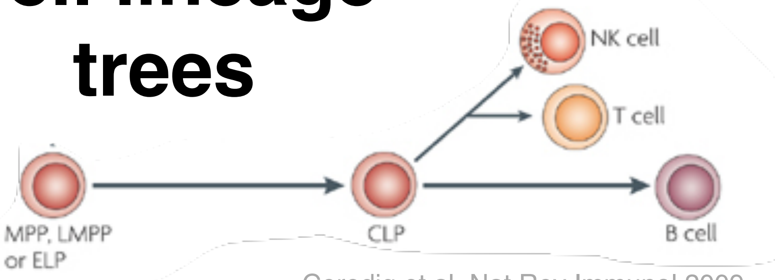
Model systems

Host/pathogen co-evolution



Legendre et al. Syst Biol 2002

Cell lineage trees



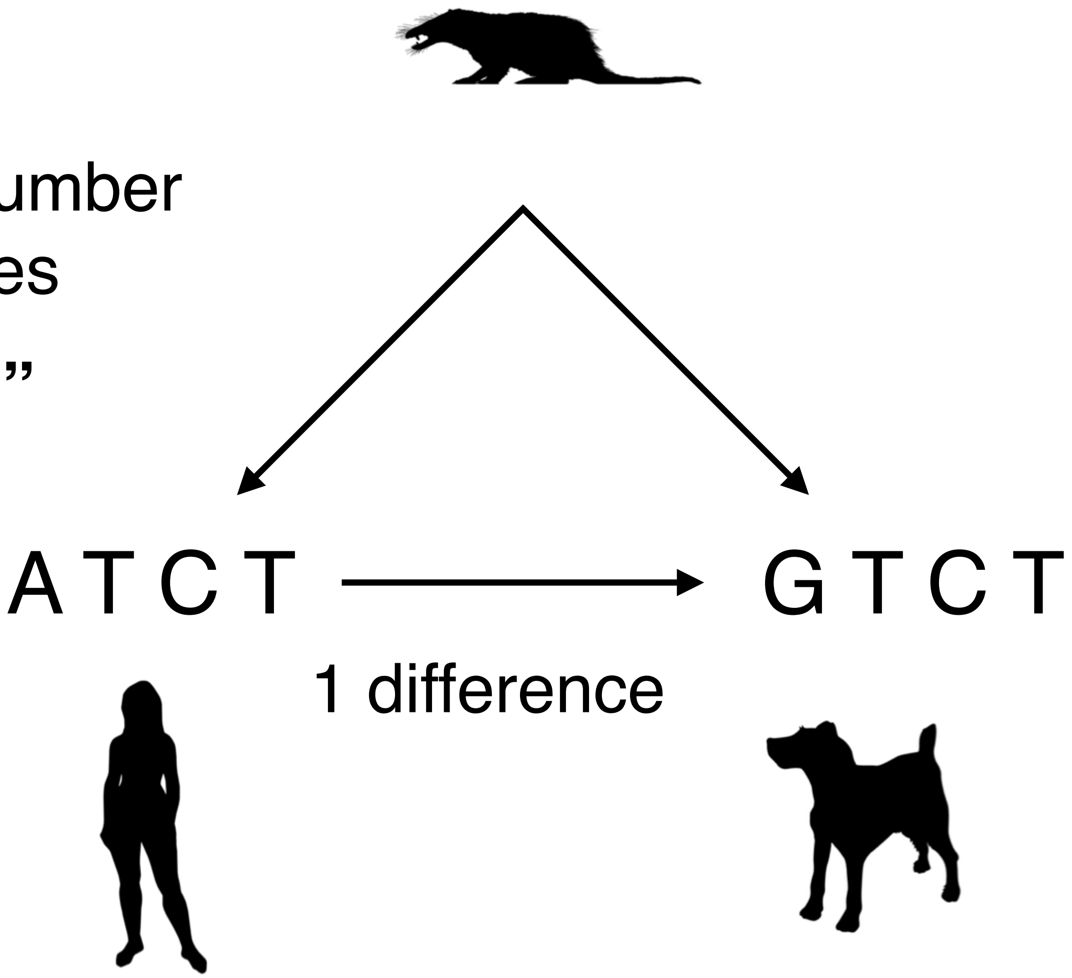
Ceredig et al. Nat Rev Immunol 2009

Pairwise distance estimation & Markov models

How to measure the amount of evolution?

Idea #1:
count the number of differences

“similarity”



A T C T

G T C T

1 difference

If we normalise with respect to length:

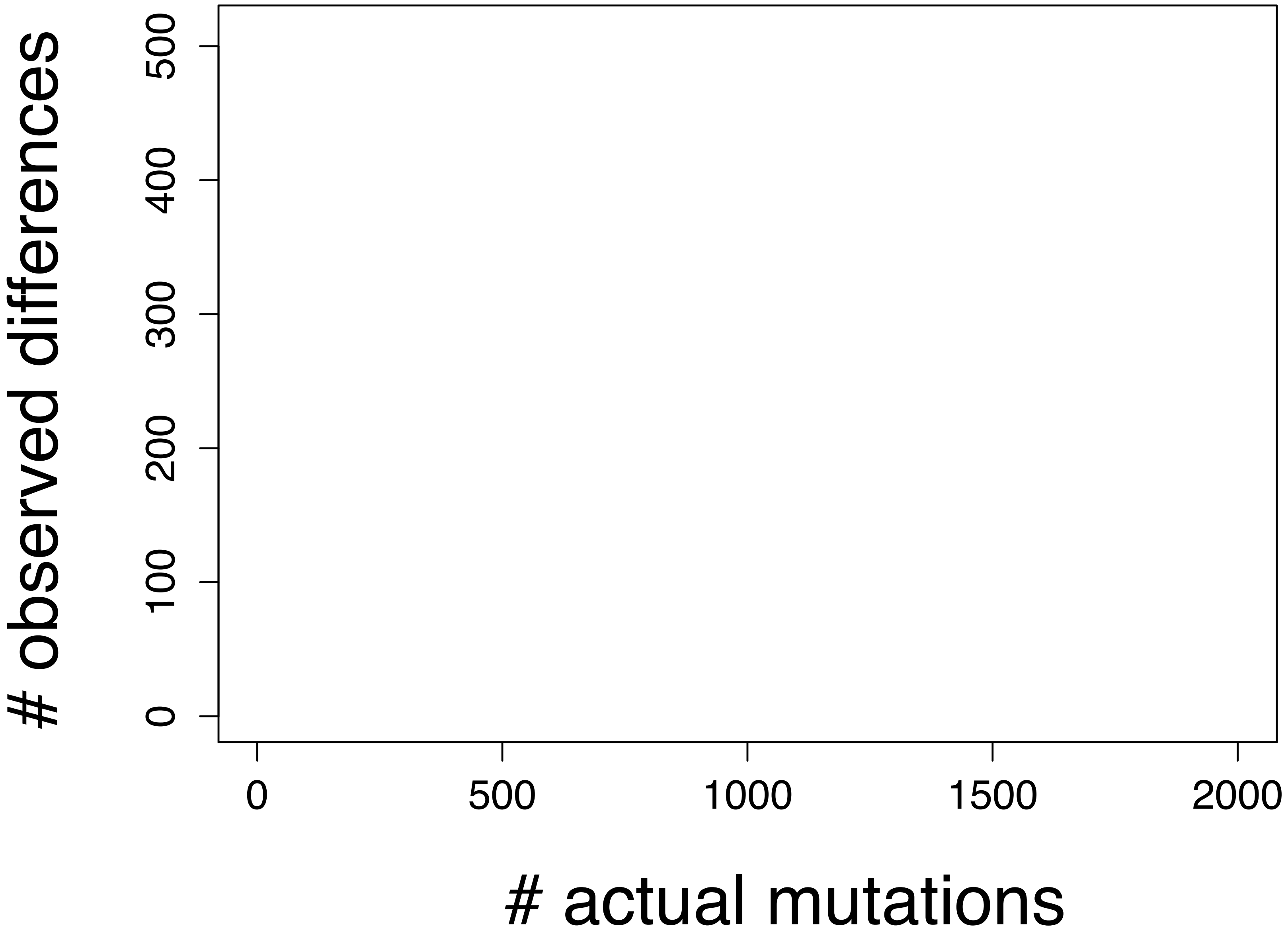
25% differences
(75% “identity”)

Problem with counting differences

		actual # of substitutions	*# of observed differences*
→	A T C T		
→	A T C T	0	0
	G T C T	1	1
	G T C T	1	1
	C T C T	2	1
	A T C T	3	0

Let's simulate sequence evolution

- Generate a random sequence of 0 and 1 of length 1000
- Repeat 2000 times:
    - Mutate a random position in the sequence
    - Compute the number of difference between the resulting sequence and the original sequence and store this number in a table
- Plot the values stored in the table


0
100
200
300
400
500
# observed differences
0
500
1000
1500
2000
# actual mutations

Markov model of evolution

- Every site evolves independently, with prob. of mutation only depending on present state (no memory).
- Probabilities of mutation at a given distance are expressed by transition matrix.

$$M^1 = \begin{array}{|c|c|c|c|c|} \hline \swarrow & A & C & G & T \\ \hline A & 0.900 & 0.033 & 0.033 & 0.033 \\ \hline C & 0.033 & 0.900 & 0.033 & 0.033 \\ \hline G & 0.033 & 0.033 & 0.900 & 0.033 \\ \hline T & 0.033 & 0.033 & 0.033 & 0.900 \\ \hline \end{array}$$

After “one unit” of evolution, the probability that an A mutates into a C is given by the corresponding entry in the matrix:

$$p(A \rightarrow C \mid d=1) = M^1[A \rightarrow C] = 0.033$$

$M^2=$

↙	A	C	G	T
A	0.900	0.033	0.033	0.033
C	0.033	0.900	0.033	0.033
G	0.033	0.033	0.900	0.033
T	0.033	0.033	0.033	0.900

x

↙	A	C	G	T
A	0.900	0.033	0.033	0.033
C	0.033	0.900	0.033	0.033
G	0.033	0.033	0.900	0.033
T	0.033	0.033	0.033	0.900

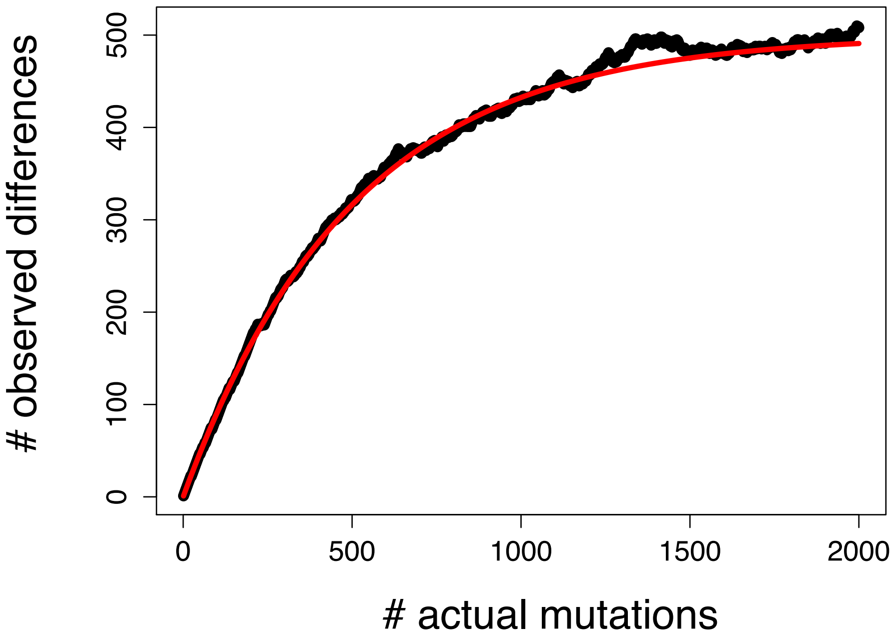
$M^2=$

↙	A	C	G	T
A	0.813	0.062	0.062	0.062
C	0.062	0.813	0.062	0.062
G	0.062	0.062	0.813	0.062
T	0.062	0.062	0.062	0.813

$$\begin{aligned} M^2[A,A] = {} & M^1[A\rightarrow A] * M^1[A,\rightarrow A] + \\ & M^1[A\rightarrow C] * M^1[C\rightarrow A] + \\ & M^1[A\rightarrow G] * M^1[G\rightarrow A] + \\ & M^1[A\rightarrow T] * M^1[T\rightarrow A] \end{aligned}$$

