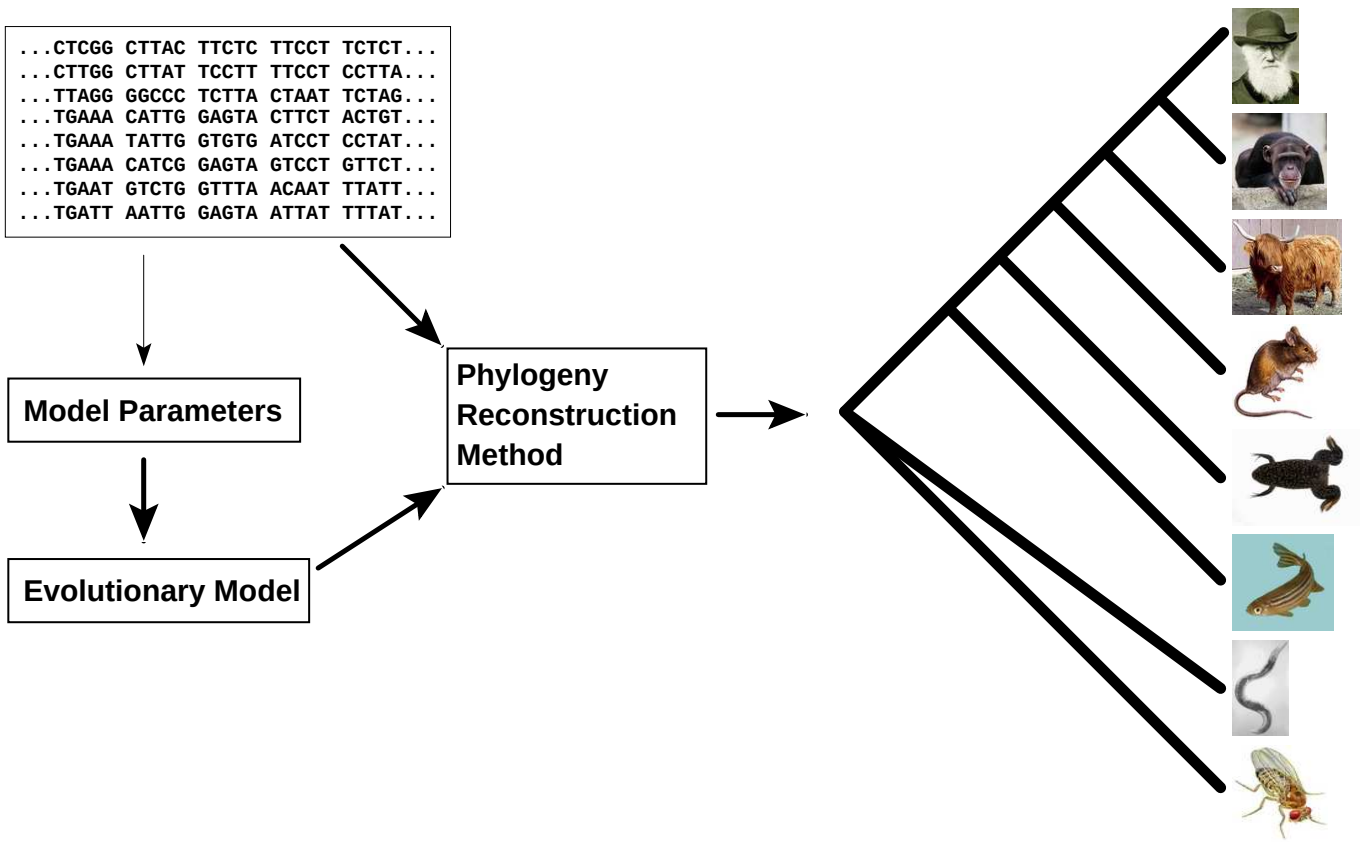
Introductory Workshop to ML Tree Reconstruction

Heiko A. Schmidt

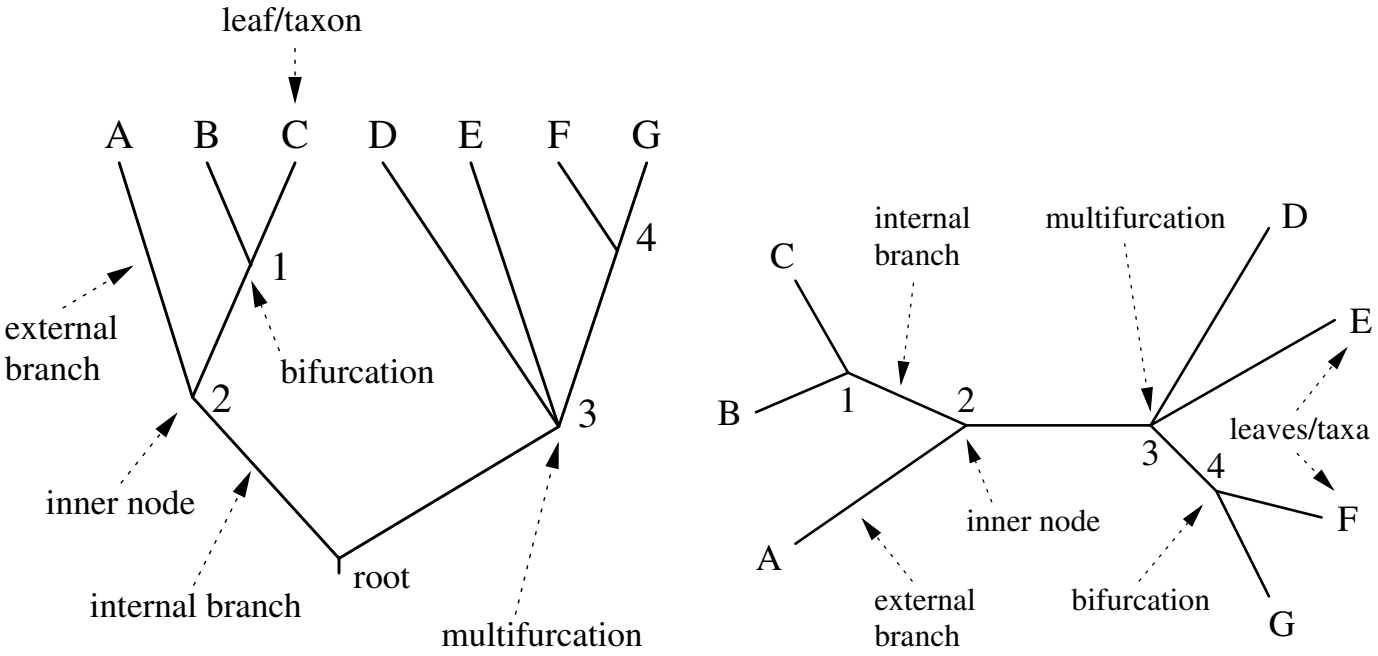
Center for Integrative Bioinformatics Vienna (CIBIV)
Max F. Perutz Laboratories (MFPL)
Vienna, Austria
heiko.schmidt@univie.ac.at

