Paper Review

- Reminder to email me your paper selection (finlay.maguire@dal.ca)
  - Due by midnight TODAY

Bayesian Methods

The Story So Far

- Maximum likelihood: Probabilistic approach that aims to find the model parameters that maximize the probability of the data
  - Which combination of tree, branch lengths, and substitution model is best?

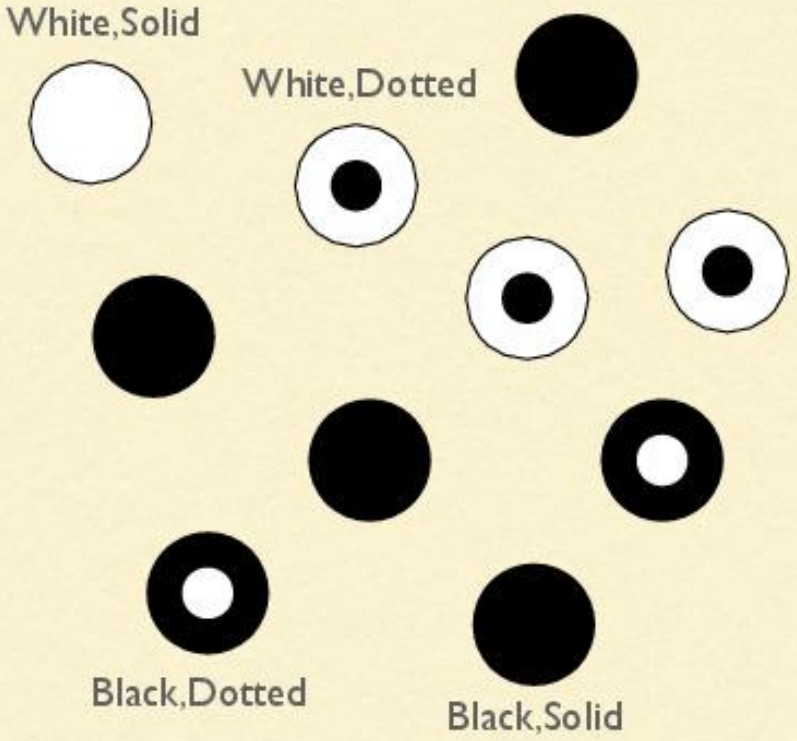
- Limitation: Maximum likelihood seeks out the best answer, but what about other models that are nearly as good?

Bayesian Methods

- What are the relative probabilities of alternative models, given the data?

- Assign weightings to different models based on their likelihood – these are Bayesian posteriors

Joint probabilities



10 marbles in a bag
Sampling with replacement

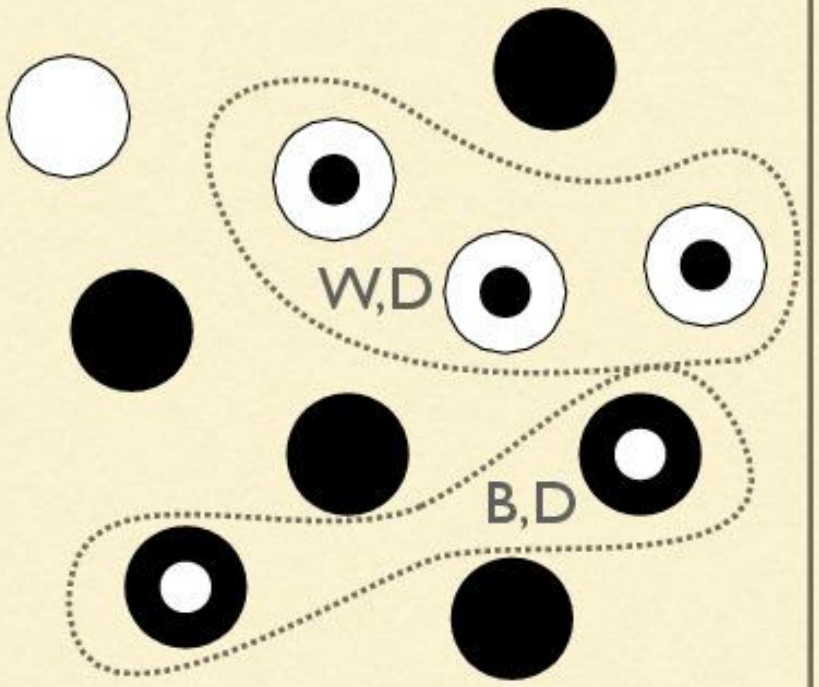
Pr(B,S) = 0.4

Pr(W,S) = 0.1

Pr(B,D) = 0.2

Pr(W,D) = 0.3

Marginal probabilities



Marginalizing over color yields the total probability that a marble is dotted (D)

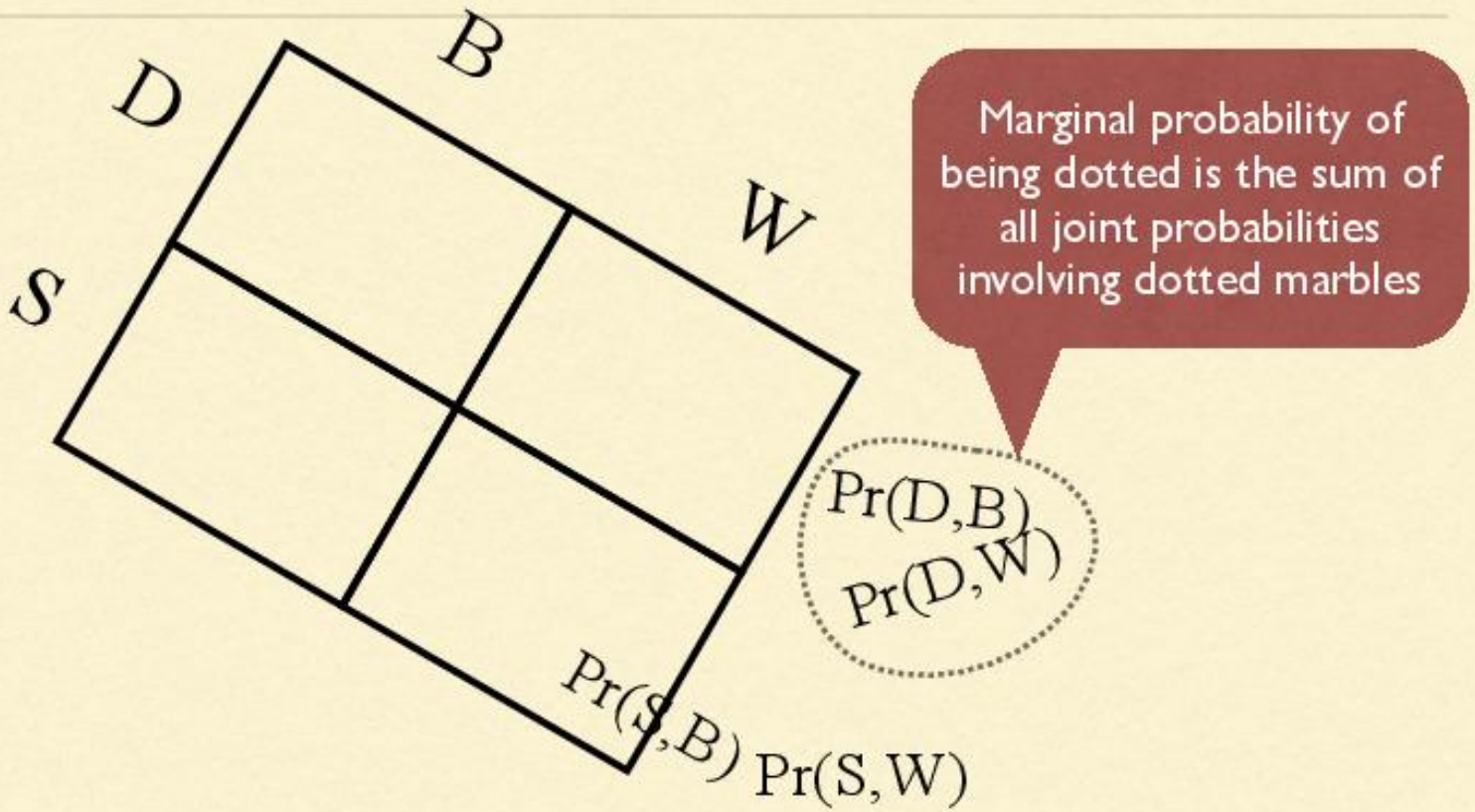
$$\Pr(\mathbf{D}) = \Pr(B,\mathbf{D}) + \Pr(W,\mathbf{D})$$
$$= 0.2 + 0.3$$
$$= 0.5$$

Marginalization involves summing all joint probabilities containing D

Marginalization

	B	W
D	Pr(D,B)	Pr(D,W)
S	Pr(S,B)	Pr(S,W)