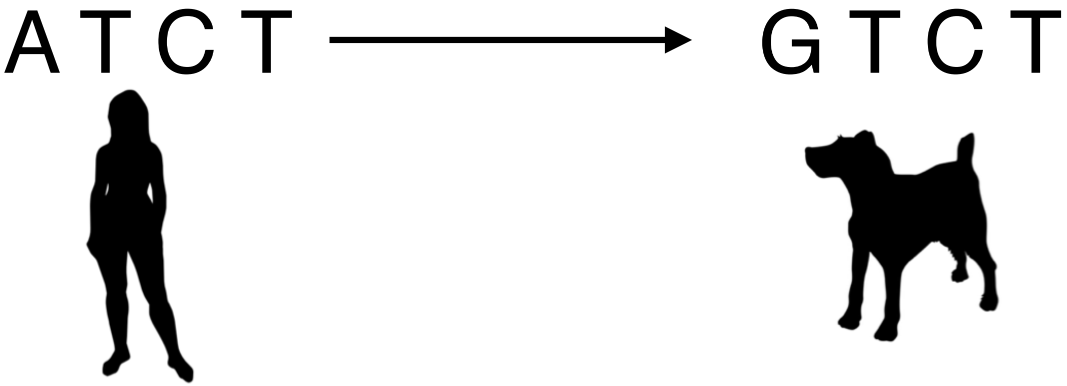
$M^\infty=$

↙	A	C	G	T
A	0.250	0.250	0.250	0.250
C	0.250	0.250	0.250	0.250
G	0.250	0.250	0.250	0.250
T	0.250	0.250	0.250	0.250



Distance estimation

- Now that we have a model and data, how can we estimate the distance?



A T C T → G T C T

$$M^1 = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & 0.900 & 0.033 & 0.033 & 0.033 \\ C & 0.033 & 0.900 & 0.033 & 0.033 \\ G & 0.033 & 0.033 & 0.900 & 0.033 \\ T & 0.033 & 0.033 & 0.033 & 0.900 \end{array}$$

Let’s assume that the distance is 1 “unit” of evolution. Given the original sequence ATCT, the likelihood of observing GTCT is:

$$M^1[A,G] * M^1[T,T] * M^1[C,C] * M^1[T,T] = 0.0243$$

is that likely?!

A brief extrapolation on maximum likelihood (ML) parameter estimation

Ronald A. Fisher

The likelihood function

L(parameter ; data) = p(data l parameter)

ML estimator: choose parameter that maximises the likelihood function!

Note that L() is not a probability distribution (does not sum to 1)

Another example:

Unfair coin. data: T T T T H T H

Model: $p(x{=}T \mid \theta) = \theta$

$p(x{=}H \mid \theta) = 1-\theta$

$L(\theta; \text{data}) = \theta*\theta*\theta*\theta*(1-\theta)*\theta*(1-\theta)$

Unfair coin. data: T T T T H T H

Model: $p(x=T \mid \theta) = \theta$

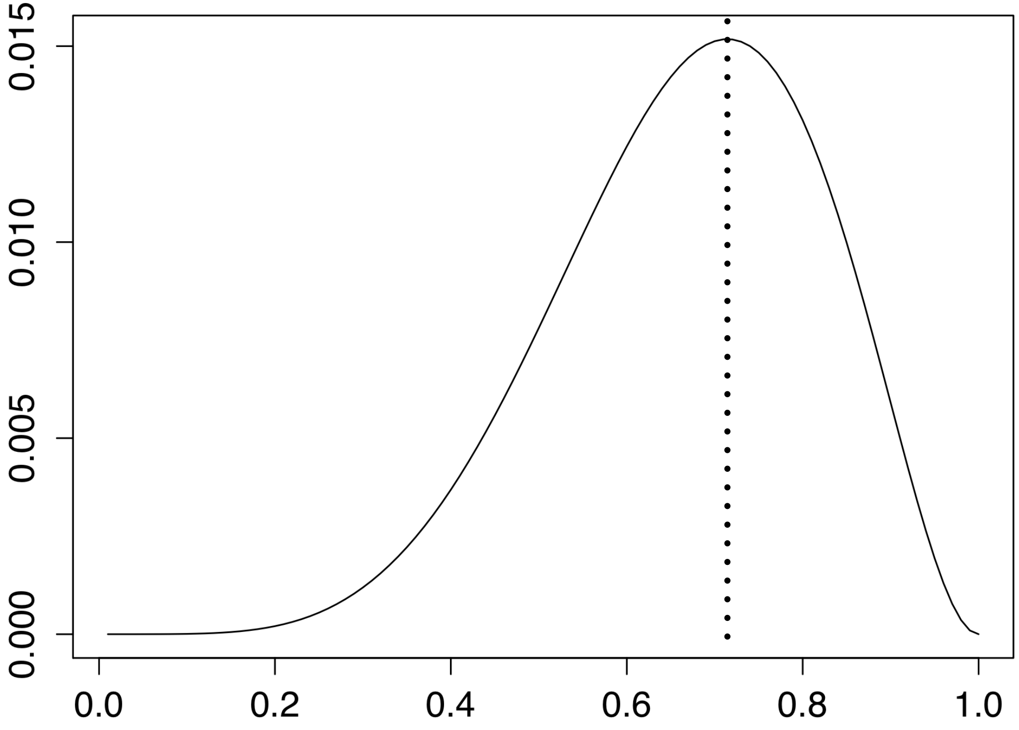
$p(x=H \mid \theta) = 1-\theta$

$L(\theta; data) = \theta*\theta*\theta*\theta*(1-\theta)*\theta*(1-\theta)$

$= \theta^5*(1-\theta)^2$

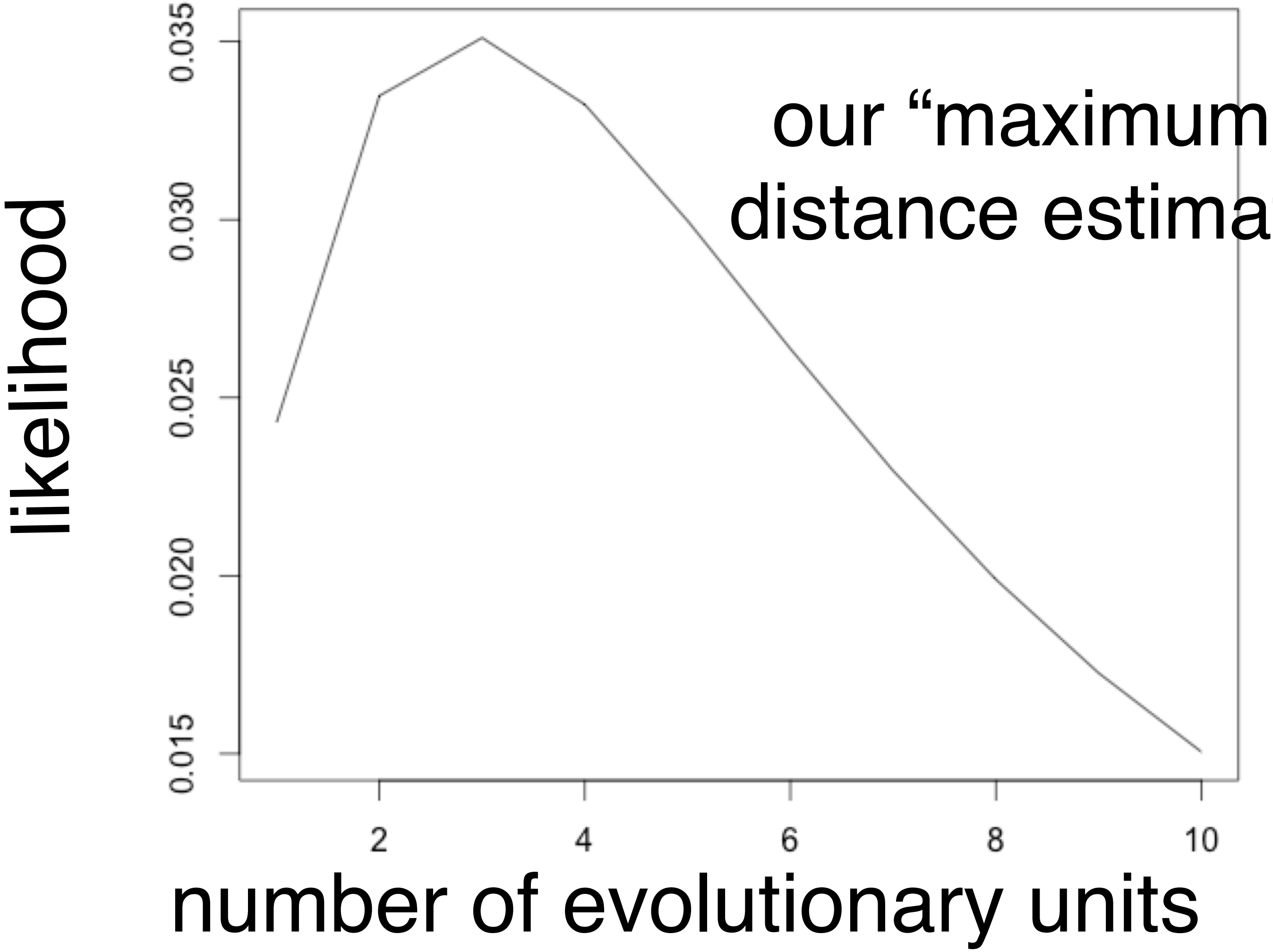
Maximise L(θ; data)

