Module 3: Phylogenomics

- Phylogenetic inference using maximum-likelihood and Bayesian analyses
- Phylogenomics and challenges of whole genome inferences

Patrícia H. Brito

UCIBIO
Applied Molecular Biosciences Unit

NOVA SCHOOL OF SCIENCE & TECHNOLOGY
DEPARTMENT OF LIFE SCIENCES

Who am I?

- Microbial genome evolution
- Evolutionary genomics
- Phylogenomics
- Population genomics
- Molecular Evolution
- Evolutionary Medicine

Patrícia H. Brito, Ph.D.

Computational Biologist

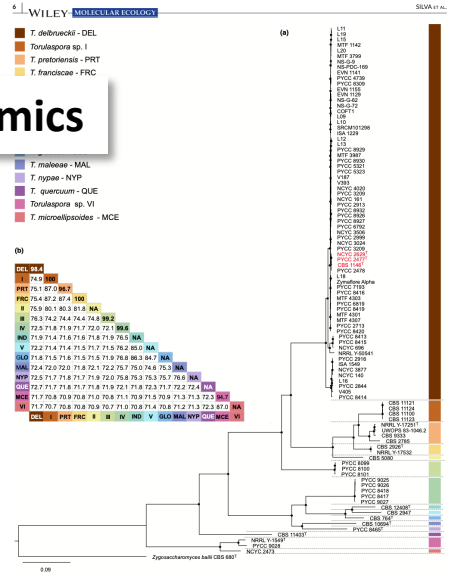
Yeast Genomics LAB

UCIBIO, NOVA Science and Technology, UNL

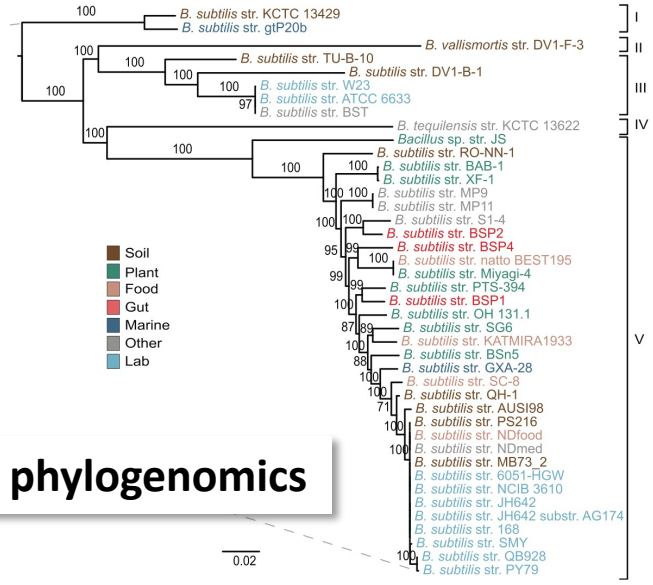
https://yeastgenomicslab.wordpress.com

phbrito@fct.unl.pt

Yeast phylogenomics

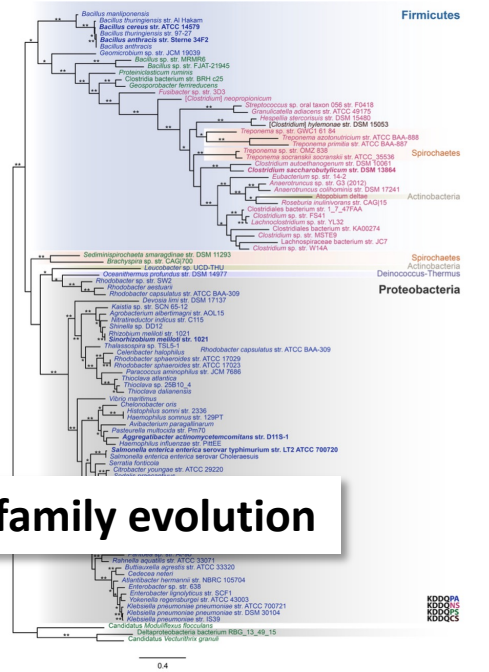


Silva et al 2022 Molecular Ecology



Bacterial phylogenomics

Brito *et al* 2018 GBE



Protein family evolution

Torcato et al 2019 JBC

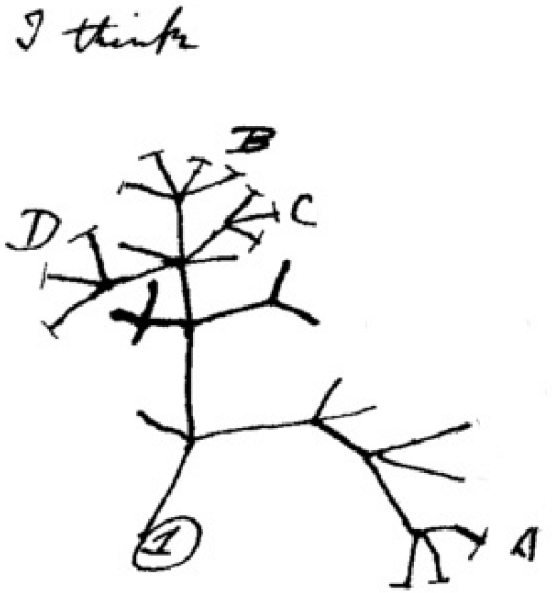
UCIBIO Applied Molecular Biosciences Unit

NOVA SCHOOL OF SCIENCE & TECHNOLOGY
DEPARTMENT OF LIFE SCIENCES

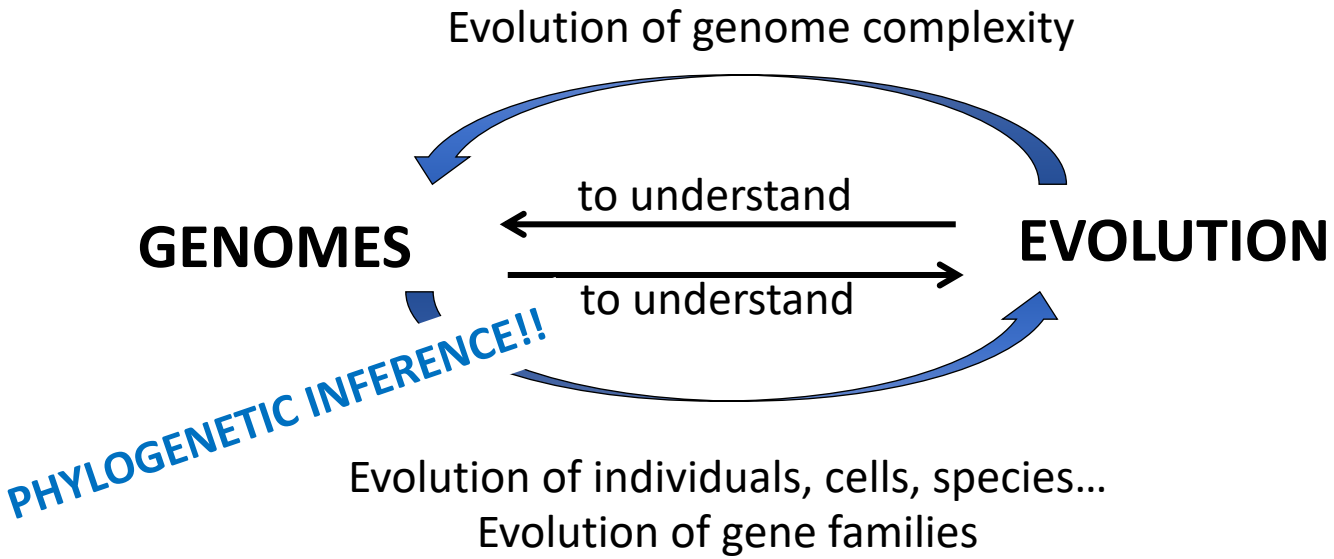
Phylogenetic inference – what it is and why do we do it?

"The affinities of all the beings of the same class have sometimes been represented by a great tree. I believe this simile largely speaks the truth."

Charles Darwin, O*rigin of Species, 1859*

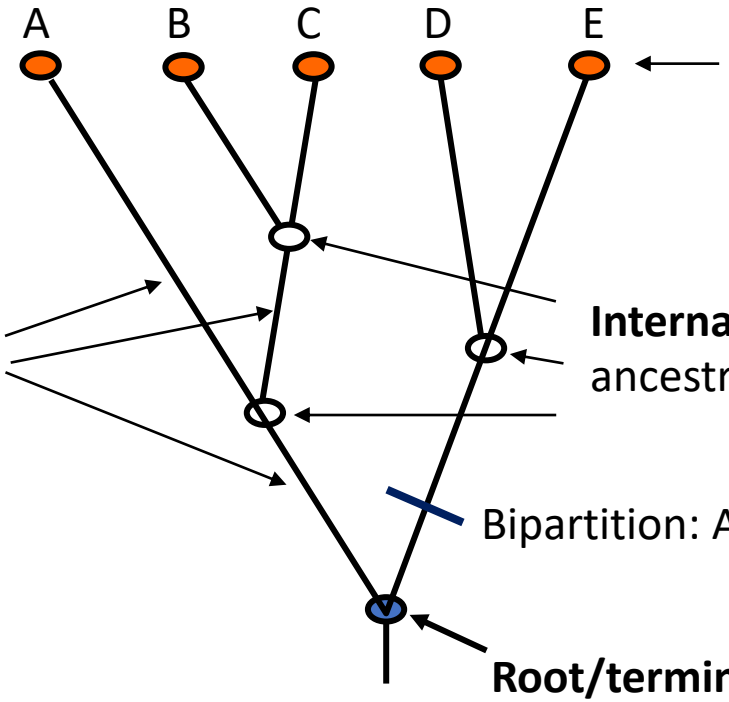


Evolution is descent with modification from a common ancestor



Phylogenetic inference – terminology

Topology



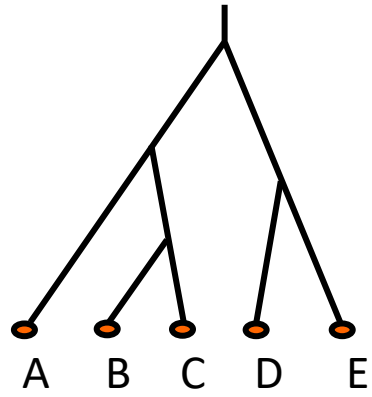
Terminal leaf/node/species/OTU/taxon
(extant organisms)