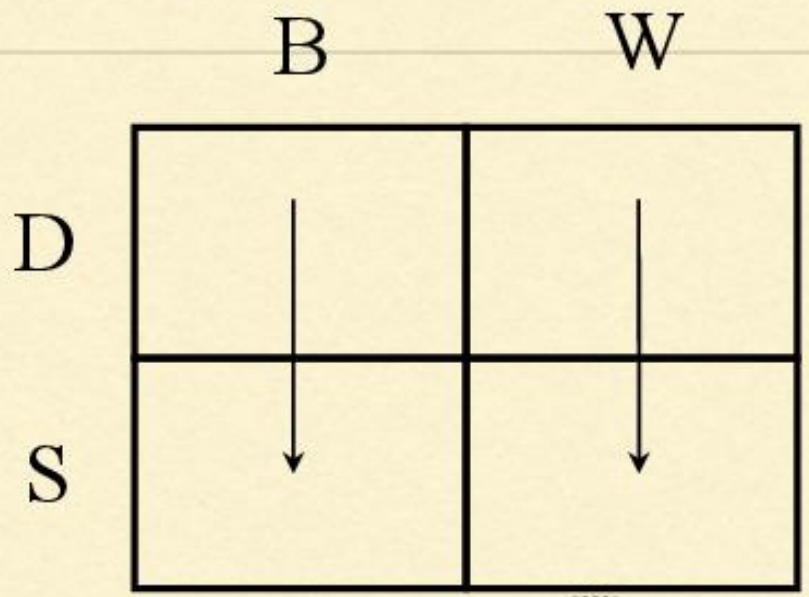
Marginalizing over colors



Marginalization

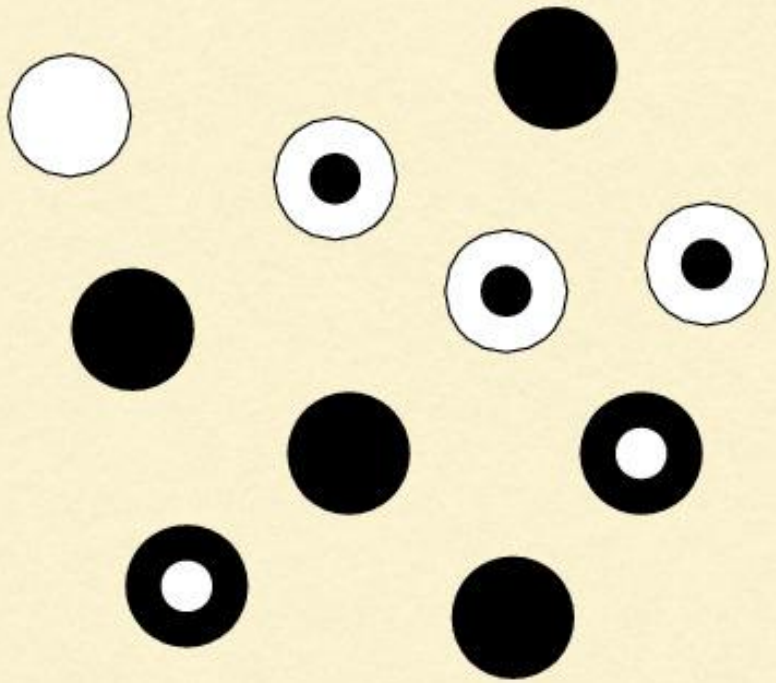
	B	W
D	Pr(D,B)	Pr(D,W)
S	Pr(S,B)	Pr(S,W)

Marginalizing over "dottedness"



Marginal probability of being a white marble

Bayes' rule



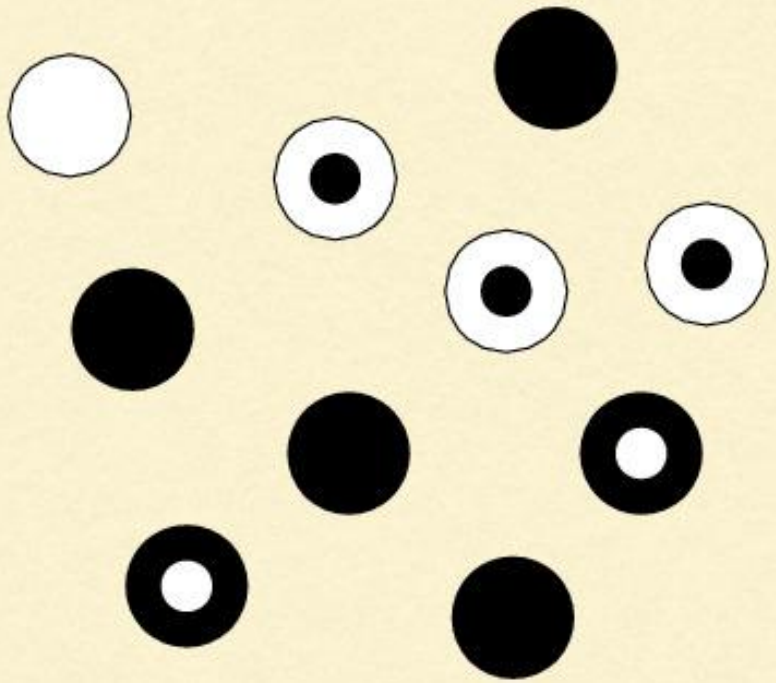
The joint probability Pr(B,D) can be written as the product of a *conditional probability* and the *probability of that condition*

Pr(B,D) → Pr(B|D) Pr(D)

Pr(B,D) → Pr(D|B) Pr(B)

Either B or D can be the condition

Bayes' rule



Equate the two ways of writing Pr(B,D)

$$\Pr(B|D)\,\Pr(D) = \Pr(D|B)\,\Pr(B)$$

Divide both sides by Pr(D)

$$\frac{\Pr(B|D)\,\cancel{\Pr(D)}}{\cancel{\Pr(D)}} = \frac{\Pr(D|B)\,\Pr(B)}{\Pr(D)}$$

Bayes' rule

$$\Pr(B|D) = \frac{\Pr(D|B)\,\Pr(B)}{\Pr(D)}$$

Bayes' rule (variations)

$$\Pr(B|D) = \frac{\Pr(D|B)\Pr(B)}{\Pr(D)}$$

$$= \frac{\Pr(D|B)\Pr(B)}{\Pr(B,D) + \Pr(W,D)}$$

Pr(*D*) is the **marginal probability** of being dotted
To compute it, we **marginalize over colors**

Bayes' rule (variations)

$$\Pr(B|D) = \frac{\Pr(D|B)\Pr(B)}{\Pr(B,D) + \Pr(W,D)}$$

$$= \frac{\Pr(D|B)\Pr(B)}{\Pr(D|B)\Pr(B) + \Pr(D|W)\Pr(W)}$$

$$= \frac{\Pr(D|B)\Pr(B)}{\sum_{\theta \in \{B,W\}} \Pr(D|\theta)\Pr(\theta)}$$

Bayes' rule in statistics

Likelihood of hypothesis θ

Prior probability of hypothesis θ

$$\Pr(\theta|D) = \frac{\Pr(D|\theta)\Pr(\theta)}{\sum_\theta \Pr(D|\theta)\Pr(\theta)}$$

Posterior probability of hypothesis θ

Marginal probability of the data (marginalizing over hypotheses)

Calculating Posterior Probabilities

$$P(M \mid D) = \frac{P(D \mid M) \times P(M)}{P(D)}$$

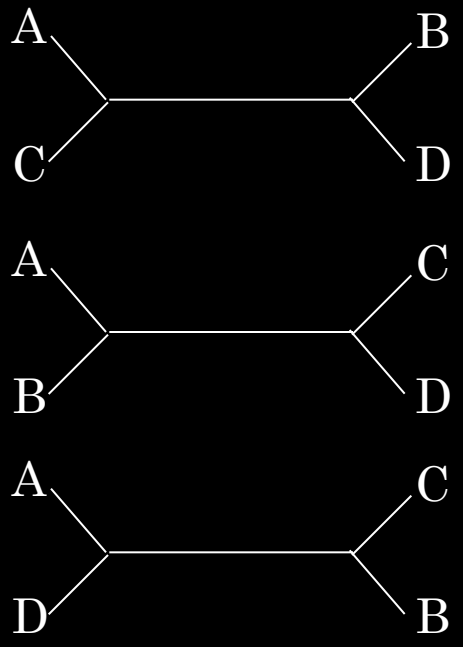
Posterior probability of a model, given the data

Normalized Likelihood

Prior probability

Prior Probability

What is the initial weighting of models?

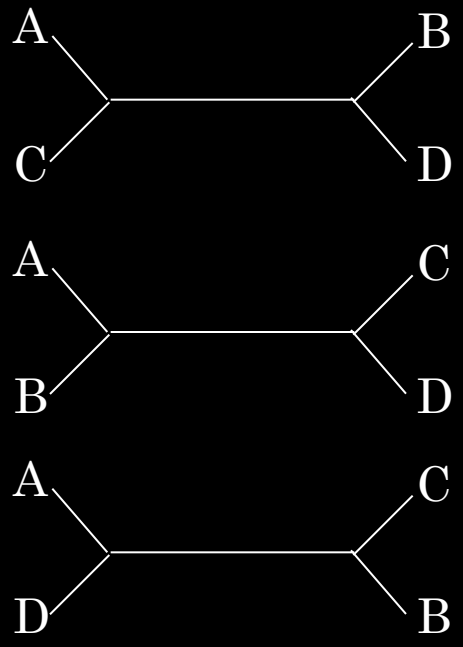


1/3

1/3

1/3

Flat



2/3

1/6

1/6

Informative

Why Bayesian?

