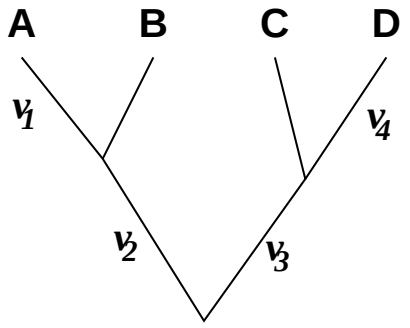
Distance-based methods

seq 1	A	G	C	T	T	A	C	C	T	G	T	T	A	C	T
seq 2	C	G	T	A	A	A	T	T	T	C	C	C	G	A	T
seq 3	C	G	C	A	A	G	T	T	T	C	C	C	G	A	T
seq 4	C	A	C	T	T	A	T	T	A	G	T	C	A	A	C

$$\Downarrow (d_{ij})_{i,j=1,\ldots,4}$$

	seq 1	seq 2	seq 3	seq 4
seq 1	0	11	11	8
seq 2	11	0	2	10
seq 3	11	2	0	9
seq 4	8	10	9	0

Distance Methods: Aim



Aim: Find branch lengths v_b such that the sum of the branch lengths connecting any two leaves gets close to the measured distances between all pairs of leaves. That is, for instance

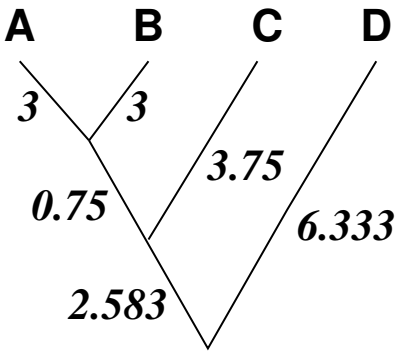
$$d_{A,D}^{measured} = v_1 + v_2 + v_3 + v_4$$

Distance Methods: UPGMA

One possibility are clustering methods like UPGMA = Unweighted Pair Group Methods using Arithmetic means.

	A	B	C	D
A	0	6	7	13
B	6	0	8	14
C	7	8	0	11
D	13	14	11	0

$\Rightarrow$

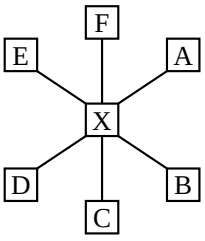


Note:

- In the reconstructed rooted tree all sequences are equally far away from the root.
- If the distance matrix does not comply to this, UPGMA will likely reconstruct the wrong tree.

Distance Methods: Neighbor Joining (NJ)

A widely used distance method is Neighbor-Joining:



1 begin with a star tree with N leaves:

2 For each taxon i compute the net divergence r

$$r_i = \sum_{k=1}^{N} d_{ik}$$

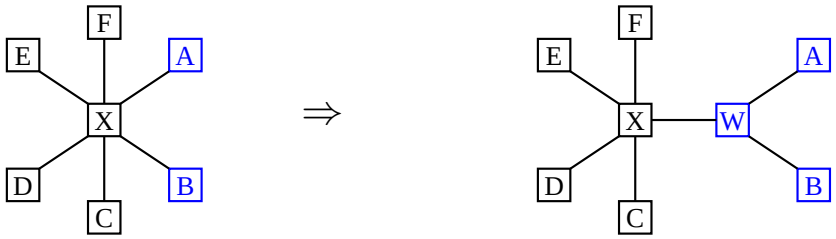
For all pairs of taxa i, j, compute rate-corrected distances

$$M_{ij} = d_{ij} - \frac{r_i + r_j}{N - 2}$$

3 Choose the pair (A, B) that minimizes this equation.

Distance Methods: Neighbor Joining (NJ)

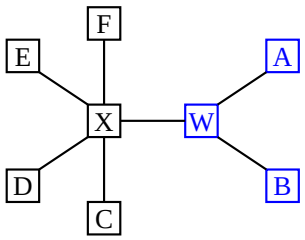
4 cluster (A, B) and define an interior node W representing them:



5 compute the branch lengths for the external edges:

$$\begin{aligned} v_{AW} &= \frac{1}{2}\left(d_{AB} + \frac{r_A - r_B}{N-2}\right) \\ v_{BW} &= d_{AB} - v_{AW}. \end{aligned}$$

Distance Methods: Neighbor Joining (NJ)



6 compute the distances of W to the remaining $N-2$ leaves k:

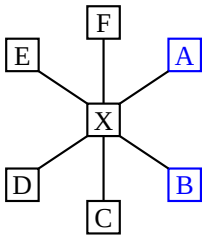
$$d_{Wk} = \frac{1}{2}\left(d_{Ak} + d_{Bk} - d_{AB}\right))$$

