Bioinformatics — Lecture 4
Phylogeny reconstruction
(EG Chps. 14, 15; MM Chps. 8)

Krzysztof Bartoszek

Linköping University

krzysztof.bartoszek@liu.se

19 XI 2024 (R35)

Today

Additional reading

DEKM R. Durbin, S. Eddy, A. Krogh, G. Mitchison. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge, 1998, Cambridge University Press.

F J. Felsenstein. Inferring Phylogenies, 2004, Sinauer.

O.G O. Gascuel (Editor). Mathematics of Evolution & Phylogeny, Oxford, 2005, Oxford University Press.

HRS D. H. Huson, R. Rupp, C. Scornavacca. Phylogenetic Networks, Cambridge, 2010, Cambridge University Press.

LSV P. Lemey, M. Salemi, A–M. Vandamme (Editors). The Phylogenetic Handbook, Cambridge, 2009, Cambridge University Press.

Y Z. Yang. Computational Molecular Evolution, Oxford, 2006, Oxford University Press.

Input and output

INPUT: A multiple sequence alignment of sequences

OUTPUT: A graph representing the common history of (similarities between) these sequences

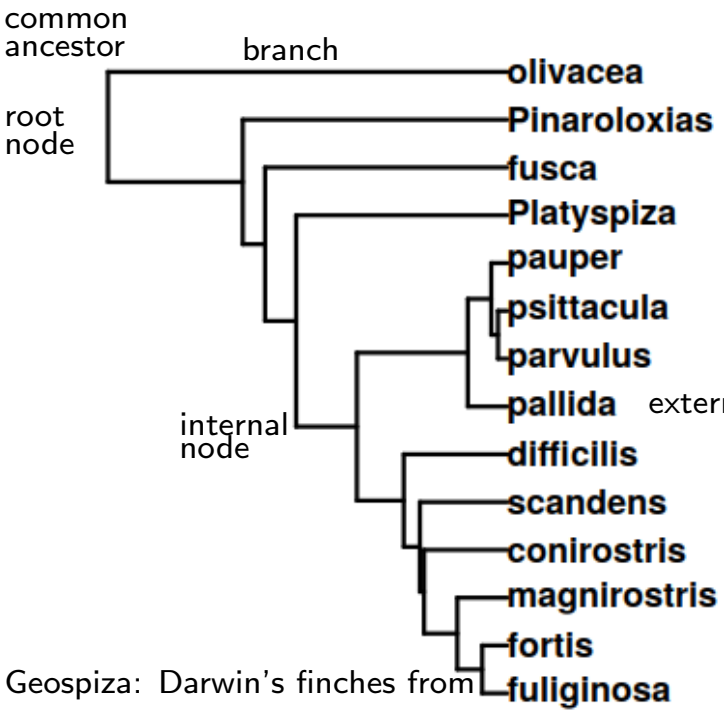
https://biology.stackexchange.com/questions/46481/darwins-first-sketch-of-a-phylogenetic-tree
from Remi.b, CC BY-SA

Types of input data

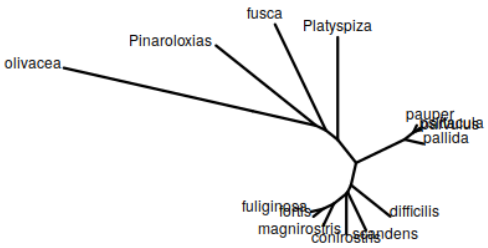
- Sequence data (multiple alignments):
  - constant columns carry no information
    - Amino acid sequences
    - Nucleotide sequences
      - Coding regions (selection)
        - 3rd position versus 1st, 2nd
      - Non–coding regions
      - Mitochondrial DNA
- Morphological data
- Joint sequence and morphological data
- From
  - different species
  - different individuals
  - different genes
  - different viral strains

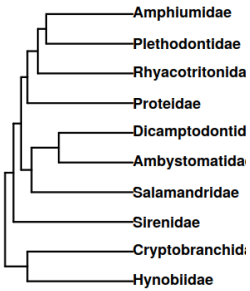
See also http://darwin-online.org.uk/content/frameset?itemID=CUL-DAR121.-&viewtype=side&pageseq=38

Rooted versus unrooted trees



Geospiza: Darwin's finches from `geiger` R package, GNU GPL

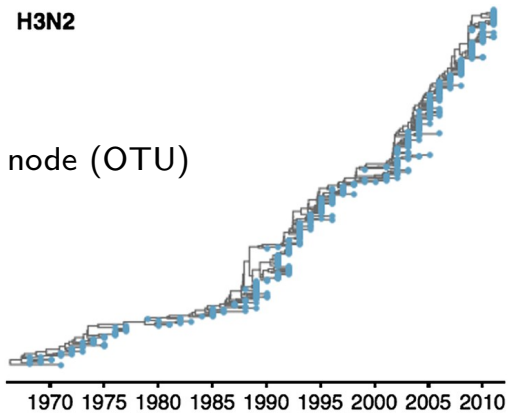


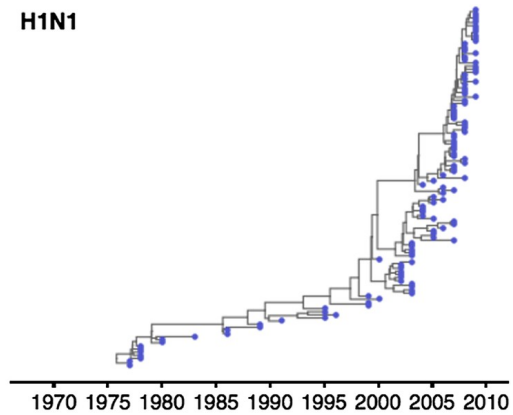


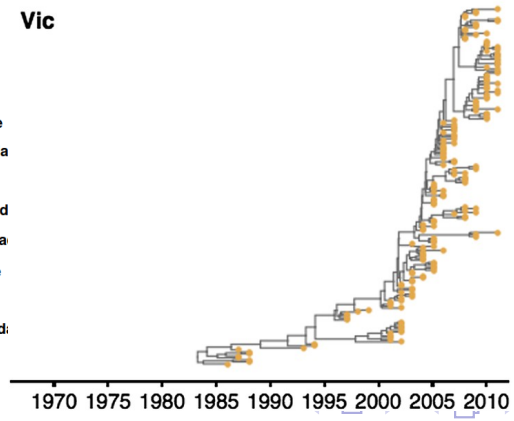
Salamander families (`geiger`, GNU GPL)

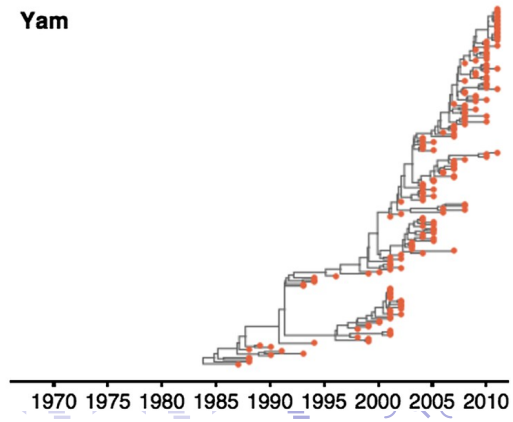
maximium–clade credibility trees

T. Bedford, M.A. Suchard, P. Lemey, G. Dudas, V. Gregory, A.J. Hay, J.W. McCauley, C.A. Russell, D.J. Smith, A. Rambaut (2014) Integrating influenza antigenic dynamics with molecular evolution eLife 3:e01914 doi: 10.7554/eLife.01914 CC BY 3.0

