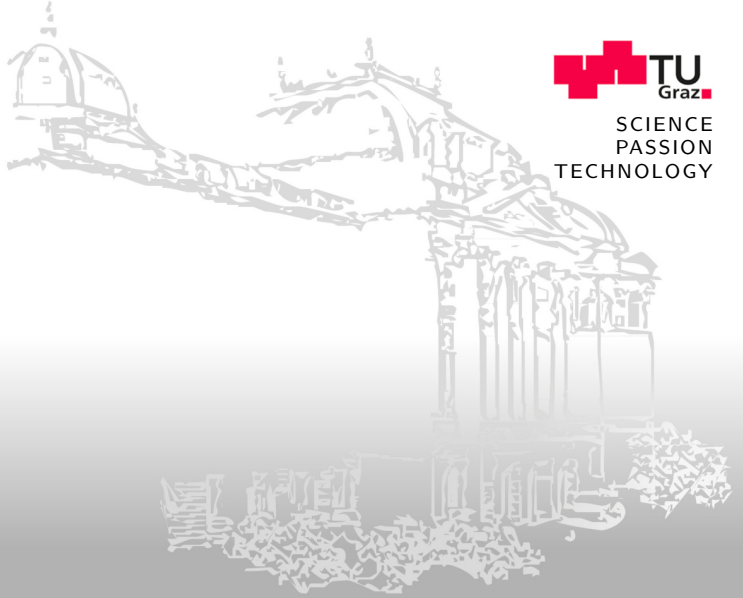


SCIENCE
PASSION
TECHNOLOGY

Phylogentik

Klaus Schliep, PhD

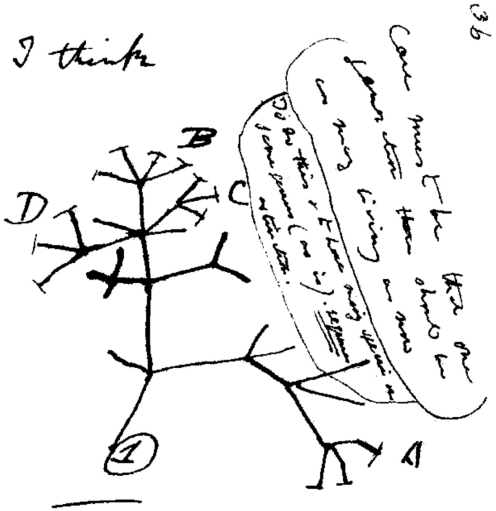
klaus.schliep@tugraz.at

Nothing in Biology Makes Sense Except in the Light of Evolution

Theodosius Dobzhansky

Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Darwin Notebook



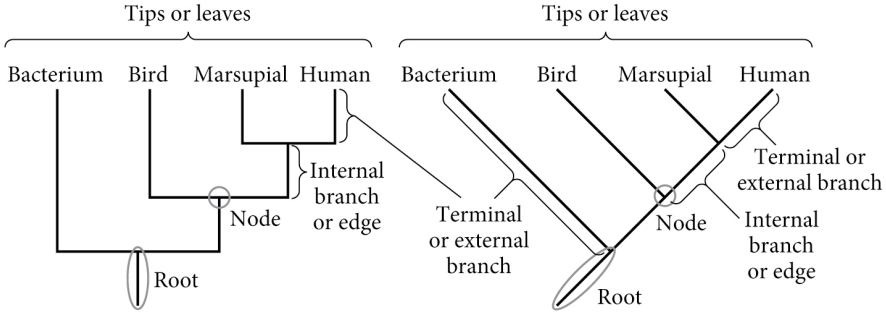
Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Was ist a Phylogenetischer Baum / Phylogenie?

