Maximum Likelihood Tree Searching

Emily Jane McTavish

Life and Environmental Sciences
University of California, Merced
ejmctavish@ucmerced.edu, twitter:snacktavish

(With thanks to Derrick Zwickl, Mark Holder, Dave Swofford and Paul Lewis for slides!)

Finding the tree with the best likelihood score is a hard problem

Finding the tree with the best likelihood score is a hard problem

- Enormous numbers of topologies to consider

Finding the tree with the best likelihood score is a hard problem

- Enormous numbers of topologies to consider
- May be multiple local optima

Finding the tree with the best likelihood score is a hard problem

- Enormous numbers of topologies to consider
- May be multiple local optima
- Need to maximize the likelihood for each topology

Enormous numbers of topologies to consider

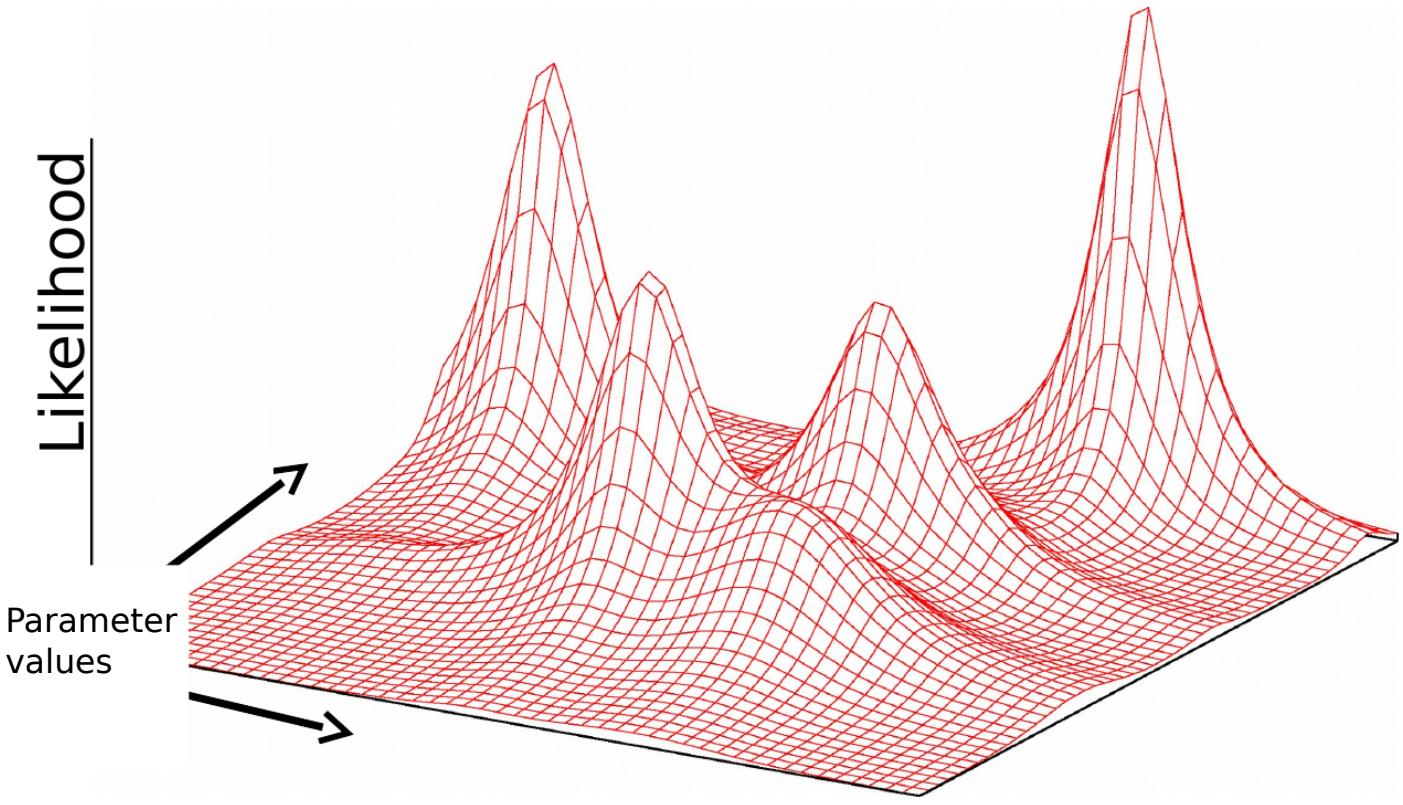
Taxa	Unrooted binary trees	Rooted binary trees
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10,395
8	10,395	135,135
9	135,135	2,027,025
10	2,027,025	3×10^{7}
15	7×10^{12}	2×10^{14}
20	2×10^{20}	8×10^{21}
50	3×10^{74}	
100	2×10^{182}	
1,000	2×10^{2860}	
10,000	8×10^{38658}	
1,000,000	1×10^{5866723}	

Enormous numbers of topologies to consider

Taxa	Unrooted binary trees	Rooted binary trees
3	1	3
4	3	15
5	15	105
6	105	945
10	2,027,025	3×10^{7}
15	7×10^{12}	2×10^{14}
20	2×10^{20}	8×10^{21}
50	3×10^{74}	
100	2×10^{182}	
1,000	2×10^{2860}	
10,000	8×10^{38658}	
1,000,000	1×10^{5866723}	

it is estimated that the there are between 10^{78} to 10^{82} atoms in the known, observable universe.

There may be multiple local likelihood optima

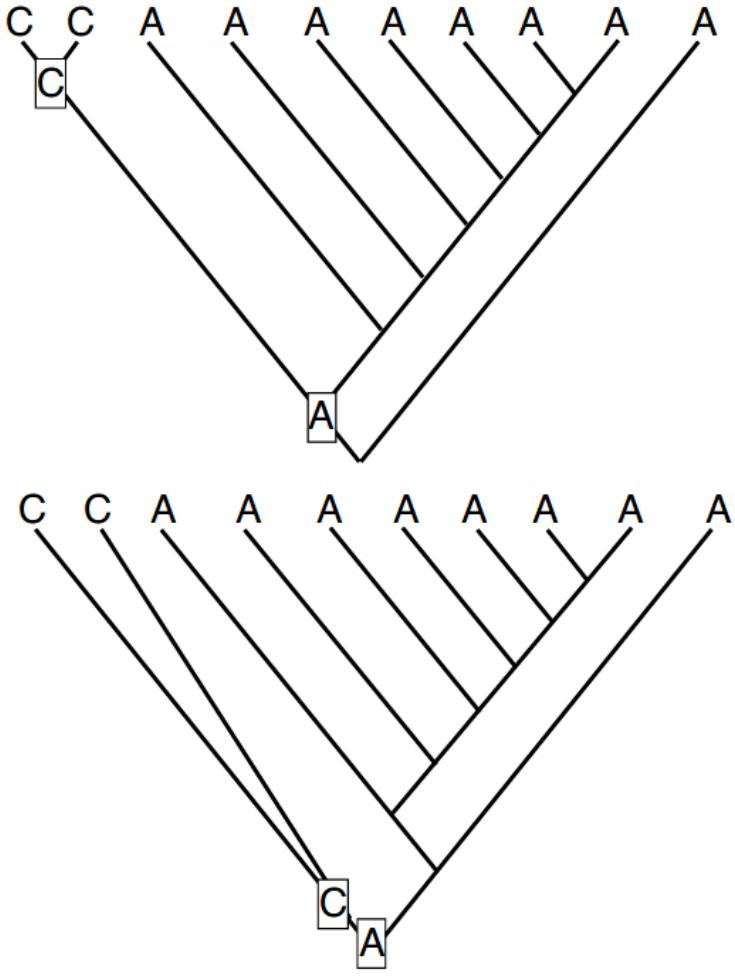


(From Zwickl)

Need to maximize the likelihood for each topology

- Update numerical parameters of the model of sequence evolution
- Branch-length parameters

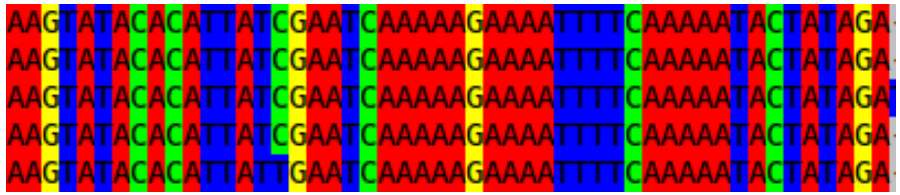
The Relevance of Branch Lengths



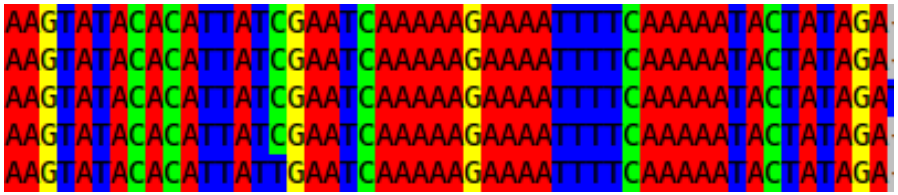
(From Swofford)

Ascertainment bias in genomic data

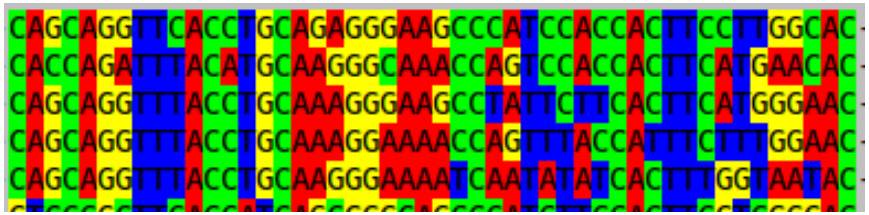
Which tree has longer branch lengths? Tree 1



Which tree has longer branch lengths? Tree 1



Tree 2



How surprised should we be to see no invariant sites?
Very surprising, unless branches are very long

How surprised should we be to see no invariant sites?
Very surprising, unless branches are very long
but only if we looked for them!

How surprised should we be to see no invariant sites?
Very surprising, unless branches are very long
but only if we looked for them!