All of the advantages of other model-based methods, plus:

(1) Explicit incorporation of prior hypotheses concerning models (e.g., “animals with fungi is much more probable than plants with fungi”)

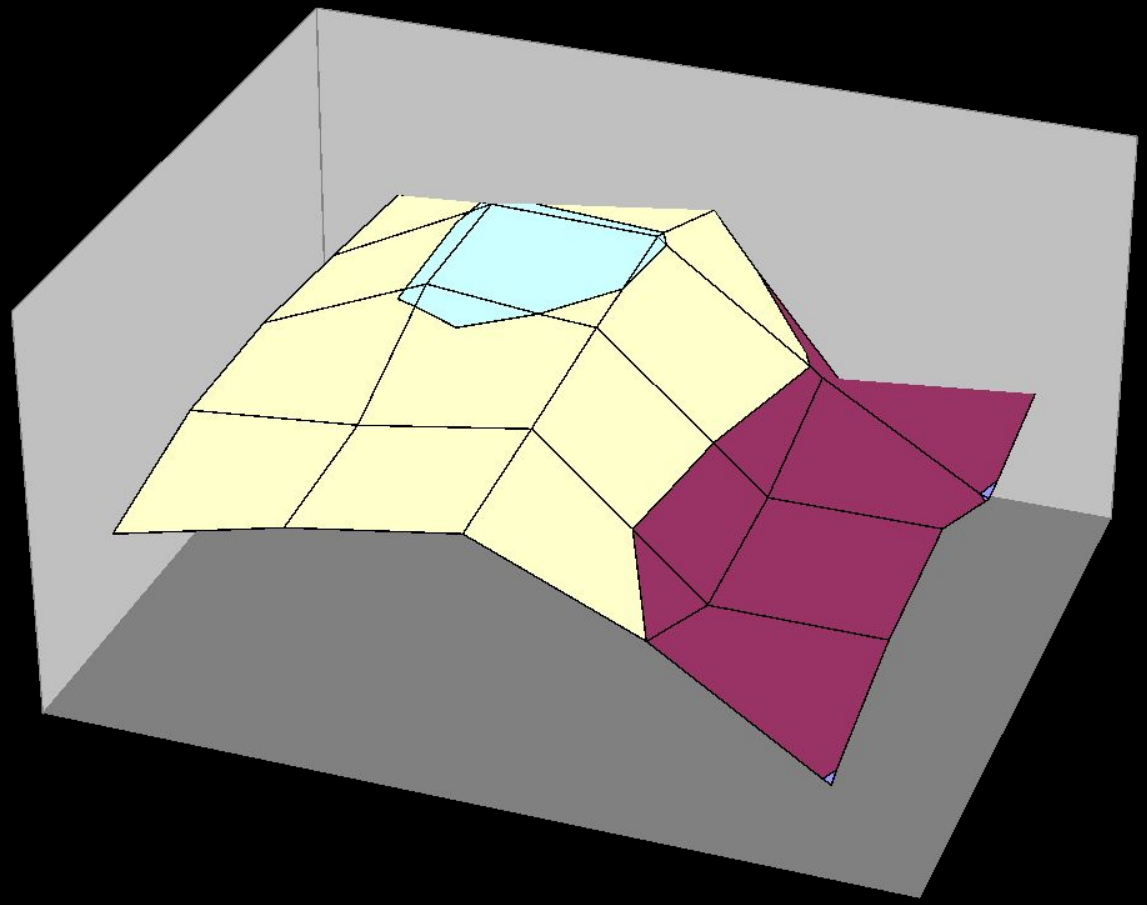
(2) Calculation of posterior probabilities: the relative ‘goodness’ of different models are taken into account

The Likelihood Surface

- For simple distributions (e.g. binomials for coin-flipping), we can analytically integrate over the entire likelihood function

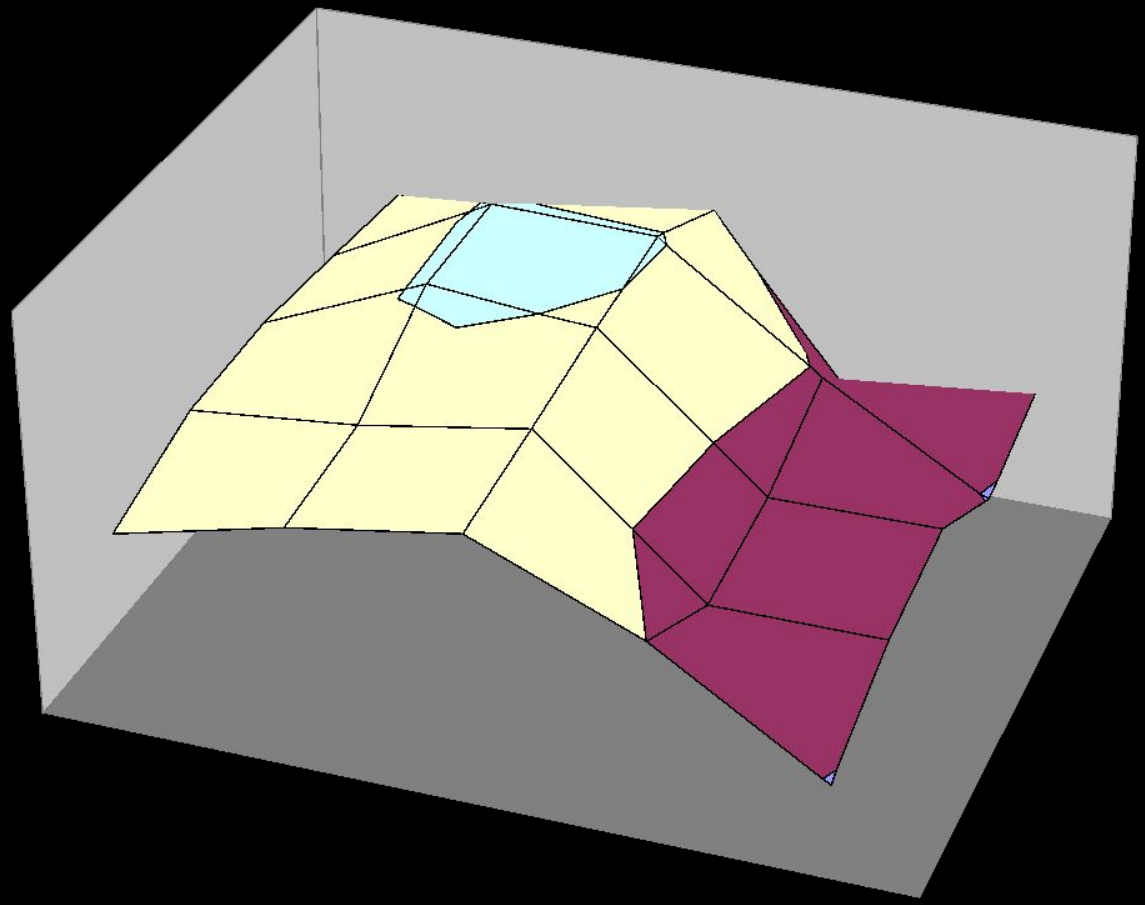
- For complex distributions (e.g. likelihoods for all trees), we cannot do this

- We could visit every point in ‘model space’ and evaluate the likelihood. But model space is colossal!



Iterative Integration: Better living through sampling

- One solution is a random walk through model space
- Randomly proposed steps can be accepted or rejected, with a preference for steps that increase the likelihood
- But we can descend the hill!



Markov chain Monte Carlo

- Markov: Previous steps must not influence future proposals and decisions

- Chain: We remember every model we visit during the sampling process. The series of models we visit is the chain

- Monte Carlo: Moves through model space are proposed at random

Procedure

(1) Start with a random model ψ

(2) Propose a change to a new model ψ'

(3) Accept the change from ψ to ψ' with probability

$$= \min\left[1, \underbrace{\frac{f(X \mid \Psi')}{f(X \mid \Psi)}}_{\text{likelihood ratio}} \times \underbrace{\frac{f(\Psi')}{f(\Psi)}}_{\text{prior ratio}} \times \underbrace{\frac{f(\Psi \mid \Psi')}{f(\Psi' \mid \Psi)}}_{\text{proposal ratio}}\right]$$

(4) Add the current tree to the chain

(5) Goto 2

Huelsenbeck et al. (2002) *Syst Biol*

Why does this help?

Cancellation of marginal likelihood

When calculating the ratio (R) of posterior densities, the marginal probability of the data cancels.

$$\frac{p(\theta^* \mid D)}{p(\theta \mid D)} = \frac{\frac{p(D \mid \theta^*)\, p(\theta^*)}{\cancel{p(D)}}}{\frac{p(D \mid \theta)\, p(\theta)}{\cancel{p(D)}}} = \frac{p(D \mid \theta^*)\, p(\theta^*)}{p(D \mid \theta)\, p(\theta)}$$