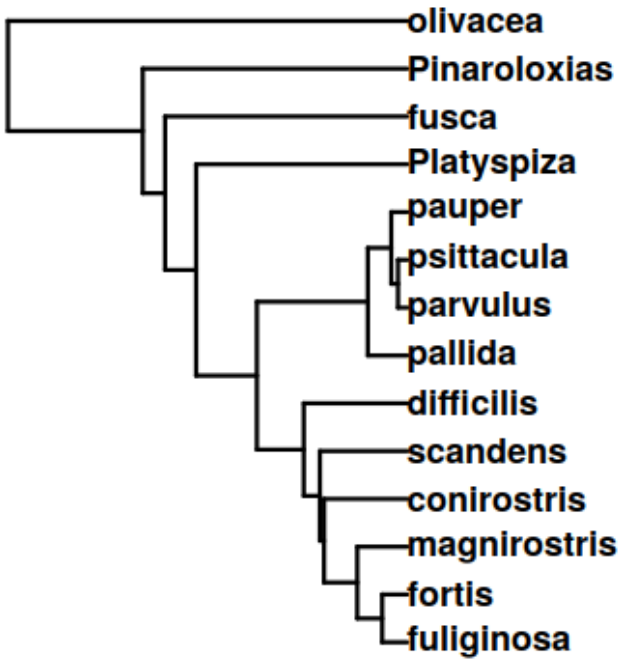






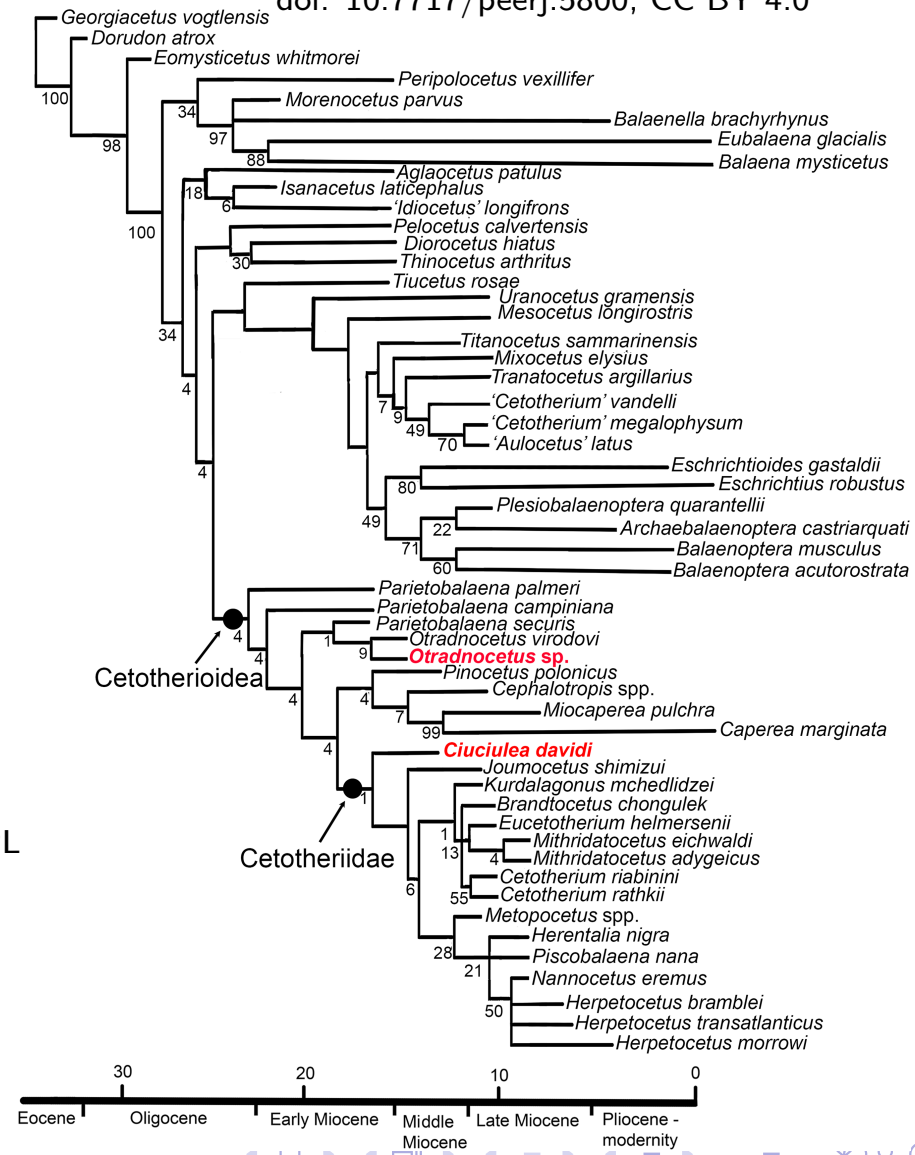


Ultrametric versus non–ultrametric

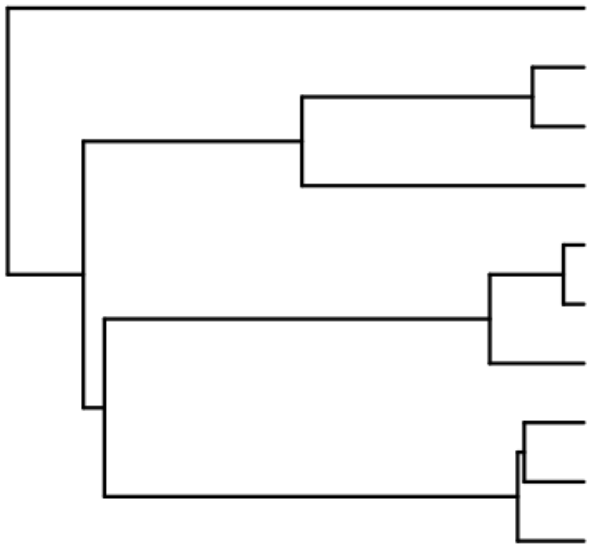


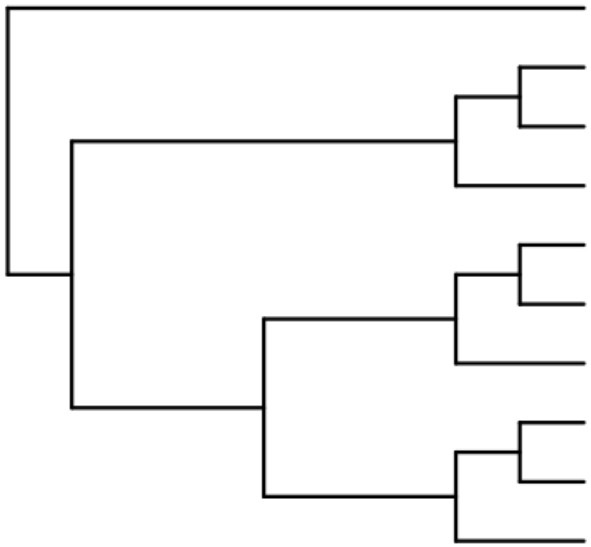
Geospiza: Darwin's finches from `geiger` R package, GNU GPL

P. Gol'din P. 2018. New Paratethyan dwarf baleen whales mark the origin of cetotheres. PeerJ 6:e5800 doi: 10.7717/peerj.5800, CC BY 4.0



Branch lengths

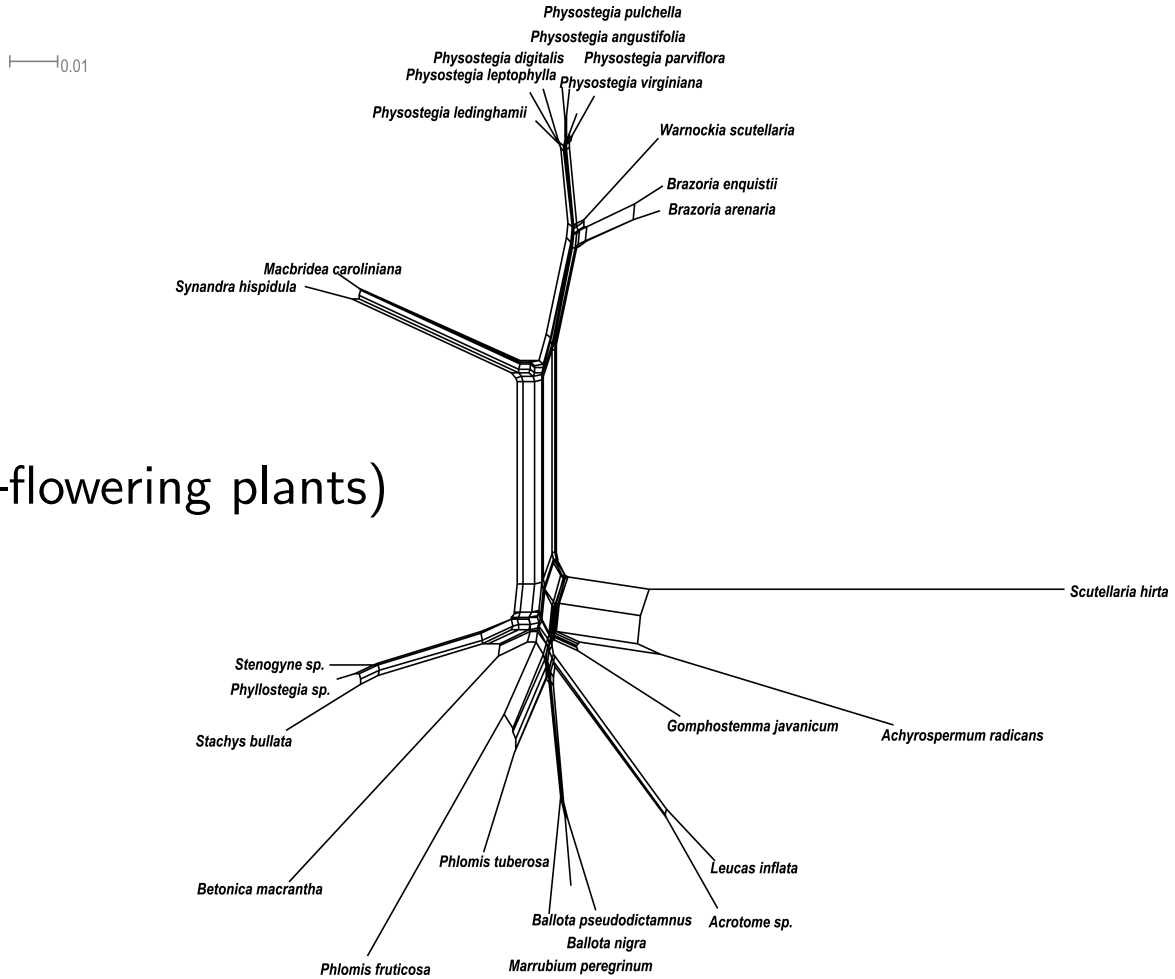




Trees simulated by `TreeSim` R package, GNU GPL-2

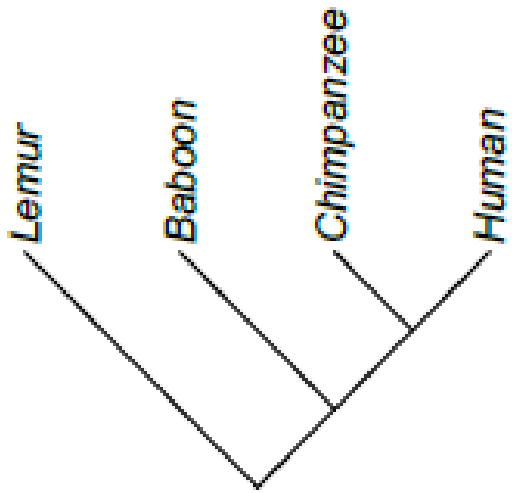
Phylogenetic networks

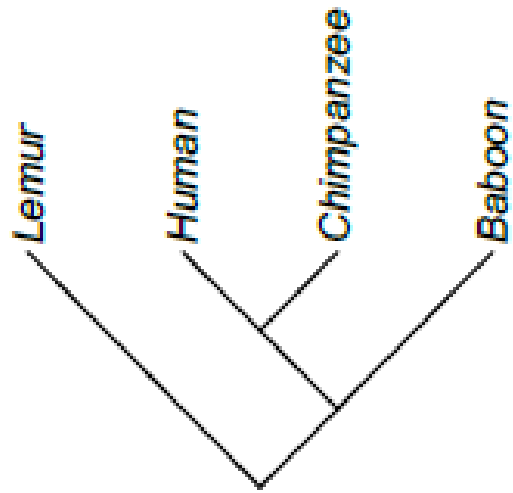
(angiosperms—flowering plants)



T. Roy, N.S. Catlin, D.M.G. Garner, P.D. Cantino, A.-C. Scheen, C. Lindqvist, 2016. Evolutionary relationships within the lamioid tribe Synandreae (Lamiaceae) based on multiple low-copy nuclear loci. PeerJ 4:e2220, doi: 10.7717/peerj.2220, CC BY 4.0

Combinatorics of trees





https://bio.libretexts.org/Learning_Objects/Laboratory_Experiments/BIOL_111_-_Laboratory_Manual/10:_Animal_Diversity-_Create_a_Phylogeny/10.03:_Understanding_phylogenetic_trees, public domain, authored, remixed, and/or curated by Alexey Shipunov.

How many trees are there on n tips?