Can correct by applying Lewis (2001) model for analysis of only variable sites implemented inference software
Based on correction for problem of not counting un-observed restriction sites in (Felsenstein, 1992)

Neat widgets created by Mark Holder:

http://phylo.bio.ku.edu/mephytis/brlen-opt.html

http://phylo.bio.ku.edu/mephytis/tree-opt.html

How do algorithms for Maximum Likelihood phylogenetics estimation solve these problems?

How do you know that you have gotten the ML tree?

How do algorithms for Maximum Likelihood phylogenetics estimation attempt to solve these problems?

How do you know that you have gotten the ML tree?
you don't!

How do algorithms for Maximum Likelihood phylogenetics estimation attempt to solve these problems?

How do you know that you have gotten the ML tree?
you don't!

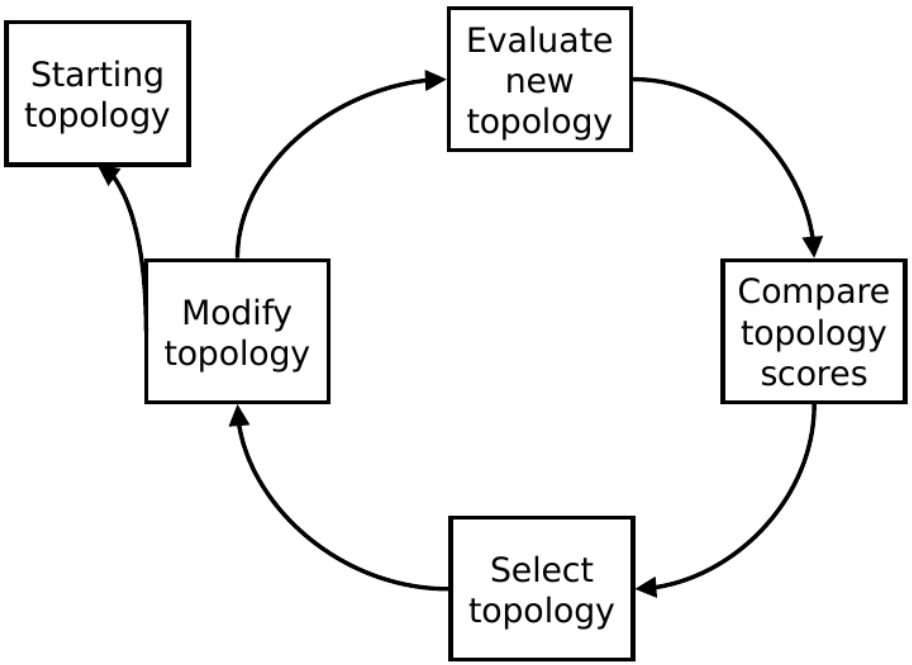
Use a heuristic search to find the best tree you can.

The general concept of heuristic tree search:

1. Start with a tree
2. Calculate the likelihood of that tree given your data (alignment)
3. Look at some trees that are similar
4. Calculate the likelihood for those trees
5. See if you did any better! Return to step 3.

Heuristic runtimes

Inference time = # of topologies to evaluate x time to evaluate each



Both are strongly a function of the # of sequences when calculating maximized likelihood

(From Zwickl)

Questions for a heuristic search:

- Where to start the search?

Questions for a heuristic search:

- Where to start the search?
- How are new trees proposed?

Questions for a heuristic search:

- Where to start the search?
- How are new trees proposed?
- How do we decide to continue looking at trees at are similar to your new tree, or to your old tree?

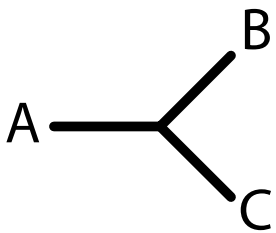
Questions for a heuristic search:

- Where to start the search?
- How are new trees proposed?
- How do we decide to continue looking at trees at are similar to your new tree, or to your old tree?
- How do you know if you are done?

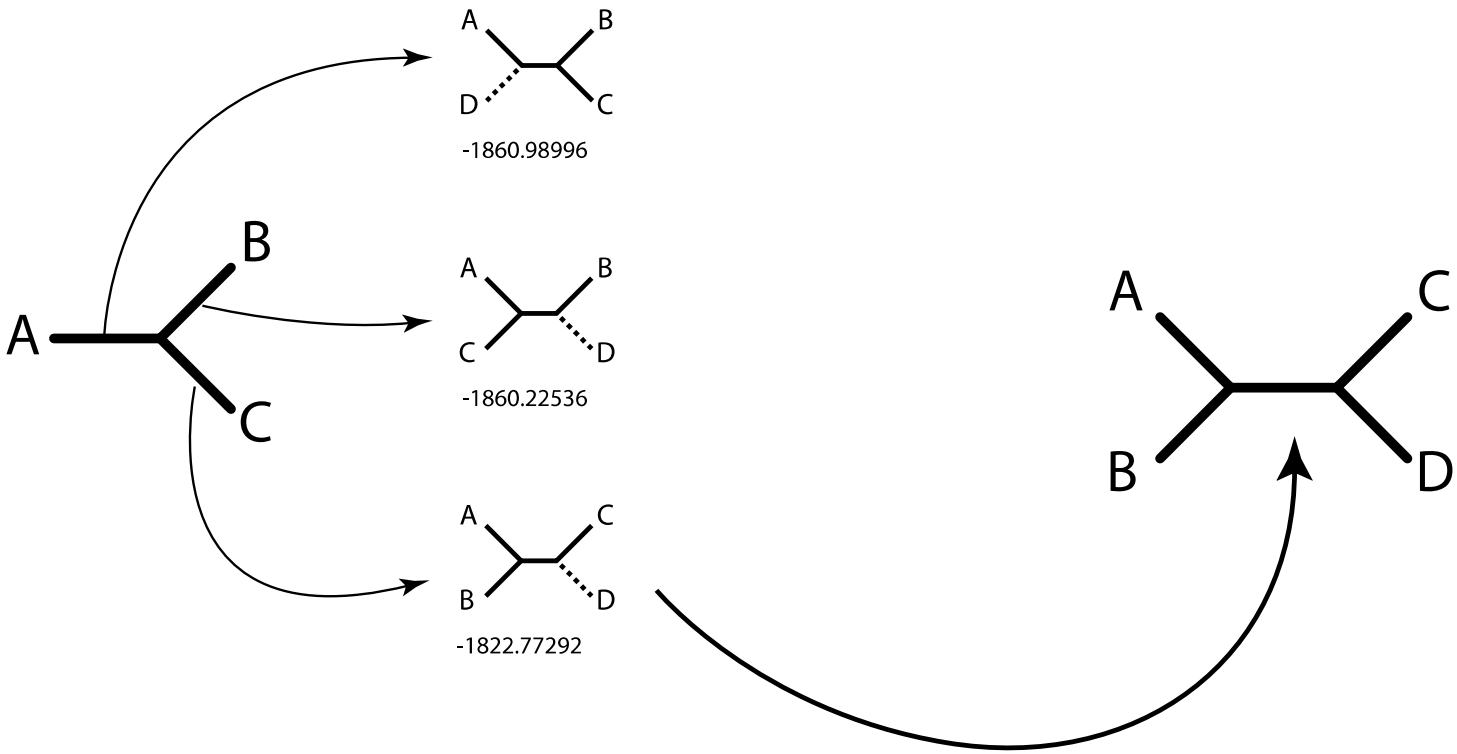
Where to start the search?

- User supplied starting tree
- Star decomposition or Stepwise Addition
- A randomly chosen tree

Stepwise addition

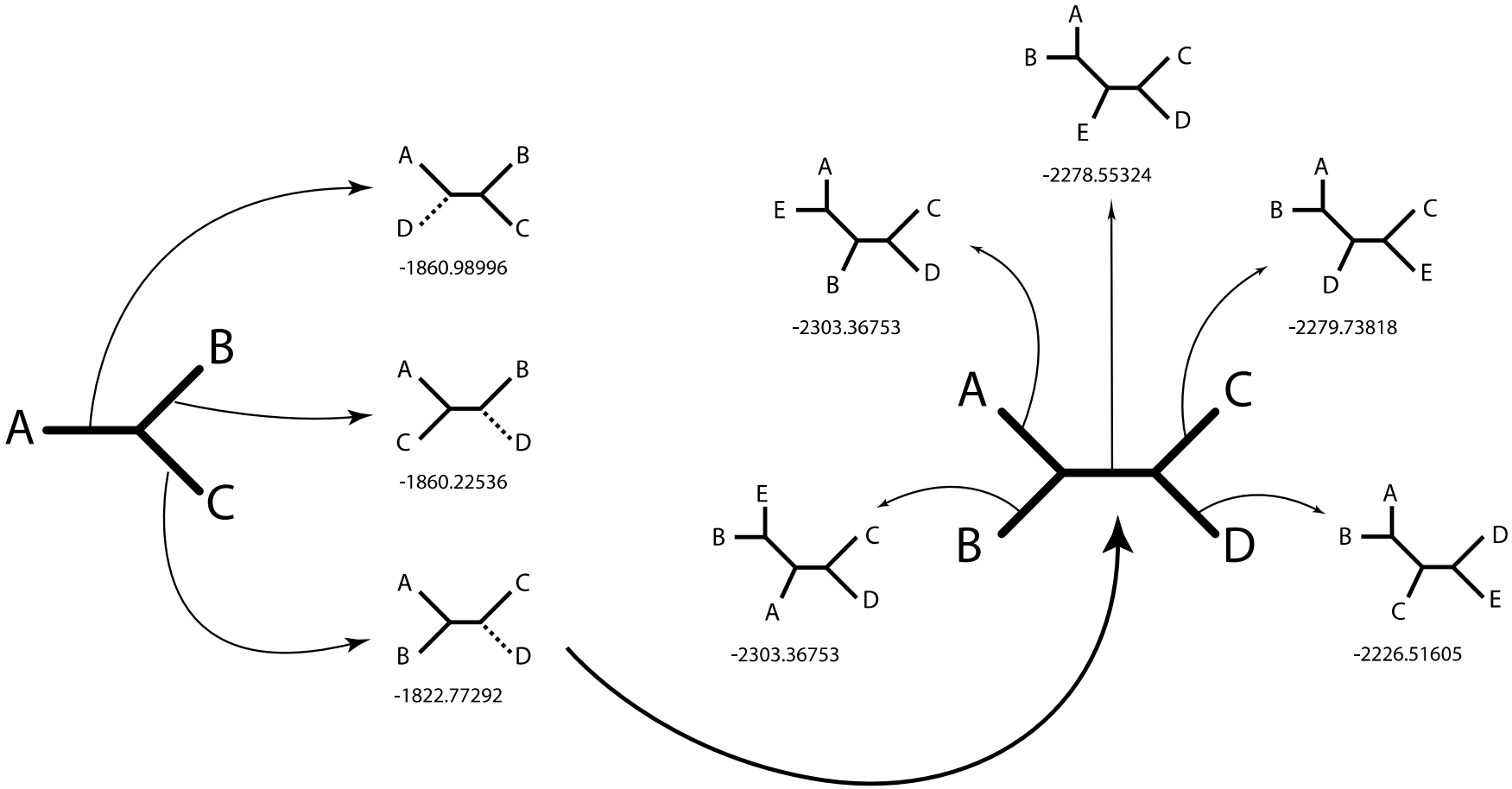


Stepwise addition



(slide from POL)