April 2008

Recap: Phylogenetic Reconstruction



Some Notation



Main Types of Phylogenetic Methods

Data	Method	Evaluation Criterion
Characters (Alignment)	**Maximum Parsimony**	Parsimony
	Statistical Approaches: Likelihood, Bayesian	Evolutionary Models
Distances	**Distance Methods**	

Introduction: ML on Coin Tossing

Given a box with 3 coins of different fairness $\left(\frac{1}{3}, \frac{1}{2}, \frac{2}{3} \text{ heads}\right)$

We take out one coin an toss 20 times:

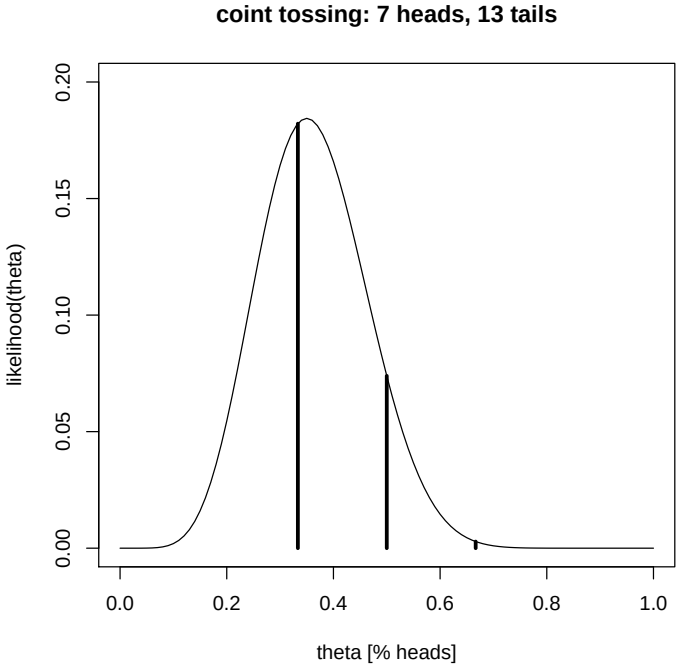
$$H, T, T, H, H, T, T, T, T, H, T, T, H, T, H, T, T, H, T, T$$

	Probability		**Likelihood**
$p(k \text{ heads in } n \text{ tosses}\|\theta)$	$\equiv$	$L(\theta\|k \text{ heads in } n \text{ tosses})$	
	$=$	$\binom{n}{k}\theta^k(1-\theta)^{n-k}$	
		(here binomial distribution)	

Aim: The ML approach searches for that parameter set θ for the generating process which maximizes the probability of our given data.

Hence, *"likelihood flips the probability around."*

Introduction: ML on Coin Tossing (Estimate)



Three coin case

$$L(\theta|7 \text{ heads in } 20) = \binom{20}{7}\theta^7(1-\theta)^{13}$$

for each coin $\theta \in \left\{\frac{1}{3}, \frac{1}{2}, \frac{2}{3}\right\}$

For infinitely many coins
$\theta = (0...1)$

ML estimate: $L(\hat{\theta}) = 0.1844$ where coin shows $\hat{\theta} = 0.35$ heads

From Coins to Phylogenies?

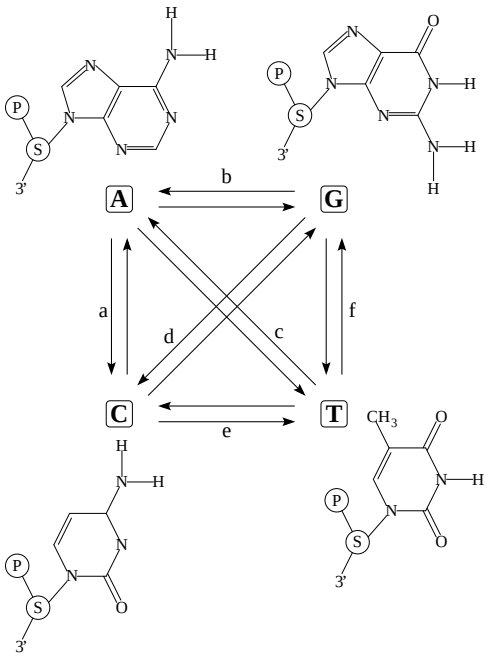
While the coin tossing example might look easy, in phylogenetic analysis, the parameter (set) θ comprises:

- evolutionary model
- its parameters
- tree topology
- its branch lengths

That means, a high dimensional optimization problem.
Hence, some parameters are often estimated/set separately.

Substitution Models

Evolutionary models are often described using a substitution rate matrix R and character frequencies Π. Here, 4×4 matrix for DNA models:



$$R = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} - & a & b & c \\ a & - & d & e \\ b & d & - & f \\ c & e & f & - \end{pmatrix} \end{array}$$

$$\Pi = (\pi_A, \pi_C, \pi_G, \pi_T)$$

From R and Π we reconstruct a substitution probability matrix P, where $P_{ij}(t)$ is the probability of changing $i \longrightarrow j$ in time t.

Modeling Evolution

- Evolution is usually modeled as a

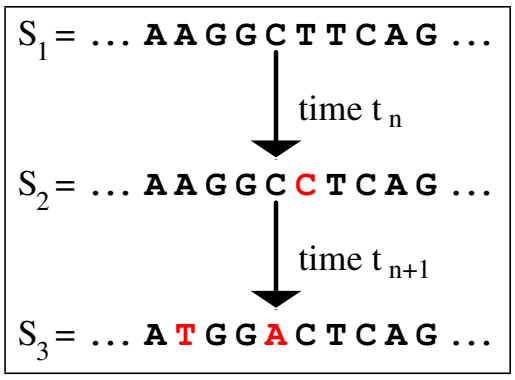
  stationary, time-reversible Markov process.

- What does that mean?

Assumptions on Evolution

Markov Process

The (evolutionary) process evolves without memory, i.e. sequence S_2 mutates to S_3 during time t_{n+1} independent of state of S_1.



Assumptions on Evolution

Stationary:
The overall character frequencies π_j of the nucleotides or amino acids are in an equilibrium and remain constant.