rooted versus unrooted

lablelled versus unlabelled

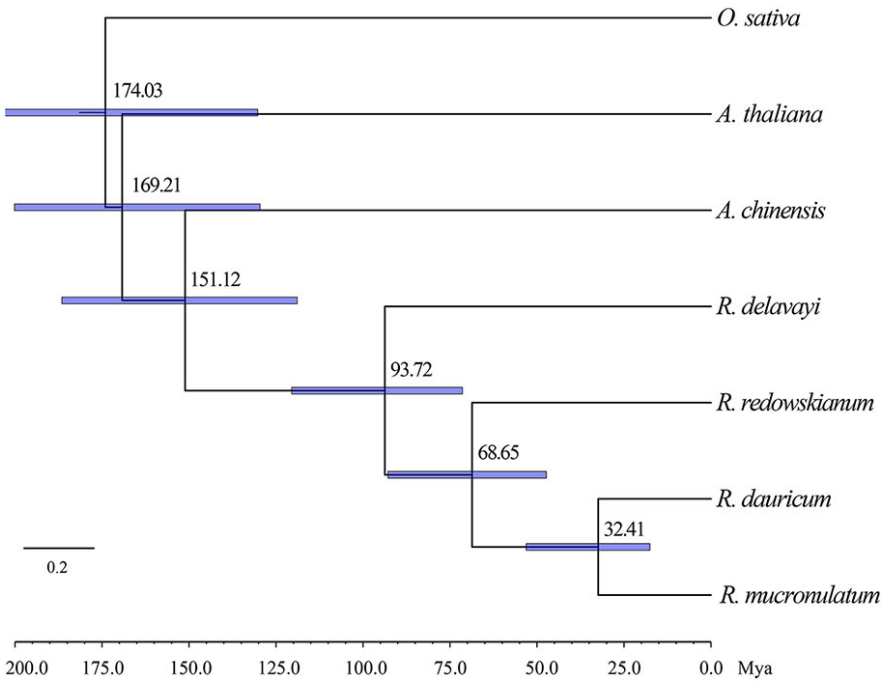
bifurcating versus multifurcating

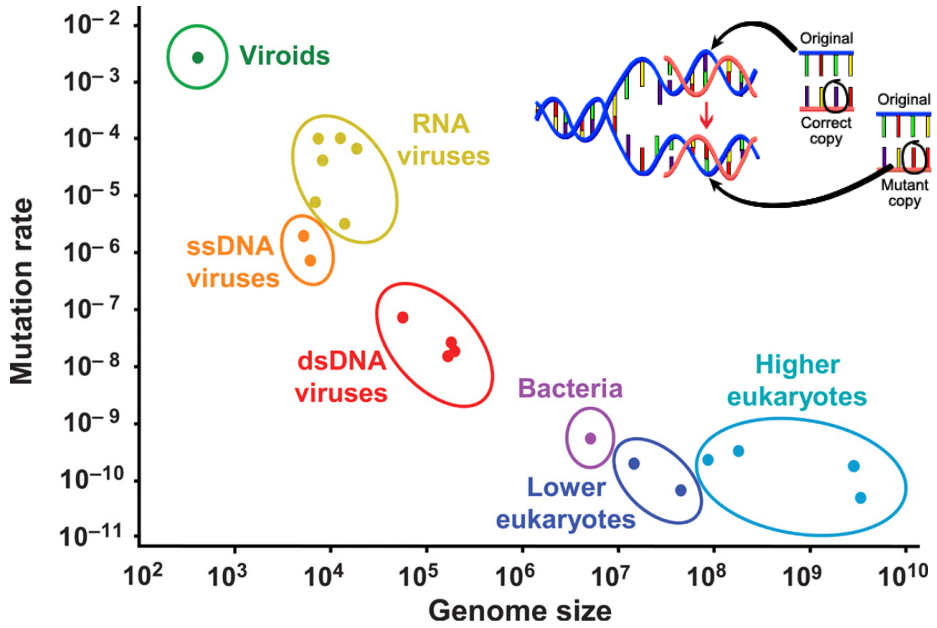
bifurcating labelled trees
rooted: $((2n-3)!)/(2^{n-2}(n-2)!)$
urooted: $(2n-5)!! = 1 \cdot 3 \cdot 5 \cdot \ldots \cdot (2n-5)$

Robinson Foulds metric: are the toplogies the same

The time scale

https://biology.stackexchange.com/questions/24398/viral-mutation-mechanism
by David/Chris, CC BY-SA 3.0





Yang B, Zhang G, Guo F, Wang M, Wang H and Xiao H (2020) A Genomewide Scan for Genetic Structure and Demographic History of Two Closely Related Species, *Rhododendron dauricum* and *R. mucronulatum* (*Rhododendron*, Ericaceae). Front. Plant Sci. 11:1093.
doi: 10.3389/fpls.2020.01093, CC BY 4.0

Rhododendron dauricum by Anneli Salo - Own work, CC BY-SA 3.0

Modelling approach (EG Ch. 14)

Model for nucleotide change on an interval

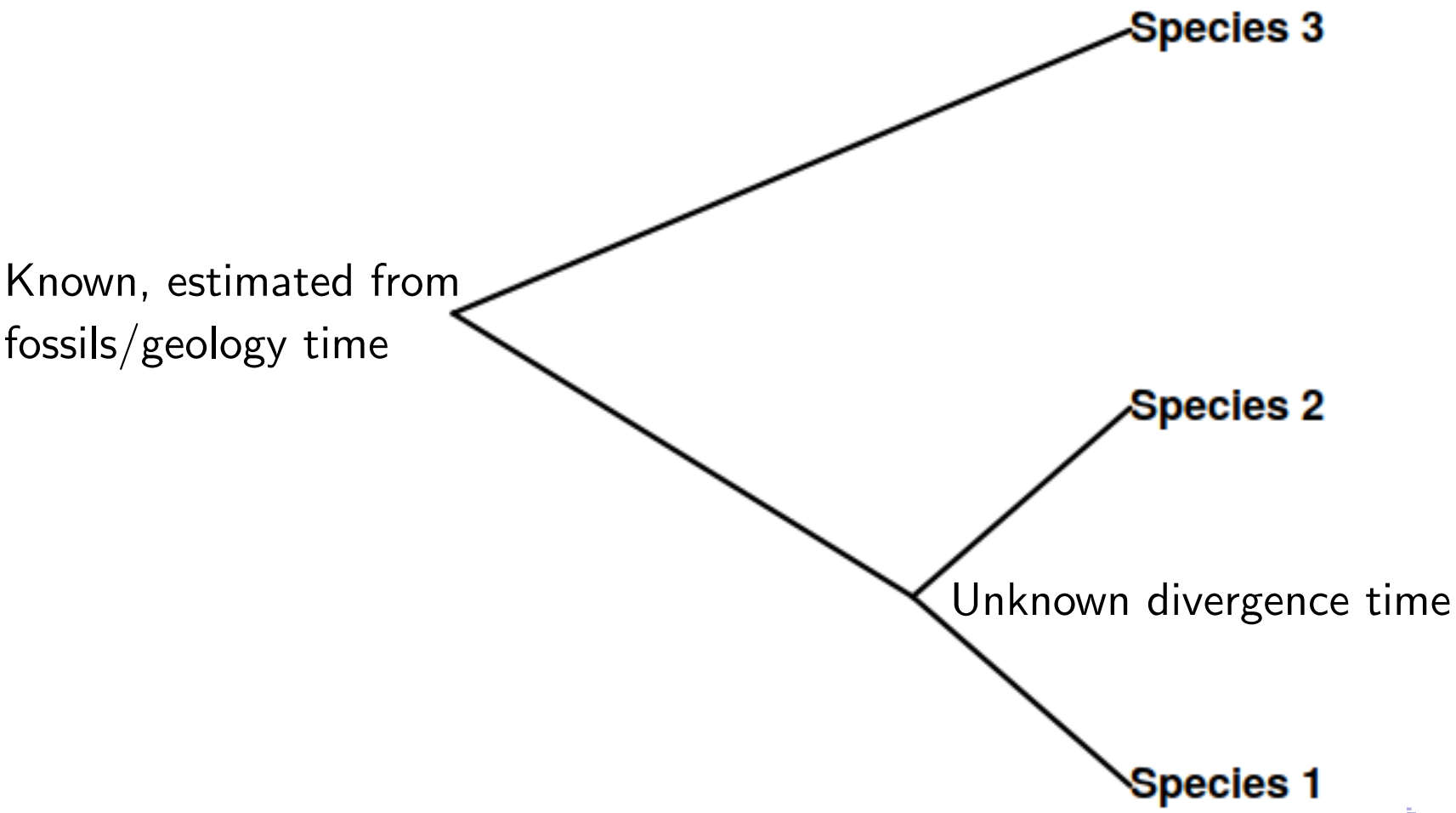
- Markov chain/process on four states

Model for time to branching

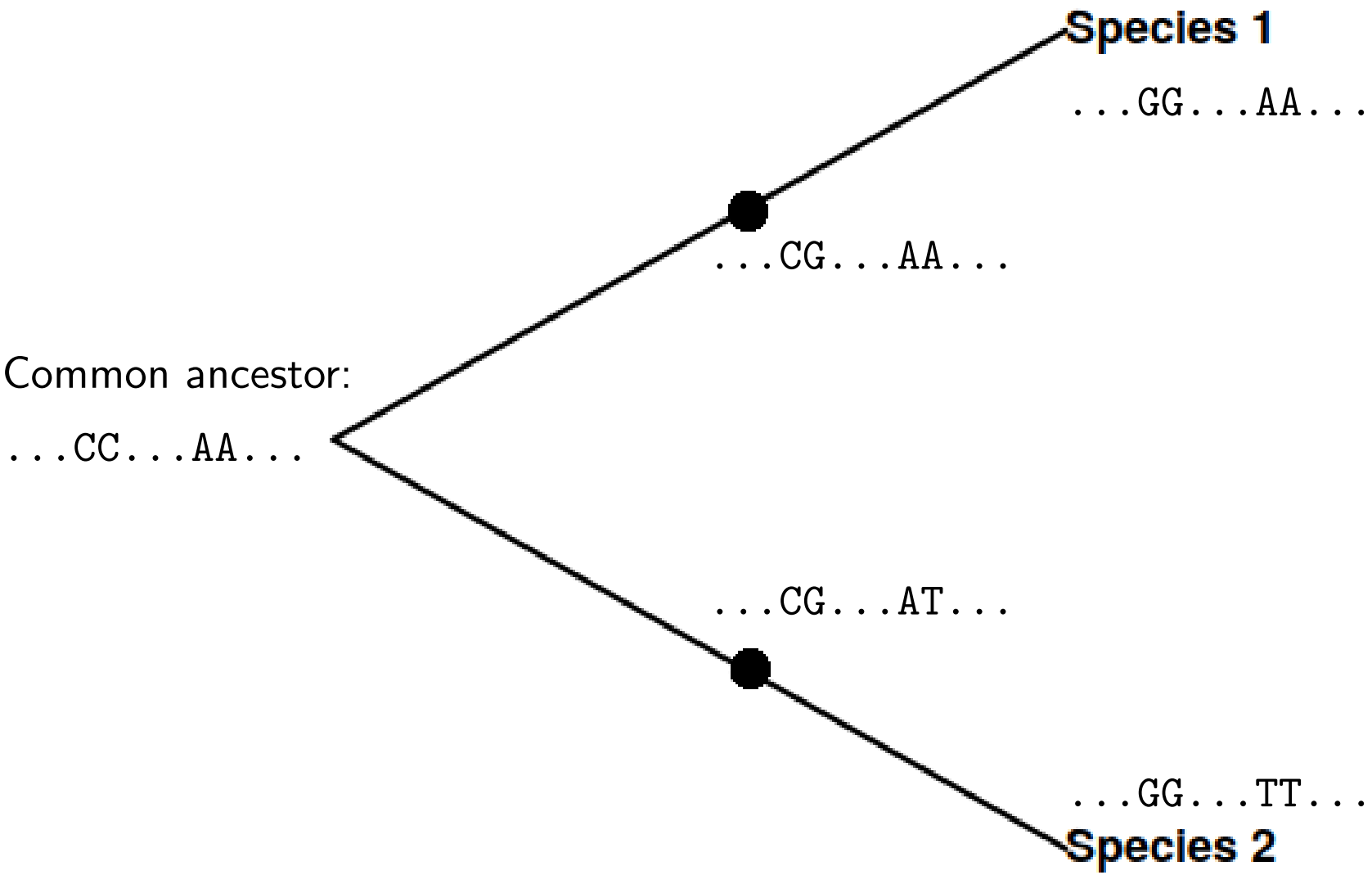
- random or fixed number of steps
- random variable (exponential, gamma)

