Maximum Likelihood Inference

Modified from
https://evomics.org/workshops/2019-workshop-on-phylogenomics-cesky-krumlov/

Why is tree inference so difficult?

- Too many trees to look at

- Too many calculations to do

Likelihood function = P(data|tree)

Key to all phylogenetics!

Too many trees

No. of binary unrooted trees with nn tips: $= 1 \times 3 \times 5 \cdots \times (2nn - 3)$

$$= \frac{(2nn - 3)!}{2^{nn-2}(nn - 3)!}$$

Tips	Binary unrooted trees
5	15
10	2,027,025
20	2.22×10^{18}
30	8.69×10^{36}
40	1.31×10^{55}
50	2.84×10^{74}
⋮	⋮

Summit@ORNL

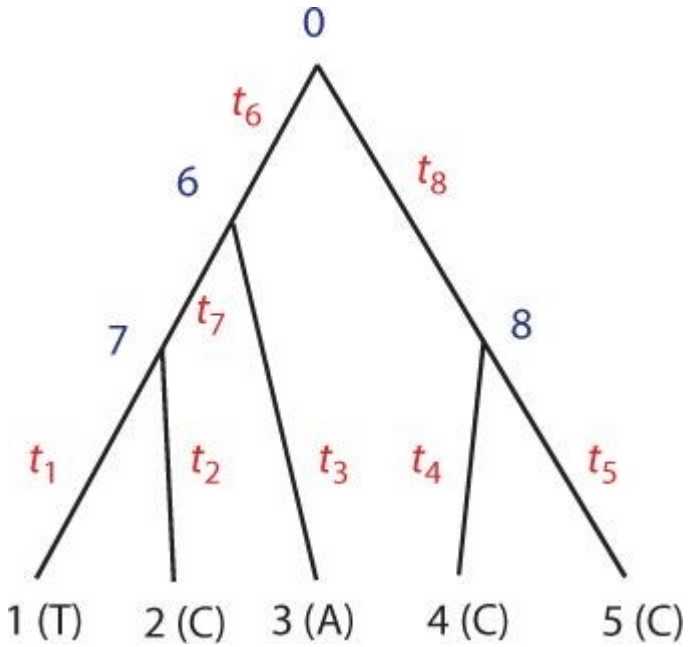
World's fastest and largest supercomputer

Peak Flops: 200.8×10^{15}

~2.07×10^{21} billion years

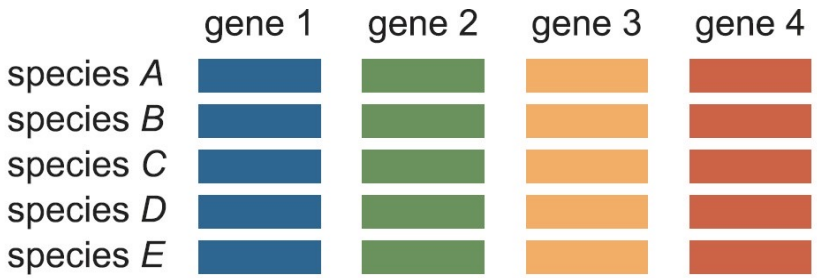
Too many calculations

- Branch length estimation
- Model parameter optimization
- ……

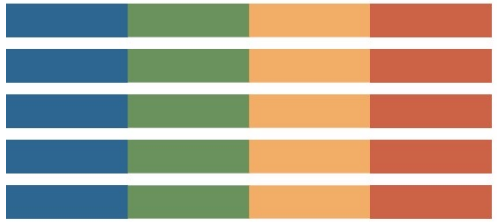


$$f(\mathbf{x}_h|\theta)=\sum_{x_0}\sum_{x_6}\sum_{x_7}\sum_{x_8}\left[\pi_{x_0}p_{x_0x_6}(t_6)p_{x_6x_7}(t_7)p_{x_7T}(t_1)p_{x_7C}(t_2)p_{x_6A}(t_3)p_{x_0x_8}(t_8)p_{x_8C}(t_4)p_{x_8C}(t_5)\right].$$

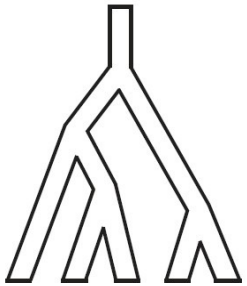
Yang Z. (2014)



concatenation



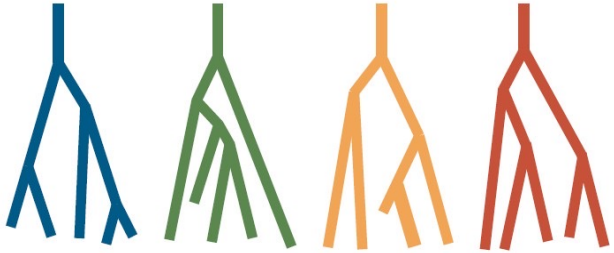
supermatrix



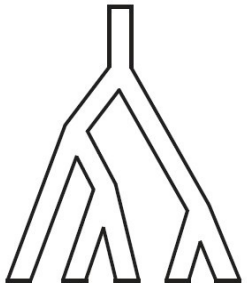
Maximum-likelihood (ML):
RAxML, IQ-TREE, PhyML, FastTree …

Bayesian Inference:
PhyloBayes, MrBayes, BEAST …

'two-step' coalescent



estimated gene trees



Substitution models

Describes how genes evolve
gives probability for each transition

DNA models:

	p	From	To T	C	A	G
JC69 (Jukes and Cantor 1969)	1	T	·	λ	λ	λ
		C	λ	·	λ	λ
		A	λ	λ	·	λ
		G	λ	λ	λ	·
K80 (Kimura 1980)	2	T	·	α	β	β
		C	α	·	β	β
		A	β	β	·	α
		G	β	β	α	·
F81 (Felsenstein 1981)	4	T	·	π_C	π_A	π_G
		C	π_T	·	π_A	π_G
		A	π_T	π_C	·	π_G
		G	π_T	π_C	π_A	·
HKY85 (Hasegawa et al. 1984, 1985)	5	T	·	$\alpha\pi_C$	$\beta\pi_A$	$\beta\pi_G$
		C	$\alpha\pi_T$	·	$\beta\pi_A$	$\beta\pi_G$
		A	$\beta\pi_T$	$\beta\pi_C$	·	$\alpha\pi_G$
		G	$\beta\pi_T$	$\beta\pi_C$	$\alpha\pi_A$	·
F84 (Felsenstein, DNAML program since 1984)	5	T	·	$(1+\kappa/\pi_Y)\beta\pi_C$	$\beta\pi_A$	$\beta\pi_G$
		C	$(1+\kappa/\pi_Y)\beta\pi_T$	·	$\beta\pi_A$	$\beta\pi_G$
		A	$\beta\pi_T$	$\beta\pi_T$	·	$(1+\kappa/\pi_R)\beta\pi_G$
		G	$\beta\pi_T$	$\beta\pi_C$	$(1+\kappa/\pi_R)\beta\pi_A$	·
TN93 (Tamura and Nei 1993)	6	T	·	$\alpha_1\pi_C$	$\beta\pi_A$	$\beta\pi_G$
		C	$\alpha_1\pi_T$	·	$\beta\pi_A$	$\beta\pi_G$
		A	$\beta\pi_T$	$\beta\pi_C$	·	$\alpha_2\pi_G$
		G	$\beta\pi_T$	$\beta\pi_C$	$\alpha_2\pi_A$	·
GTR (REV) (Tavaré 1986; Yang 1994b; Zharkikh 1994)	9	T	·	$a\pi_C$	$b\pi_A$	$c\pi_G$
		C	$a\pi_T$	·	$d\pi_A$	$e\pi_G$
		A	$b\pi_T$	$d\pi_C$	·	$f\pi_G$
		G	$c\pi_T$	$e\pi_C$	$f\pi_A$	·

Protein models:

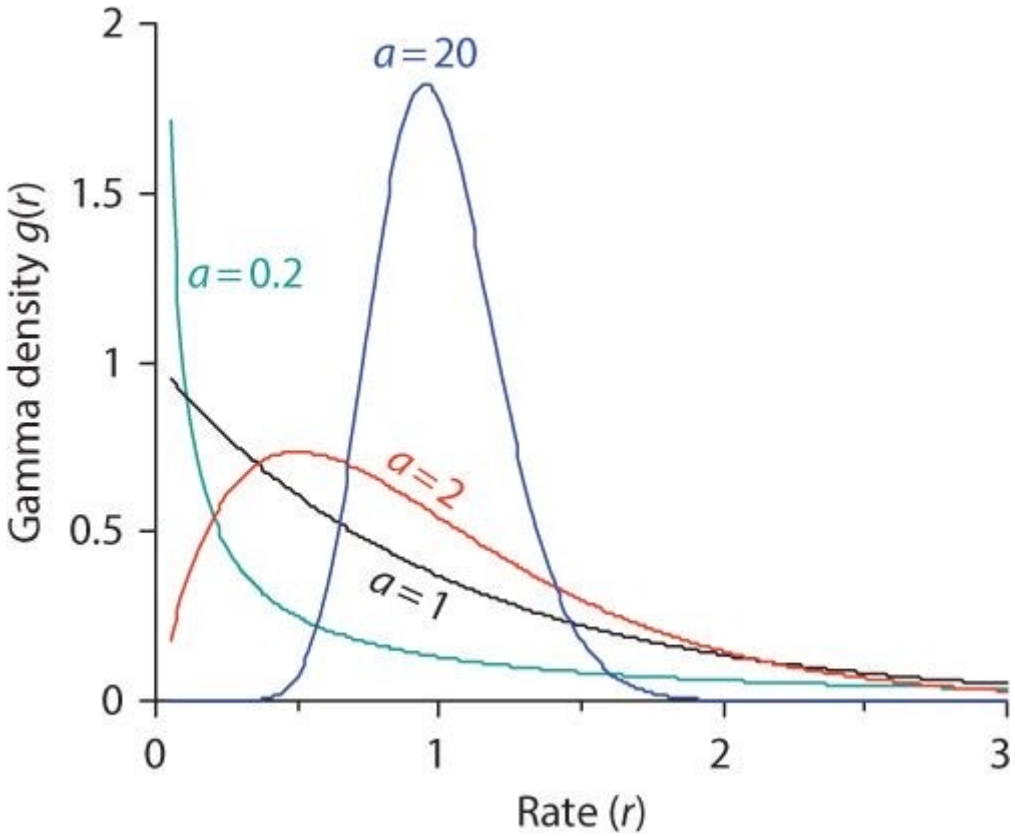
- Empirical model
  - exchangeability matrix
  - equilibrium frequencies

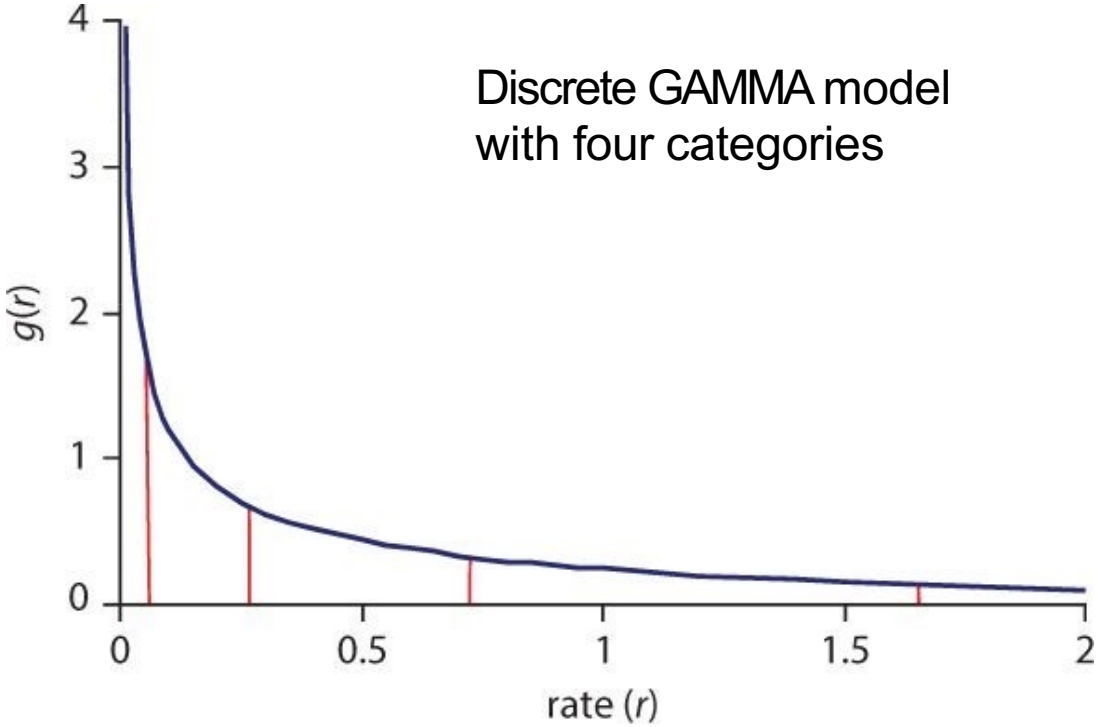
	p	From	To T	C	A	G
JC69 (Jukes and Cantor 1969)	1	T	·	λ	λ	λ
		C	λ	·	λ	λ
		A	λ	λ	·	λ
		G	λ	λ	λ	·
GTR (REV) (Tavaré 1986; Yang 1994b; Zharkikh 1994)	9	T	·	$a\pi_C$	$b\pi_A$	$c\pi_G$
		C	$a\pi_T$	·	$d\pi_A$	$e\pi_G$
		A	$b\pi_T$	$d\pi_C$	·	$f\pi_G$
		G	$c\pi_T$	$e\pi_C$	$f\pi_A$	·

Yang Z. (2014)