0.71= 5/7



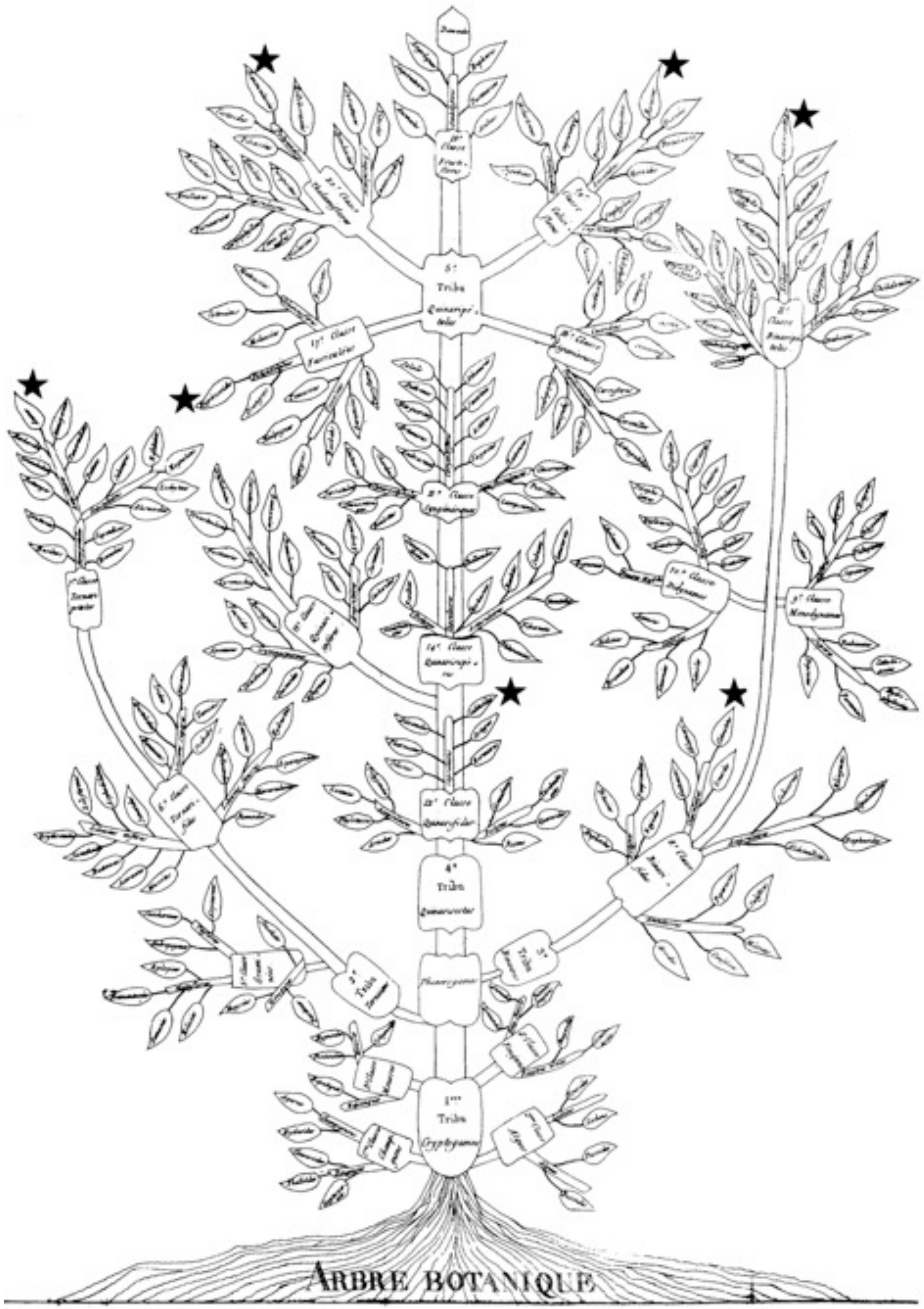
back to our example

A T C T ⟶ G T C T

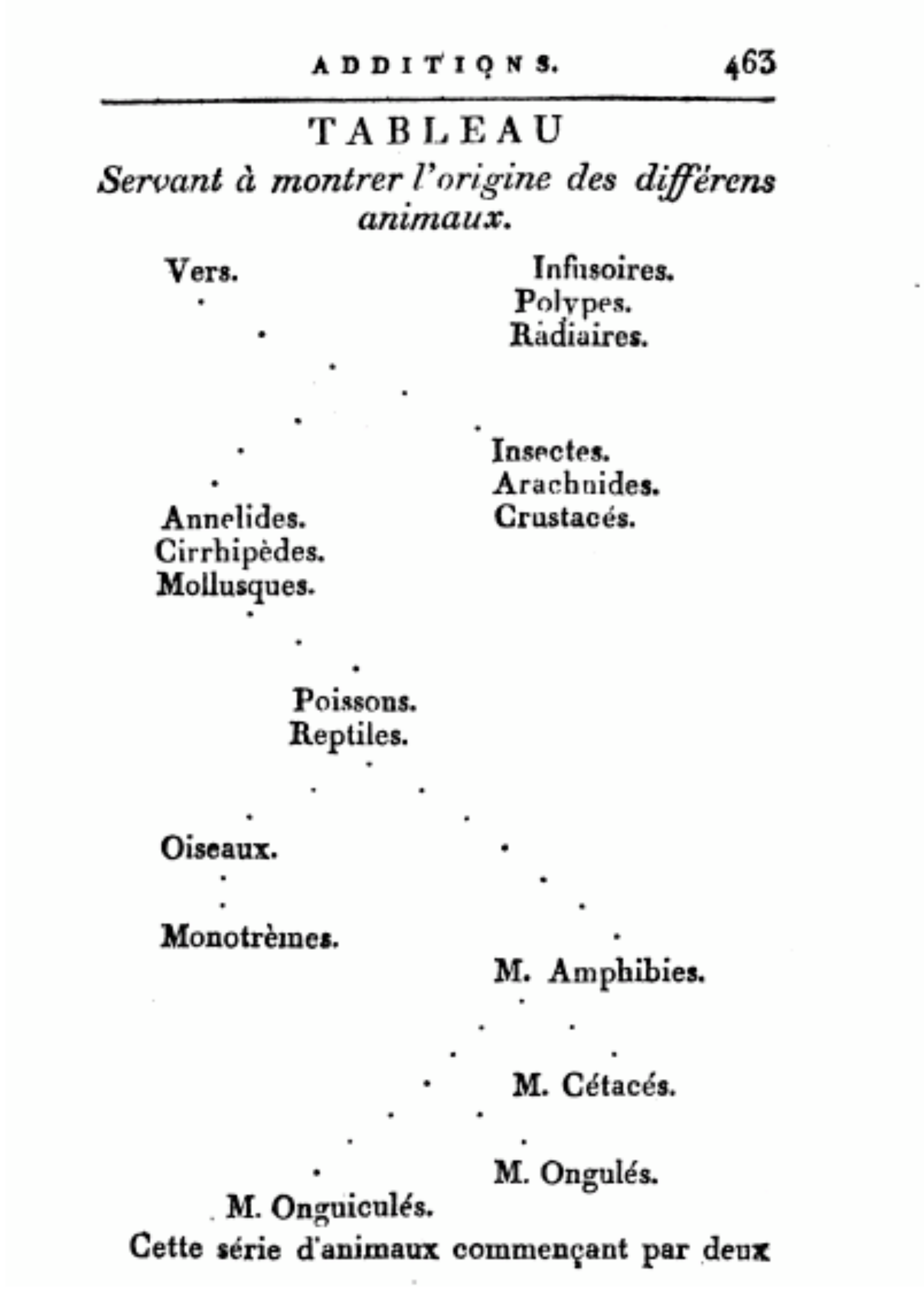


our “maximum likelihood” distance estimate is 3 units!

Augustin Augier,
Arbre Botanique
(1801)



Lamarck, Philosophie Zoologique , 1809



ADDITIONS. 463

TABLEAU

Servant à montrer l'origine des différens animaux.

Vers.

Infusoires.
Polypes.
Radiaires.

Insectes.
Arachnides.
Crustacés.

Annelides.
Cirrhipèdes.
Mollusques.

Poissons.
Reptiles.

Oiseaux.

Monotrèmes.

M. Amphibies.

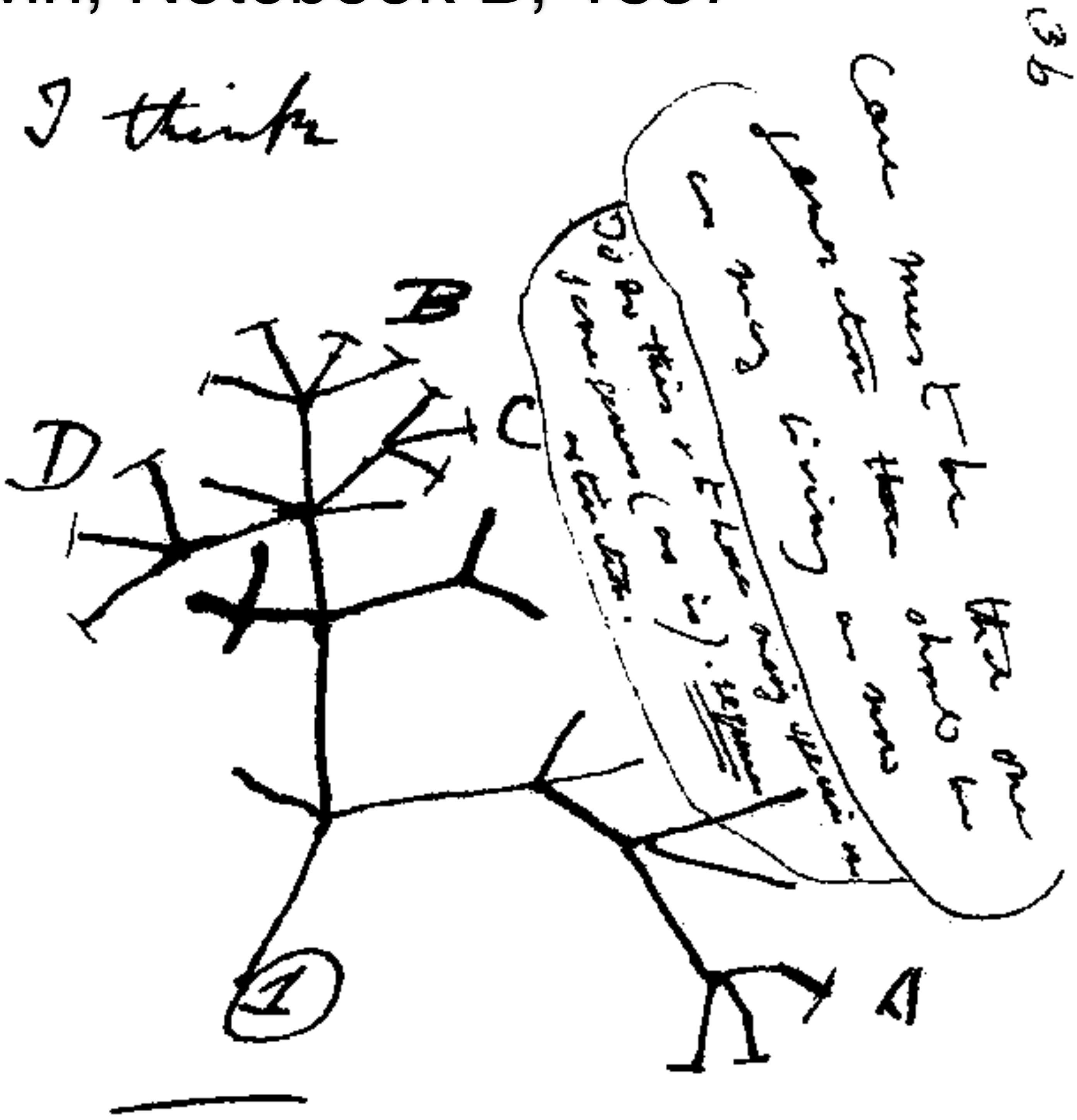
M. Cétacés.

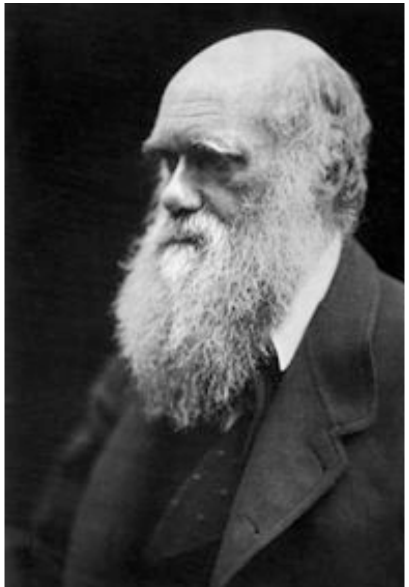
M. Ongulés.

