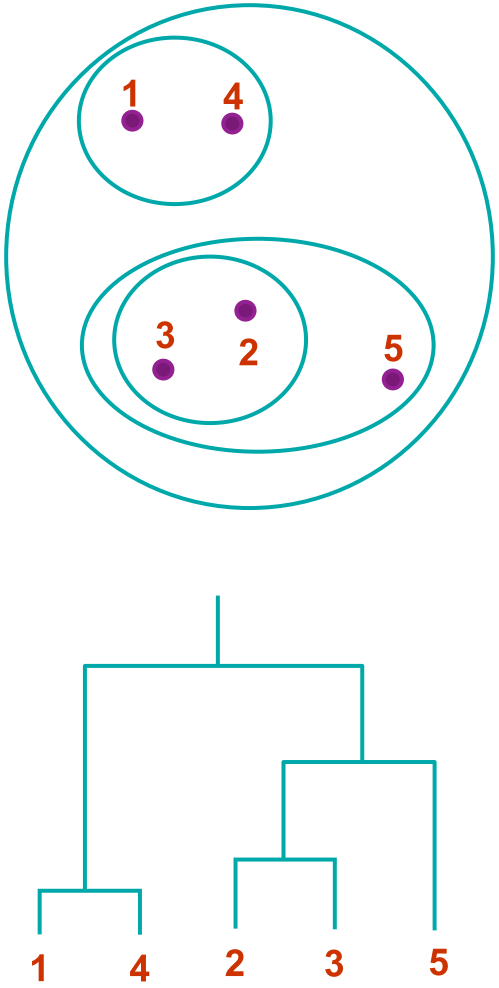
Molecular Evolution and Phylogenetic Tree Reconstruction



Orthology, Paralogy, Inparalogs, Outparalogs

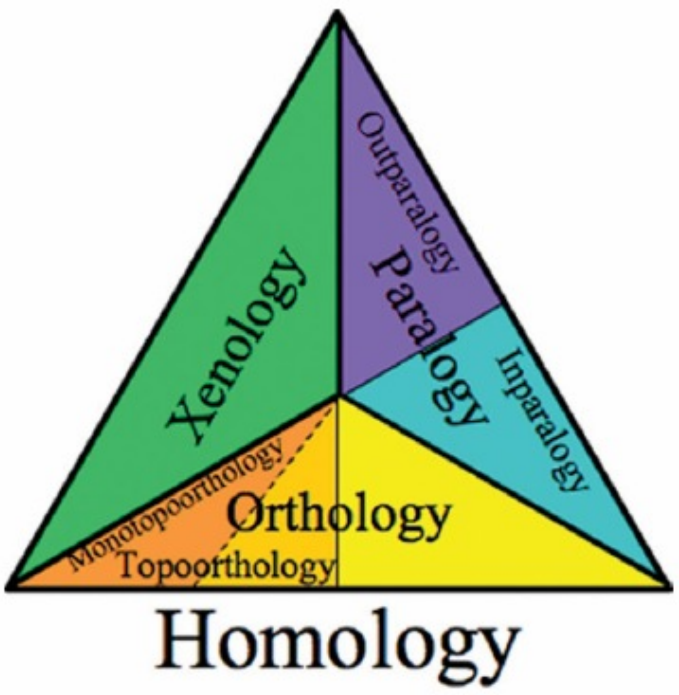
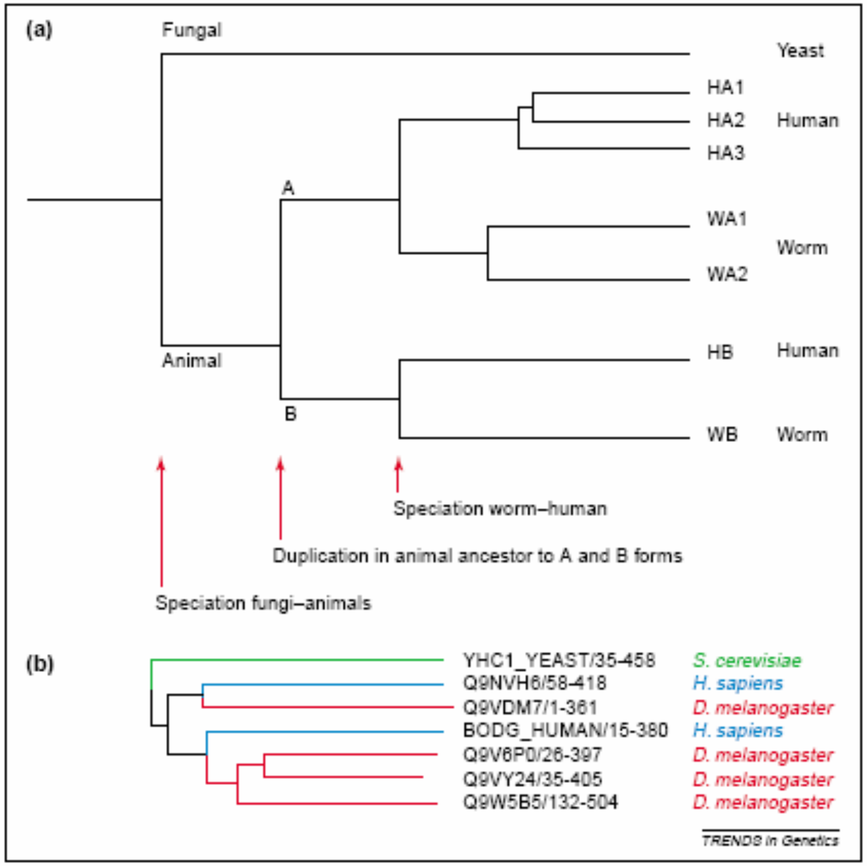


Figure 1. Refinements of homology.

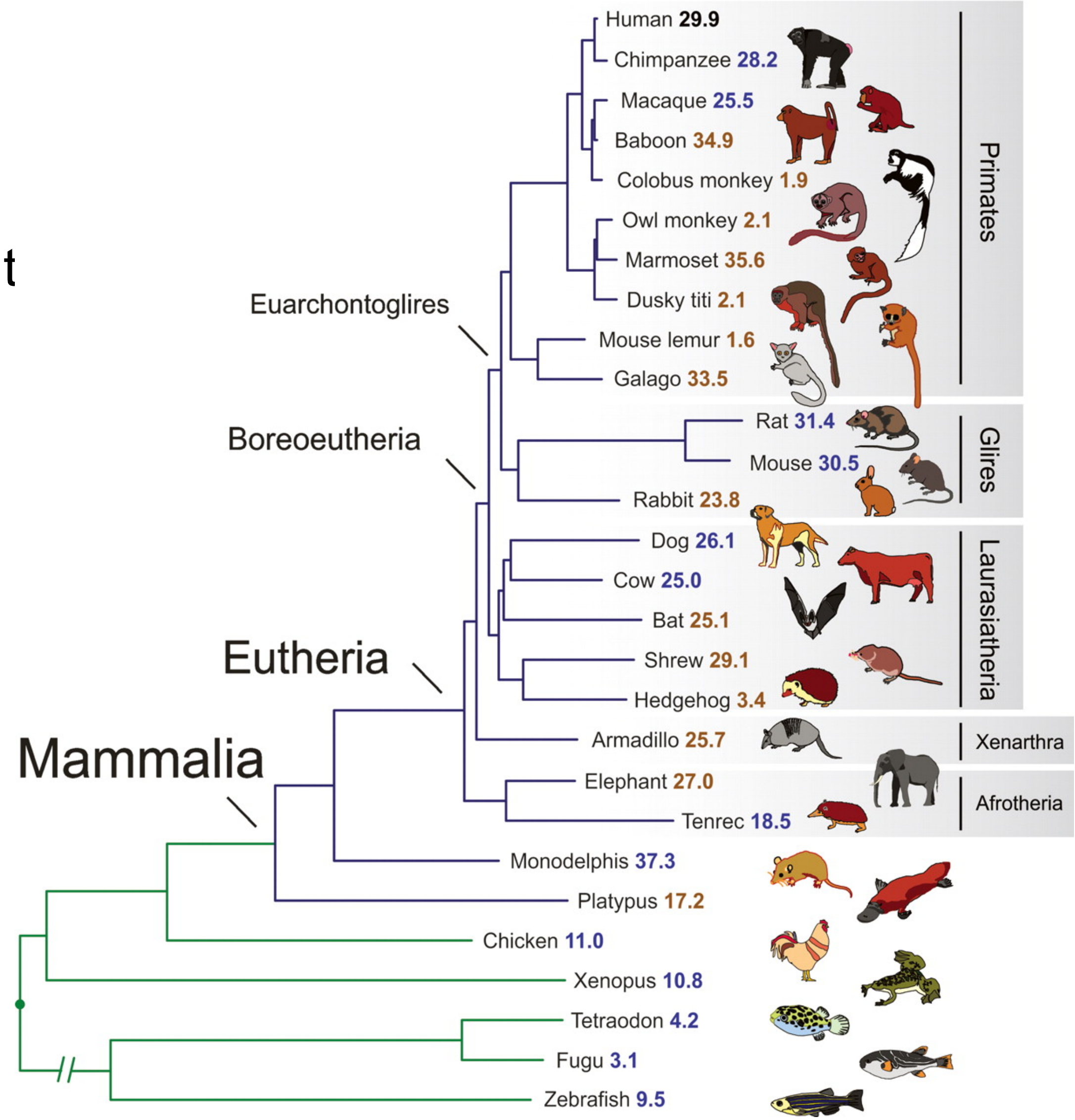


TRENDS in Genetics

Fig. 1. The definition of inparalogs and outparalogs. (a) Consider an ancient gene inherited in the yeast, worm and human lineages. The gene was duplicated early in the animal lineage, before the human–worm split, into genes A and B. After the human–worm split, the A form was in turn duplicated independently in the human and worm lineages. In this scenario, the yeast gene is orthologous to all worm and human genes, which are all co-orthologous to the yeast gene. When comparing the human and worm genes, all genes in the HA* set are co-orthologous to all genes in the WA* set. The genes HA* are hence 'inparalogs' to each other when comparing human to worm. By contrast, the genes HB and HA* are 'outparalogs' when comparing human with worm..However, HB and HA*, and WB and WA* are inparalogs when comparing with yeast, because the animal–yeast split pre-dates the HA*–HB duplication. (b) Real-life example of inparalogs: γ-butyrobetaine hydroxylases. The points of speciation and duplication are easily identifiable. The alignment is a subset of Pfam:PF03322 and the tree was generated by neighbor-joining in Belvu. All nodes have a bootstrap support exceeding 95%.