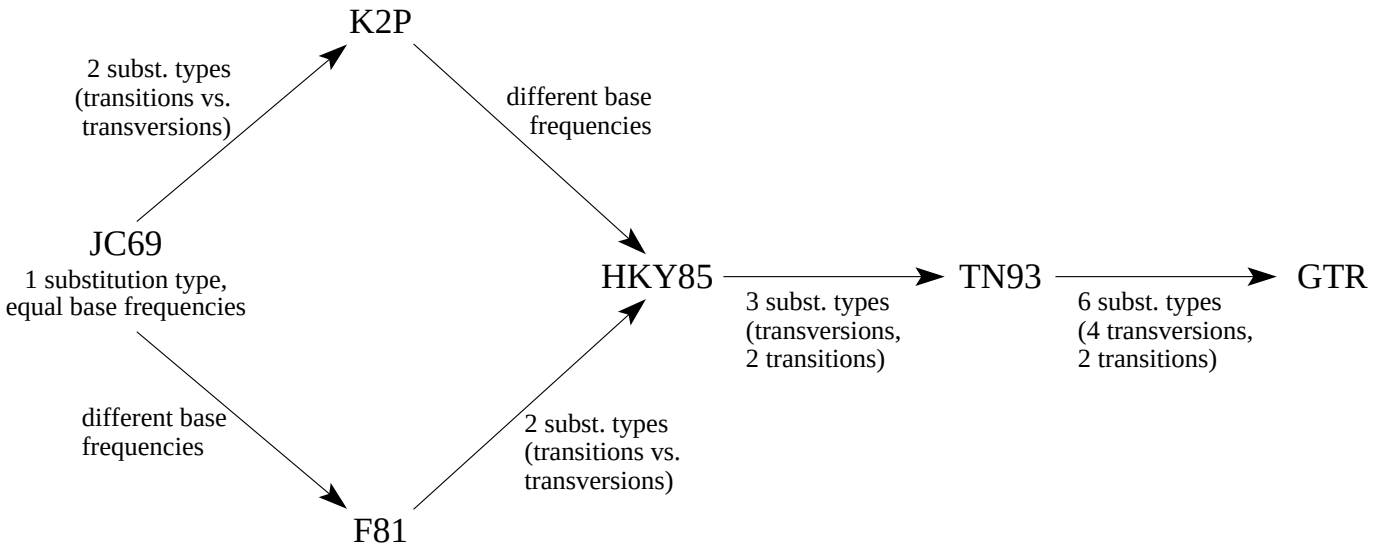
Time-Reversible:
Mutations in either direction are equally likely

$$\pi_i \cdot P_{ij}(t) = P_{ji}(t) \cdot \pi_j$$

This means a mutation is as likely as its back mutation.

$$P(i \rightarrow j) = P(i \leftarrow j) \qquad \text{(JC69)}$$

DNA models



Further modification:
rate heterogeneity: invariant sites, Γ-distributed rates, mixed.

Protein Models

Generally this is the same for protein sequences, but with 20×20 matrices. Some protein models are:

- Poisson model ("JC69" for proteins, rarely used)
- Dayhoff (Dayhoff *et al.*, 1978, general matrix)
- JTT (Jones *et al.*, 1992, general matrix)
- WAG (Whelan & Goldman, 2000, more distant sequences)
- VT (Müller & Vingron, 2000, distant sequences)
- mtREV (Adachi & Hasegawa, 1996, mitochondrial sequences)
- cpREV (Adachi *et al.*, 2000, cloroplast sequences)
- mtMAM (Yang *et al.*, 1998, Mammalian mitochondria)
- mtART (Abascal *et al.*, 2007, Arthropod mitochondria)
- rtREV (Dimmic *et al.*, 2002, reverse transcriptases)
- ...
- BLOSUM 62 (Henikoff & Henikoff, 1992) $\rightarrow$ database searching

Computing ML Distances Using $\mathbf{P_{ij}(t)}$

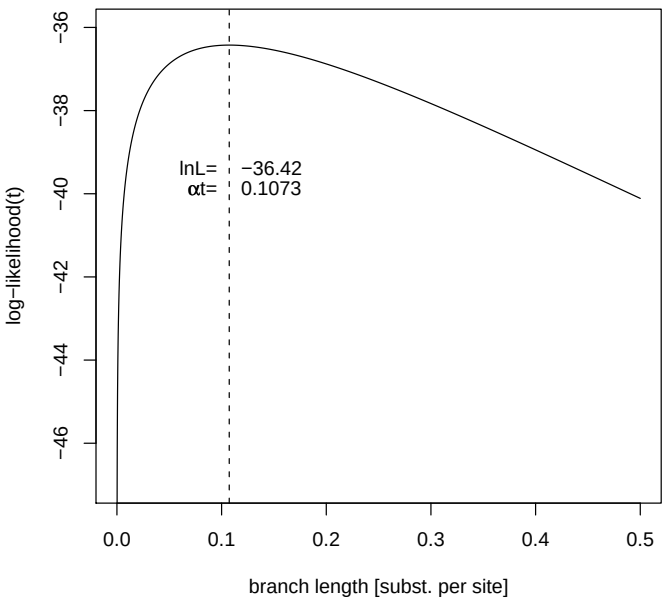
The Likelihood of sequence s evolving to s' in time t:

$$L(t|s \rightarrow s') = \prod_{i=1}^{m} \left(\Pi(s_i) \cdot P_{s_i s'_i}(t) \right)$$

Likelihood surface for two sequences under JC69:

GATCCTGAGAGAAATAAAC $= s$
GGTCCTGACAGAAATAAAC $= s'$

Note: we do not compute the probability of the distance t but that of the data $D = \{s, s'\}$.



Computing Likelihood Values for Trees

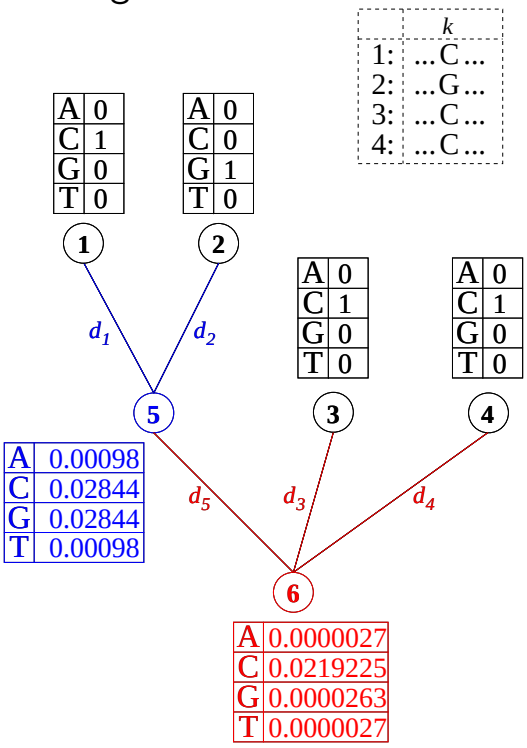
Given a tree with branch lengths and sequences for all nodes, the computation of likelihood values for trees is straight forward.
Unfortunately, we usually have no sequences for the inner nodes (ancestral sequences).
Hence we have to evaluate every possible labeling at the inner nodes:

$$L\left(\begin{smallmatrix}\mathrm{C}&&\mathrm{C}\\&\diagdown\diagup&\\\mathrm{G}&&\mathrm{C}\end{smallmatrix}\right) = L\left(\begin{smallmatrix}\mathrm{C}&&\mathrm{C}\\&\diagdown\ \mathrm{A}\ \mathrm{A}\ \diagup&\\\mathrm{G}&&\mathrm{C}\end{smallmatrix}\right) + L\left(\begin{smallmatrix}\mathrm{C}&&\mathrm{C}\\&\diagdown\ \mathrm{A}\ \mathrm{C}\ \diagup&\\\mathrm{G}&&\mathrm{C}\end{smallmatrix}\right) + \cdots + L\left(\begin{smallmatrix}\mathrm{C}&&\mathrm{C}\\&\diagdown\ \mathrm{G}\ \mathrm{C}\ \diagup&\\\mathrm{G}&&\mathrm{C}\end{smallmatrix}\right) + \cdots + L\left(\begin{smallmatrix}\mathrm{C}&&\mathrm{C}\\&\diagdown\ \mathrm{T}\ \mathrm{T}\ \diagup&\\\mathrm{G}&&\mathrm{C}\end{smallmatrix}\right)$$

for every column in the alignment. . . but there is a fast algorithm.