Ein phylogenetischer Baum beschreibt Beziehungen zwischen verschiedenen, verwandten Objekten.
Die *Objekte* können biologische Sequenzen, morphologische Messungen, Arten, Sprachen, Manuscripte, etc. sein

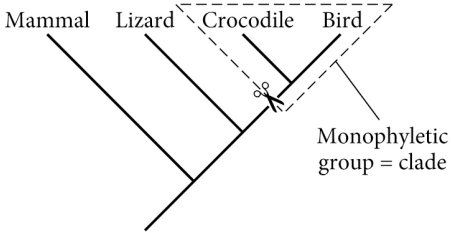
Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

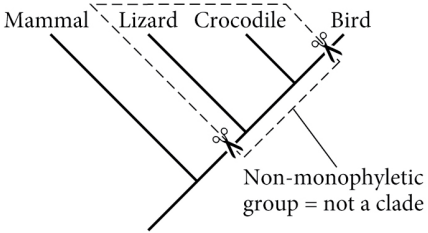
Terminology



Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

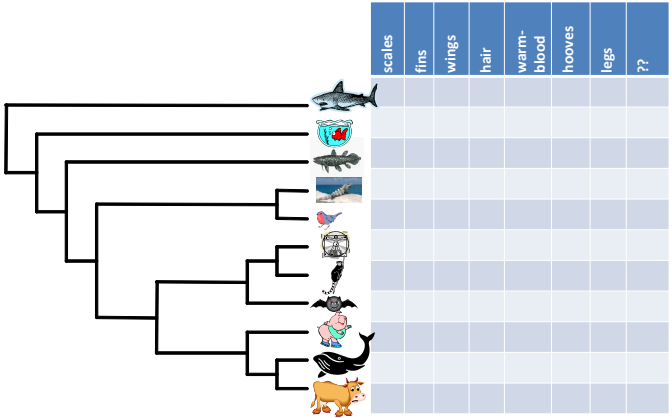
Terminology





Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Excercise

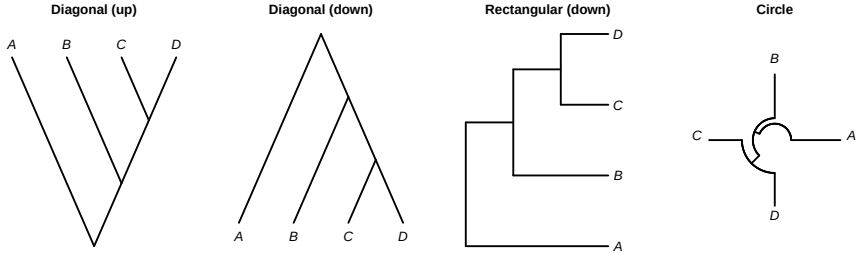


Which characters are monophyletic?

Fill the table! How many changes are for each character necessary?

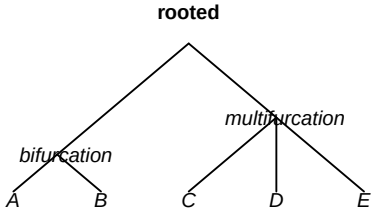
Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

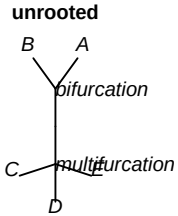
Styles



Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Birfucating trees, Polytomies

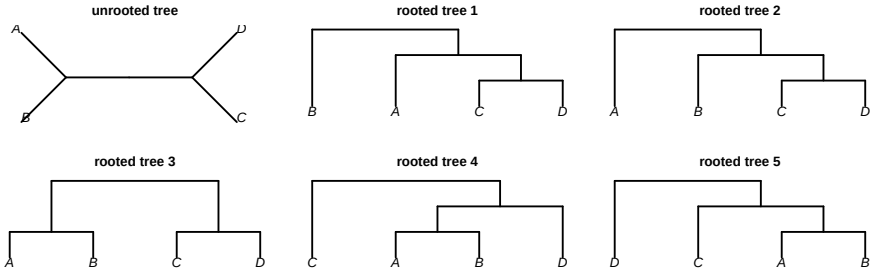




Some software only accepts bifurcating trees, i.e. trees without multifurcations / polytomies.

Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

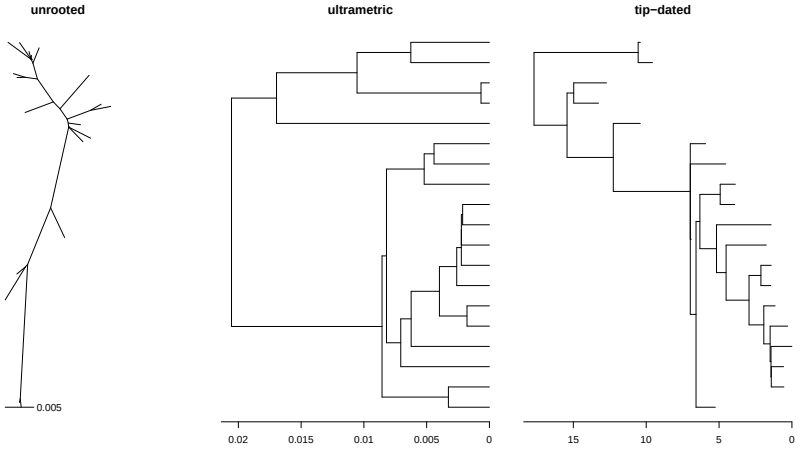
Unrooted trees



Many inference programs return unrooted trees. There are many ways to root a tree (on every internal edge).

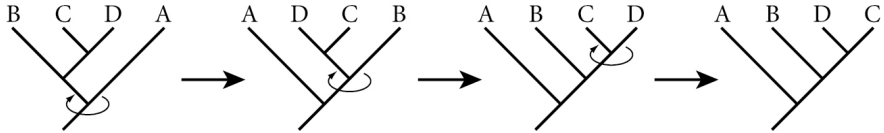
Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Unrooted, ultrametric & tipdated trees



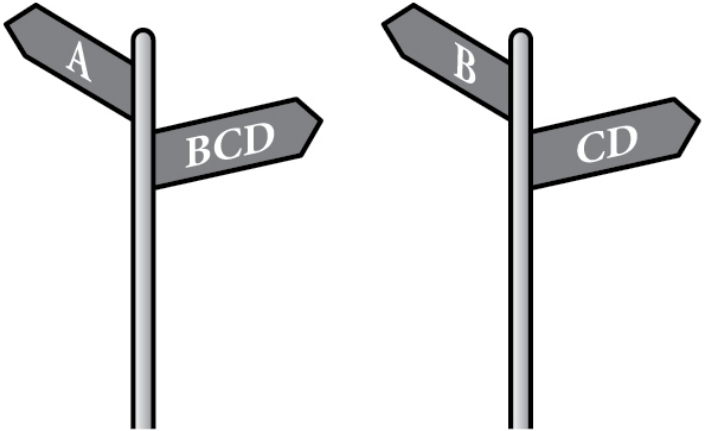
Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Topology



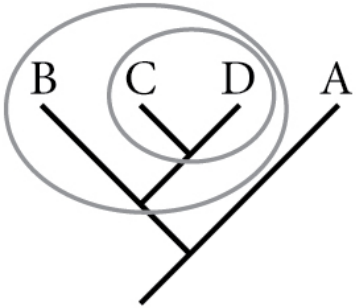
Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