Phylogenetic Trees

- Nodes: species
- Edges: time of independent evolution
- Edge length represents evolution time
    - AKA genetic distance
    - Not necessarily chronological time

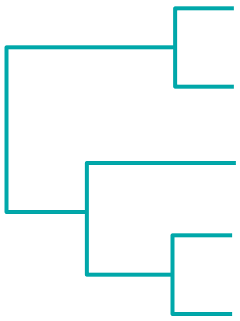


Inferring Phylogenetic Trees

Trees can be inferred by several criteria:

- Morphology of the organisms
  - *Can lead to mistakes*
- Sequence comparison

Example:



Mouse: ACAGTGACGCCCCAAACGT
Rat: ACAGTGACGCTACAAACGT
Baboon: CCTGTGACGTAACAAACGA
Chimp: CCTGTGACGTAGCAAACGA
Human: CCTGTGACGTAGCAAACGA

Distance Between Two Sequences

Basic principle:

- Distance proportional to degree of independent sequence evolution

Given sequences x^i, x^j,

d_{ij} = distance between the two sequences

One possible definition:

d_{ij} = fraction f of sites u where $x^i[u] \neq x^j[u]$

Better scores are derived by modeling evolution as a continuous change process

Molecular Evolution

Modeling sequence substitution:

Consider what happens at a position for time Δt,

- $P(t)$ = vector of probabilities of {A,C,G,T} at time t
- μ_{AC} = rate of transition from A to C per unit time
- $\mu_A = \mu_{AC} + \mu_{AG} + \mu_{AT}$ rate of transition out of A
- $p_A(t+\Delta t) = p_A(t) – p_A(t)\ \mu_A\ \Delta t + p_C(t)\ \mu_{CA}\ \Delta t + p_G(t)\ \mu_{GA}\ \Delta t + p_T(t)\ \mu_{TA}\ \Delta t$

Molecular Evolution

In matrix/vector notation, we get

$$P(t+\Delta t) = P(t) + Q\,P(t)\,\Delta t$$

where Q is the substitution rate matrix

$$Q = \begin{pmatrix} -\mu_A & \mu_{GA} & \mu_{CA} & \mu_{TA} \\ \mu_{AG} & -\mu_G & \mu_{CG} & \mu_{TG} \\ \mu_{AC} & \mu_{GC} & -\mu_C & \mu_{TC} \\ \mu_{AT} & \mu_{GT} & \mu_{CT} & -\mu_T \end{pmatrix}$$

Molecular Evolution

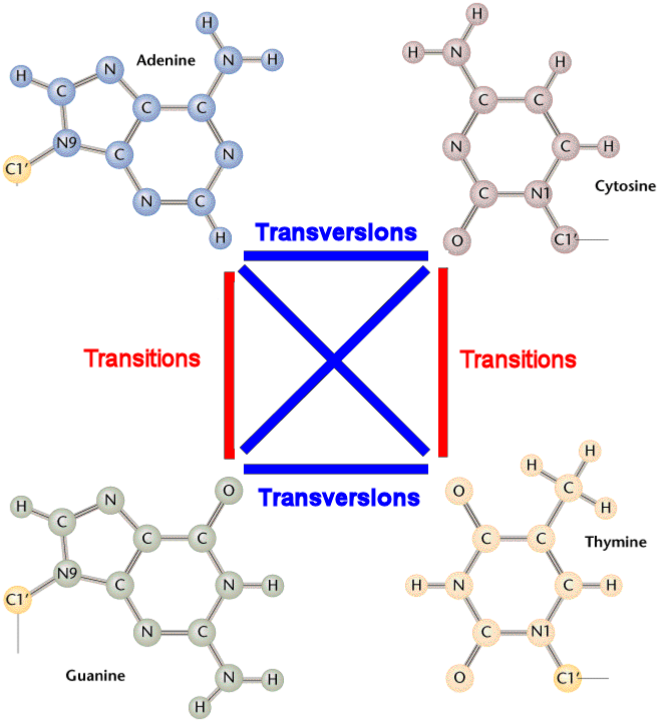
- This is a differential equation:

$$P'(t) = Q\,P(t)$$

- Q => prob. distribution over {A,C,G,T} at each position, stationary (equilibrium) frequencies π_A, π_C, π_G, π_T

- Each Q is an evolutionary model
  - Some work better than others

Evolutionary Models



- Jukes-Cantor $$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$
- Kimura $$Q = \begin{pmatrix} * & \kappa & 1 & 1 \\ \kappa & * & 1 & 1 \\ 1 & 1 & * & \kappa \\ 1 & 1 & \kappa & * \end{pmatrix}$$
- Felsenstein $$Q = \begin{pmatrix} * & \pi_T & \pi_T & \pi_T \\ \pi_C & * & \pi_C & \pi_C \\ \pi_A & \pi_A & * & \pi_A \\ \pi_G & \pi_G & \pi_G & * \end{pmatrix}$$
- HKY $$Q = \begin{pmatrix} * & \kappa\pi_T & \pi_T & \pi_T \\ \kappa\pi_C & * & \pi_C & \pi_C \\ \pi_A & \pi_A & * & \kappa\pi_A \\ \pi_G & \pi_G & \kappa\pi_G & * \end{pmatrix}$$

Estimating Distances

- Solve the differential equation and compute expected evolutionary time given sequences

$$P'(t) = Q\,P(t)$$

Jukes-Cantor:

Let $P_{AA}(t) = P_{CC}(t) = P_{CC}(t) = P_{CC}(t) = r$

$P_{AC}(t) = \ldots = P_{TG}(t) = s$

Then,

$$r'(t) = -\tfrac{3}{4}\,r(t)\,\mu + \tfrac{3}{4}\,s(t)\,\mu$$

$$s'(t) = -\tfrac{1}{4}\,s(t)\,\mu + \tfrac{1}{4}\,r(t)\,\mu$$

Which is satisfied by

$$r(t) = \tfrac{1}{4}\,(1 + 3e^{-\mu t})$$

$$s(t) = \tfrac{1}{4}\,(1 - e^{-\mu t})$$

Estimating Distances

- Solve the differential equation and compute expected evolutionary time given sequences

$$P'(t) = Q\,P(t)$$

Jukes-Cantor:

$$P = \begin{pmatrix} \frac{1}{4}+\frac{3}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} \\ \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}+\frac{3}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} \\ \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}+\frac{3}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} \\ \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}+\frac{3}{4}e^{-t\mu} \end{pmatrix}$$

Estimating Distances

Let p = probability a base is different between two sequences,

Solve to find **t**

- Jukes-Cantor $r(t) = 1 – p = ¼ (1 + 3e^{-\mu t})$

$p = ¾ – ¾ e^{-\mu t}$

$¾ – p = ¾ e^{-\mu t}$

$1 – 4p/3 = e^{-\mu t}$

Therefore,

$\mu t = - \ln(1 – 4p/3)$

Letting $d = ¾ \mu t$, denoting substitutions per site,

$$d = -\frac{3}{4}\ln(1-\frac{4}{3}p)$$

Estimating Distances

d: Branch length in terms of substitutions per site

- Jukes-Cantor

$$d = -\frac{3}{4}\ln(1 - \frac{4}{3}p)$$

- Kimura

$$d = -\frac{1}{2}\ln(1 - 2P - Q) - \frac{1}{4}\ln(1 - 2Q)$$

Simple method for building tree: UPGMA

UPGMA (unweighted pair group method using arithmetic averages)
Or the **Average Linkage Method**

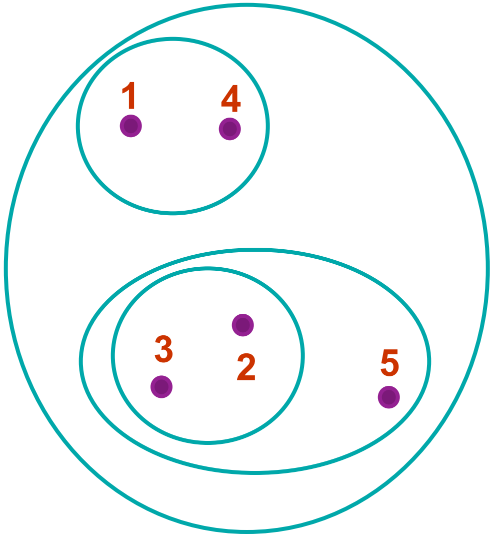
Given two disjoint clusters C_i, C_j of sequences,

$$d_{ij} = \frac{1}{|C_i| \times |C_j|} \Sigma_{\{p \in Ci,\ q \in Cj\}} d_{pq}$$