Branches

Internal nodes (these are hypothetical ancestrals of terminal nodes)

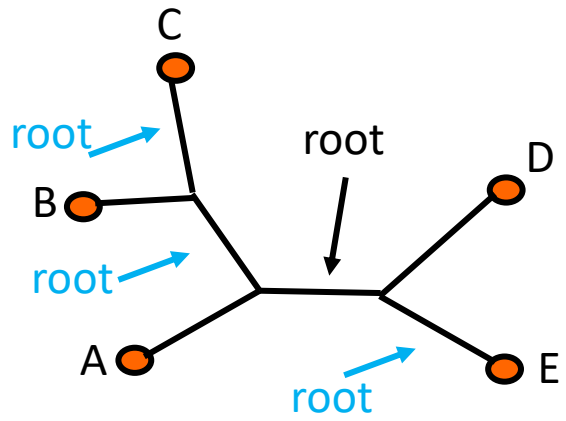
Bipartition: ABC|DE

Root/terminal node/ancestral node
MRCA



((A,(B,C)),(D,E))

Newick format



Phylogenetic inference – typical pipeline

Question:
we want to know how a particular group of species/organisms/cells relate to each other

Typical pipeline in Molecular Phylogenetics

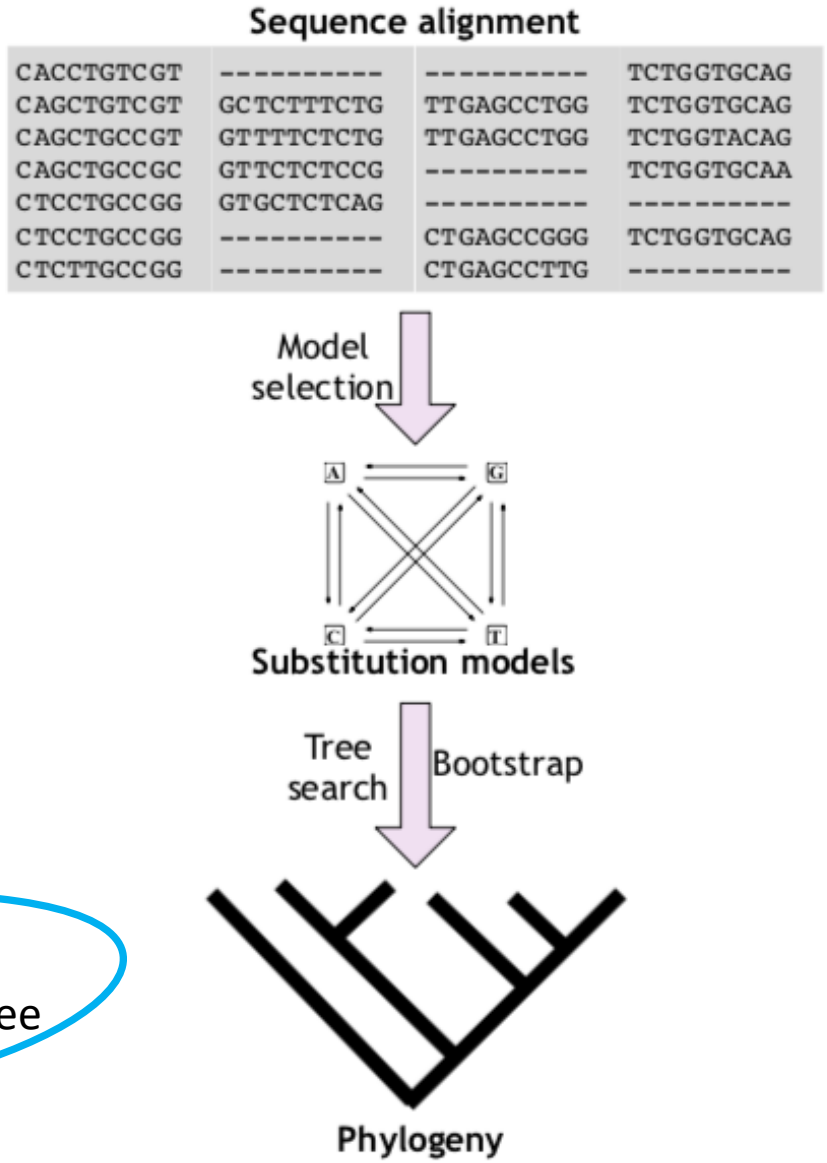
1. Choose the molecular marker (genomic region or specific data type)
2. Get the sequences of that molecular marker for all terminals in the tree
3. Choose an **optimality criterion** and an **algorithm** to estimate the gene tree

1 gene → some genes → many genes → genomes

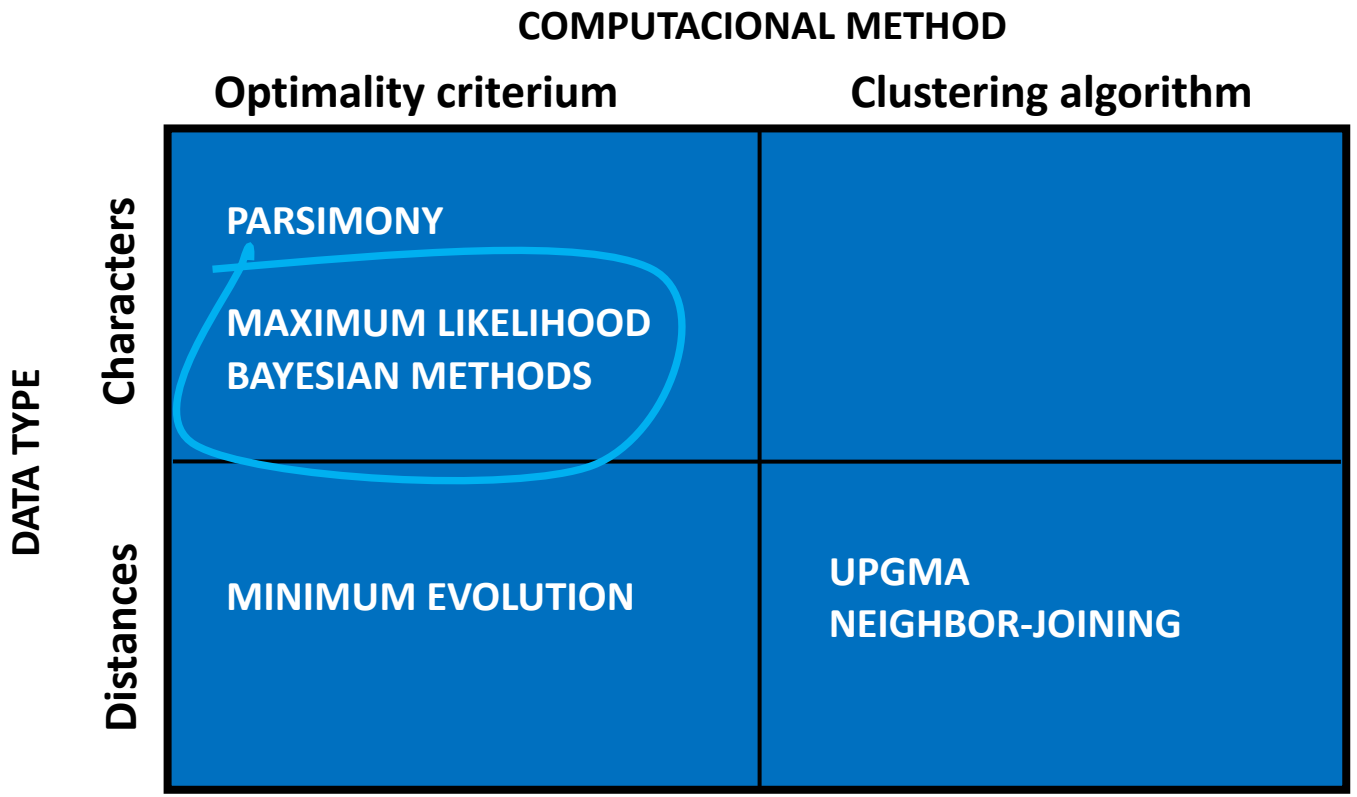
More genes → more resolution at different levels of the tree → higher support

In most analyses the implicit assumption is that all genes do in fact have the same gene tree, that these gene trees are congruent with and converge on the species tree

not always true!



Phylogenetic inference – methods overview



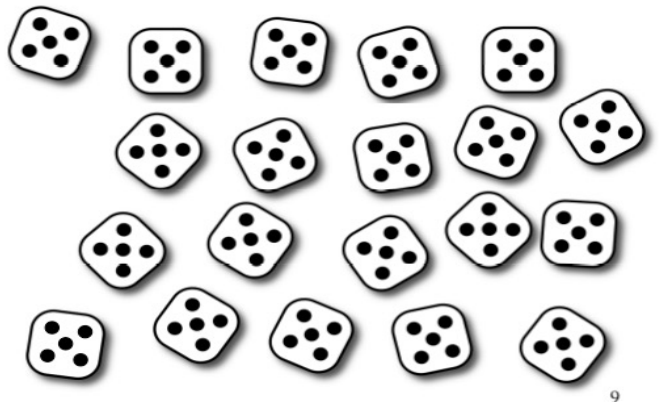
Taxa Characters

```
Species A ATGGCTATTCTTATAGTACG
Species B ATCGCTAGTCTTATATTACA
Species C TTCACTAGACCTGTGGTCCA
Species D TTGACCAGACCTGTGGTCCG
Species E TTGACCAGTTCTCTAGTTCG
```

Taxa Distances

```
           A      B      C      D      E
Species A ----
Species B 0.23   ----
Species C 0.87   0.59   ----
Species D 0.73   1.12   0.17   ----
Species E 0.59   0.89   0.61   0.31   ----
```

Maximum Likelihood



$$\text{Prob(obs}|\text{fair dice)} = \left(\frac{1}{6}\right)^{20} = \frac{1}{3,656,158,440,062,976}$$