Stepwise addition



(slide from POL)

Stepwise addition

- Greedy, but can introduce a new taxon on the path between taxa that have already been joined.
- The tree can depend on the input order of the taxa
- Number of trees scored for N taxa :

$$\begin{aligned} \#\text{ trees scored} &= \sum_{i=3}^{N-1} (2i-3) \\ &= (N-1)(N-3) \end{aligned}$$

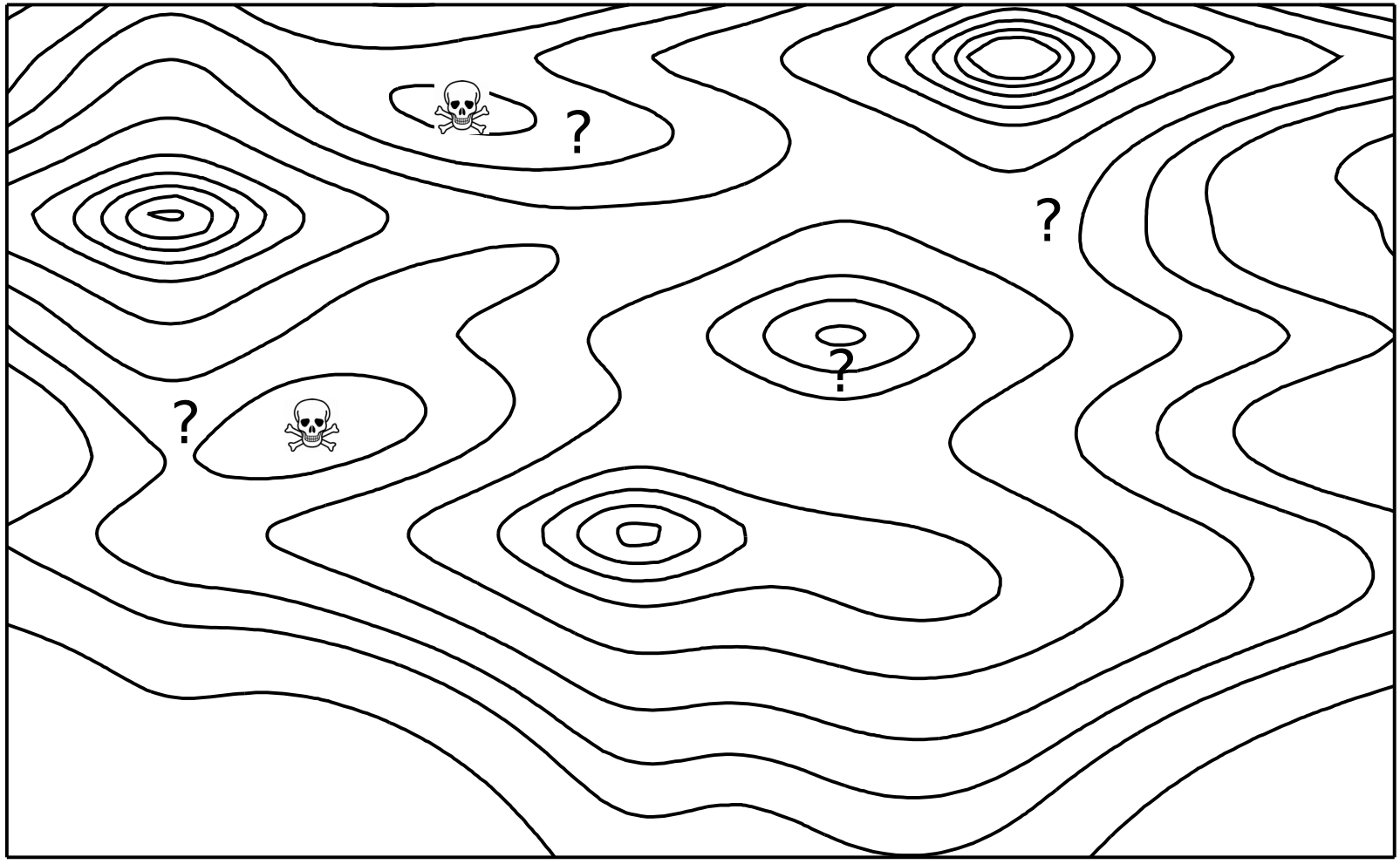
Thus, stepwise addition is $O(N^2)$. For N=10:

$$63 = 3 + 5 + 7 + 9 + 11 + 13 + 15$$

Does your starting tree matter?

- Can help escape local optima
- When data is uninformative, bias in starting tree can affect estimate

