Maximum Likelihood in Molecular phylogeny

Heiko A. Schmidt

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

Main Types of Phylogenetic Methods

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

Introduction: ML on Coin Tossing

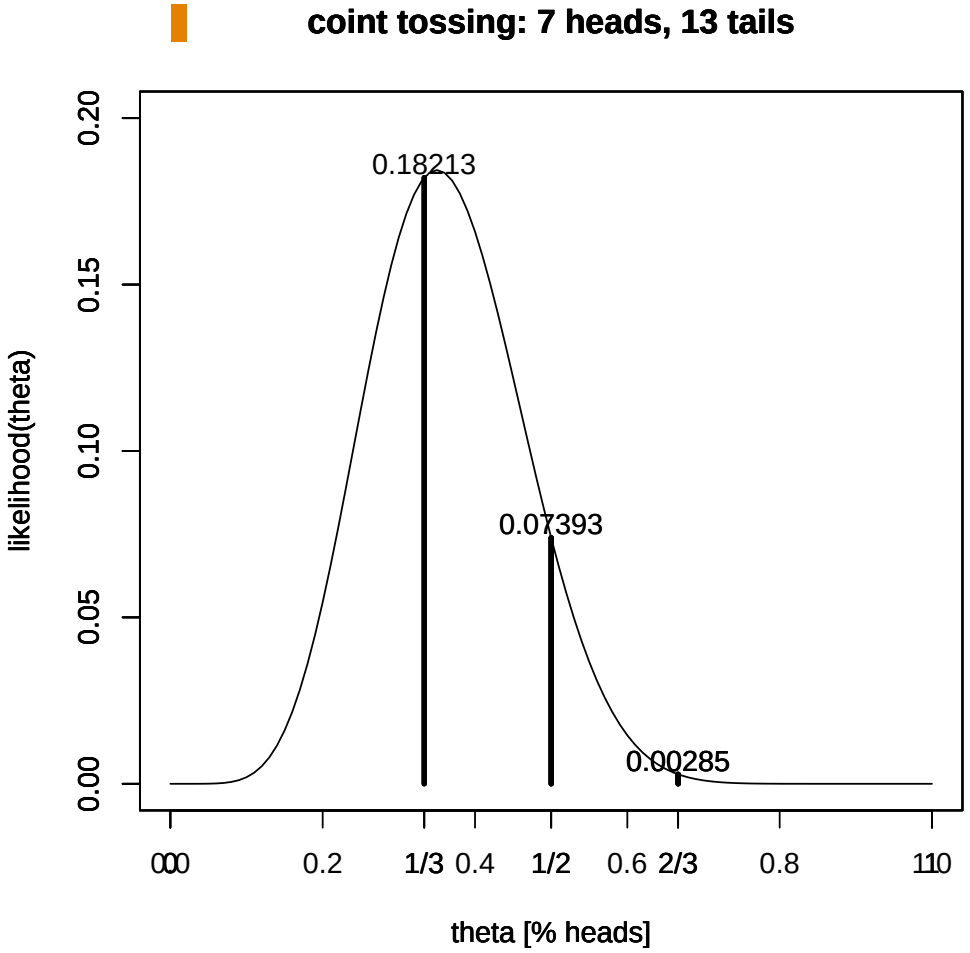
Given a box with 3 coins of different fairness $\left(\frac{1}{3}, \frac{1}{2}, \frac{2}{3}\right.$ heads$)$

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**
$p(k$ 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)

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 $\theta \in \{\frac{1}{3}, \frac{1}{2}, \frac{2}{3}\}$

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?

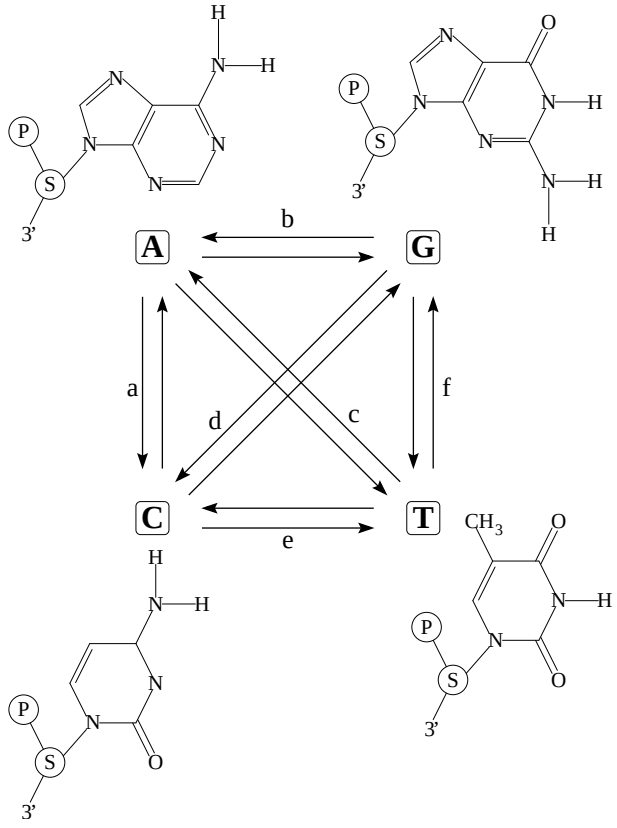
While the coin tossing example might look easy, in phylogenetic analysis, the parameter (set) θ comprises:

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

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

Substitution Models

Evolutionary models are often described using substitution rate matrices. Here, 4×4 matrix for DNA models:



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

DNA Substitution Models

The rate matrix R and the character frequencies Π

$$R = \begin{pmatrix} - & a & b & c \\ a & - & d & e \\ b & d & - & f \\ c & e & f & - \end{pmatrix} \qquad \Pi = (\pi_A, \pi_C, \pi_G, \pi_T)$$

are transformed into the instantaneous rate matrix Q

$$Q = \begin{pmatrix} \bullet_A & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & \bullet_C & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & \bullet_G & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & \bullet_T \end{pmatrix} \qquad \begin{aligned} \bullet_A &= -(a\pi_C + b\pi_G + c\pi_T) \\ \bullet_C &= -(a\pi_A + d\pi_G + e\pi_T) \\ \bullet_G &= -(b\pi_A + d\pi_C + f\pi_T) \\ \bullet_T &= -(c\pi_A + e\pi_C + f\pi_G) \end{aligned}$$

where the row sums are zero.

How to Get Substitution Probabilities?

Given now the instantaneous rate matrix Q, we can compute a substitution probability matrix P

$$P(t) = e^{Qt}$$

With this matrix P we can compute probability values for changes over a time t.

Time t is normalized by a factor μ such that

$$\sum_{i \neq j} \pi_i P_{ij}(1) = 0.01 \quad \text{or} \quad \sum_{i \neq j} Q_{ij} = \sum_{i=j} Q_{ij} = 1$$

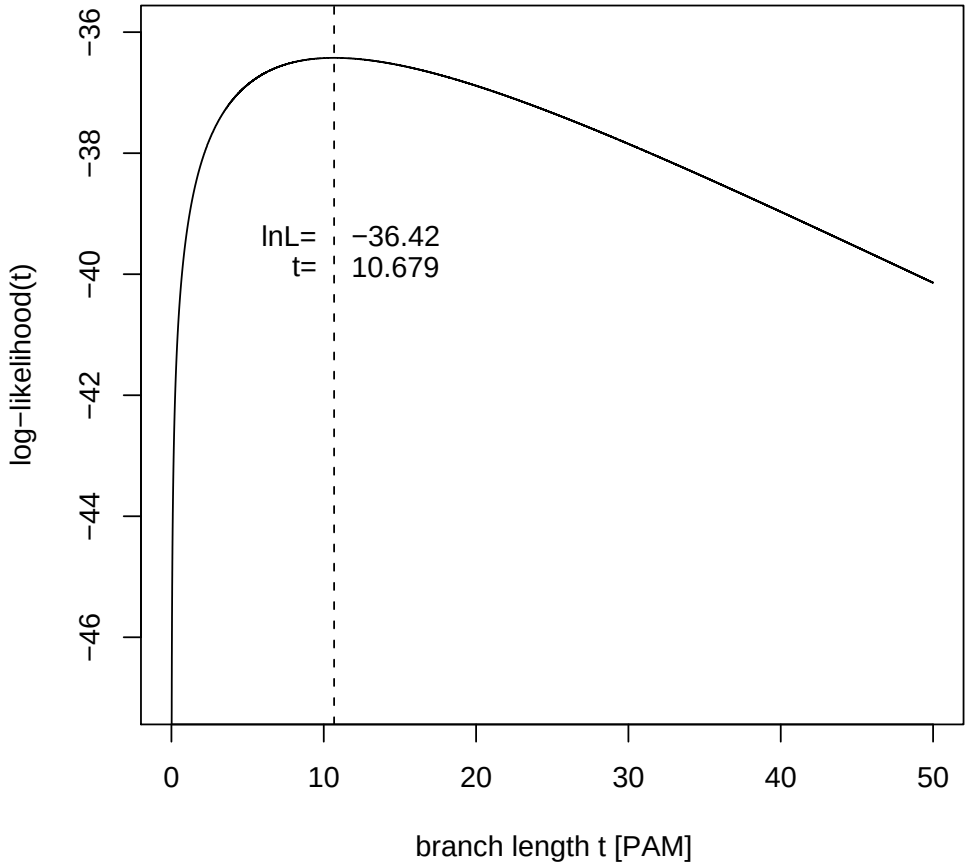
Computing ML Distances Using $\mathbf{P_{ij}(t)}$