Posterior odds

Apply Bayes' rule to both top and bottom

Likelihood ratio

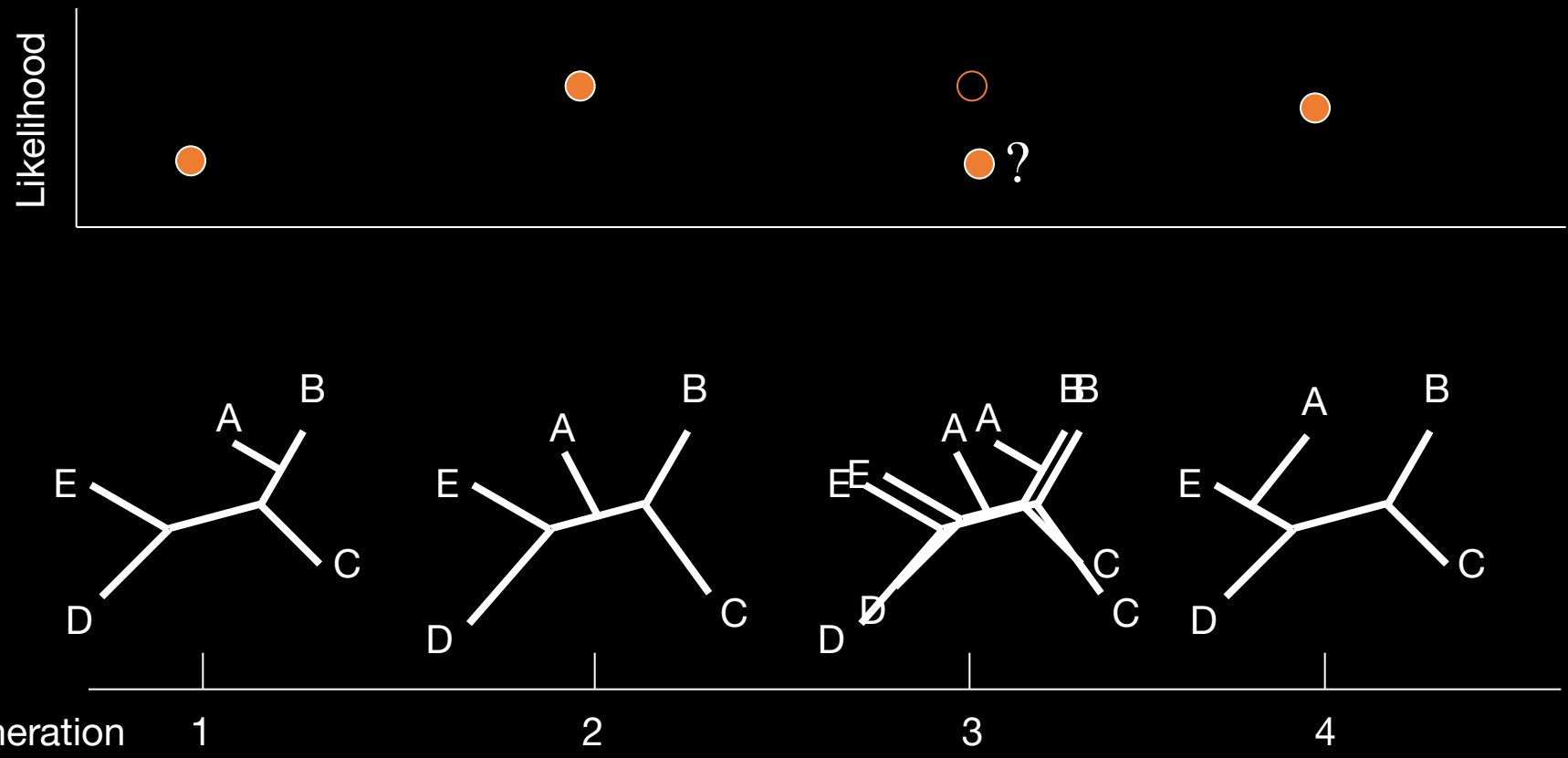
Prior odds

Goto 2???

- In theory (assuming certain basic properties of the chain), MCMC will sample every point in likelihood space in proportion to its posterior probability

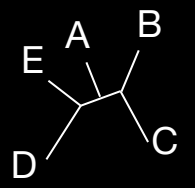
- IF the chain is run for an infinite number of iterations

MCMC in Practice



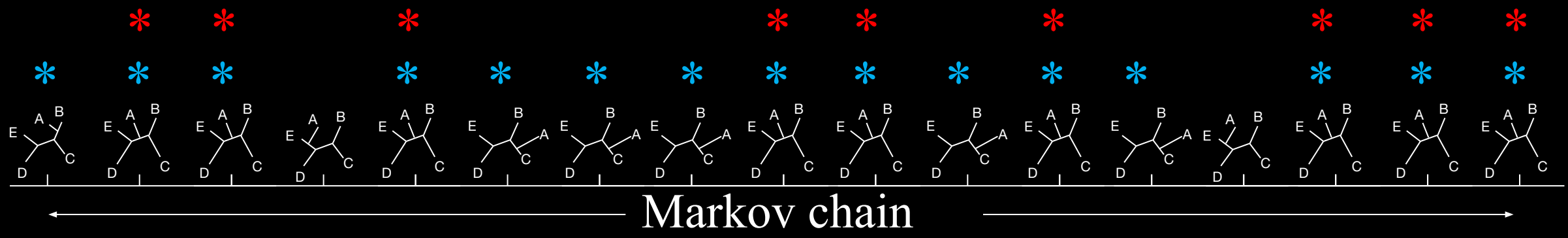
Posterior Probability

If no *a priori* preference is given to specific trees, the posterior probability of trees and bipartitions is equal to their frequency in the Markov chain

Posterior (

) = 9 / 17 ~ 0.53

Posterior (ABC | DE) = 15 / 17 ~ 0.88



Posteriors on TREES

- Simply the frequency of the tree (integrated over all branch lengths) in the Markov chain

- Addresses all phylogenetic hypotheses at once

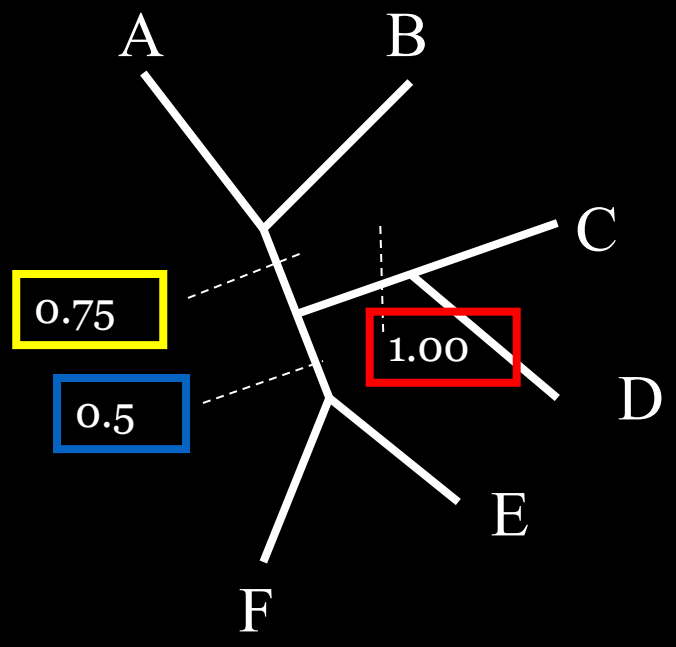
Posteriors on TREES

- Can be very unstable!
- Practical example:
  - 30-sequence alignment
  - 3,000,000 iteration chain
  - 30,000 trees saved in chain (1 / 100 thinning)
  - >25,000 different trees!

- Most-frequent tree sampled twice, so posterior = (2/30,000)

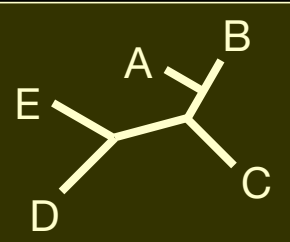
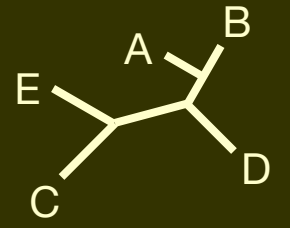
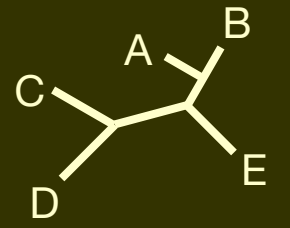
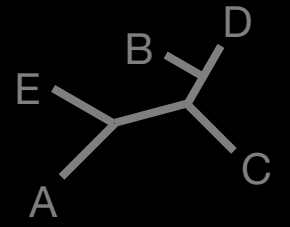
Posteriors on SPLITS

- Far more stable (independent evaluation of tree features)
- Lose information about dependencies within tree



Interpreting Posteriors

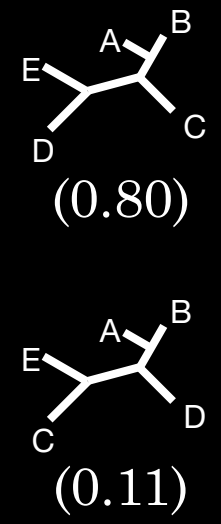
- ‘Confidence intervals’ of models
  - Rank the models in decreasing order of PP, and take the set that corresponds to the top x% (e.g., the top 95%)
  - May include multiple trees or splits, but will certainly exclude a lot more

Tree	Posterior	Cumulative
	0.80	0.80
	0.11	0.91
	0.05	0.96
	0.02	0.98

Interpreting Posteriors

Bayes factors: the ratio of posterior probabilities for two hypotheses (models) H_1 and H

$$B(x) = \frac{P(H_1|x)}{P(H_2|x)}$$



$= 7.2$

B(x)	Interpretation
1-3	Barely worth mentioning
3-20	Positive evidence
20-150	Strong preference
150+	Very strong preference