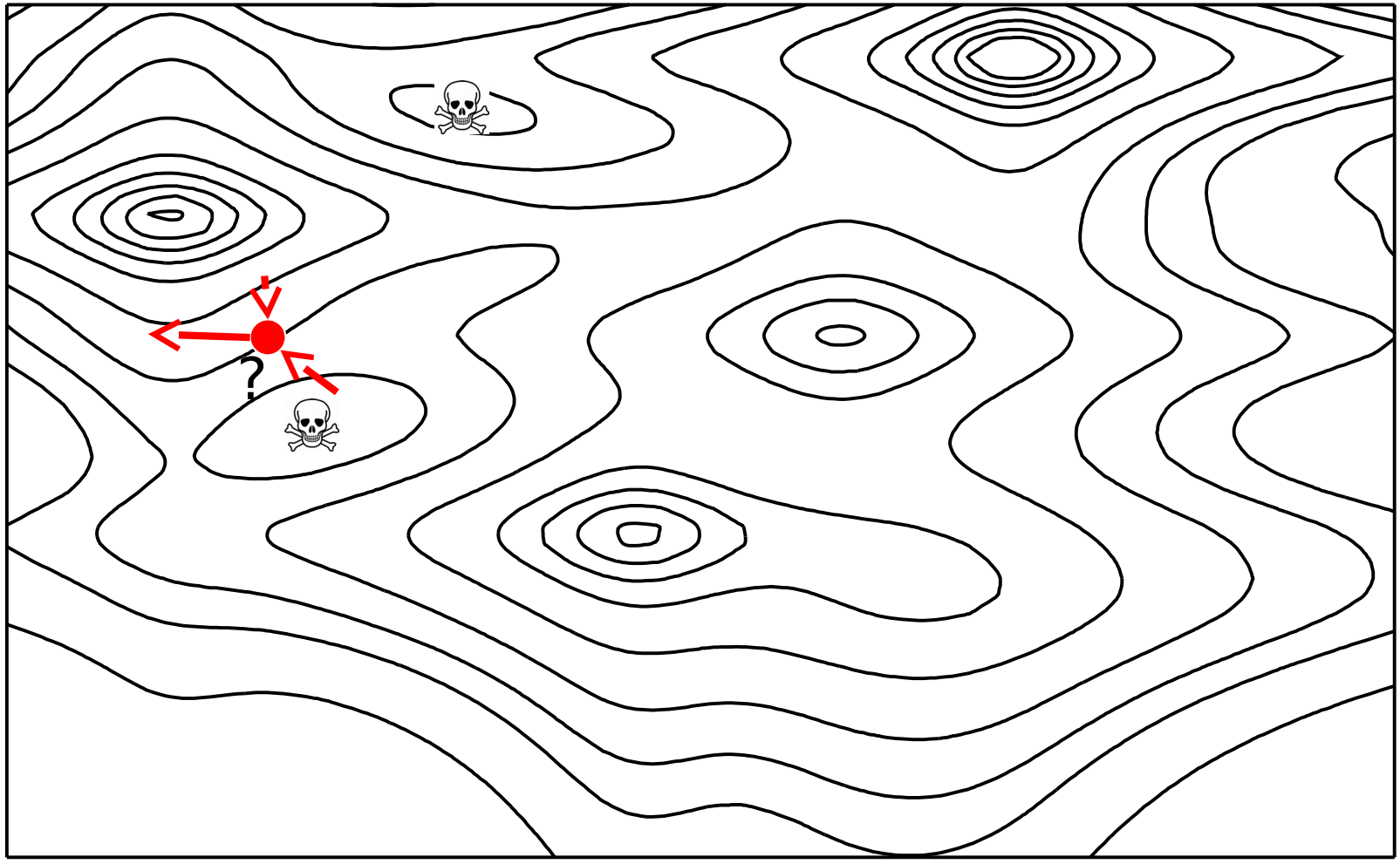
Heuristics: starting point



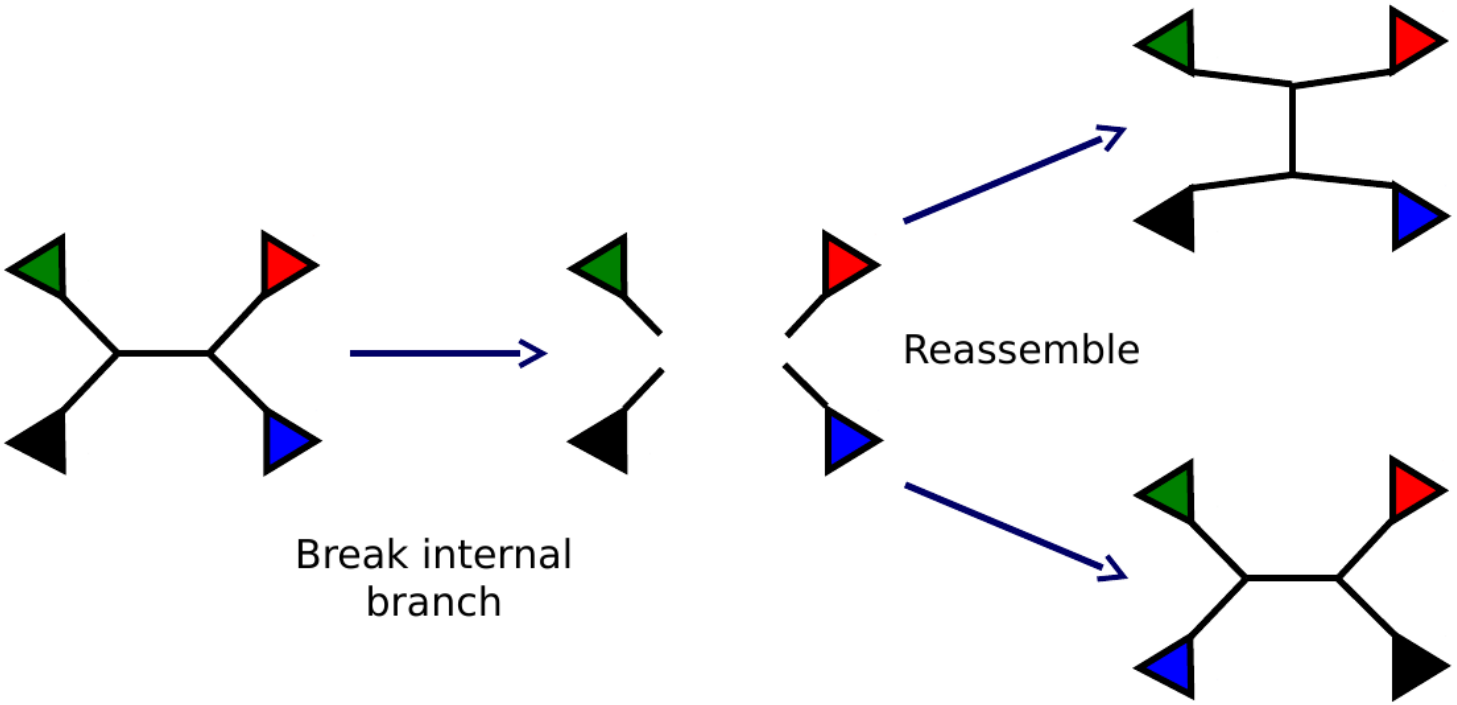
How are new topologies proposed?

- Branch swapping and tree rearrangement

Heuristics: proposing new values

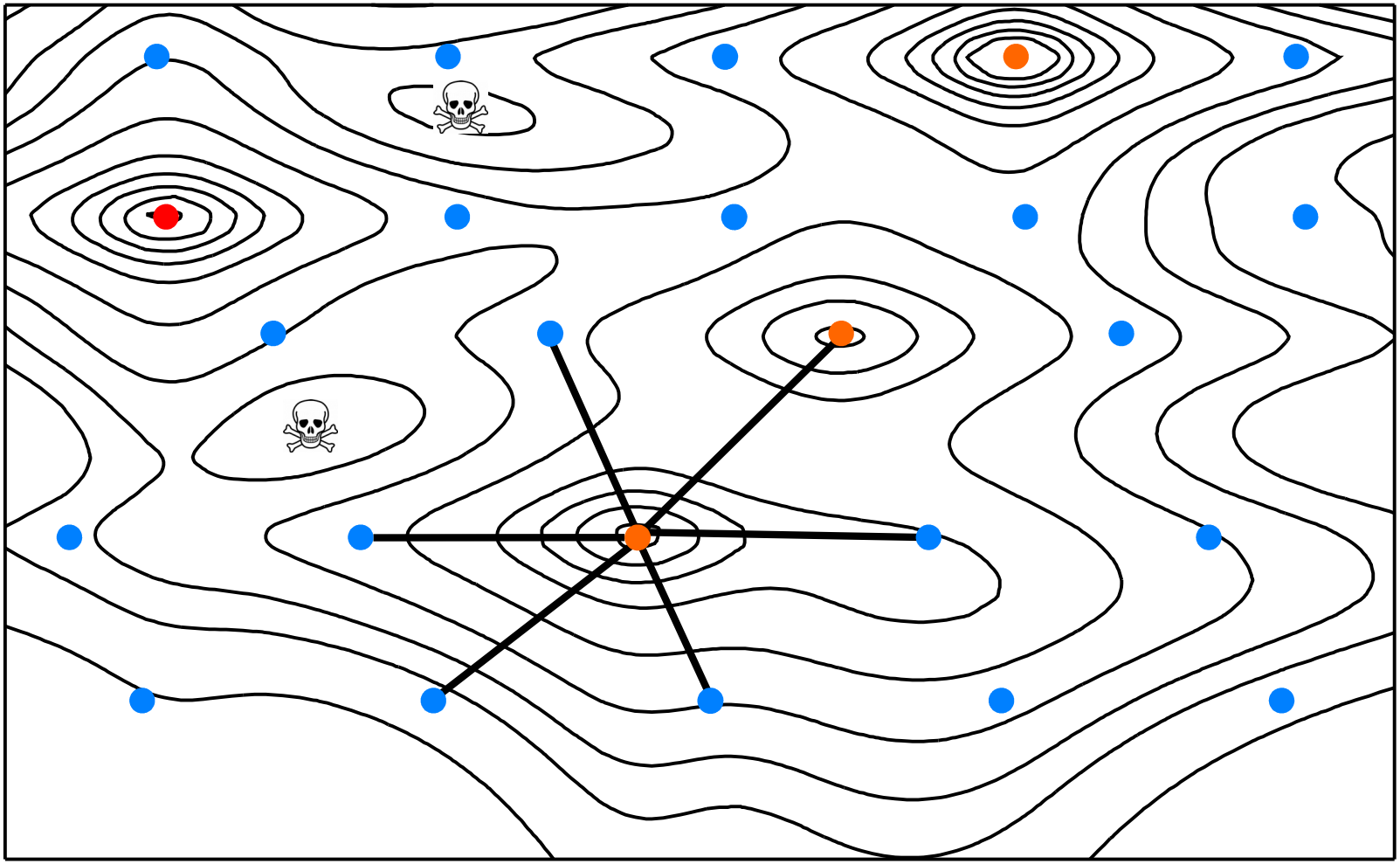


Nearest neighbor interchange (NNI)



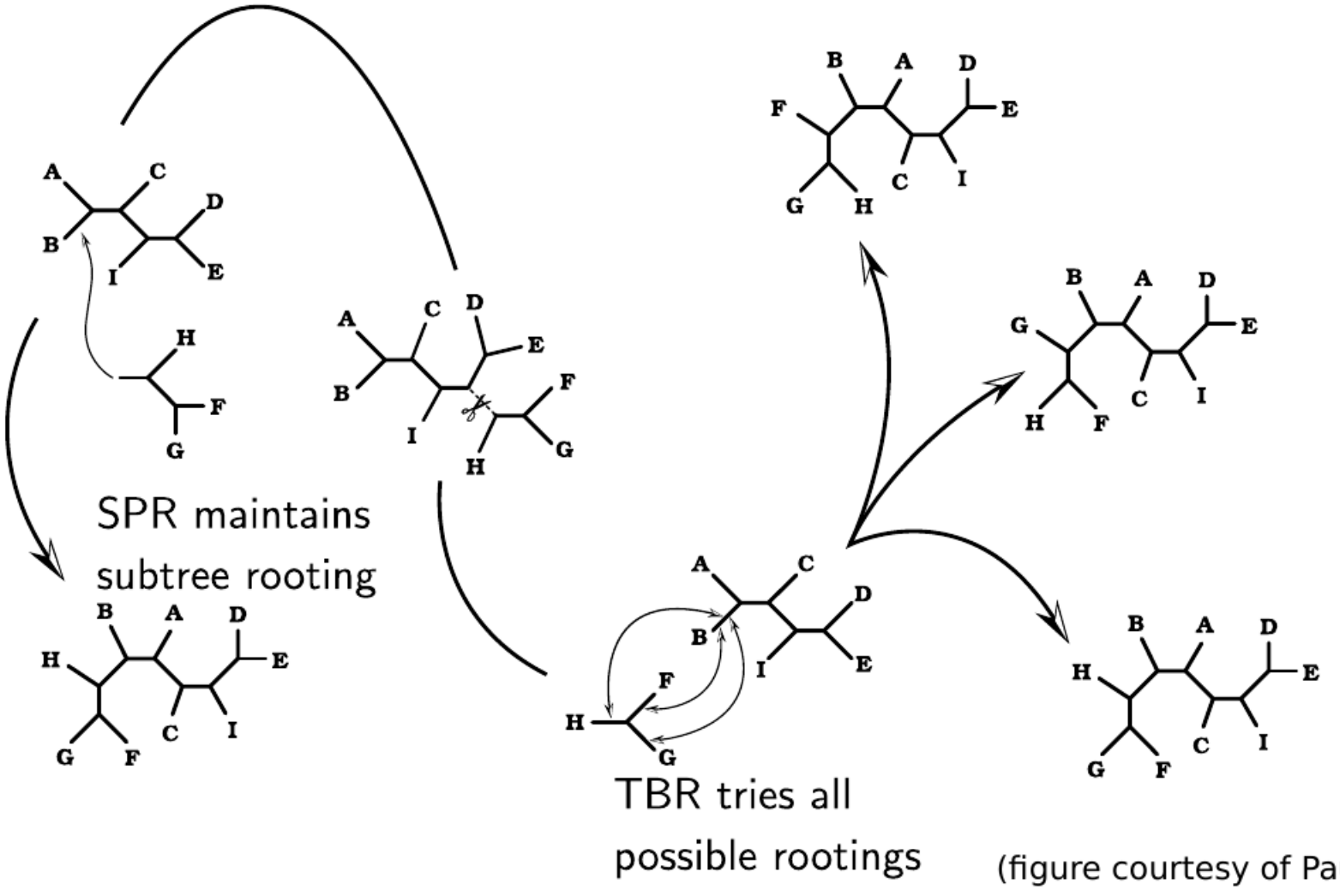
(From Zwickl)

