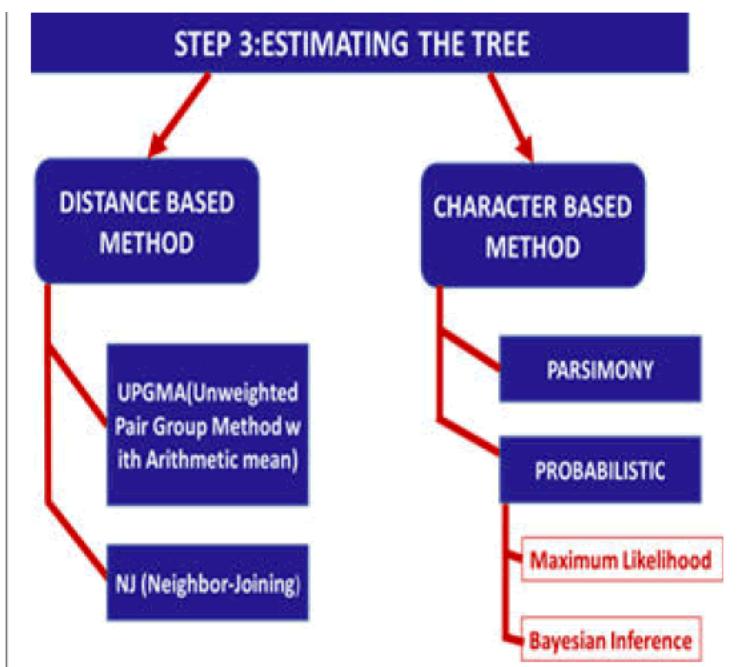
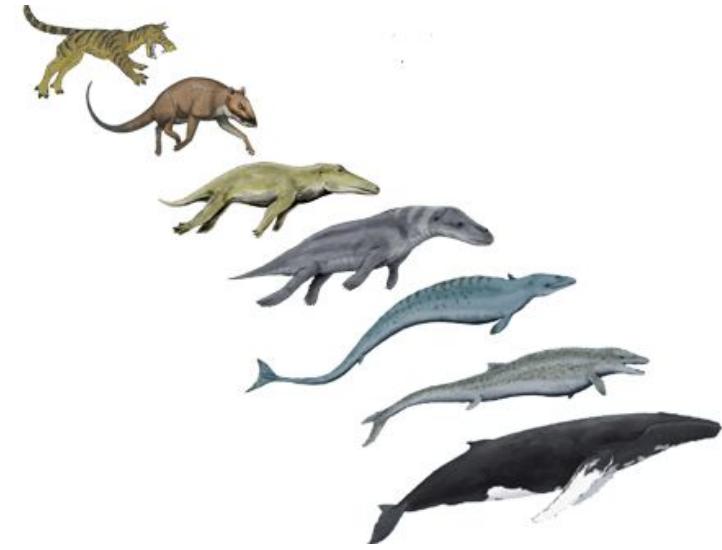
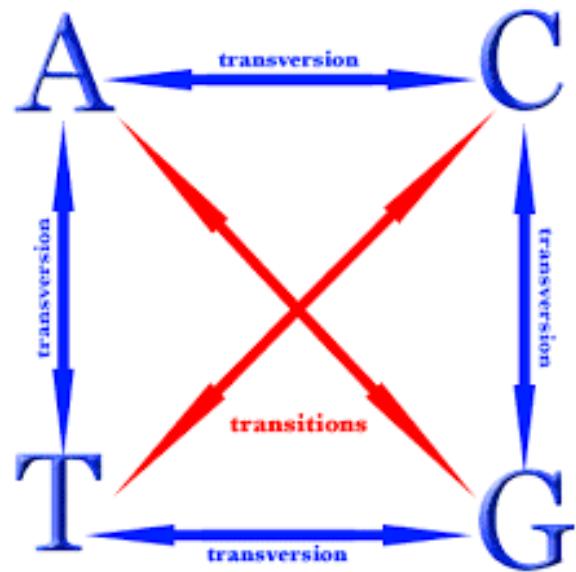


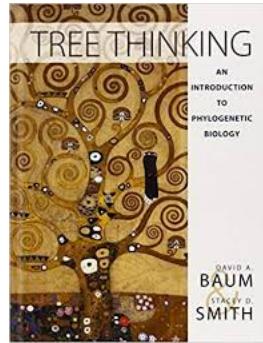
Maximum-Likelihood and Bayesian Inference

22 Oct 2020

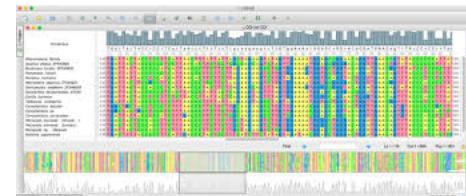


Covered on Tuesday

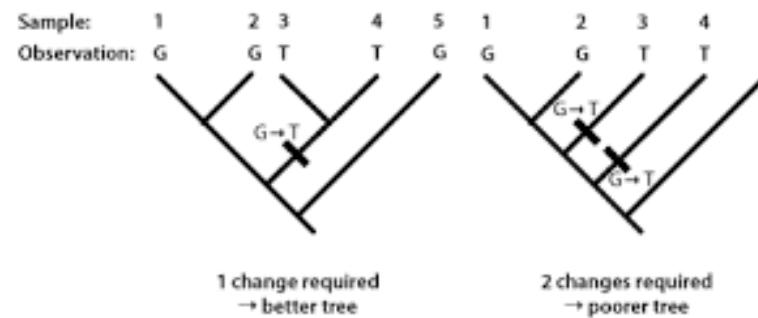
- Tree thinking



- Sequence alignment



- Parsimony for tree inference



Topics to cover today

Maximum Likelihood

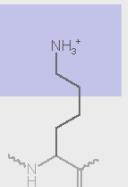
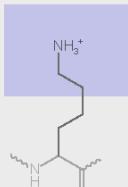
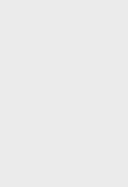
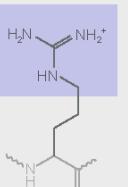
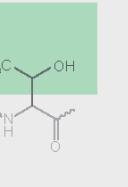
- Specifying models of molecular evolution
- Maximum-likelihood inference
- Bootstrap support

Bayesian Inference

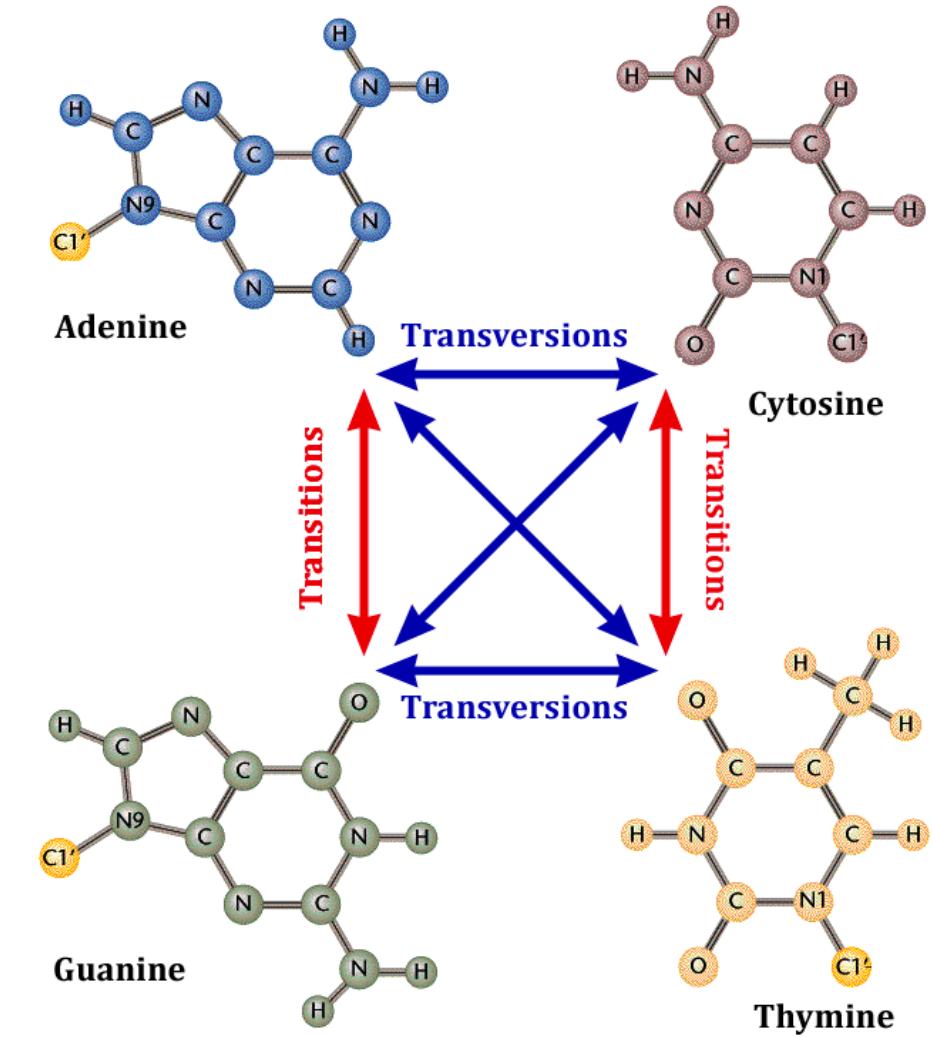
- Bayes theorem
- Bayesian phylogenetics and MCMC

How do nucleotides change?

- Change at any given site is independent of the base in its prior iteration

DNA level	Point mutations				
	No mutation	Silent	Nonsense	Missense	
				conservative	non-conservative
TTC	TTT	ATC	TCC	TGC	
AAG	AAA	UAG	AGG	ACG	
Lys	Lys	STOP	Arg	Thr	
					
DNA level					
mRNA level					
protein level					

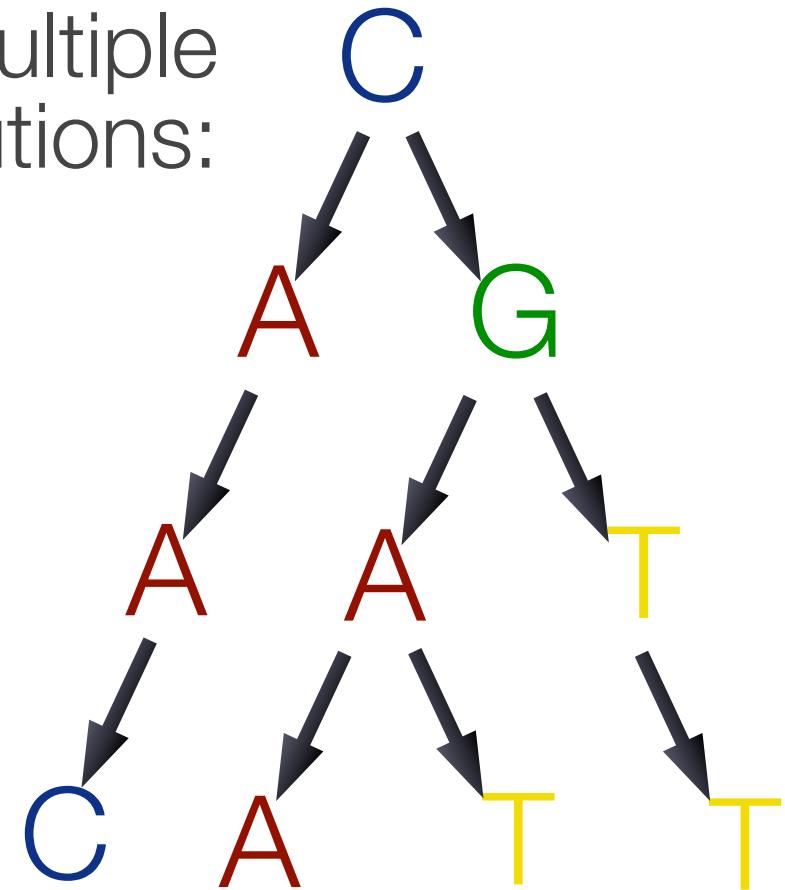
Legend: basic (purple), polar (green)



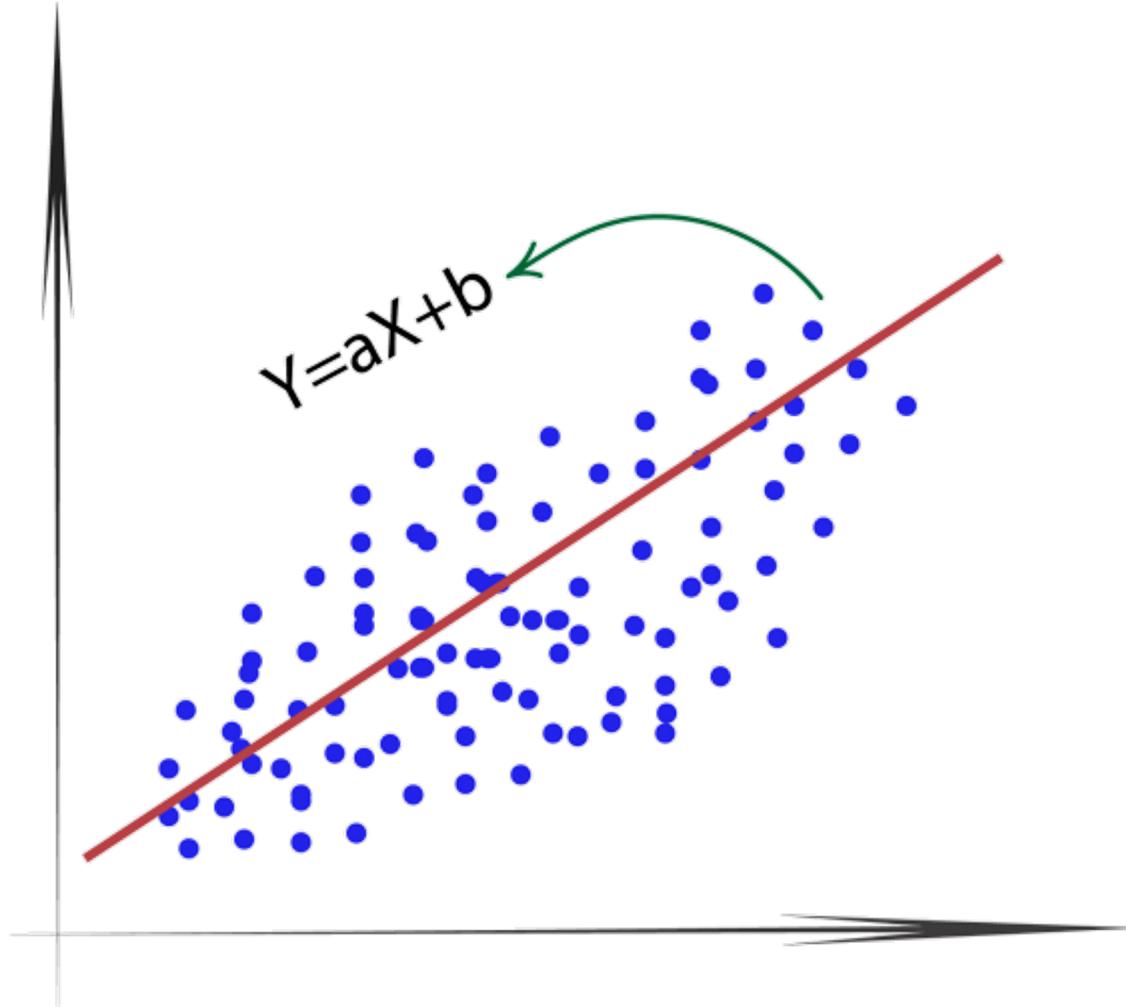
Multiple substitutions

- Given enough time, may be multiple changes
- 25% of nucleotide sites are expected to be identical by chance
- Models vary in complexity from very simple to extremely complex
- Model choice is important
- Model must suit the data

multiple
substitutions:

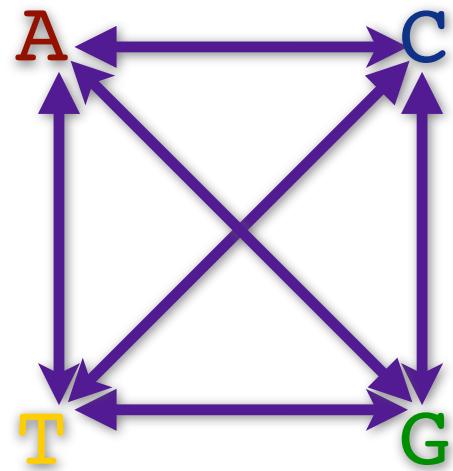


A model?