7 continue at step 2 with the reduced set of leaves

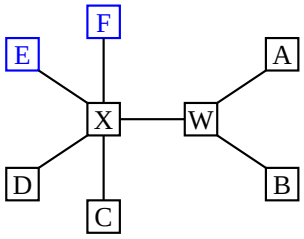
Distance Methods: The NJ Tree Step-by-step

The algorithm is repeated until the tree is fully resolved:

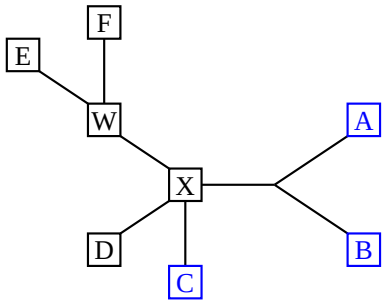
1.



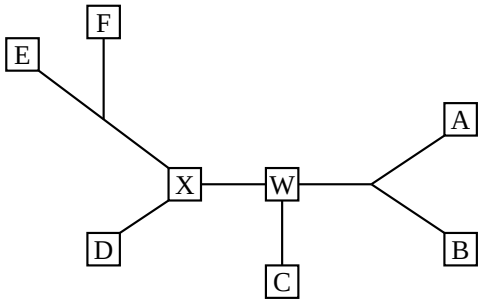
2.



3.



4.

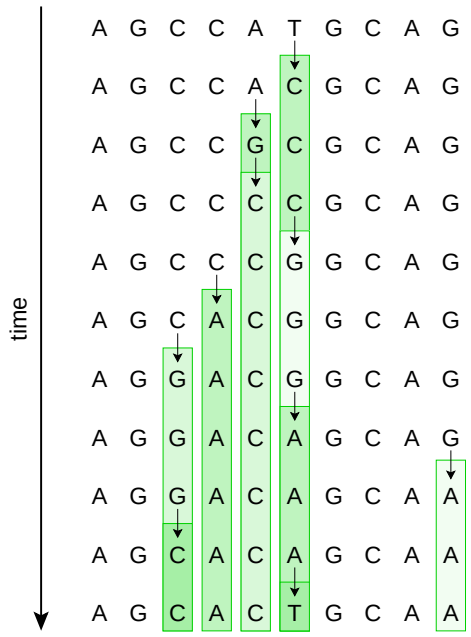


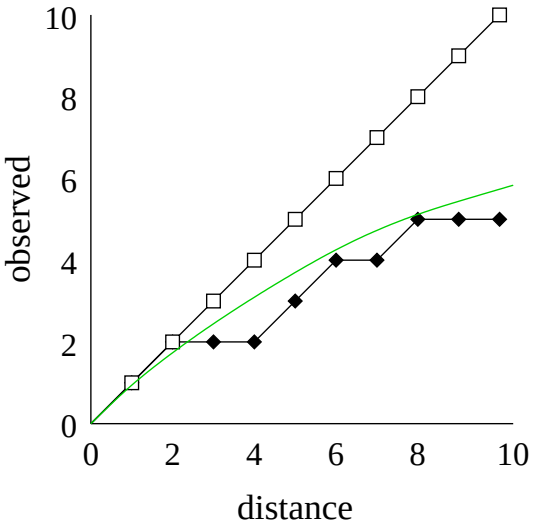
The most simple tree: How to get distances?

The most simple tree could be seen as two sequences and the distance between them.

Distances can be computed in various ways. . .

Jukes-Cantor Correction for Multiple Mutations





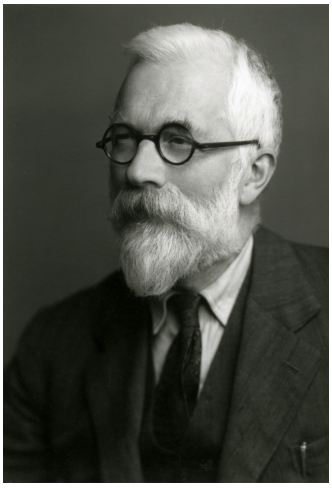
The substitution process is commonly modeled as a Markov process.

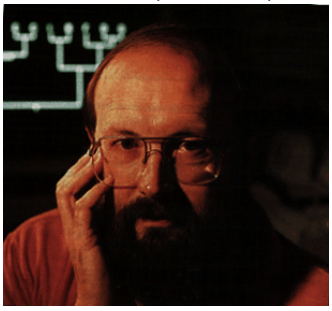
The most simple tree: How to get distances?

The most simple tree could be seen as two sequences and the distance between them.

Distances can be computed in various ways. . .

Usually via Maximum Likelihood (ML).