Daughter branches inherit ancestral nucleotide

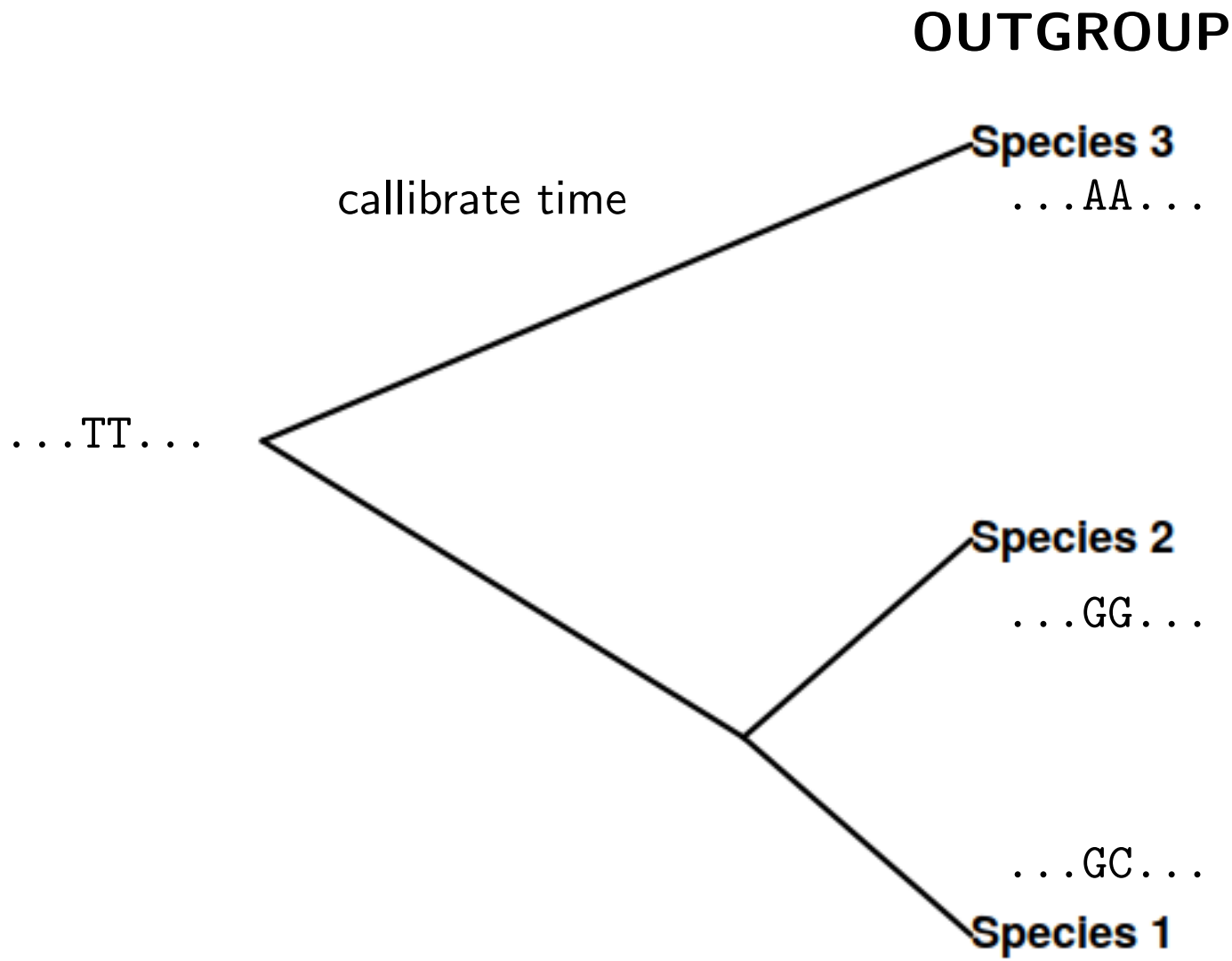
See also MM Fig. 8–6A



See also MM Fig. 8–6B



See also MM Fig. 8–6C



Discrete–time models (Markov chain)

Nucleotide transition matrix:

$$\mathbf{P} = \begin{bmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{bmatrix}$$

Initial distribution: $\vec{\phi}_0 = (\phi_{0,A}, \phi_{0,C}, \phi_{0,G}, \phi_{0,T})^T$

Distribution after n steps: $\vec{\phi}_n^T = \phi_{n-1}^T \mathbf{P} = \ldots = \phi_0^T \mathbf{P}^n$

Stationary distribution $\vec{\phi}_{d*}^T = \phi_{d*}^T \mathbf{P}$

Continuous–time models (EG Ch. 11.7)

Nucleotide instantaneous transition rate matrix (generator):

$$\mathbf{Q} = \begin{bmatrix} q_{AA} & q_{AC} & q_{AG} & q_{AT} \\ q_{CA} & q_{CC} & q_{CG} & q_{CT} \\ q_{GA} & q_{GC} & q_{GG} & q_{GT} \\ q_{TA} & q_{TC} & q_{TG} & q_{TT} \end{bmatrix}$$

s.t. $q_{ii} = -\sum_{i \neq j} q_{ij}$
Transition probability function (*forward Kolmogorov equation*):

$$\frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} = -\mathbf{Q}\mathbf{P}(t)$$

Distribution after time t: $\vec{\phi}_n = \mathbf{P}(t)\vec{\phi}_0$
Times between mutations are exponential
Stationary distribution $\vec{\phi}_{c^*}^T = \vec{\phi}_{c^*}^T \mathbf{P}(t)$

Jukes–Cantor model (EG Ch. 14.2.1, 14.3.1)

$$\mathbf{P} = \begin{bmatrix} 1-3\alpha & \alpha & \alpha & \alpha \\ \alpha & 1-3\alpha & \alpha & \alpha \\ \alpha & \alpha & 1-3\alpha & \alpha \\ \alpha & \alpha & \alpha & 1-3\alpha \end{bmatrix} \quad \mathbf{Q} = \begin{bmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{bmatrix}$$

1 free parameter

$P(X_n^d = X_0^d) = (1/4) + (3/4) \cdot (1-4\alpha)^n$

$P(X_n^d = i) = (1/4) - (1/4) \cdot (1-4\alpha)^n) \quad i \neq X_0^d$

$$P_{ii}(t) = (1/4) + (3/4)e^{-4\alpha t} \qquad P_{ij}(t) = (1/4) - (1/4)e^{-4\alpha t} \quad i \neq j,$$

$$\vec{\phi}^T_{JCd^*} = (1/4, 1/4, 1/4, 1/4)^T = \vec{\phi}^T_{JCc^*}$$

Kimura models (EG Ch. 14.2.2, 14.2.3, 14.3.2)

$$\mathbf{P} = \begin{bmatrix} 1-\alpha-2\beta & \beta & \alpha & \beta \\ \beta & 1-\alpha-2\beta & \beta & \alpha \\ \alpha & \beta & 1-\alpha-2\beta & \beta \\ \beta & \alpha & \beta & 1-\alpha-2\beta \end{bmatrix}$$

$$\mathbf{Q} = \begin{bmatrix} \alpha+2\beta & \beta & \alpha & \beta \\ \beta & \alpha+2\beta & \beta & \alpha \\ \alpha & \beta & \alpha+2\beta & \beta \\ \beta & \alpha & \beta & \alpha+2\beta \end{bmatrix}$$

2 free parameters

Transitions: α (purine by purine/pyramidine by pyramidine, $A \leftrightarrow G$, $C \leftrightarrow T$)

Transversion: β (purine $\leftrightarrow$ pyramidine, $\{A, G\} \leftrightarrow \{C, T\}$)

$$\vec{\phi}^T_{Kd^*} = (1/4, 1/4, 1/4, 1/4)^T = \vec{\phi}^T_{Kc^*}$$

Kimura models (EG Ch. 14.2.2, 14.2.3, 14.3.2)

$$\mathbf{P_{3ST}} = \begin{bmatrix} 1-\alpha-\beta-\gamma & \beta & \alpha & \gamma \\ \beta & 1-\alpha-\beta-\gamma & \gamma & \alpha \\ \alpha & \gamma & 1-\alpha-\beta-\gamma & \beta \\ \gamma & \alpha & \beta & 1-\alpha-\beta-\gamma \end{bmatrix}$$

$$\mathbf{P_{Eq14.13}} = \begin{bmatrix} 1-\alpha-2\gamma & \gamma & \alpha & \gamma \\ \delta & 1-\alpha-2\delta & \delta & \alpha \\ \alpha & \gamma & 1-\alpha-2\gamma & \gamma \\ \delta & \alpha & \delta & 1-\alpha-2\delta \end{bmatrix}$$

and others

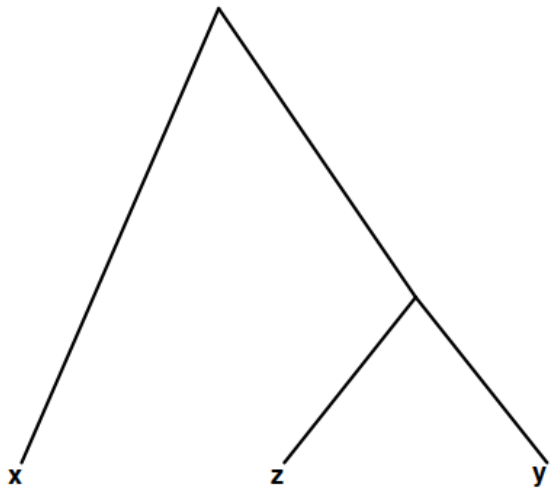
Felsenstein models (EG Ch. 14.2.4, 14.3.3)

Probability/rate of substitution is proportional to the stationary distribution of the substituting nucleotide.

$$\mathbf{P} = \begin{bmatrix} 1-u+u\phi_*^A & u\phi_*^C & u\phi_*^G & u\phi_*^T \\ u\phi_*^A & 1-u+u\phi_*^C & u\phi_*^G & u\phi_*^T \\ u\phi_*^A & u\phi_*^C & 1-u+u\phi_*^G & u\phi_*^T \\ u\phi_*^A & u\phi_*^C & u\phi_*^G & 1-u+u\phi_*^T \end{bmatrix}$$

$$\text{continuous}: \begin{cases} P_{ii}(t) = e^{-ut} + (1-e^{-ut})\phi_*^i \\ P_{ij}(t) = (1-e^{-ut})\phi_*^j \quad i \neq j \end{cases}$$

Distance functions (EG Eq. (15.1))



A tree can be thought of as a distance function on the set of tips
$d(y,z) = 2$, $d(x,y) = d(x,z) = 4$
ultrametric tree: all tips are contemporary *or*
for all triplets of tips x, y, z:
two of the distances are equal: $d(x,y) = d(x,z)$
and are greater than the third: $d(y,z) \leq d(x,y)$, $d(y,z) \leq d(x,z)$

UPGMA algorithm (EG Ch. 15.3, MM p.177)

Unweighted pair group method with arithmetic mean
Given any ultrametric distance a unique tree can be derived

distance between two clusters G_u and G_v of tips is

$$d(G_u, G_v) = \frac{1}{|G_u| \cdot |G_v|} \sum_{u \in G_u} \sum_{v \in G_v} d(u, v)$$