M. Onguiculés.

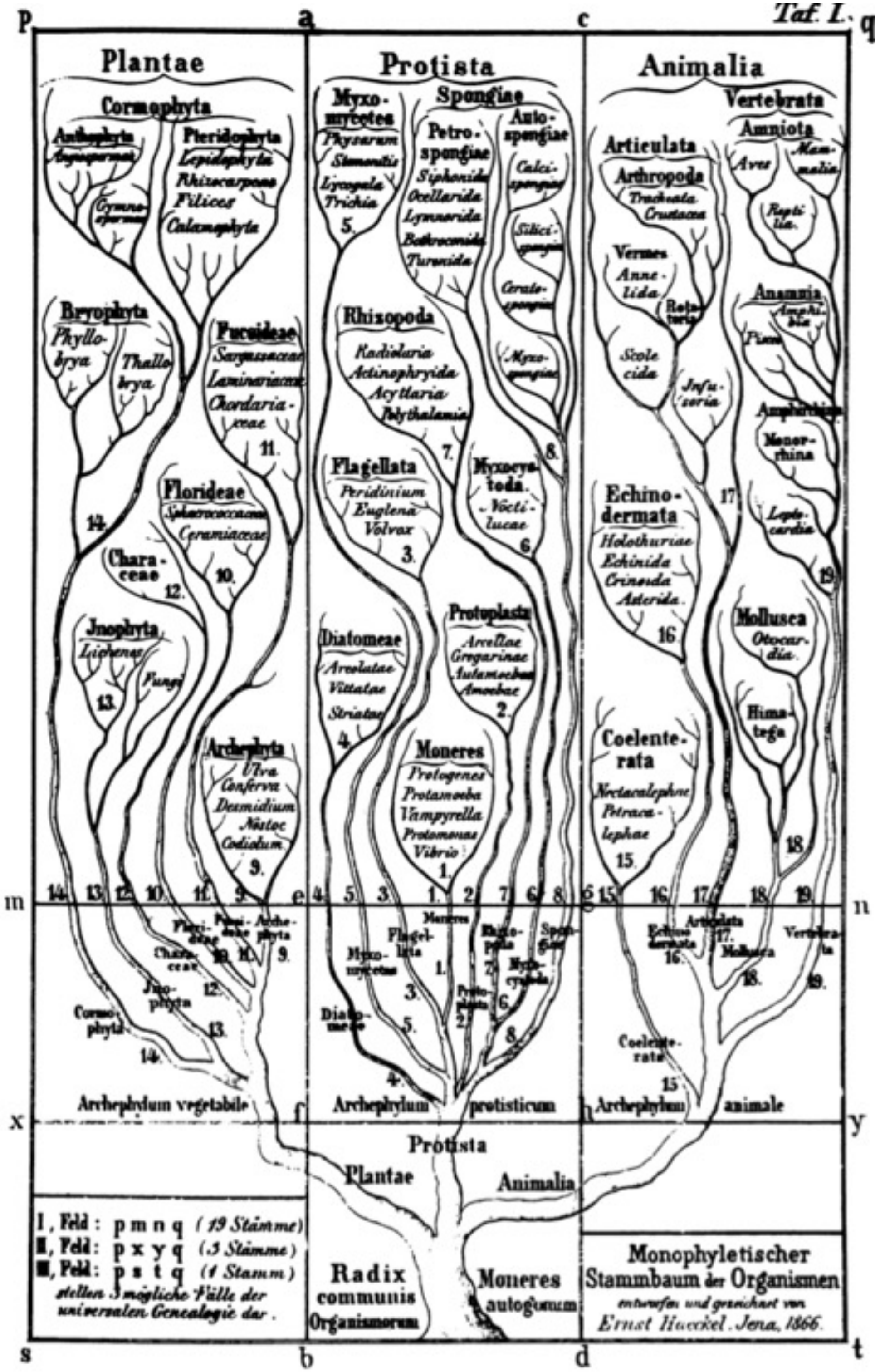
Cette série d'animaux commençant par deux

Darwin, Notebook B, 1837


36
I think
Case must be that one generation then should be as many living as now
To do this & to have many species in same genus (as is). requires extinction.
A
B
C
D
1



Wikipedia

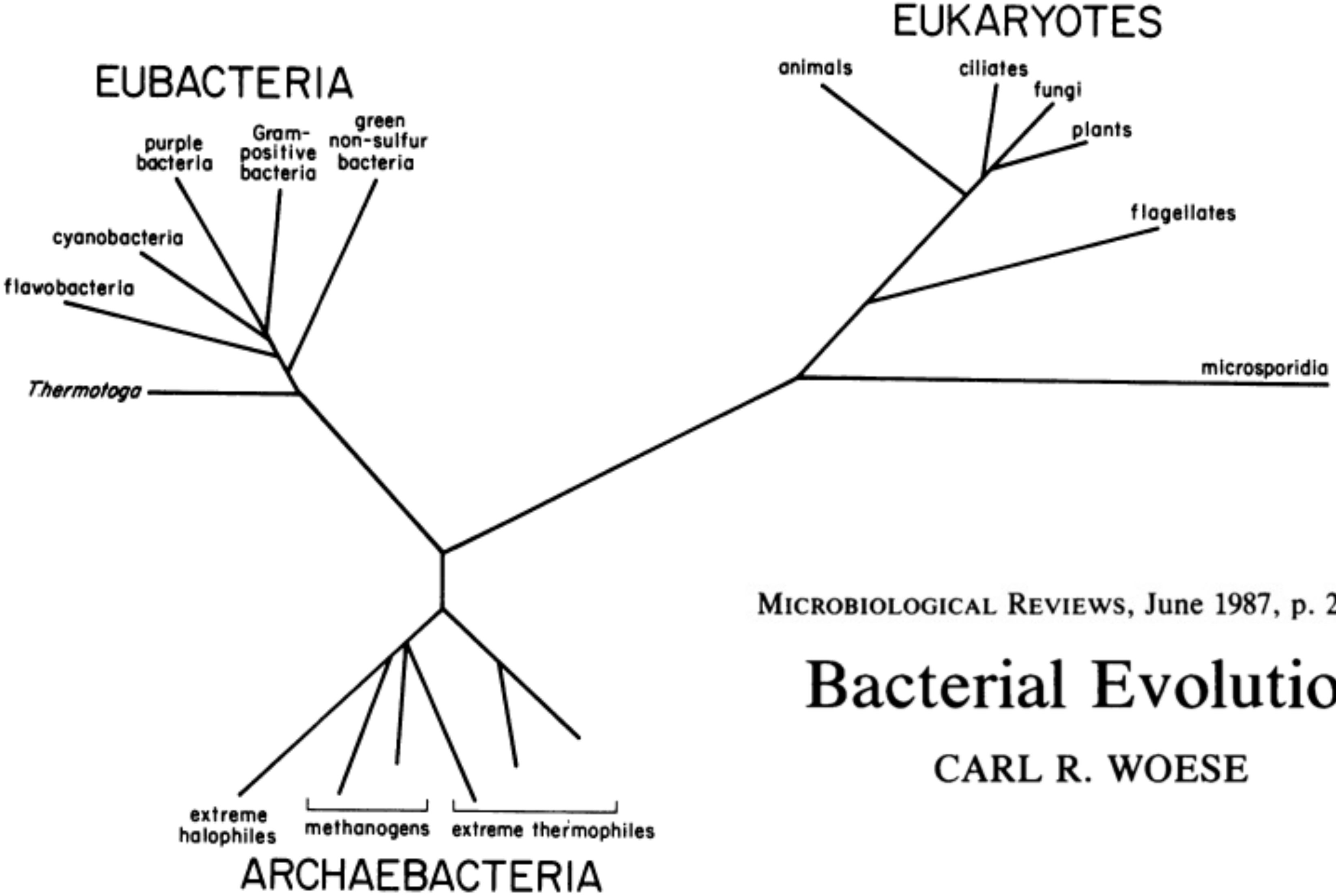


Monophyletischer Stammbaum der Organismen

entworfen und gezeichnet von

Ernst Haeckel. Jena, 1866.

16S rRNA was used by Woese (1987) to group early life forms into three kingdoms



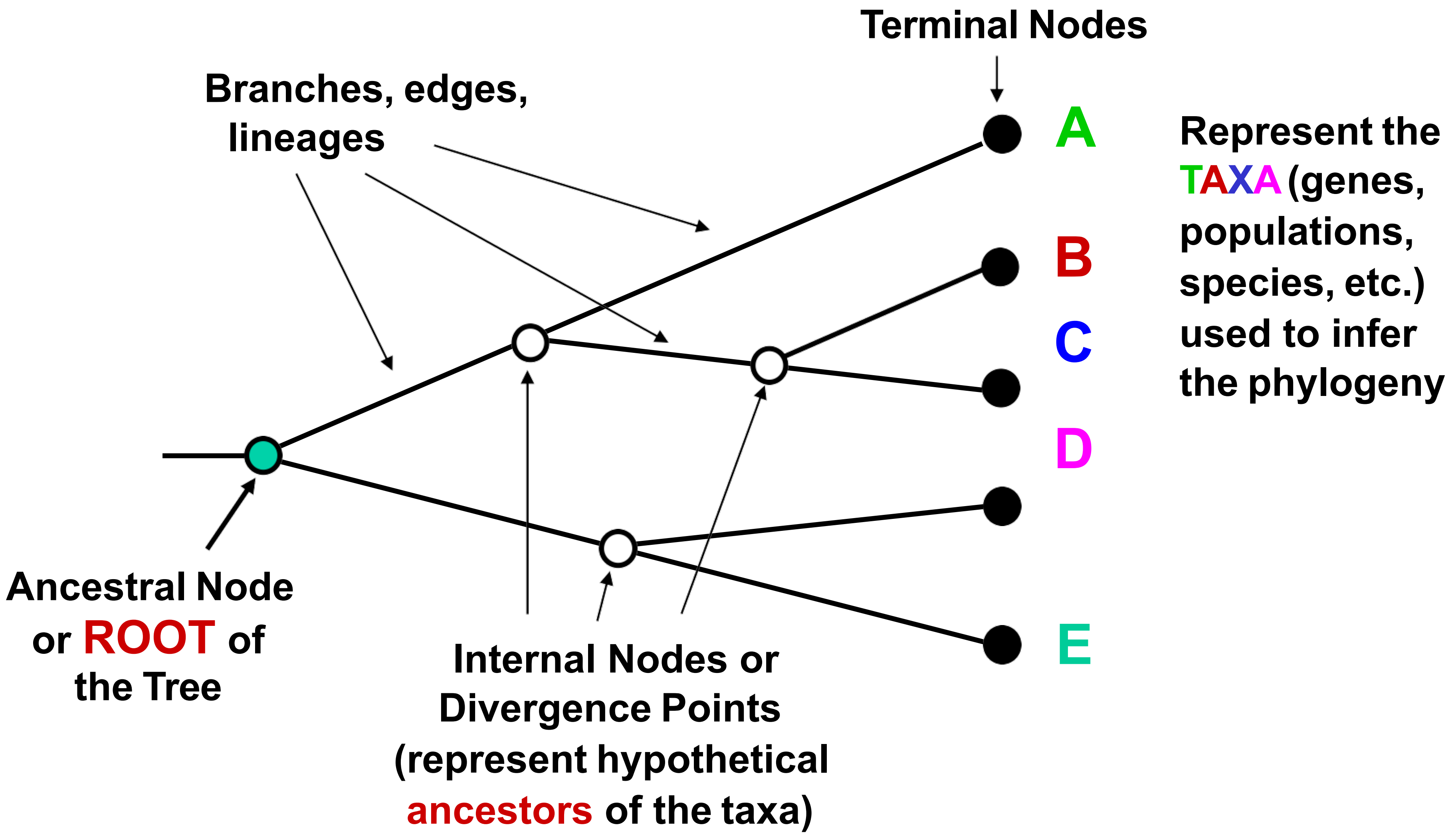
MICROBIOLOGICAL REVIEWS, June 1987, p. 221–271

Bacterial Evolution

CARL R. WOESE

Tree terminology
and
“tree thinking”

Common Phylogenetic Tree Terminology



Terminal Nodes

Branches, edges, lineages

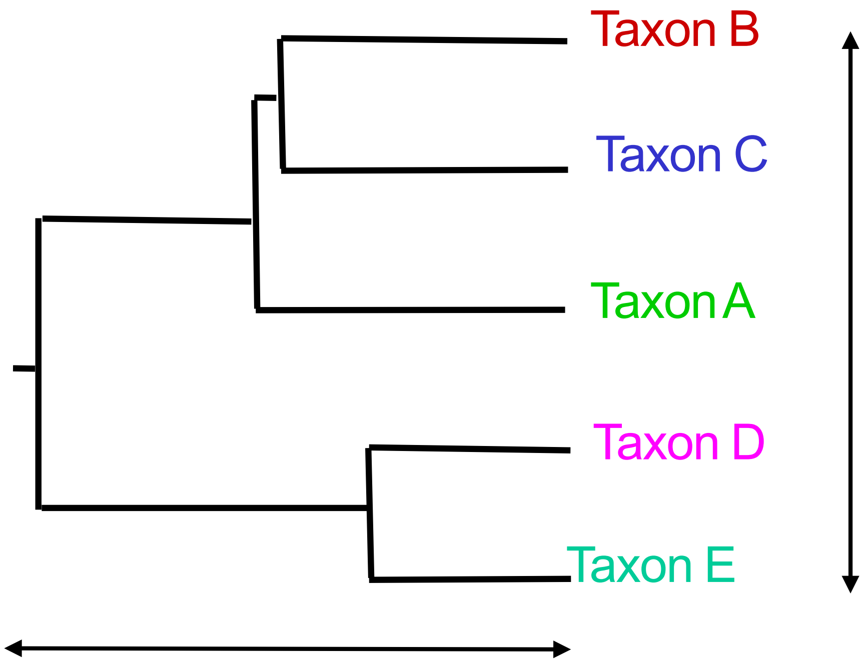
Represent the TAXA (genes, populations, species, etc.) used to infer the phylogeny

Ancestral Node or ROOT of the Tree

Internal Nodes or Divergence Points (represent hypothetical ancestors of the taxa)

Slide credit: HK

Phylogenetic trees diagram the *evolutionary relationships* between the taxa



No meaning to the spacing between the taxa, or to the order in which they appear from top to bottom.