GAMMA model for rate variation

Alpha is the shape paramter

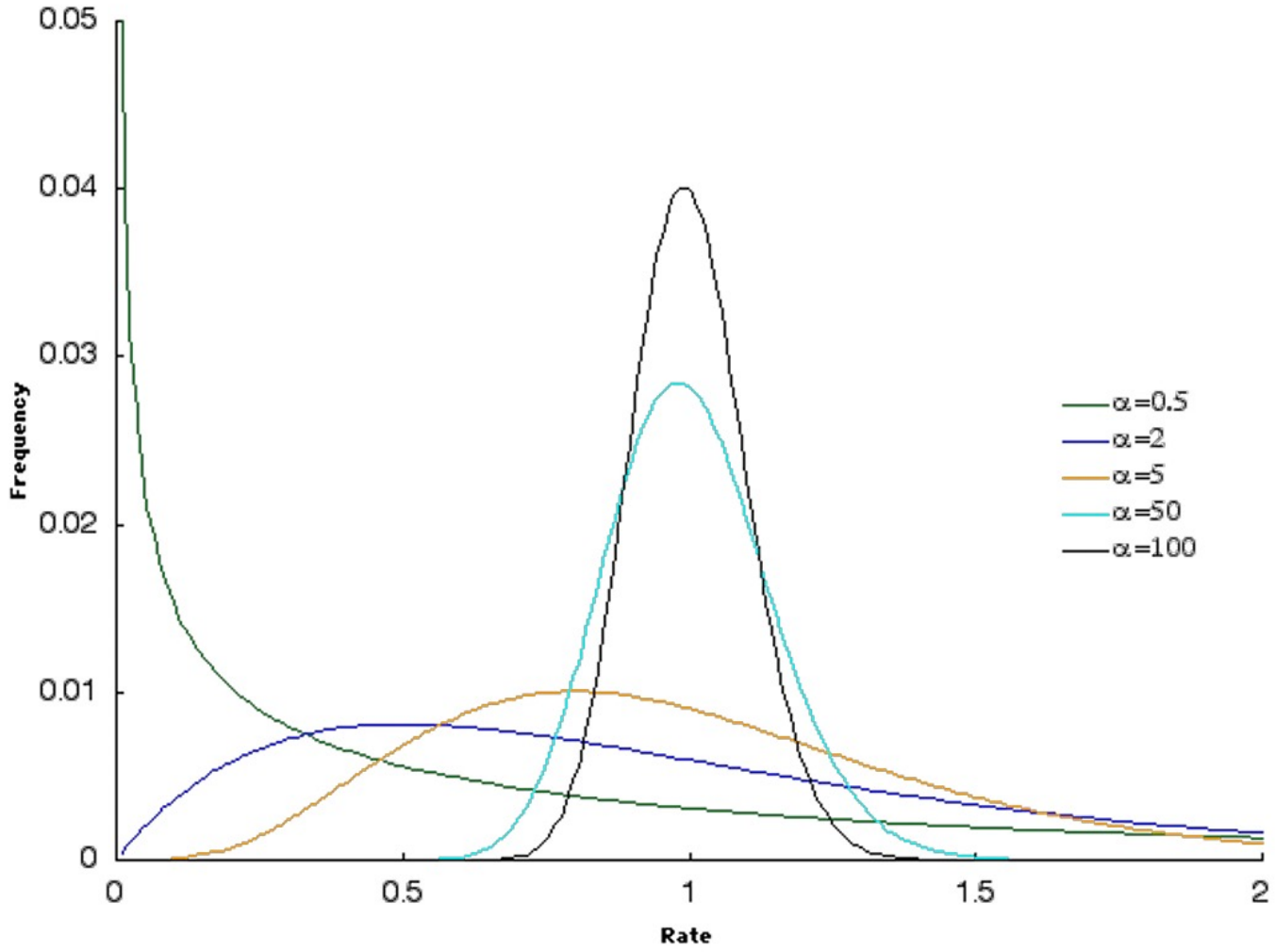




Discrete GAMMA model
with four categories

The basic idea is that the rate at each site is drawn independently from a distribution of rates. The most widely used choice is the Gamma distributiont

Yang Z. (2014)



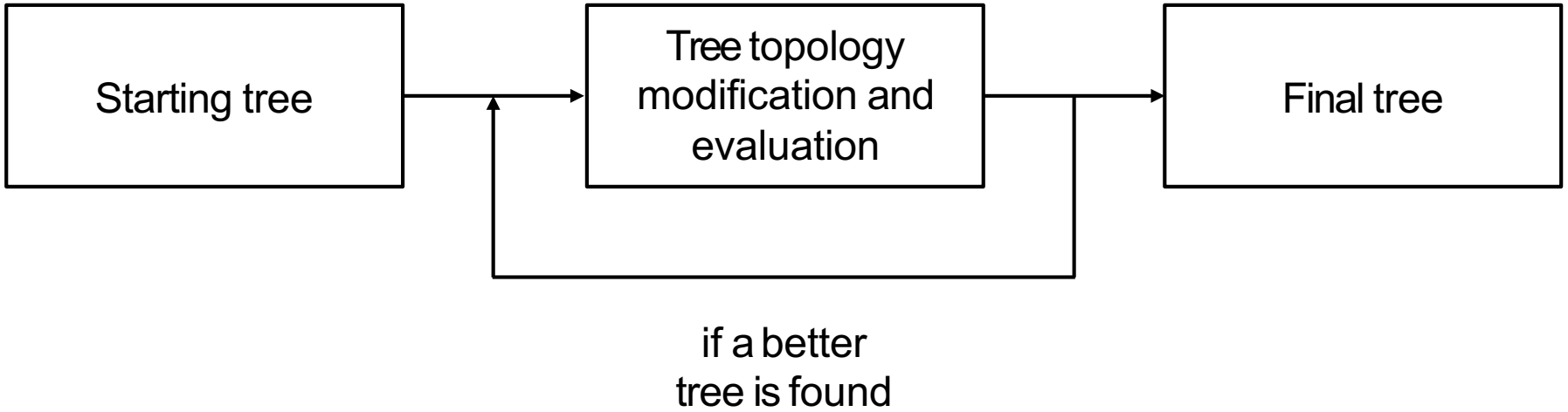
Gamma distributions are governed by two parameters: a shape parameter (Alpha) and a scale parameter (Beta). The mean of a G-distribution is equal to the product of these, ab.

we set the mean of the G-distribution equal to 1 by constraining b = 1/a. By varying this single parameter, a, the distribution can take on a variety of different shapes.

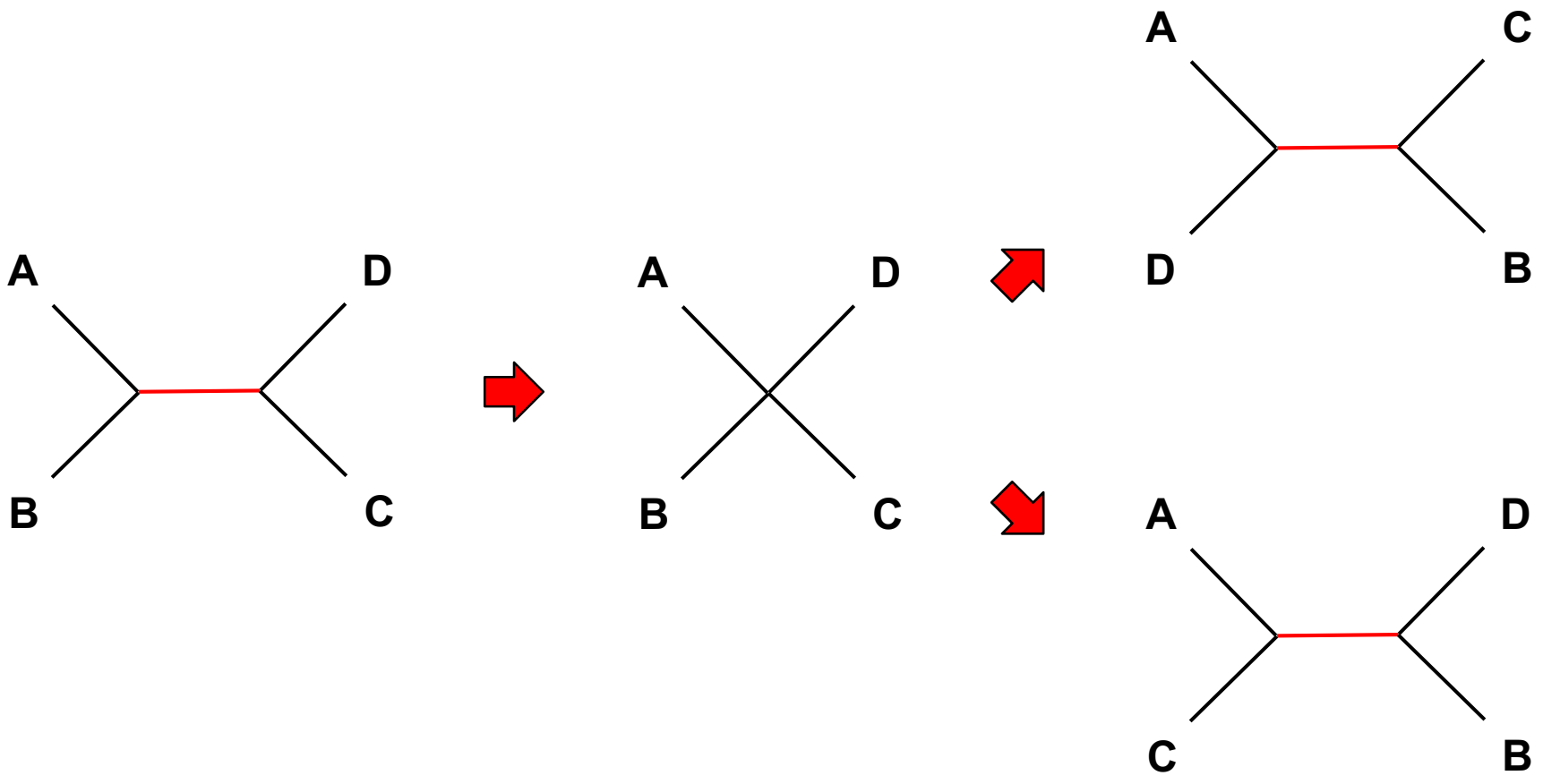
Fast phylogenetic approaches

- Too many trees to look at
  - Heuristic search of the tree space
- Too many calculations to do
  - Approximate likelihood calculation
- Other techniques for fast phylogenetics