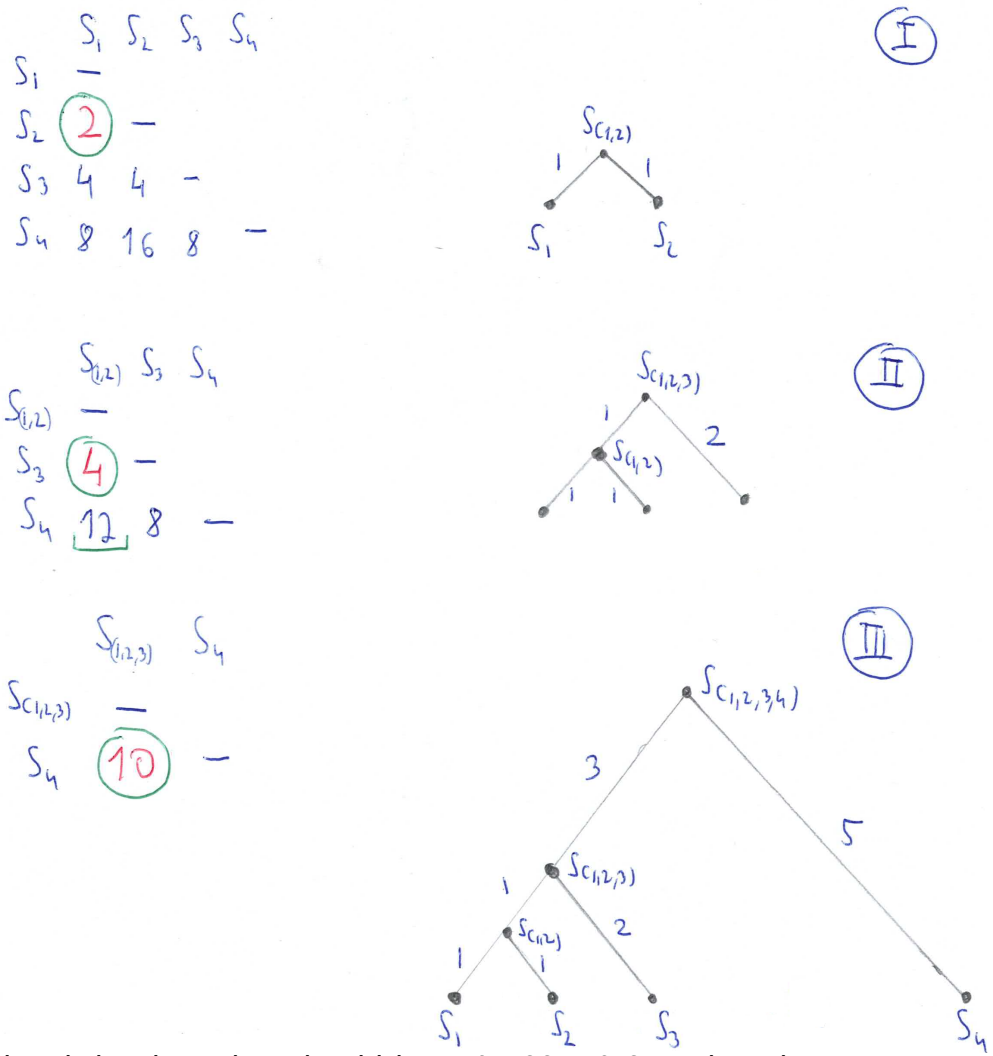
distance update:

$$d(G_u \cup G_v, G_w) = \frac{|G_u| d(G_u, G_w) + |G_v| d(G_v, G_w)}{|G_u| + |G_v|}$$

UPGMA algorithm (EG Ch. 15.3, MM p.177)

1. Join the closest two sequences to form a cluster
2. Recalculate the evolutionary distances between the cluster and the remaining sequences.
3. Join the closest two sequences or join the closest cluster and sequence
4. Recalculate the evolutionary distances between the clusters and the remaining sequences
5. Repeat steps 3 and 4 until all sequences are connected in a single cluster.

UPGMA algorithm (see also MM Fig. 8–13 A–D)



Typo in MM Fig. 8–13 D, root's daughther branches should have 0.4225, 0.2775 lengths.

Neighbour joining algorithm (EG 15.4, MM p. 188)

Given any tree–derived distance $d(\cdot, \cdot)$ a tree can be constructed

1. For all pairs of tips (x, y) calculate

$$\delta(x, y) = (n - 4)d(x, y) - \sum_{z \neq, x, y} (d(x, z) + d(y, z))$$

δ is **NOT** a distance as it can take *negative* values.

2. Find the pair, (x, y) with lowest value of δ. They have to be neighbours in d. Create a new node clustering them.

Neighbour joining algorithm (EG 15.4, MM p. 188)

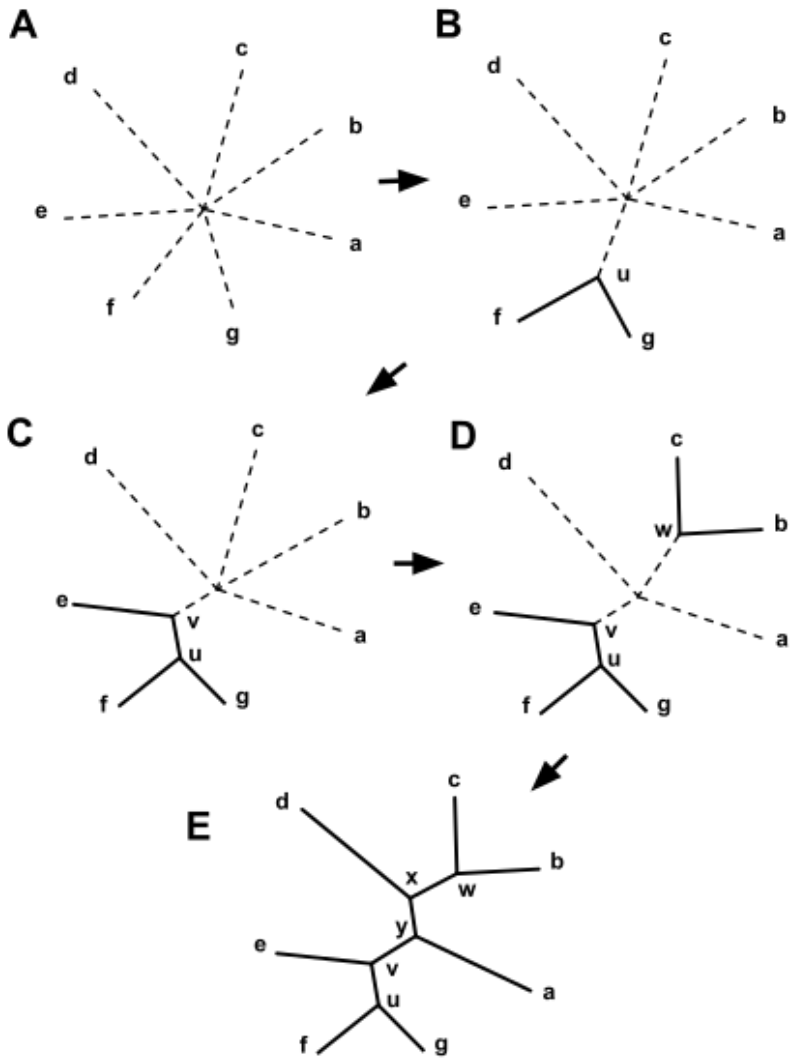
3. Calculate the distance from each of the taxa outside of this pair, z to the new node as
$d(\{x, y\}, z) = (d(x, z) + d(y, z) - d(x, y))/2$.

4. Calculate the distance from each of the taxa in the pair to this new node, using any outside z, as
$d(x, \{x, y\}) = (d(x, z) - d(y, z) + d(x, y))/2$,
$d(y, \{x, y\}) = (d(y, z) - d(x, z) + d(x, y))/2$.

5. Repeat steps 1.–4. using the new node $\{x, y\}$ instead of the nodes x and y, remember $n := n - 1$

https://en.wikipedia.org/wiki/Neighbor_joining

Example



Parsimony (EG Ch. 15.6)

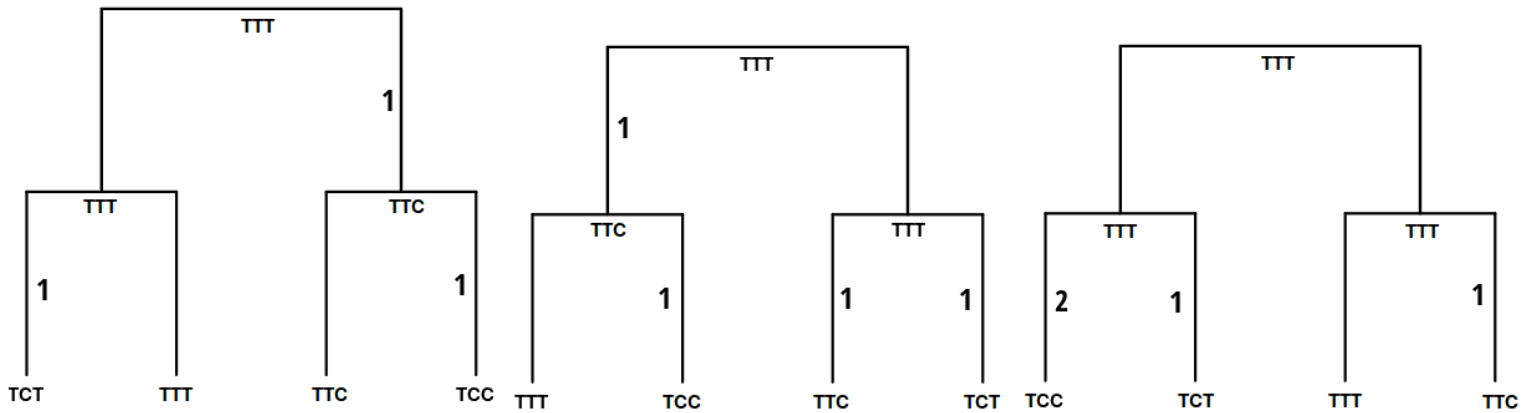
To each tree a *cost* is assigned.

Optimal tree: minimizing this cost. Common cost: each nucleotide substitution has *unit* cost.

Step 1 List all possible topologies

Step 2 For each tree find labeling of internal nodes minimizing cost (*Fitch's algorithm*)

Step 1 is impossible for larger clades—*heuristics*



Fitch's algorithm (DEKM Ch. 7.4)

Initialize: $C = 0$, $k = 2n - 1$

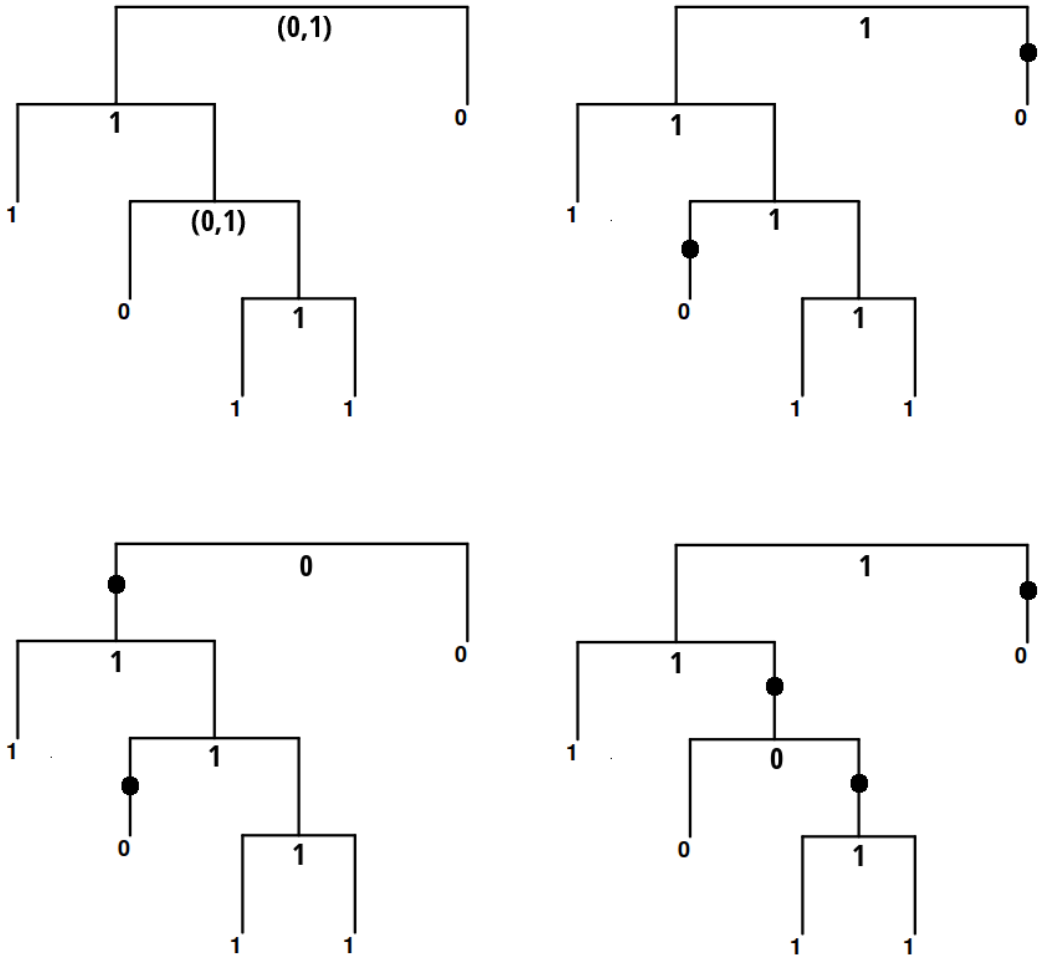
Recursion: To obtain the set R_k:

If k is a leaf node:
Set $R_k = x_u^k$
If k is not a leaf node:
Compute R_i, R_j for the daughter nodes i, j of k and set $R_k = R_i \cap R_j$ if this intersection is not empty, or else set $R_k = R_i \cup R_j$ and $C++$

Termination: Minimal cost of tree$= C$

Fitch, M.W. (1971) Defining the course of Evolution: Minimum change for a specific tree topology. *Syst. Zool.* 20:406–416

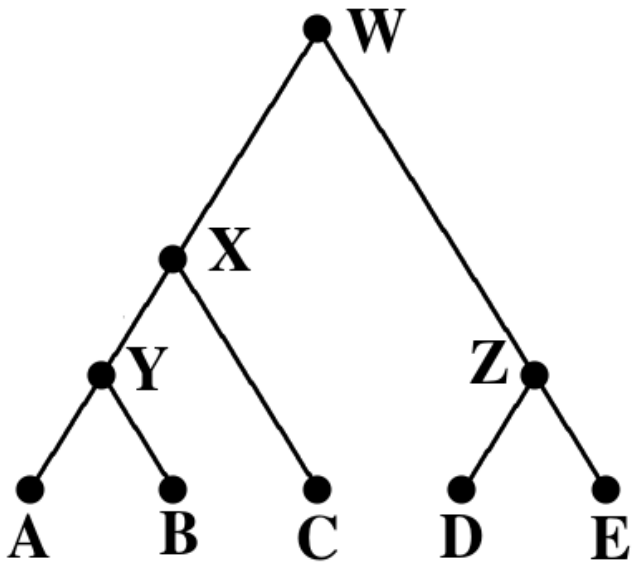
See also DEKM Fig. 7.10



Maximum likelihood (EG Ch. 15.7)

Model of nucleotide substitutions

$$\phi_W P_{WX}(d_{WX}) P_{WZ}(d_{WZ}) P_{XY}(d_{XY}) P_{XC}(d_{XC}) P_{YA}(d_{YA}) P_{YB}(d_{YB}) \times P_{ZD}(d_{ZD}) P_{ZE}(d_{ZE})$$



Maximum likelihood

Likelihood calculated for all possible combinations of internal node values (Felsenstein's pruning algorithm).

Independent columns assumption: repeat for all columns of the alignment and take product

Repeat for all possible topologies (heuristics)

Optimize for evolutionary model parameters

Optimization over: model parameters (numerical), branch lengths (numerical), topology (special)

Felsenstein's pruning algorithm (LSV Ch. 6.3.1, DEKM p. 201)

$s \in \{A, C, G, T\}$
i inner node
$L^i(s)$: likelihood of subtree rooted at i with nucleotide s at node i

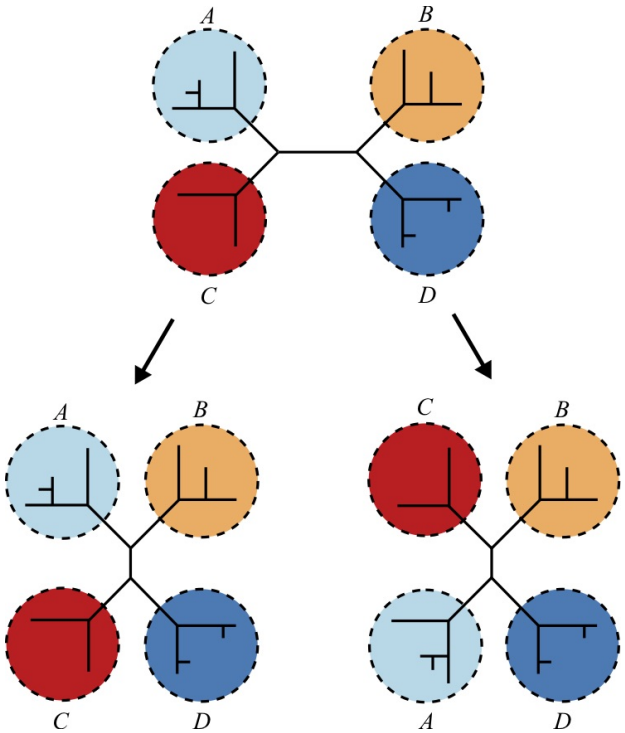
$$L^i(s) = \prod_{o:\text{daughter of } i} \left[\sum_{x \in \{A,C,G,T\}} P_{sx}(t_o) L^o(x) \right]$$

and for a leaf j

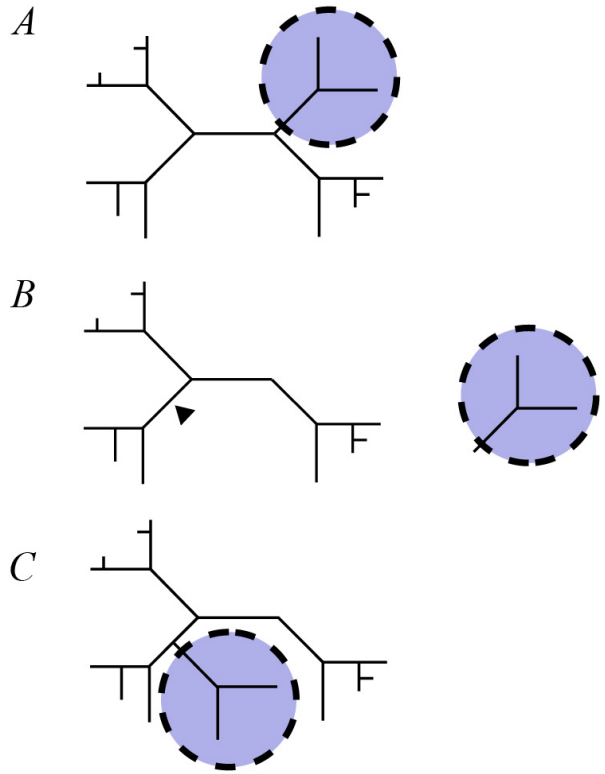
$$L^j(s) = \begin{cases} 1 & \text{if } s \text{ is leaf } j\text{'s state} \\ 0 & \text{otherwise} \end{cases}$$

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* 17:368–376

Moving in tree space (see also LSV Fig. 6–5)



Nearest Neighbor Interchange



Subtree Pruning and Regrafting

Applied Bioinformatics,Chapter 4:Multiple Sequence Alignments, Molecular Evolution, and Phylogenetics by David A. Hendrix, by David A. Hendrix, CC BY 4.0

Model selection (LSV Ch 10, EG Ch 5.9.4)

Models: Competing phylogenies, evolutionary models e.t.c.

Nested models: R^2, likelihood ratio

$$R^2 = RSS_{\text{model}}/RSS_{\text{under null model}}$$

$$2(\mathcal{L}_1 - \mathcal{L}_0) \sim \chi^2_{K_1 - K_0}$$

Information criteria (LSV Ch 10, EG Ch 5.9.4)

The lower the better

Akaike Information Criterion (AIC), corrected (AIC_c)for sample size (n)

$$AIC = -2\mathcal{L} + 2K \qquad AIC_c = AIC + \frac{2K(K+1)}{n-K-1}$$

Bayesian/Schwarz Information Criterion (BIC):

$$BIC = -2\mathcal{L} + K \log n$$

Bayes factor (difficult to compute, BIC approximation for logs)

$$B_{ij} = \frac{P(Data|M_i)}{P(Data|M_j)}$$

We have R competing models, $i = 1, \ldots R$.

$$\Delta AIC_{c_i} = AIC_{c_i} - AIC_{c_{\min}} \qquad \Delta AIC_i = AIC_i - AIC_{\min}$$

$0 < \Delta < 4$ or $0 < \Delta < 7$: plausible

$\Delta > 14$: implausible

$\Delta AIC_i \leq 2$	substantial support (evidence)
$4 \leq \Delta AIC_i \leq 7$	considerably less support
$\Delta AIC_i > 10$	essentially no support

Fig. 2 of K. P. Burnham, D. R. Anderson, K. P. Huyvaert, 2011. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. Behav. Ecol. Sociobiol. 65:23–35 doi:10.1007/s00265-010-1029-6

p. 271 of K. P. Burnham, D. R. Anderson, 2004. Multimodel Inference: Understanding AIC and BIC in Model Selection. Socio. Meth. Res. 33(2):261–304 doi:10.1007/10.1177/0049124104268644

AIC: contains large scaling constants, e.g., $AIC_1 = 300000$, $AIC_2 = 300020$
only ΔAIC_i: interpretable as strength of evidence (Burnham & Anderson 2004)

Model averaging (LSV Ch 10)

We have a large number of competing models

$$\Delta\{A/B\}IC_i = \{A/B\}IC_i - \min\{A/B\}IC$$

$$\text{weight of model } i \; w_i = \frac{\exp(-1/(2\Delta_i)}{\sum_r \exp(-1/(2\Delta_r)}$$

Model–averaged estimate of numerical parameter

$$\hat{\theta} = \frac{\sum_i w_i I_\theta(M_i)\hat{\theta}_i}{\sum_i w_i I_\theta(M_i)}$$

where

$$I_\theta(M_i) = \begin{cases} 1 & \text{if parameter } \theta \text{ belongs to model } M_i \\ 0 & \text{otherwise} \end{cases}$$

Split significance: bootstrapping (EG 15.9.2, MM. p186)

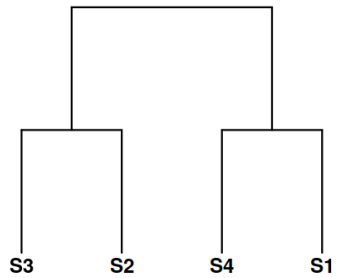
Data	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
S_1	T	A	T	A	T	G	C	G	C	C
S_2	T	A	A	T	A	G	C	G	C	C
S_3	T	A	T	A	A	G	G	G	C	C
S_4	T	A	T	A	A	G	G	G	C	C

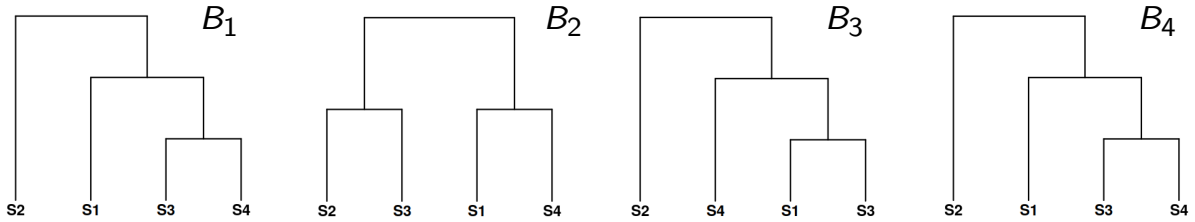
B_1	P_4	P_4	P_2	P_7	P_1	P_3	P_1	P_{10}	P_7	P_4
S_1	A	A	A	C	G	T	G	C	C	A
S_2	T	T	A	C	G	A	G	C	C	T
S_3	A	A	A	G	G	T	G	C	G	A
S_4	A	A	A	G	G	T	G	C	G	A