This dimension either can have no scale (for 'cladograms'), can be proportional to genetic distance or amount of change (for 'phylograms' or 'additive trees'), or can be proportional to time (for 'ultrametric trees' or true evolutionary trees).

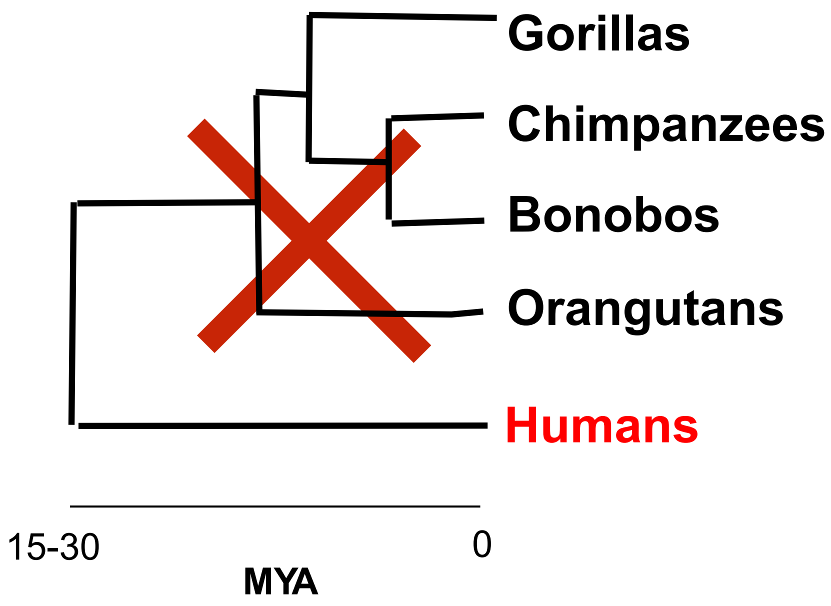
((A,(B,C)),(D,E)) = The above phylogeny as nested parentheses

These say that B and C are more closely related to each other than either is to A, and that A, B, and C form a **clade** that is a **sister group** to the clade composed of D and E. If the tree has a time scale, then D and E are the most closely related.

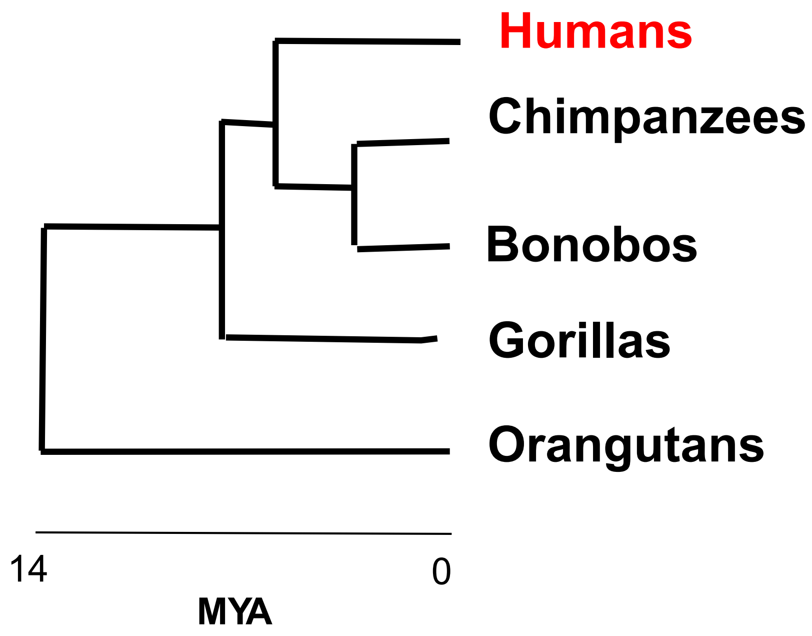
Slide credit: HK

Which species are the closest living relatives of modern humans?

morphological tree



mitochondrial DNA

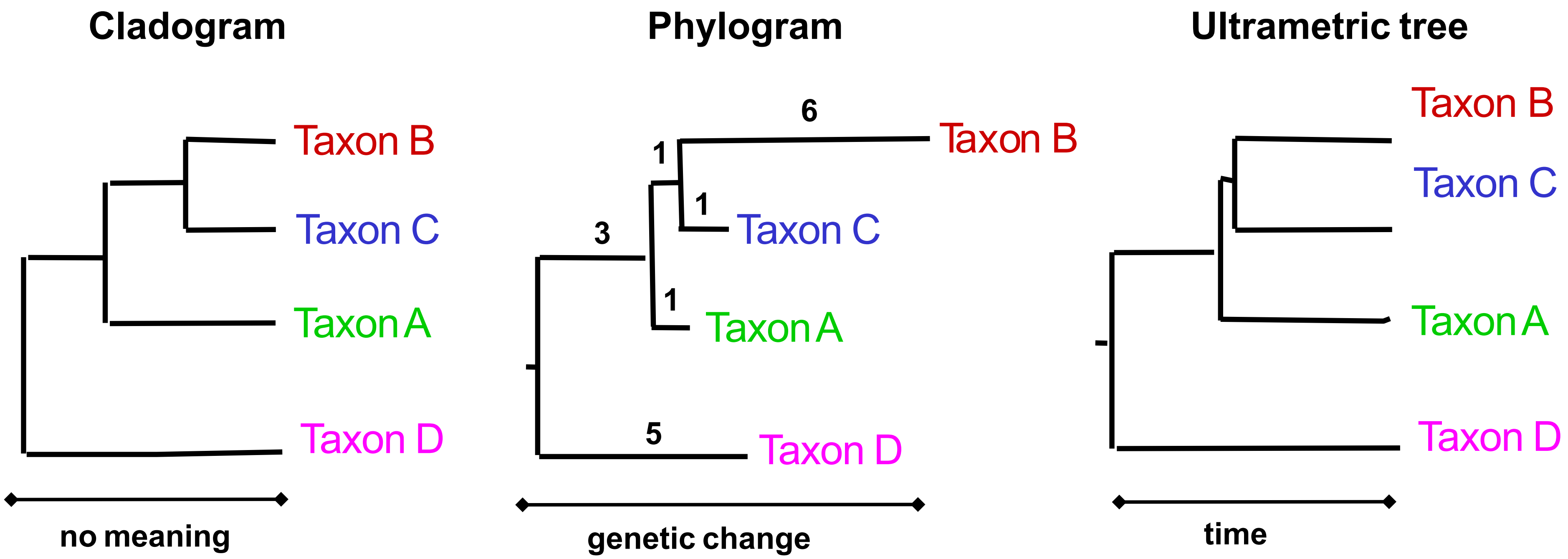


The pre-molecular view was that the great apes (chimpanzees, gorillas and orangutans) formed a clade separate from humans, and that humans diverged from the apes at least 15-30 MYA.

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.

Slide credit: HK

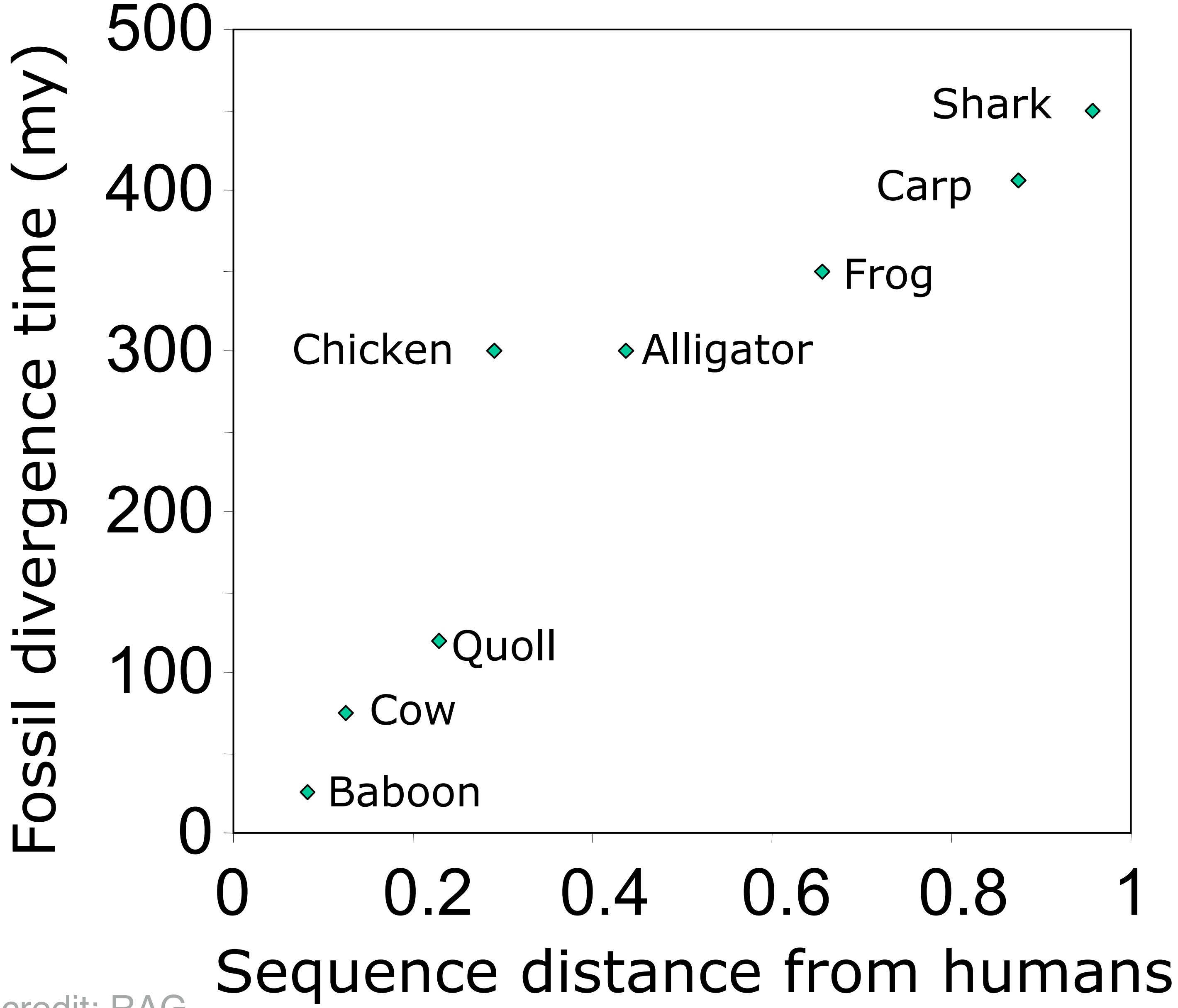
Three types of trees

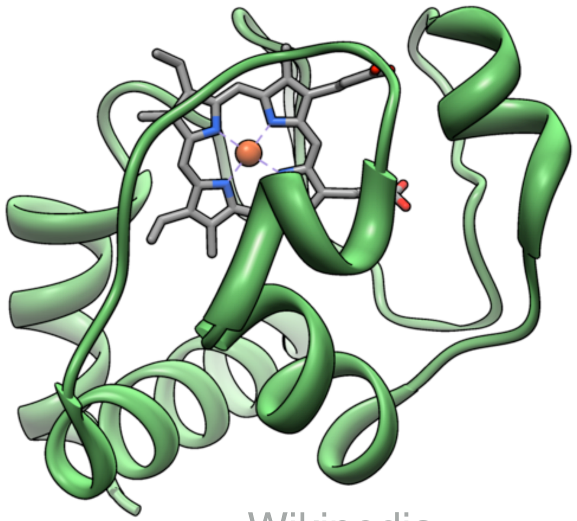


All show the same evolutionary relationships, or branching orders, between the taxa.