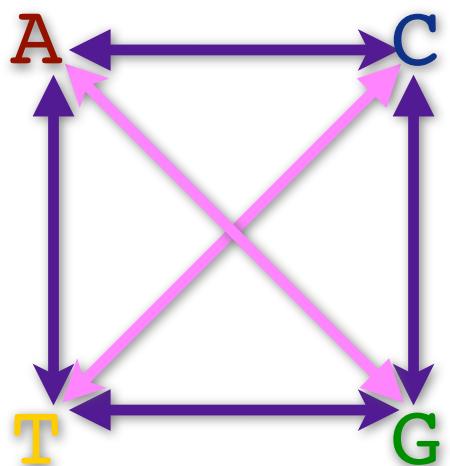
Underlying models

Jukes Cantor



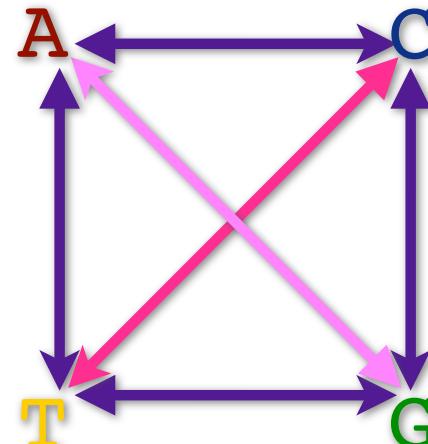
- All bases evolve independently
- All bases are equal frequency
- Each base can change with equal probability

Kimura



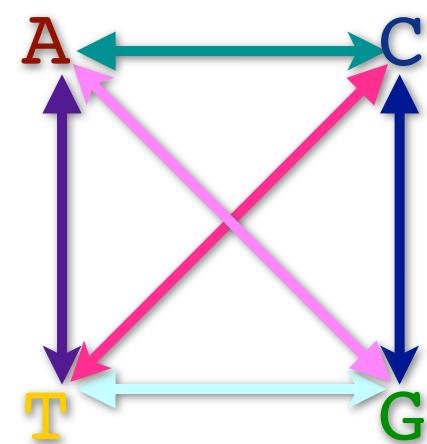
- All bases evolve independently
- All bases are equal frequency
- Transitions and transversions evolve at different rates

TrN



- All bases evolve independently
- All bases at unequal frequency
- Transitions and transversions evolve at two different rates

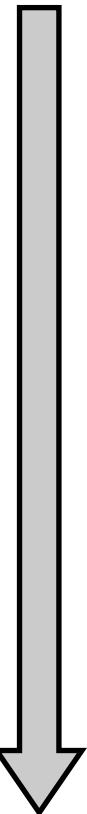
GTR



- All bases evolve independently
- All bases at unequal frequency
- All changes occur at different rates

Model hierarchy

Simple



Complex

base frequencies are equal and
all substitutions are equally likely
(Jukes-Cantor)



base frequencies are equal but transitions and
transversions occur at different rates
(Kimura 2 parameter)

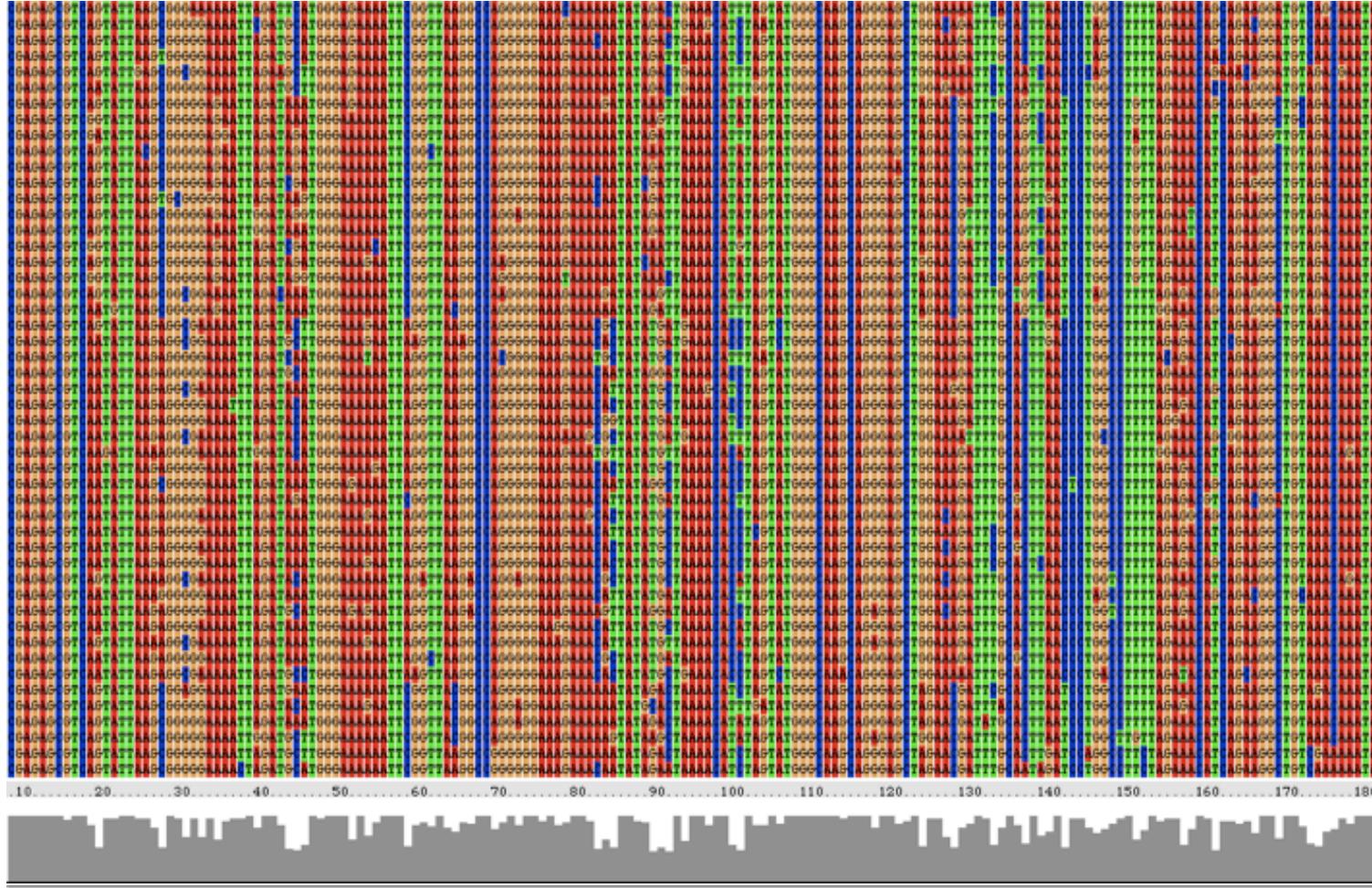


unequal base frequencies and transitions and
transversions occur at different rates
(Hasegawa-Kishino-Yano)



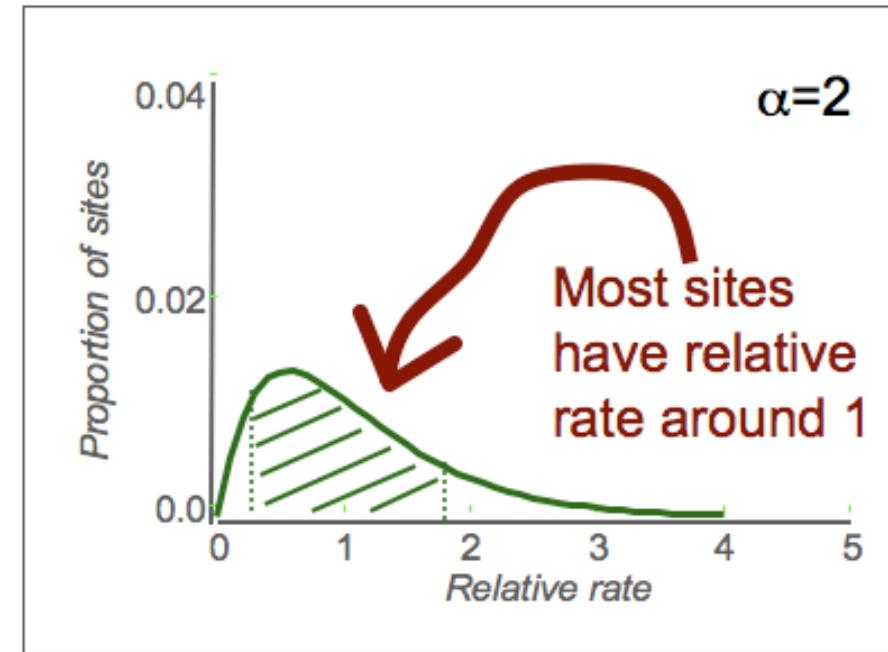
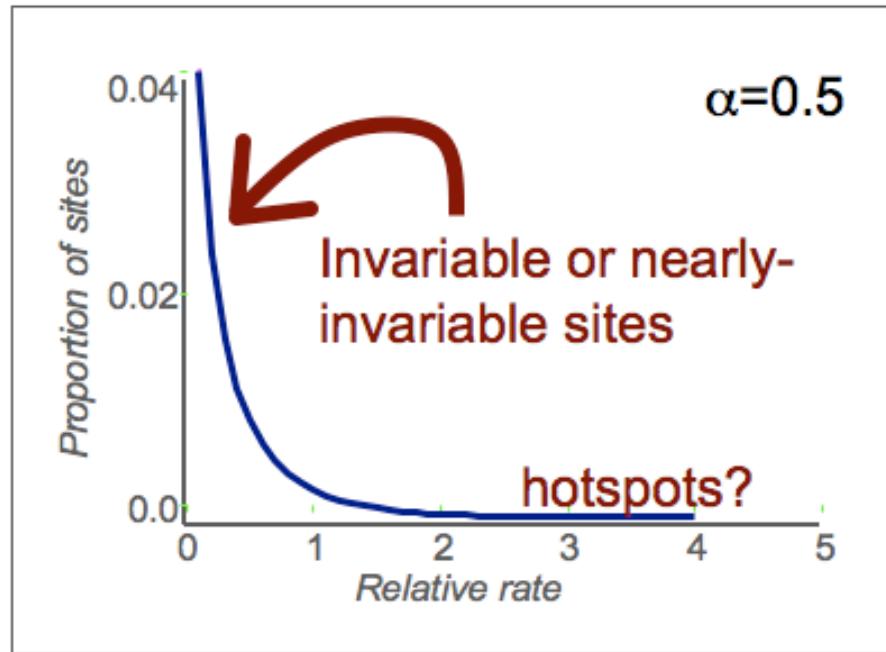
unequal base frequencies and all
substitution types occur at different rates
(General Reversible Model)

Rate Heterogeneity between sites



- Most models assume rate heterogeneity
 - 3rd position wobble
 - Hypervariable vs invariant sites
- Use a gamma distribution to model heterogeneity
 - Usually 4-8 discrete categories of rates

Rate heterogeneity

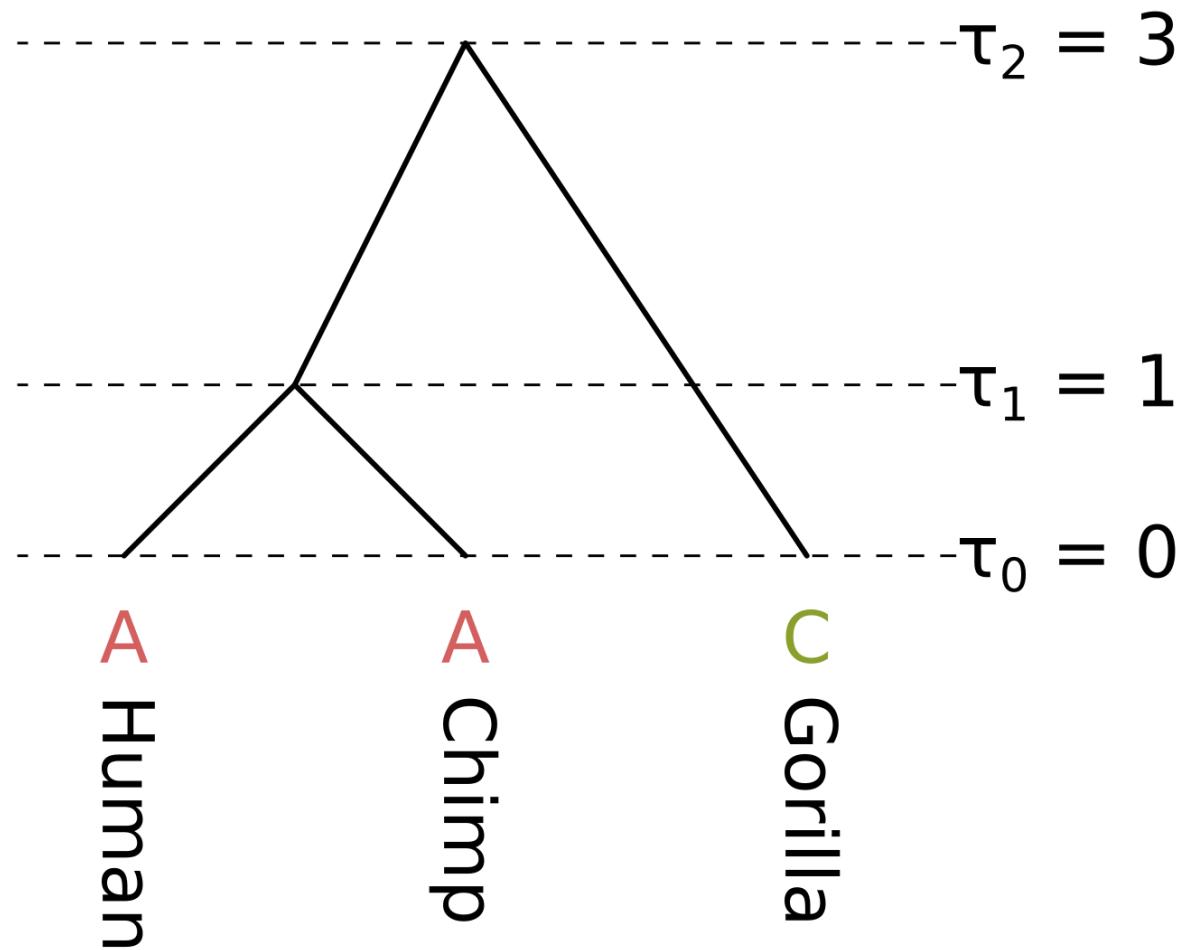


- Strong rate heterogeneity $\alpha < 1$
- Weak rate heterogeneity $\alpha > 1$

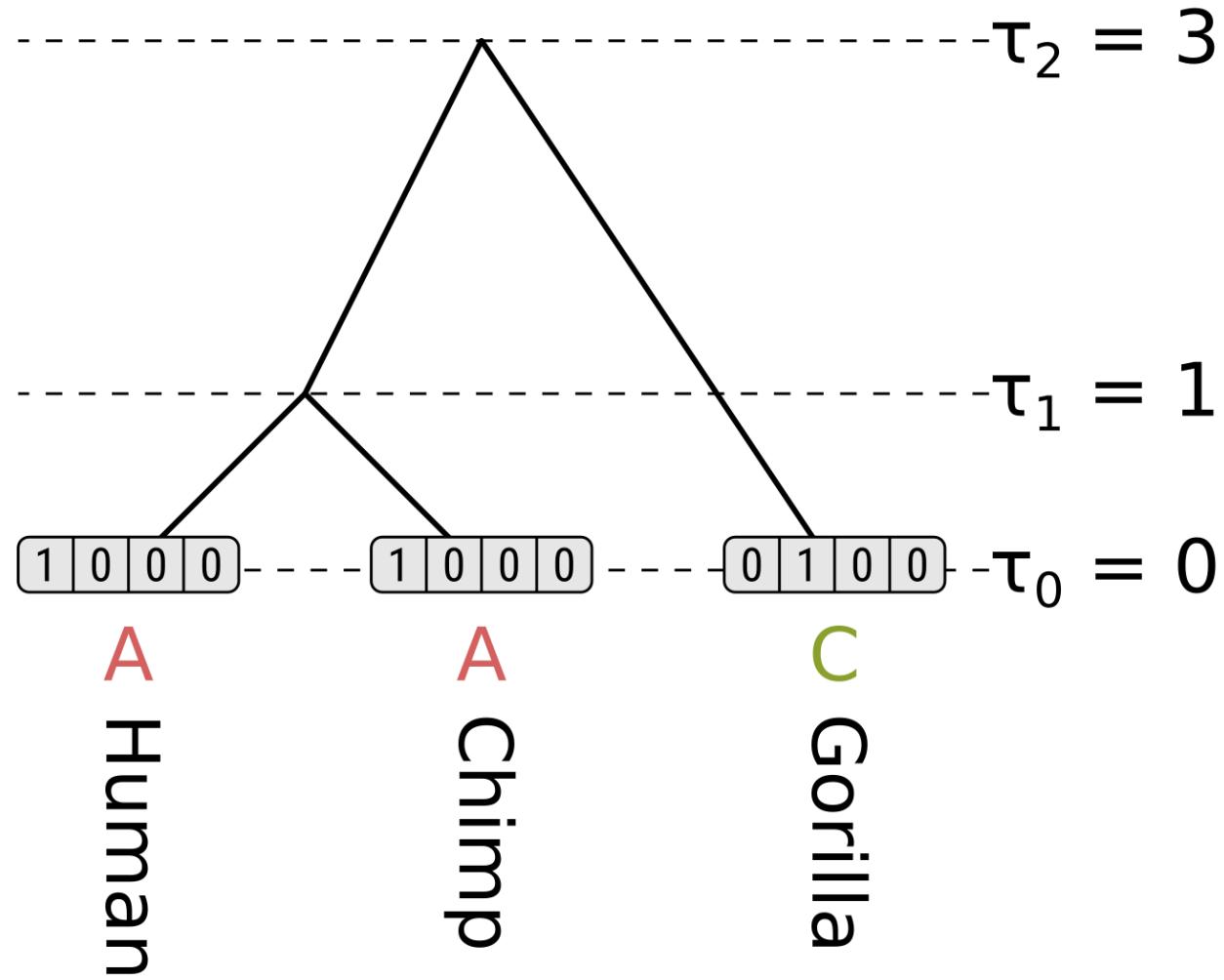
Maximum likelihood

- Closely related to the more common concept of probability
- Considered the most statistically valid approach for molecular phylogenetics
 - Along with Bayesian
- Incorporates detailed models of molecular evolution

Parsimony v. Likelihood



Parsimony v. Likelihood



Parsimony v. Likelihood

For the human and chimp branches, these will be (to four decimal places):

$$P_{xx}(0.1) = \frac{1}{4} \left(1 + 3e^{-\frac{4}{3}0.1} \right) = 0.9064$$

$$P_{xy}(0.1) = \frac{1}{4} \left(1 - e^{-\frac{4}{3}0.1} \right) = 0.0312$$

For the HC branch, these will be:

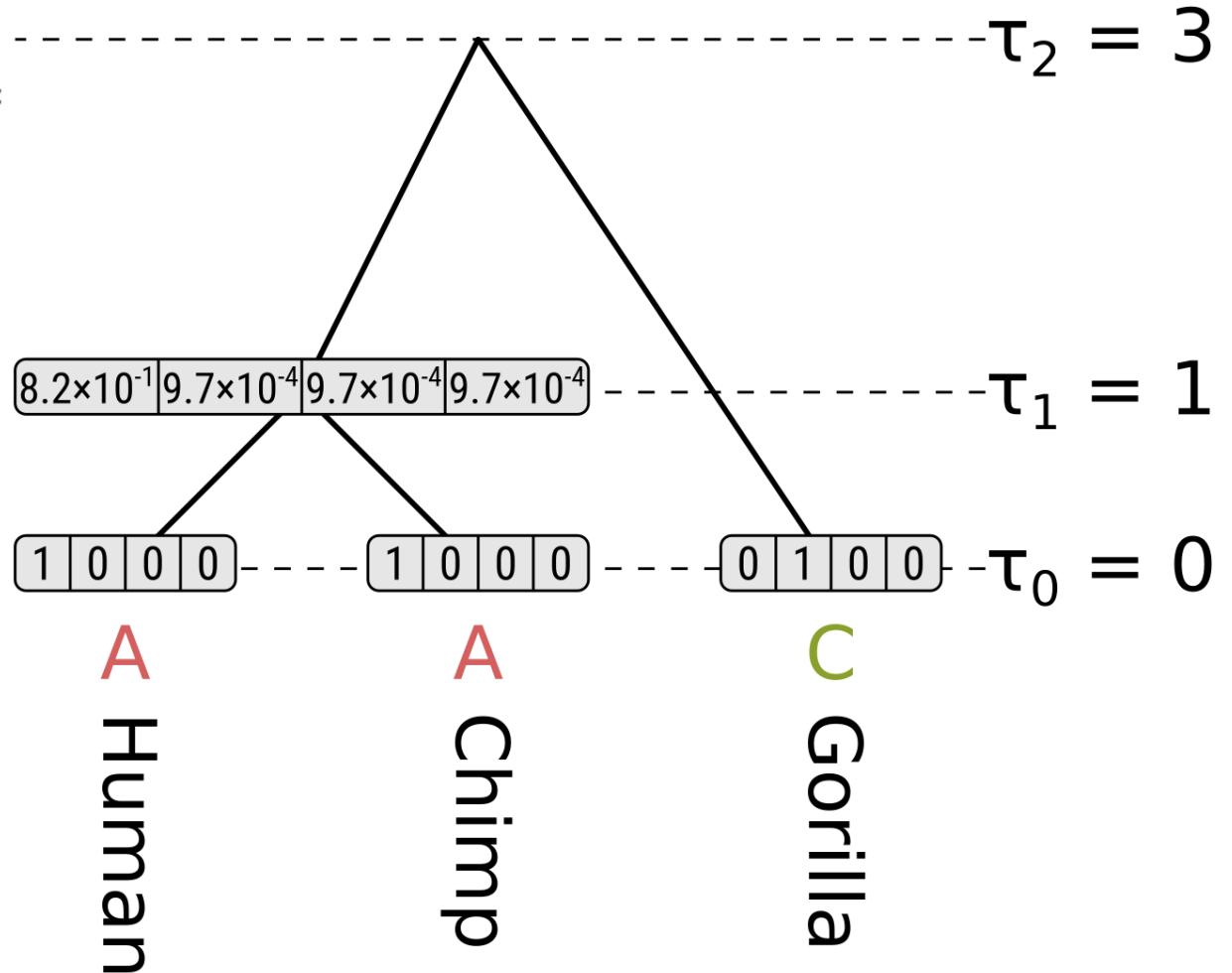
$$P_{xx}(0.2) = \frac{1}{4} \left(1 + 3e^{-\frac{4}{3}0.2} \right) = 0.8245$$

$$P_{xy}(0.2) = \frac{1}{4} \left(1 - e^{-\frac{4}{3}0.2} \right) = 0.0585$$

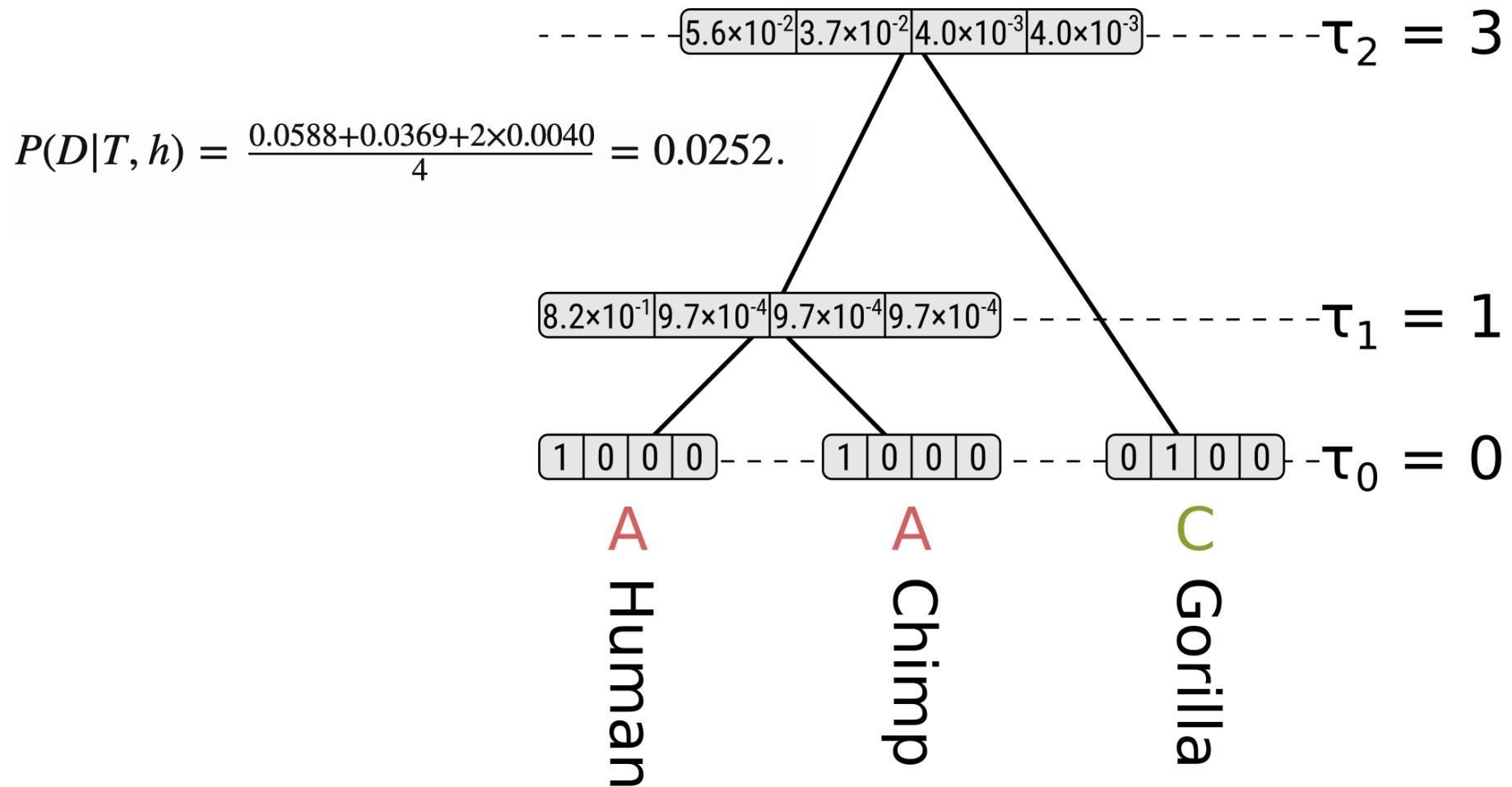
For the gorilla branch, these will be:

$$P_{xx}(0.3) = \frac{1}{4} \left(1 + 3e^{-\frac{4}{3}0.3} \right) = 0.7528$$

$$P_{xy}(0.3) = \frac{1}{4} \left(1 - e^{-\frac{4}{3}0.3} \right) = 0.0824$$

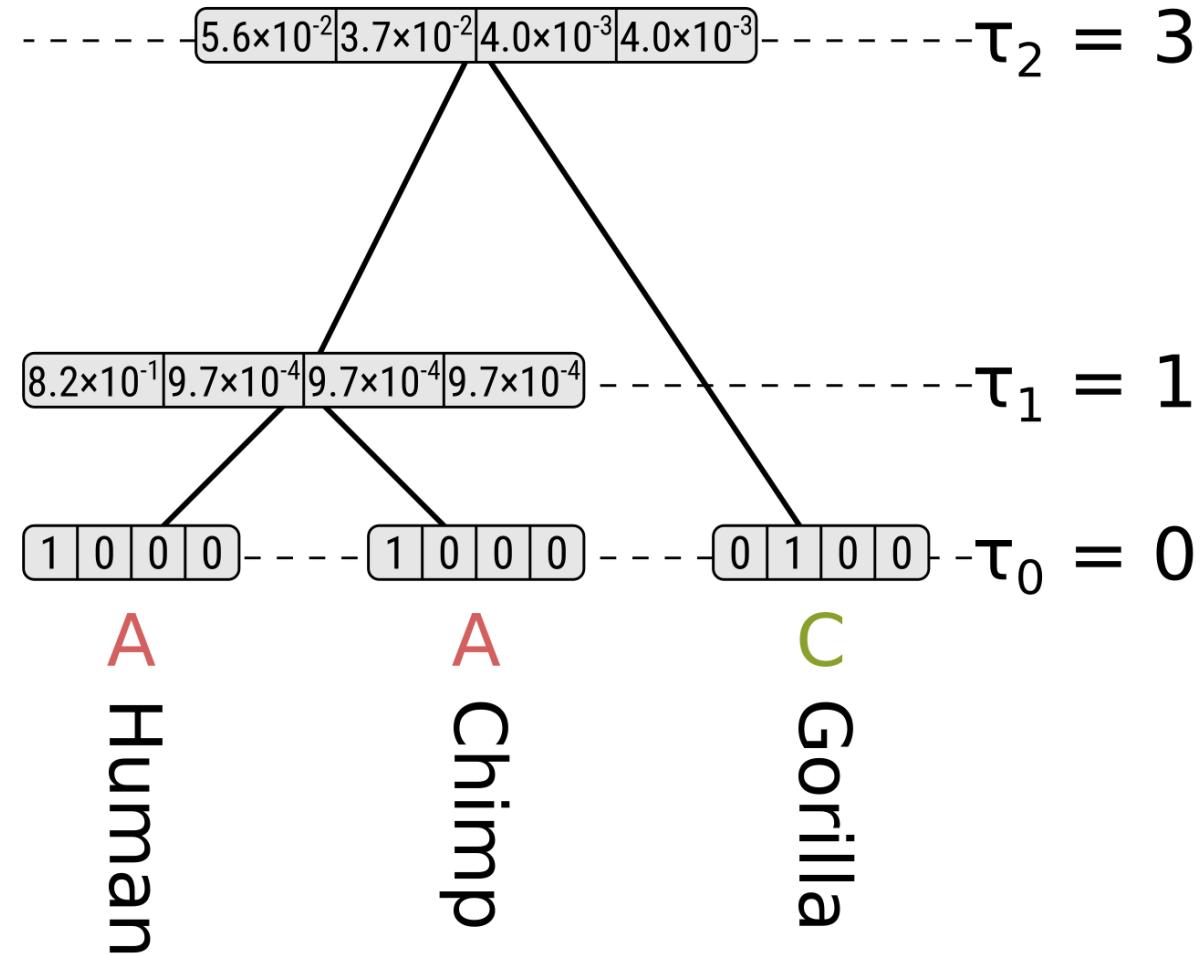


Parsimony v. Likelihood



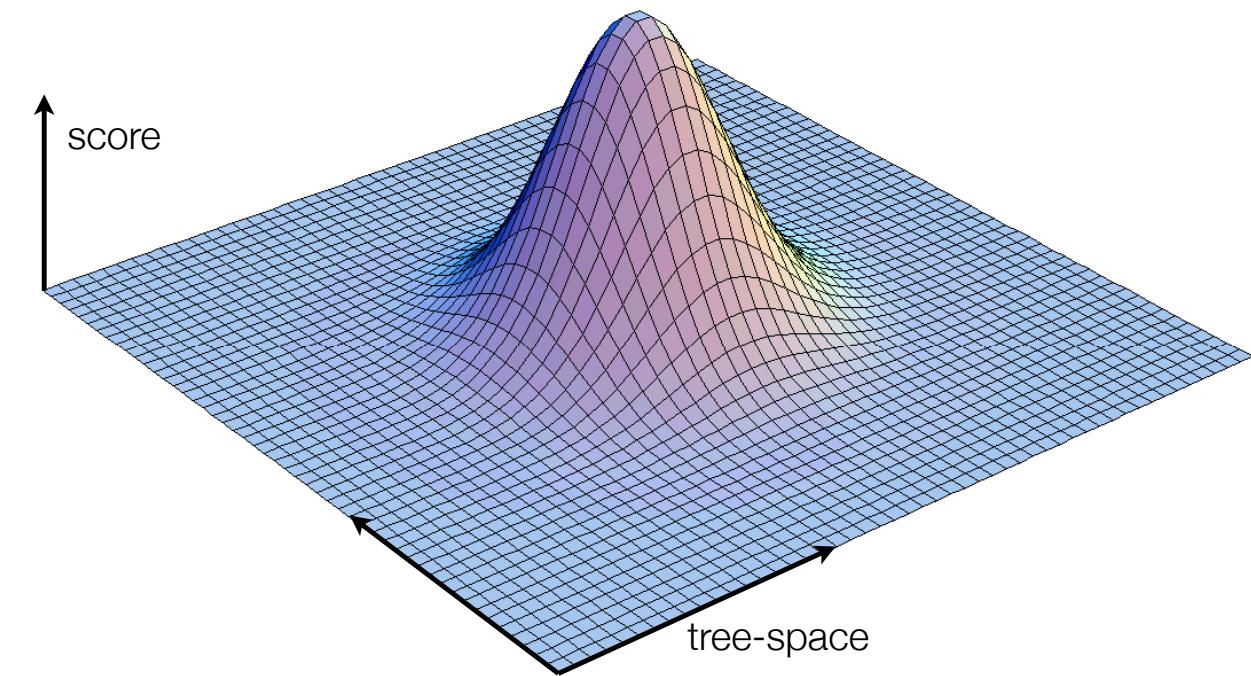
Parsimony v. Likelihood

Parsimony score would be?



Tree space

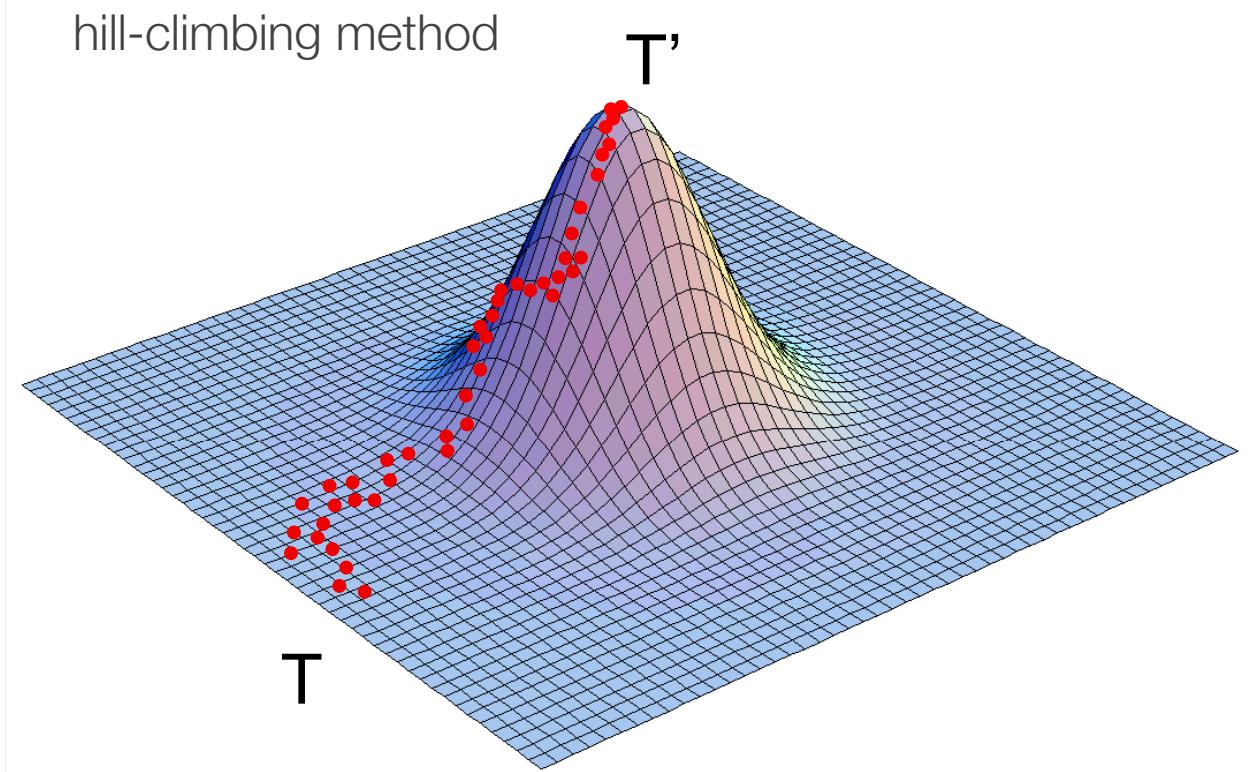
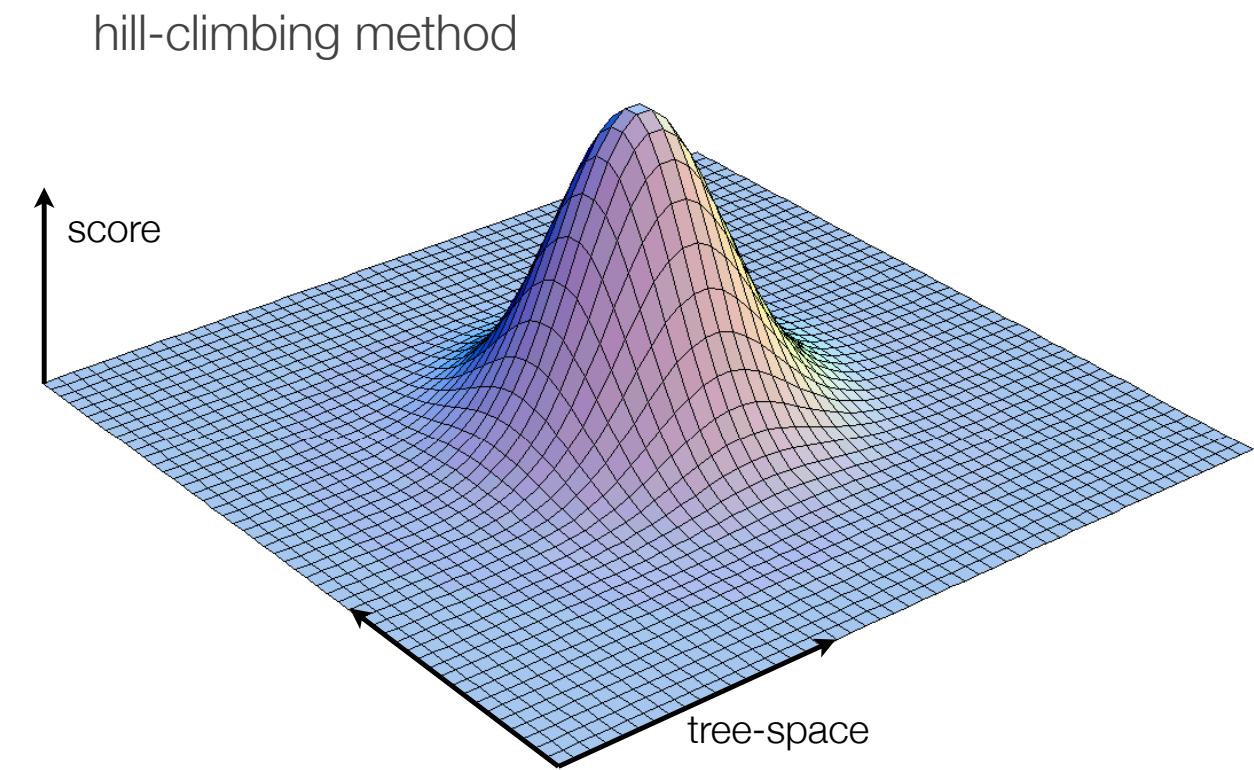
hill-climbing method