Likelihoods of Trees (Single alignment column, given tree)

For a single alignment column and a given tree:



Likelihoods of nucleotides i at inner nodes:

$$L_5(i) = [P_{iC}(d_1) \cdot L(C)] \cdot [P_{iG}(d_2) \cdot L(G)]$$

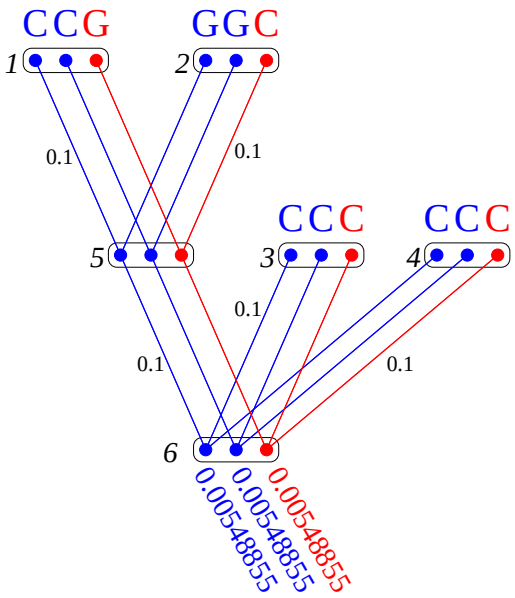
$$L_6(i) = \prod_{v=\{2,3,4\}} \left[\sum_{j=\{ACGT\}} P_{ij}(d_v) \cdot L_v(j) \right]$$

Site-Likelihood of an alignment column k:

$$L^{(k)} = \sum_{i=\{ACGT\}} \pi_i \cdot L_6(i) = 0.005489$$

with all $d_x = 0.1$ and $P_{ij}(0.1) = \begin{cases} .91 & i \neq j \\ .03 & i = j \end{cases}$ (JC)

Likelihoods of Trees (multiple columns)



Considering this tree with $n = 3$ sequences of length $m = 3$ the tree likelihood of this tree is

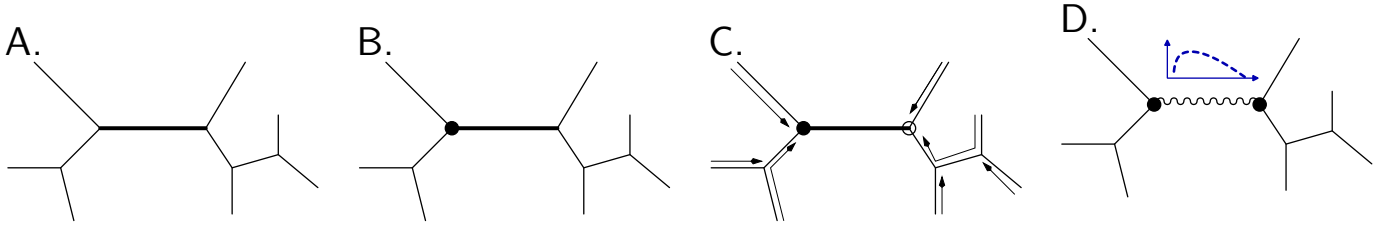
$$\mathcal{L}(T) = \prod_{k=1}^{m} L^{(k)} = 0.005489^2 \cdot 0.005489$$
$$= 0.0000001653381$$

or the log-likelihood

$$\ln \mathcal{L}(T) = \sum_{k=1}^{m} \ln L^{(k)} = -15.61527$$

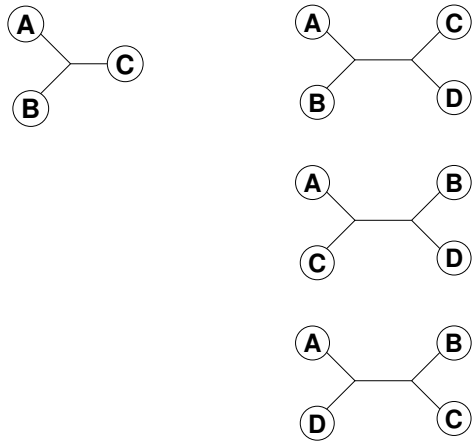
Adjusting Branch Lengths Step-By-Step

To compute optimal branch lengths do the following. Initialize the branch lengths.
Choose a branch (A.). Move the virtual root to an adjacent node (B.). Compute all partial likelihoods recursively (C.). Adjust the branch length to maximize the likelihood value (D.).



Repeat this for every branch until no better likelihood is gained.

Number of Trees to Examine...