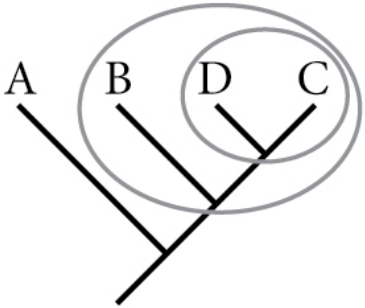
Topology



Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

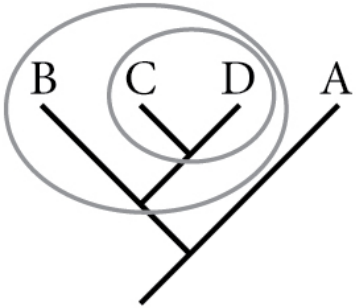
Topology

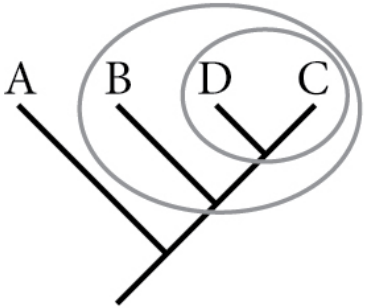




Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Newick






```
( ( B, ( C,   D )), A);
```

```
( A, ( B, ( C, D ) ) );
```

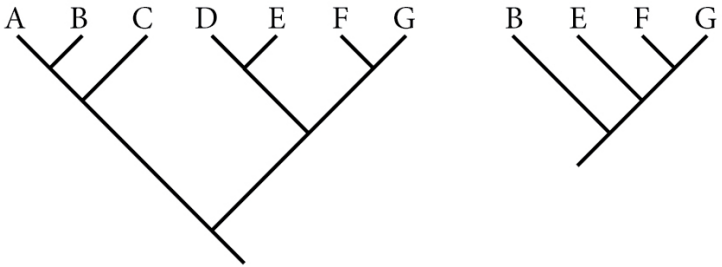
Newick format: Parentheses enclose elements (terminal or internal branches), separated by commas that emerge from the same internal node

Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Newick's lobster house

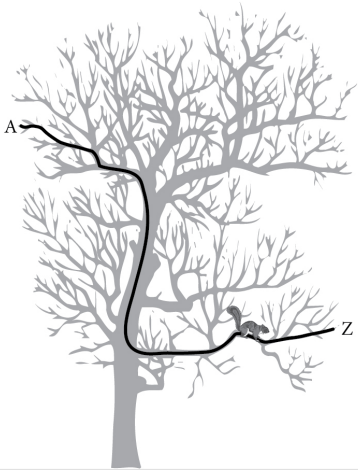
Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Merging and Pruning



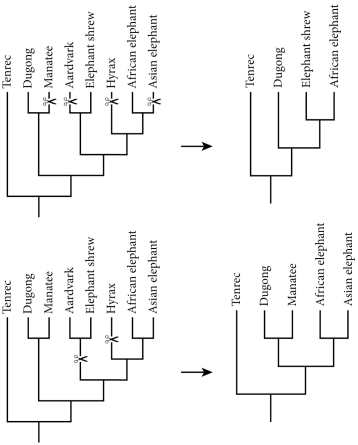
Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Merging and Pruning



Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Merging and Pruning



Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Online Quiz

Die Webseite https://klash.shinyapps.io/tree_thinking/ enthält einige kleine Quizzes.

Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

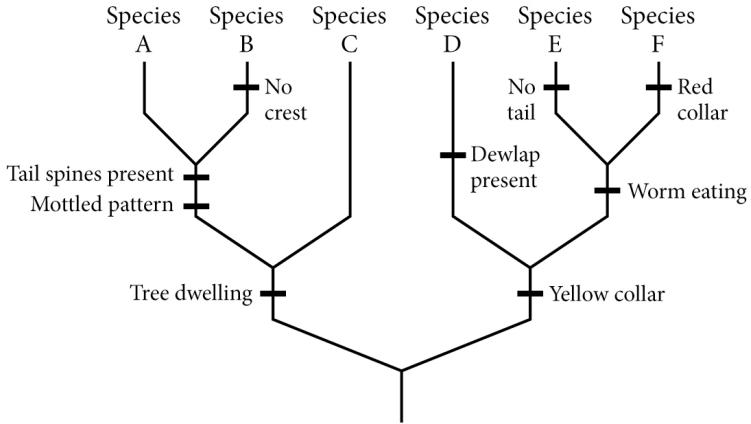
Character Evolution

TABLE 4.1 Characters and character states in lizards

	Character	Ancestral state	Derive state
	Crest on head	Absent	Present
	Colored collar	Absent	Present
	Preferred prey	Insects	Worms
	Pattern on back	Stripes	Mottled
	Tail spines	Absent	Present
	Habitat	Ground dwelling	Tree dwelling
	Tail spots	Present	Absent
	Dewlap (flap of skin under chin)	Absent	Present

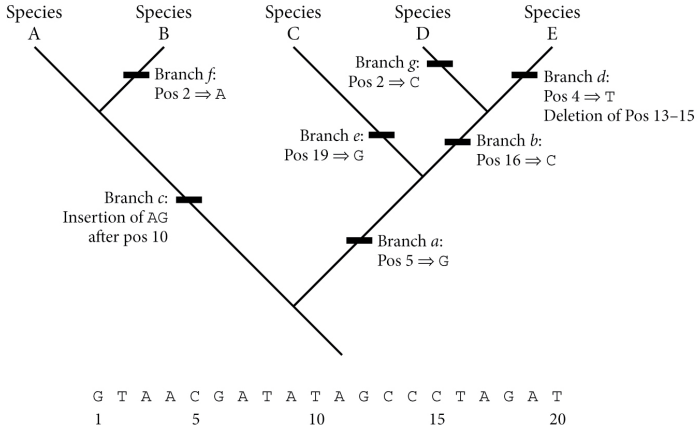
Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Character Evolution



Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Character Evolution (DNA)



Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Software for tree visualisation

- Dendroscope https://www.wsi.uni-tuebingen.de/lehrstuehle/algorithms-in-bioinformatics/software/dendroscope/
- TreeView https://code.google.com/archive/p/treeviewx/
- figtree https://github.com/rambaut/figtree/releases
- R packages: ape https://CRAN.R-project.org/package=ape, phytools https://CRAN.R-project.org/package=phytools & ggtree https://bioconductor.org/packages/ggtree/
- Python: ETE http://etetoolkit.org/
- Online: phylo.io http://phylo.io/

Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Schätzung phylogenetetischer Bäume

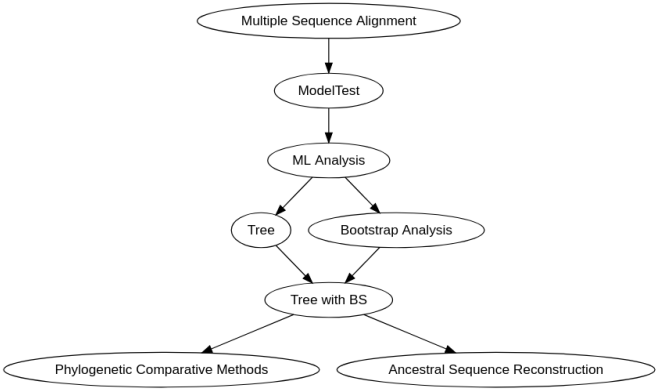
Methoden für die Inferenz von Phylognien falle in 3 Gruppen

- Distanz Methoden, z.B. UPGMA, Neighbor Joining (NJ)
  Vorteil: schnell, Nachteil: keine Aussagen über vorherige Zustände (ancestral state) möglich.
- Parsimony
  Nur für nah verwandte Sequenzen empfehlenswert.
  ("long branch attraction", "Felsenstein zone")
- Maximum Likelihood und Bayesianische (MCMC) Methoden
  Probabilistische und statistisch etabliert, aber teilweise sehr langsam

ML und Bayesianische Methoden werden heutzutage bevorzugt.

Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Usual Workflow (Maximum Likelihood)



Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

A tree for the number of trees

# species unrooted trees	# of trees	# species rooted trees
3	1	2
4	3	3
5	15	4
6	105	5
7	945	6
8	10,395	7
9	135,135	8
10	2,027,025	9
11	34,459,425	10
12	654,729,075	11
13	13,749,310,575	12
14	316,234,143,225	13
15	7,905,853,580,625	14
50	$2.753 \cdot 10^{76}$	49

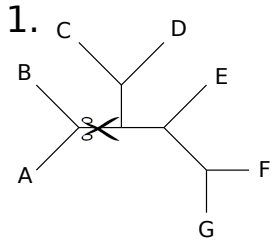
Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Tree rearrangements

- The strategy of evaluating the maximum parsimony criterion for all trees (brute force) in order to find the best tree topology is in most cases highly impracticable.
- Instead, (local) tree rearrangements are used to search locally within the tree space. The idea behind such a heuristic is to use a starting tree and search locally for improved scores (parsimony, maximum likelihood, Least-Squares), until no further rearrangements can lead to a tree with a better score.

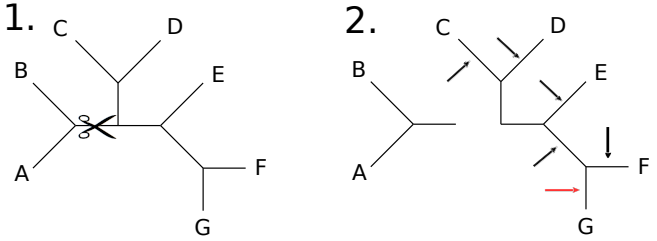
Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Subtree pruning and regrafting



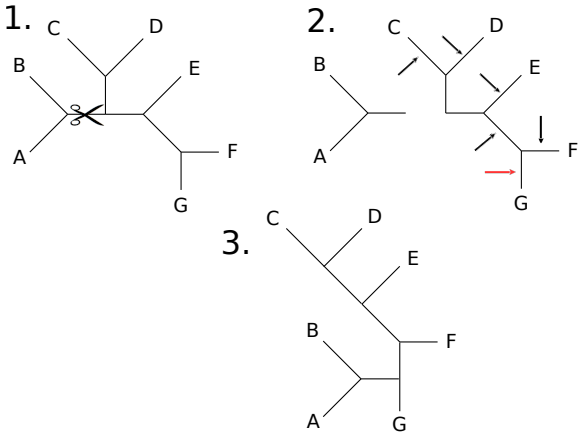
Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Subtree pruning and regrafting



Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Subtree pruning and regrafting



Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

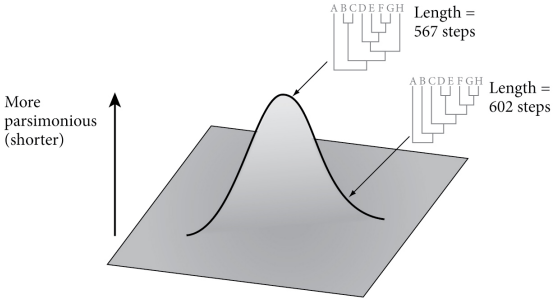
Tree space

- Tree rearrangements only guarantee to find a local optimum!
- To better explore the tree space often the search with the rearrangements is started from different starting trees. These can be random trees, random addition trees or for example the current best trees which was itself pertubated with several tree rearrangements.

Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Tree space

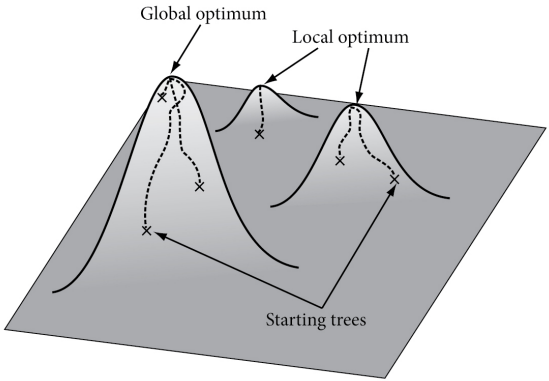
Visualization of tree space with the best (shortest) trees sit at the peak



Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Tree space

Searching the tree space with multple optima.



Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Substitution models

The K80, K2P, or Kimura two parameter model.
This model has a different rate for transitions (C$\longleftrightarrow$T and A$\longleftrightarrow$G) and transversions (all other substitutions).

$$\begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{pmatrix} -(\alpha+2\beta) & \alpha & \beta & \beta \\ \alpha & -(\alpha+2\beta) & \beta & \beta \\ \beta & \beta & -(\alpha+2\beta) & \alpha \\ \beta & \beta & \alpha & -(\alpha+2\beta) \end{pmatrix} \end{array}$$

Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Common substitution models

Model:	Summary
JC69 (Jukes & Cantor 1969)	Equal rates; equal nucleotide frequencies.
K80 (Kimura 1980)	Different rates for transitions and transversions.
F81 (Felsenstein 1981)	Equal rates; different nucleotide frequencies.
HKY85 (Hasegawa et al. 1985)	Different rate for transitions and transversions; unequal base frequencies.
F84 (Felsenstein 1984)	Similar to HKY85 but with two additional rates (for pyrimidines and purines).
TN93 (Tamura & Nei 1993)	Two types of transversions; one type of transition; unequal base frequencies.
GTR (several refs.)	All rates different; all base frequencies different.

Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Model Selection

Use programs like ModelTest, ProtTest to choose the best fitting substitution model!

	Model	df	logLik	AIC	AICw	AICc	AICcw	BIC
1	JC	91.00	-54303.67	108789.35	0.00	108794.77	0.00	109341.20
2	JC+I	92.00	-50672.85	101529.71	0.00	101535.25	0.00	102087.63
3	JC+G	92.00	-48684.10	97552.19	0.00	97557.74	0.00	98110.11
4	JC+G+I	93.00	-48588.86	97363.73	0.00	97369.39	0.00	97927.71
5	F81	94.00	-54212.64	108613.27	0.00	108619.06	0.00	109183.32
6	F81+I	95.00	-50548.97	101287.94	0.00	101293.86	0.00	101864.05
7	F81+G	95.00	-48500.49	97190.99	0.00	97196.90	0.00	97767.10
8	F81+G+I	96.00	-48401.46	96994.92	0.00	97000.96	0.00	97577.10
9	HKY	95.00	-51275.86	102741.72	0.00	102747.64	0.00	103317.83
10	HKY+I	96.00	-47450.80	95093.59	0.00	95099.64	0.00	95675.77
11	HKY+G	96.00	-44893.04	89978.08	0.00	89984.13	0.00	90560.26
12	HKY+G+I	97.00	-44762.63	89719.27	0.00	89725.44	0.00	90307.51
13	GTR	99.00	-50758.41	101714.83	0.00	101721.26	0.00	102315.20
14	GTR+I	100.00	-47079.80	94359.60	0.00	94366.16	0.00	94966.03
15	GTR+G	100.00	-44746.72	89693.44	0.00	89700.00	0.00	90299.87

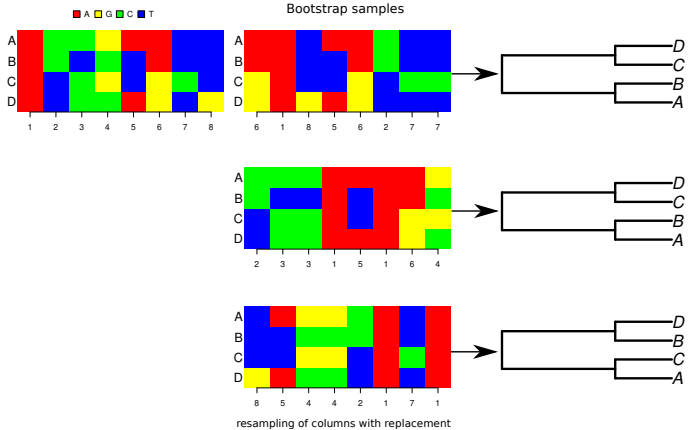
Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Bootstrap