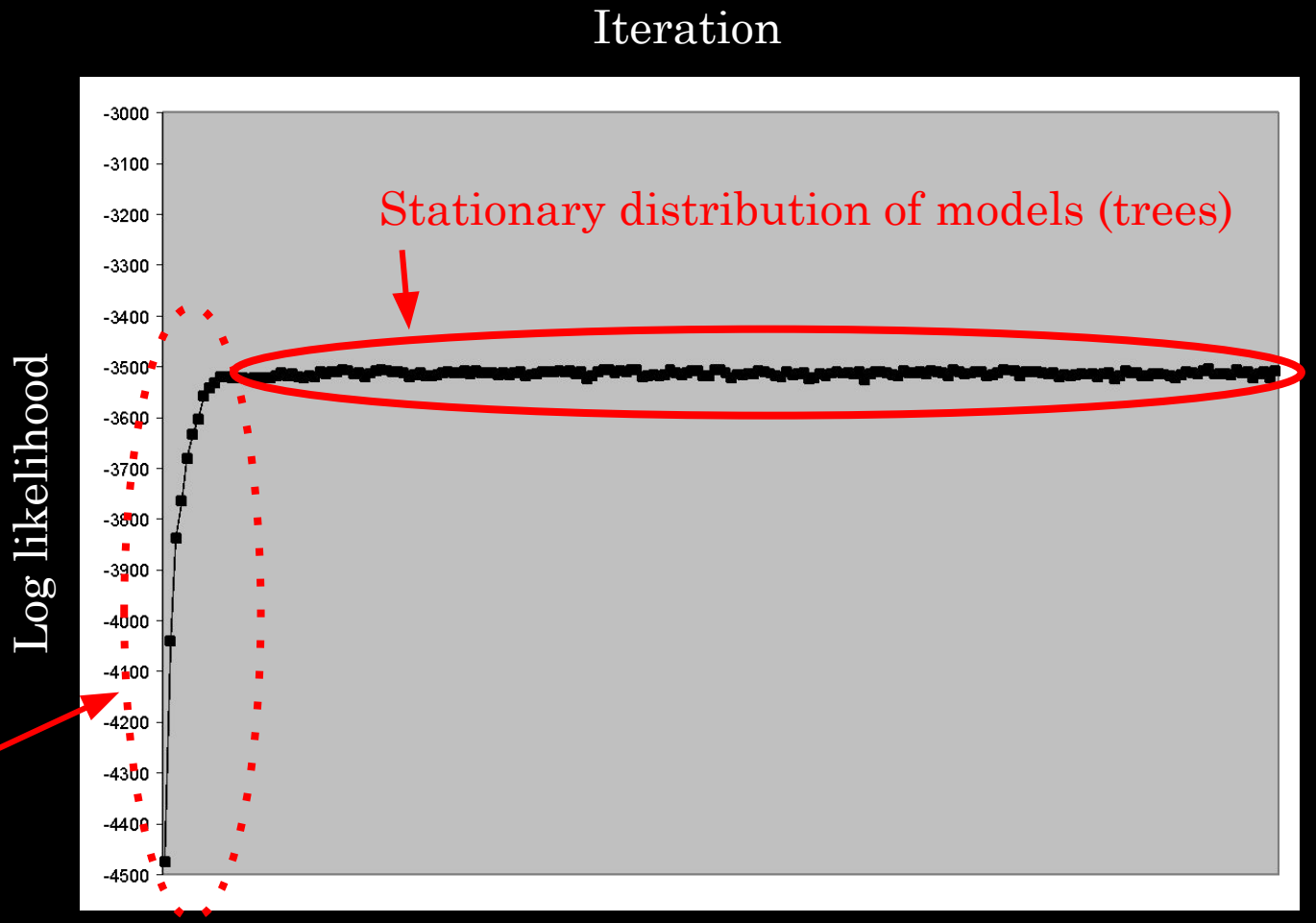
Kaas and Raftery (1995) *Journal of the American Statistical Association*

Markov chains in action!

Evaluate progress using e.g. a log-likelihood plot

Burn-in phase:
low-posterior models that arise from poor initial model parameters

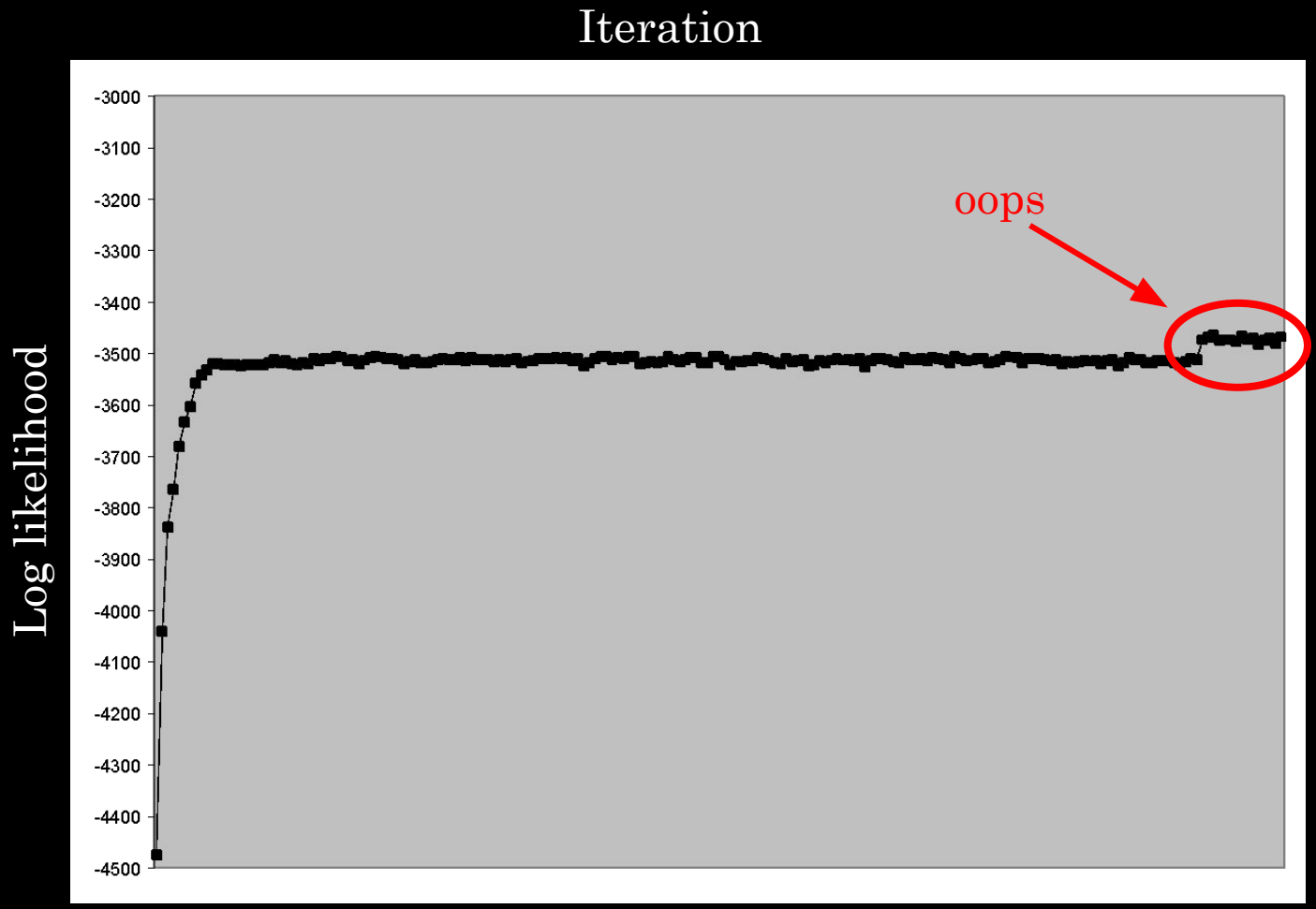
Stationary phase: sampling from models with high posterior probability



Markov chains in action!

However, problems can arise

More-complex models are susceptible to sudden increases in posterior after long periods of sampling



A couple of solutions

- Metropolis-coupled MCMC: heated chains
  - Cold chain collects samples of the posterior distribution
  - Multiple heated chains are more likely to accept moves to lower-probability trees
  - Chains can SWAP
- Replicated runs, assess convergence of chains

More-Complex Models

Variable rates along branches

OPEN ACCESS Freely available online

PLOS BIOLOGY

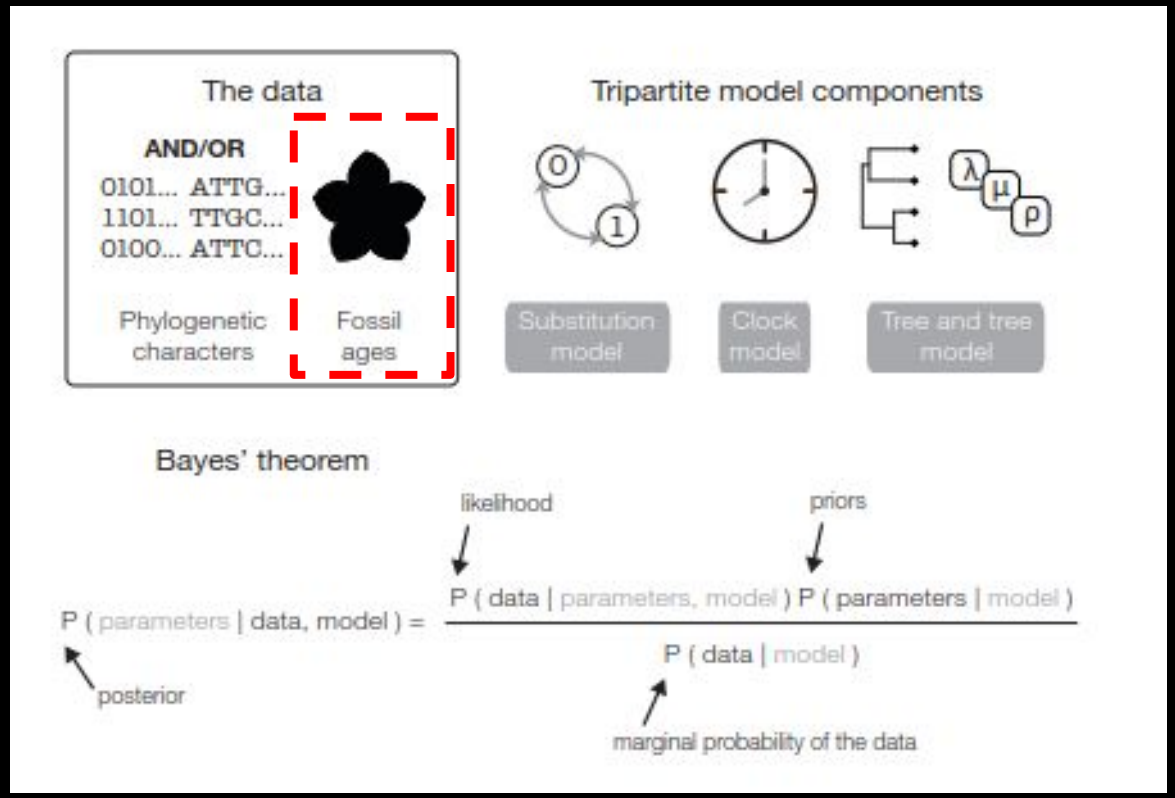
Relaxed Phylogenetics and Dating with Confidence