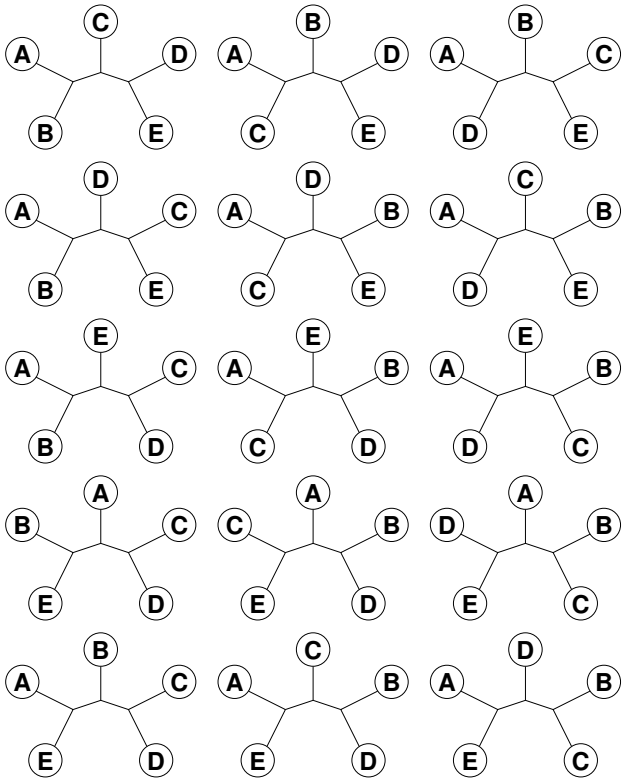


$$B(n) = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

$$B(10) = 2027025$$

$$B(55) = 2.98 \cdot 10^{84}$$

$$B(100) = 1.70 \cdot 10^{182}$$



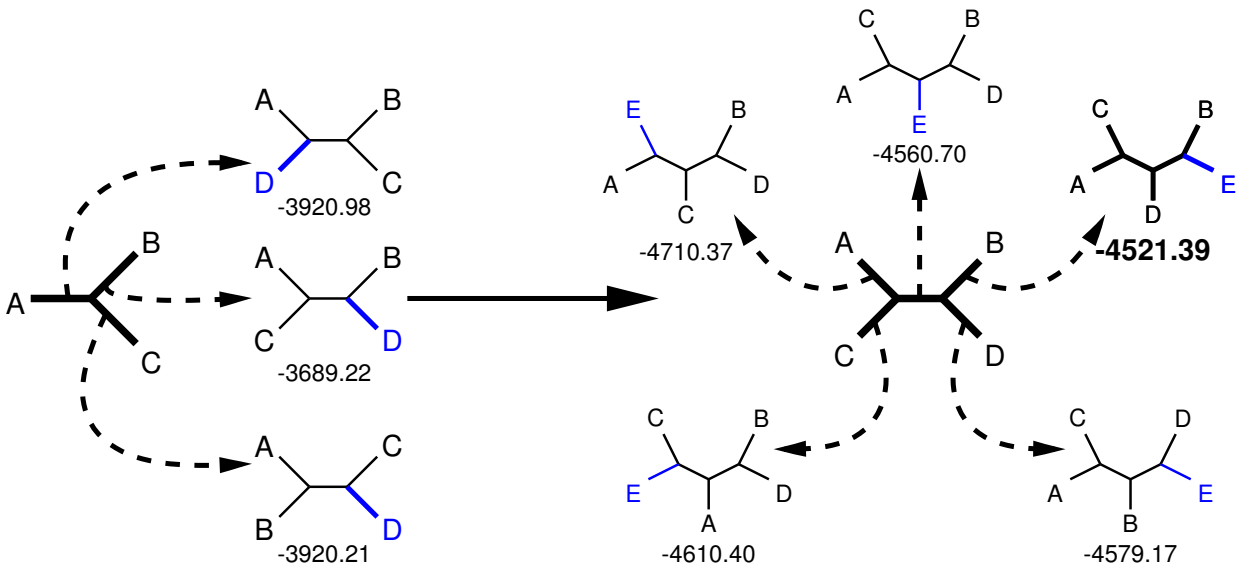
Finding the ML Tree

Exhaustive Search: guarantees to find the optimal tree, because all trees are evaluated, but not feasible for more than 10-12 taxa.

Branch and Bound: guarantees to find the optimal tree, without searching certain parts of the tree space – can run on more sequences, but often not for current-day datasets.

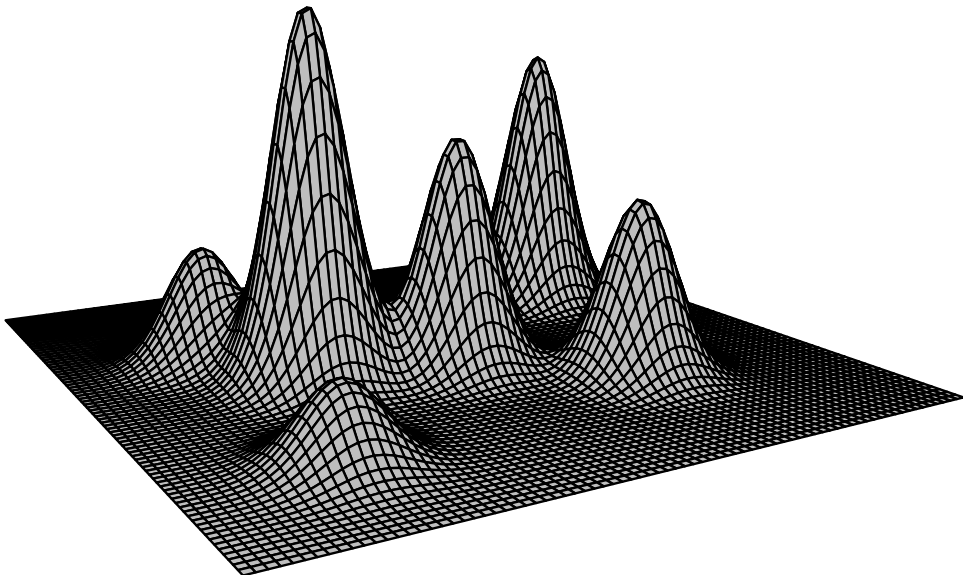
Heuristics: cannot guarantee to find the optimal tree, but are at least able to analyze large datasets.

Build up a tree: Stepwise Insertion



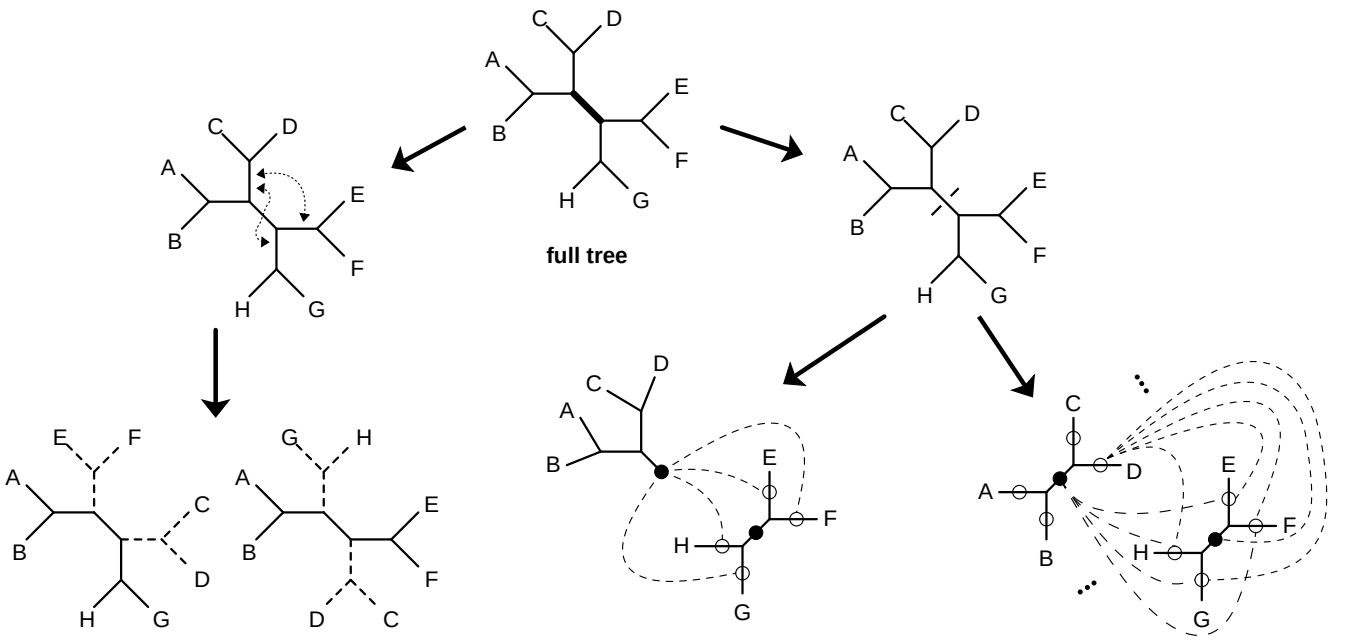
Local Maxima

What if we have multiple maxima in the likelihood surface?