NNI Treespace



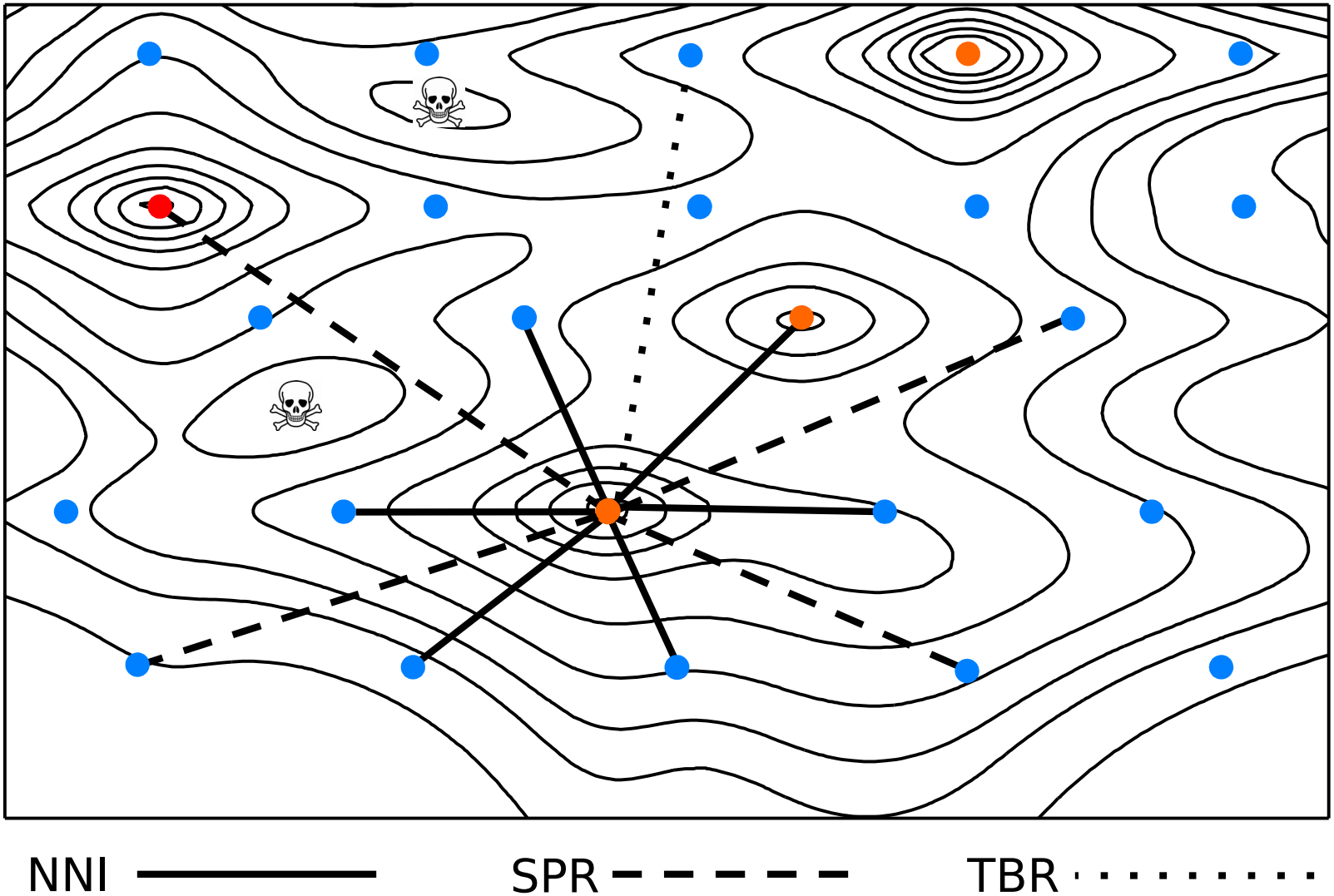
NNI

Subtree Pruning Regrafting (SPR) Tree Bisection Reconnection (TBR)



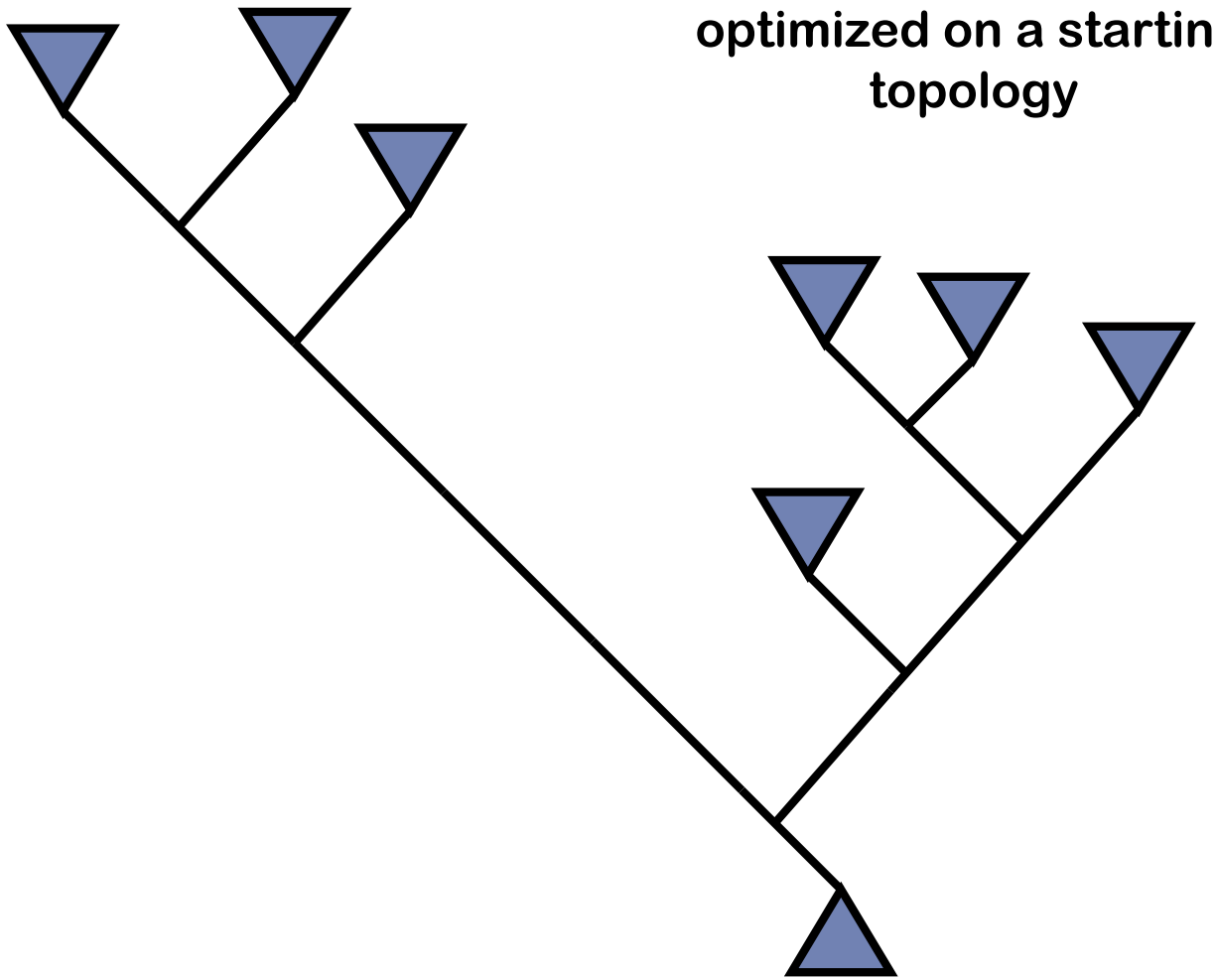
(figure courtesy of Paul Lewis)

SPR/TBR moves in NNI treespace

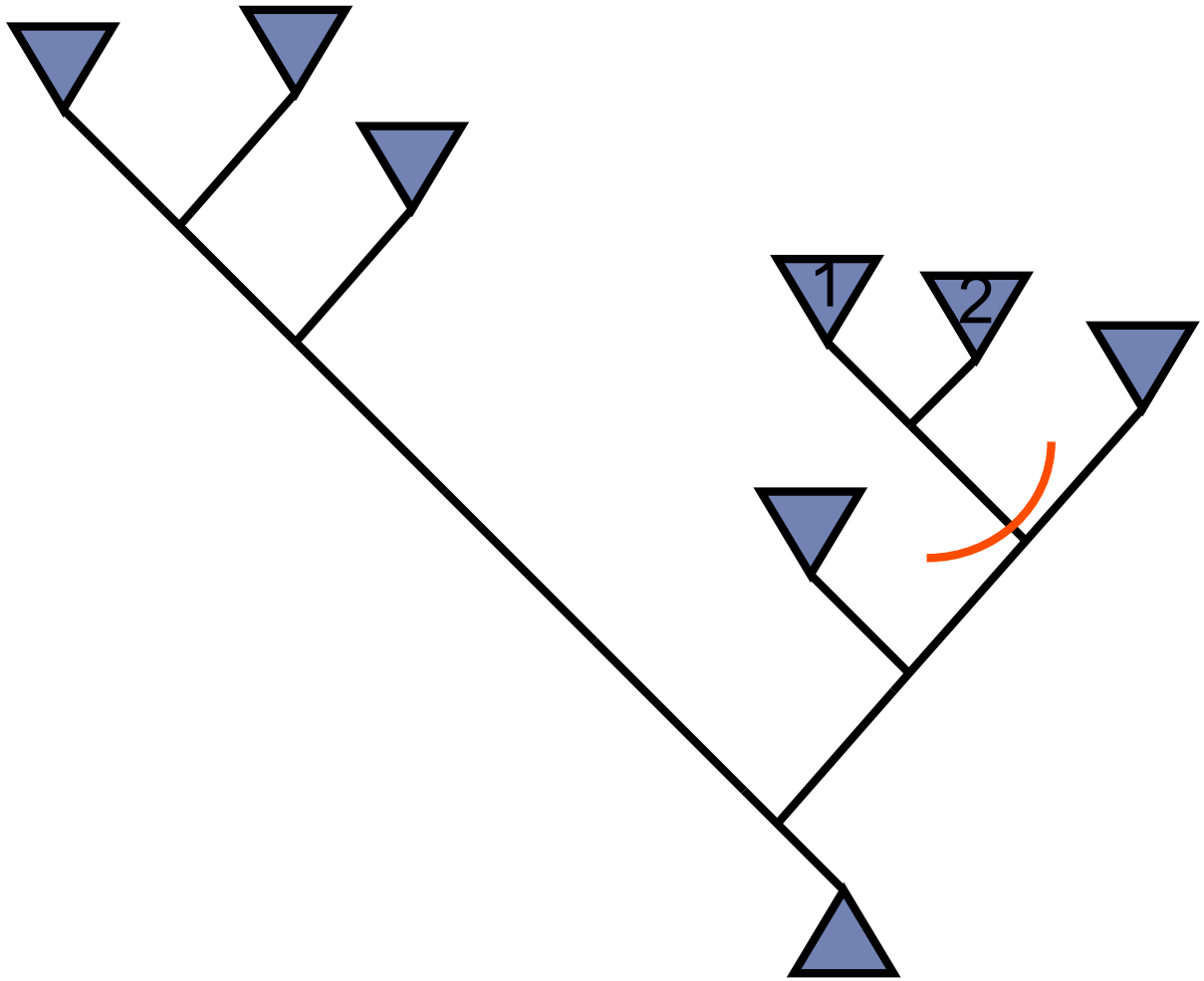


Re-arranging your tree requires updating branch lengths and evolutionary model parameters.