Ronald Fisher (1890-1962)


Joe Felsenstein (born 1942)

Introduction: ML on Coin Tossing

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

We take out one coin and toss 20 times:

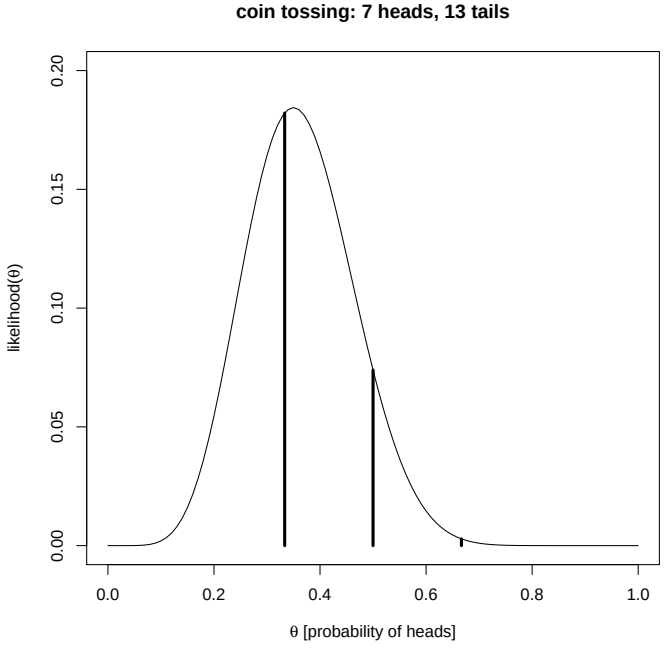
$$H, T, T, H, H, T, T, T, T, H, T, T, H, T, H, T, T, H, T, T$$

	Probability		**Likelihood**
$p(k$ heads in n tosses$\|\theta)$	$\equiv$	$L(\theta\|k$ heads in n 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.

Modeling Evolution

- Evolution is usually modeled as a

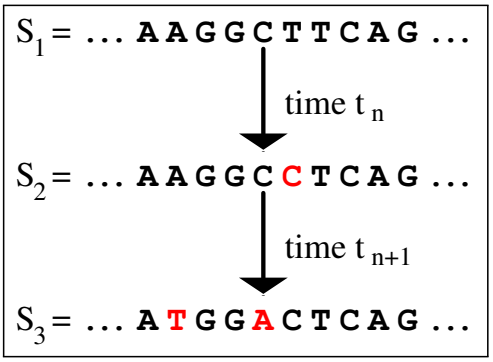
  stationary, time-reversible Markov process.

- 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

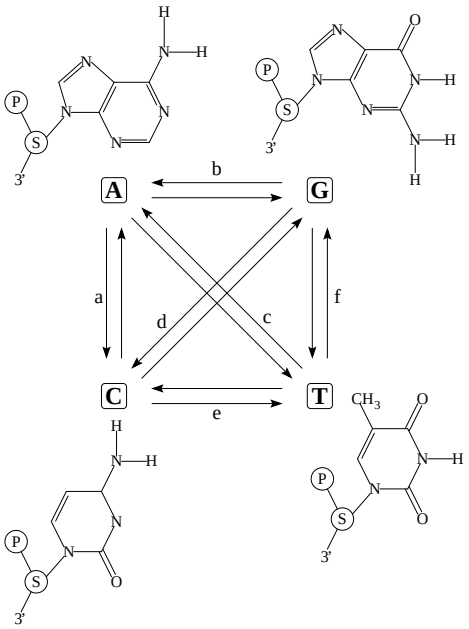
$$\pi_i \cdot P_{ij}(t) = P_{ji}(t) \cdot \pi_j$$