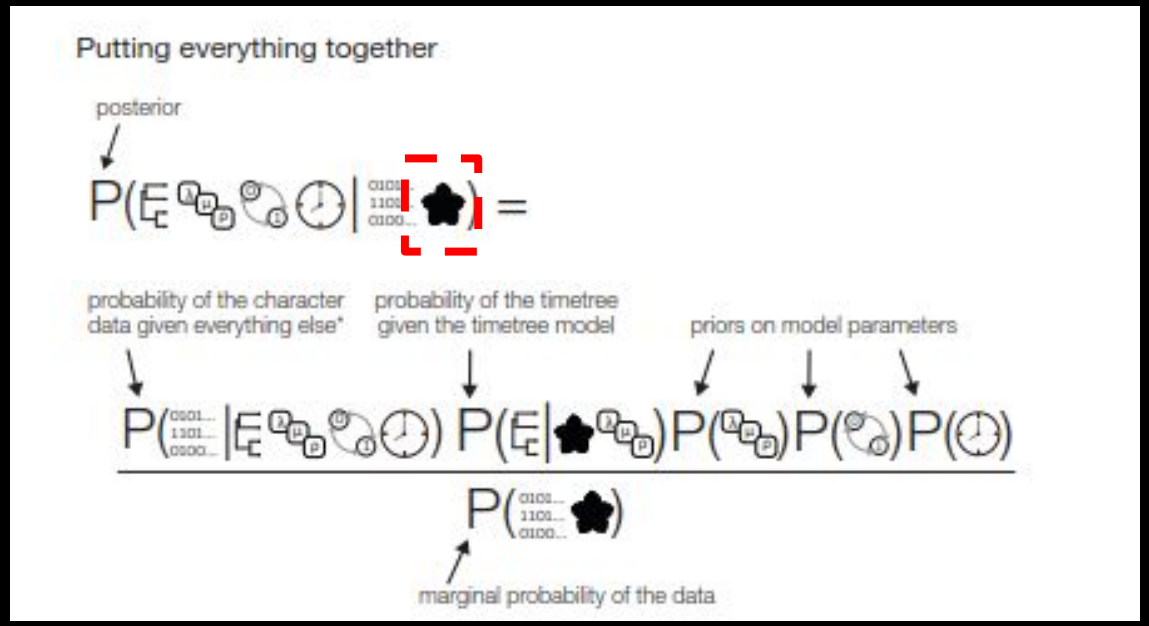
Alexei J. Drummond, Simon Y. W. Ho, Matthew J. Phillips, Andrew Rambaut*

Department of Zoology, University of Oxford, Oxford, United Kingdom

In phylogenetics, the unrooted model of phylogeny and the strict molecular clock model are two extremes of a continuum. Despite their dominance in phylogenetic inference, it is evident that both are biologically unrealistic and that the real evolutionary process lies between these two extremes. Fortunately, intermediate models employing

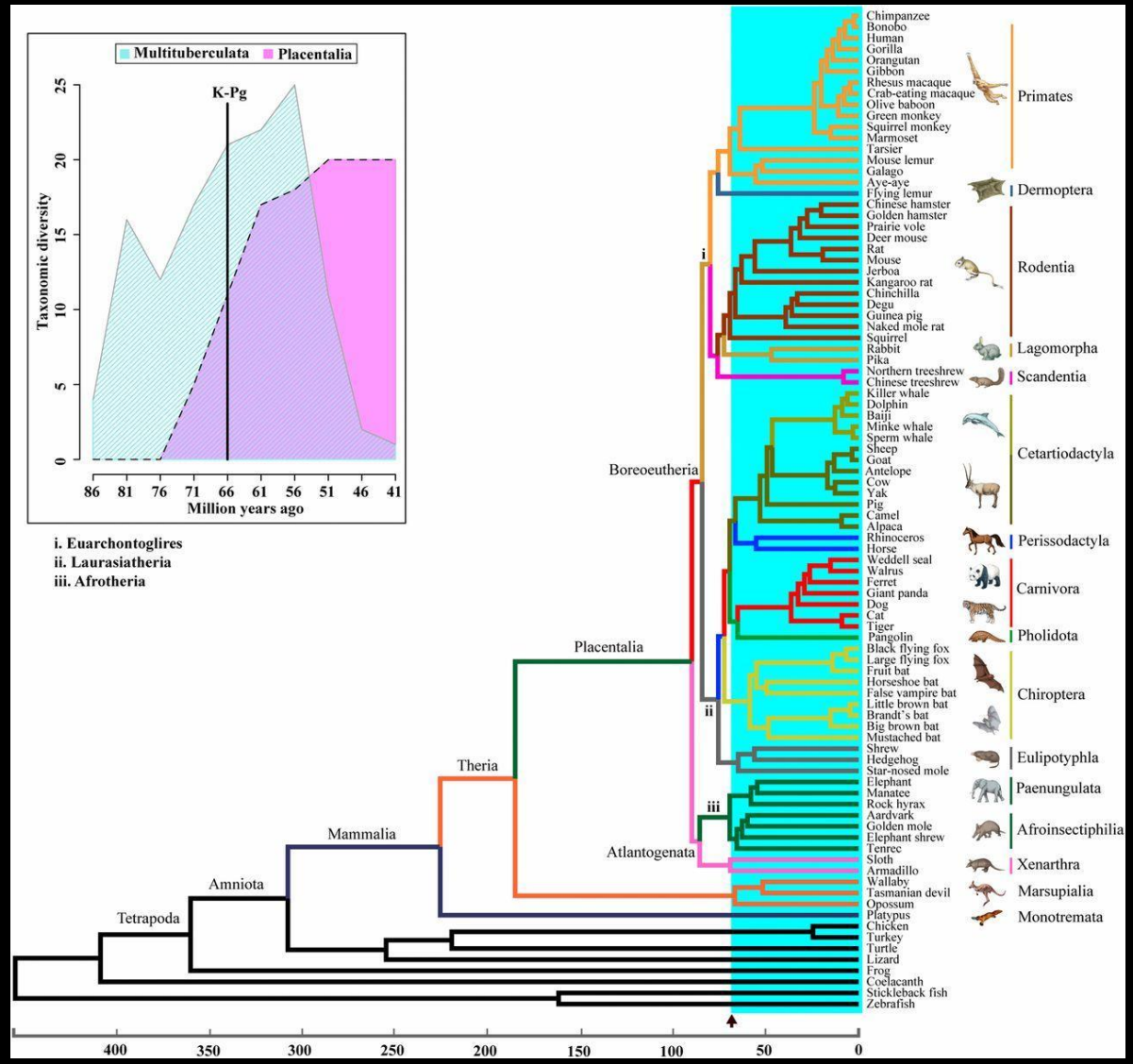
Other models and data: Estimating divergence times with fossils as calibration points





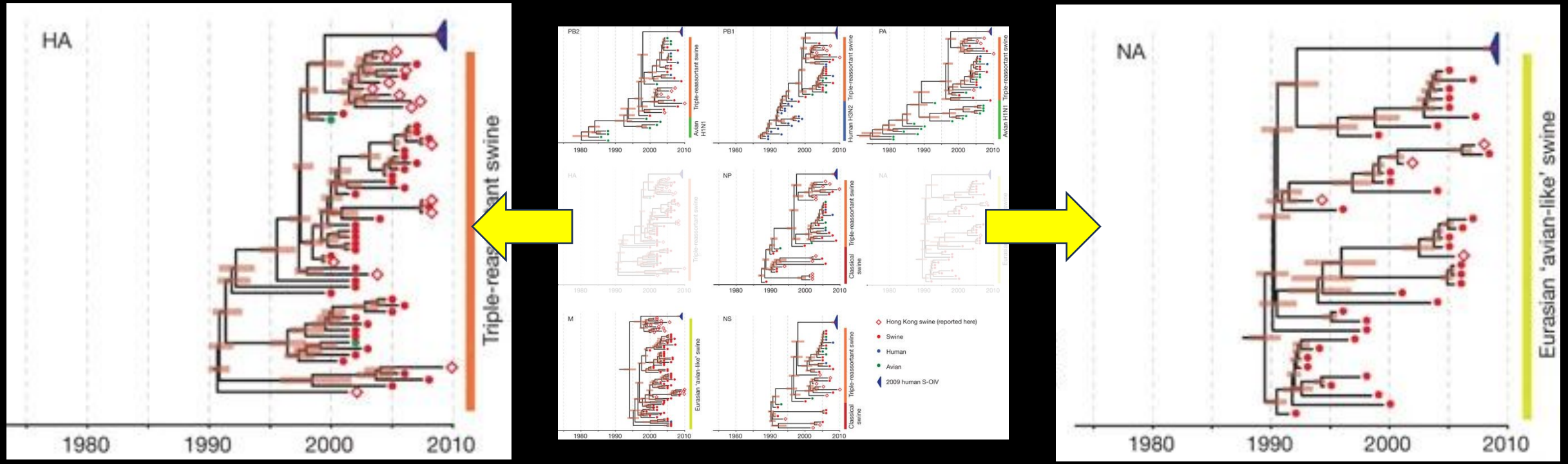
Warnock and Wright (2020) *Elements of Palaeontology*
Accessible version: https://ecoevorxiv.org/repository/view/4372/s