Claim that if $C_k = C_i \cup C_j$, then distance to another cluster C_l is:

$$d_{kl} = \frac{d_{il}\ |C_i| + d_{jl}\ |C_j|}{|C_i| + |C_j|}$$

Algorithm: Average Linkage



Initialization:

Assign each x_i into its own cluster C_i

Define one leaf per sequence, height 0

Iteration:

Find two clusters C_i, C_j s.t. d_{ij} is min

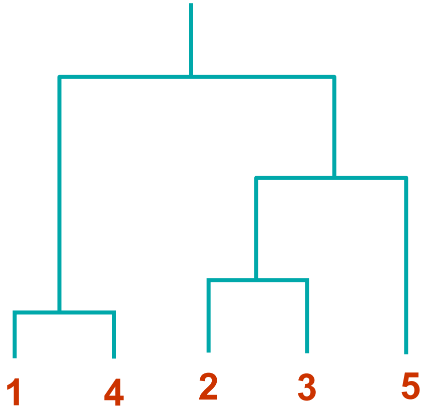
Let $C_k = C_i \cup C_j$

Define node connecting C_i, C_j, and place it at height $d_{ij}/2$

Delete C_i, C_j

Termination:

When two clusters i, j remain, place root at height $d_{ij}/2$



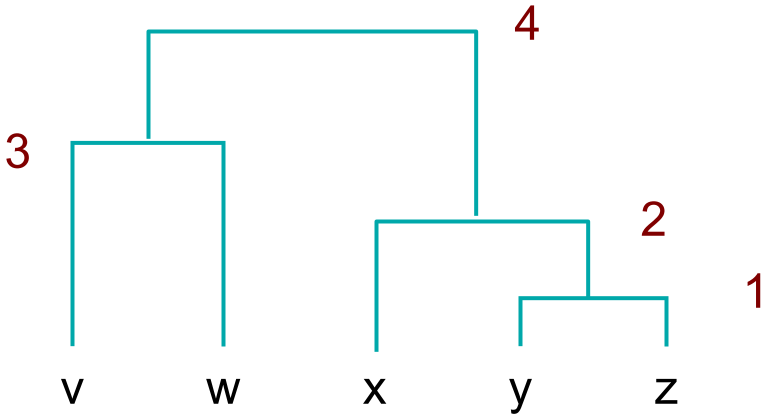
Average Linkage Example

	v	w	x	y	z
v	0	6	8	8	8
w		0	8	8	8
x			0	4	4
y				0	2
z					0

	v	w	xyz
v	0	6	8
w		0	8
xyz			0

	vw	xyz
vw	0	8
xyz		0

	v	w	x	yz
v	0	6	8	8
w		0	8	8
x			0	4
yz				0



Ultrametric Distances and Molecular Clock

Definition:

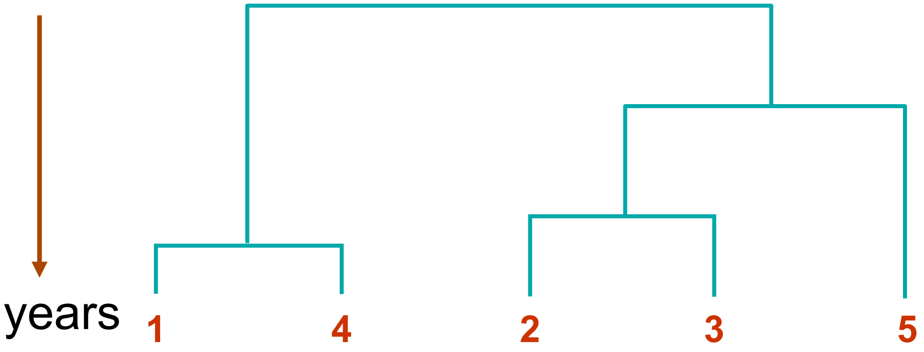
A distance function d(.,.) is ultrametric if for any three distances $d_{ij} \leq d_{ik} \leq d_{ij}$, it is true that

$$d_{ij} \leq d_{ik} = d_{jk}$$

The Molecular Clock:

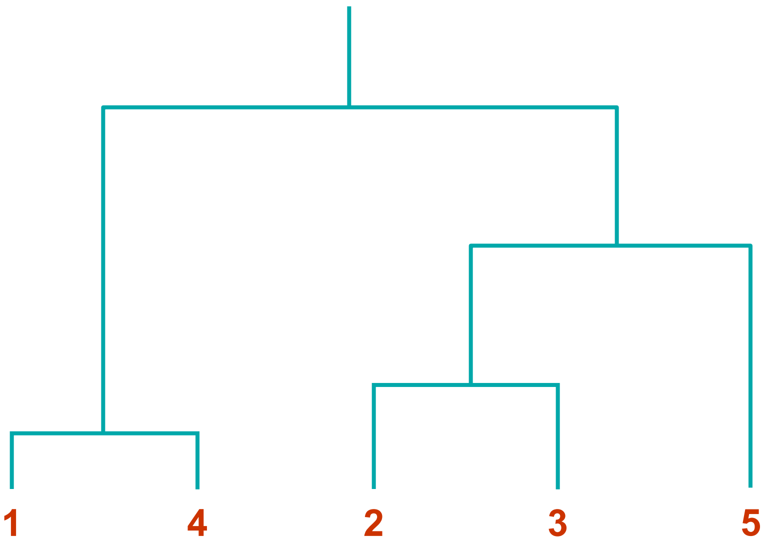
The evolutionary distance between species x and y is 2× the Earth time to reach the nearest common ancestor

That is, the molecular clock has constant rate in all species



The molecular clock results in ultrametric distances

Ultrametric Distances & Average Linkage



Average Linkage is guaranteed to reconstruct correctly a binary tree with ultrametric distances

Proof: Exercise

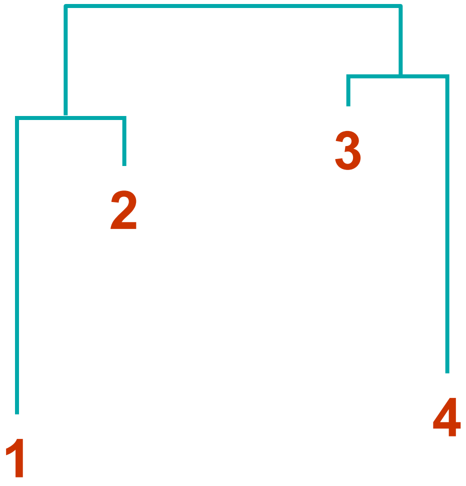
Weakness of Average Linkage

Molecular clock: all species evolve at the same rate (Earth time)

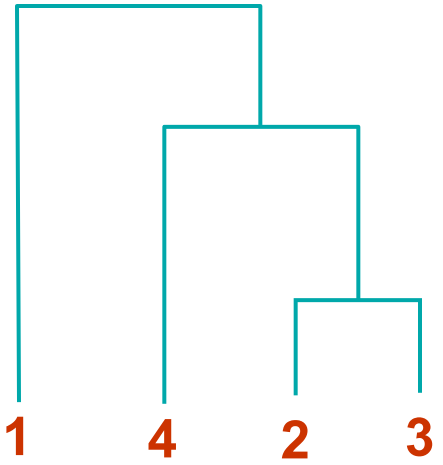
However, certain species (e.g., mouse, rat) evolve much faster

Example where UPGMA messes up:

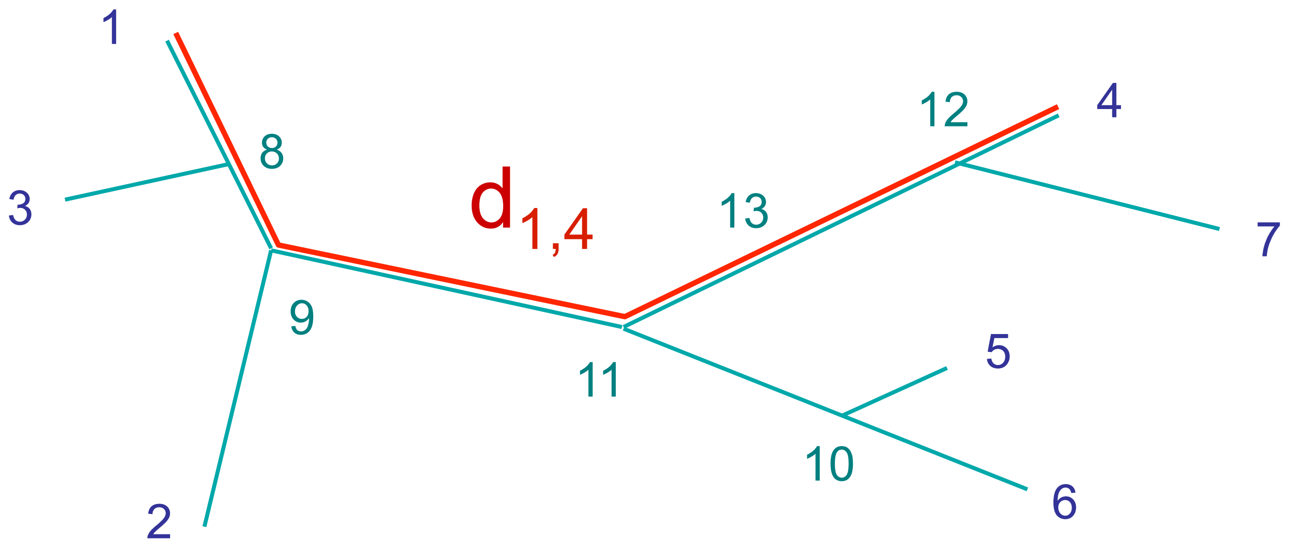
Correct tree



AL tree



Additive Distances

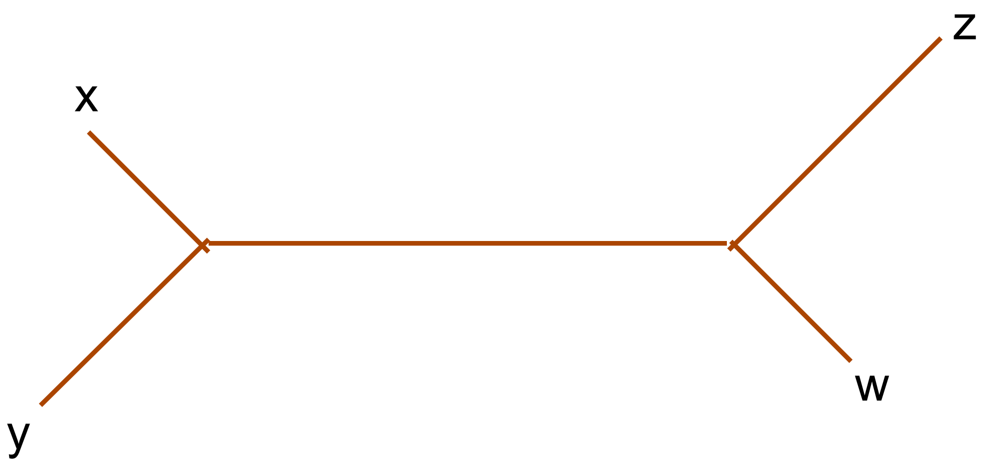


Given a tree, a distance measure is **additive** if the distance between any pair of leaves is the sum of lengths of edges connecting them

Given a tree T & additive distances d_{ij}, can uniquely reconstruct edge lengths:

- Find two neighboring leaves i, j, with common parent k
- Place parent node k at distance $d_{km} = ½ (d_{im} + d_{jm} – d_{ij})$ from any node m ≠ i, j

Additive Distances



For any four leaves x, y, z, w, consider the three sums

$$
\begin{aligned}
&d(x, y) + d(z, w)\\
&d(x, z) + d(y, w)\\
&d(x, w) + d(y, z)
\end{aligned}
$$

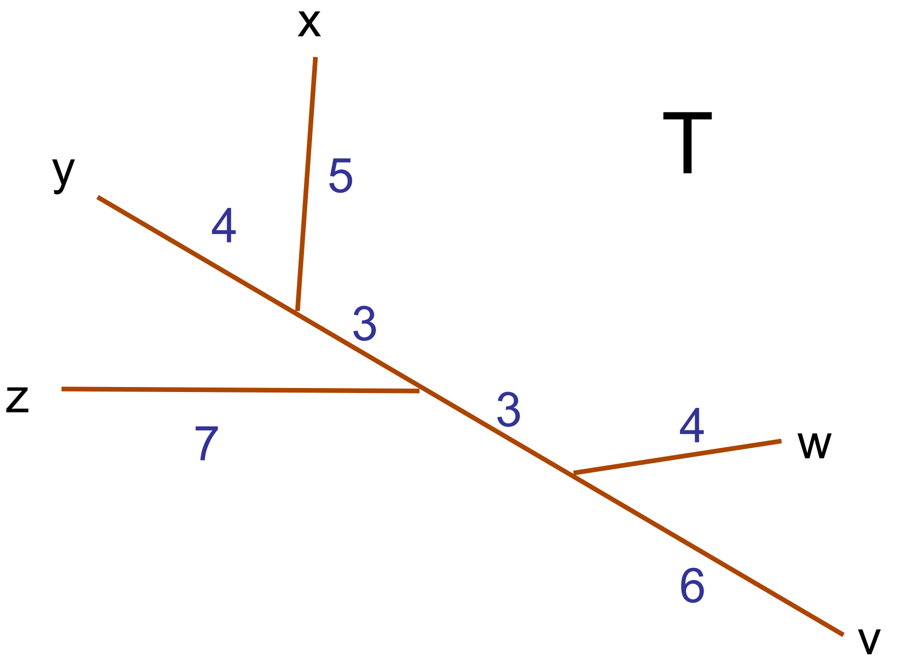
One of them is smaller than the other two, which are equal

$$d(x, y) + d(z, w) < d(x, z) + d(y, w) = d(x, w) + d(y, z)$$

Reconstructing Additive Distances Given T

D

	v	w	x	y	z
v	0	10	17	16	16
w		0	15	14	14
x			0	9	15
y				0	14
z					0

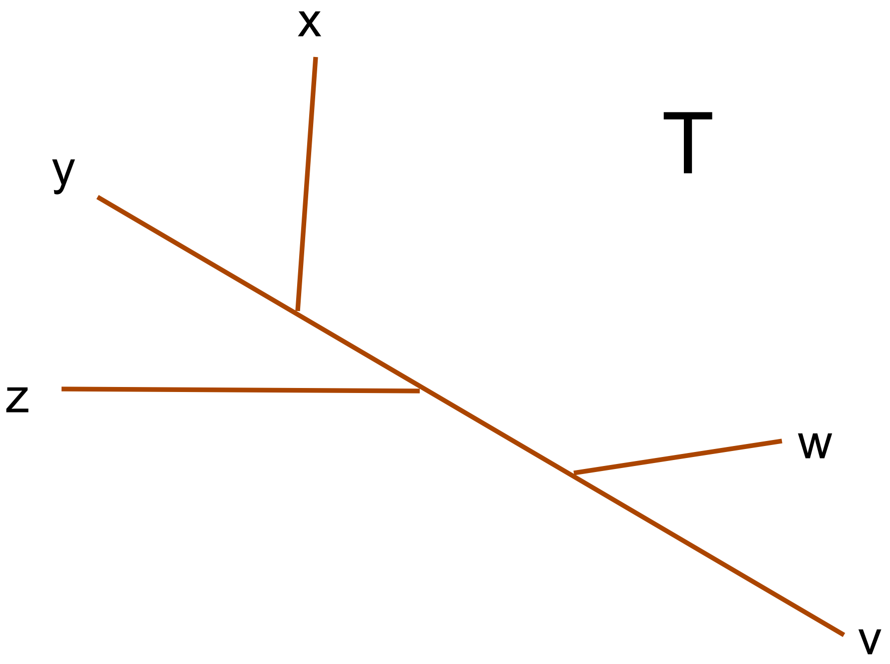


If we know T and D, but do not know the length of each leaf, we can reconstruct those lengths

Reconstructing Additive Distances Given T

D

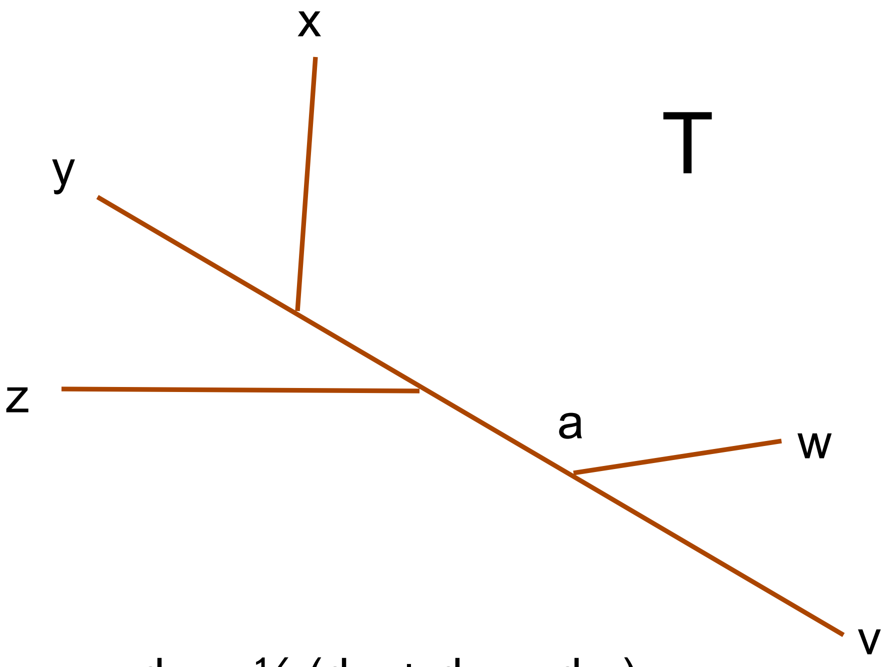
	v	w	x	y	z
v	0	10	17	16	16
w		0	15	14	14
x			0	9	15
y				0	14
z					0