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

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

Substitution Models

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

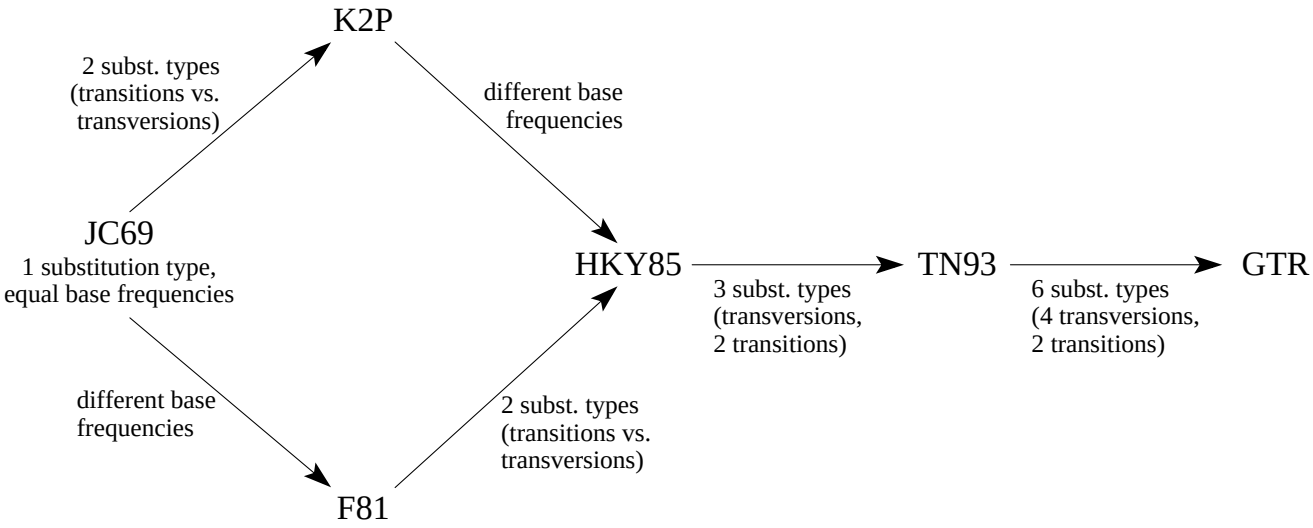


$$R = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} - & a & b & c \\ a & - & d & e \\ b & d & - & f \\ c & e & f & - \end{pmatrix} \end{array}$$

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

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

Relations between 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, chloroplast 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)~~ $\rightarrow$ for database searching

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

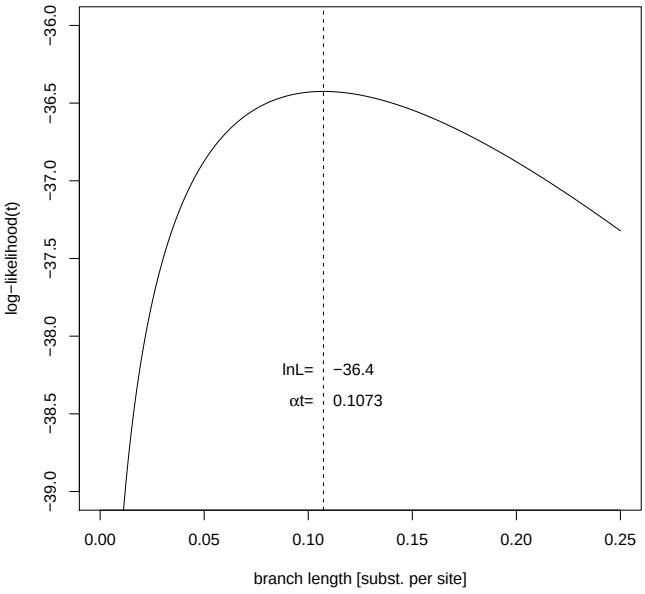
The Likelihood of sequence s evolving to s' in time t:

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

Likelihood surface for two sequences under JC69:

TGATCCTGAGTGAACTAAGC $= s'$
TGGTCCTGACTGAACTAAGC $= s$

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



Computing Likelihood Values for Trees

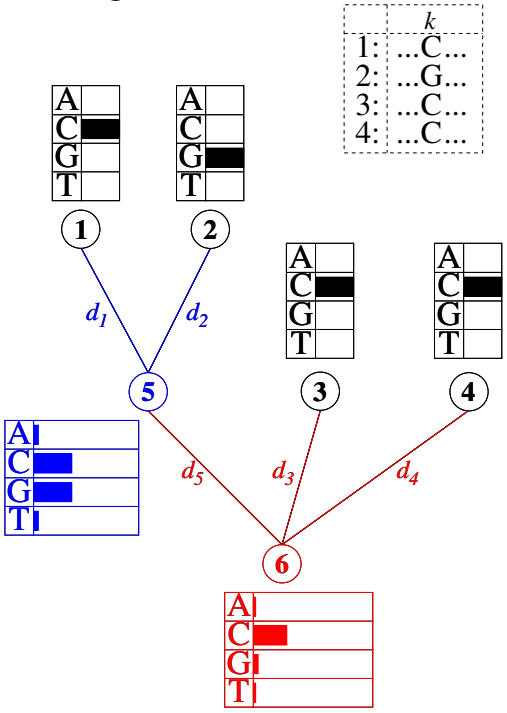
Given a tree with branch lengths and sequences for all nodes, the computation of likelihood values for trees is straight forward.
Unfortunately, we usually have no sequences for the inner nodes (ancestral sequences).
Hence we have to evaluate every possible labeling at the inner nodes:

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

for every column in the alignment... but there is a faster algorithm.
(Peeling Algorithm by Felsenstein, 1981)

Likelihoods of Trees (Single alignment column, given tree)

For a single alignment column and a given tree:



Likelihoods of nucleotides i at inner nodes:

$$L_5(i) = [P_{iC}(d_1) \cdot L_1(C)] \cdot [P_{iG}(d_2) \cdot L_2(G)]$$

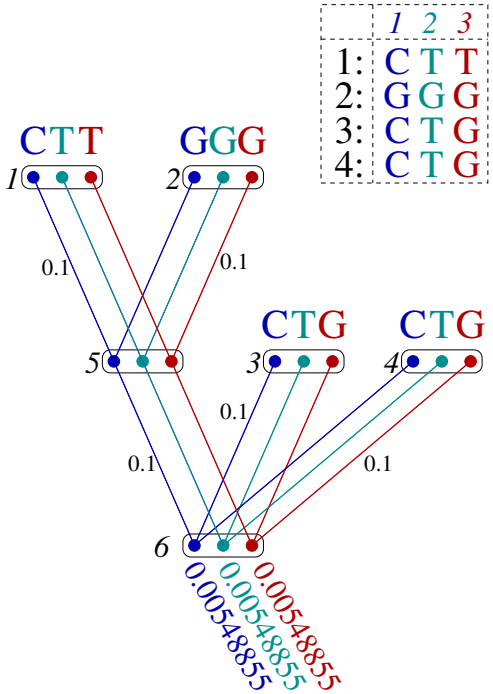
$$L_6(i) = \prod_{v=\{3,4,5\}} \left[\sum_{j=\{ACGT\}} P_{ij}(d_v) \cdot L_v(j) \right]$$

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 $P_{ij}(0.1) = \begin{cases} .9068 & i = j \\ .0313 & i \neq j \end{cases}$ (JC)

Likelihoods of Trees (multiple columns)



Considering this tree with $n = 4$ sequences of length $m = 3$ the tree likelihood of this tree is

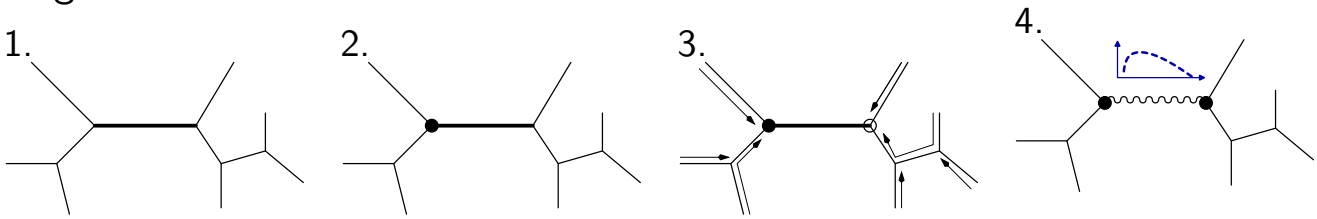
$$\begin{aligned}\mathcal{L}(T) &= \prod_{k=1}^{m} L^{(k)} \\ &= 0.0054886 \cdot 0.0054886 \cdot 0.0054886 \\ &= 0.0000001653381\end{aligned}$$

or the log-likelihood

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

Adjusting Branch Lengths Step-By-Step

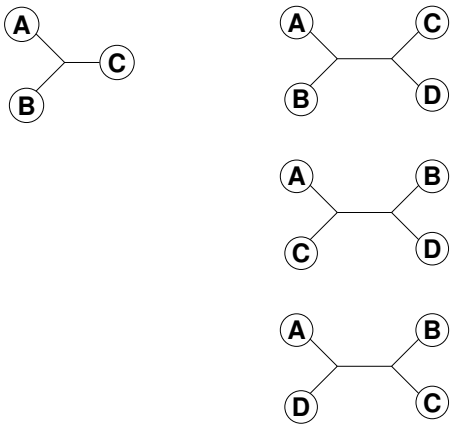
To compute optimal branch lengths do the following. Initialize the branch lengths.
(1.) Choose a branch. (2.) Move the virtual root to an adjacent node. (3.) Compute all partial likelihoods recursively. (4.) Adjust the branch length to maximize the likelihood value.



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

This is based on the Pulley-Principle (Felsenstein, 1981) which states that the root can be moved on the tree but the likelihood doesn't change.

Number of Trees to Examine...