Estimating ancient divergence times using the fossil record for calibration



Liu et al. (2017) Proc Natl Acad Sci

Estimating recent divergence times using samples at or near internal nodes



Smith et al. (2009) *Nature*

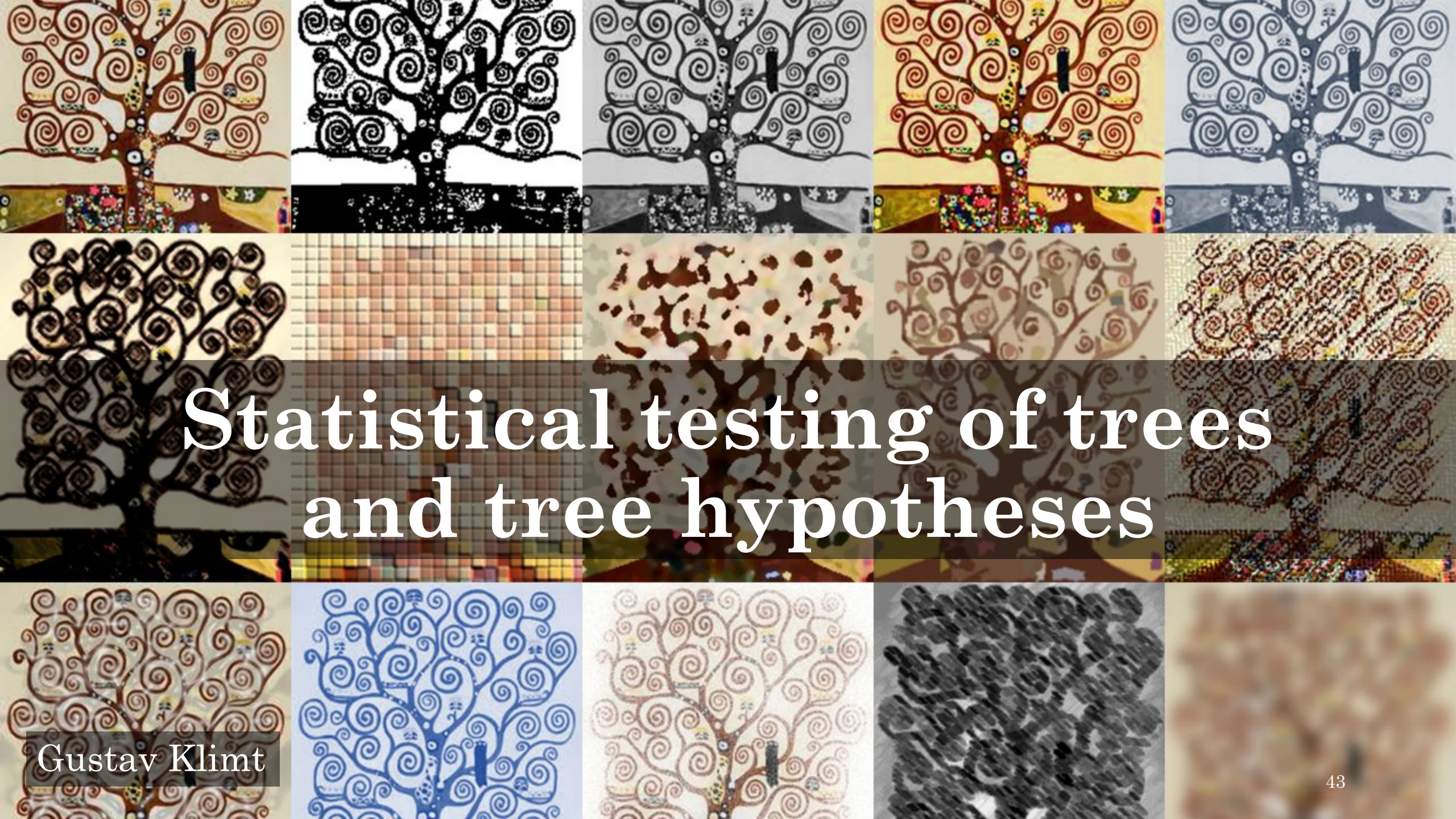
Conclusion

- Parsimony: simple but potentially slow, and not a reasonable assumption in most cases
- Distance: Can be quite fast, often used to build starting trees for other methods
- Maximum Likelihood: Model based, highly accurate to the limits of your data and model
- Bayesian: Probability distributions based on likelihoods, not just the best* tree
- Next up: How much do I trust my trees?

Implementations

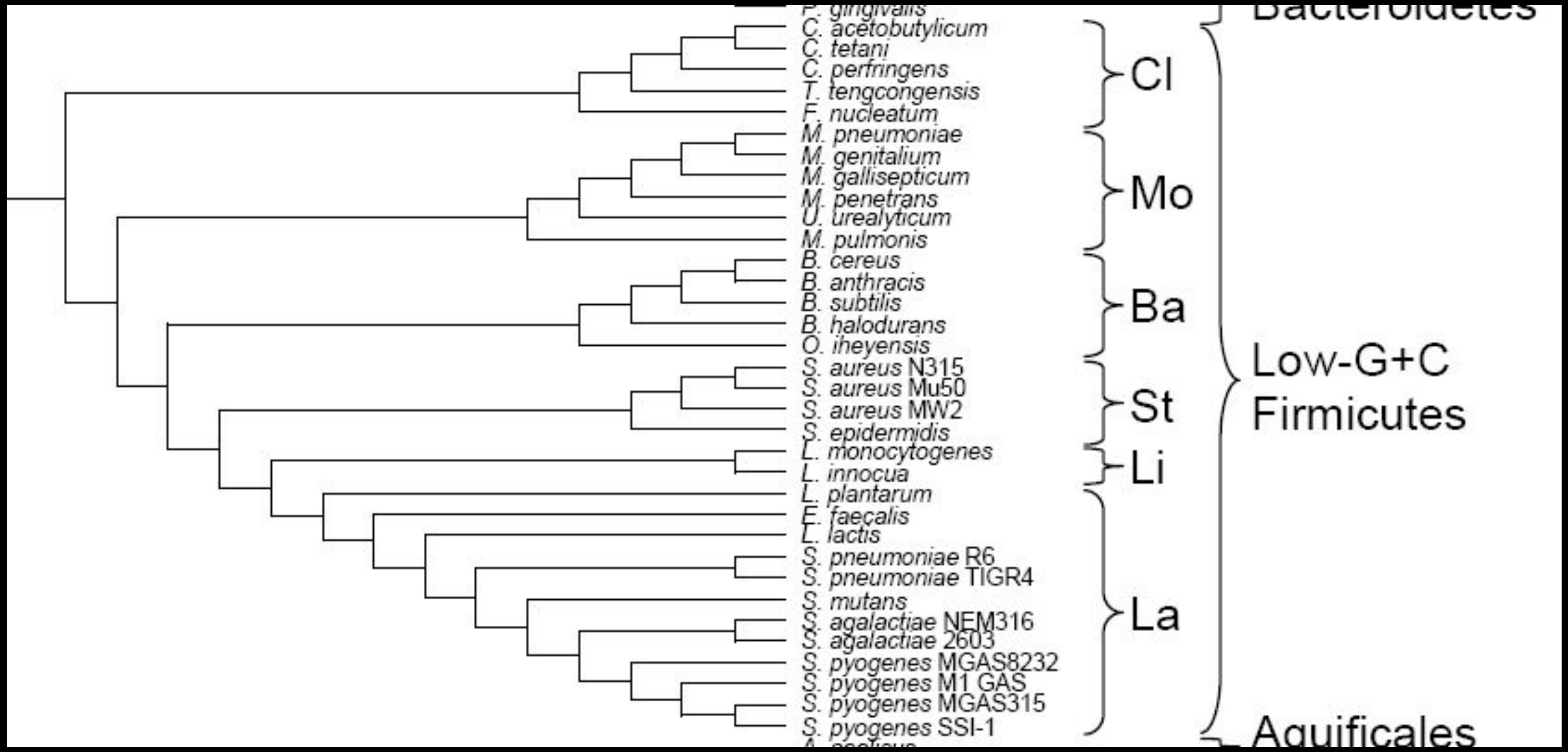
- RevBayes: https://revbayes.github.io/
- BEASTX: https://beast.community/
- BEAST2: https://www.beast2.org/

Höhna et al. (2016) *Systematic Biology*
Drummond and Rambaut (2007) *BMC Evolutionary Biology*
Bouckaert R et al. (2019) *PLoS Computational Biology*



Statistical testing of trees and tree hypotheses

Gustav Klimt



is a hypothesis

…but what is the strength of support for this hypothesis?