Slide credit: HK

Evidence for the Molecular Clock: Cytochrome c





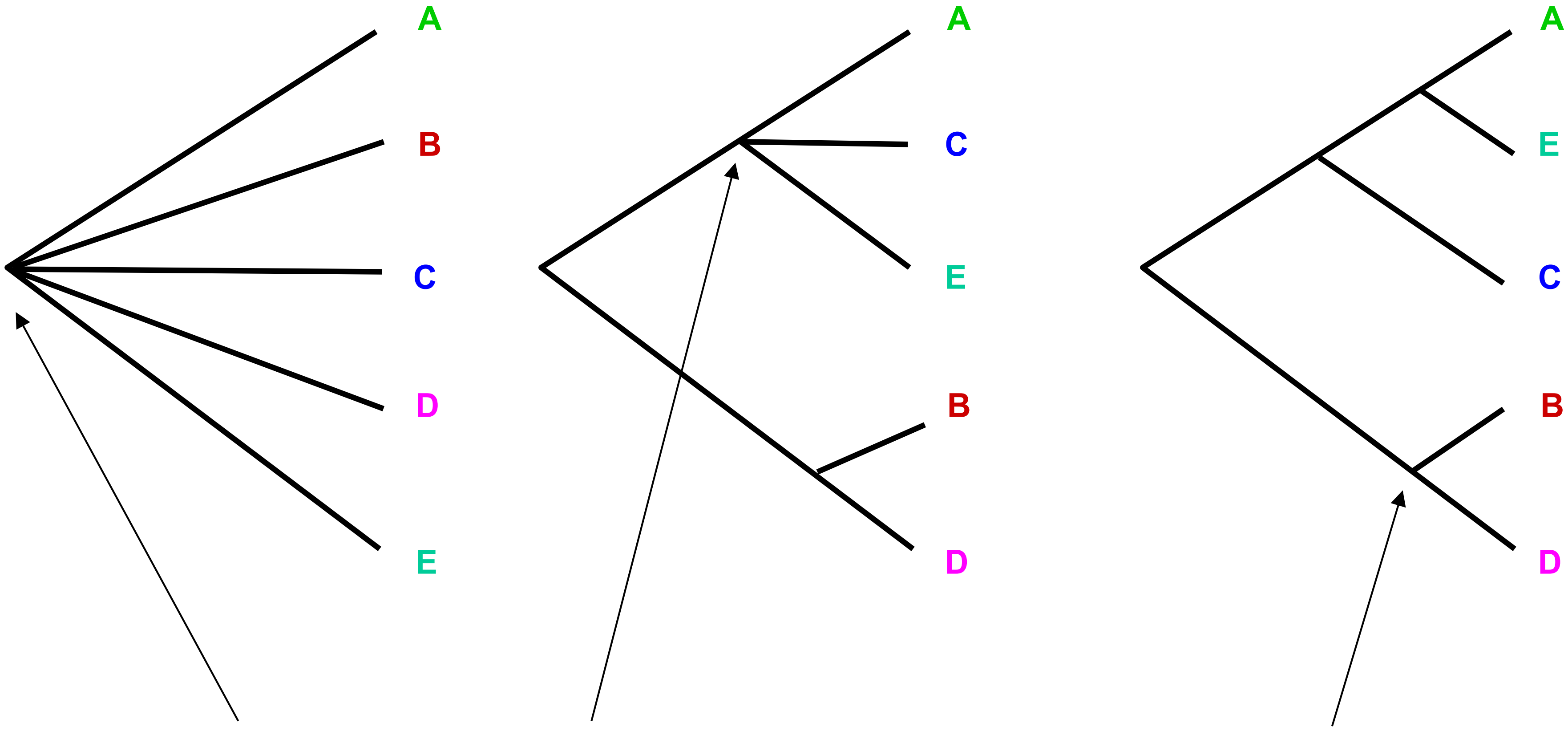
Wikipedia

Slide credit: RAG

Completely unresolved or "star" phylogeny

Partially resolved phylogeny

Fully resolved, bifurcating phylogeny

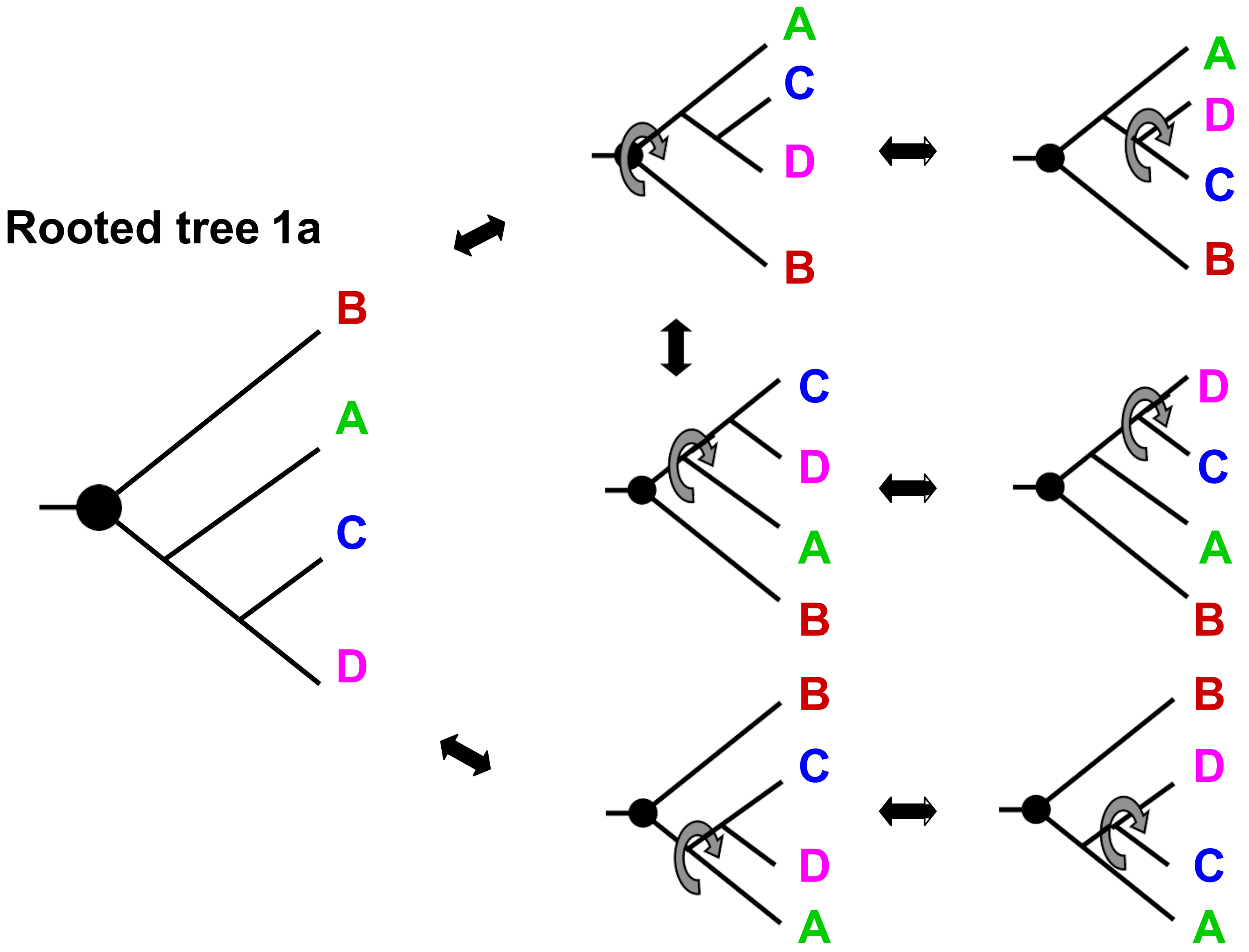


Polytomy or multifurcation (trifurchation)

A bifurcation

Slide credit: HK

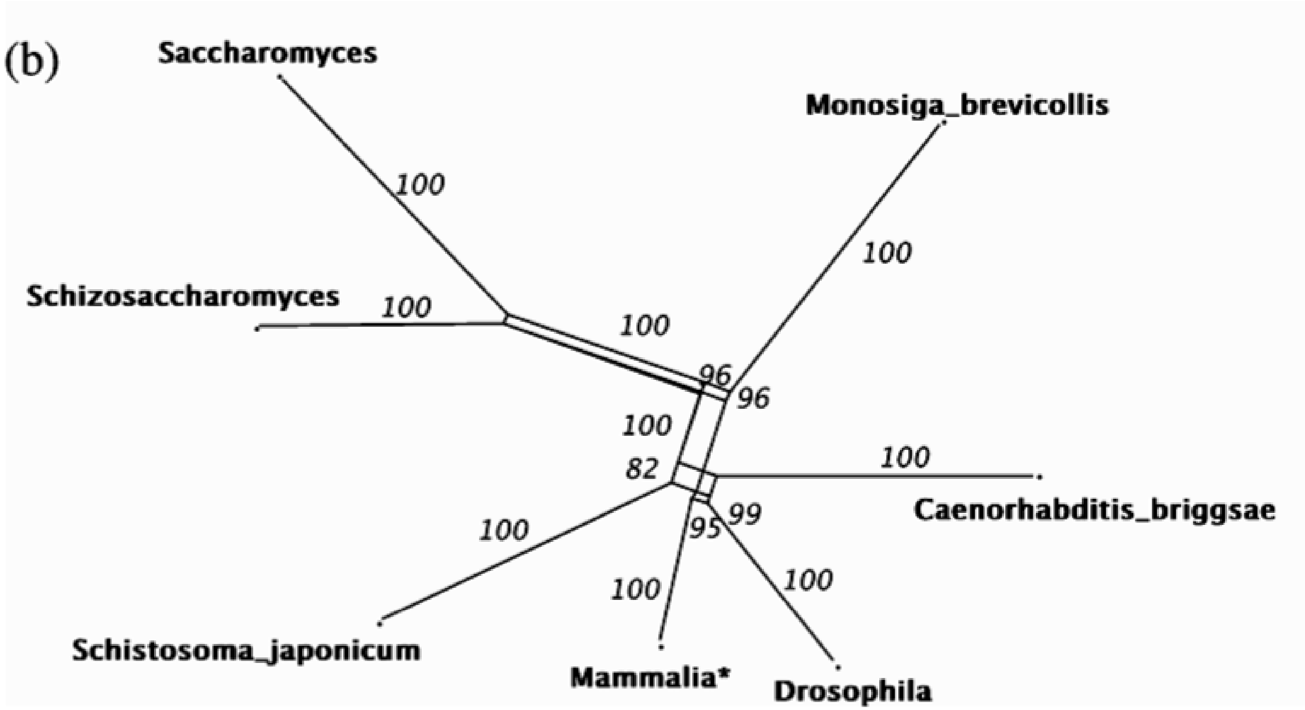
All of these rearrangements show the same evolutionary relationships between the taxa



Slide credit: HK

Limits of the tree representation

- Some events, such as hybridization, recombination, or combinations of lateral gene transfers, are poorly represented by trees.