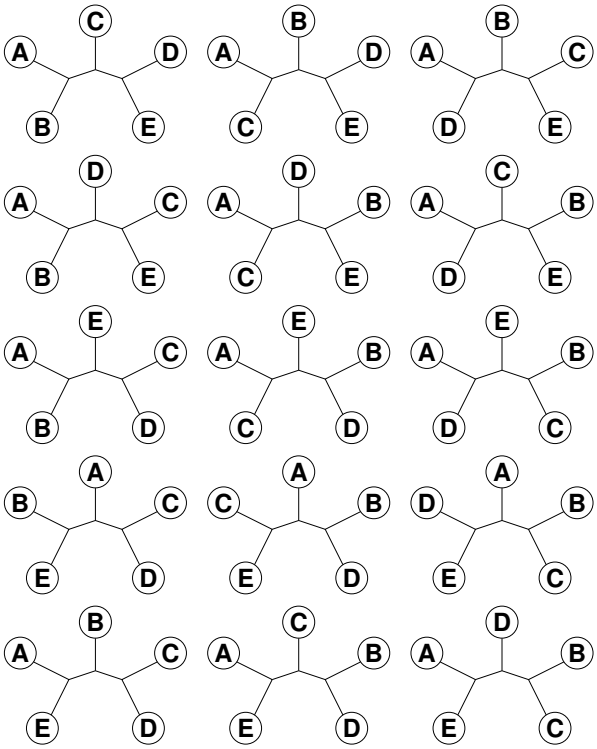


$$B(n) = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

$$B(10) = 2027025$$

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

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



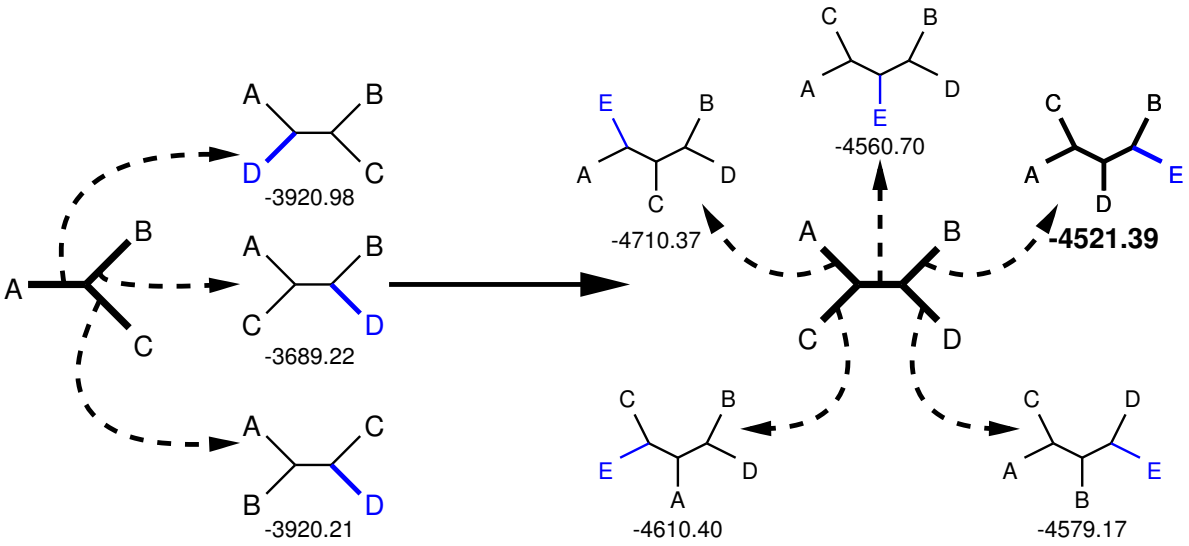
Finding the ML Tree

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.

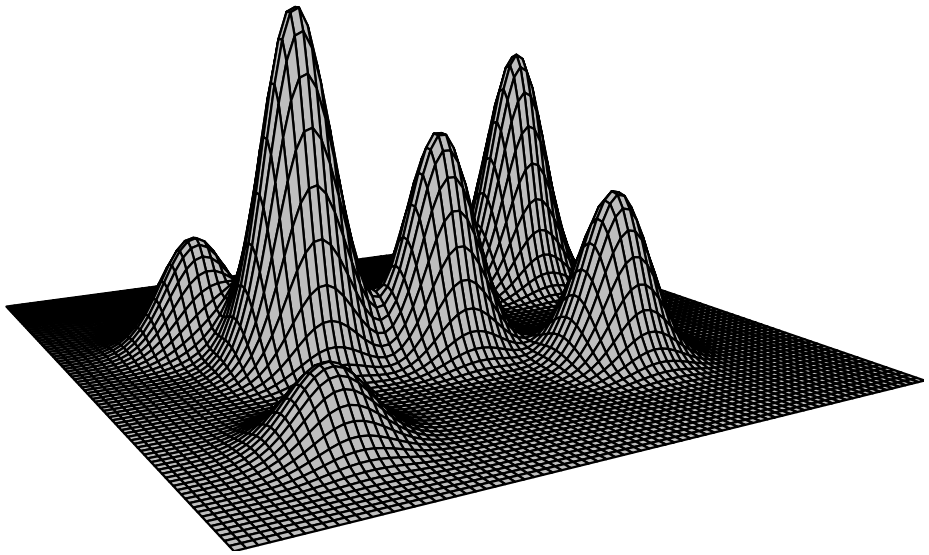
Build up a tree: Stepwise Insertion



Is also used for other (non-ML) methods like parsimony.