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

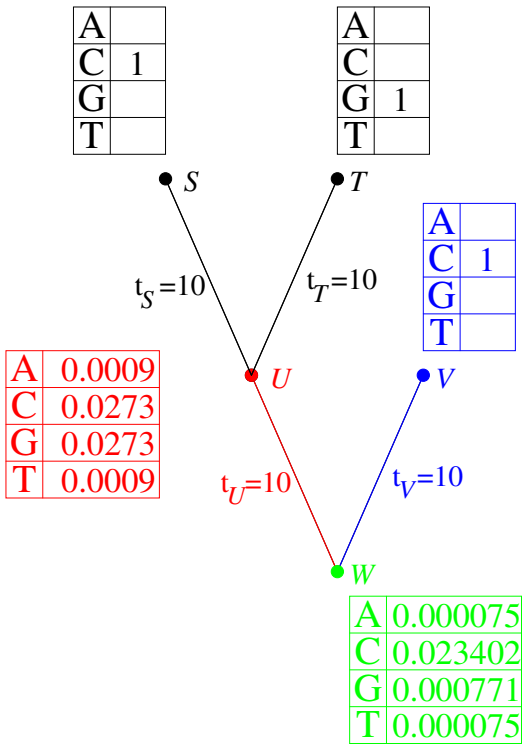
$$L(t|s \to s') = \prod_{i=1}^{m} \Big(\Pi(s_i) \cdot P_{s_i s'_i}(t) \Big)$$

Likelihood surface for two sequences under JC69:

GATCCTGAGAGAAATAAAC
GGTCCTGACAGAAATAAAC



Likelihoods of Trees (Single column, given tree)



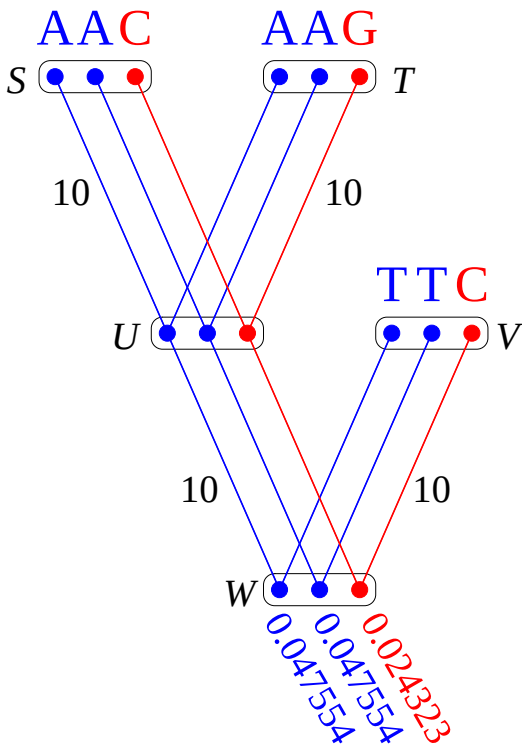
Likelihoods of nucleotides at inner nodes:

$$L_W(i) = \left[\sum_{\substack{u= \\ ACGT}} P_{iu}(t_U) L_U(u)\right] \cdot \left[\sum_{\substack{v= \\ ACGT}} P_{iv}(t_V) L_V(v)\right]$$

Site-Likelihood of an alignment column k:

$$L^{(k)} = \sum_{\substack{i= \\ ACGT}} \pi_i \cdot L_W(i) = 0.024323$$

Likelihoods of Trees (multiple columns)



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

$$\mathcal{L}(T) = \prod_{k=1}^{m} L^{(k)} = 0.047554^2 \cdot 0.024323$$

$$= 0.000055$$

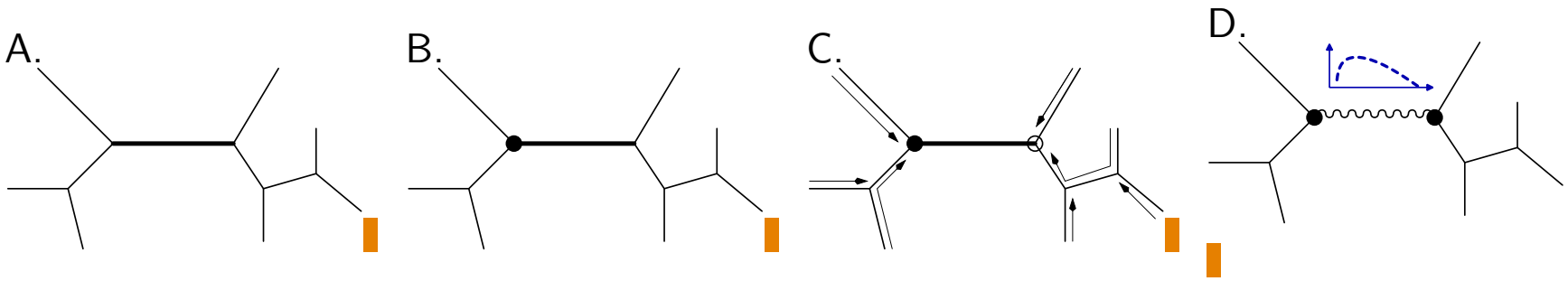
or the log-likelihood

$$\ln \mathcal{L}(T) = \sum_{k=1}^{m} \ln L^{(k)} = -9.80811$$

Adjusting Branch Lengths Step-By-Step

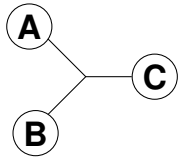
To compute optimal branch lengths do the following. Initialize the branch lengths.