Prob (obs|non-fair dice) = $1^{20} = 1$

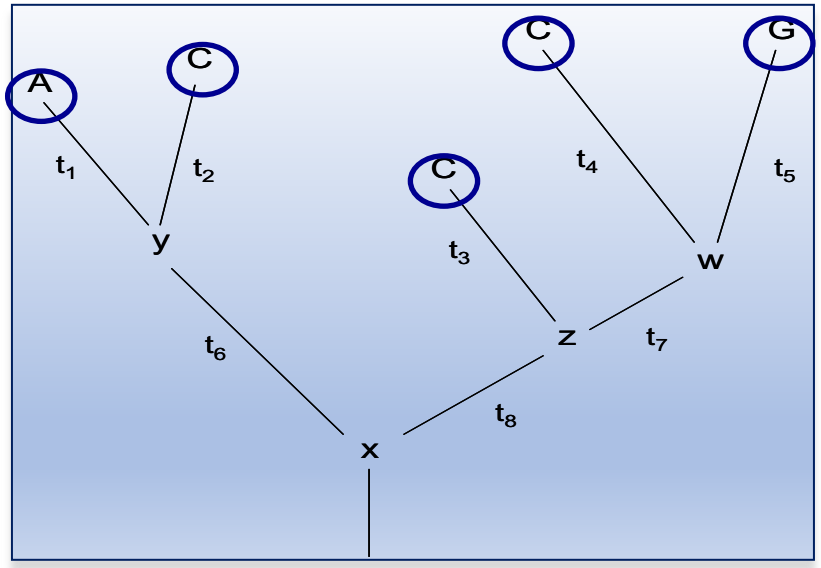
observation

hypothesis

→ The same observed result leads to very different likelihoods depending on the assumed model (hypothesis)

→ The model that assumes "non-fair dice" is the one that gives higher probability of observing that result, and as such it is the most likely model.

Maximum Likelihood of a Tree (topology & branch lengths)



$$L = \Pr ob(D \mid T) = \prod_{i+1}^{m} \Pr ob(D^{(i)} \mid T)$$

data

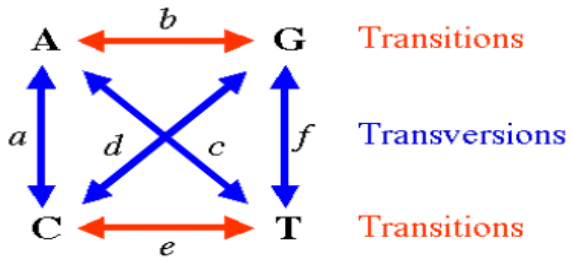
tree

data at the i^{th} site

Two assumptions:

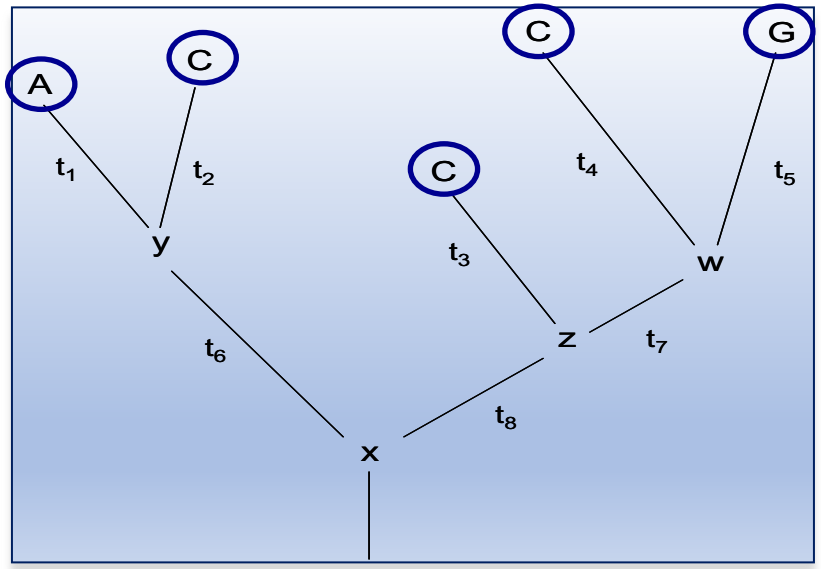
1. Evolution in different sites is independent
2. Evolution in different lineages is independent

$$\Pr ob(x) \quad \Pr ob(y \mid x, t_6) \quad \Pr ob(A \mid y, t_1) \quad \Pr ob(C \mid y, t_2)$$
$$\Pr ob(z \mid x, t_8) \quad \Pr ob(C \mid z, t_3)$$
$$\Pr ob(w \mid z, t_7) \quad \Pr ob(C \mid w, t_4) \quad \Pr ob(G \mid w, t_5)$$



Designation	Rate params	Base frequencies	Number of free params
JC	*a=b=c=d=e=f*	$\pi_A = \pi_C = \pi_G = \pi_T$	1
K80, K2P	*a=c=d=f, b=e*	$\pi_A = \pi_C = \pi_G = \pi_T$	2
TrNef	*a=c=d=f, b, e*	$\pi_A = \pi_C = \pi_G = \pi_T$	3
K81, K3ST	*a=f, b=e, c=d*	$\pi_A = \pi_C = \pi_G = \pi_T$	3
TVMef	*a, c, d, f, b=e*	$\pi_A = \pi_C = \pi_G = \pi_T$	5
TIMef	*a=f, c=d, b, e*	$\pi_A = \pi_C = \pi_G = \pi_T$	4
SYM	*a, b, c, d, e, f*	$\pi_A = \pi_C = \pi_G = \pi_T$	6
F81	*a=b=c=d=e*	$\pi_A, \pi_C, \pi_G, \pi_T$	4
HKY	*a=c=d=f, b=e*	$\pi_A, \pi_C, \pi_G, \pi_T$	5
TrN	*a=c=d=f, b, e*	$\pi_A, \pi_C, \pi_G, \pi_T$	6
K81uf	*a=f, b=e, c=d*	$\pi_A, \pi_C, \pi_G, \pi_T$	6
TVM	*a, c, d, f, b=e*	$\pi_A, \pi_C, \pi_G, \pi_T$	8
TIM	*a=f, c=d, b, e*	$\pi_A, \pi_C, \pi_G, \pi_T$	7
GTR, REV	*a, b, c, d, e, f*	$\pi_A, \pi_C, \pi_G, \pi_T$	9

Maximum Likelihood of a Tree (topology & branch lengths)



$$L = \Pr ob(D \mid T) = \prod_{i+1}^{m} \Pr ob(D^{(i)} \mid T)$$

data

tree

data at the ith site

$$\Pr ob(x) \quad \Pr ob(y \mid x, t_6) \quad \Pr ob(A \mid y, t_1) \quad \Pr ob(C \mid y, t_2)$$
$$\Pr ob(z \mid x, t_8) \quad \Pr ob(C \mid z, t_3)$$
$$\Pr ob(w \mid z, t_7) \quad \Pr ob(C \mid w, t_4) \quad \Pr ob(G \mid w, t_5)$$

For 1 nucleotide at ancestral position x, and for 1 tree

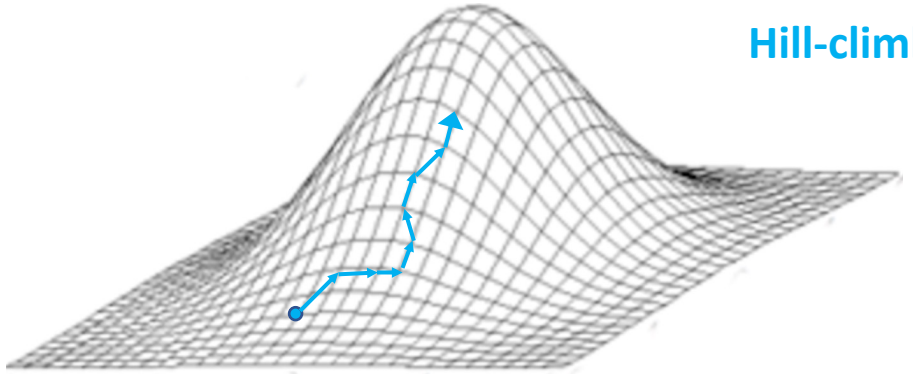
$$\Pr ob\left(D^{(i)} \middle| T\right) = \sum_x \sum_y \sum_z \sum_w \Pr ob\left(A, C, C, C, G, x, y, z, w \middle| T\right)$$

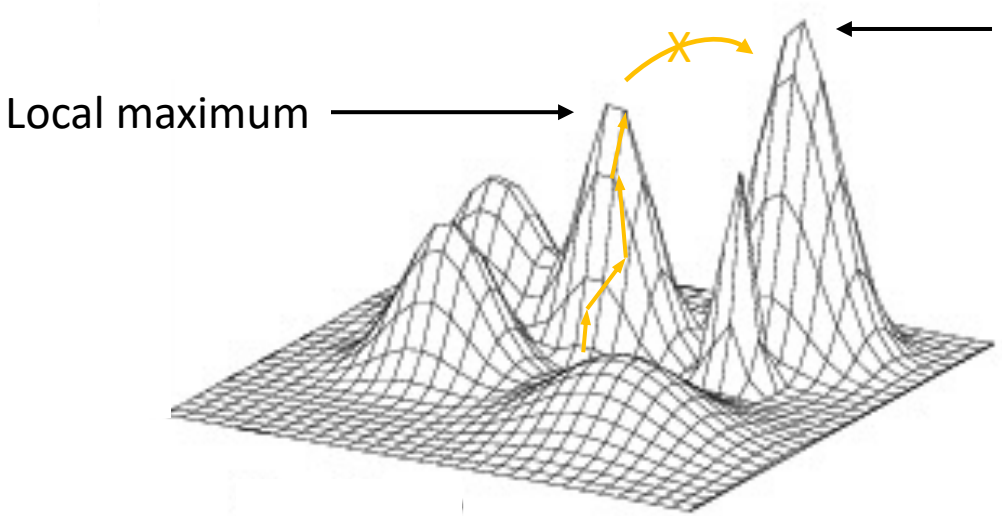
→ Probability of data at the ith site given a certain tree

Phylogenetic inference is a NP-complete problem where exhaustive searches for datasets of 10+ terminals are practically impossible → **Heuristic methods**