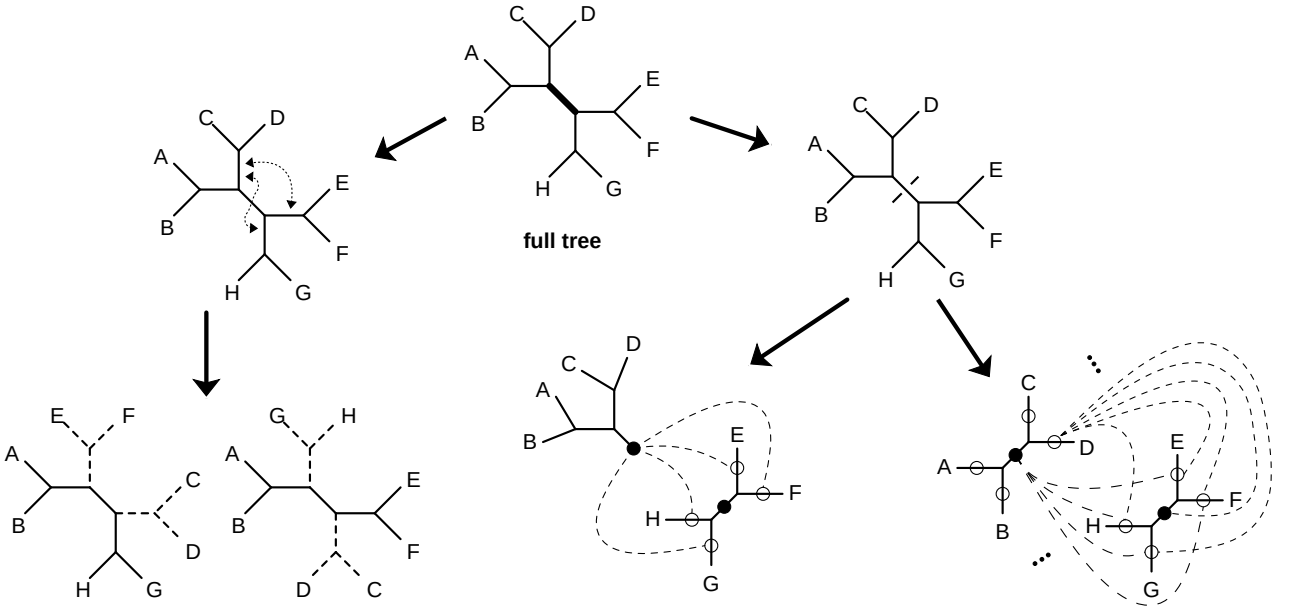
Local Maxima

What if we have multiple maxima in the likelihood surface?



Tree rearrangements to escape local maxima.

Tree Rearrangements: Scanning a Tree's Neighborhood



From a current tree construct other trees by rearranging its subtrees and evaluate all resulting trees. Repeat with the best tree found, until no better tree can be found. This also used for other (non-ML) methods, like parsimony.

How reliable is the reconstructed tree:

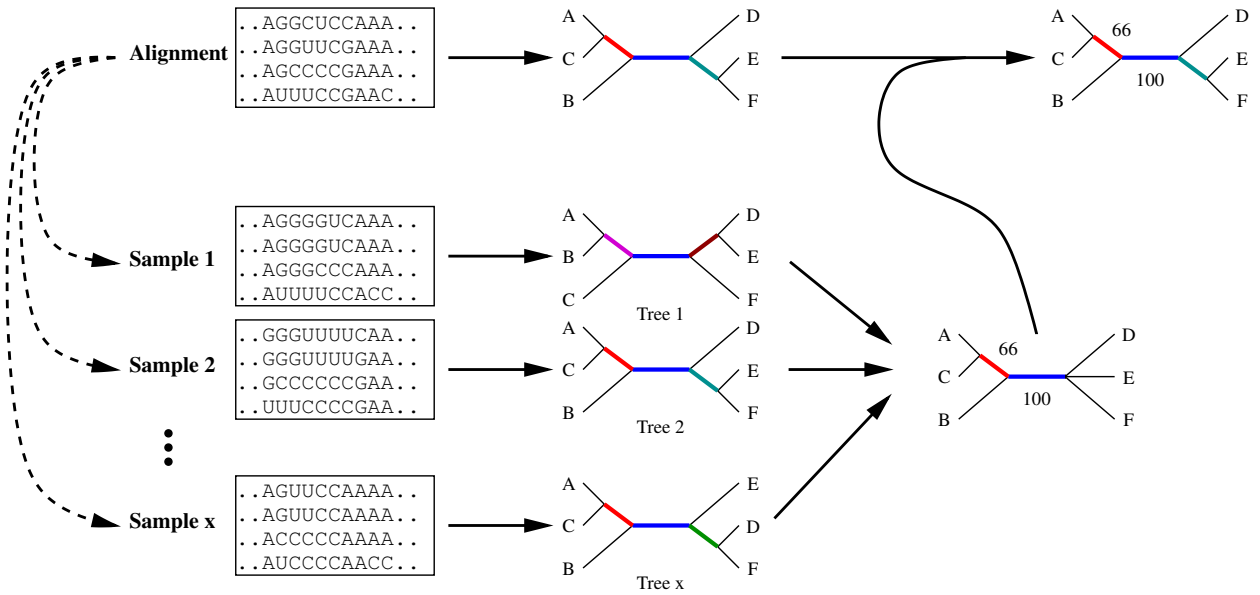
- Usually programs deliver a single (best) tree, but without confidence values for the subtrees.
- How can we assess reliability for the subtree?