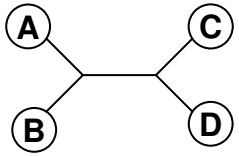
Choose a branch (A.). Move the virtual root to an adjacent node (B.). Compute all partial likelihoods recursively (C.). Adjust the branch length to maximize the likelihood value (D.).



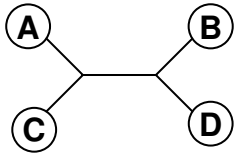
Repeat this for every branch until no better likelihood is gained.

Number of Trees to Examine. . .

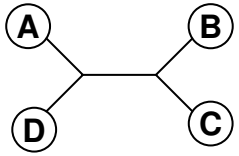




$L_{AB|CD}$



$L_{AC|BD}$



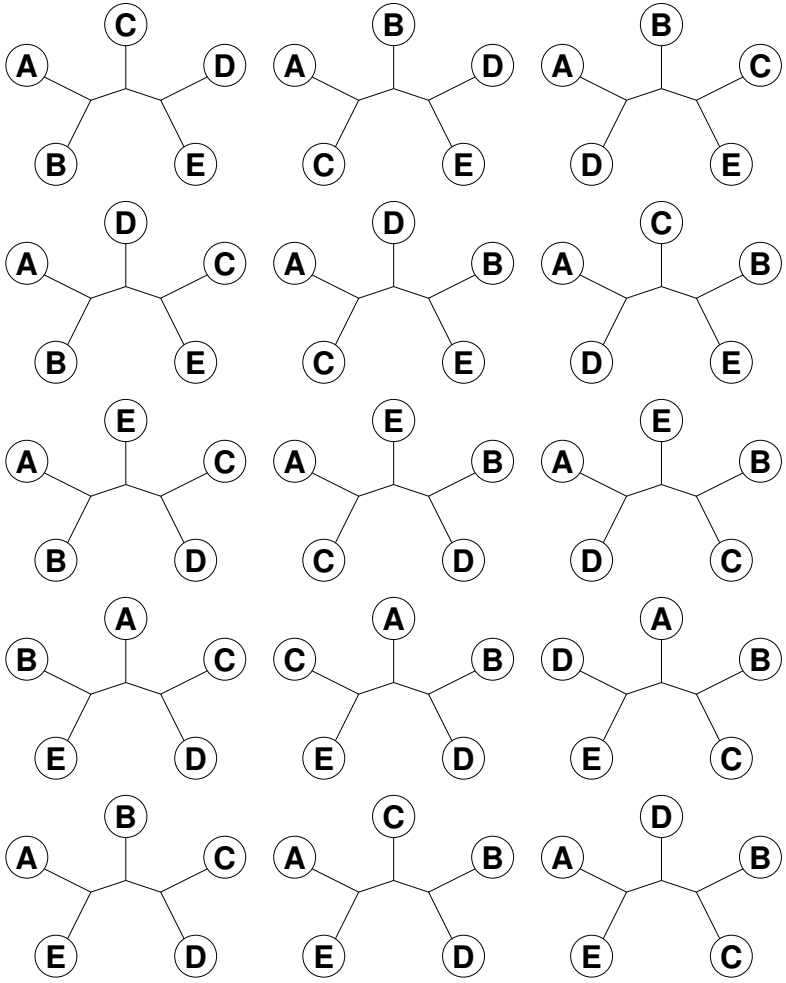
$L_{AD|BC}$

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

$$B(10) = 2027025$$

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

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



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.