Tree rearrangements to escape local maxima.

Tree Rearrangements: Scanning a Tree's Neighborhood



Search Strategy of IQPNNI

Concept: BioNJ tree + randomization + fastNNI

1. Start with (fast) BioNJ tree.
2. Do fastNNIs to optimize trees, i.e., evaluate all NNIs simultaneously and then accept all best ones which are non-conflicting. (after first round, identical to PHYML).
3. Remove randomly a certain amount of taxa and re-insert them by a fast and rough quartet-based method. (some randomization)
4. Repeat (2)-(3) until stop criterion is met.

Pro: Can evade local optima,
offers automatic stopping criterion,
hints when search didn't run enough,
numerically optimized ML computation,
offers codon models

Con: slower than PhyML/RAxML

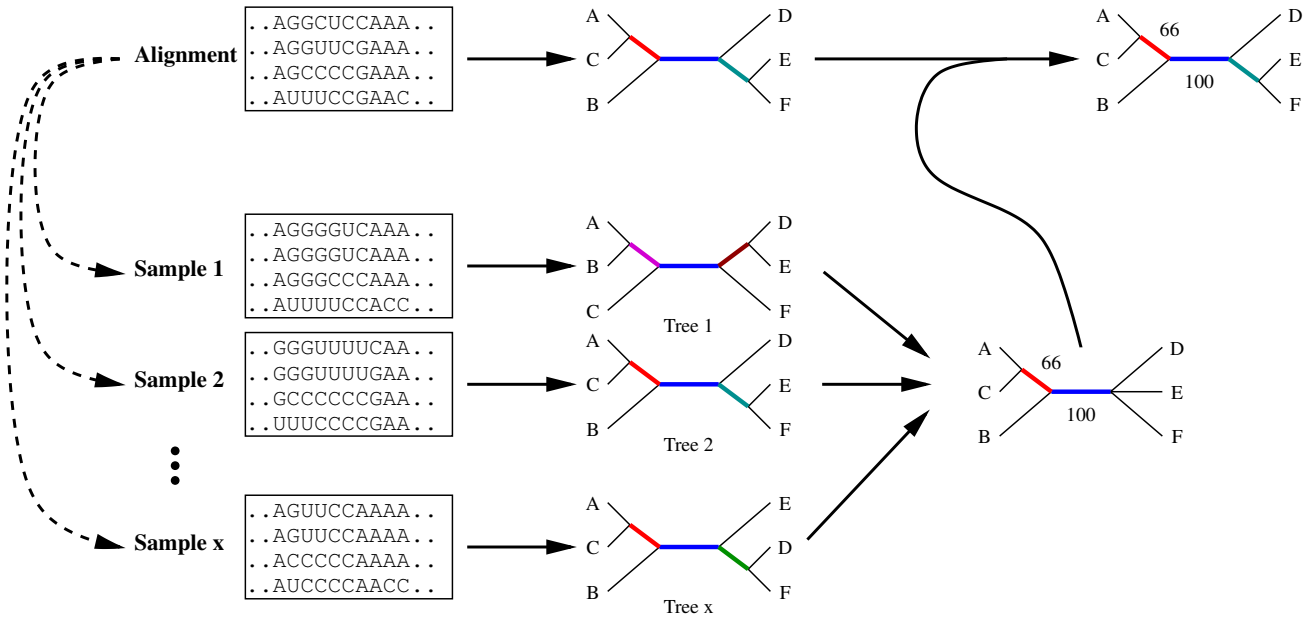
ML-based programs:

- IQPNNI
- RAxML
- PhyML
- GARLI
- TREE-PUZZLE
- dnaml (PHYLIP)
- fastDNAml
- MetaPiga
- SSA
- nucml, protml (MOLPHY)
- `http://evolution.genetics.washington.edu/phylip/software.html`

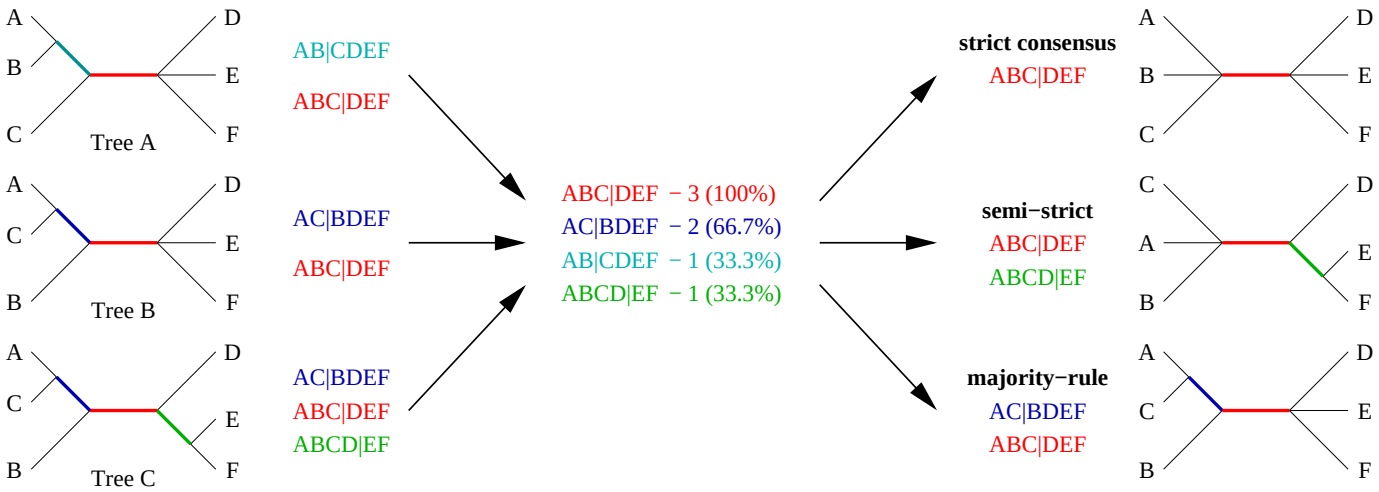
How reliable is the reconstructed tree:

- Usually programs deliver a single tree, but without confidence values for the subtrees.
- How can we assess reliability for the subtree?