Reconstructing Additive Distances Given T

D

	v	w	x	y	z
v	0	10	17	16	16
w		0	15	14	14
x			0	9	15
y				0	14
z					0



D_1

	a	x	y	z
a	0	11	10	10
x		0	9	15
y			0	14
z				0

$$d_{ax} = ½\,(d_{vx} + d_{wx} – d_{vw})$$

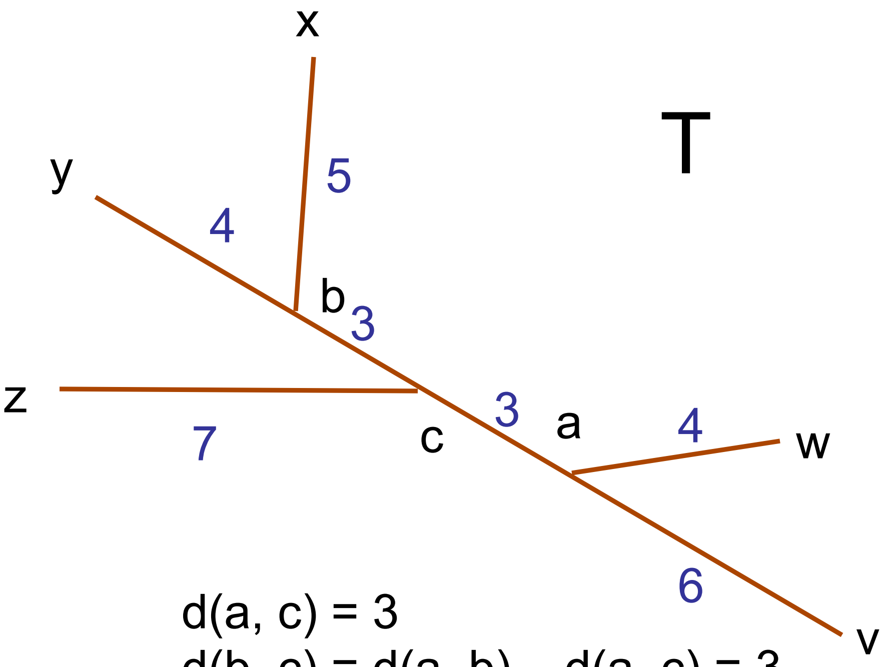
$$d_{ay} = ½\,(d_{vy} + d_{wy} – d_{vw})$$

$$d_{az} = ½\,(d_{vz} + d_{wz} – d_{vw})$$

Reconstructing Additive Distances Given T

D_1

	a	x	y	z
a	0	11	10	10
x		0	9	15
y			0	14
z				0



D_2

	a	b	z
a	0	6	10
b		0	10
z			0

D_3

	a	c
a	0	3
c		0

d(a, c) = 3
d(b, c) = d(a, b) – d(a, c) = 3
d(c, z) = d(a, z) – d(a, c) = 7
d(b, x) = d(a, x) – d(a, b) = 5
d(b, y) = d(a, y) – d(a, b) = 4
d(a, w) = d(z, w) – d(a, z) = 4
d(a, v) = d(z, v) – d(a, z) = 6

Correct!!!

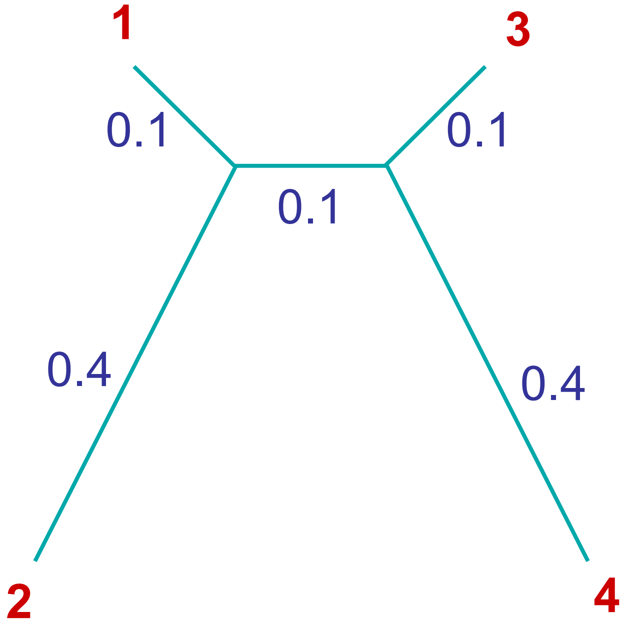
Neighbor-Joining

- Guaranteed to produce the correct tree if distance is additive
- May produce a good tree even when distance is not additive

Step 1: Finding neighboring leaves

Define

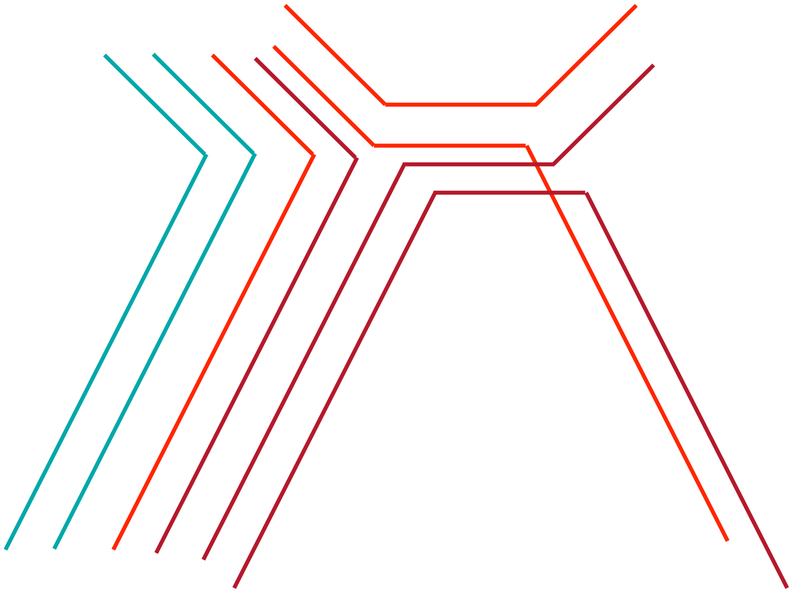
$D_{ij} = (N - 2)\, d_{ij} - \sum_{k \neq i} d_{ik} - \sum_{k \neq j} d_{jk}$

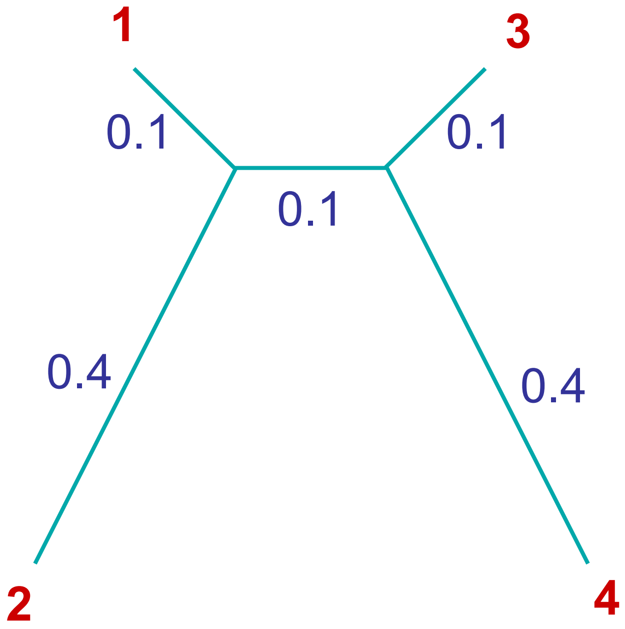


Claim: The above "magic trick" ensures that i, j are neighbors if D_{ij} is minimal

Neighbor-Joining

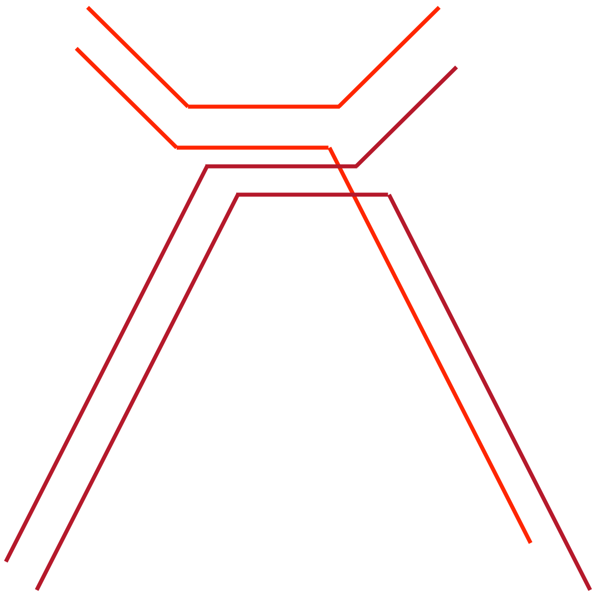
$$D_{ij} = (N - 2)\, d_{ij} - \sum_{k \neq i} d_{ik} - \sum_{k \neq j} d_{jk}$$