General principle of the bootstrap:

1. Estimate a parameter from the data: $\hat{\theta}$
2. Resample with replacement the data.
3. Estimate the parameter $\hat{\theta}^*$ for this “bootstrap” sample.
4. Repeat steps 2 and 3 many times.
5. Assess the distribution of the $\hat{\theta}^*$'s: this gives an estimate of the error of $\hat{\theta}$ This procedure "mimicks" the process of sampling repeatedly a distribution, and makes no assumption on the shape of this distribution.

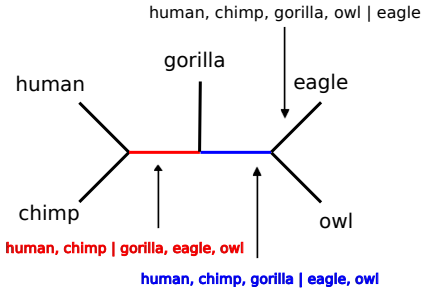
Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at



Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

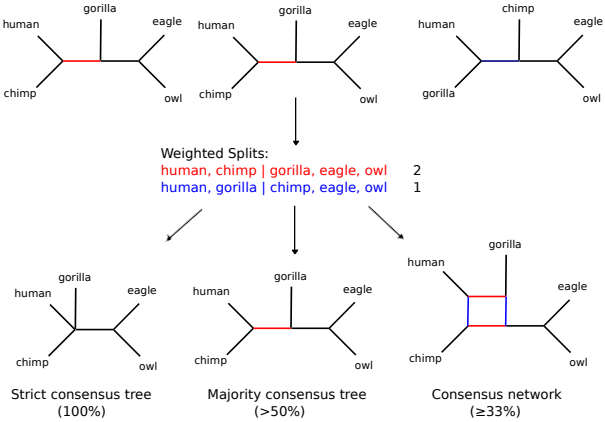
Splits (Definition)

- A split is a bipartition of the taxa (labels, species) into two sets
- A bipartition of one taxa vs. the rest is known as a trivial split
- A split corresponds to an edge in a phylogenetic tree, removing the edge from a tree creates two sub-trees



Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Consensus Trees



Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Software

General software for estimating distance an ML trees

- PHYLIP https://evolution.genetics.washington.edu/phylip.html
- Paup* https://paup.phylosolutions.com/
- R package phangorn https://CRAN.R-project.org/package=phangorn

Fast and more specialized Likelihood tools:

- RAxML https://cme.h-its.org/exelixis/web/software/raxml/
- iqtree http://www.iqtree.org/

Bayesian phylogentic software:

- RevBayes https://revbayes.github.io/
- BEAST2 https://paup.phylosolutions.com/

Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at

Übung

- Erstellen Sie ein Alignment aus den Sequenzen des COX1 Gen cats_dogs_COX1.fas mit *clusta_lomega*, *mafft* oder *muscle5*.
- Erstellen Sie mit Hilfe dieses Alignments mit iqtree (`http://iqtree.cibiv.univie.ac.at/`) einen phylogenetischen Baum. Der Einfachheit wegen bleiben wir bei den Standardeinstellungen.
- Welches phylogenitische Model wurde ausgewählt?
- Sind die Gruppen (Hunde und Katzen) monophyletisch?
- Benutze figtree im Terminal um ein schönere Abbildung mit Bootstrapwerten zu erstellen.

Klaus Schliep, PhD, UBT
klaus.schliep@tugraz.at