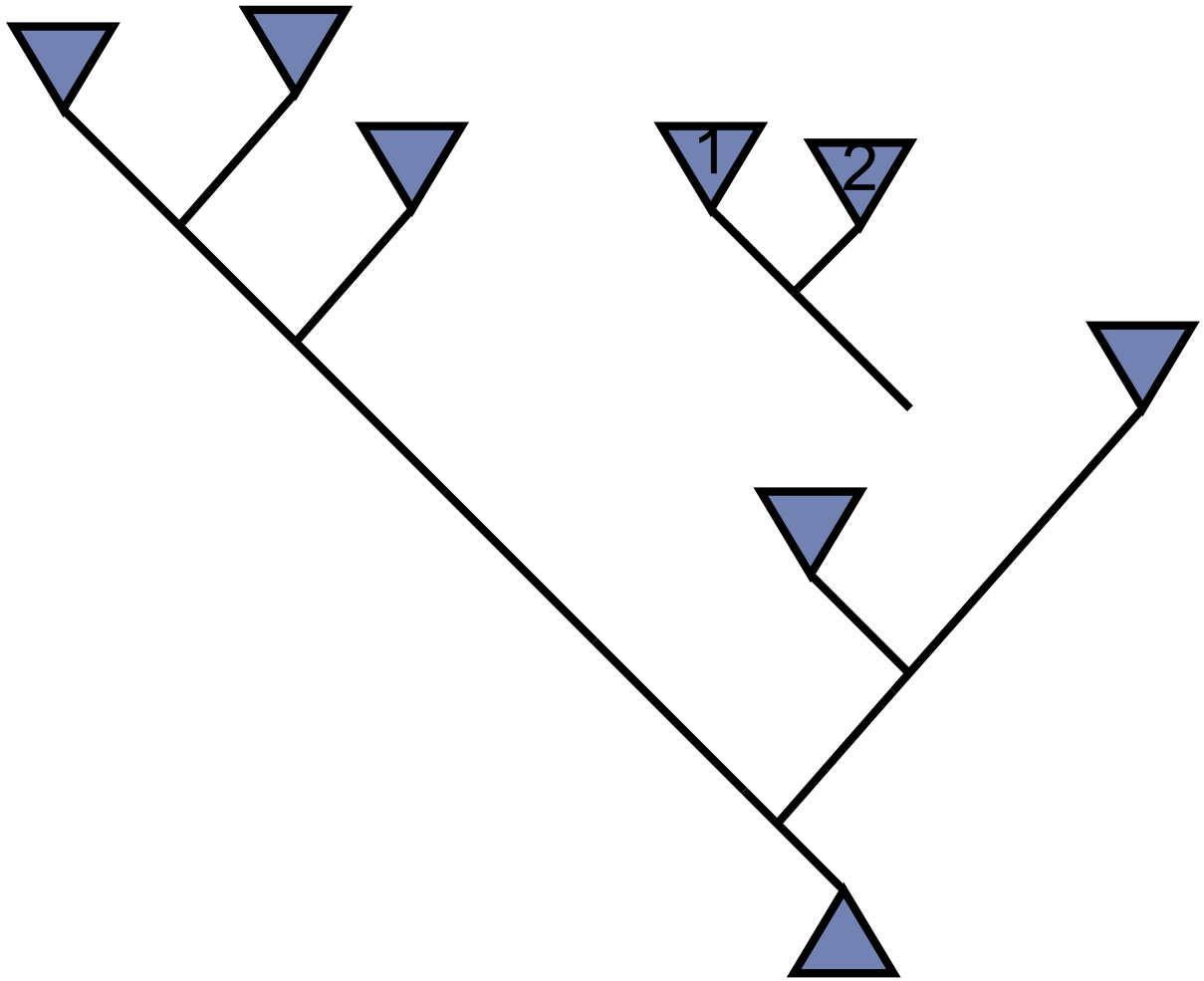
Searching with approximate likelihoods



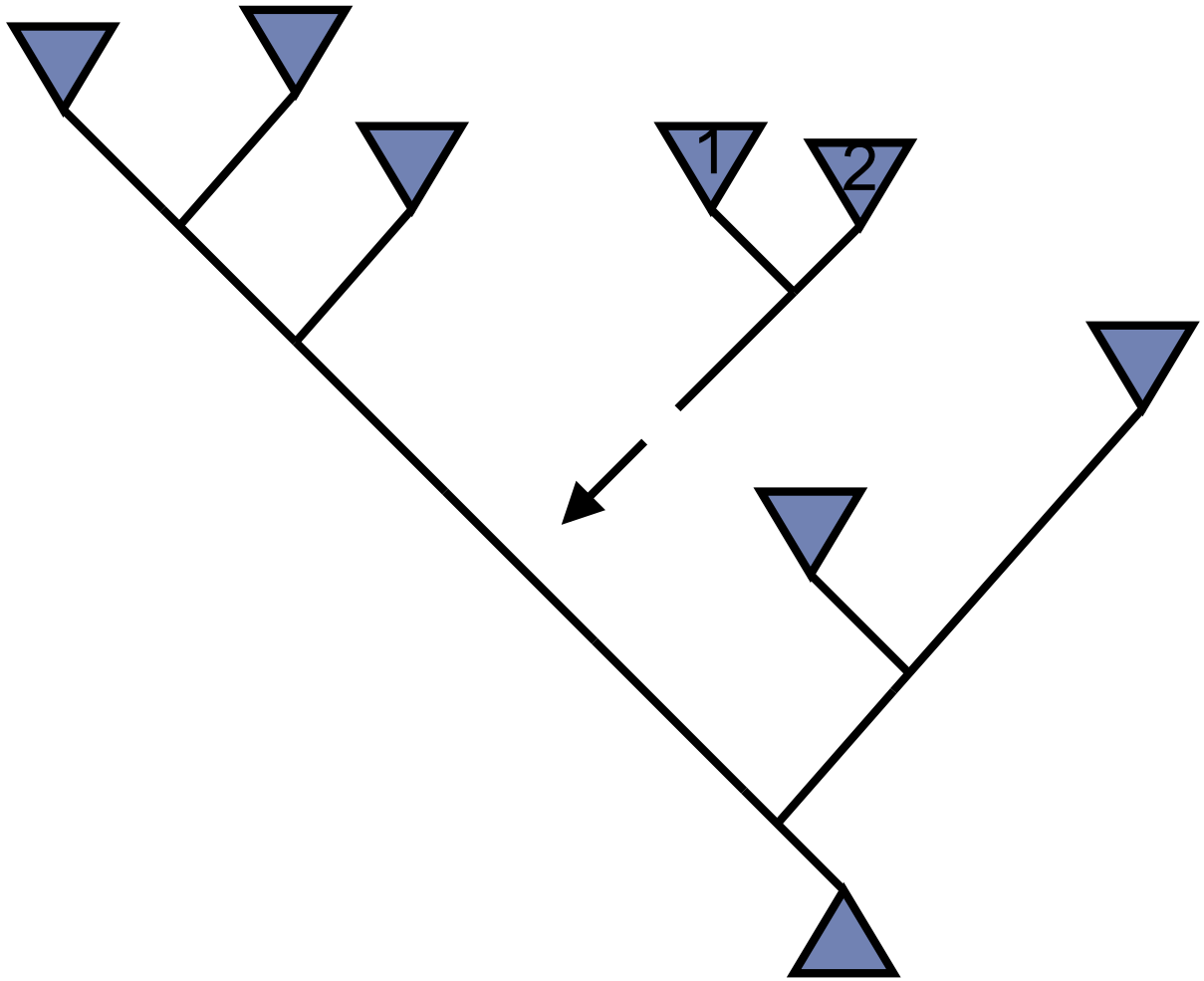
Altering the tree: subtree pruning-regrafting (SPR)