Number of taxa *T*	Number of unrooted bifurcating trees *B(T)*
3	1
4	3
5	15
6	105
7	945
8	10,395
9	135,135
10	2,027,025
22	3×10^{23}
50	3×10^{74}

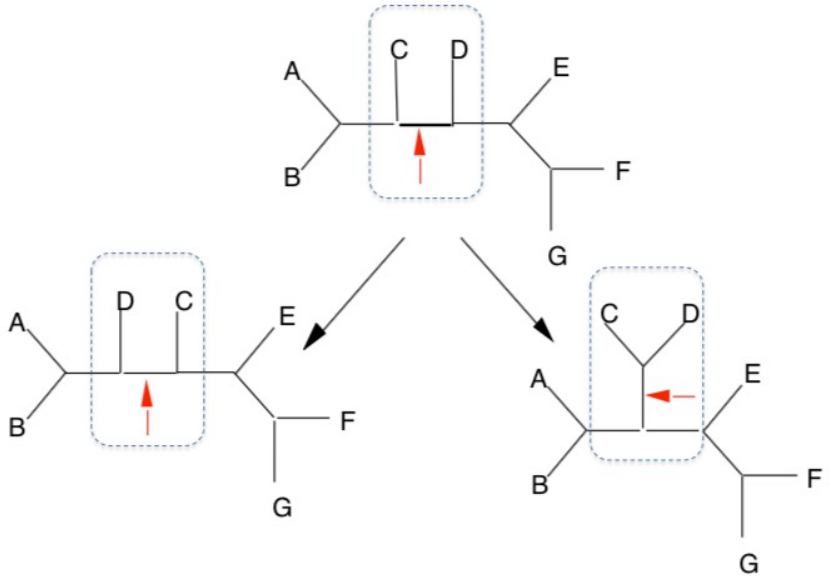
Hill-climbing algorithms



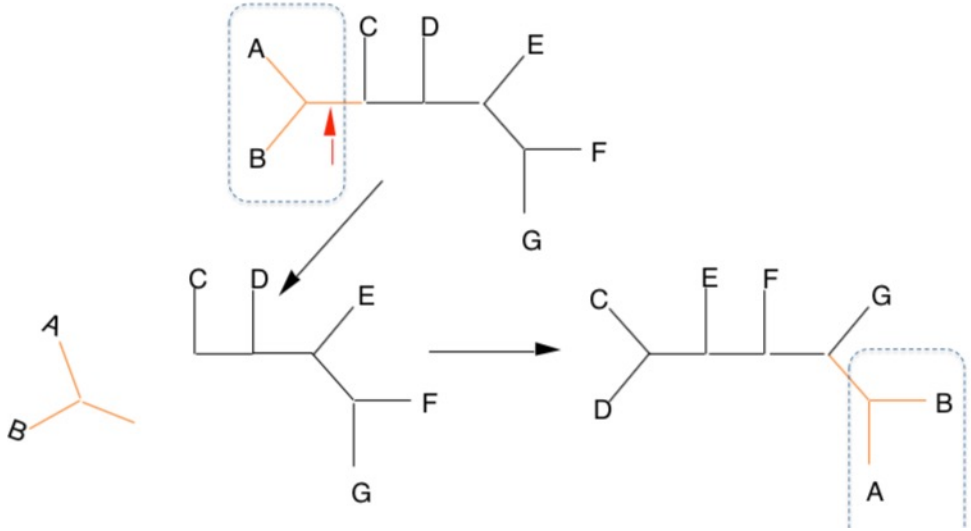


Phylogenetic tree space

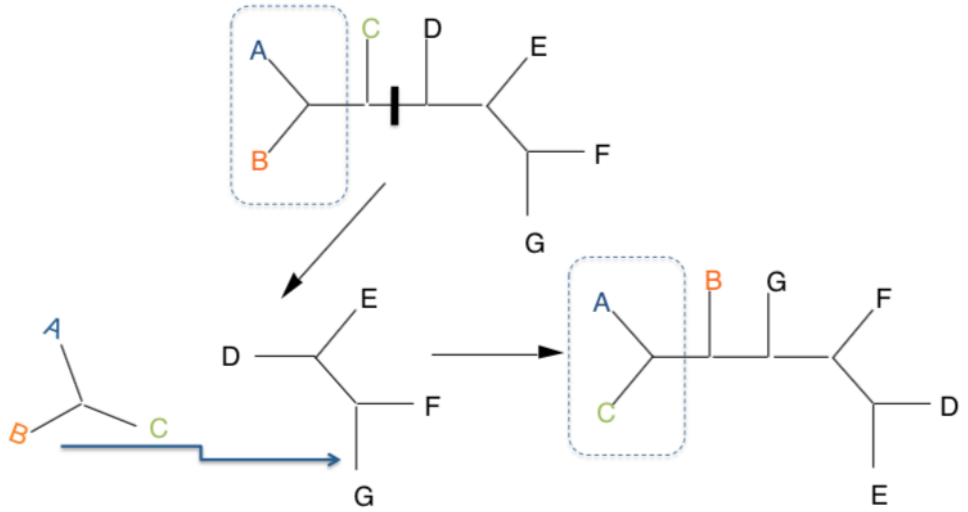
Nearest neighbor interchange (NNI)



Subtree pruning and regrafting (SPR)



Tree bisection and reconnection (TBR)

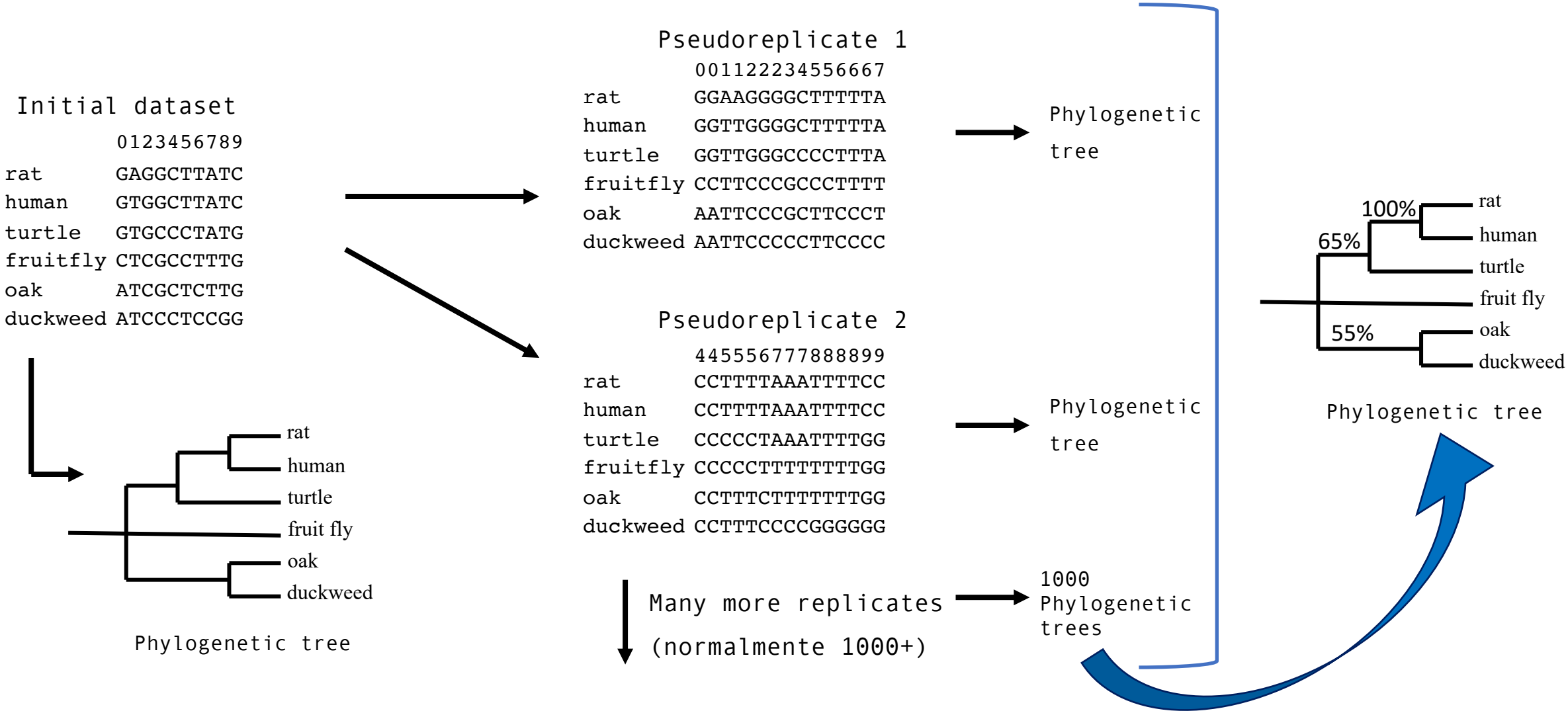


- **Branch swapping**

  NNI < SPR < TBR

- **Multiple replicates with random starting points**

Bootstrap resampling method used to estimate branch support on a phylogenetic tree.
Provides an indication of the robustness (confidence) in each bipartition.



Bayes' theorem

Rev. Thomas Bayes
(1701-1761)

Prior probability

Maximum likelihood

$$f(p \mid D) = \frac{f(p) f(D \mid p)}{f(D)}$$

Posterior probability

Data probability, sum of all (pi|D), normalizing constate that ensures that the posterior probability integrates to 1.