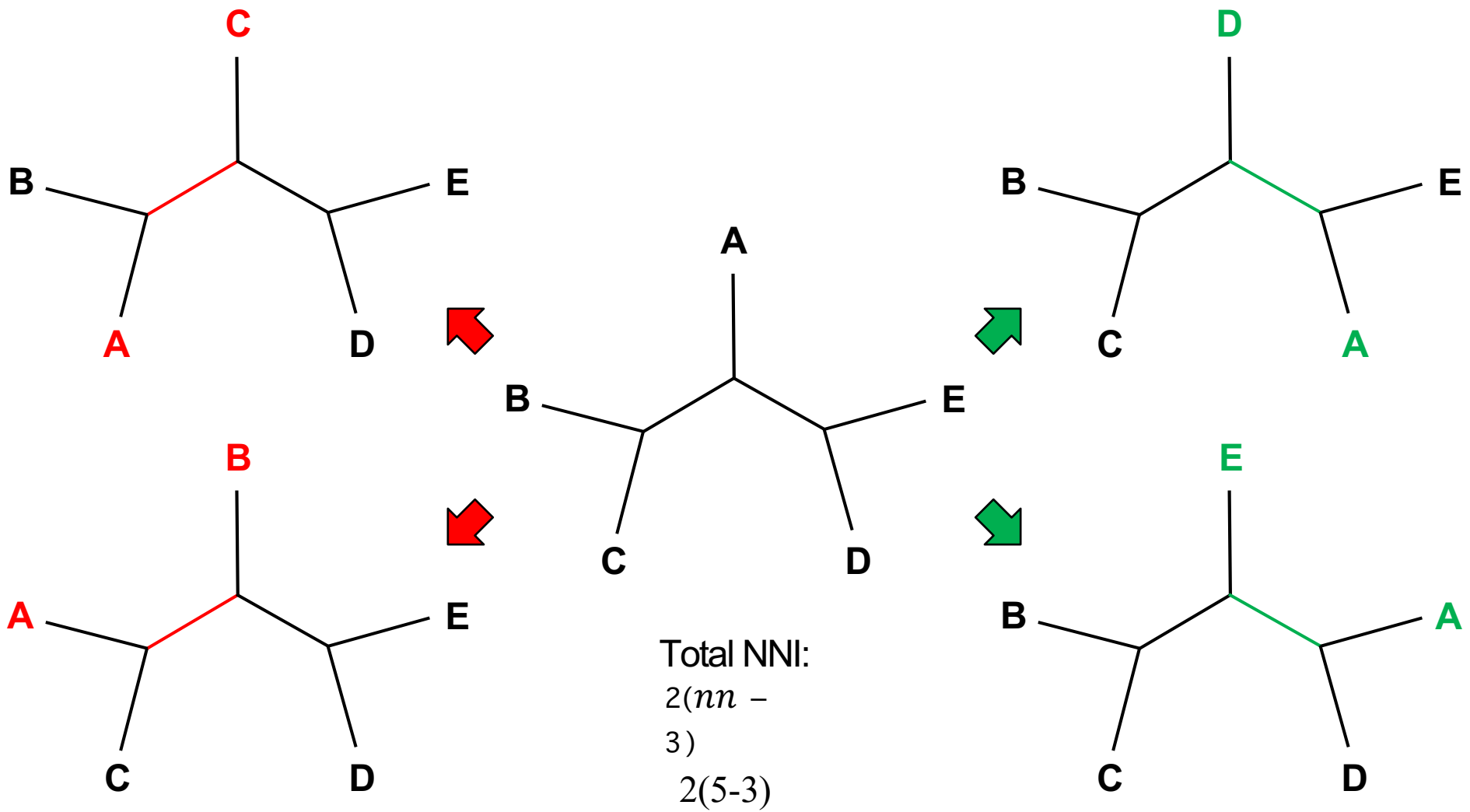
Heuristic tree search



Nearest Neighbor Interchange (NNI)



Nearest Neighbor Interchange (NNI)



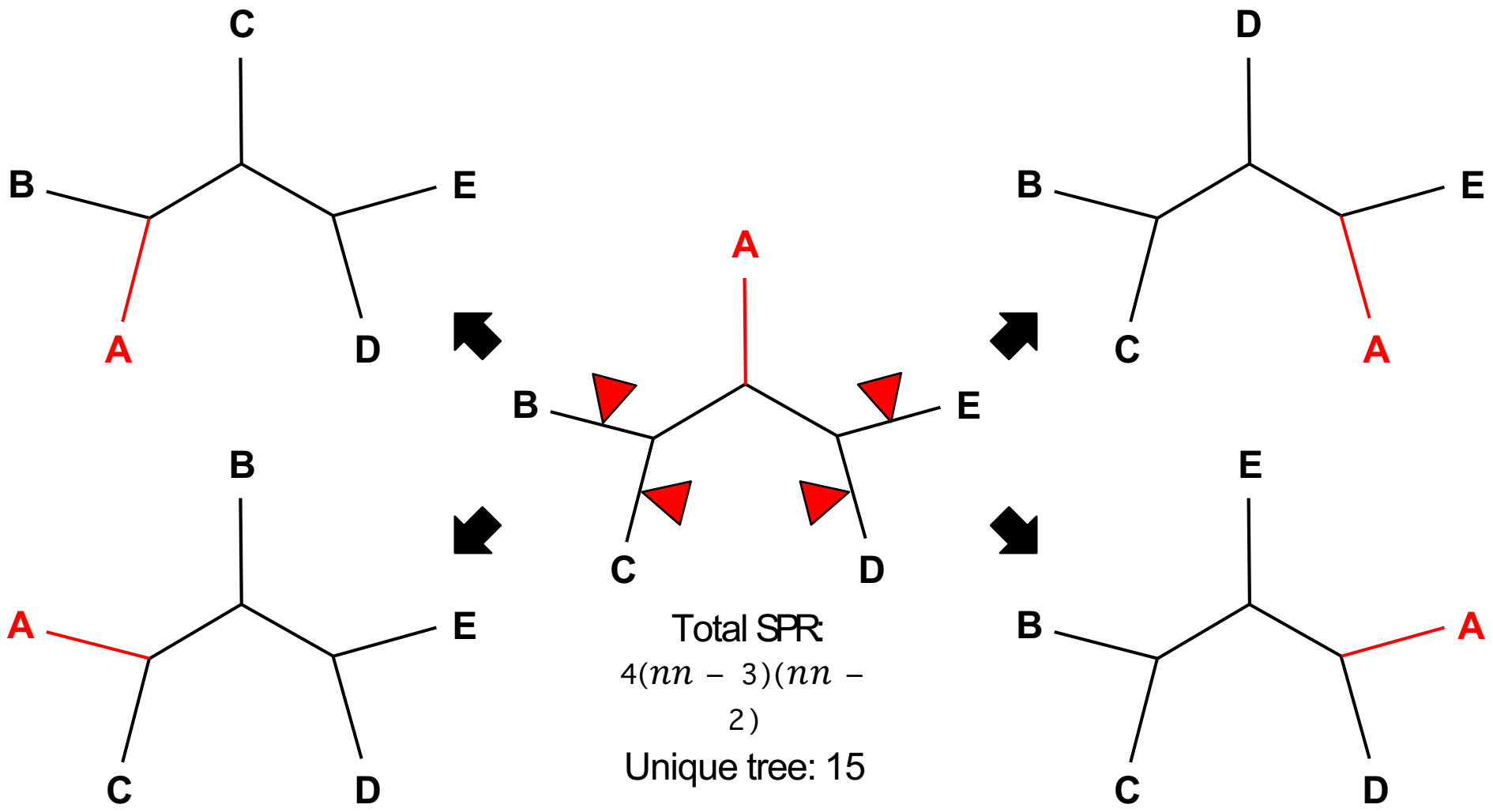
Total NNI:

$2(nn - 3)$

$2(5\text{-}3)$

4

Subtree Pruning and Re-grafting (SPR)