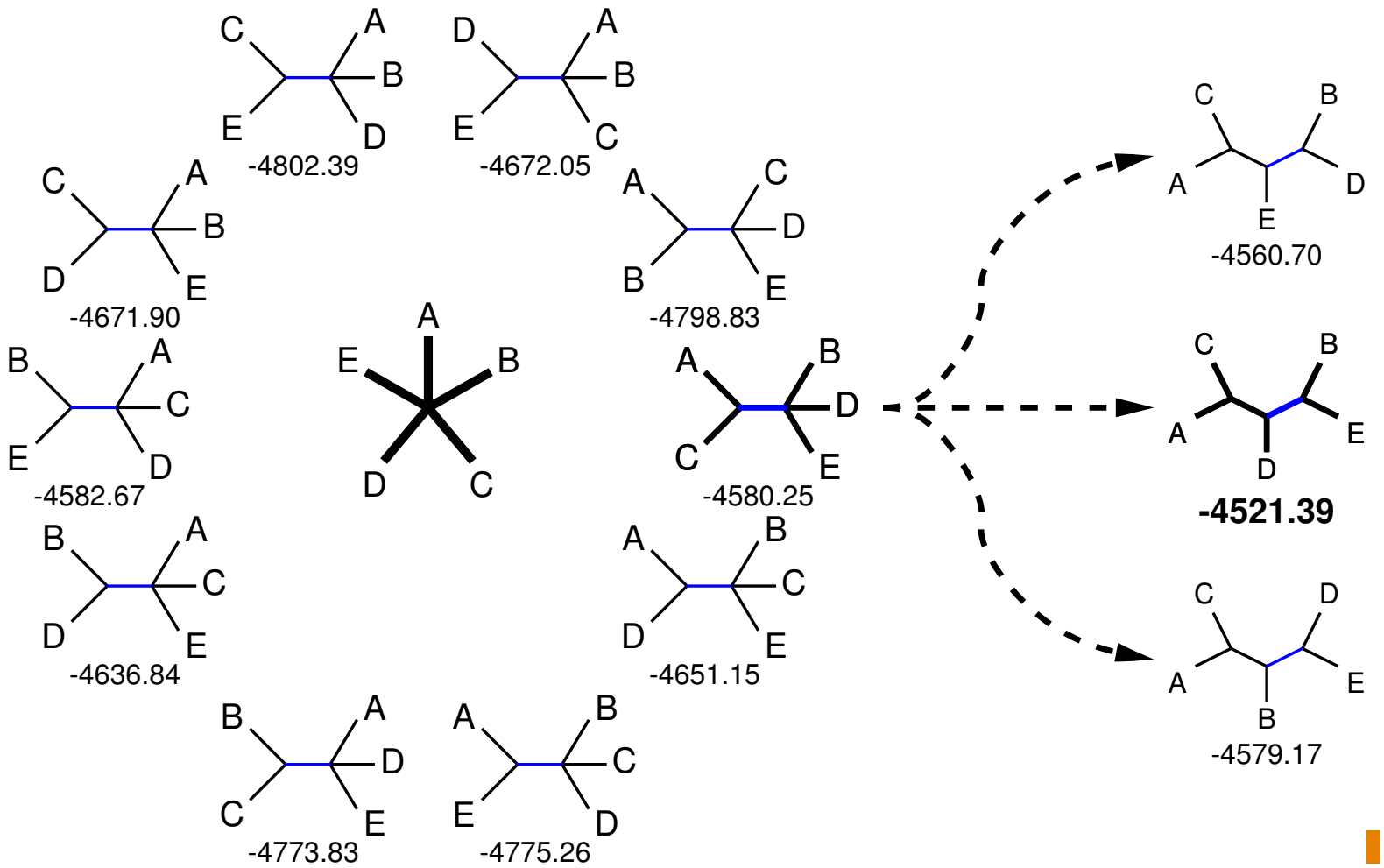
Heuristics: cannot guarantee to find the optimal tree, but is at least able to analyze large datasets. Typically involve tree optimization procedures like NNI, SPR, or TBR.

'Solution'