Bayesian phylogenetic inference

Rev. Thomas Bayes
(1701-1761)

Given:

τ = phylogenetic tree (topology + branch lengths)

X = data (aligned molecular data)

The posterior probability of tree is:

$$f(\tau_i \mid X) = \frac{f(X \mid \tau_i) f(\tau_i)}{\sum_{j=1}^{T} f(X \mid \tau_j) f(\tau_j)}$$

Max. likelihhod

Prior probability

- Uniform dist – topology
- Exponential dist – branch lengths
- Gamma dist – rate variation
- Dirichlet dist – allele frequency

1 tree

All possible trees

Typically, impossible to estimate. But by using **MCMC chains** to sample the posterior distribution we do not need to estimate this quantity

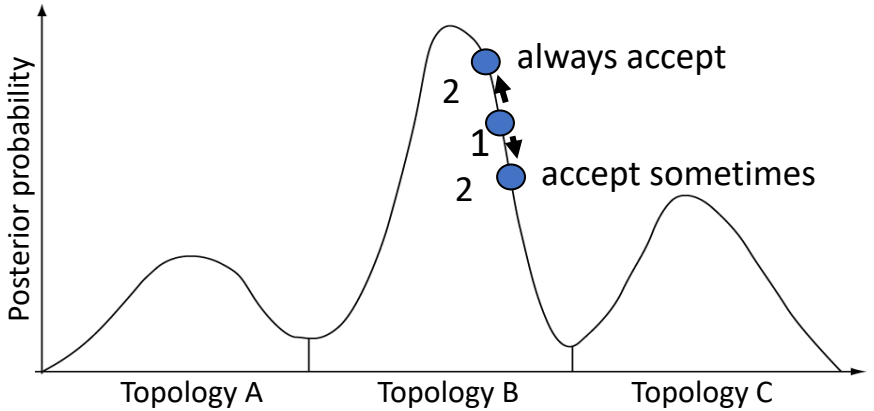
Markov Chain Monte Carlo sampling (MCMC)

Not a "hill-climbing" method!

MCMC approximates $f(\tau_i \mid X)$ by sampling a high number of trees τ_i from the posterior distribution. The trees with higher probabilities are the ones most likely to be sampled during the MCMC sampling process. Therefore, MCMC focuses most of the sampling effort on sampling the distribution of interest - **the proportion of time that MCMC method samples a give region of the parameter space is proportional to the posterior distribution of that region.**

Strategy for running an MCMC:

1. Start at a random point
2. Make a small-scale change
3. Estimate the ratio (r) of the probabilities of the new and the original state:
   If r > 1 -> accept change
   If r < 1 -> accept change with probability r
4. Back to step 2



Metropolis-Hasting decision criteria:

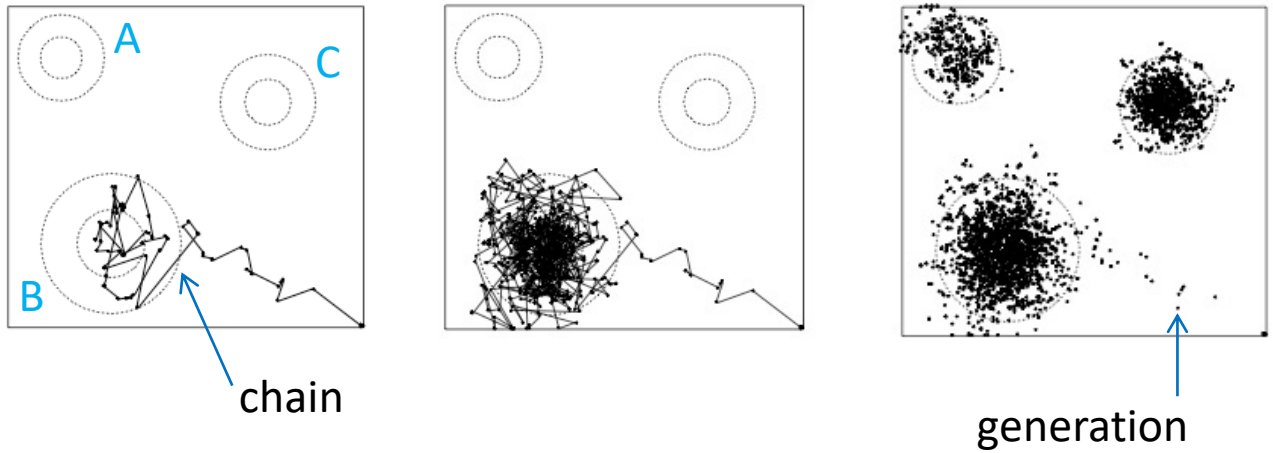
$$r = \min\left(1, \frac{f(\tau^*)}{f(\tau)} \frac{f(X \mid \tau^*)}{f(X \mid \tau)} \frac{f(\tau \mid \tau^*)}{f(\tau^* \mid \tau)}\right)$$

Hastings ratio

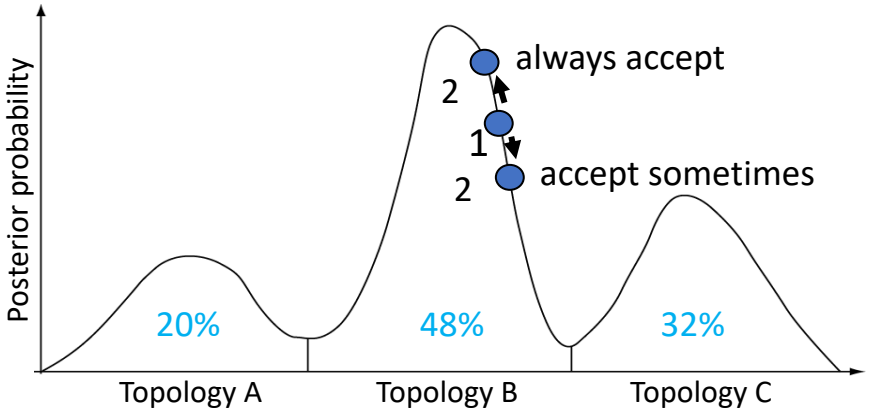
Max. likelihood ratio

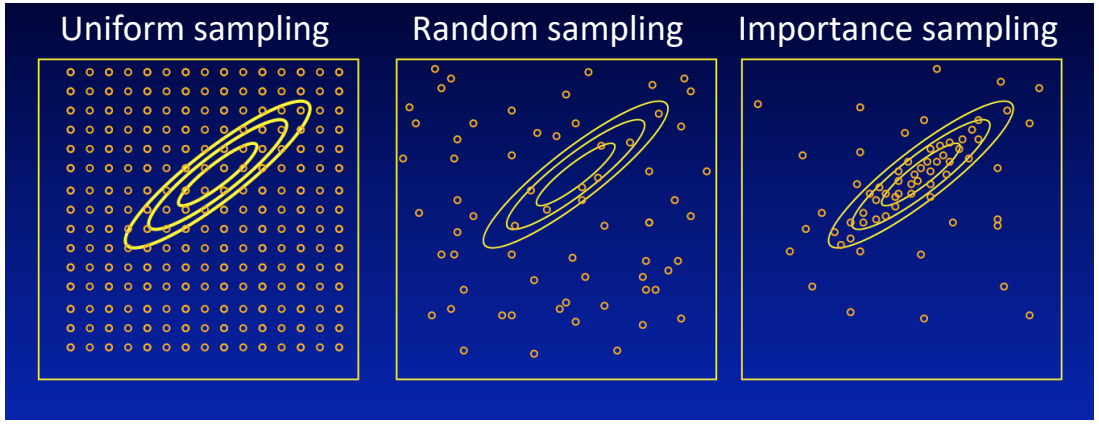
Ratio of prior probabilities

MCMC …



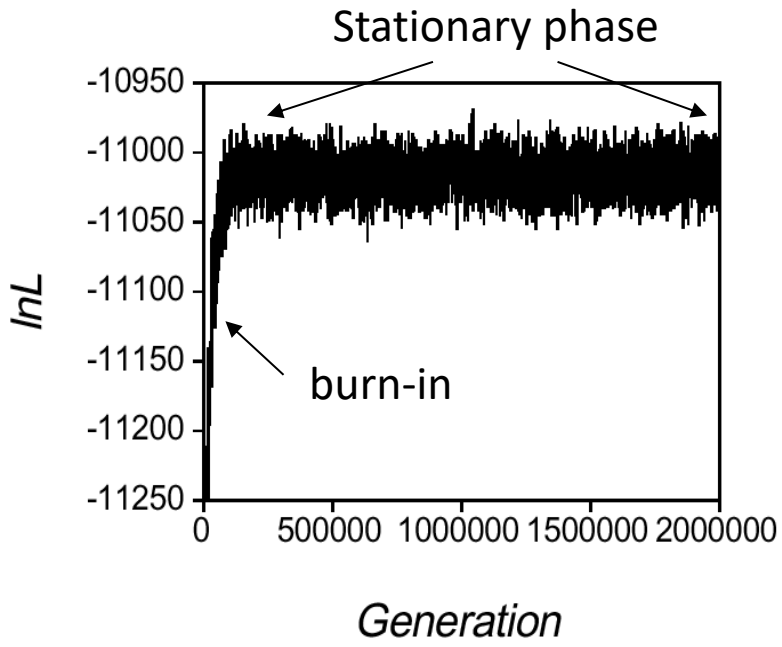
adapted from Ronquist and P. Lewis

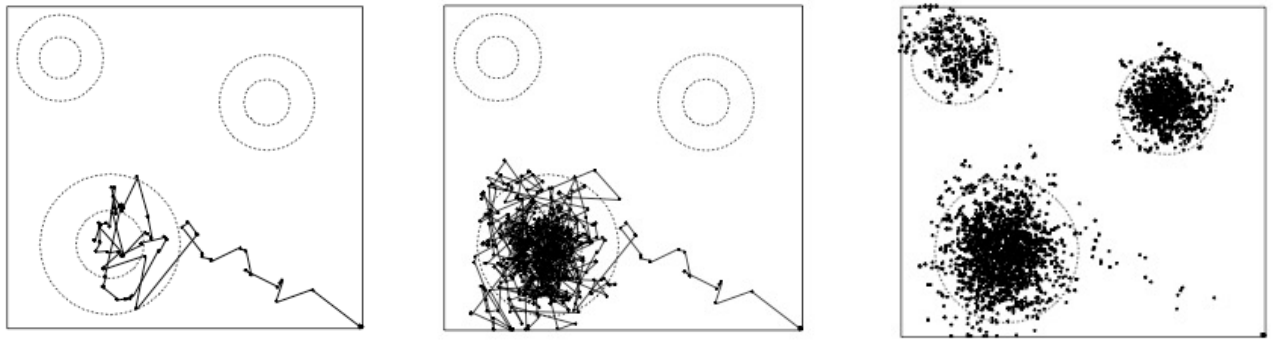




adapted from J. Felsenstein

The Trace Plot





Assessing **Convergence**:

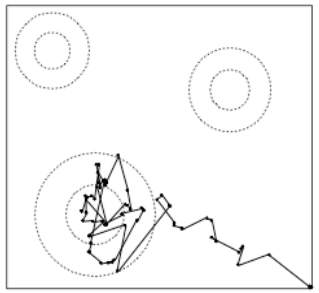
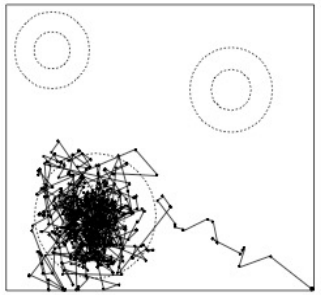
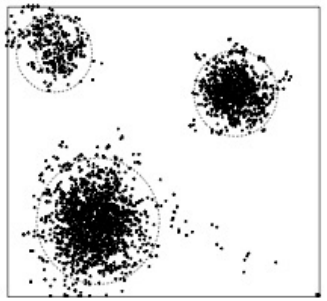
1. Check for the plateau in the trace plot
2. Look at sampling behavior within the run (autocorrelation times, effective sample size etc)
3. Compare independent runs with different, randomly chosen starting points

Summarizing sampled topologies

List of sampled topologies

τ^*

Results are summarized with credibility intervals and majority consensus trees

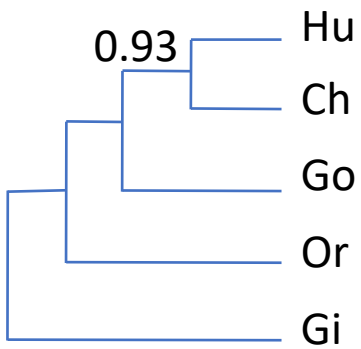




i	τ_i	$f(\boldsymbol{X}\|\tau_i)$
1	(Gi,Hu,((Ch,Go),Or))	0.000
2	(Gi,(Hu,(Ch,Go)),Or)	0.026
3	(Gi,(Hu,Or),(Ch,Go))	0.000
4	(Gi,((Hu,Or),Go),Ch)	0.000
5	(Gi,((Hu,Or),Ch),Go)	0.001
6	(Gi,Hu,((Ch,Or),Go))	0.000
7	(Gi,(Hu,Go),(Ch,Or))	0.000
8	(Gi,((Hu,Go),Ch),Or)	0.037
9	(Gi,((Hu,Go),Or),Ch)	0.000
10	(Gi,(Hu,(Ch,Or)),Go)	0.001
11	(Gi,Hu,(Ch,(Go,Or)))	0.001
12	(Gi,(Hu,(Go,Or)),Ch)	0.001
13	(Gi,(Hu,Ch),(Go,Or))	0.004
14	(Gi,((Hu,Ch),Go),Or)	0.919
15	(Gi,((Hu,Ch),Or),Go)	0.009

The posterior probability of a clade is simply the sum of the posterior probabilities of all trees that contain that clade.

Example:

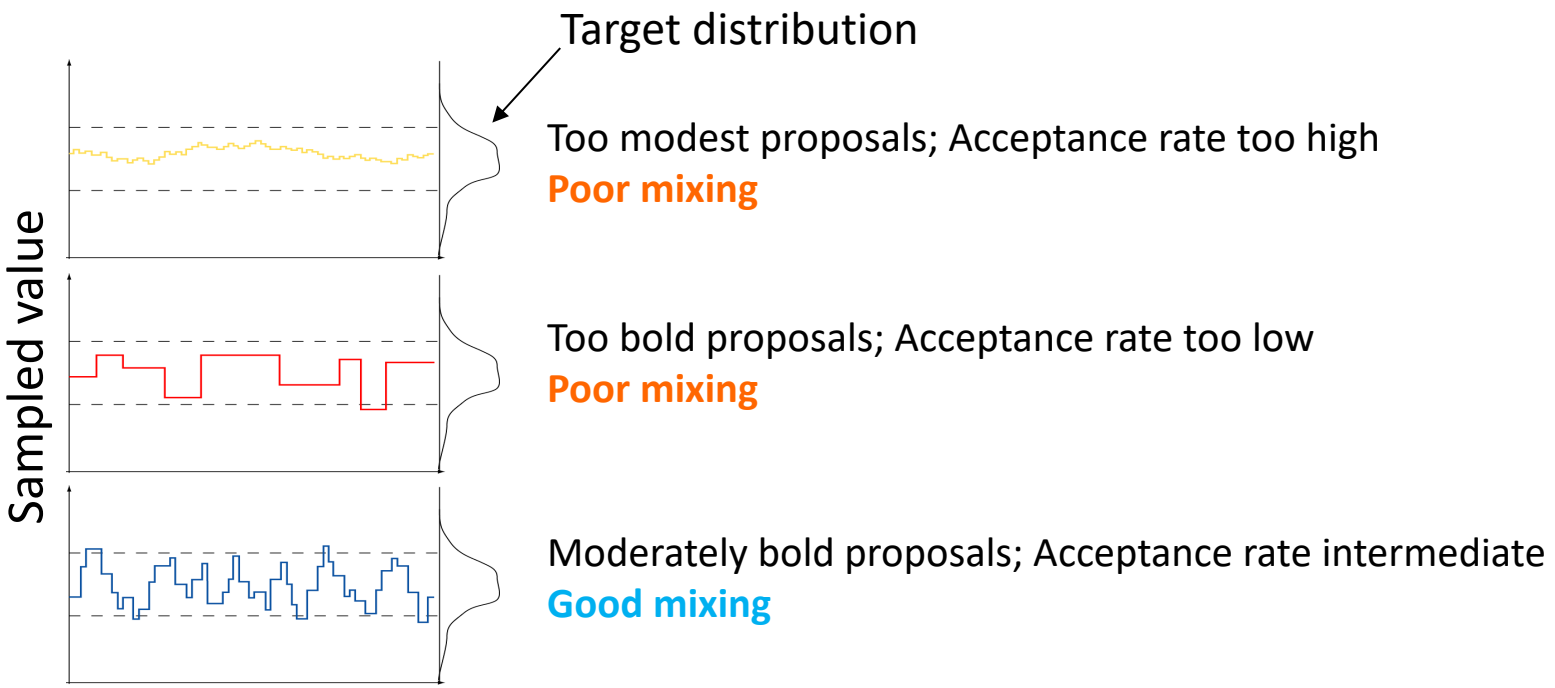
- A credible 95% interval for these topologies includes trees 14 and 8 trees -> $f(X|\tau_i)=0.956$
- The probability of the Human-Chimp clade is T13 + T14 + T15 = 0.932



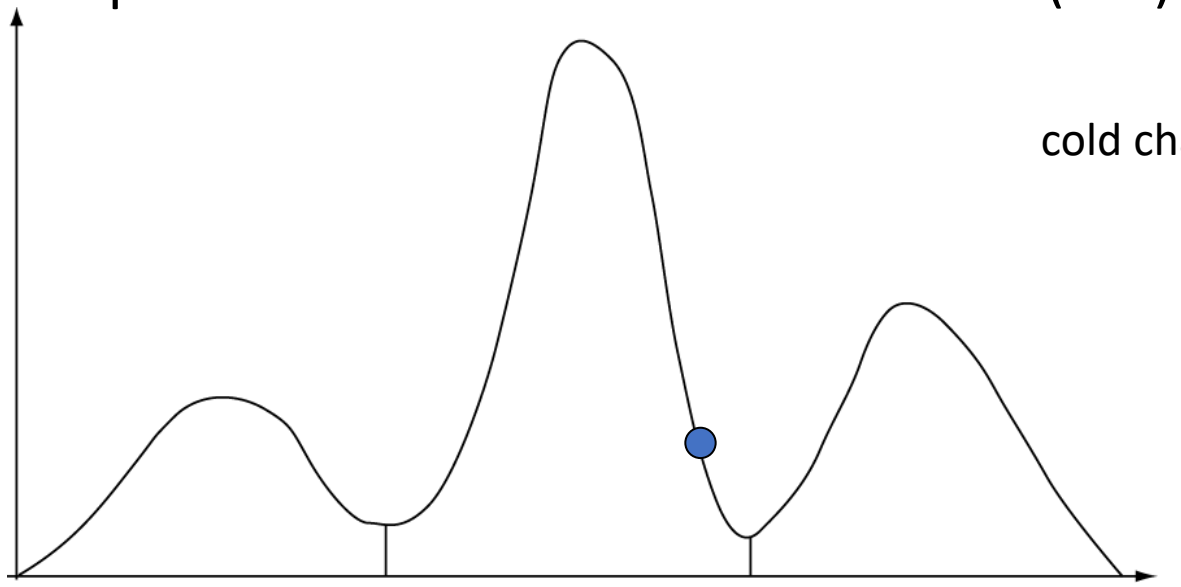
A majority rule consensus tree is formed by combining all the clades with the highest posterior probability that are compatible

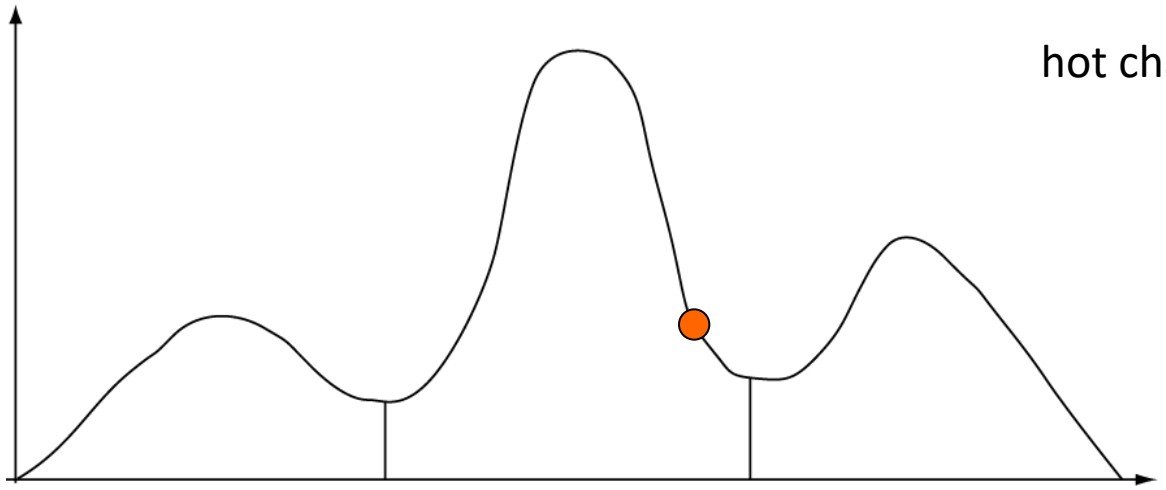
Mixing refers to how often proposed changes to parameters are accepted during the MCMC run. High acceptance rate means chain is making too small moves. Low acceptance rate means proposed changes are too large.
Optimal acceptance rate: 20-60 percent.