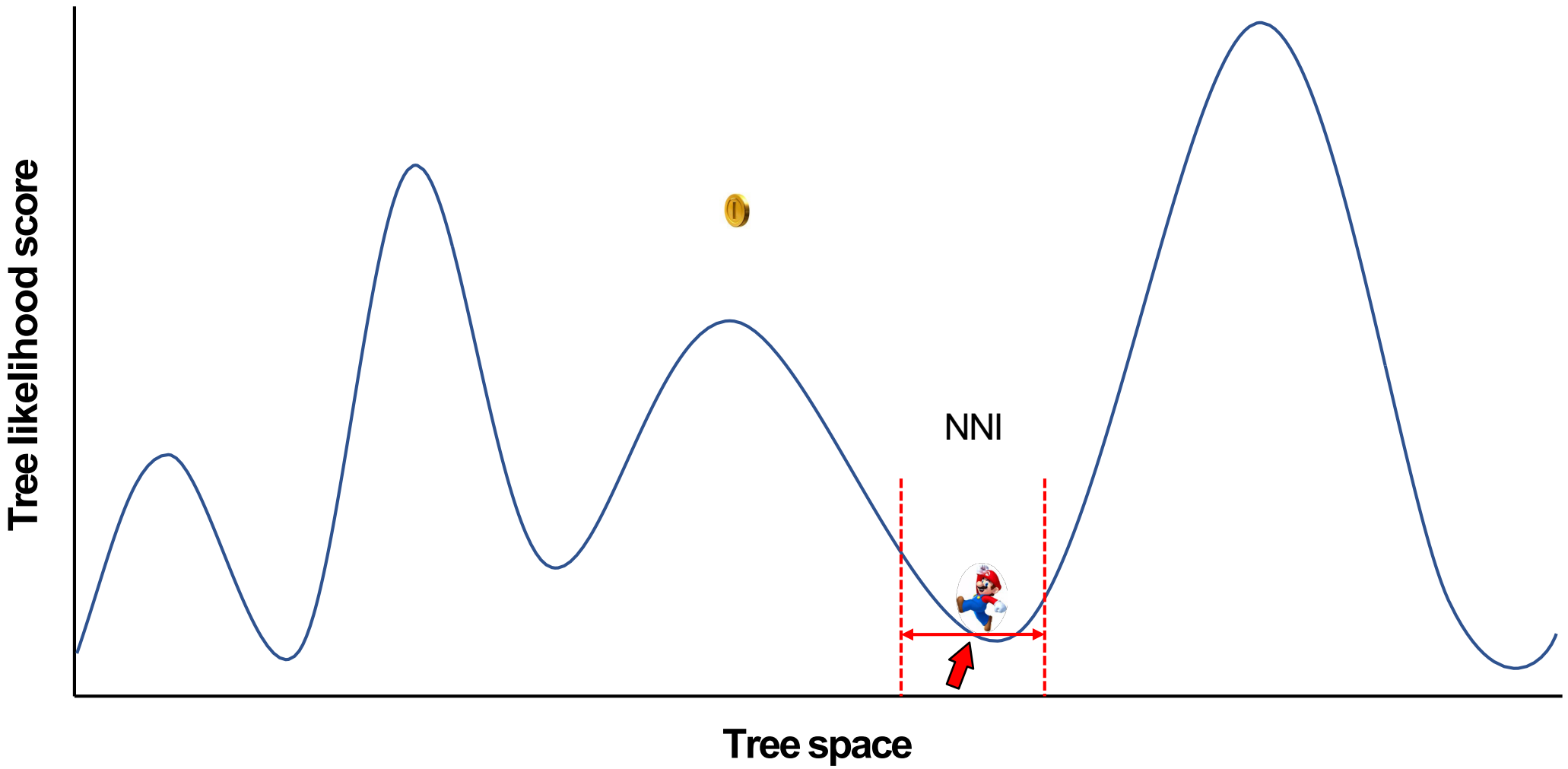
Total SPR:

$4(nn - 3)(nn - 2)$

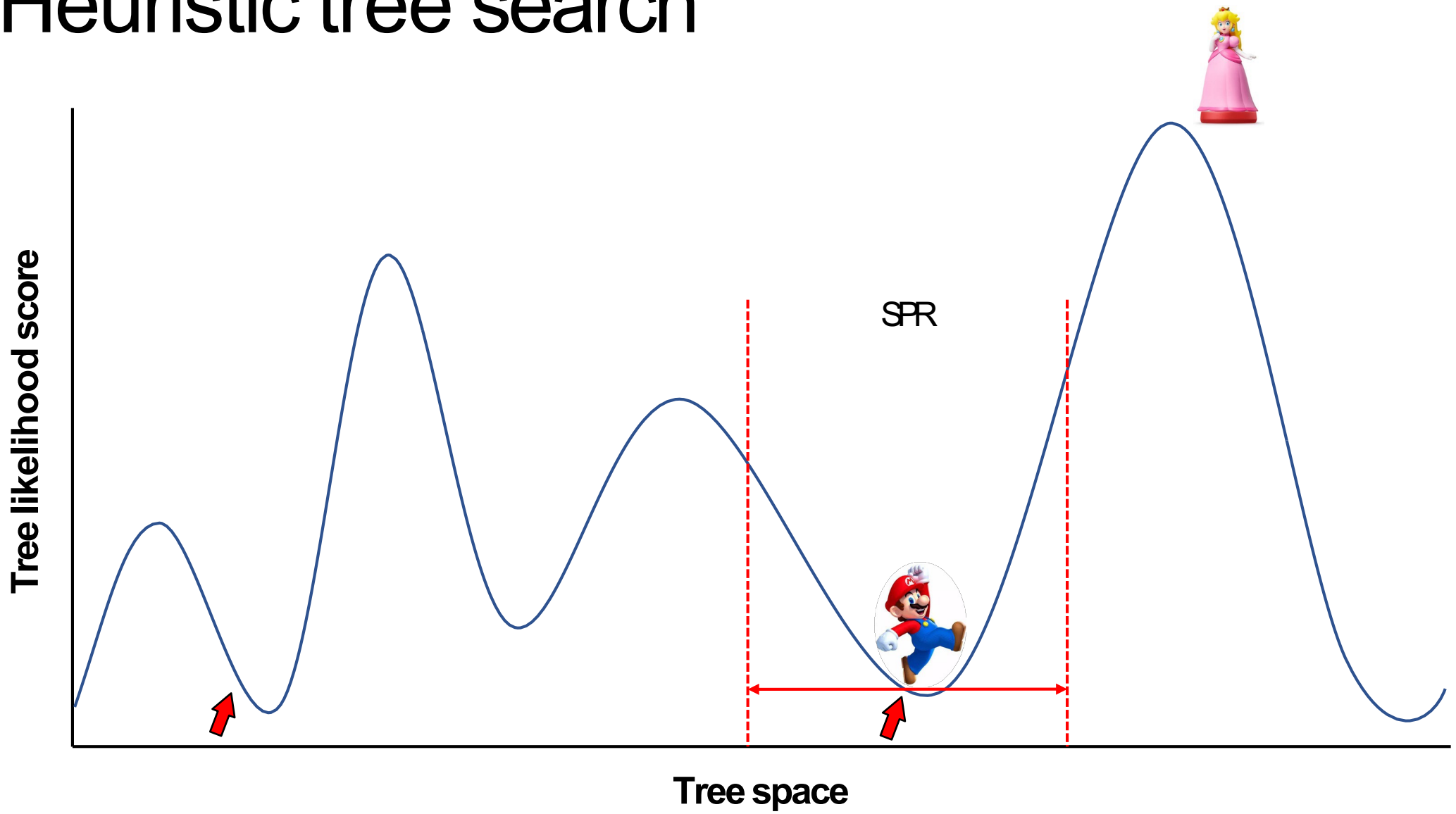
Unique tree: 15

Heuristic search of tree space

Heuristic tree search



Heuristic tree search



Approximate likelihood calculation

- Global optimization vs. local optimization

- Exhaustive optimization vs. approximate optimization
  - Diminished return from extra efforts
  - Subsequent topological changes can invalidate extra efforts