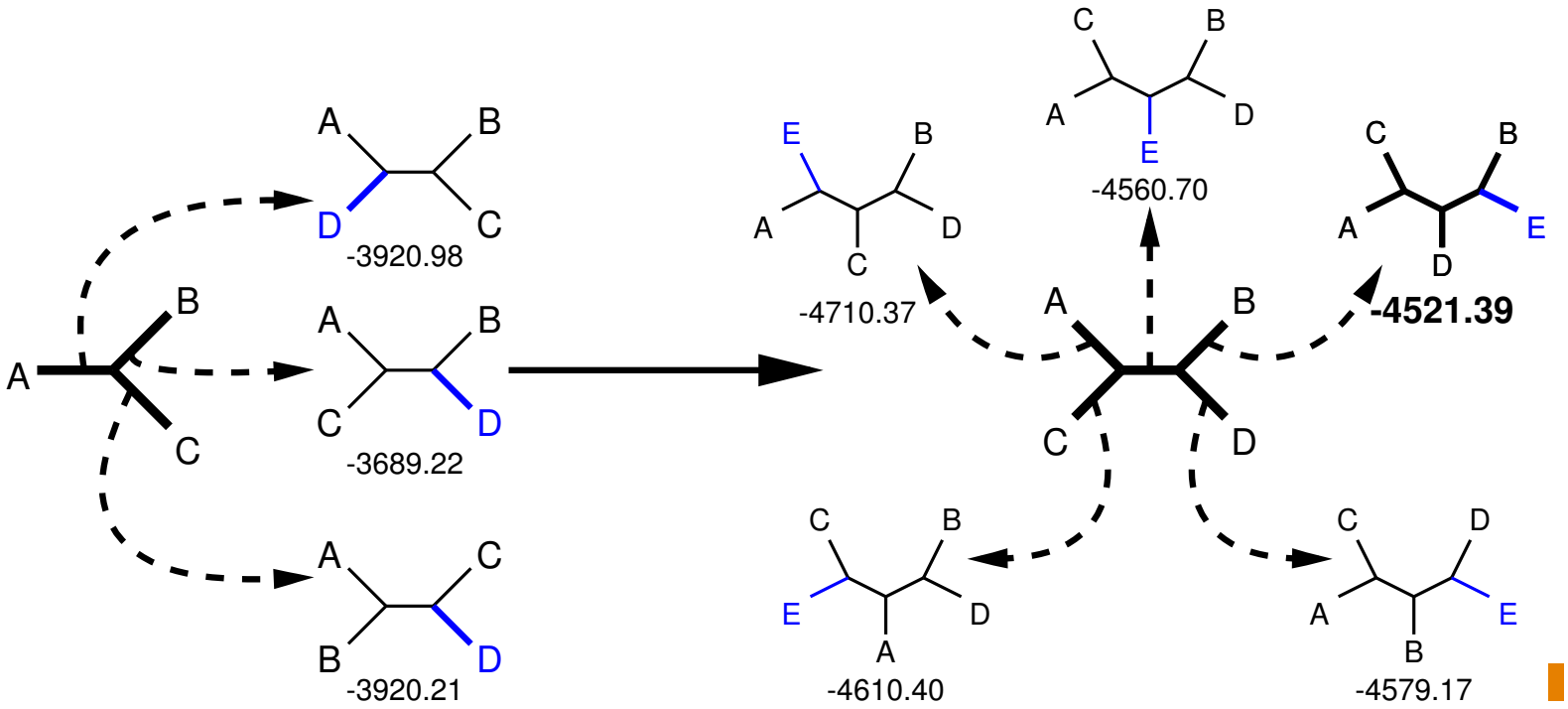
Heuristics are used to reduce the number of trees to be examined. Some well-known methods are:

- stepwise insertion (DNAML from PHYLIP, fastDNAml, PAUP*)
- star decomposition (PROTML, NUCML from MOLPHY)
- Bayesian sampling from the universe of trees (MrBayes)
- Genetic Algorithms (MetaPIGA)
- quartet puzzling (TREE-PUZZLE, Qstar)