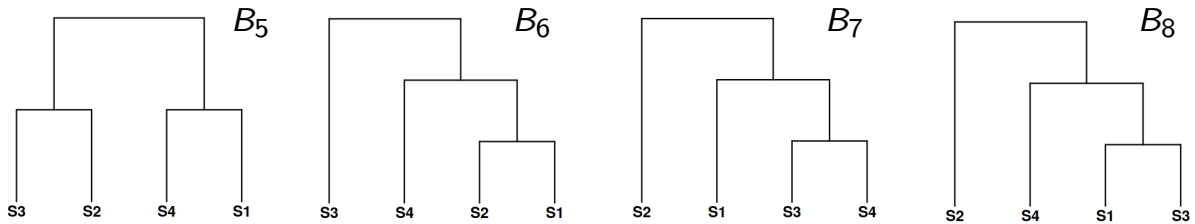
B_2	P_5	P_9	P_{10}	P_{10}	P_3	P_8	P_3	P_5	P_7	P_1
S_1	G	C	C	C	T	G	T	A	C	T
S_2	G	C	C	C	A	G	A	T	C	T
S_3	G	C	C	C	T	G	T	A	G	T
S_4	G	C	C	C	T	C	A	A	G	T

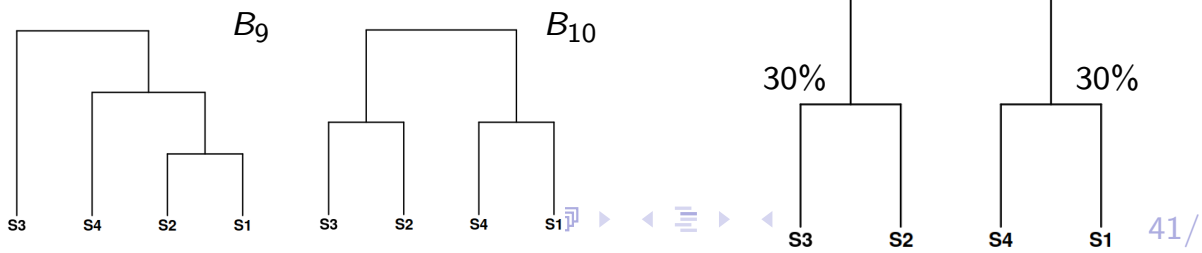
B_3	P_4	P_3	P_1	P_7	P_3	P_{10}	P_8	P_9	P_9	P_3
S_1	A	T	T	C	T	C	G	C	C	T
S_2	T	A	T	C	A	C	G	C	C	A
S_3	A	T	T	G	T	C	G	C	C	T
S_4	A	A	T	G	A	C	G	C	C	A

B_4	P_9	P_{10}	P_1	P_1	P_6	P_5	P_4	P_5	P_5	P_2
S_1	C	C	T	T	G	T	A	T	T	A
S_2	C	C	T	T	G	A	T	A	A	A
S_3	C	C	T	T	G	A	A	A	A	A
S_4	C	C	T	T	G	A	A	A	A	A



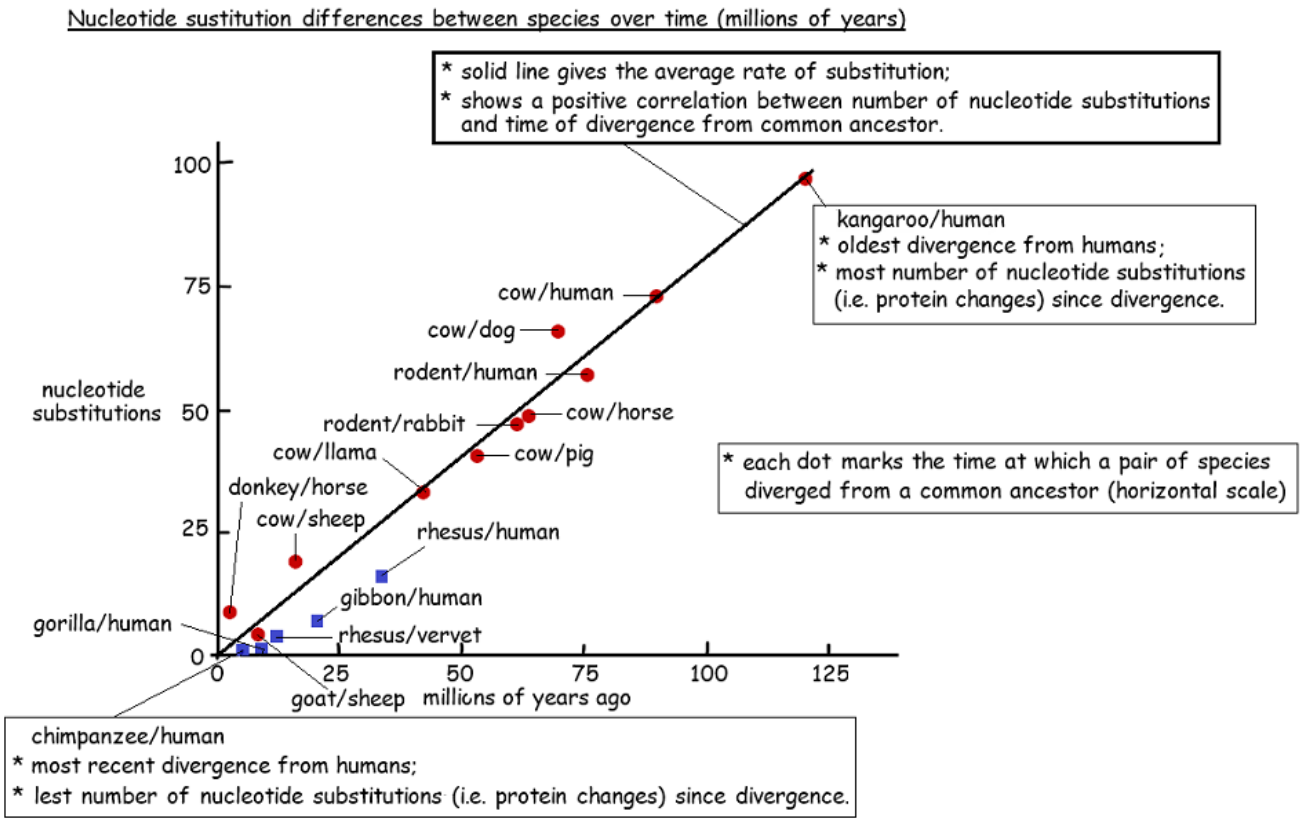






Molecular clock hypothesis (MM p. 162)

Amino acid (accepted) substitution rates (NOT mutation rates) are constant in time



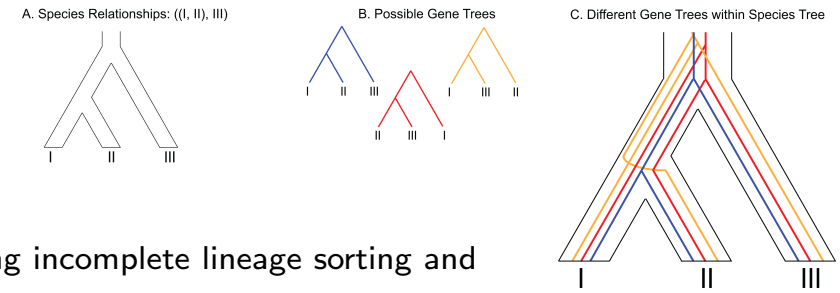
https://www.mrgscience.com/topic-54-cladistics.html, CC BY-NC-SA 3.0

Gene trees versus species trees (Y Ch 3.1.4)

Species tree: phylogeny relating a group of species
Gene tree: phylogeny relating (specific) gene sequences from a group of species
A clade of species can generate many conflicting gene trees due to

1. random errors, limited sequence data
2. lateral gene transfer, esp. near root
3. gene duplications and losses
4. ancestral polymorphisms (diversity within a species) for closely related species

Species tree: genome–wide DNA



M.A. Campbell, T.J. Buser, M.E. Alfaro ME, J.A. López. 2020. Addressing incomplete lineage sorting and paralogy in the inference of uncertain salmonid phylogenetic relationships. PeerJ 8:e9389, doi:10.7717/peerj.9389, CC BY 4.0.

Consensus tree (F Ch 30)

Summarize the information contained in a set of trees all on the same set of species

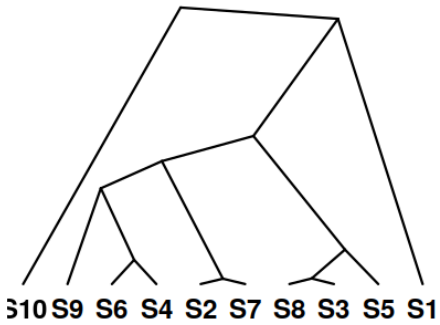
Monophyletic group: tip descendants of an internal node

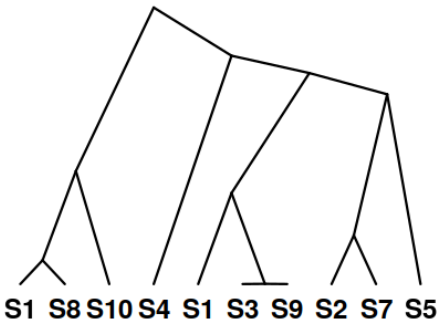
Strict consensus: tree containing monophyletic groups present in all trees

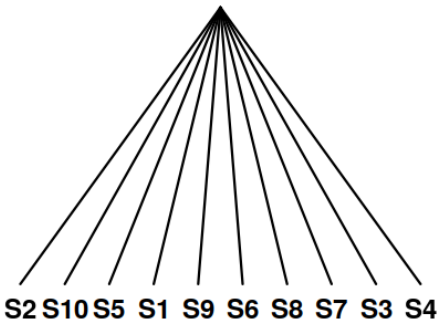
Majority–rule consensus: tree containing monophyletic groups present in the majority of trees

Multiple other rules

Strict consensus (see also F Figs. 30.1–3)

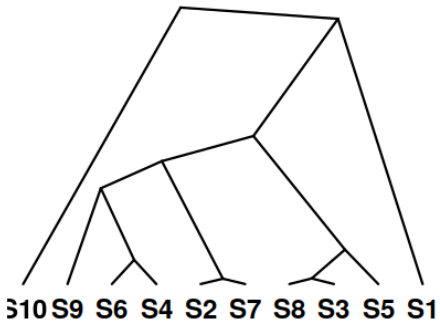


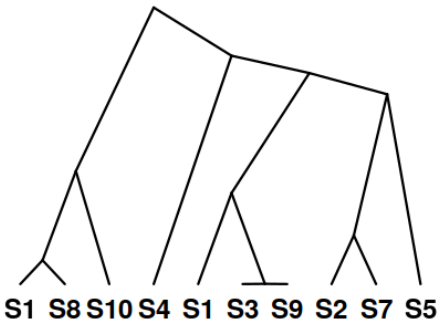


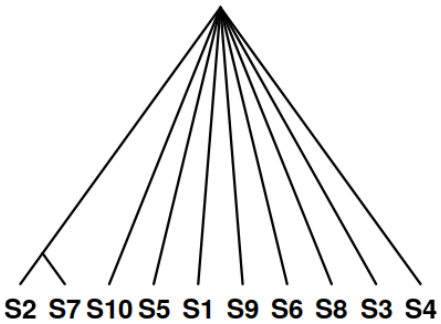



```
ape::consensus(...,p=1,rooted=TRUE)
```

Majority consensus (see also F Figs. 30.1, 30.4)








```
ape::consensus(...,p=0.5,rooted=TRUE)
```