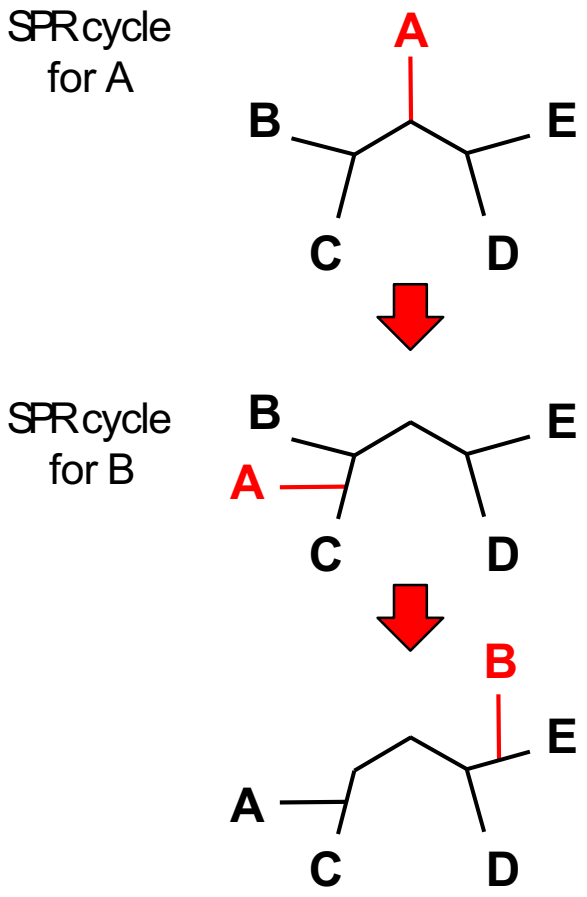
Other techniques for fast phylogenetics

- GAMMA vs CAT
- Fast approaches for node support
- Parallelization
- …

RAxML - Randomized Axelerated Maximum Likelihood

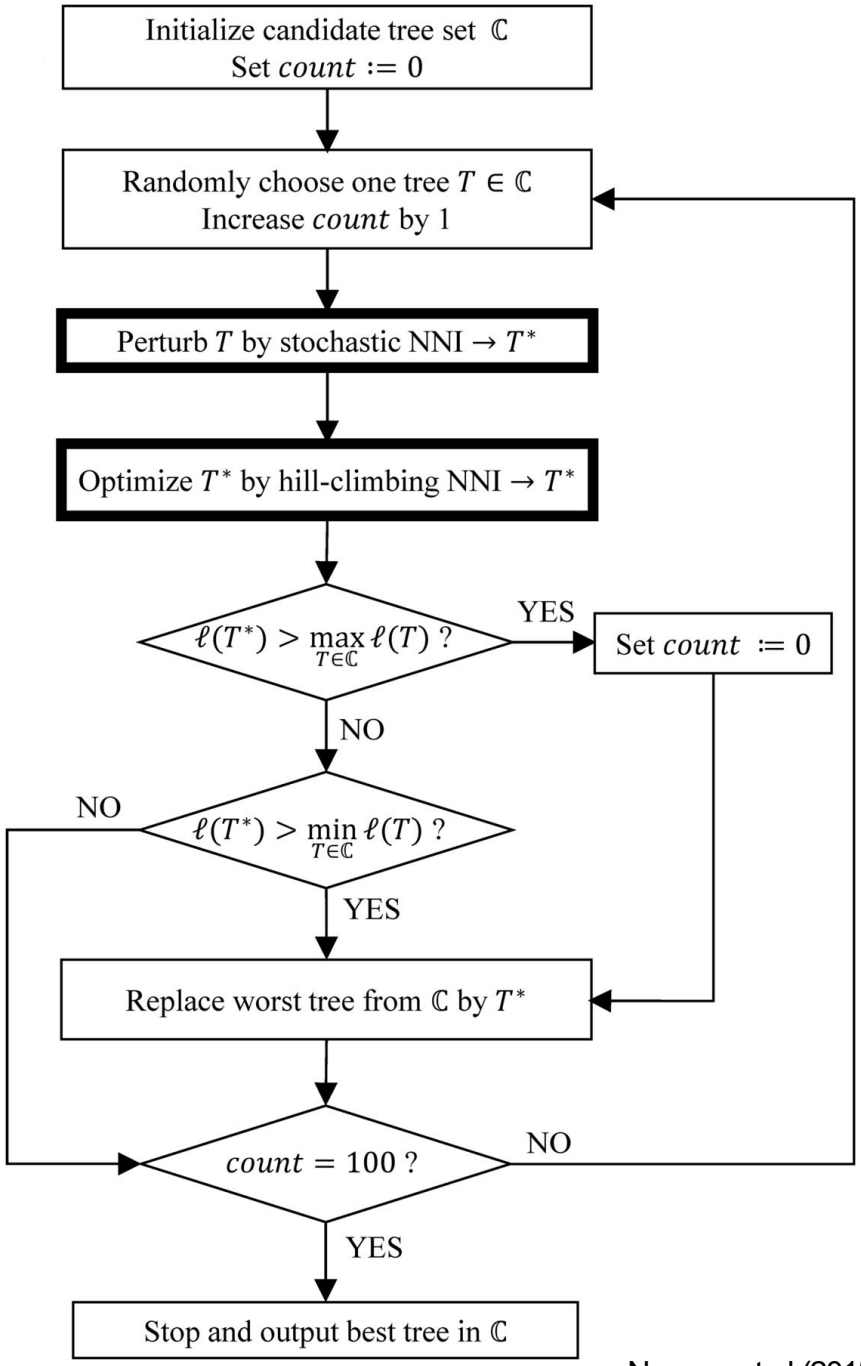
- Parsimony starting tree:
  - Parsimony is connected with ML
  - Speed and randomization!

- Lazy SPR
  - Only pre-scoring during one SPR iteration
  - SPRs leading to better scores are immediately implied
  - Dynamic adjustment of Lazy SPR radius