Star Decomposition

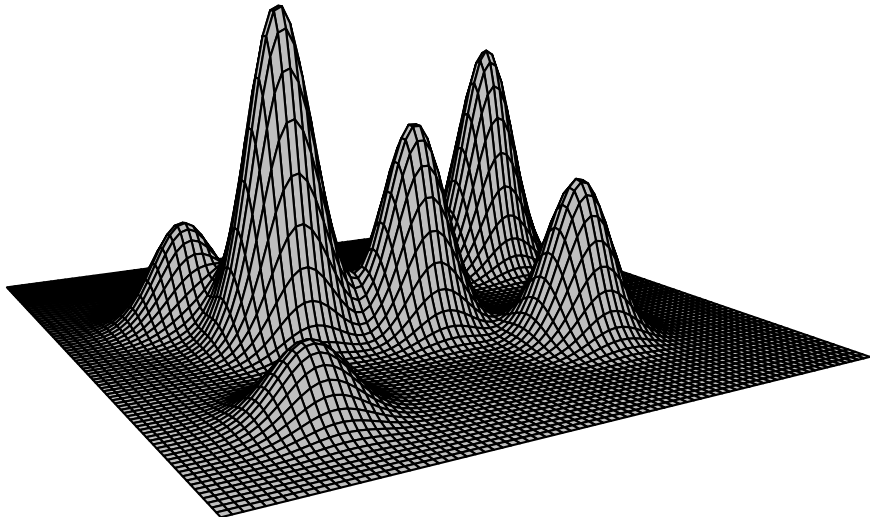


Stepwise Insertion



Refinements

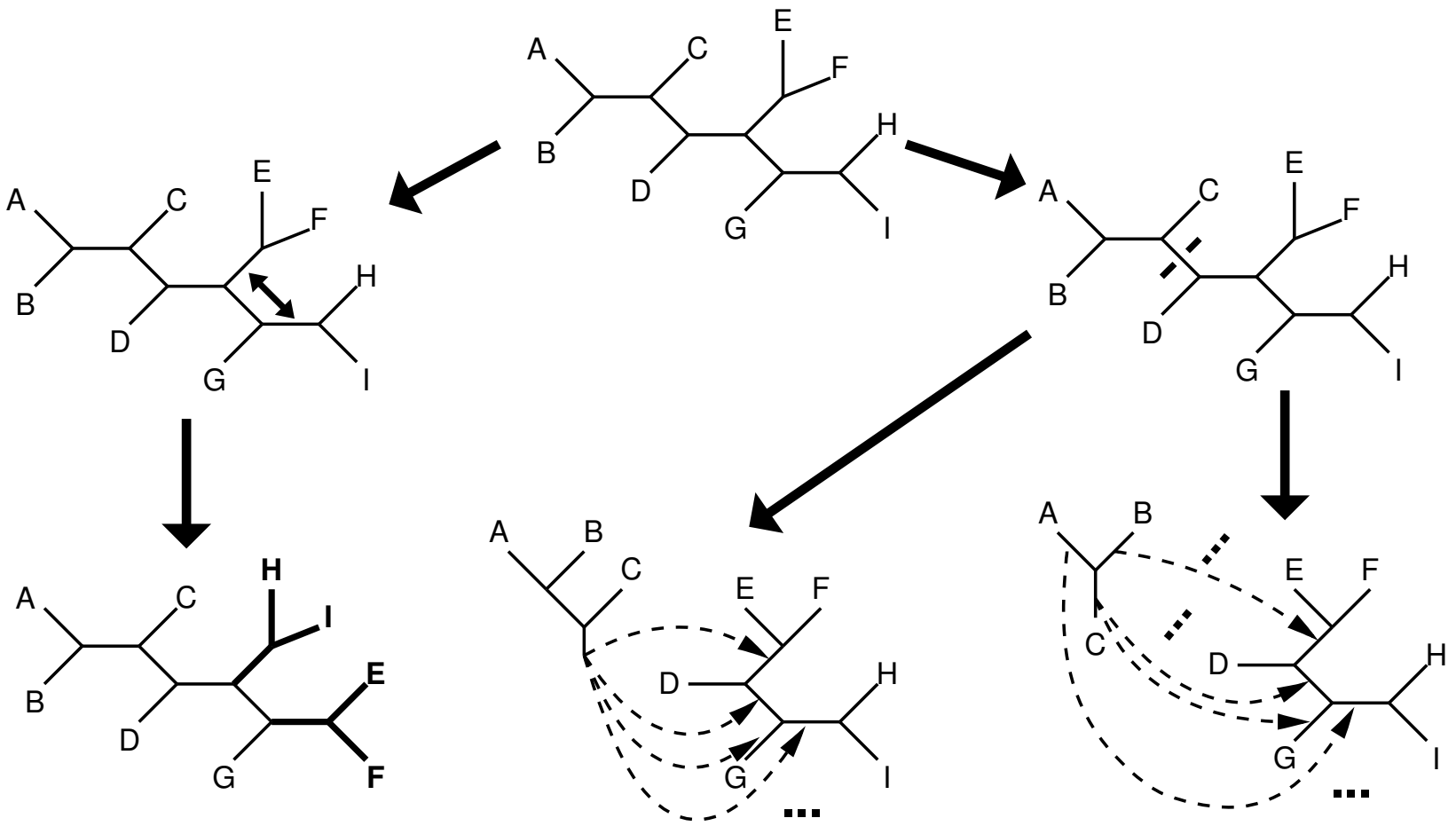
What if we have multiple maxima in the likelihood surface?



Common optimization techniques:

- Global rearrangements, SPR (DNAML from PHYLIP)
- Local rearrangements, NNI
- fast NNI (PHYML, IQPNNI)
- lazy NNI (RAxML)
- TBR (PAUP*)

Tree Optimizations



Nearest Neighbor Interchange

Possible NNI trees = O(n)

subtree pruning + regrafting

Possible SPR trees = O(n*n)

tree-bisection + reconnection

Possible TBR trees = O(n*n*n)

Posterior Probabilities and Empirical Bayes

- We can now reconstruct ML trees, but how comparable are the likelihoods, how reliable the groupings?

- Branch reliability can be checked, support values computed using:
  - Bootstrapping, Jackknifing alignment columns + consensus.
  - Randomizing input orders in stepwise insertions (TREE-PUZZLE).

Posterior Probabilities and Empirical Bayes

- Problem: How different are likelihoods? Just from the value of likelihoods one often cannot tell whether they are significantly different.
- Nomalization: Posterior probabilities are computed:

$$p_i = \frac{L_1}{\sum_n L_n}$$

- Usage:
  - Which sites along an alignment support a tree most?
  - Are there sites/partitions not supporting a tree?
  - Which model of evolution (e.g. dependent, independent) is supported by which site/partition? (PAML)
  - Is a site fast/medium/slowly evolving? (PAML, TREE-PUZZLE)
  - Constructing confidence sets on posterior tree likelihoods (MrBayes)

LRT – Likelihood Ratio Test (1)

The Likelihood function offers a natural way of comparing nested evolutionary hypothesis using the Likelihood Ratio (LR) statistics:

$$\Delta = 2(\ln L_1 - \ln L_0)$$

L_1 maximum likelihood under the more parameter-rich, complex model (alternative hypothesis, H_1)

L_0 maximum likelihood under the less parameter-rich simple model (Null-hypothesis, H_0)

If the models are nested, i.e., H_0 is a special case of H_1 and the Null-hypothesis (H0) is correct, Δ is asymptotically χ^2-distributed with the number of degrees of freedom equal to the difference in number of free parameters between the two models.