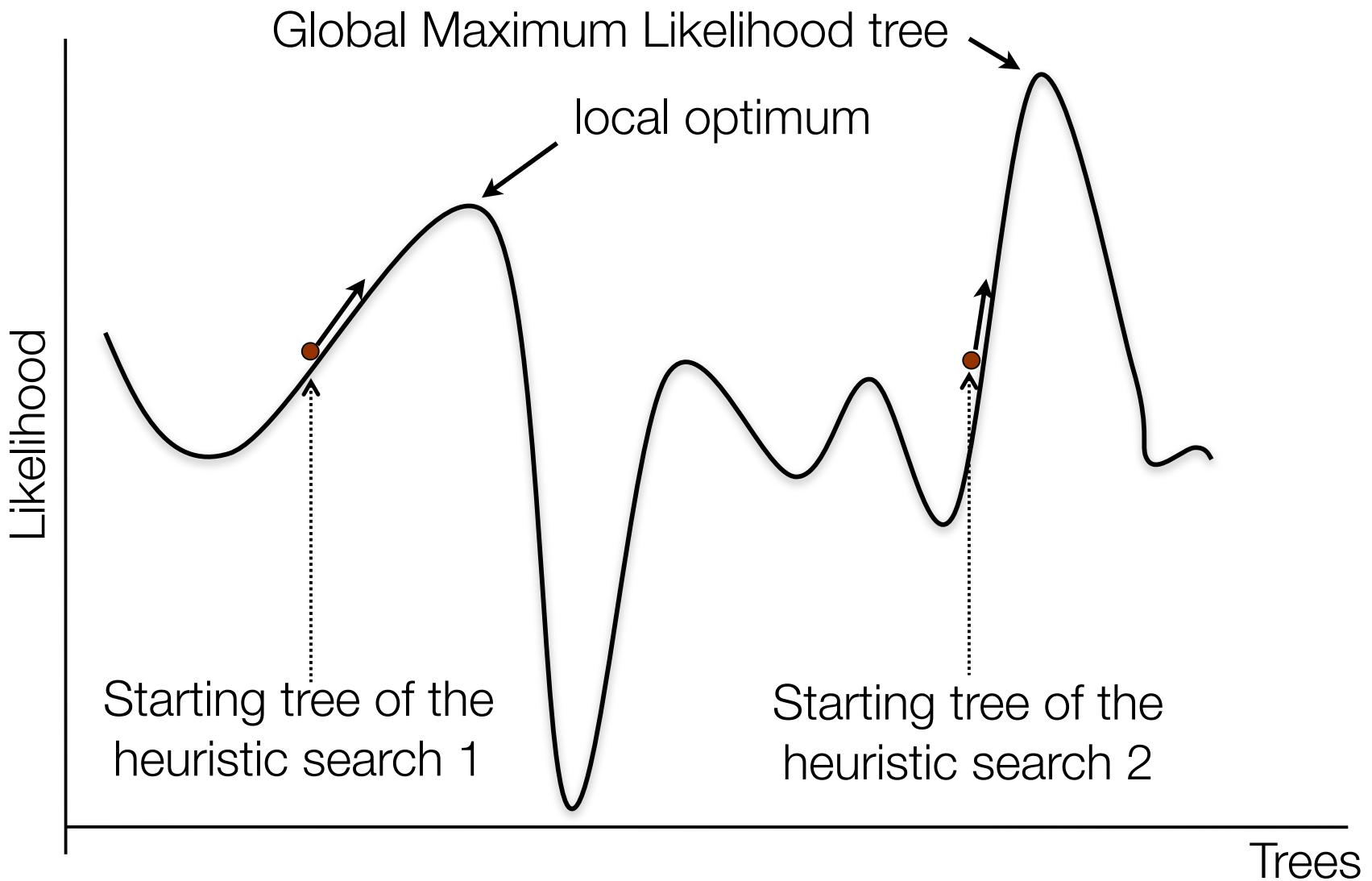
ML search

- Start with a tree (T)
- Perturb with branch-swapping (T')
- Calculate likelihood of tree
- If T' has better score than T , continue walk; if not, try different perturbations
- For each tree, need to optimize parameters of the model

Tree space



Tree space



Programs for inferring trees

- Garli
 - Limited GUI but full functioning command line
 - Complete control of model selection
 - Great for small data sets <300 tips and one gene, but slow for many genes and tons of taxa
- RAxML
 - Great for large data sets; also have a RAXML NG for NGS data
 - Limitation in the number of models of molecular evolution
 - Command line, but there is a RAxML BlackBox
- Mega
 - GUI version, easy to learn
 - However not widely accepted in publications (not all reasons are supported)
 - May not work very well for large data sets, but fast on small data sets

What kind of data do we need?

- Alignment files in either fasta or phylip format

Fasta

Phyllum

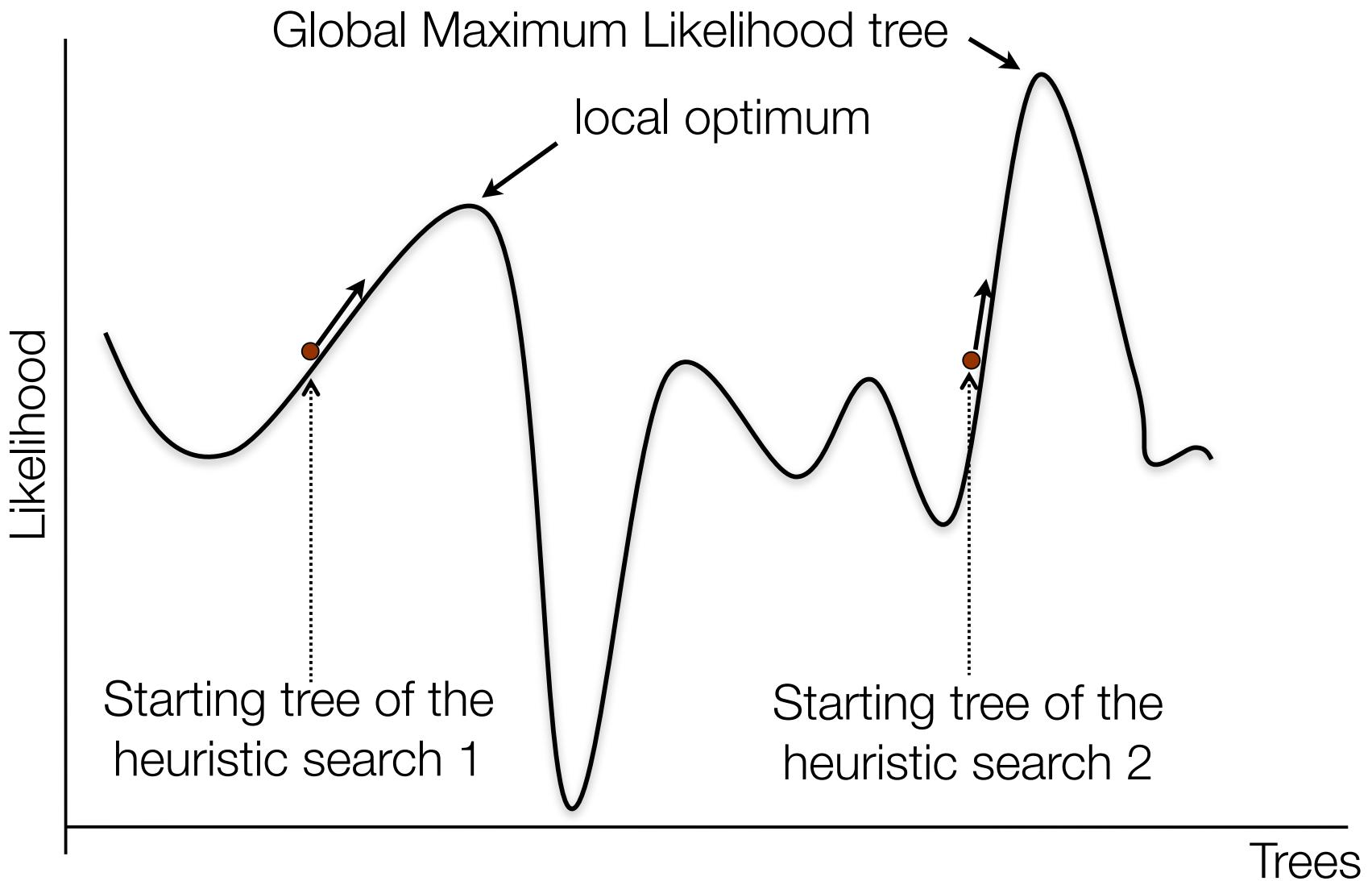
		atpB.phy	Evaluation (3 days left)				
L:	C:	(none) ◊	Unicode (UTF-8) ◊	Unix (LF) ◊	Saved: 5/14/14, 9:18:38 AM	502,603 / 718 / 340	100% ◊
1	338 1467						
2	Oltmannsiellopsis_viridis	AAAAAAAACTGGTAAGTAGTACAAATTATGGCCGTTCTGACTGTAAATTCTAGCGGATCAATGCCAACATTACGGCAAT					
3	Spirodela_polyrhiza	CAAATAAATCCTACTGGTACTCGGTTTCCAATTGGAGAAAAAAACCTGGCGCTTCGCTCAATTATTGCCAGATTGATGCTGTTTCCCCGGTAAAGCCAAATTATTAATGCTTGATG					
4	Pinus_leiophylla_var_chihuahuana	GAAAATTCCTTGTGGCTTCCGGCTGAGAAAGAAATCTGGGACGTATTGCTCAATTCTGGCGGACTTGTGGATCTTCTCCAGGTATATGCTTAAATTATTTACCA					
5	Hesperaloe_parviflora	AAAATCACTCTACTGGCTTCCGGTTCACACTGGAGAAAAAAACCTGGCGCTTCGCTCAATTATTGCCGCGACTTGTGGATCTTCTCCGGCGGAAGATCTGCTTAATTATTAACGGCTTGTGAT					
6	Didierea_magadascariensis	AAAAAAACCTGGGCTTCGCTCAATTCTGGCGCTTCGCTCAATTATTGCCGCGACTTGTGGATCTTCTCCGGCGGAAGATCTGCTTAATTATTAACGGCTTGTGAT					
7	Isoetes_flaccida	AAAAAAACCTGGGCTTCGCTCAATTCTGGCGCTTCGCTCAATTATTGCCGCGACTTGTGGATCTTCTCCGGCGGAAGATCTGCTTAATTATTAACGGCTTGTGAT					
8	Ostreococcus_tauri	CAAGACATTTGGTGCATCAATTCTGGCGCTTCGCTCAATTATTGCCGCGACTTGTGGATCTTCTCCGGCGGAAGATCTGCTTAATTATTAACGGCTTGTGAT					
9	Eucalyptus_grandis	AGATAATCCTACTCTGGCTCTGGAGATTCCACACTGGAAAAAAACCTGGGACTTCTCAATTCTGGCGCTTCGCTCAATTATTGCCGCGACTTGTGGATCTTCTCCGGCGGAAGATCTGCTTAATTATTAACGGCTTGTGAT					
10	Gossypium_hirsutum	AAAATAATCCTACTCTGGCTCTGGAGATTCCACACTGGAAAAAAACCTGGGACTTCTCAATTCTGGCGCTTCGCTCAATTATTGCCGCGACTTGTGGATCTTCTCCGGCGGAAGATCTGCTTAATTATTAACGGCTTGTGAT					
11	Aethionema_cordifolium	AGATAATACTTCACTTCACTGGGGTTCAACACTGGTCTGCAATTCTGGGACTTCTCAATTCTGGCGCTTCGCTCAATTATTGCCGCGACTTGTGGATCTTCTCCGGCGGAAGATCTGCTTAATTATTAACGGCTTGTGAT					
12	Ptilidium_pulcherrimum	AGAGAAATTCCTACTTCTGGGGTCTGGGGCTTCAACACTGGTCTGCAATTCTGGGACTTCTCAATTCTGGCGCTTCGCTCAATTATTGCCGCGACTTGTGGATCTTCTCCGGCGGAAGATCTGCTTAATTATTAACGGCTTGTGAT					
13	Ficus_sp	AGAGAAATTCCTACTTCTGGGGTCTGGGGCTTCAACACTGGTCTGCAATTCTGGGACTTCTCAATTCTGGCGCTTCGATCAATTCTGGGACTTCTGGGGGAAGATCTGCTTAATTATTAACGGCTTGTGAT					
14	Lotus_japonicus	ATTAATAATCCTACCGGCTCAAGGTTCTGCTTCAAAAGAAAATACCTGGGCGCTTGTGGCCAAATTATGGCCGCTGATGTTACTTCTCCGGGAAAGATCTGCTTAATTACAGCTCTTAATGTTAAA					
15	Acorus_calamus	AGAGAAATTCCTACTTCTGGGTTCTGGGGCTTCAACTGGTCAAGGAAAATACCTGGGCGCTTGTGGCCAAATTCTGGCCGCAACTTGTGGCTTCTGGGGTAAAGG					
16	Gunnera_manicata	AGAAATTAATCCTACTTCTGGGGTCTGGGGCTTCAACACTGGTCAAGGAAAATACCTGGGCGCTTGTGGCTCAATTCTGGGCGGAAAGTCTGCTTAATTATTAACGGCTTGTGAT					
17	Marchantia polymorpha	AAAACAATTTTTCTTTGGTGTACGTTCTGGGGCTTCAACTGGTCAAGGAAAATACCTGGGCGCTTGTGGCTCAATTCTGGGCGGAAAGTCTGCTTAATTATTAACGGCTTGTGAT					
18	Chaetosphaeridium_globosum	ATAAAATCTCTGGTGTACCTGGGGCTTCAACTGGGCGCTTCAACTGGTCAAGGAAAATACCTGGGCGCTTGTGGCTCAATTCTGGGCGGAAAGTCTGCTTAATTATTAACGGCTTGTGAT					
19	Hevea_brasiliensis	AGATAATTCCTACTGGCTGGGGATCTGGGGCTTCAACTGGGCGCTTCAACTGGTCAAGGAAAATACCTGGGCGCTTGTGGCTCAATTCTGGGCGGAAAGTCTGCTTAATTATTAACGGCTTGTGAT					
20	Chtoranthus_spicatus	AGAGAAATTCCTACTGGCTGGGGTTCACACTGGTCAAGGAAAATCTGGGGCTTCACTGGTCAATTCTGGGGCTTGTGGCTCAATTCTGGGGCTTGTGGCTCAATTATTAACGGCTTGTGAT					

- Either nucleotide or amino acids; specify the appropriate model
 - Missing data/gaps are treated as N's
 - Should specify an outgroup, but will run without it
 - If you specify multiple, but they are not monophyletic, the first will be used

Once you hit go, what happens?

- RAxML generates starting trees using stepwise addition order parsimony
 - Taxa are inserted into the tree one at a time
- After all taxa added, subtree pruning re-grafting (SPR)
 - Similar to TBR as done in TNT
- Other programs us a Neighbor Joining distance tree as the starting tree
 - Parsimony subtrees seen as an advantage because each start is a distinct and different starting point, which helps search tree space
 - Why is this good?

Tree space

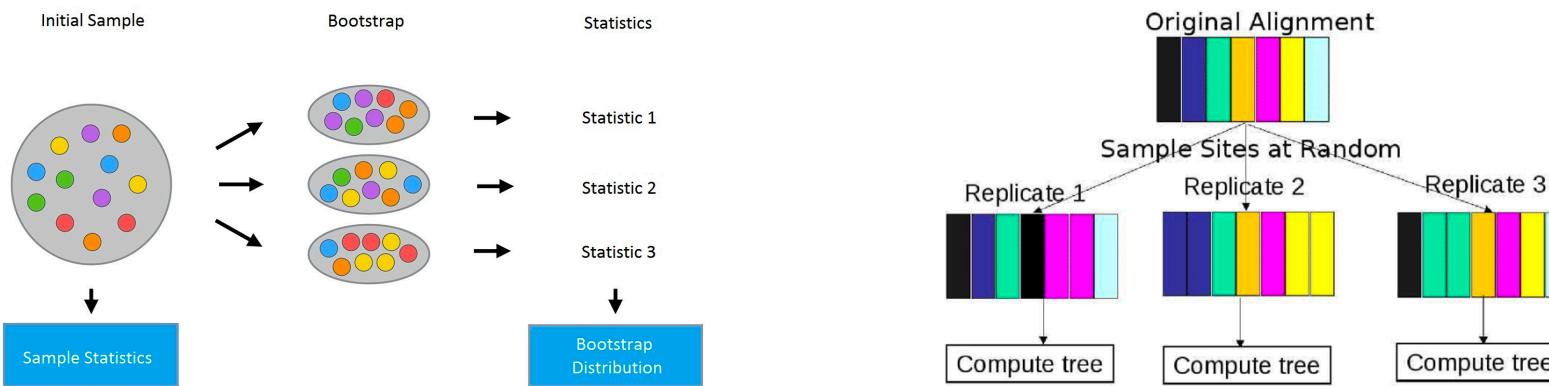


Scope of the question and data

- RAxML was designed for large data sets
 - Reason the models of molecular evolution are constrained to GTR
- For broad taxonomic questions
 - chloroplast genes (slower evolving than nuclear)
 - Amino acid/protein alignments – avoid the third position wobble
- Questions below the species level
 - Can use SNP data from RAD-Seq or Genome Resequencing
 - Models of molecular evolution are different, no invariant sites

How do we tell support for a topology?