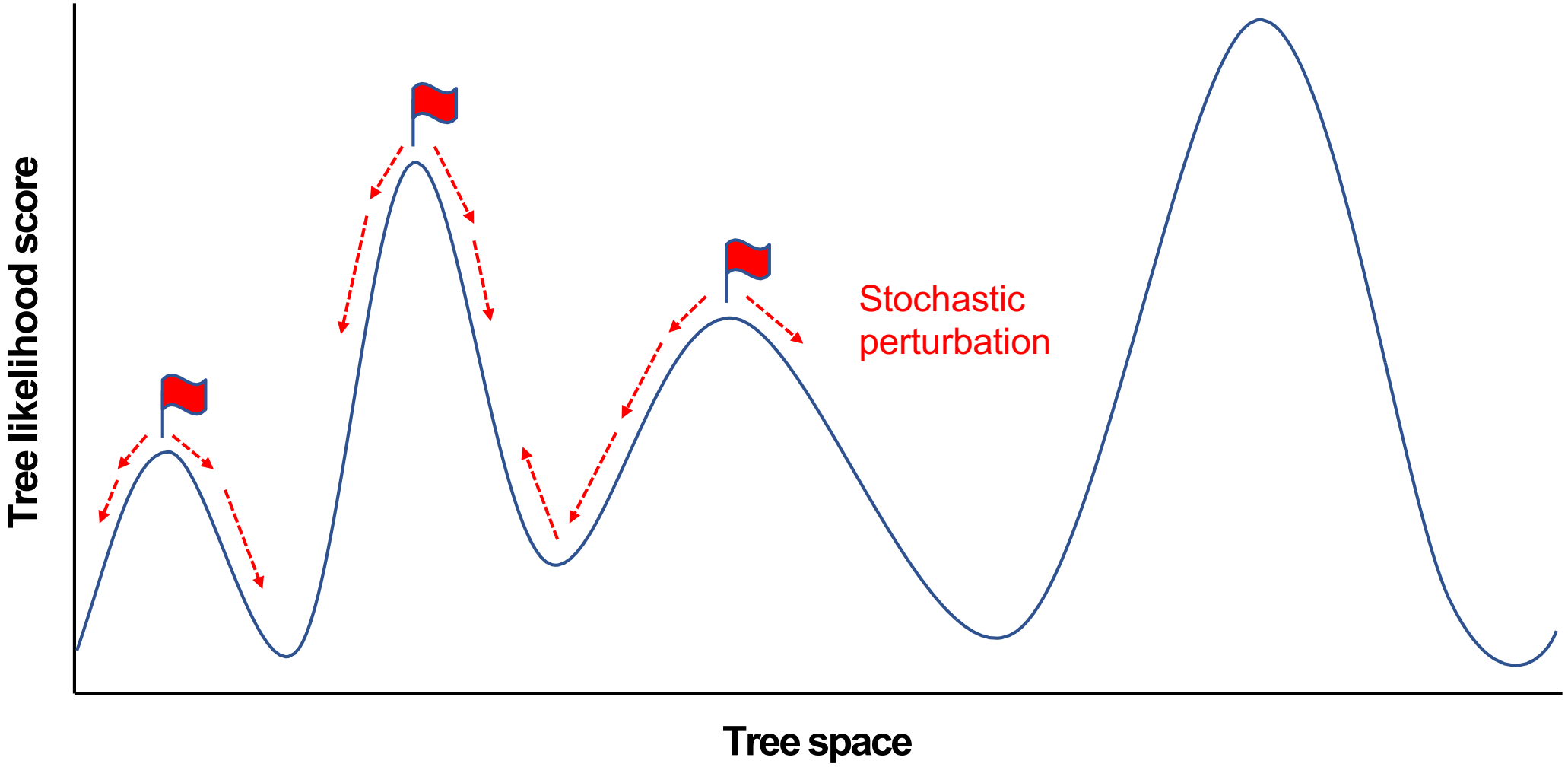
IQ-TREE

- A pool of starting trees
- A pool of candidate trees
- NNI- instead of IQP-based perturbation
- Simultaneous NNI modifications
  - Reduced NNI neighborhood



Nguyen et al (2015) Mol. Biol. Evol.

Escape from local optima: IQ-TREE

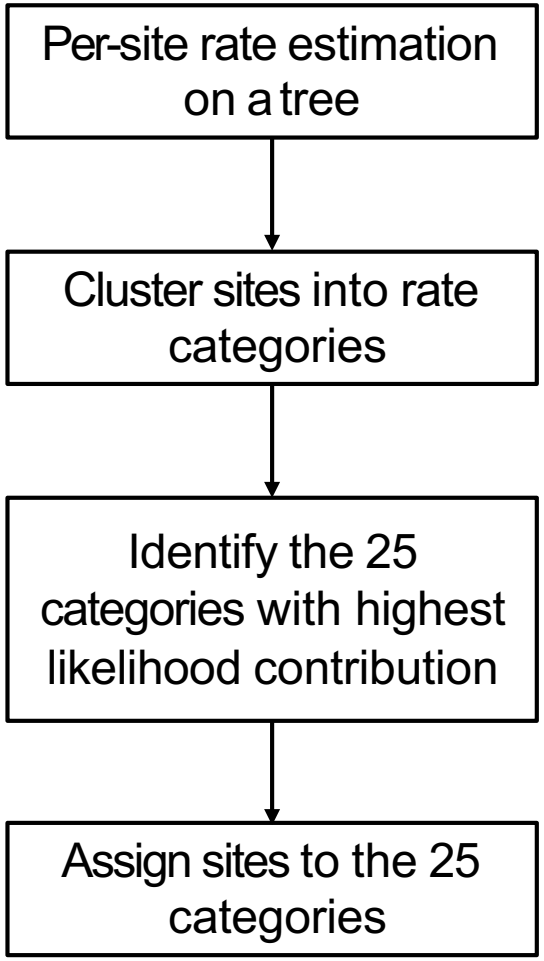


Other techniques for fast phylogenetics

- GAMMA vs CAT

GAMMA vs CAT

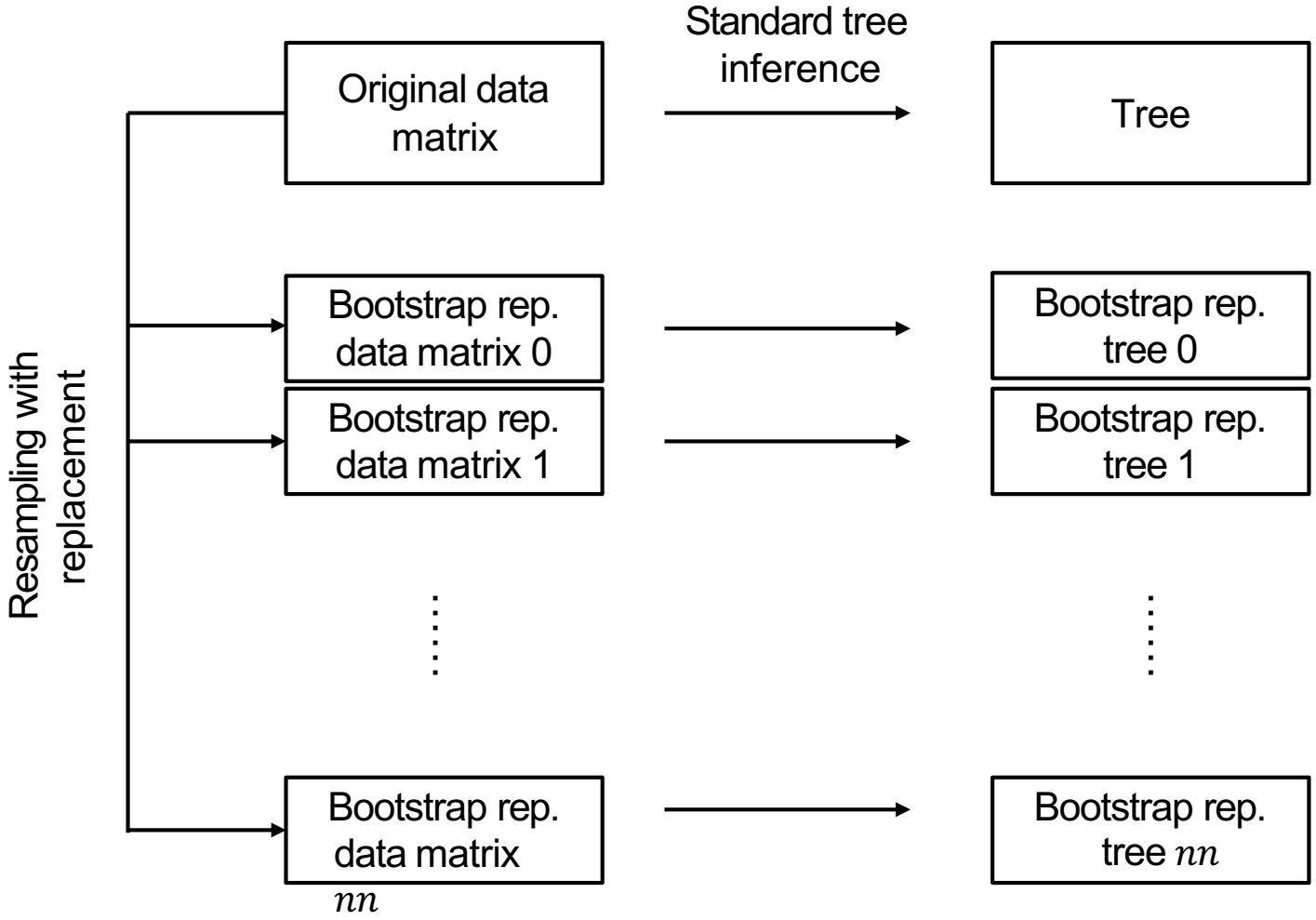
- GAMMA
  - model rate heterogeneity among sites using the gamma distribution
  - each site has certain probability belonging to each rate category
- CAT
  - assign sites into fixed number of rate categories
  - each site belongs to a specific rate category