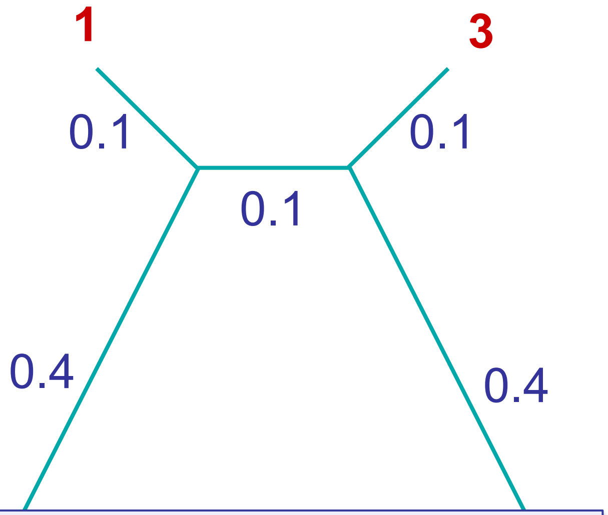




Neighbor-Joining

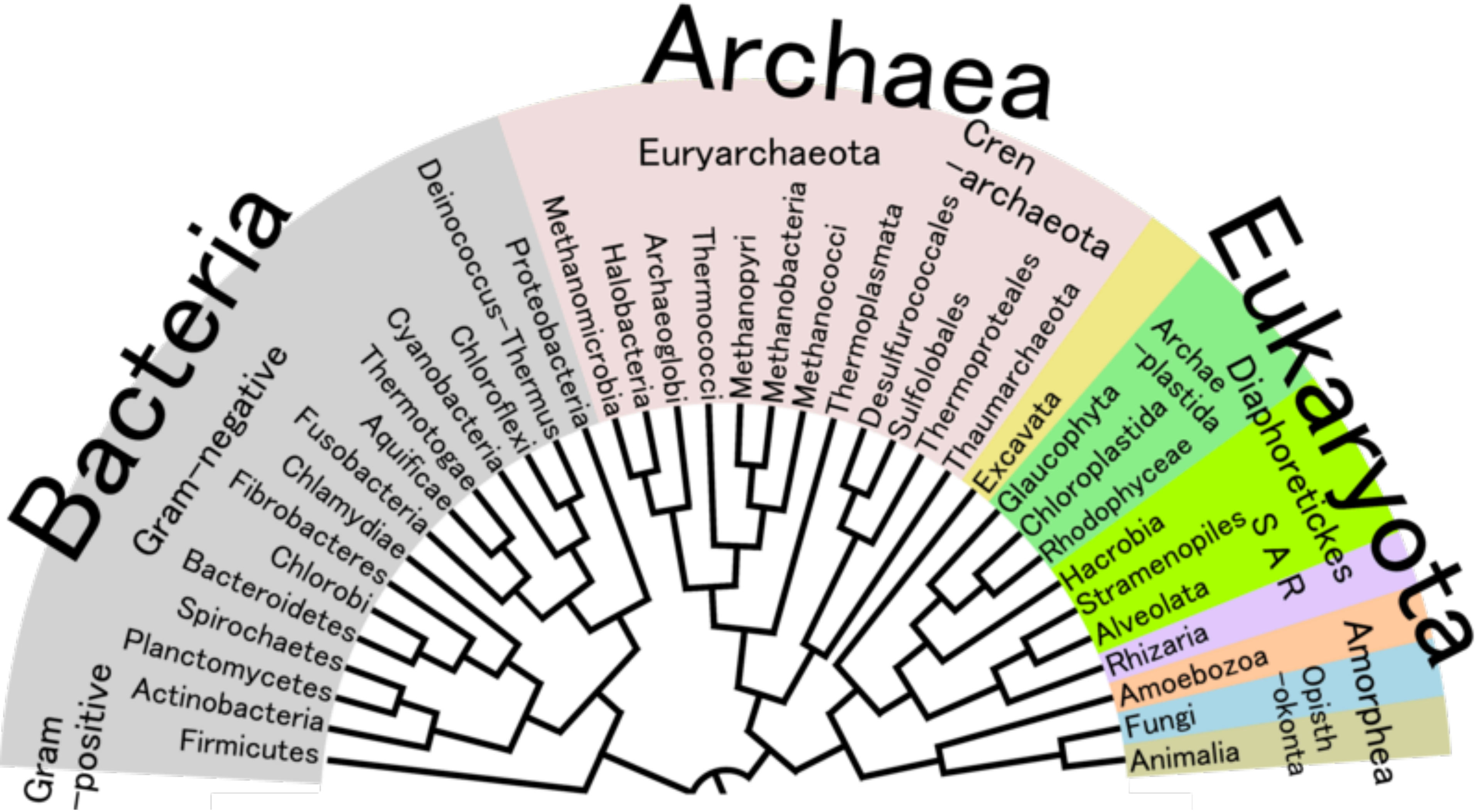
$D_{ij} = (N – 2)\, d_{ij} – \sum_{k\neq i} d_{ik} – \sum_{k\neq j} d_{jk}$



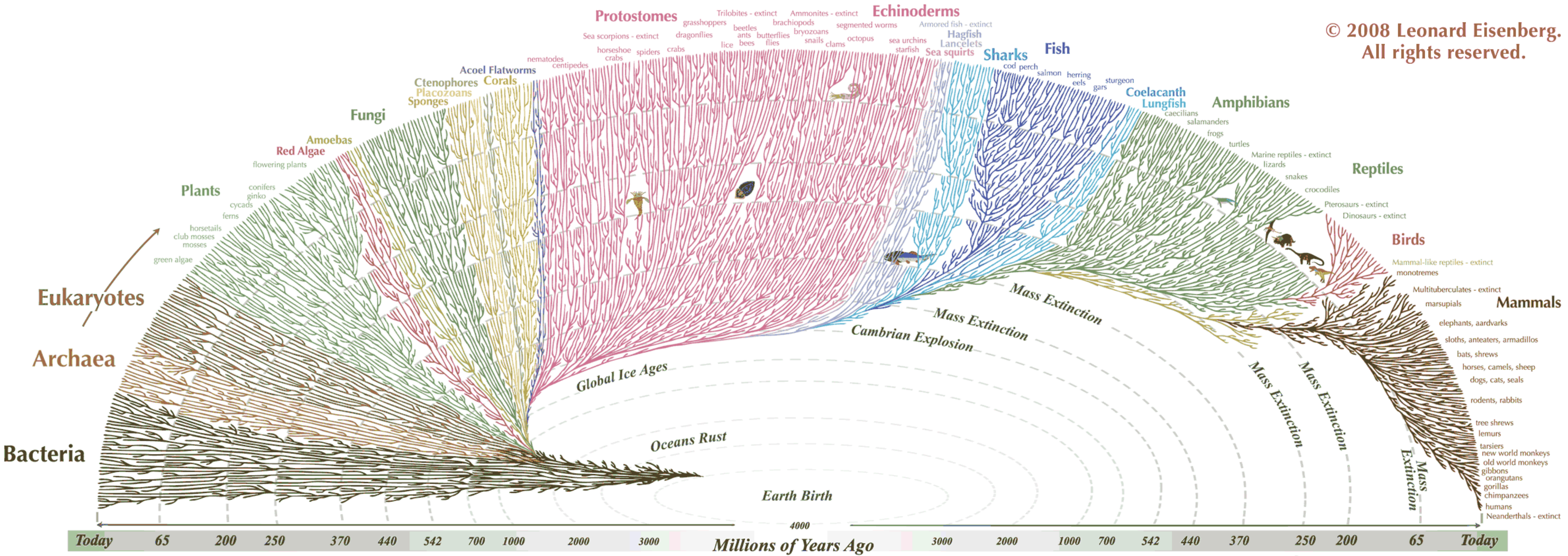


- All leaf edges appear negatively exactly twice
- All other edges appear negatively once for every path from each of the two leaves i, j, to leaves k ≠ i, j

Some Trees

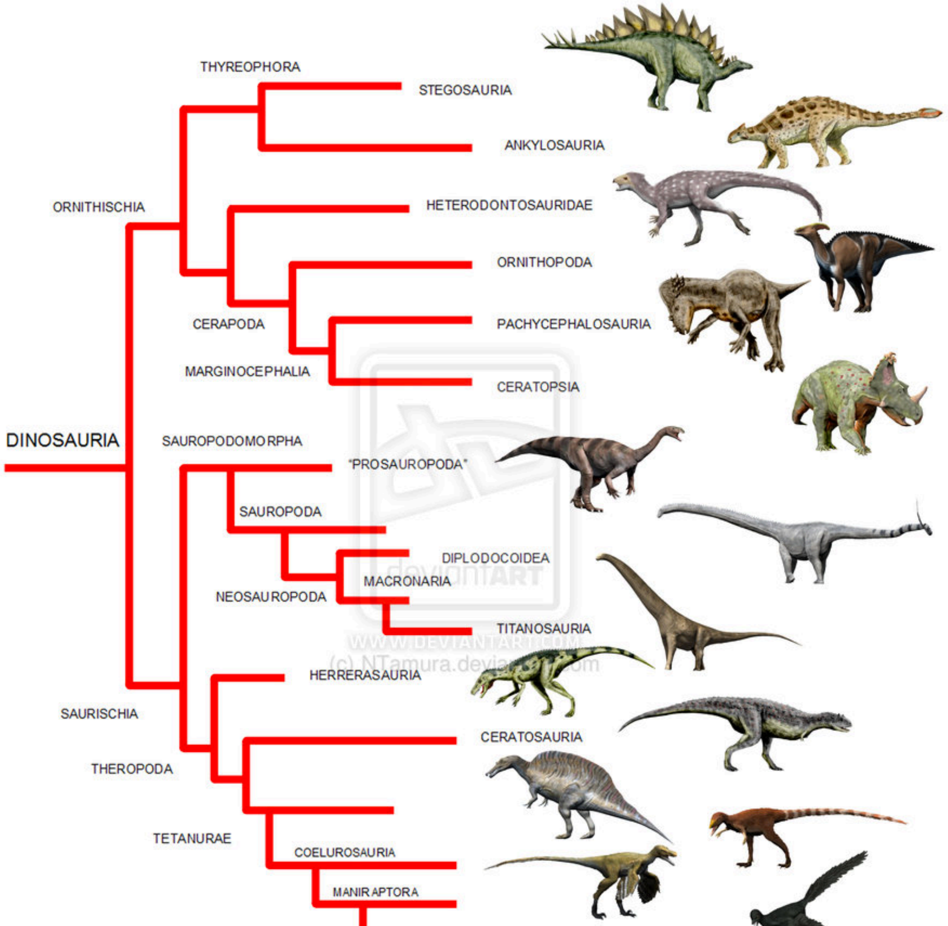

Bacteria
Gram-positive
Firmicutes
Actinobacteria
Planctomycetes
Spirochaetes
Bacteroidetes
Chlorobi
Fibrobacteres
Chlamydiae
Fusobacteria
Gram-negative
Aquificae
Thermotogae
Cyanobacteria
Chloroflexi
Deinococcus-Thermus
Proteobacteria
Archaea
Euryarchaeota
Methanomicrobia
Halobacteria
Archaeoglobi
Thermococci
Methanopyri
Methanobacteria
Methanococci
Thermoplasmata
Desulfurococcales
Sulfolobales
Thermoproteales
Cren-archaeota
Thaumarchaeota
Eukaryota
Excavata
Glaucophyta
Chloroplastida
Rhodophyceae
Archae-plastida
Diaphoretickes
Hacrobia
Stramenopiles
Alveolata
S A R
Rhizaria
Amoebozoa
Fungi
Animalia
Opisth-okonta
Amorphea