LRT – Likelihood Ratio Test (2)

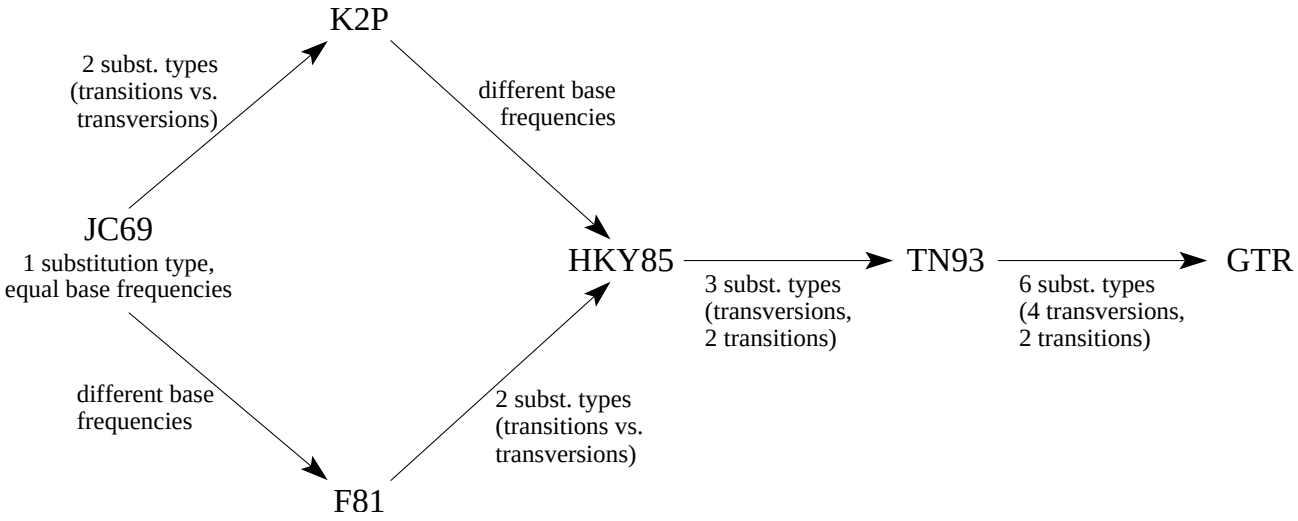
- If the LRT is significant (i.e., $p < 0.05$ or $p < 0.01$): the use of the additional parameters in the alternative model H_1 increases the likelihood significantly.

- If Δ is close to zero, that is, $p > 0.05$: the alternative hypothesis H_1 does not fit the data significantly better than H_0, that means using the additional parameters of H_1 does not explain the data better.

- Only nested models can be tested:
  One model (H_0, Null-model, constraint model) is nested in another model (H_1, alternative, unconstraint model) if the model H_0 can be produced by restricting parameters in model H_1.

LRT – Typical cases of nested models

- Different levels of evolutionary models:



- *rate-homogeneous models* (H_0) are nested in *rate-heterogeneous models* (H_1)
- A tree assuming *molecular clock* (H_0) are nested its *non-clock* version (H_1)

How to Compare Tree Topologies

Since tree topologies are no normal parameters, they are generally not nested in each other. Hence, other methods have to be used.

Usually, the alignment (or the site-likelihoods) are bootstrapped, i.e., sampled with replacement, to produce a distribution of likelihoods which is then used for

- pairwise tests against the best trees (Kishino-Hasegawa test)
- multiple tests among user defined trees (Shimodaira-Hasegawa test)

to decide which trees are significantly worse.

Methods to derive distributions for testing

- Bootstrapping from the alignment columns
- Bootstrapping from the alignment columns' site-likelihood
- Parametric bootstrap, i.e., simulation with fixed model (parameters) on a given tree, to produce alignments.

Overview over Likelihood-based Analyses

- Comparing hypothesis with Likelihood-Ratio-Test (=LRT)
  - different models of evolution (ModelTest)
  - testing molecular clock assumption and root position (TREE-PUZZLE)
- Parameter estimation (TREE-PUZZLE, PAUP)
- Testing for phylogenetic content (TREE-PUZZLE)
- Comparing/testing different tree topologies with Kishino-Hasegawa test, Shimodaira-Hasegawa test (TREE-PUZZLE)
- Constructing confidence sets on posterior likelihoods (MrBayes)

Software using Likelihood

- `dnaml` (PHYLIP), fastDNAml (stepwise insertion+SPR)
- IQPNNI (BioNJ-tree + randomization + fast NNI)
- PAUP* (several methods)
- PHYML (BioNJ-tree + fast NNI)
- RAxML (MP-tree + lazy NNI)
- MetaPIGA (optimizing trees with genetic algorithms)
- TREE-PUZZLE (quartet puzzling with random orders)

- PAML (Model analysis on given trees)
- MrBAYES ('Random' MCMC sampling from the universe trees and parameters)