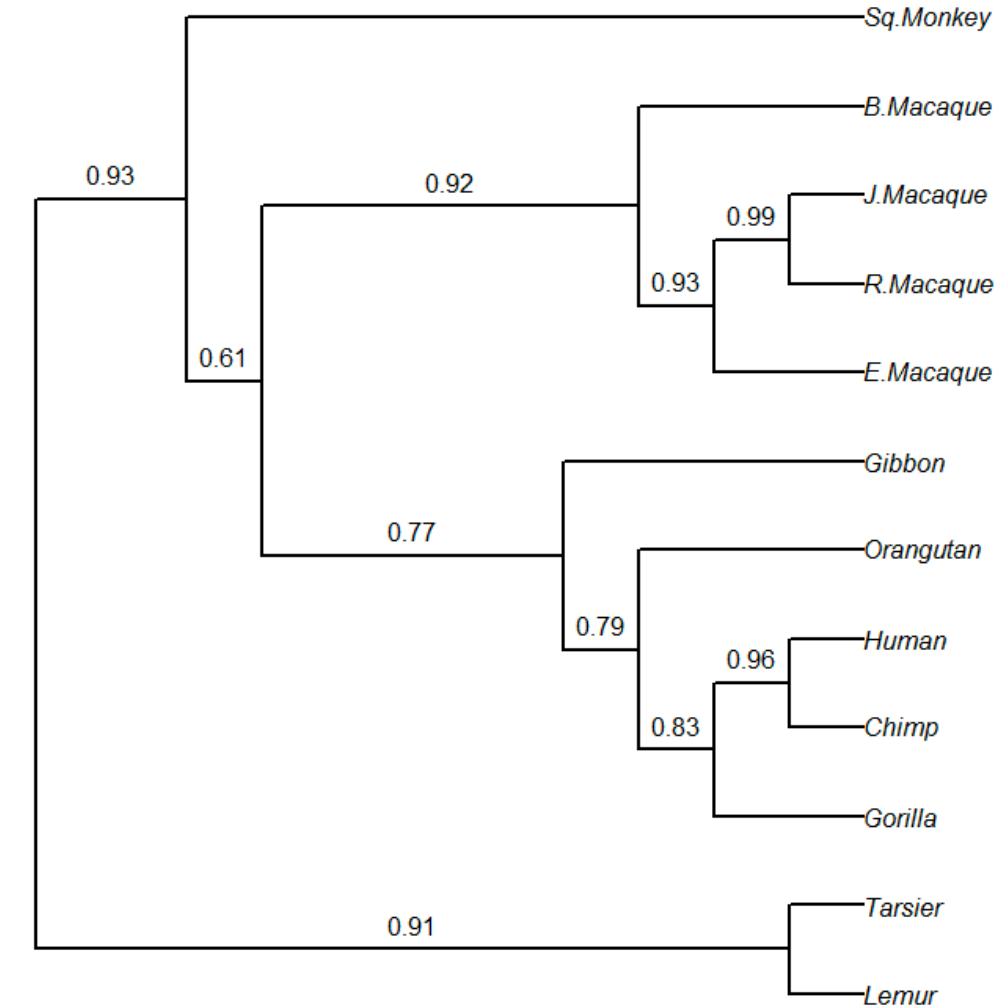
- Most analyses use bootstrap support



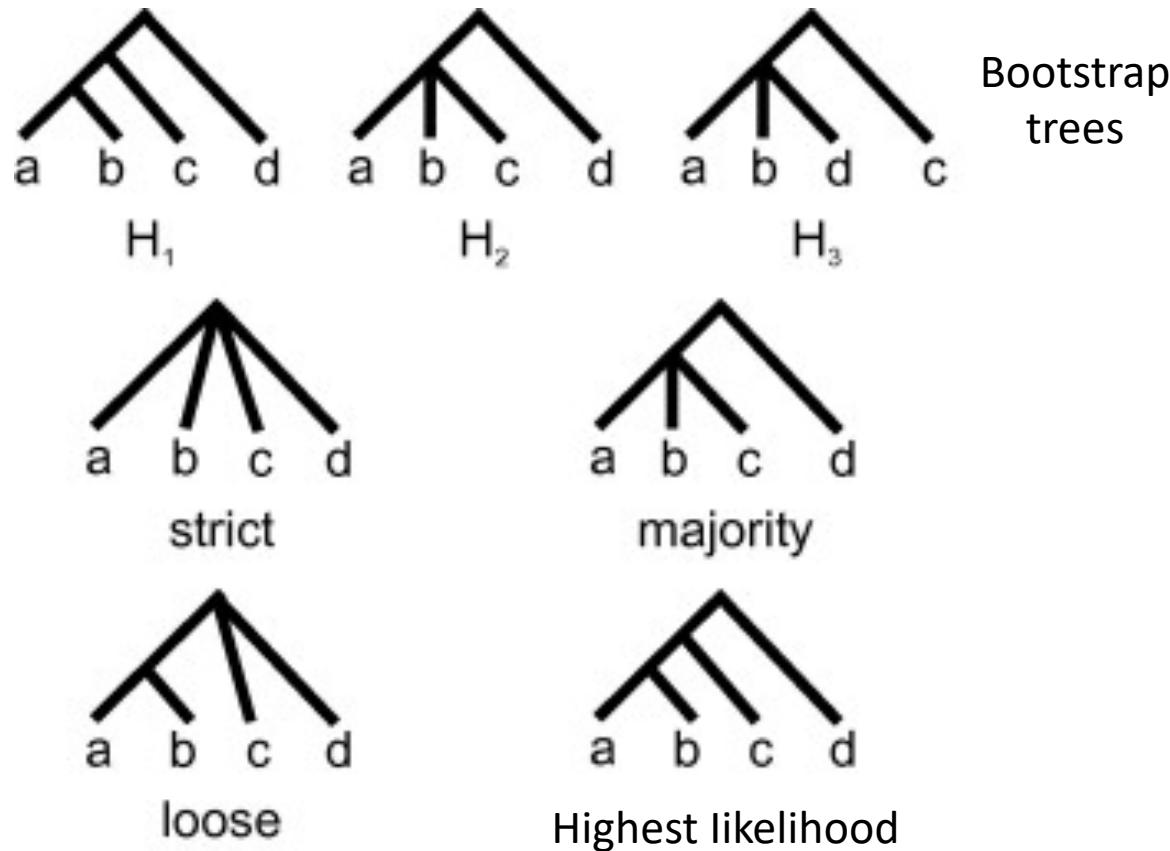
- In our case, we will sample with replacement the original alignment to generate 1000 alignments of the same length
- How often are the same taxa related to each other out of the 1000 replicates
- Usually a value of $BS > 70$ or more is considered strong support

What do we consider good support?

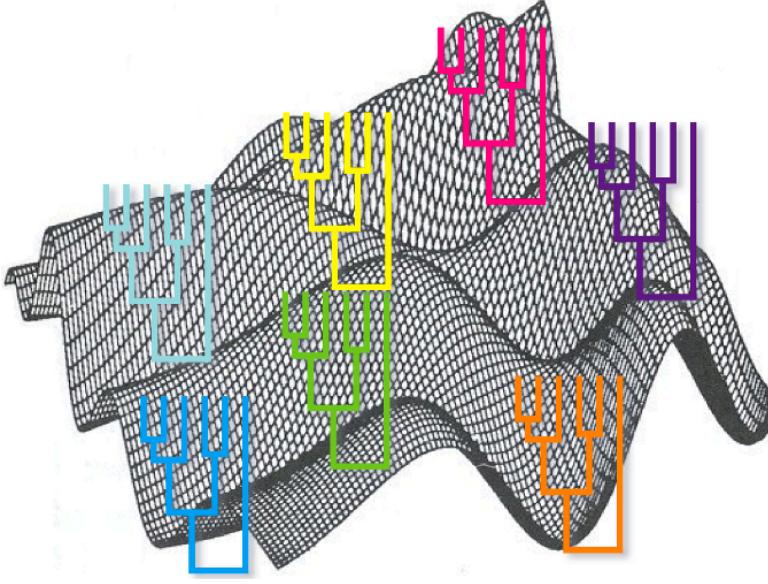
- If there is strong support for relationships in the original data, it will come through in the bootstrap support



Support for topology



- If topology is different between analyses, need to look at how well-supported relationships are
- The “best” tree may not be well-supported at
- Polytomies?



Bayesian approaches

Can we apply prior knowledge to inform our analyses?

ML vs Bayesian

- Maximum likelihood
 - Search for the tree that maximizes the chance of seeing the data
 $(P(\text{Data} \mid \text{Tree}))$
 - Data are random variables but parameters are fixed
- Bayesian
 - Search for the tree that maximizes the chance of seeing the tree given the data $(P(\text{Tree} \mid \text{Data}))$
 - Data are random variables and so are the model parameters
- If priors are well-behaved, then both ML and Bayesian inference converge on the same value (i.e. tree topology)

Expected outcomes

- For any ML analysis, we can expect a tree with the highest likelihood score
- For Bayesian any given run (sampler) may work well in most cases, all runs will fail in some cases, and not guaranteed to work for any particular case
- When do we know the run provides an accurate approximation for a given analysis?
 - We Never do

Bayes Theorem

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

Posterior probability Likelihood Prior probability of A

Prior of B, normalizing constant

The diagram illustrates the components of Bayes' Theorem. At the top, three labels are positioned above the formula: 'Posterior probability' points to $P(A|B)$, 'Likelihood' points to $P(B|A)$, and 'Prior probability of A' points to $P(A)$. Below the formula, an arrow points upwards to the term $P(B)$, which is labeled 'Prior of B, normalizing constant'.

Bayes theorem rewritten

$$P(\text{tree} | \text{data}) = \frac{P(\text{data} | \text{tree}) P(\text{tree})}{P(\text{data})}$$

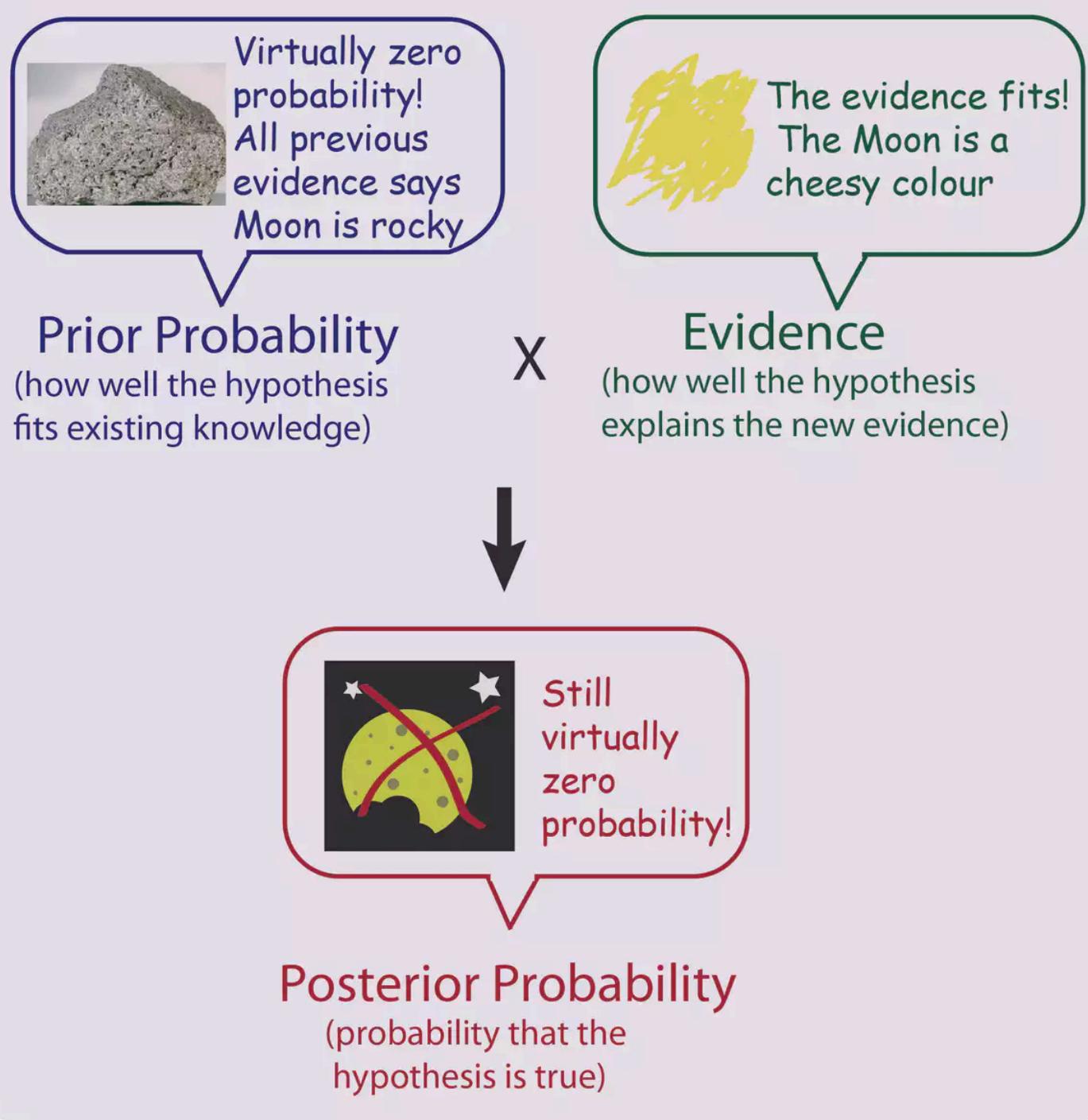
- $P(\text{Data})$ generally not computable
- Parameter space includes
 - Tree topology
 - Branch lengths
 - Substitution model parameters

Example



Hypothesis: The moon is made of cheese

New Evidence: I see the moon is yellow



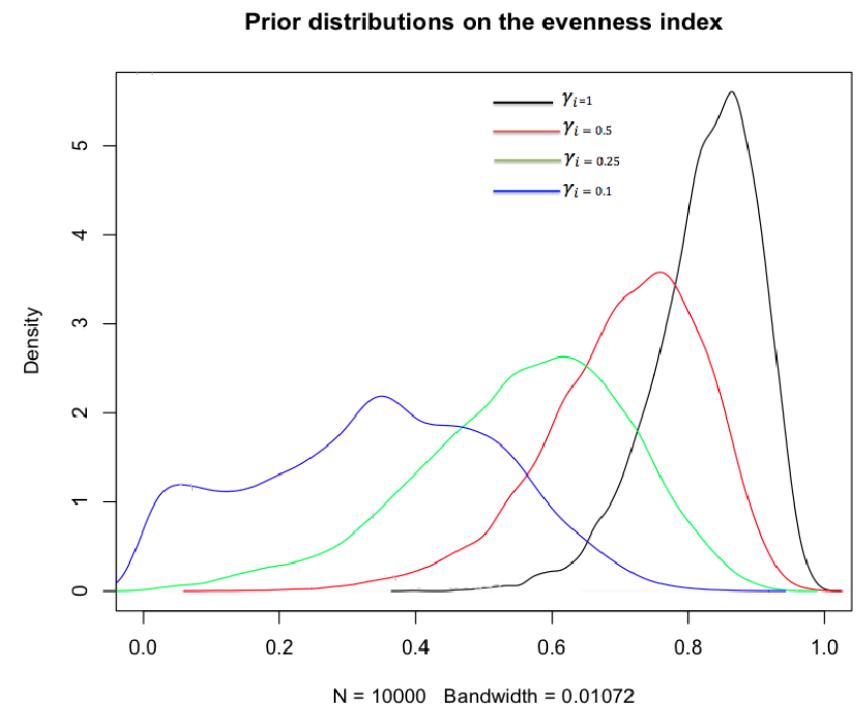
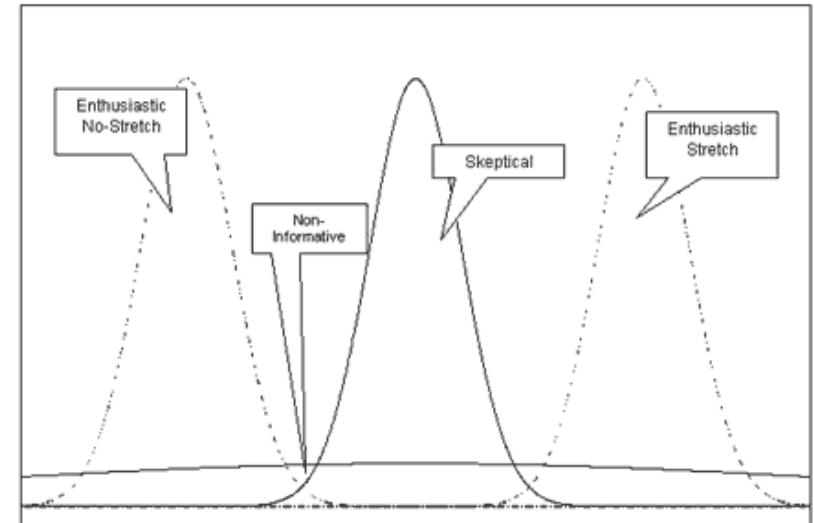
Bayesian inference

<http://theconversation.com/bayes-theorem-the-maths-tool-we-probably-use-every-day-but-what-is-it-76140>

Benefits of stretching before exercise?

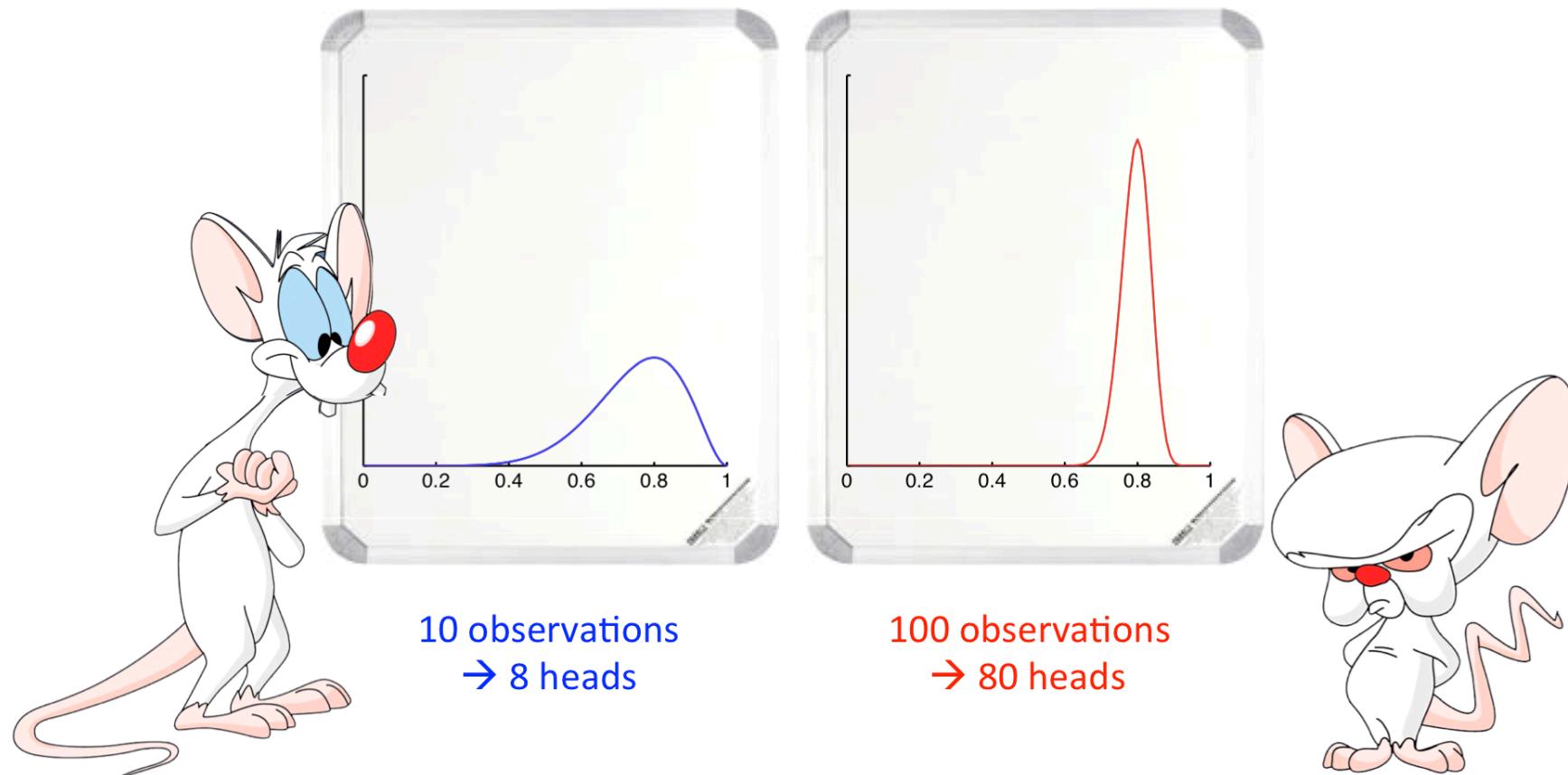
Priors

- Priors can benefit the analysis
 - Incorporate previous information
 - Make the analysis more conservative
- Priors can be detrimental if the data is not very informative
- Probability distributions of parameters
 - Reflect the action of random forces or reflect your uncertainty
 - What are the likely values?