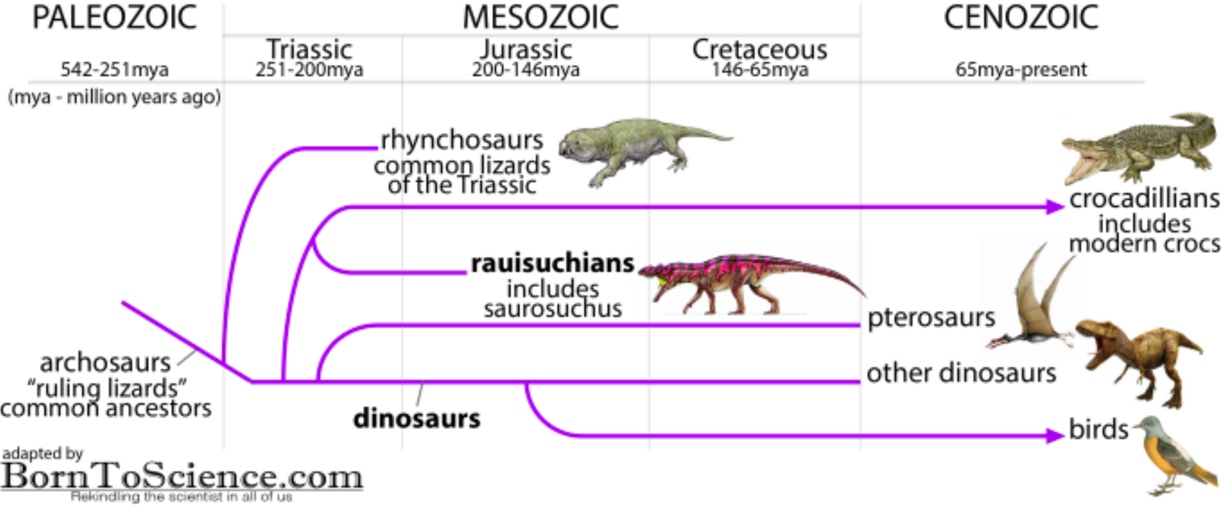
Some Trees



All the major and many of the minor living branches of life are shown on this diagram, but only a few of those that have gone extinct are shown. Example: Dinosaurs - extinct