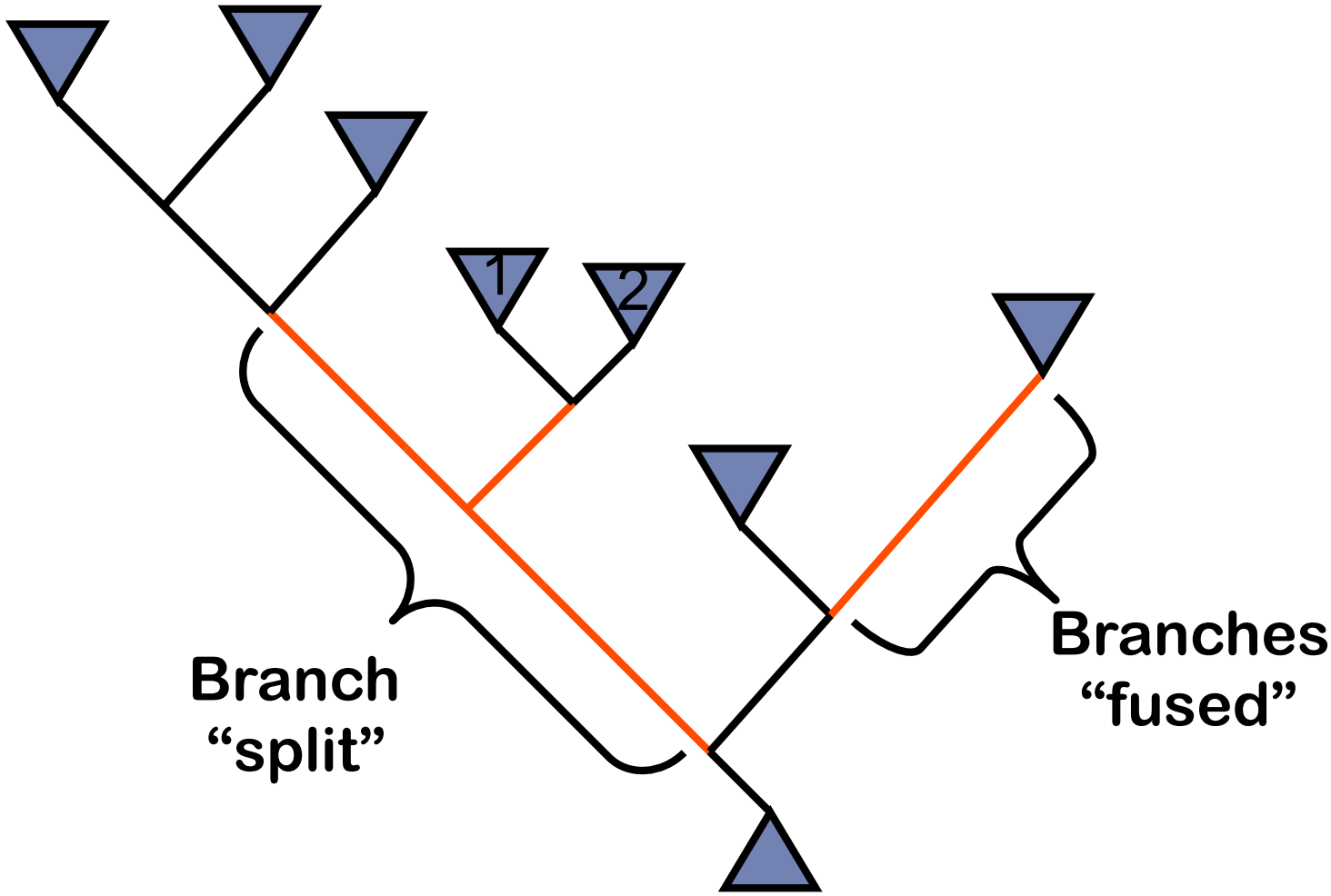
Altering the tree: subtree pruning-regrafting (SPR)



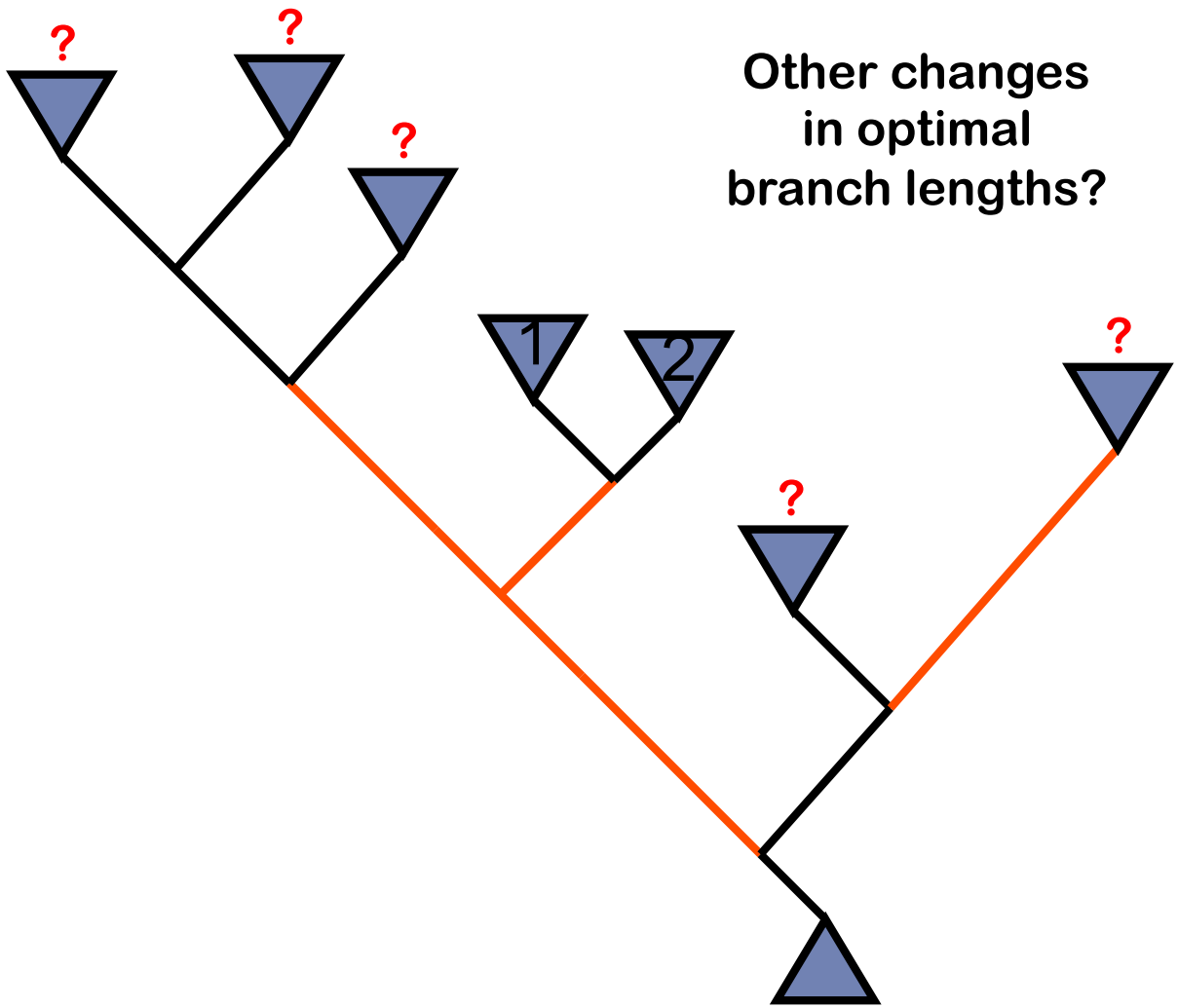
Altering the tree: subtree pruning-regrafting (SPR)



Scoring and optimizing the new topology



Scoring and optimizing the new topology



Localizing branch length optimization important for speed of analysis

How do you decide if you should accept a new tree?

- Hill climbing: likelihood score is better (RAxML)
- Computational analog of evolution by natural selection (Garli)

How do you know if you are done?

- Stop tree search when likelihood stops improving.

How do you know if you are done?

- Stop tree search when likelihood stops improving.
- Searches are stochastic, so there is no guarantee that any search finds the true maximum likelihood topology and parameter values!

How do you know if you are done?

- Stop tree search when likelihood stops improving.
- Searches are stochastic, so there is no guarantee that any search finds the true maximum likelihood topology and parameter values!
- Continue searching until you run at least one additional search that finds the same topology as the best overall result.

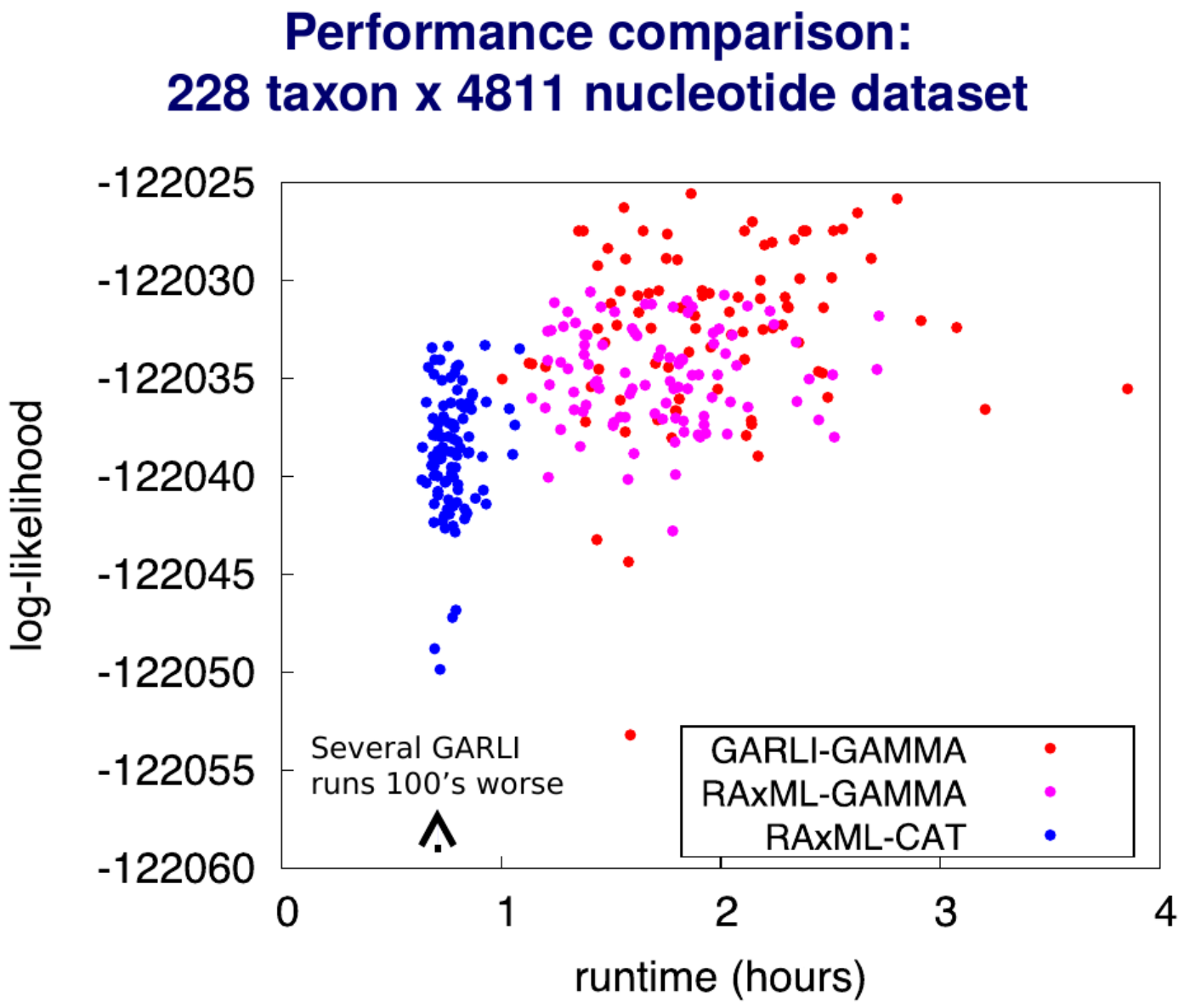
In lab today we will discuss and apply two software packages that estimate ML trees