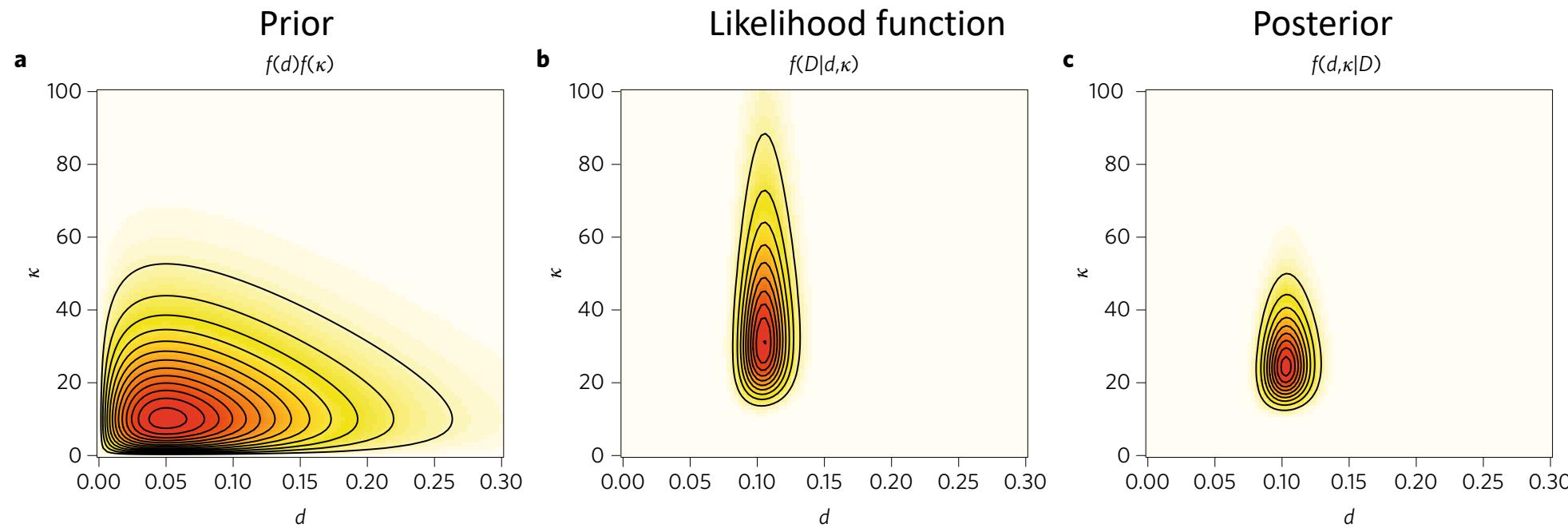


How much “faith” do we put in our priors?

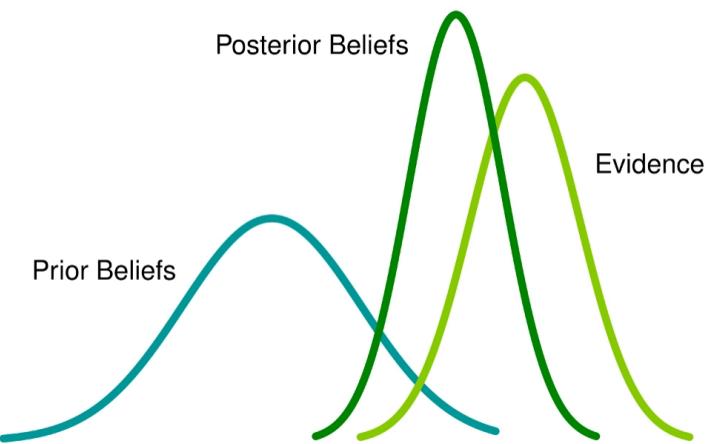
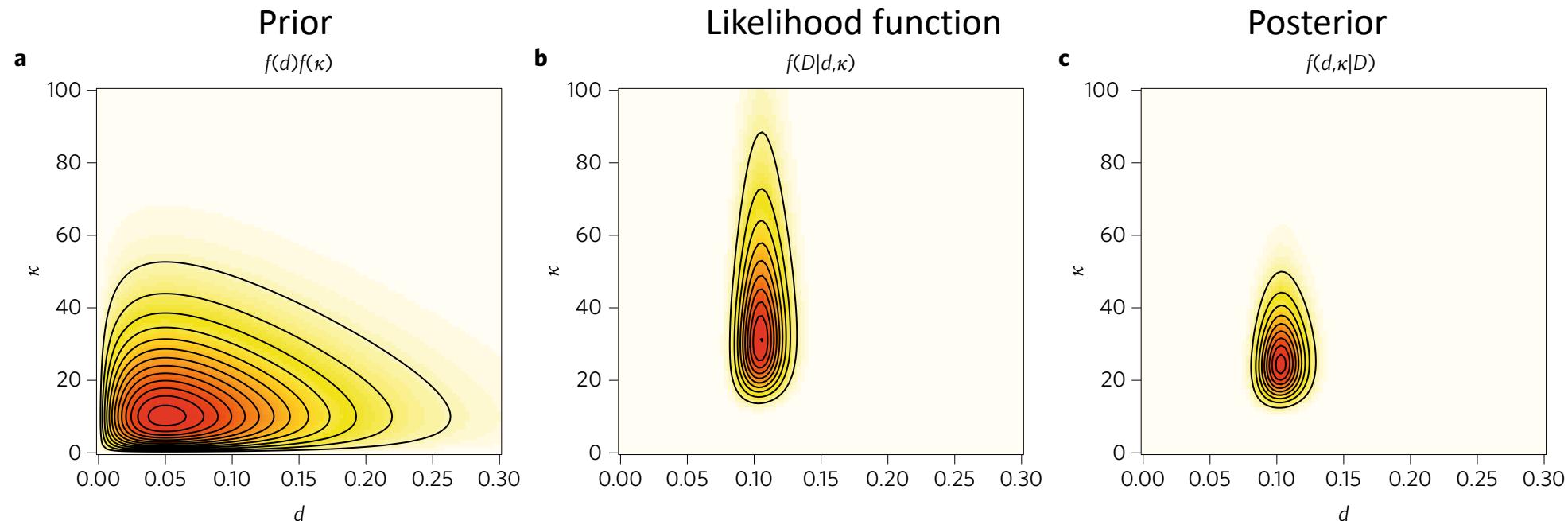
Precision-weighting: who are you going to believe more?



Impact of prior on posterior

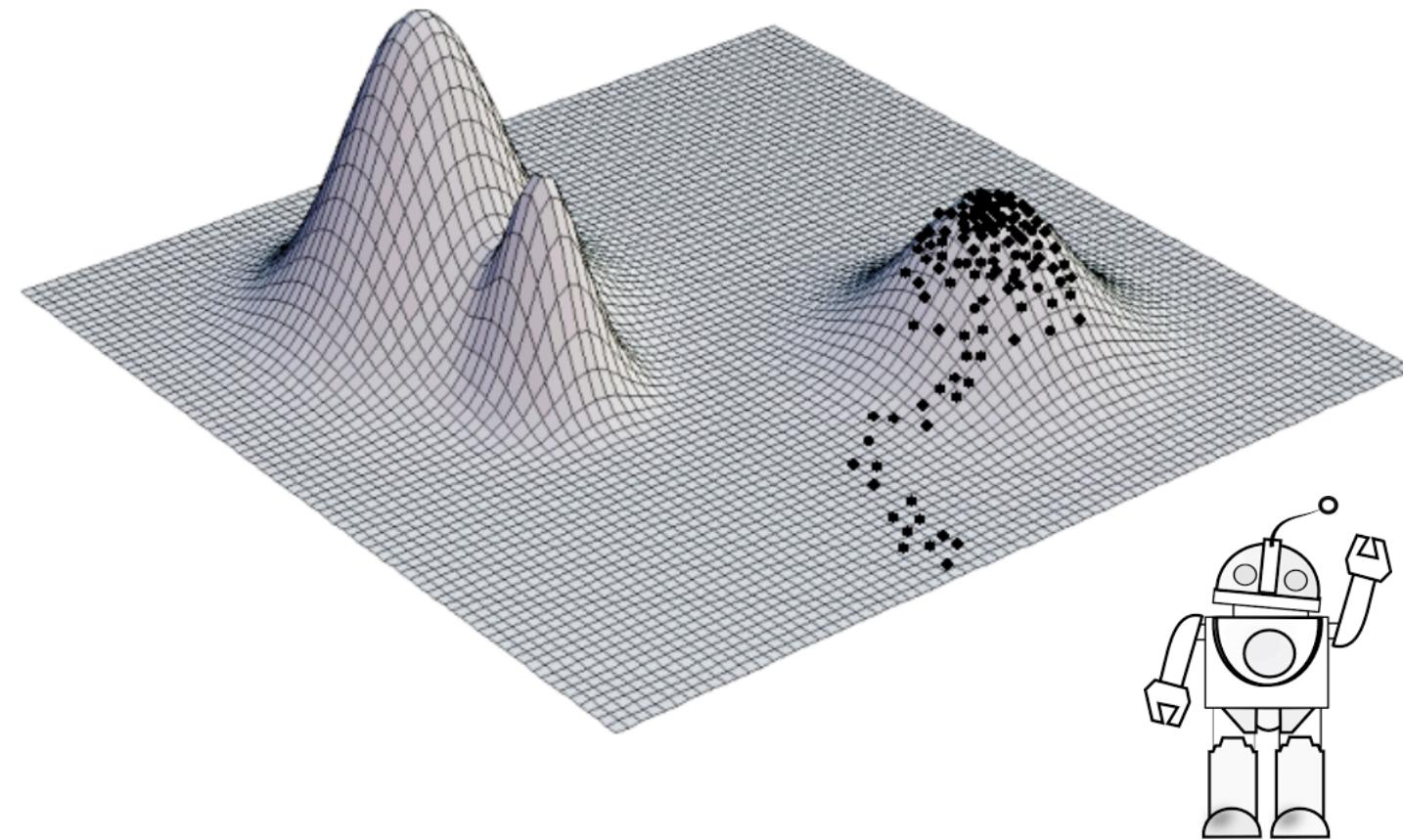


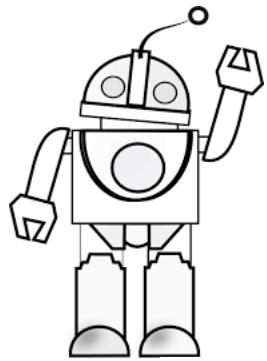
Impact of prior on posterior



Tree space may be complex

Stuck at a local optimum





Markov Chain Monte Carlo (MCMC)



Proposal 1: 2 steps uphill → Accepted

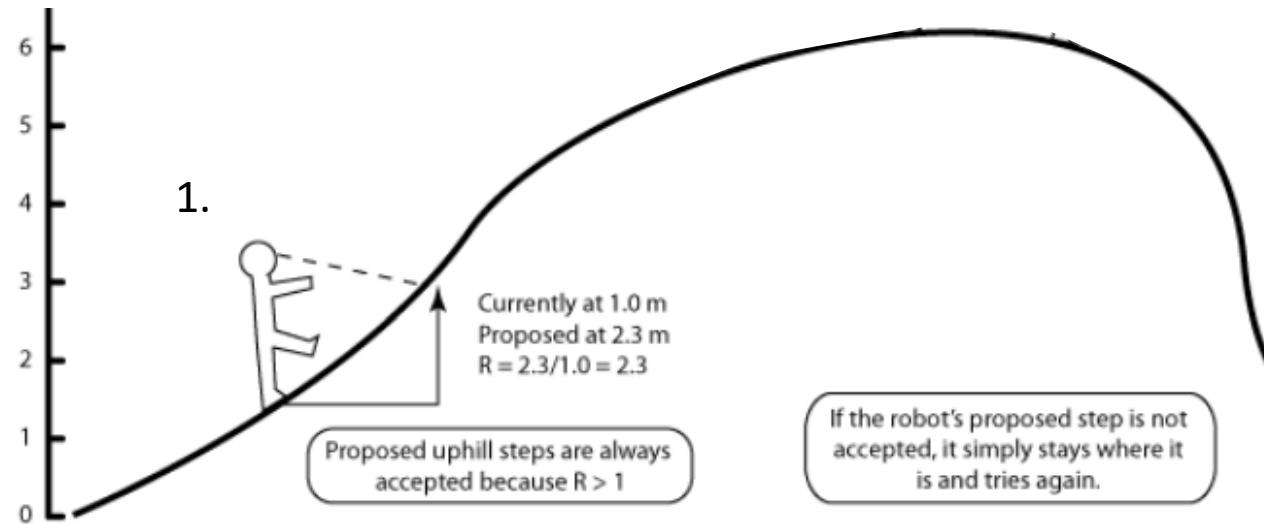
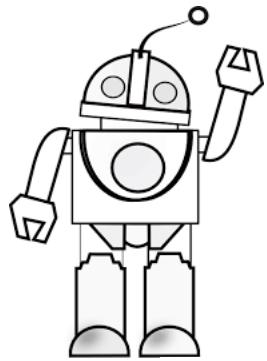


Illustration of MCMC method process (Lewis, 2011)



Markov Chain Monte Carlo (MCMC)



Proposal 1: 2 steps uphill → Accepted

Proposal 2: 1 step downhill → Accepted

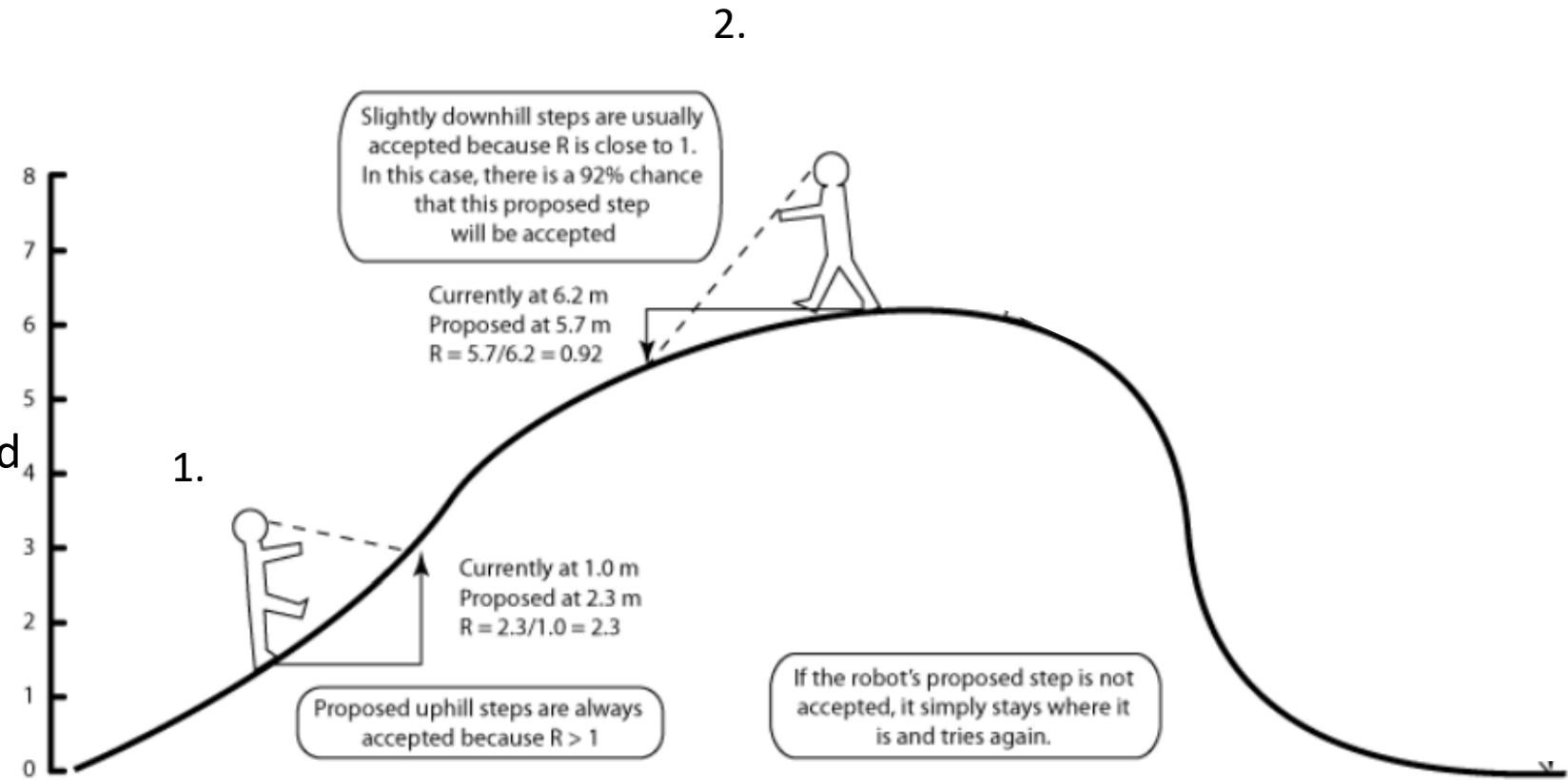
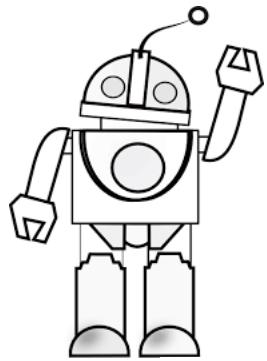