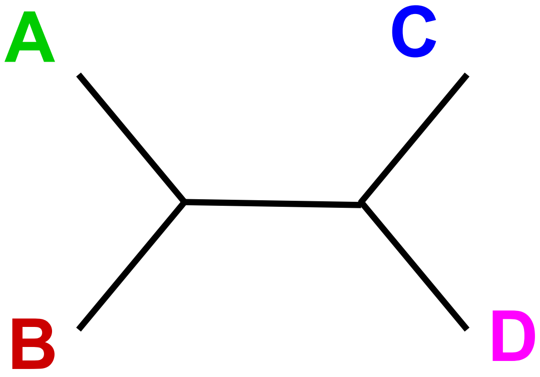
concatenated alignment of 146 genes, and ML distances under a JTT + F + Γ model.

Huson and Bryant, MBE 2006

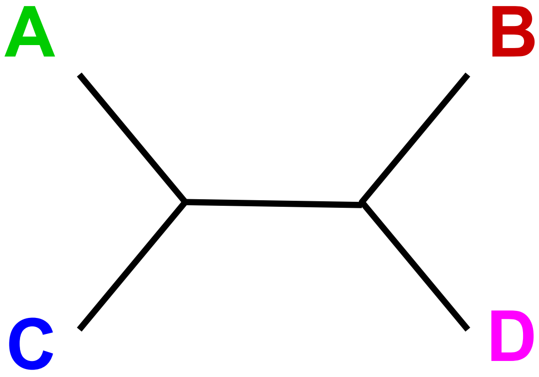
How to infer trees?

There are three possible unrooted trees for four taxa (A, B, C, D)

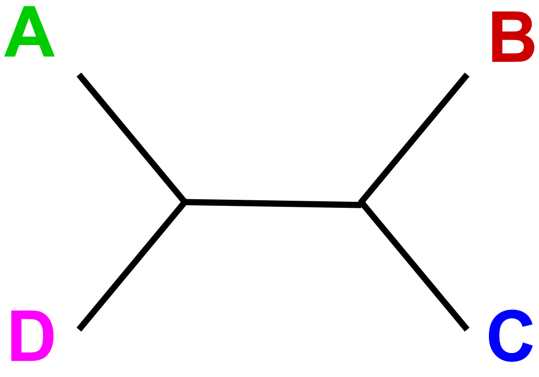
Tree 1



Tree 2

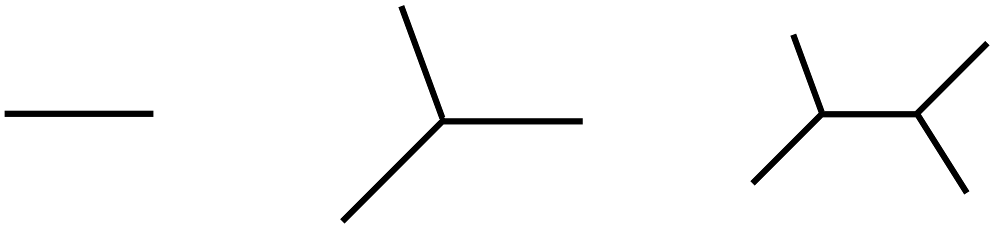


Tree 3



Slide credit: HK

How many branches are there in an unrooted bifurcating tree of *n* taxa?



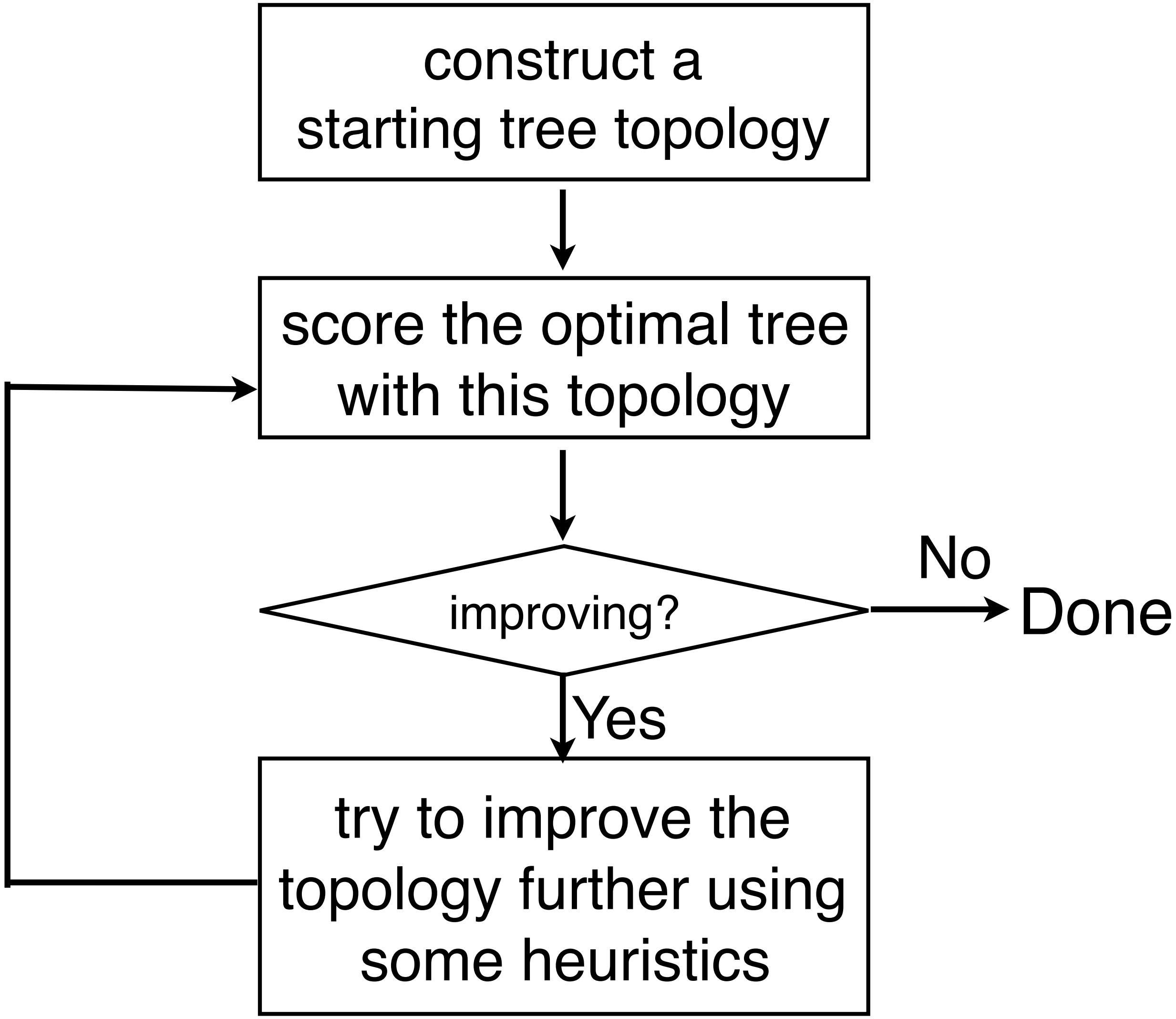
# taxa	2	3	4	n
# branches	1	3	5	2*n*-3

How many topologies?

Number of “taxa”	Number of binary trees Unrooted	Rooted
3	1	

Slide credit: RAG

General Scheme to Build Phylogenetic Trees


construct a starting tree topology
score the optimal tree with this topology
improving?
No
Done
Yes
try to improve the topology further using some heuristics

- UPGMA

- Maximum parsimony
- Least squares fit of distances
- Maximum likelihood

- Swap leaves
- Swap subtrees

Distance matrix

Recall that have learned how to estimate the evolutionary distance between pairs of sequences:

	Mouse	Rat	Human	Swine	Chimp
Mouse		0.12	0.38	0.28	0.38
Rat	0.12		0.32	0.45	0.52
Human	0.38	0.32		0.38	0.08
Swine	0.28	0.45	0.38		0.33
Chimp	0.38	0.52	0.08	0.33	

Slide credit: RAG

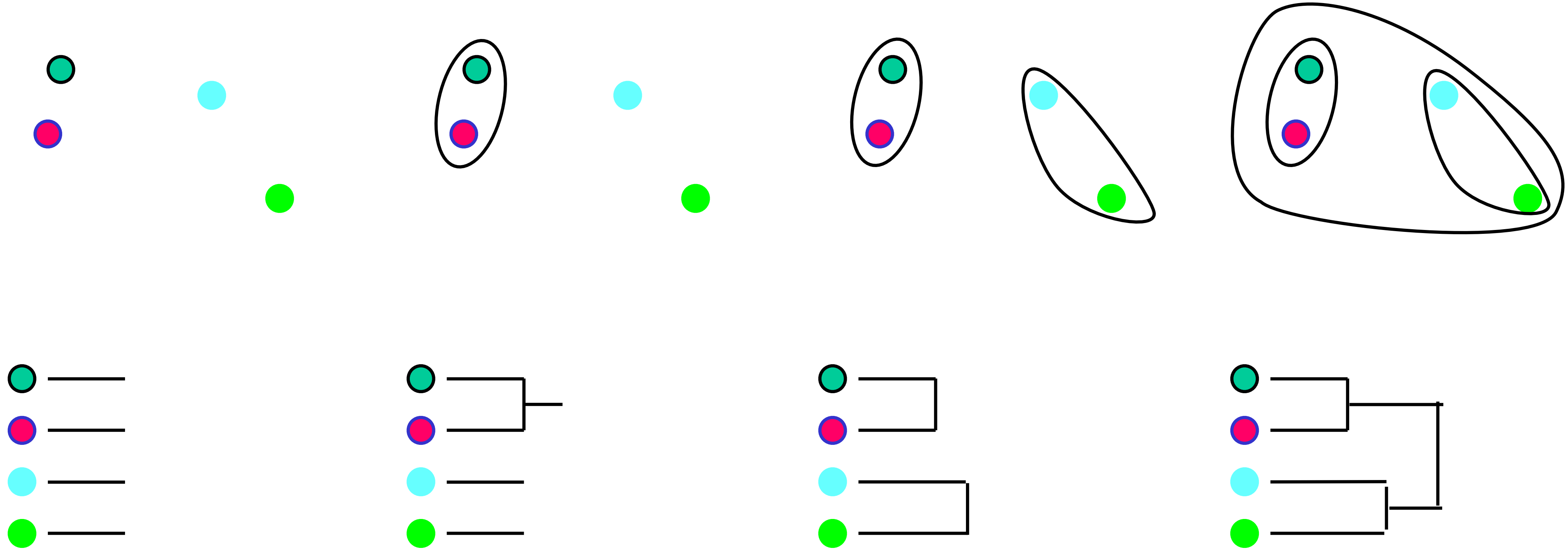
UPGMA

(Unweighted pair group method using arithmetic averages)

Recursively group the closest two remaining leaves

distance [mutations/site]

	●	🔴	🔵	🟢
●		0.1	0.3	0.4
🔴			0.3	0.3
🔵				0.2
🟢				

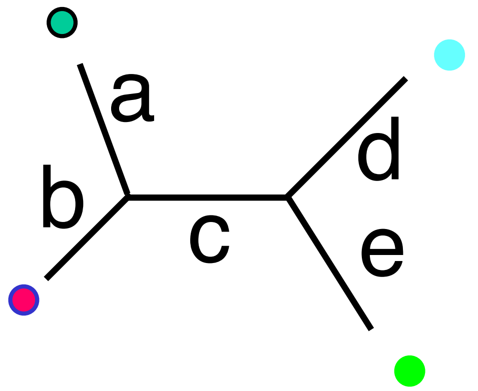


Slide credit: RAG

Distance Trees as Scoring Scheme

distance [mutations/site]

	●	🔴	🔵	🟢
●		0.1	0.3	0.4
🔴			0.3	0.3
🔵				0.2
🟢				



$a+b = 0.1$

$a+c+d = 0.3$

$a+c+e = 0.4$

$b+c+d = 0.3$

$b+c+e = 0.3$

$d+e = 0.2$

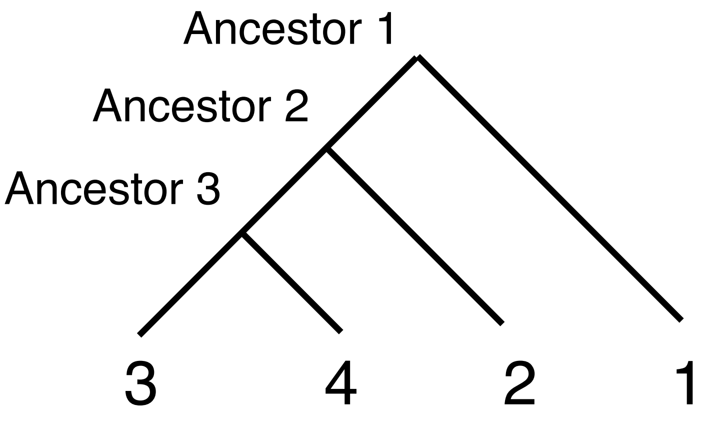
5 unknown
6 equations

→ overdetermined
→ minimise error

Maximum Parsimony

- Occam's Razor: the simplest explanation is the likeliest.
- Scoring scheme: minimum number of change in (discrete) characters.