- Garli (Zwickl, 2006)
  - Stochastic, genetic algorithm-like approach
  - Computational analog of evolution by natural selection.
  - Not currently under active development, but valuable for understanding tree searching
- RAxML (Stamatakis, 2006)
  - Hill-climbing algorithm
  - GTR+CAT approximation major speedup over GTR+G
    - For modeling rate heterogeneity across very large trees (e.g., hundreds of taxa), and is not recommended for smaller trees.
    - Different than Lartillot CAT model using empirical amino acid profiles (named independently around same time)

ML tree inference software:
For small datasets (< 50 taxa), all of the ML tree inference programs perform well
For large datasets (hundreds of sequences):

- PAUP* is very rigorous, but slowest
- RAxML is generally the fastest
- GARLI often has a slight edge over RAxML in optimality (although often more variability)

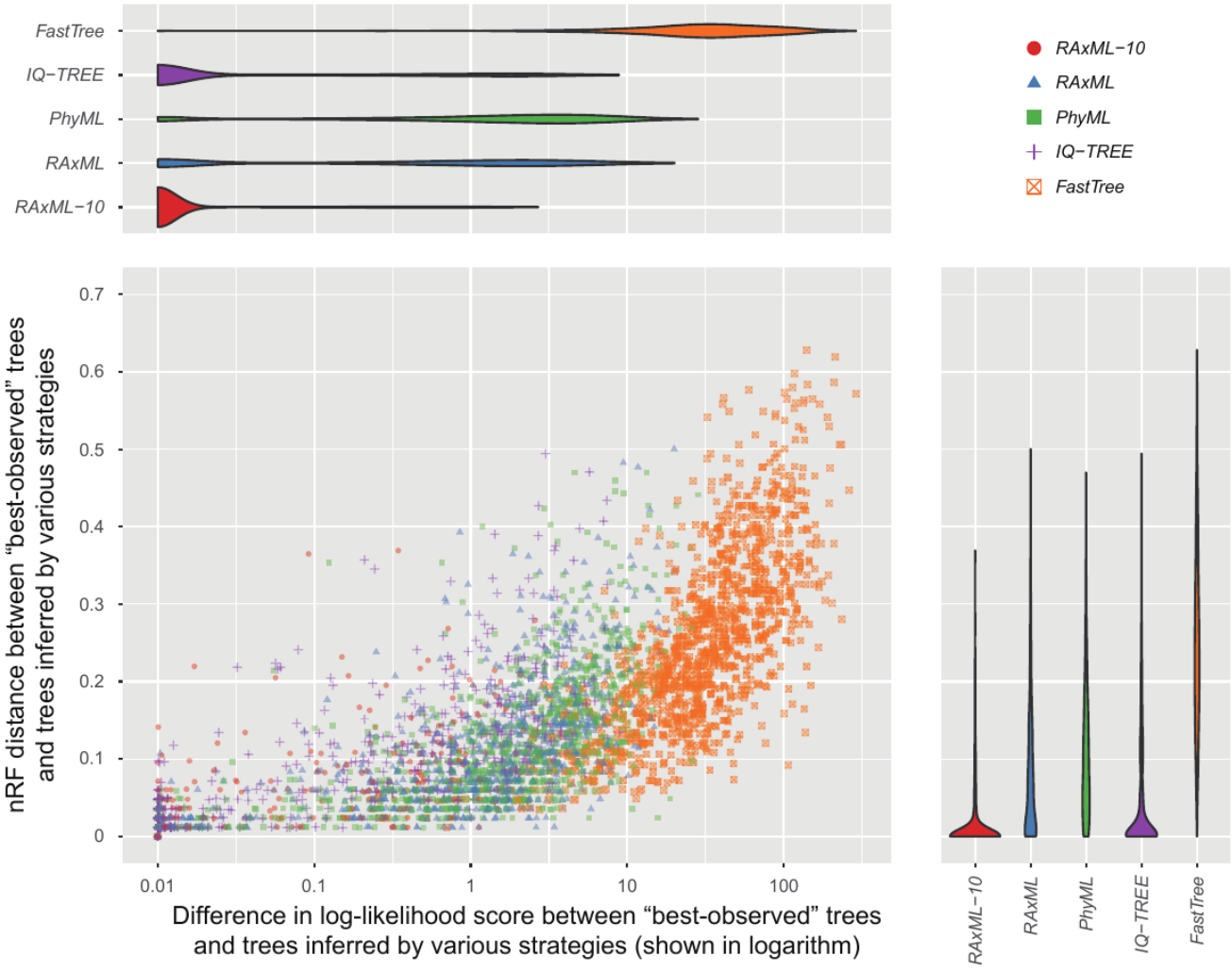
Simulations by Zwickl (Garli)



ML tree inference software:
For VERY large datasets (1000+sequences):

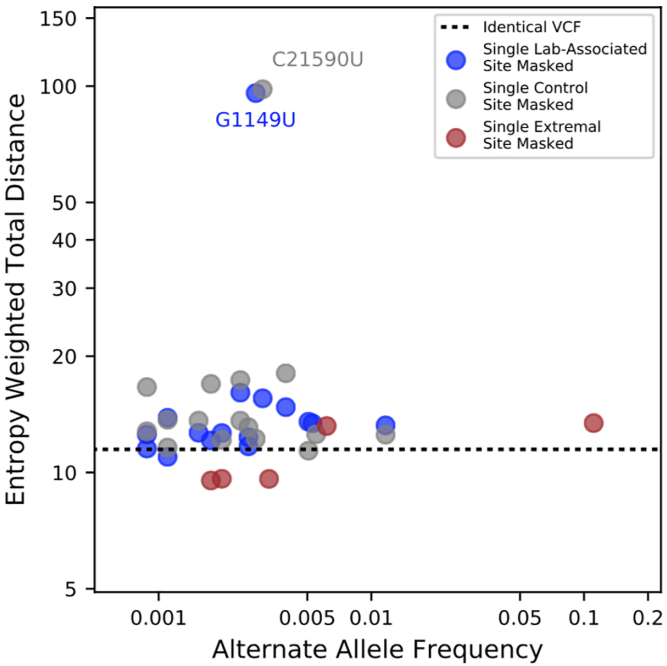
- RAxML/EXaML (Kozlov et al., 2015) is very efficient, especially with multiple runs
- IQ-TREE (Nguyen et al., 2015) also fast and relatively accurate
- FASTTREE(Price et al., 2009) is very fast, but (excessive) tradeoffs with accuracy (per Zhou et al. (2017))

Figure 3



Log-likelihood score differences between inferred trees and "best-observed" trees plotted against topological distances.(Zhou et al., 2017)

Example of local optima:



Phylogenies made after removing two variants, one control and one lab-associated, are outliers for entropy-weighted total distance (Fig 6, S4 Fig) and other tree distance statistics (S3 Fig). In each case, however, the likelihood of the tree produced from the full dataset is actually higher (S5 Table), suggesting that our tree-building method discovered a different locally optimal but less favorable topology rather than a dramatic impact of each site individually. These results suggest higher level uncertainty in the tree topology largely independent of the effects of lab-associated variants.

Fig 6. The relationship between alternate allele frequencies of lab-associated variants and effect of masking on inferred tree topology. Entropy-weighted total distances relative to the reference maximum likelihood phylogeny are shown for phylogenies constructed after masking individual sites. Blue points correspond to sites with lab-specific alternate alleles, grey points correspond to control sites with parsimony scores of 1 and similar alternate allele frequencies to the sites with lab-specific alternate alleles, and brown points correspond to non-lab-specific extremal sites. The black horizontal line indicates the entropy-weighted total distance value for a maximum likelihood phylogeny constructed from an alignment identical to that of the reference phylogeny. Two outliers, C21590U (control) and G1149U (lab-associated), have outsize effects on inferred tree topology.
https://doi.org/10.1371/journal.pgen.1009175.g006

Summary

- For >15 sequences, an unfathomably large number of trees are possible.
- We have to rely on heuristics that are not guaranteed to find the actual ("global") optimal solution.
- We have control on how thorough our searches are
- You should conduct multiple searches to look for evidence that you are not finding trees which are local optima.

Questions?

Measuring differences between trees:

- Robinson–Foulds (RF) distance: (A + B) where A is the number of splits in the first tree but not the second tree and B is the number of splits second tree but not the first tree.
- Weighted RF distance: RF distance weighted by edge lengths

Computer lab takehomes:

- Perform ML phylogenetics search
- Compare searches
- Work with variety of phylogenetic software and file formats
- Bootstrapping
- Consequences of model misspecification

Felsenstein, J. (1992). Phylogenies from Restriction Sites: A Maximum-Likelihood Approach. *Evolution*, 46(1):159–173.

Kozlov, A. M., Aberer, A. J., and Stamatakis, A. (2015). ExaML version 3: a tool for phylogenomic analyses on supercomputers. *Bioinformatics*, 31(15):2577–2579.

Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic biology*, 50(6):913–925.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, 32(1):268–274.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2009). FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution*, 26(7):1641–1650.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690.

Zhou, X., Shen, X.-X., Hittinger, C. T., and Rokas, A. (2017). Evaluating Fast Maximum Likelihood-Based Phylogenetic Programs Using Empirical Phylogenomic Data Sets. *bioRxiv*, page 142323.

Zwickl, D. J. (2006). GARLI—genetic algorithm for rapid likelihood inference. *See http://www. bio. utexas. edu/faculty/antisense/garli/Garli. html*.