The time it takes for a MCMC to obtain an adequate sample of the posterior depends on its mixing behavior

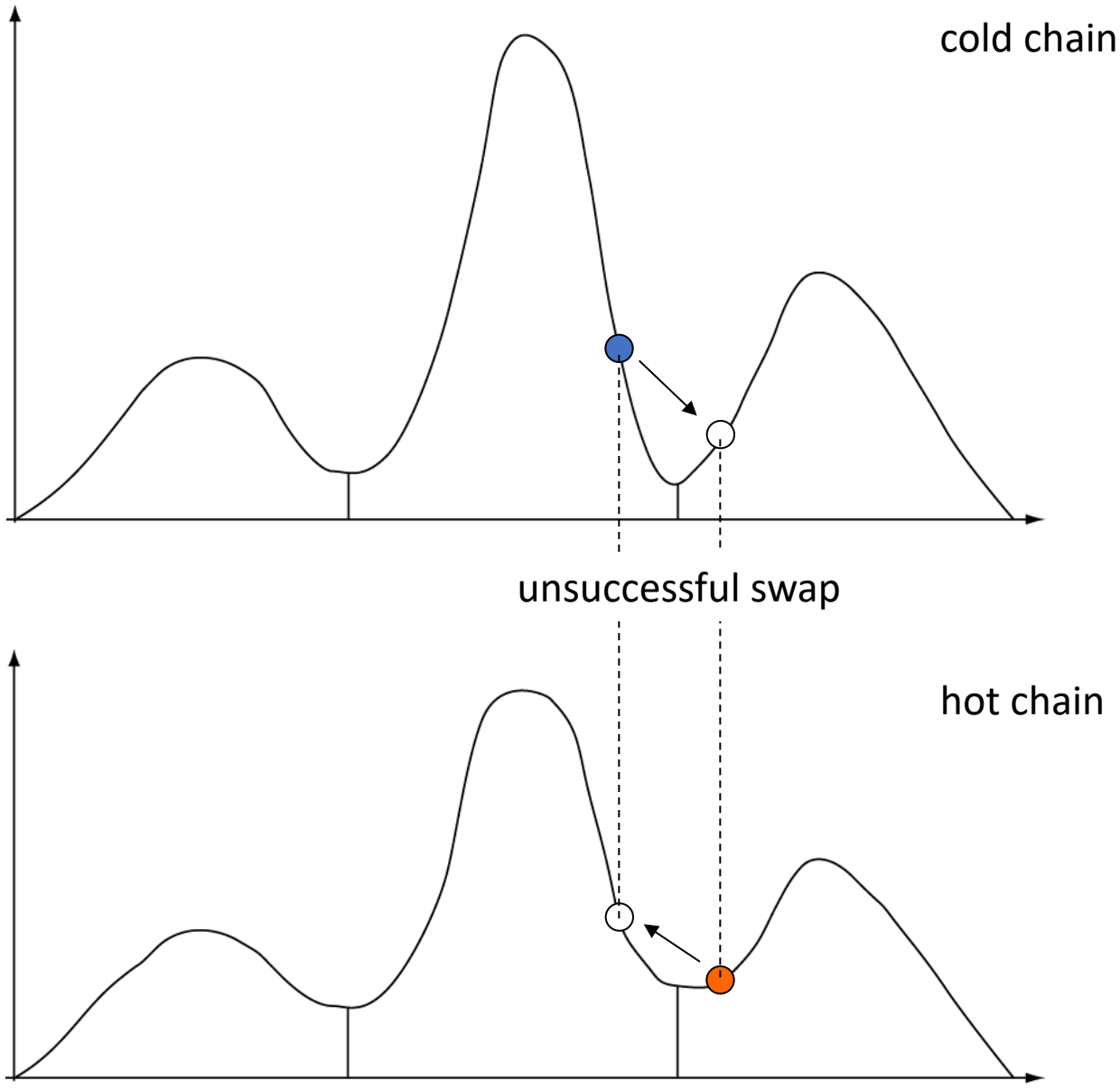


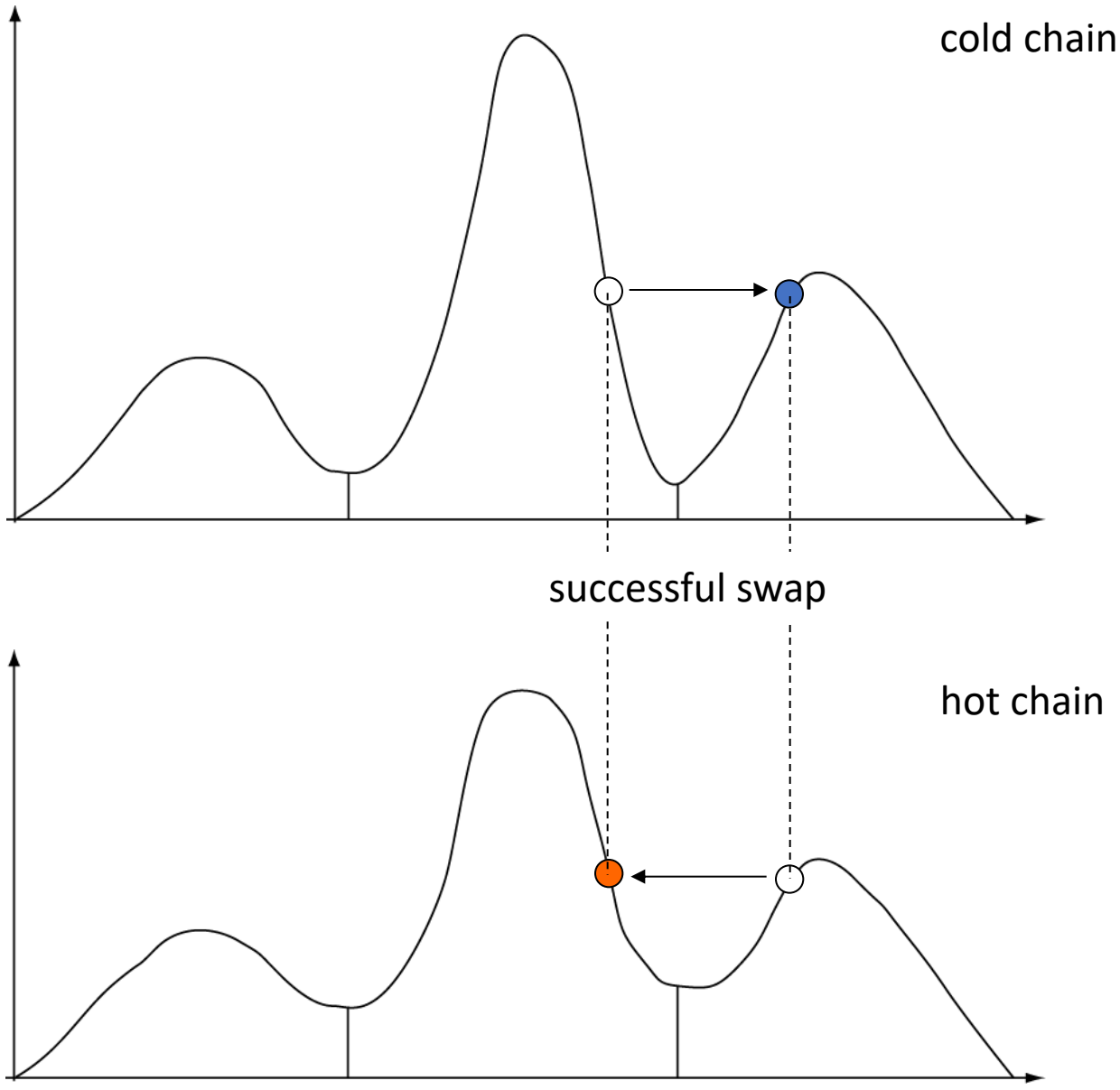
Metropolis-**c**oupled **M**arkov **c**hain **M**onte **C**arlo aka $(MC)^3$