Bayesian phylogenetics

Posterior support for each phylogeny

Consensus tree

Prior assumptions on phylogeny, models, model parameters

MCMC moves between topologies

Constant rate birth–death model

Birth rate: λ, death rate: μ

Single particle dynamics

Step 1 Particle lives for an exponenital($\lambda + \mu$) time

Step 2 With probability $\mu/(\mu + \lambda)$ dies, with probability $\lambda/(\mu + \lambda)$ splits into descendants according to some distribution

can generalize to time dependent rates

Generalized single particle dynamics

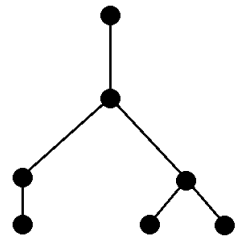
Step 1 Particle lives for a random time

Step 2 Produces offspring ($= 0$ is death) according to some distribution

Characterizing trees: balance indices (F Ch 33)

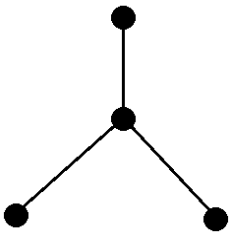
Colless': (binary) sum of balances (absolute difference between number of leaf descendents in left and right node) for each internal node

$$\sum_{v \text{ internal}} |L_v - R_v|$$



Sackin's: sum of distances from root of leaves of tree

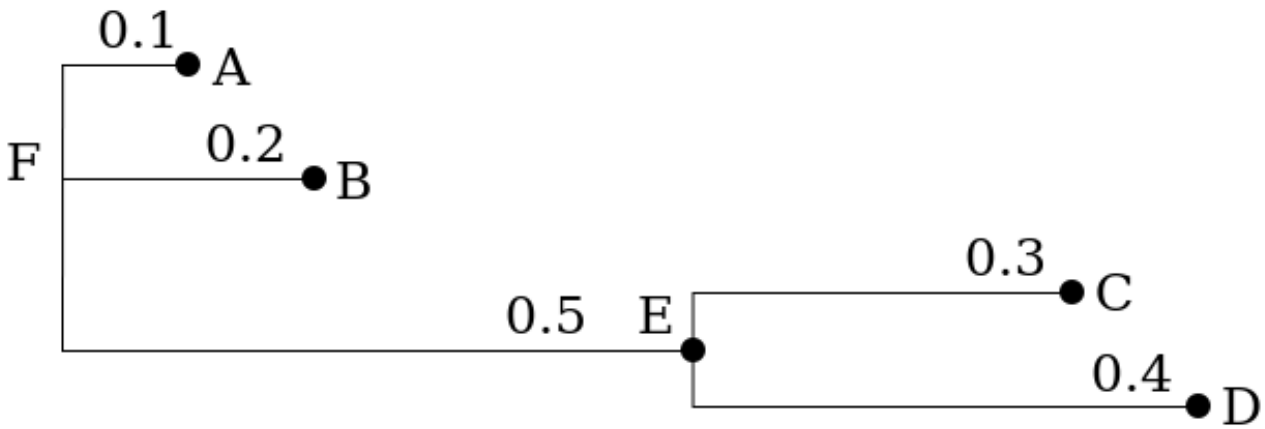
Cophenetic:



$$\sum_{i,j \text{ leaf}} \phi_{ij}, \ \phi_{ij} : \text{distance from root to mrca of } i, \ j$$

Quartet Index: number of B_4 quartets in tree

Newick format



ould be represented in Newick format in several ways

```
(,,(,));                                no nodes are named
(A,B,(C,D));                            leaf nodes are named
(A,B,(C,D)E)F;                          all nodes are named
(:0.1,:0.2,(:0.3,:0.4):0.5);            all but root node have a distance to parent
(:0.1,:0.2,(:0.3,:0.4):0.5):0.0;        all have a distance to parent
(A:0.1,B:0.2,(C:0.3,D:0.4):0.5);        distances and leaf names (popular)
(A:0.1,B:0.2,(C:0.3,D:0.4)E:0.5)F;      distances and all names
((B:0.2,(C:0.3,D:0.4)E:0.5)A:0.1)F;     a tree rooted on a leaf node (rare)
```

https://en.wikipedia.org/wiki/Newick_format by Quantling (Raster: Lee Newberg; Vector: Beao), public domain

Nexus files

```
#NEXUS
Begin data;
Dimensions ntax=4 nchar=15;
Format datatype=dna missing=? gap=-;
Matrix
Species1   atgctagctagctcg
Species2   atgcta??tag-tag
Species3   atgttagctag-tgg
Species4   atgttagctag-tag
;
End;
```

Basic blocks [edit]

TAXA block

The TAXA block contains information about taxa.

DATA block

The DATA block contains the data matrix (e.g. sequence alignment).

TREES block

The TREES block contains phylogenetic trees described using the Newick format, e.g. `((A,B),C);`:

```
#NEXUS
BEGIN TAXA;
  TAXLABELS A B C;
END;

BEGIN TREES;
  TREE tree1 = ((A,B),C);
END;
```

https://en.wikipedia.org/w/index.php?title=Nexus_file&oldid=764543064

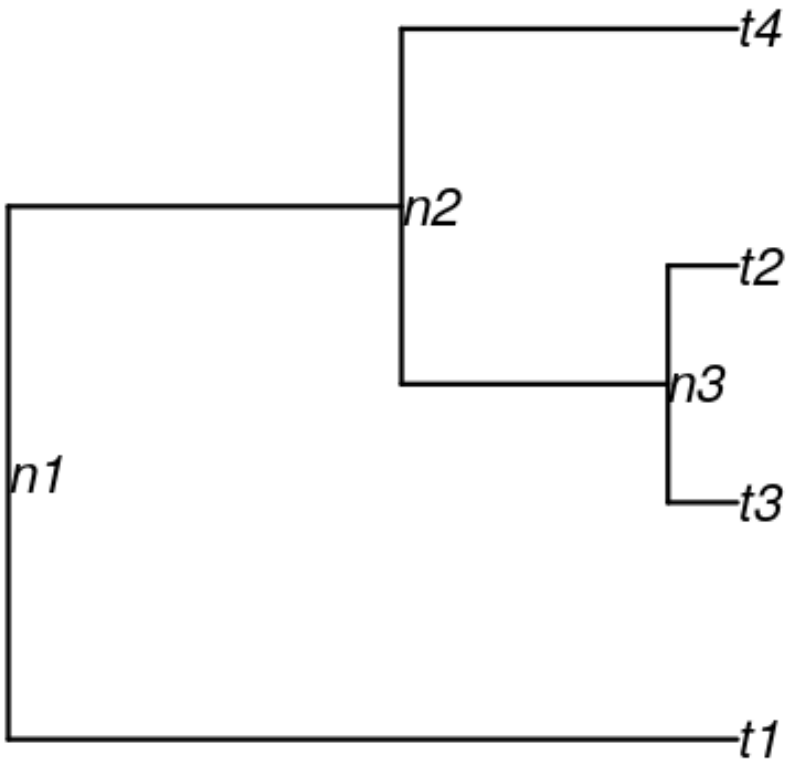
Ape's phylo format

A `list` with four fields (basic)

1. `$edge`: a matrix of edges, first column starting node, second column ending node
2. `$edge.length`: a vector of edge lengths, order as in field edge (optional)
3. `$tip.label`: a vector of tip names
4. `$Nnode`: number of internal nodes

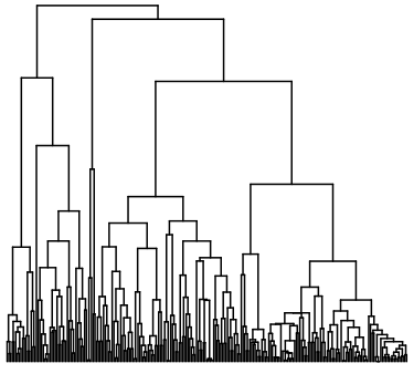
Extra `$node.label`: a vector of internal node names (optional)

Phylo format example

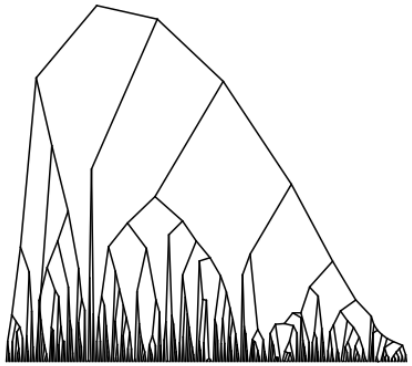



```
> phyltree$edge
     [,1] [,2]
[1,]    5    2
[2,]    5    6
[3,]    6    7
[4,]    7    1
[5,]    7    4
[6,]    6    3
> phyltree$edge.length
[1] 2.0104388 1.0844612 0.7351889 0.1907887 0.1907887 0.9259776
> phyltree$tip.label
[1] "t3" "t1" "t4" "t2"
> phyltree$node.label
[1] "n1" "n2" "n3"
> phyltree$Nnode
[1] 3
```

Graphical presentation (see also HRS Fig 13–2

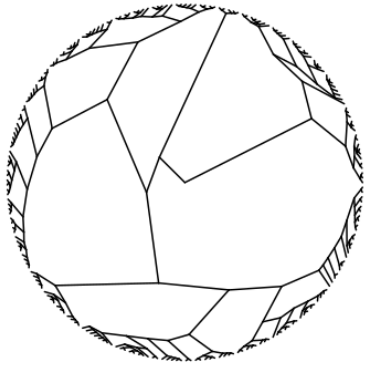


`type="phylogram"` (default)

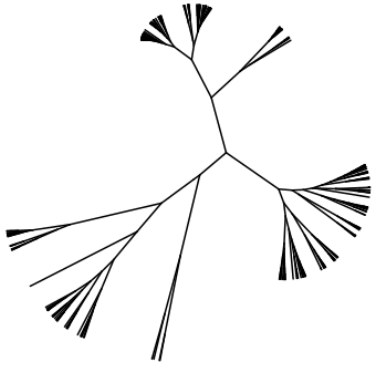


`type="cladogram"`

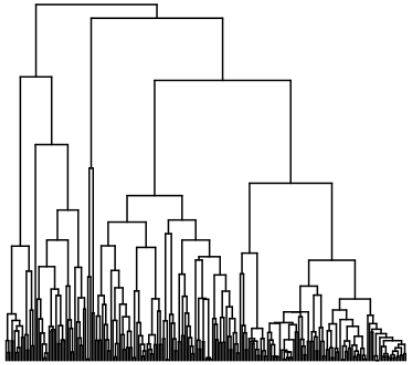
`type="fan"`



`type="radial"`



`type="unrooted"`



`type="tidy"`

primates phylogeny from `geiger` R package, GNU GPL, `ape::plot.phylo(..., type=,...)`

Phylogenetic software (a primer)

TreeView: a very simple treeviewer `taxonomy.zoology.gla.ac.uk/rod/treeview.html`
ClustalW: multiple sequence alignment `www.clustal.org`

PAUP*: Parsimony/ML phyl. inference `paup.phylosolutions.com`

Phylip: phyl. inference by J. Felsenstein

`evolution.genetics.washington.edu/phylip.html`

RAxML: ML phyl. inference `sco.h-its.org/exelixis/web/software/raxml`
PHyML: ML phyl. inference `www.atgc-montpellier.fr/phyml`

MrBayes: Bayesian phyl. inference (F. Ronquist NRM)

`mrbayes.sourceforge.net`

BEAST: Bayesian phyl. inference `beast.community`, `www.beast2.org`

See also

`cran.r-project.org/web/views/Phylogenetics.html`

`evolution.gs.washington.edu/phylip/software.html`

Questions?