Illustration of MCMC method process (Lewis, 2011)



Markov Chain Monte Carlo (MCMC)



Proposal 1: 2 steps uphill → Accepted

Proposal 2: 1 step downhill → Accepted

Proposal 3: 10 steps downhill → Rejected

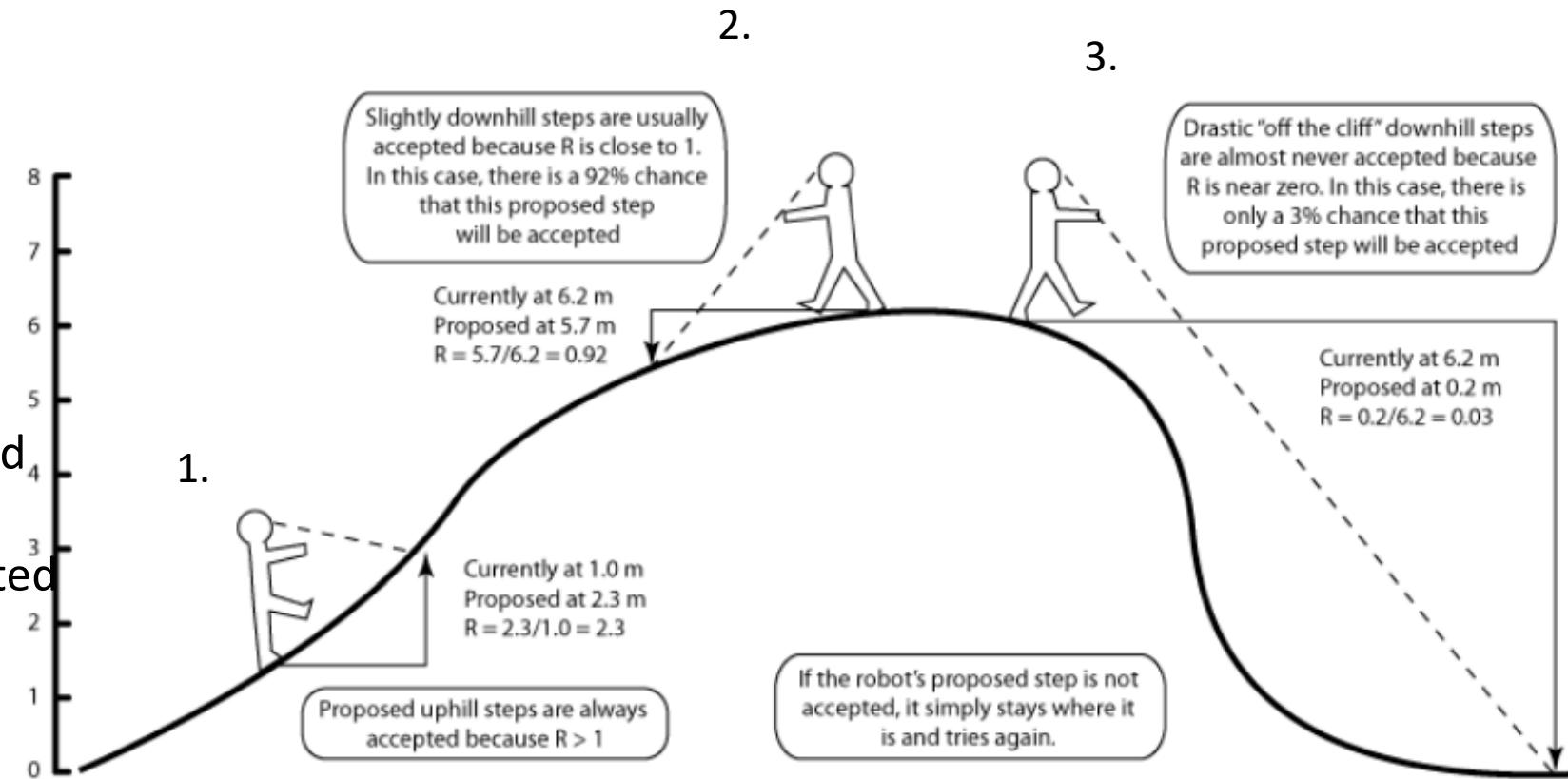
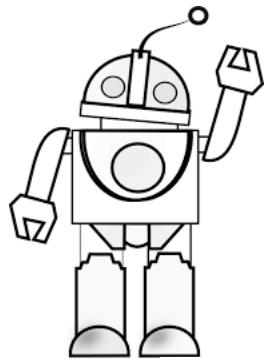


Illustration of MCMC method process (Lewis, 2011)



Markov Chain Monte Carlo (MCMC)



Proposal 1: 2 steps uphill → Accepted

Proposal 2: 1 step downhill → Accepted

Proposal 3: 10 steps downhill → Rejected

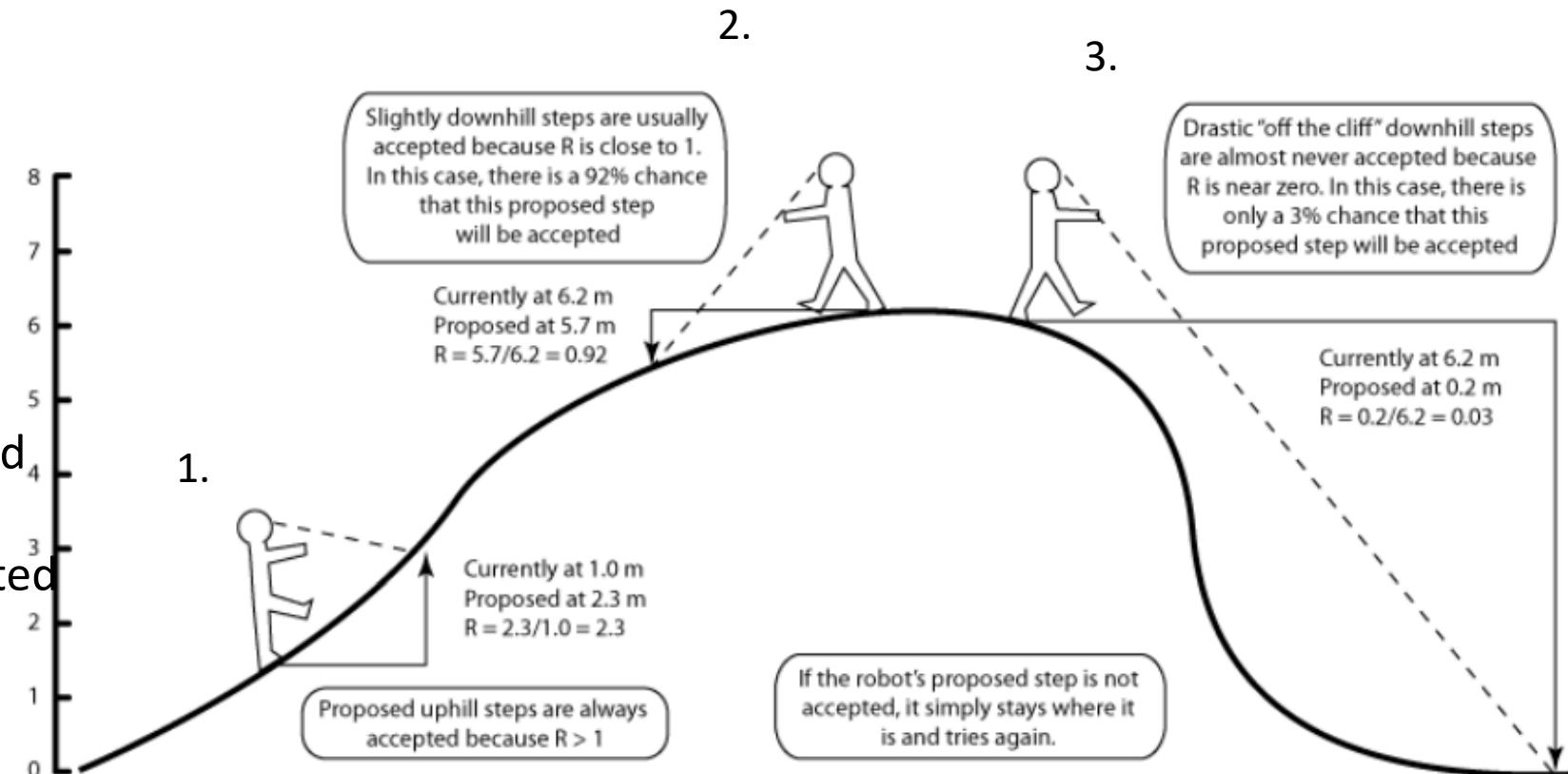
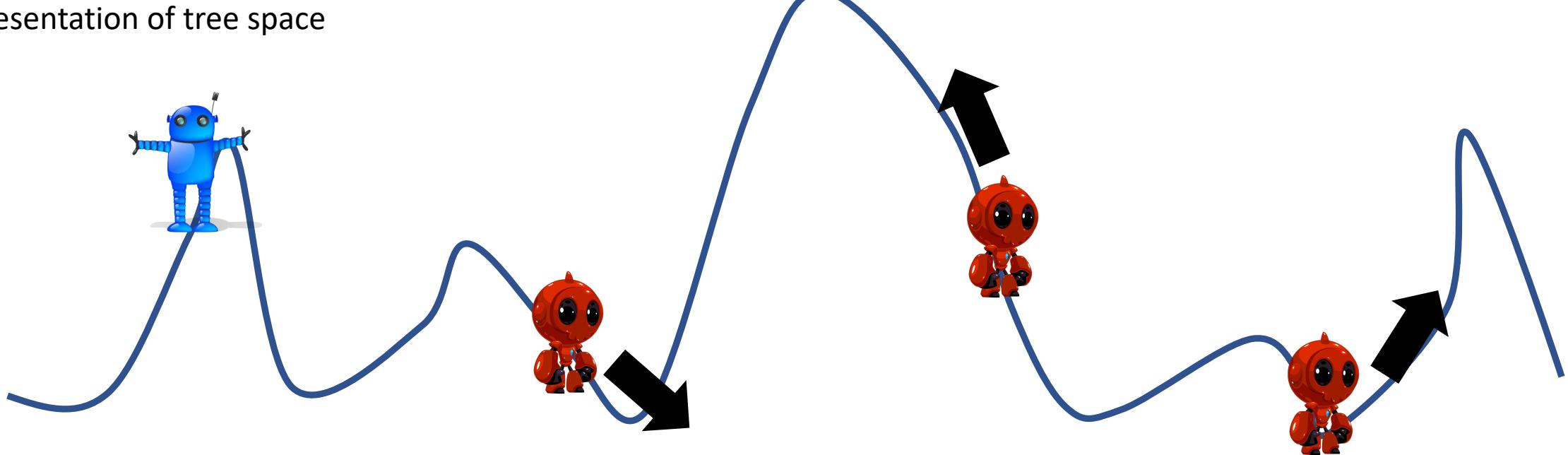


Illustration of MCMC method process (Lewis, 2011)

Markov Chain Monte Carlo (MCMC)

- Four chains (robots) exploring tree space
- 1 cold chain never takes a step with a worse score, 3 heated chains vary in chance of taking a worse step

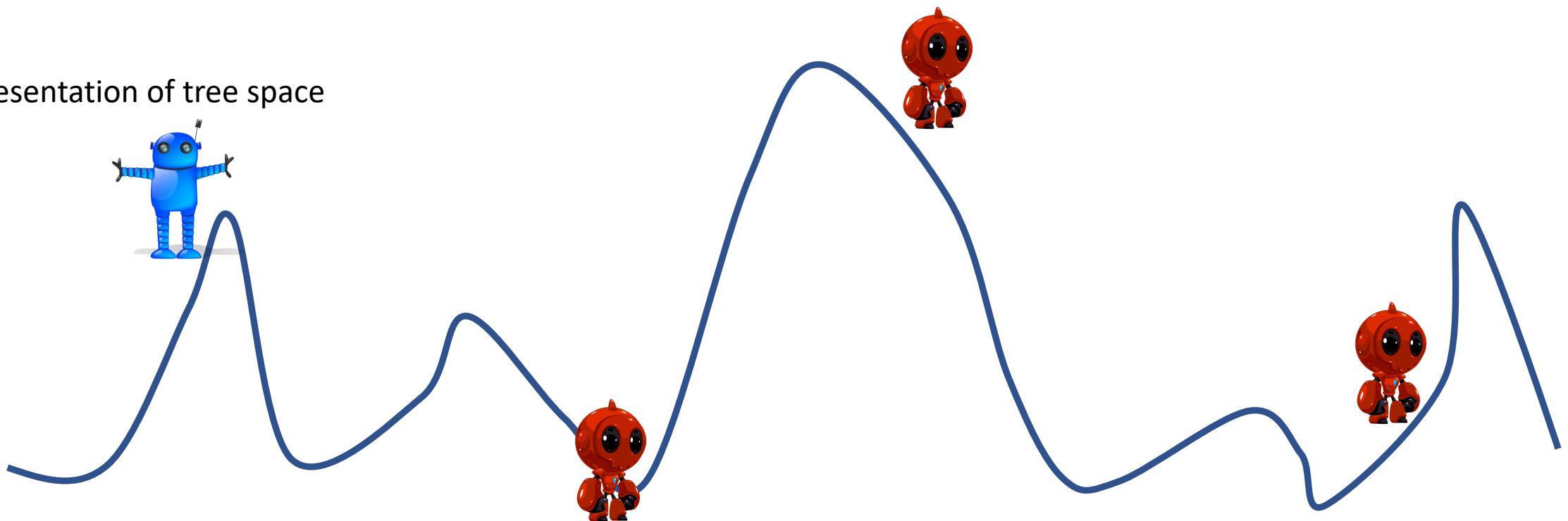
Representation of tree space



Markov Chain Monte Carlo (MCMC)

- Four chains (robots) exploring tree space
- The cold chain is the recorder, and if one of the hot chains finds a higher scoring place, they will switch positions

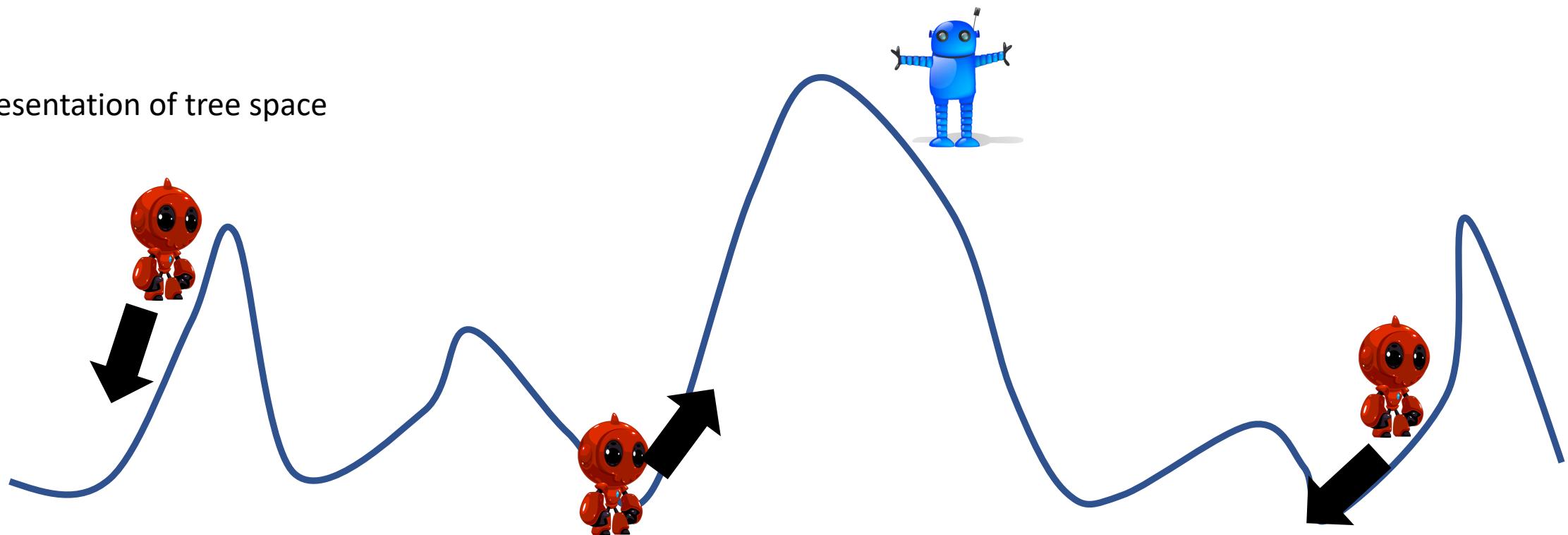
Representation of tree space



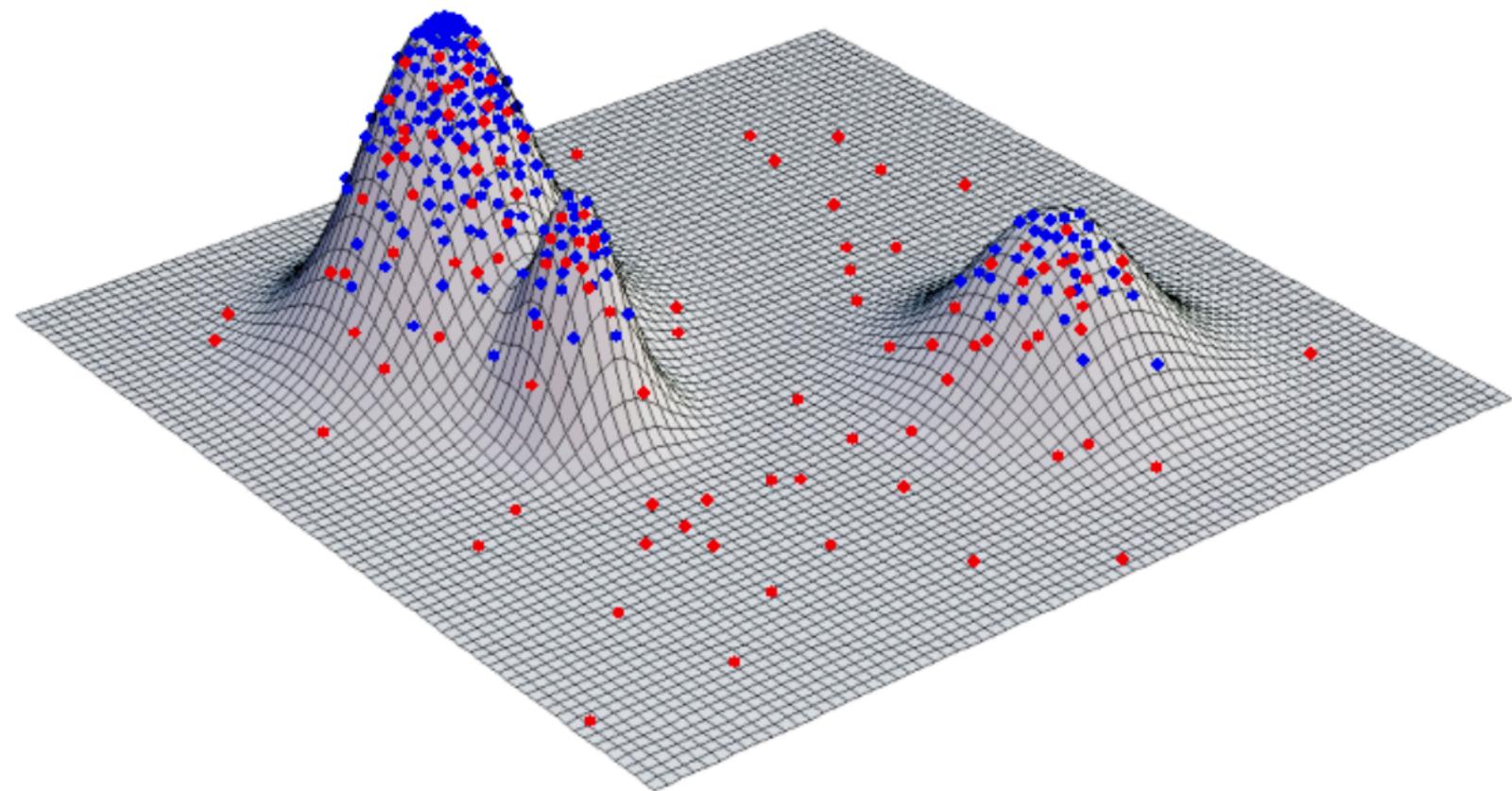
Markov Chain Monte Carlo (MCMC)