Bootstrap and Consensus Tree

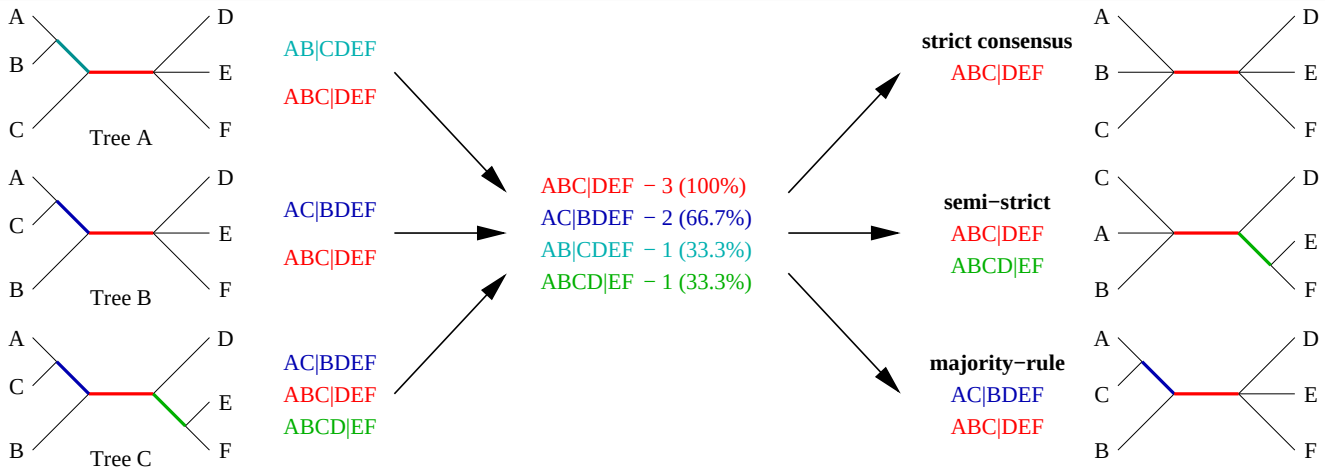
- Bootstrapping creates many pseudo-alignments by sampling alignment columns with replacement from the original alignment.
- From the pseudo-alignment we reconstruct trees.
- From the trees we collect and count all splits.
- From the splits we construct a consensus tree.

- **Definition:** A split $\mathcal{A}|\mathcal{B}$ in the tree is the bipartition of the leaves/taxa into two subsets $\mathcal{A}$ and $\mathcal{B}$ induced by removing an edge or branch from the tree.
- **Definition:** A trivial split is a split induced by an external branch, because they have to be present in every tree. Otherwise a leaf would not be connected to the tree.
- **Definition:** Two splits $\mathcal{A}|\mathcal{B}$ and $\mathcal{C}|\mathcal{D}$ are compatible, i.e. not contradictory, if at least one intersection of $\mathcal{A} \cap \mathcal{C}$, $\mathcal{A} \cap \mathcal{D}$, $\mathcal{B} \cap \mathcal{C}$, $\mathcal{B} \cap \mathcal{D}$ is empty.

Estimating Confidence: The Bootstrap



Summarizing Trees: Consensus Methods



- Strict consensus: contains all splits occuring in all input tree.
- Semi-strict consensus: contains all splits which are not contradicted by any tree.
- Majority consensus M_ℓ: contains all splits which occur in more than ℓ input trees, where $\ell \geq 50\%$ typically exactly 50%.
- Majority Rule extended (MRe): starting from the most to the least frequent splits one collects compatible splits. If a split is incompatible to the already collected ones, it is discarded and the next is examined.

Towards Phylogenomics

- In the past often sequences of single genes were used as representative for the species they originates from to reconstruct the tree of the species history - the speciestree.
- However, the reconstructed tree reflects the gene's history - thus, the genetree.
- With the advent of genomics data, often sequence data for several genes per species are available.
- This approach is based on the hope, that the majority of genes evolved according to the species history, averaging out the genes which did not.

Reconstruct a speciestree from multi-gene data