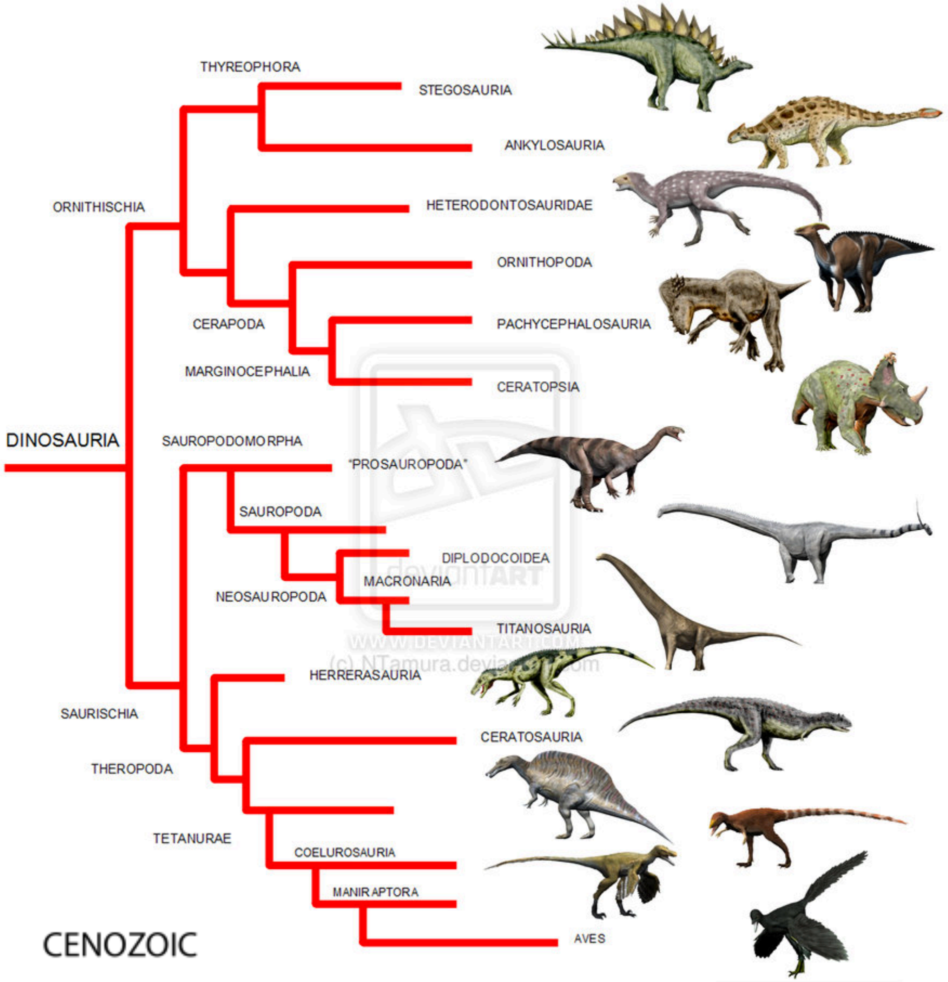
Some Trees

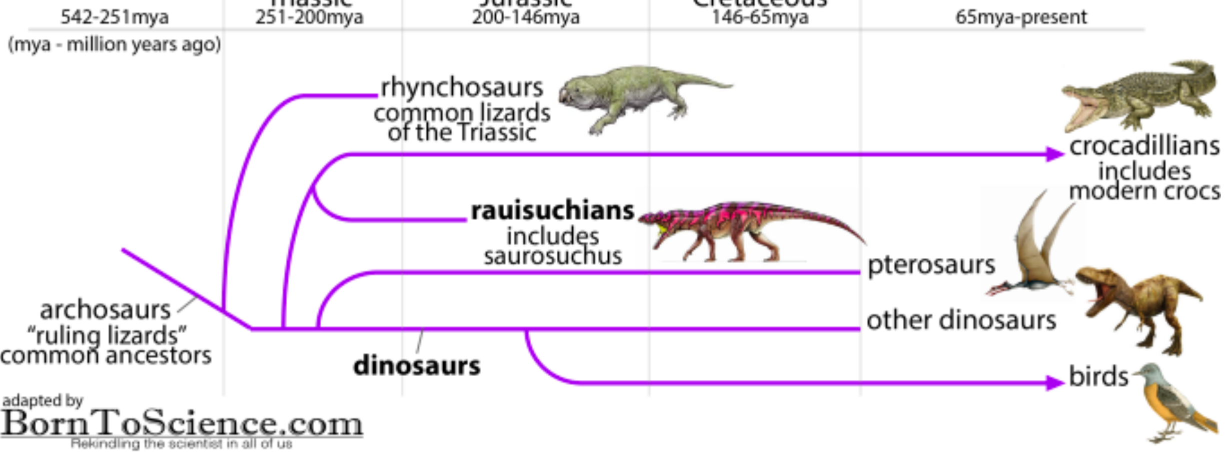




adapted by
BornToScience.com
Rekindling the scientist in all of us

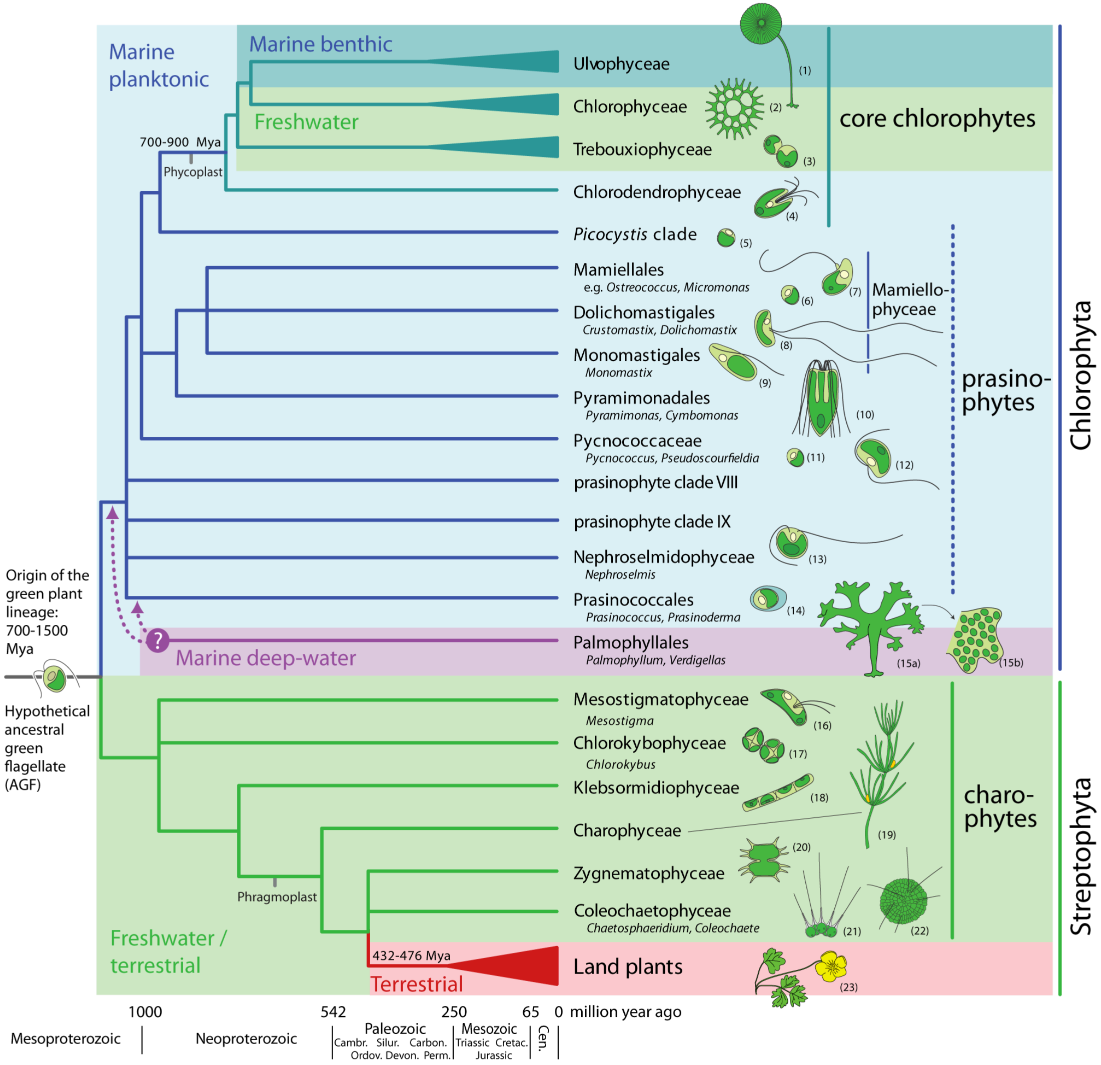
Some Trees



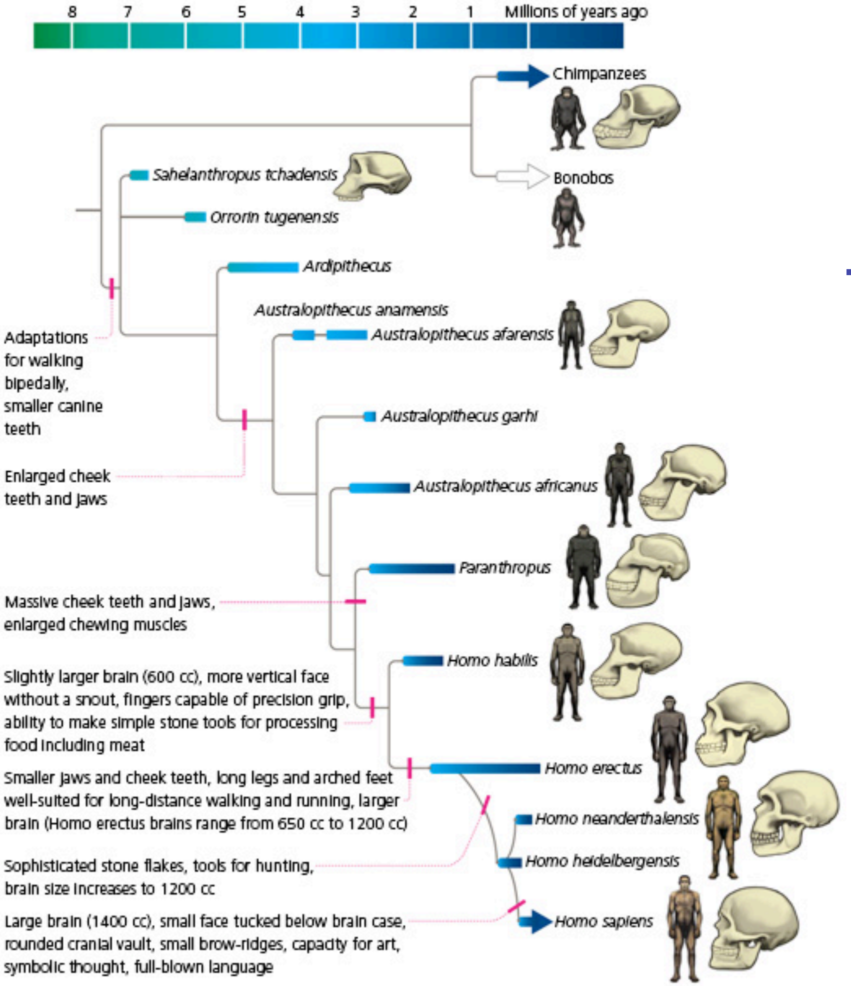


adapted by
BornToScience.com
Rekindling the scientist in all of us

Some Trees


Marine planktonic
Marine benthic
Freshwater
700-900 Mya
Phycoplast
Ulvophyceae (1)
Chlorophyceae (2)
Trebouxiophyceae (3)
Chlorodendrophyceae (4)
core chlorophytes
Picocystis clade (5)
Mamiellales
e.g. *Ostreococcus, Micromonas* (6) (7)
Mamiello-phyceae
Dolichomastigales
Crustomastix, Dolichomastix (8)
Monomastigales
Monomastix (9)
Pyramimonadales
Pyramimonas, Cymbomonas (10)
Pycnococcaceae
Pycnococcus, Pseudoscourfieldia (11) (12)
prasinophyte clade VIII
prasinophyte clade IX
Nephroselmidophyceae
Nephroselmis (13)
Prasinococcales
Prasinococcus, Prasinoderma (14)
prasino-phytes
Palmophyllales
Palmophyllum, Verdigellas (15a) (15b)
Marine deep-water
?
Chlorophyta
Origin of the green plant lineage: 700-1500 Mya
Hypothetical ancestral green flagellate (AGF)
Mesostigmatophyceae
Mesostigma (16)
Chlorokybophyceae
Chlorokybus (17)
Klebsormidiophyceae (18)
Charophyceae (19)
Zygnematophyceae (20)
Coleochaetophyceae
Chaetosphaeridium, Coleochaete (21) (22)
Phragmoplast
charo-phytes
Streptophyta
Freshwater / terrestrial
432-476 Mya
Terrestrial
Land plants (23)
1000
542
250
65
0 million year ago
Mesoproterozoic
Neoproterozoic
Paleozoic
Cambr. Silur. Carbon.
Ordov. Devon. Perm.
Mesozoic
Triassic Cretac.
Jurassic
Cen.

Some Trees



Adaptations for walking bipedally, smaller canine teeth

Enlarged cheek teeth and jaws

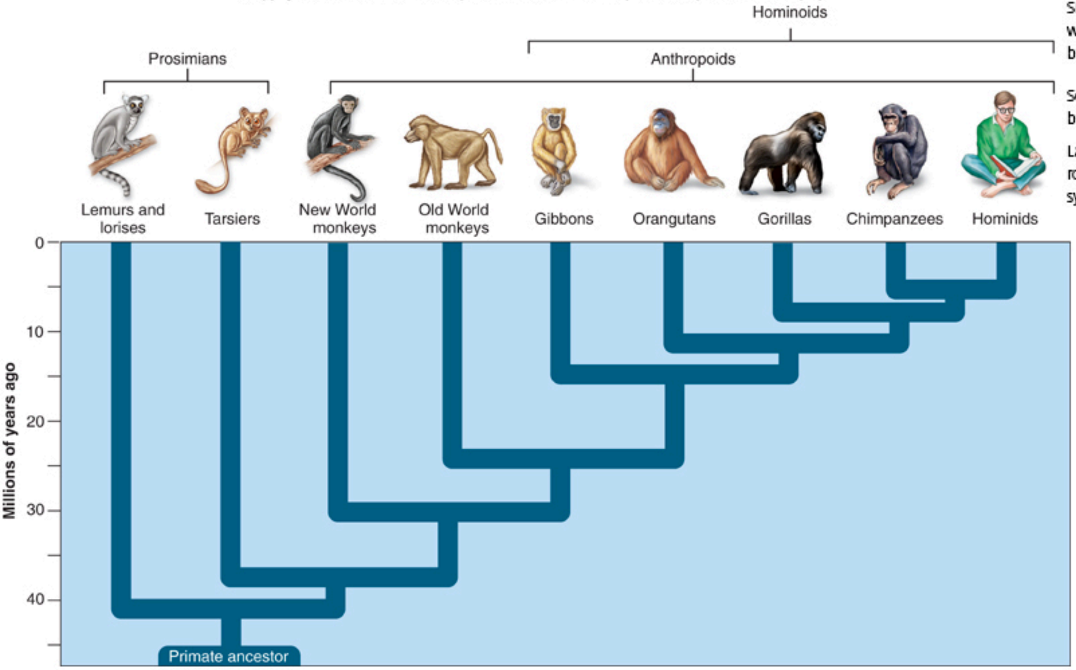
Massive cheek teeth and jaws, enlarged chewing muscles

Slightly larger brain (600 cc), more vertical face without a snout, fingers capable of precision grip, ability to make simple stone tools for processing food including meat

Smaller jaws and cheek teeth, long legs and arched feet well-suited for long-distance walking and running, larger brain (Homo erectus brains range from 650 cc to 1200 cc)

Sophisticated stone flakes, tools for hunting, brain size increases to 1200 cc

Large brain (1400 cc), small face tucked below brain case, rounded cranial vault, small brow-ridges, capacity for art, symbolic thought, full-blown language


Hominoids
Prosimians
Anthropoids
Lemurs and lorises
Tarsiers
New World monkeys
Old World monkeys
Gibbons
Orangutans
Gorillas
Chimpanzees
Hominids
0
10
20
30
40
Millions of years ago
Primate ancestor