- Four chains (robots) exploring tree space
- The cold chain is the recorder, and if one of the hot chains finds a higher scoring place, they will switch positions

Representation of tree space

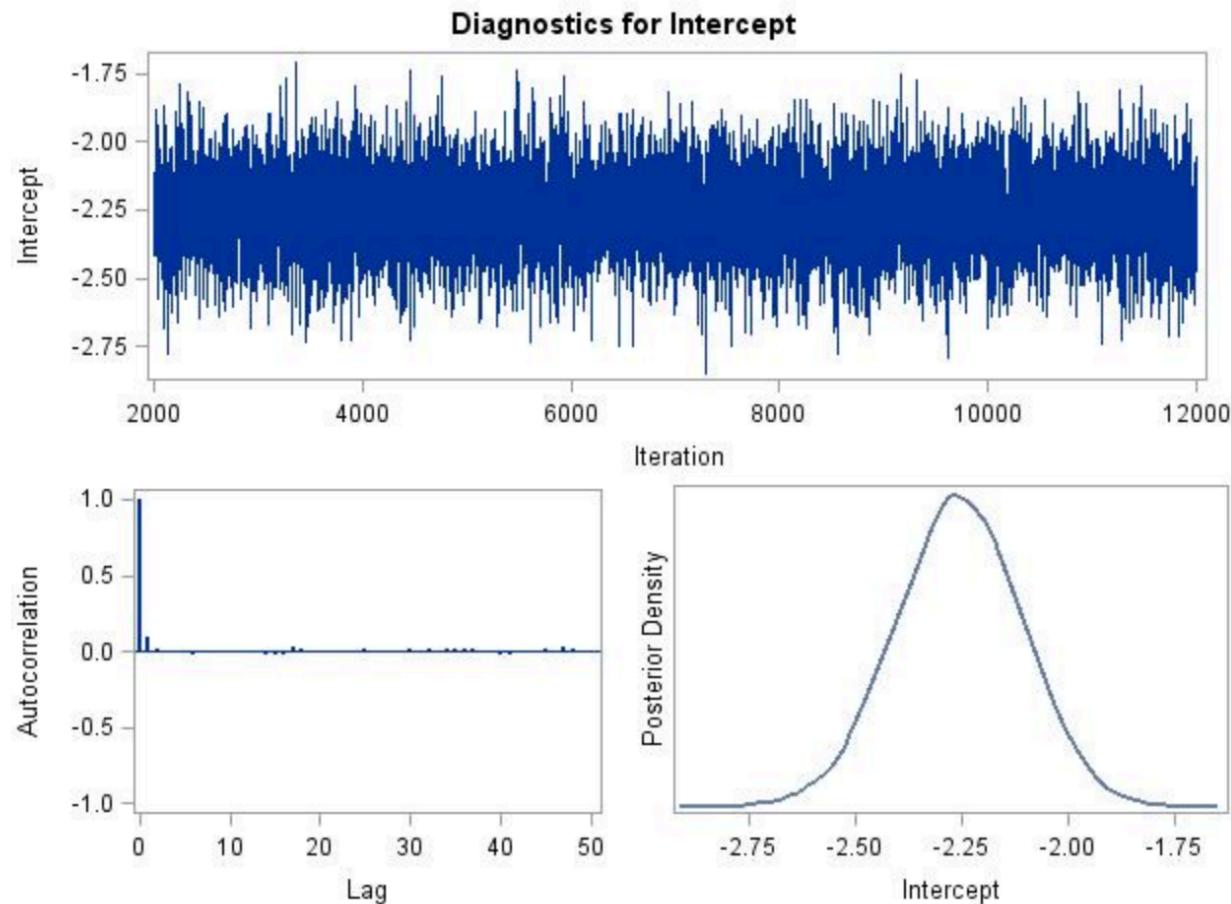


Exploration of tree space



MCMC

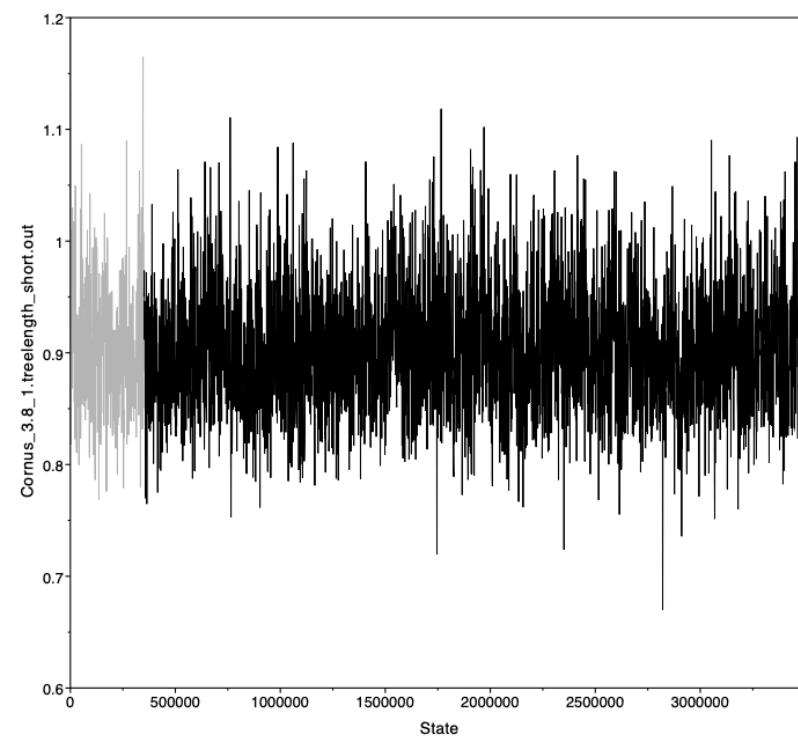
- Yields much larger sample of trees than ML because it produces one tree for every generation versus one tree per tree search
- Trees produced are highly auto-correlated
 - Sample relatively infrequently and discard trees early in the process
 - Millions of generations required compared to 1,000 bootstrap replicates to explore tree space



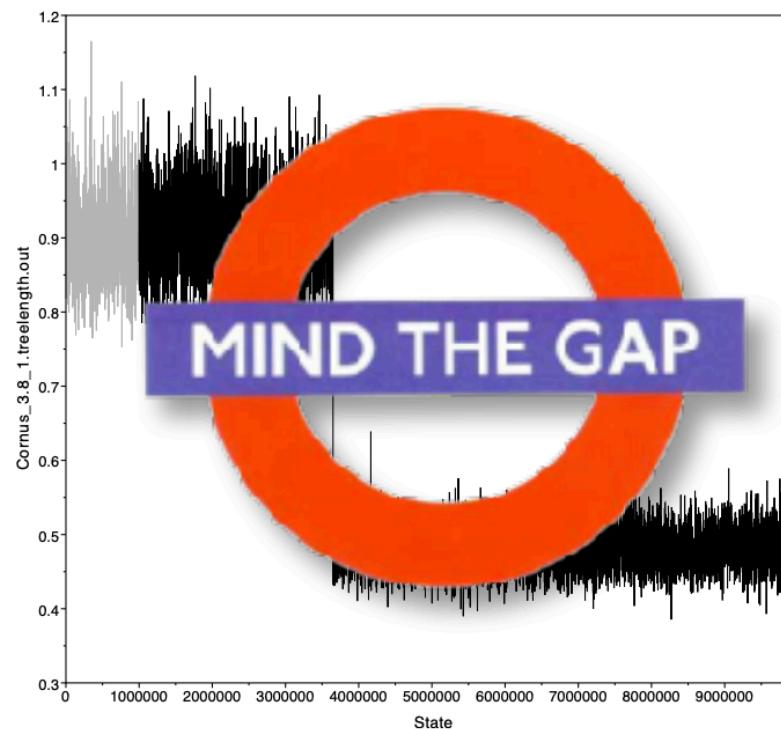
How the run may look

Tracer run after a Bayesian run

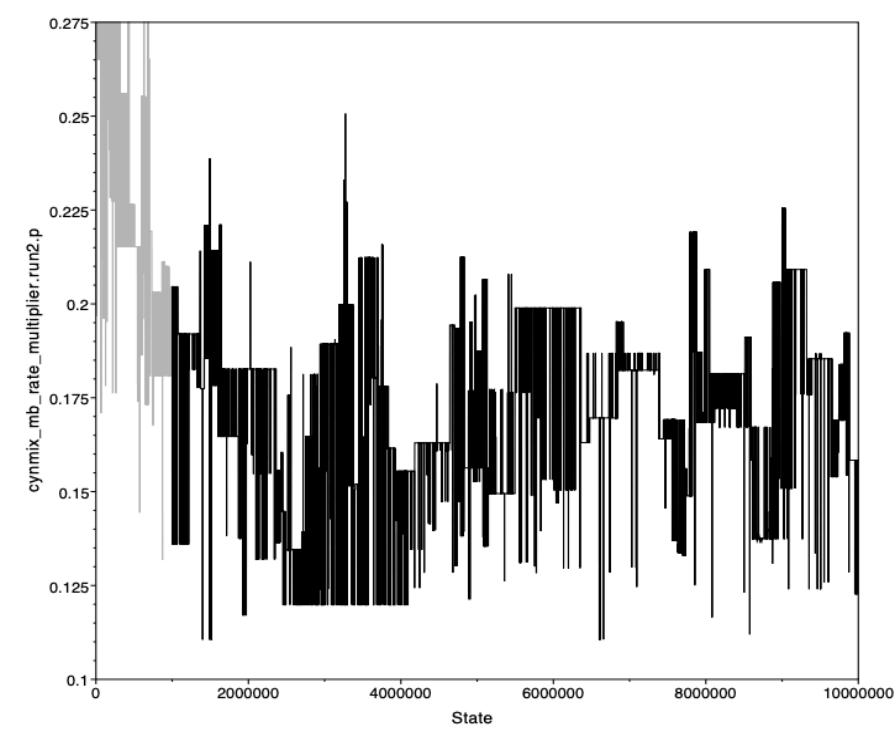
all looks good...



until it doesn't



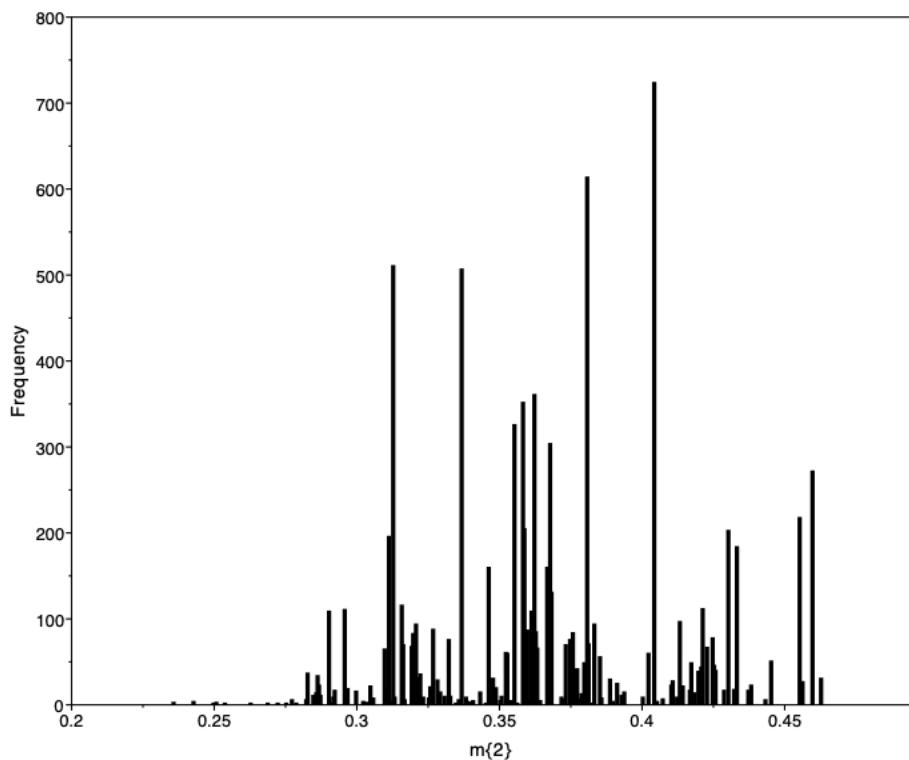
bad mixing



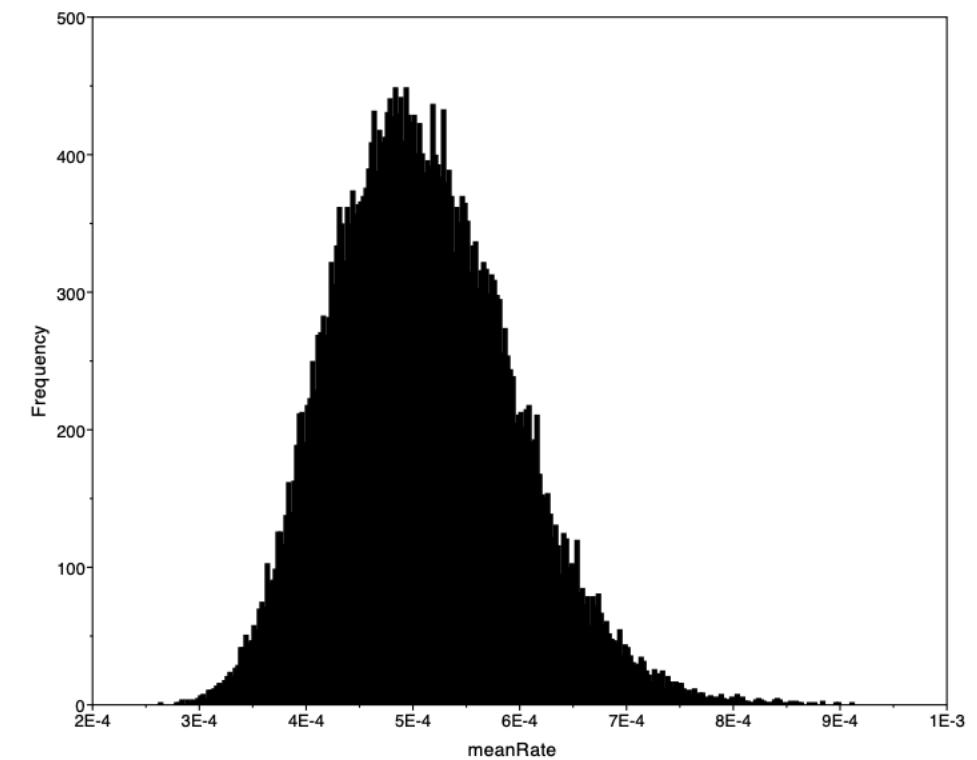
How the run may look

Posterior estimates of parameters

bad mixing



better mixing

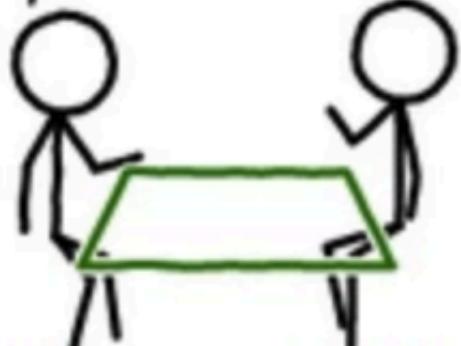


Bayesian humor

BAYESIAN VERSUS FREQUENTIST

I DON'T KNOW WHY WE
STILL HAVE THIS
"DEBATE" EACH YEAR.

WANT TO TOSS
FOR IT?
I KNOW
ITS A BIT PASSÉ
SURE



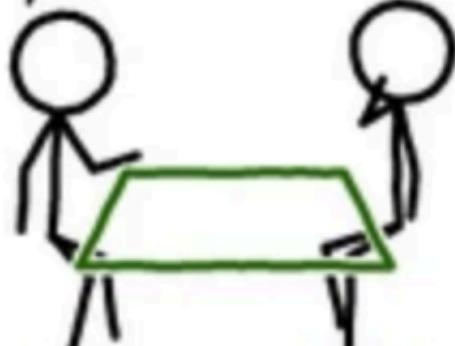
BAYESIAN

FREQUENTIST

MAYBE IT IS
MAYBE IT ISN'T

WAIT A MINUTE -
IS THAT YOUR
BIASED COIN?

HMM TRICKY



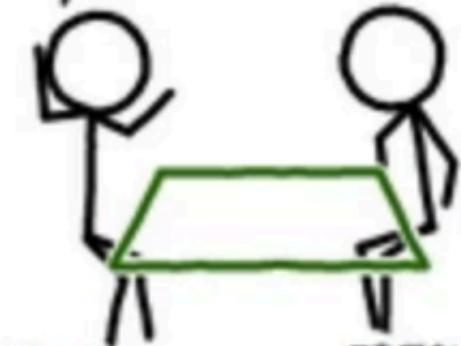
BAYESIAN

FREQUENTIST

WELL OK THEN
TAKING ALL THESE
THINGS INTO
CONSIDERATION . . .
HEADS!

AHA! PRIOR
CONSTRUCTION!
BAYES WINS AGAIN!

DAMN



BAYESIAN

FREQUENTIST

REVBAYES #005