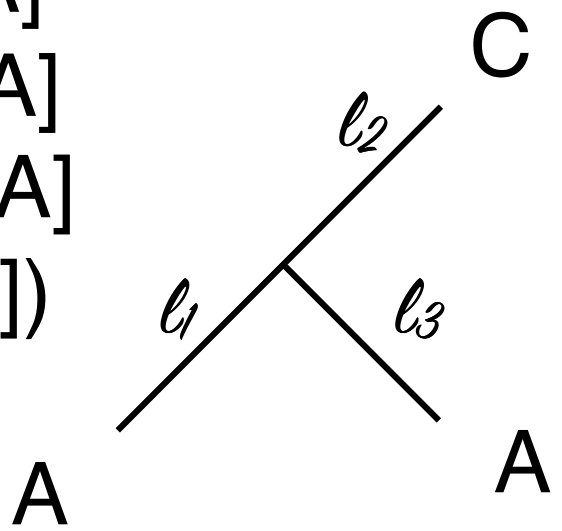
	Number of feet	Tail?	Favorite food	Can fly?
Species 1	0	Y	Carnivore	No
Species 2	4	Y	Herbivore	No
Species 3	2	Y	Herbivore	Yes
Species 4	6	Y	Herbivore	Yes
Ancestor 1	2	Y	Carnivore	No
Ancestor 2	2	Y	Herbivore	No
Ancestor 3	2	Y	Herbivore	Yes



Maximum Likelihood

The best tree is the one that maximizes the probability that the observed sequences would result **IF** the tree were correct.

$$
\begin{aligned}
\text{likelihood} = f(A)^{*}\ (&\\
&M^{l1}[A\text{->}A]^{*}M^{l2}[A\text{->}C]^{*}M^{l3}[A\text{->}A]\\
+\ &M^{l1}[A\text{->}C]^{*}M^{l2}[C\text{->}C]^{*}M^{l3}[C\text{->}A]\\
+\ &M^{l1}[A\text{->}G]^{*}M^{l2}[G\text{->}C]^{*}M^{l3}[G\text{->}A]\\
+\ &M^{l1}[A\text{->}T]^{*}M^{l2}[T\text{->}C]^{*}M^{l3}[T\text{->}A])
\end{aligned}
$$



Note: here we arbitrarily start from the bottom left taxon (which has an A) as the "root" of the tree, but the likelihood is the same no matter what starting point we choose. This is a property of our Markov model, which is called "reversible".

Slide credit: RAG

Discussion on methods

	👍	👎
Distance	- Fast & scalable - Statistically consistent (converges to true tree if the model is correct)	- Does not use all information available optimally
Parsimony	- Fast - Intuitive	- Statistically inconsistent - Lowly regarded in the phylogenetic community
Likelihood	- Statistically consistent - Statistically efficient - Highly regarded in the phylogenetic community	- Slow

Confidence

Measuring confidence with the Bootstrap

Our data is limited, representing (infinitesimally) small fraction of an ideal, infinite set
What is the uncertainty due to our limited data?

Scenario: take 100 same-size sets of data from infinite set

Calculate tree for each set

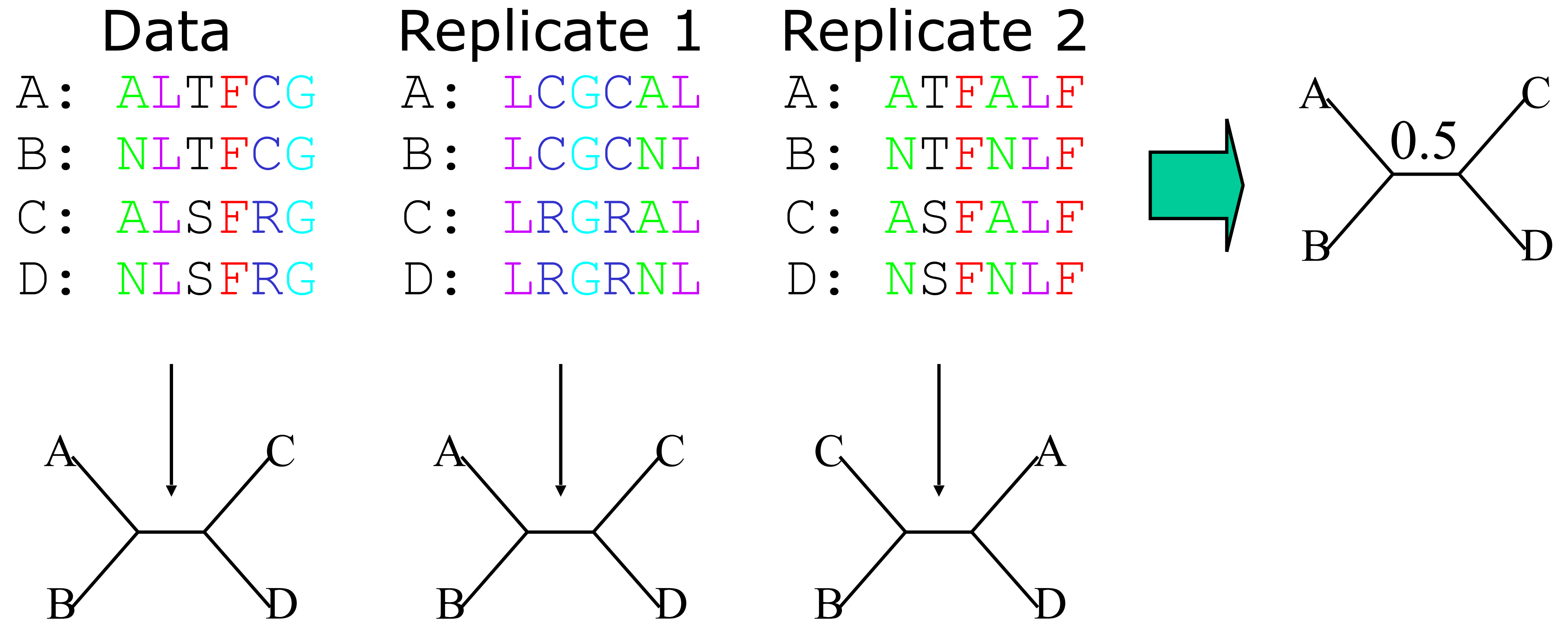
See what is consistent across trees

Bootstrap value: what fraction of all trees have a given node

Measuring confidence with the Bootstrap

But we don’t have access to infinite set!

Best we can do: use the set we have to represent the sets we don’t have: **Bootstrapping**

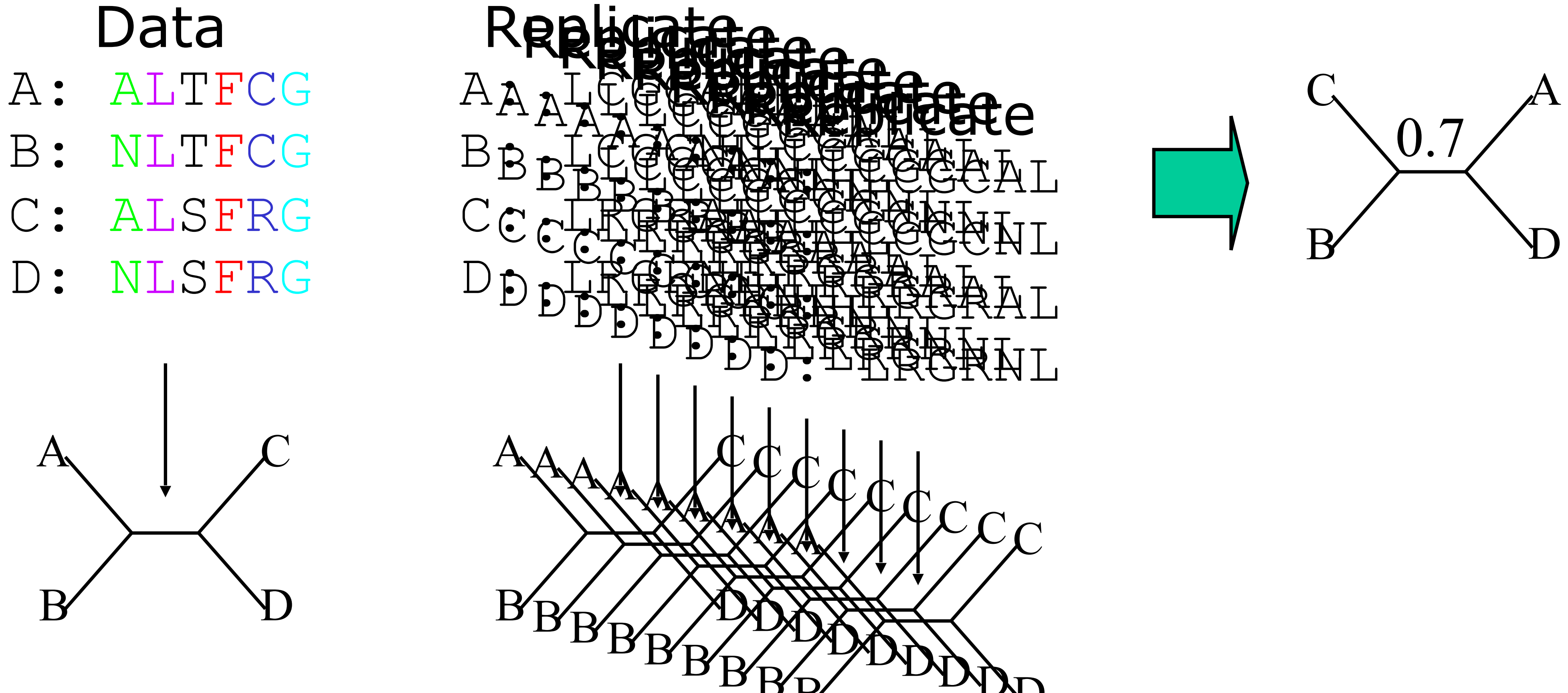


Data	Replicate 1	Replicate 2
A: ALTFCG	A: LCGCAL	A: ATFALF
B: NLTFCG	B: LCGCNL	B: NTFNLF
C: ALSFRG	C: LRGRAL	C: ASFALF
D: NLSFRG	D: LRGRNL	D: NSFNLF

Measuring confidence with the Bootstrap

But we don’t have access to infinite set!

Best we can do: use the set we have to represent the sets we don’t have: **Bootstrapping**



Data

A: ALTFCG
B: NLTFCG
C: ALSFRG
D: NLSFRG

Questions?

🌍 http://lab.dessimoz.org

@cdessimoz