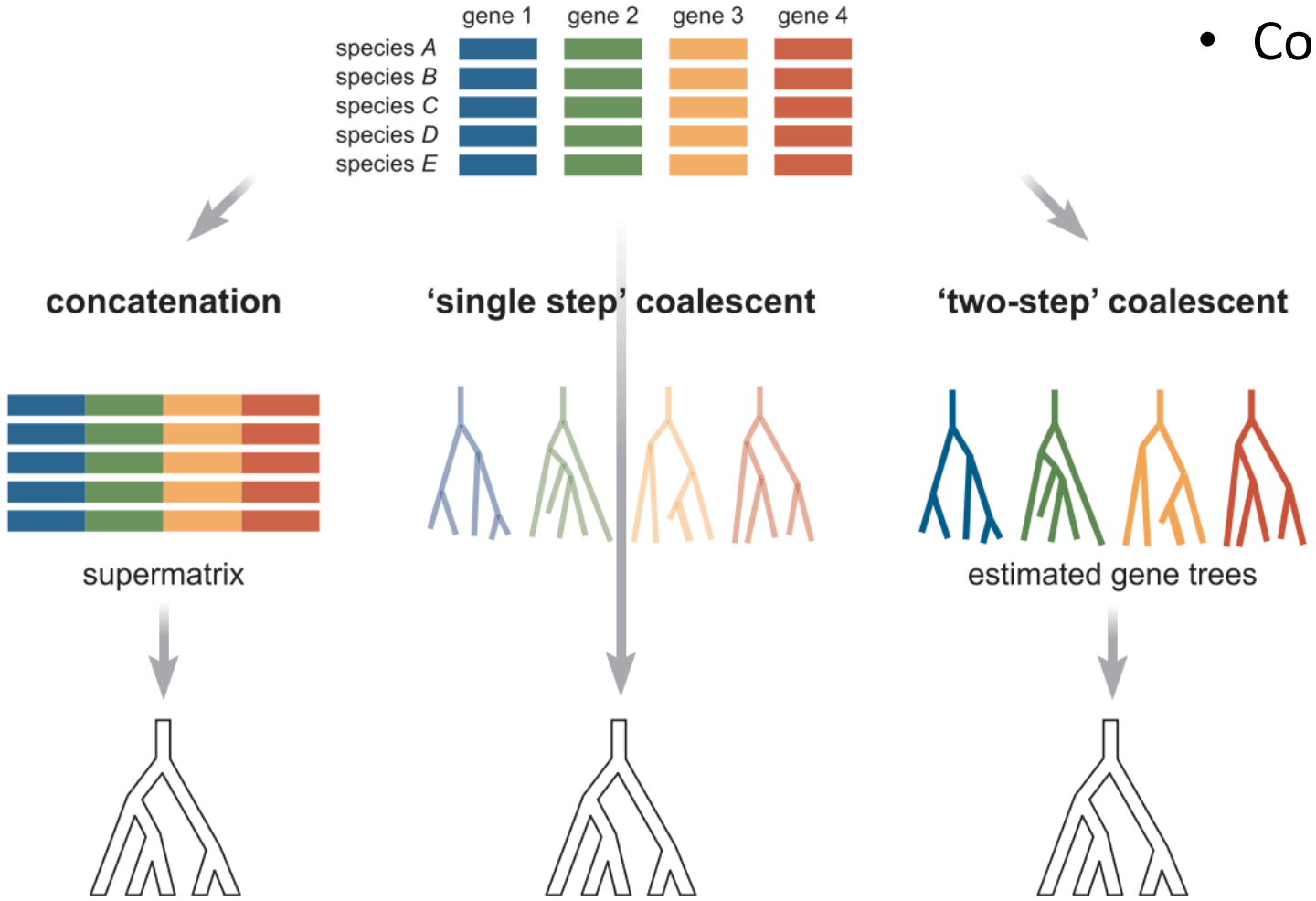
Slides from Ronquist


cold chain
unsuccessful swap
hot chain

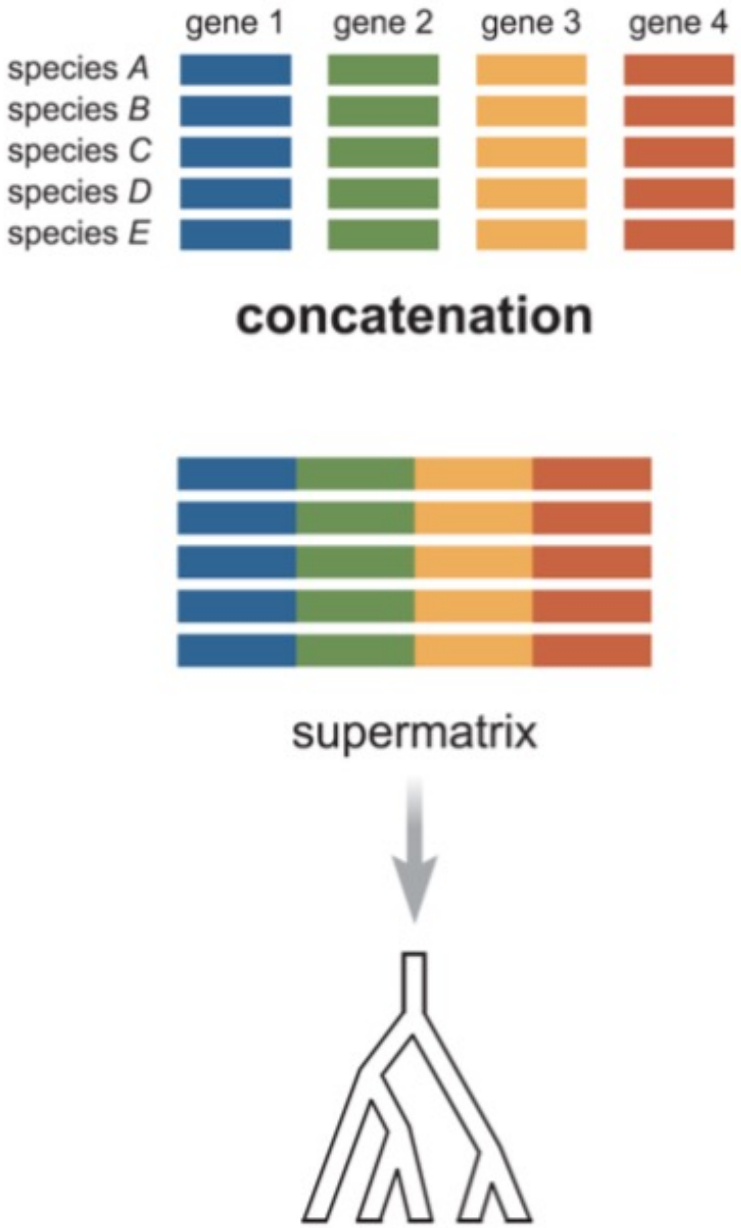

cold chain
successful swap
hot chain

Inference of large phylogenies

- Large datasets
- Many calculations
- Complex data



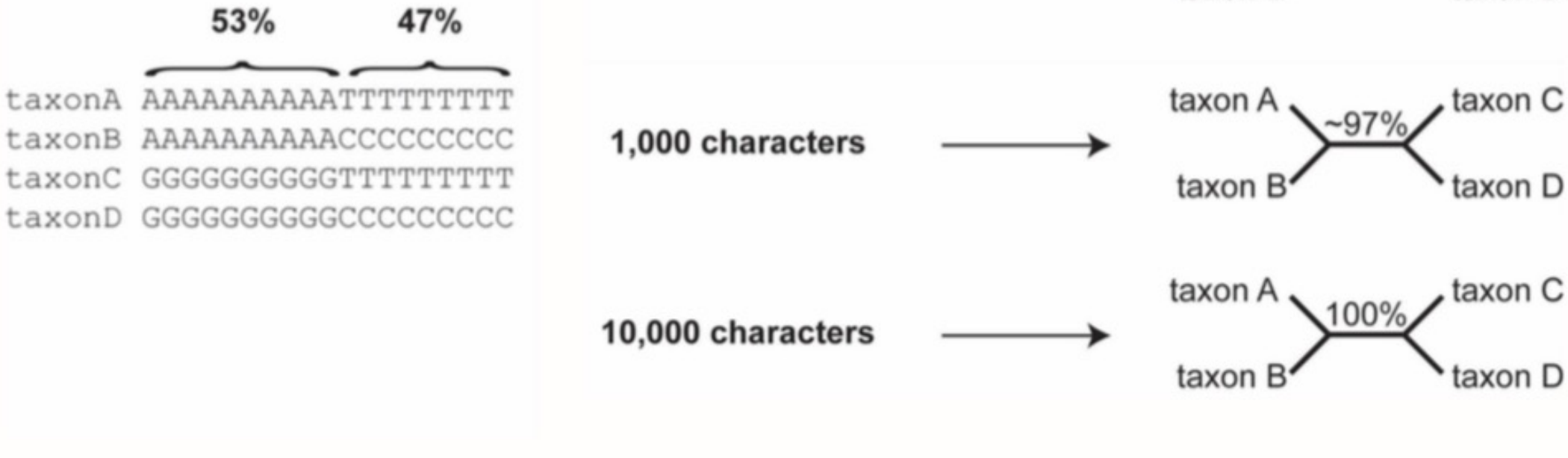
Liu *et al* 2015



Inference of large phylogenies

- The implicit assumption is that all genes do in fact have the same gene tree, that these gene trees are congruent with and converge on the **species tree**
- The use of many genes eliminates **stochastic error** (e.g. insufficient sequence length) and **systematic error** (some gene trees my depart from model assumptions)
- If we add extra requirements such as **single copy orthologs** and **core genes** then we might also reduce/eliminate biological causes of incongruence between gene tree and species tree

Things to be aware of… analysis of large concatenated datasets may lead to **misleading bootstrap support**



Rokas and Carroll (2006) Plos Biol.

Things to be aware of… sometimes all gene trees differ from each other and from the concatenation phylogeny!!

doi:10.1038/nature12130

Inferring ancient divergences requires genes with strong phylogenetic signals

Leonidas Salichos[1] & Antonis Rokas[1]

To tackle incongruence, the topological conflict between different gene trees, phylogenomic studies couple concatenation with practices such as rogue taxon removal or the use of slowly evolving genes. Phylogenomic analysis of 1,070 orthologues from 23 yeast genomes identified 1,070 distinct gene trees, which were all incongruent with the phylogeny inferred from concatenation. Incongruence severity increased for shorter internodes located deeper in the phylogeny. Notably, whereas most practices had little or negative impact on the yeast phylogeny, the use of genes or internodes with high average internode support significantly improved the robustness of inference. We obtained similar results in analyses of vertebrate and metazoan phylogenomic data sets. These results question the exclusive reliance on concatenation and associated practices, and argue that selecting genes with strong phylogenetic signals and demonstrating the absence of significant incongruence are essential for accurately reconstructing ancient divergences.

Salichos and Rokas Nature 2013

Some recipes for handling incongruence in concatenation analysis:

- Remove all sites containing gaps
- Remove fast-evolving or unstable species
- Selecting genes that recover specific clades
- Selecting the most slow-evolving genes
- Selecting genes whose bootstrap consensus trees have high average support
- Multiple searches using distinct starting trees

Strategy:

- Apply different phylogenetic methods (different optimality criteria/approaches)
- Assess conflict across gene trees
- Investigate alternative hypotheses for branches showing conflict/assess sensitivity of results

References

The Phylogenetic Handbook, A practical approach to phylogenetic analysis and hypothesis testing. Ed, Philippe Lemey, Marco Salemi, Anne-Mieke Vandamme. 2009

Inferring Phylogenies, Joe Felsenstein 2004

Software

Iqtree http://www.iqtree.org/

MrBayes http://mrbayes.scs.fsu.edu

RaxML https://cme.h-its.org/exelixis/web/software/raxml/