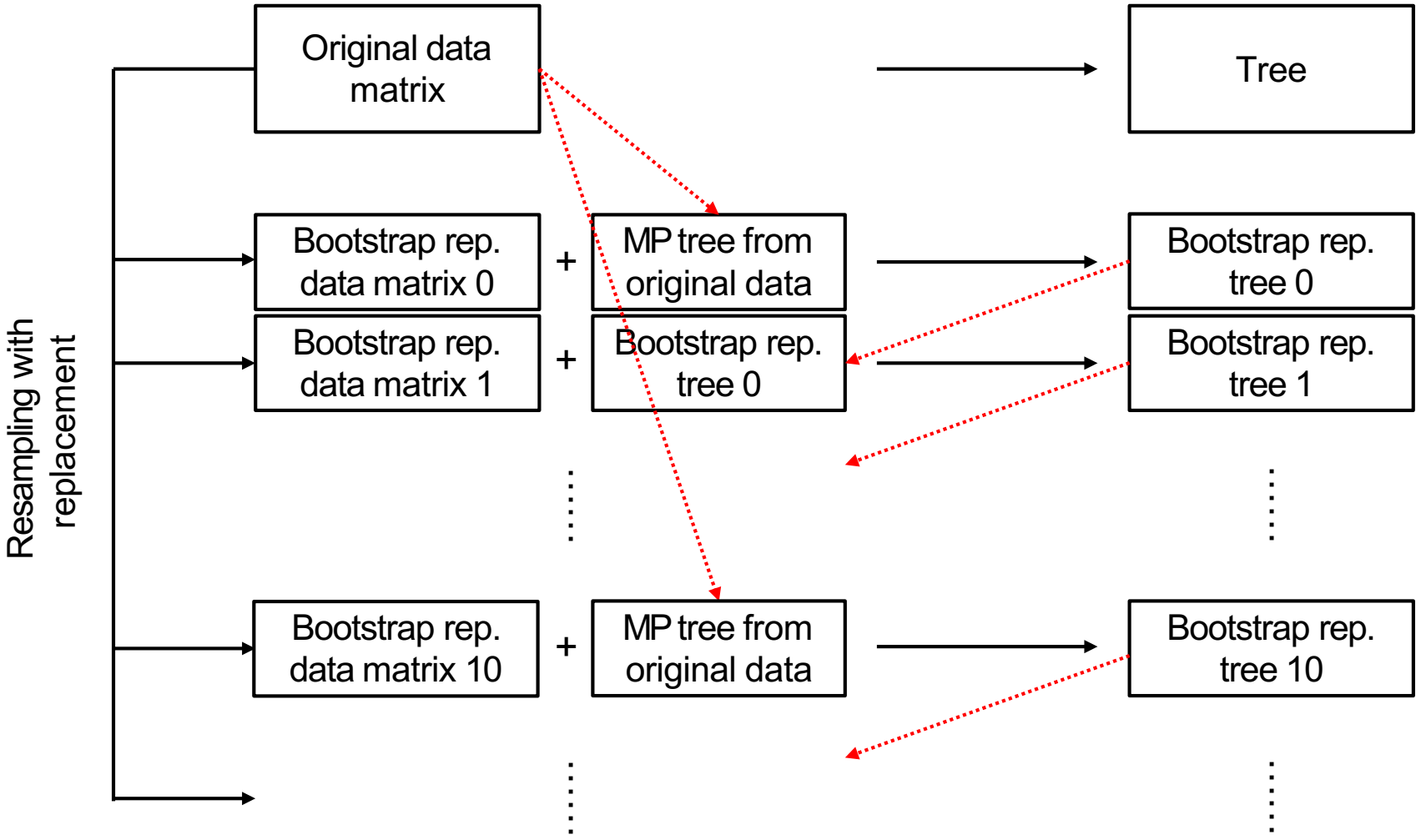
Other techniques for fast phylogenetics

- GAMMA vs CAT
- Fast approaches for node support

Standard Bootstrap



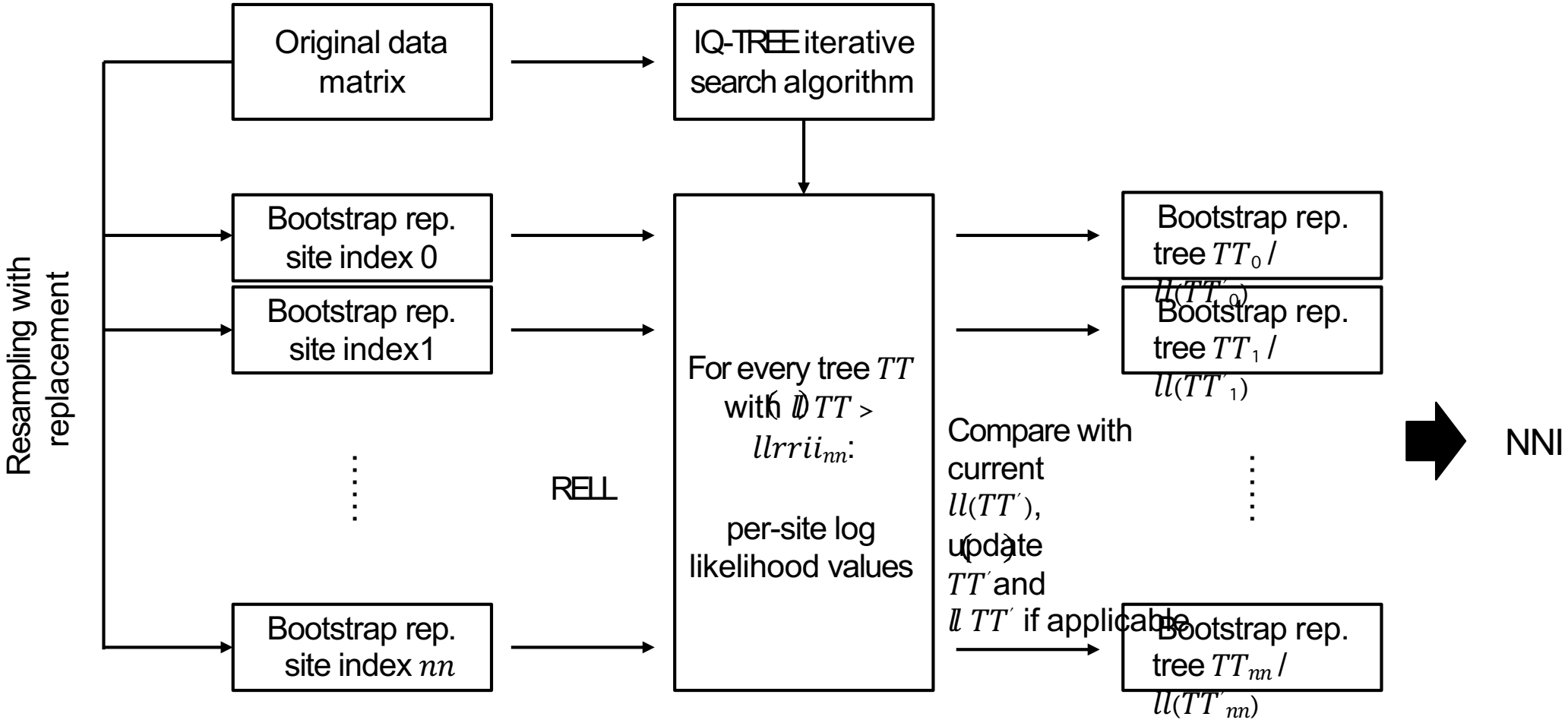
Rapid bootstrap (RAxML)



Additional shortcuts:

- LSR radius randomly chosen between 5 and 15;
- 2 iterations of LSR;
- More aggressive subtree skipping;
- Thorough optimization for best 5 instead of 20;

Ultra-fast bootstrap (IQ-TREE)



Other techniques for fast phylogenetics

- GAMMA vs CAT
- Fast approaches for node support
- Parallelization

Parallelization

- Multi-threading and MPI
- Parallel tree searches:
  - MPI: RAxML, PhyML, IQ-TREE
- Likelihood calculation
  - RAxML/IQ-TREE

Important remarks

- Different software can give different scores on the same tree
- Trees with the same topology can still get different scores

Recommendations:

- Multiple searches using distinct starting trees

- More than one phylogenetic software
  - SPR-based software for trees with many taxa