Estimating Confidence: The Bootstrap



Summarizing Trees: Consensus Methods



Summarizing Trees: Consensus Methods

- Majority-based: (Sorted splits added in descending order)
  - Strict consensus: all splits found in all trees
  - Semi-Strict consensus: all splits uncontradicted in all trees
  - Majority Rule Consensus M_ℓ: all splits found in more than fraction ℓ of the trees (typically $\ell = 0.5$).
  - Relative Majority Consensus: all splits even below 0.5 down to the first incongruence.
  - Majority Rule extended (MRe): incompatible splits are discarded and all added that are compatible with incorporated splits.
- Adams consensus: reflects common nestings (hard to interpret)

Quartet Puzzling

The Quartet Puzzling algorithm implemented in the TREE-PUZZLE program is a three step procedure:

maximum-likelihood step: compute ML trees for all quartets of an alignment.

puzzling step: compose intermediate tree from quartet trees (this is done multiple times).

consensus step: construct a majority rule consensus tree from the intermediate trees and evaluate the branch lengths.

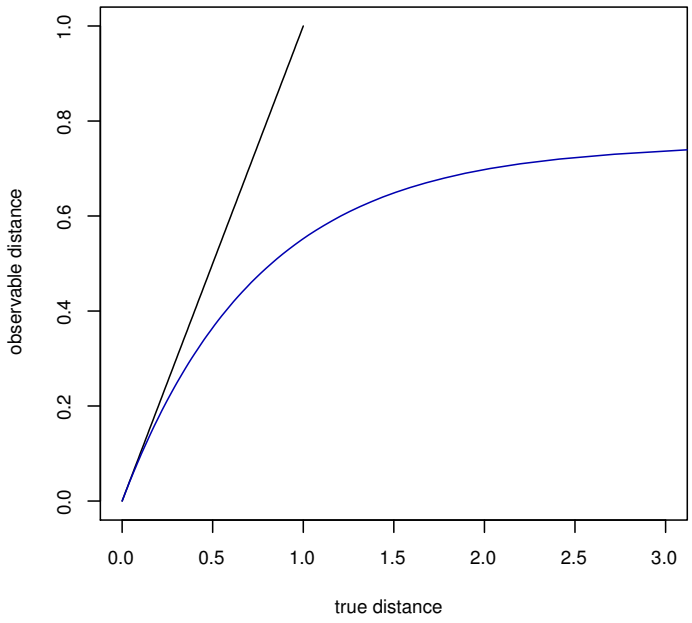
Phylogenetic Information

The information about the true tree, might be obscured or unextractable from an alignment due to

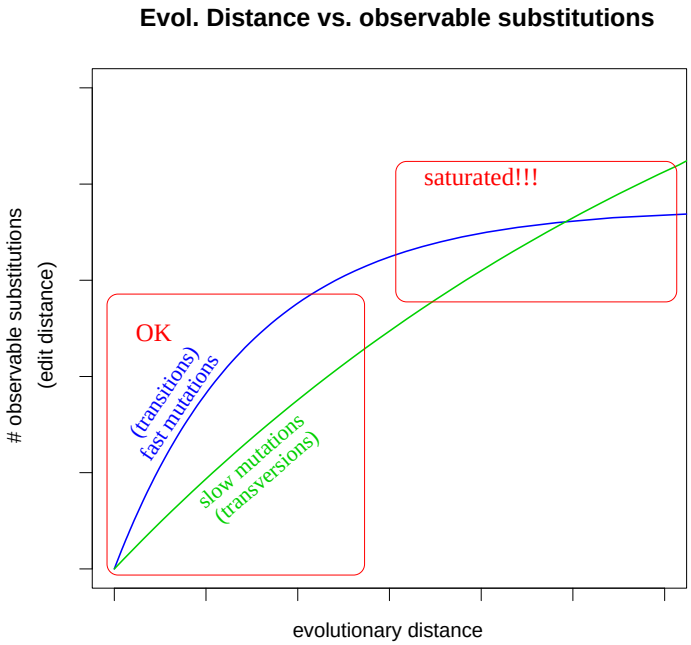
- too similar sequences
  (no differences $\longrightarrow$ no information)
- sequences are to divergent
  (saturated sequences $\longrightarrow$ information drowned in noise)

Are there ways to check for this?

Reminder: Jukes-Canter Correction for Multiple Hits



Plotting Substitutions vs. Distance for DNA



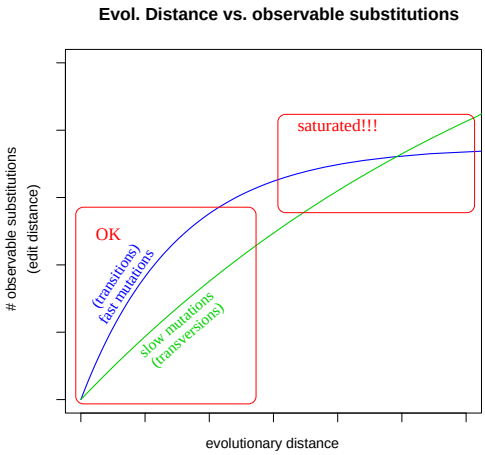
Transitions (ts) usually occur much more often the transversions (tv).
Thus, the ts-curve rises faster but reaches the plateau earlier.

The tv-curve can only 'overtake' the ts-curve if the latter is quite saturated!

Saturation Plots for DNA

Saturation Plots can be created as follows

- Take every pair of sequences
  - Count the number of observable substitutions (e.g., transitions, transversions)
  - Compute the distances of the sequence pair (e.g., with ML)
- . . . and plot the evolutionary distance (x-axis) against the observed substitutions (y-axis) for each class of mutations.



Software for DNA Saturation Plots

Saturation Plots can be created using

- **Windows:** DAMBE (Xia and Xie, 2001)

```
→ Graphics menu
→ Transition and transversion vs. divergence
```

- **All OS:** TREE-PUZZLE with the `-wtstv` option, plotting the data in `*.tstv` with a few lines in the `R` program (`www.r-project.org`):

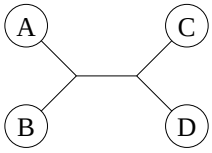
```
tstvtab = read.table("ali.phy.tstv", header=T)  # read data
attach(tstv)                                    # use headers as names
pdf(file="tstv.pdf")                            # open PDF file
maxsubst=max(ts,tv)                             # find maximum
plot(distance,ts,col=2,ylab="observed substitutions",ylim=c(0,maxsubst))
points(distance,tv,col=3)                       # plot
dev.off()                                       # close PDF file
detach(tstvtab)                                 # release names
q()                                             # quit R program
```

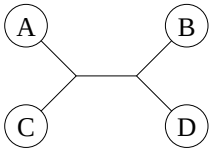

Likelihood Weights, Posterior Prob., and Empirical Bayes

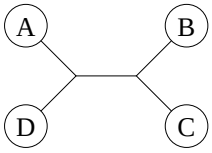
- We can compute a likelihood value for a tree based given an alignment and model... (cf. the *lecture on ML methods*).
- Problem: How different are the likelihoods?
  Just from the value of likelihoods one often cannot tell whether they are significantly different.
- Normalization: Posterior probabilities are computed:

$$p_i = \frac{L_i}{\sum_n L_n}$$

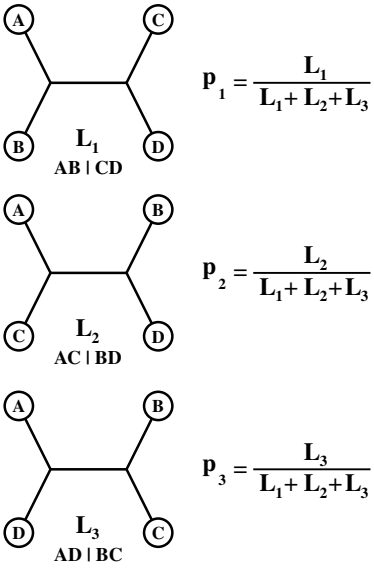
- We can use that on the three different quartet topologies to assess phylogenetic information in our data.







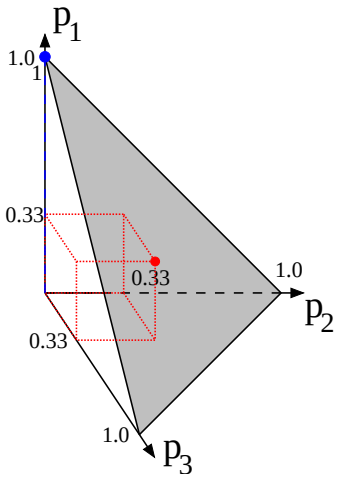
Plotting Posteriors: Likelihood Mapping

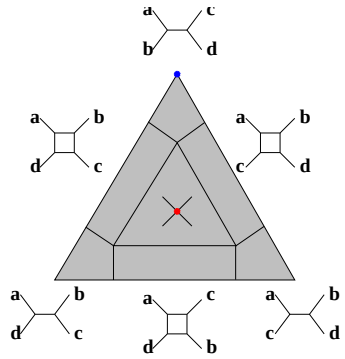


$$p_1 = \frac{L_1}{L_1 + L_2 + L_3}$$

$$p_2 = \frac{L_2}{L_1 + L_2 + L_3}$$

$$p_3 = \frac{L_3}{L_1 + L_2 + L_3}$$

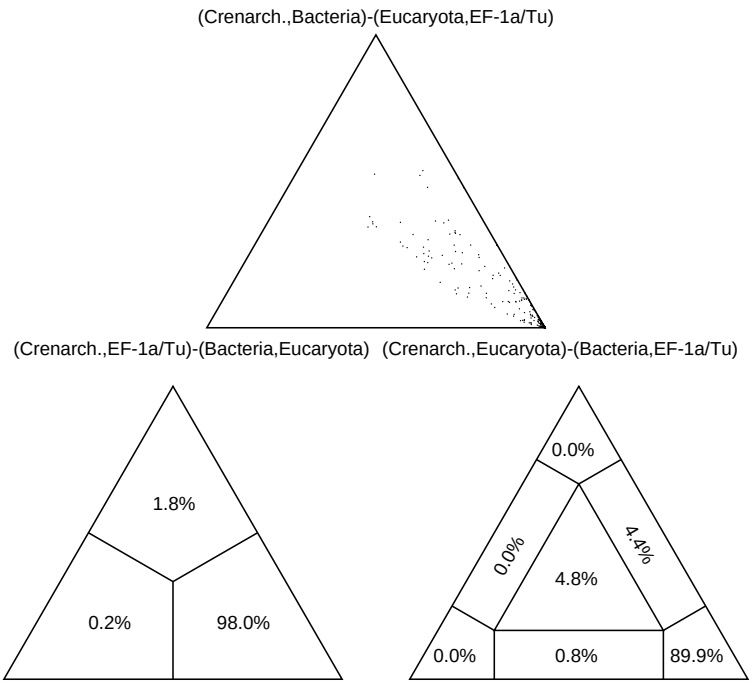




Since $p_1 + p_2 + p_3 = 1$, 3D points (p_1, p_2, p_3) fall into a triangular (simplex).

If we repeat this for all quartets (or a large random subset) in a dataset we can assess the amount of phylogenetic signal in the dataset.

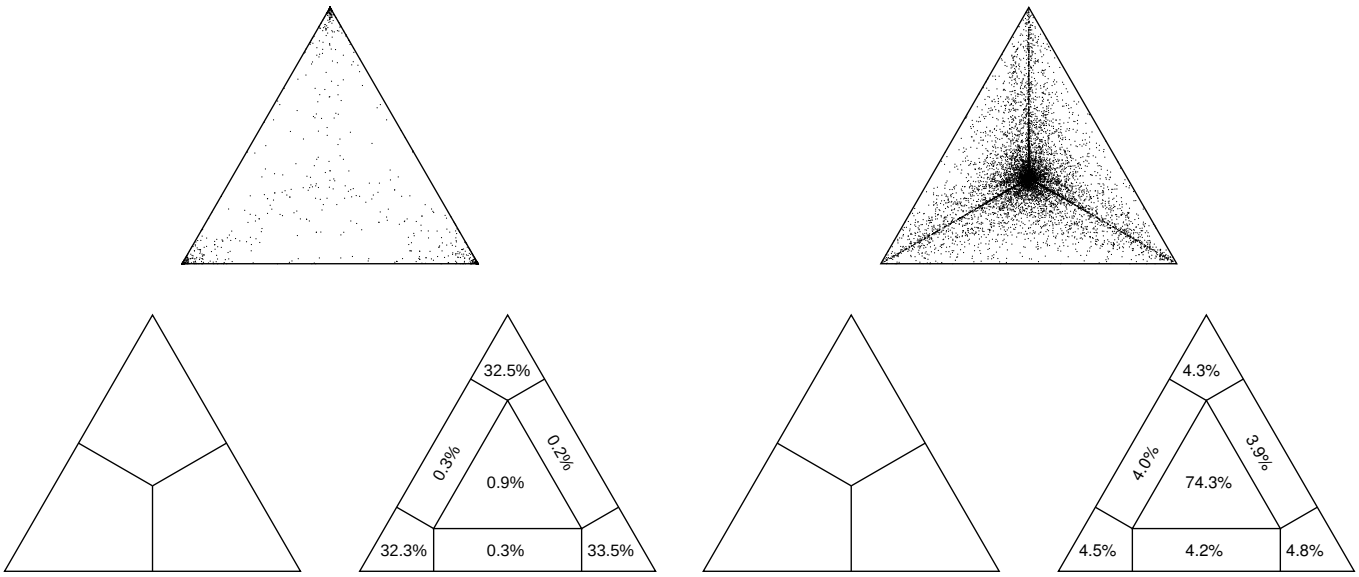
Likelihood Mapping (Cluster Analysis)



The Simplex Plot can visualize the relationship among (4) sets of taxa.

The taxa/sequences are assigned to four sets (A,B,C,D) one for each leaf of a quartet tree.

Likelihood Mapping (Information Content)



The Simplex Plot can also visualize the information content in an alignment.

By not assigning taxa to clusters, four are chosen randomly for each leaf. We have to add the percentages in the corners (resolved) or rectangles (partly resolved), respectively. Center means unresolved.

Exercises:

the exercises can be found at

`http://www.cibiv.at/~hschmidt/ML-short-course-2008`