Beiko et al. (2005) *PNAS*

Significant Significance Questions

1. Do the data (that's usually the alignment) strongly support the relationships in the tree?

2. Is the recovered tree statistically better than all other possible trees?

3. Is a *tree* really the best explanation of the data?

Significant Significance Questions

1. Do the data (that's usually the alignment) strongly support the relationships in the tree?

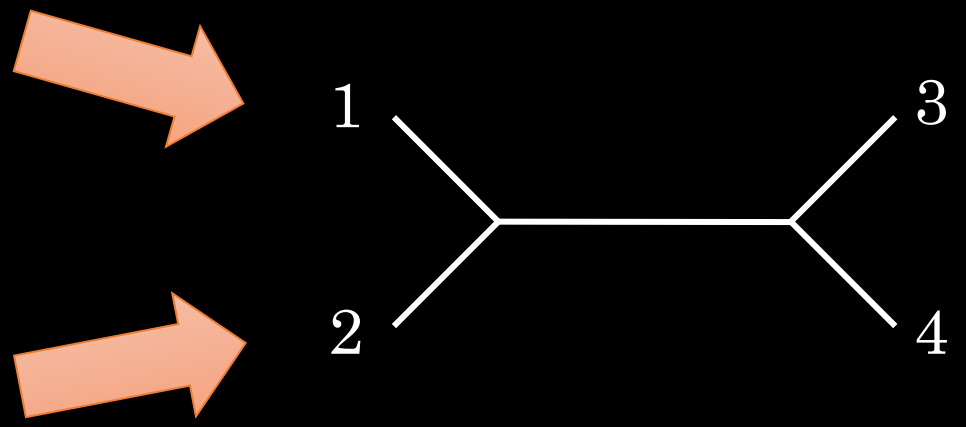
(Bayesian posteriors, for example)

Why ask these awkward questions?

Ask for a tree, get a tree

```
1 ACCGAGCAA
2 ACCGAGCAA
3 ACCGAGCAA
4 ACCGAGCAA
```

```
1 ACCGAATGA
2 ACCGAGCAG
3 GTTAGGCAG
4 GTTAGATGA
```



Problems with datasets

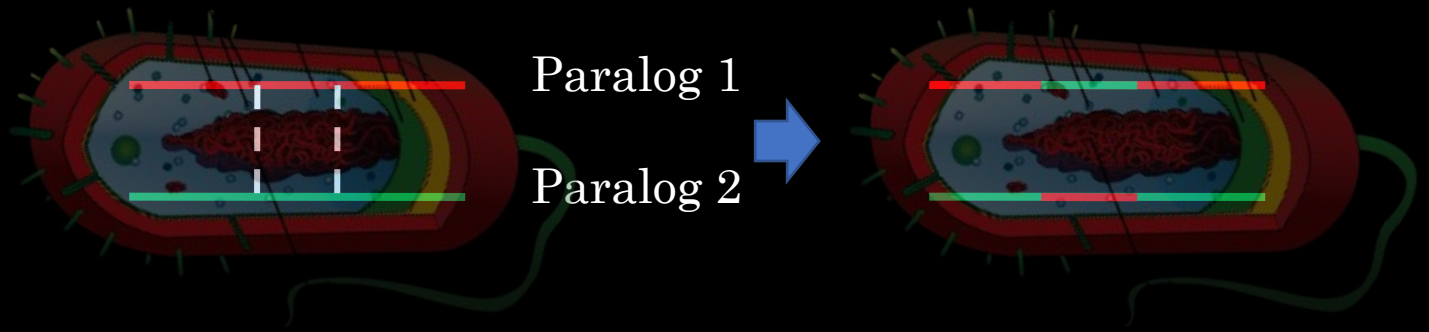
- Signal saturation: too many substitutions (and multiple substitutions!) between sequences

- Lack of signal: some short branches in the tree may lack supporting data or be sufficiently ancient to have been erased

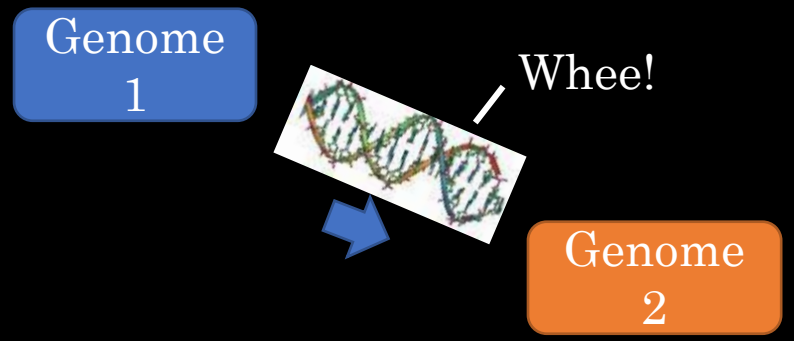
- Misleading signals due to “stochastic errors”, biases, and convergence

And reticulate evolution

Gene conversion / recombination



Lateral gene transfer (one or more genes)



Addressing significance questions

1. Strength of support – resampling, subsampling, and simulation

2. Better than alternatives – Bayesian, paired-site comparisons, testing alternative types of model

3. Treelike signal – phylogenetic networks

(1) Tree support

The Nonparametric bootstrap

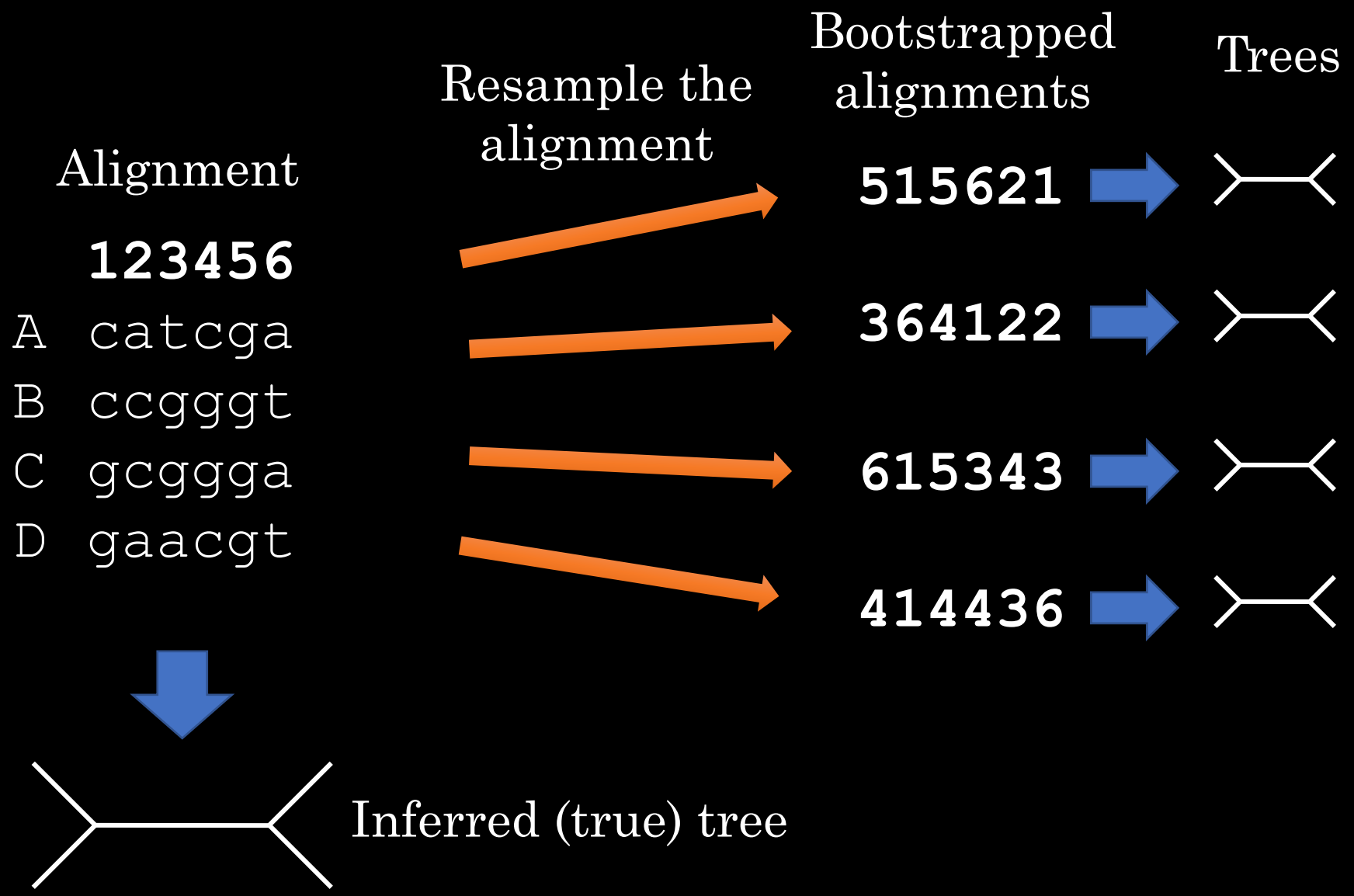
- Resample from the distribution of data points (alignment columns) and see whether we get the same answer
- Do this a bunch of times (100-ish)
- Map the results onto the original tree

Generating bootstrap replicates

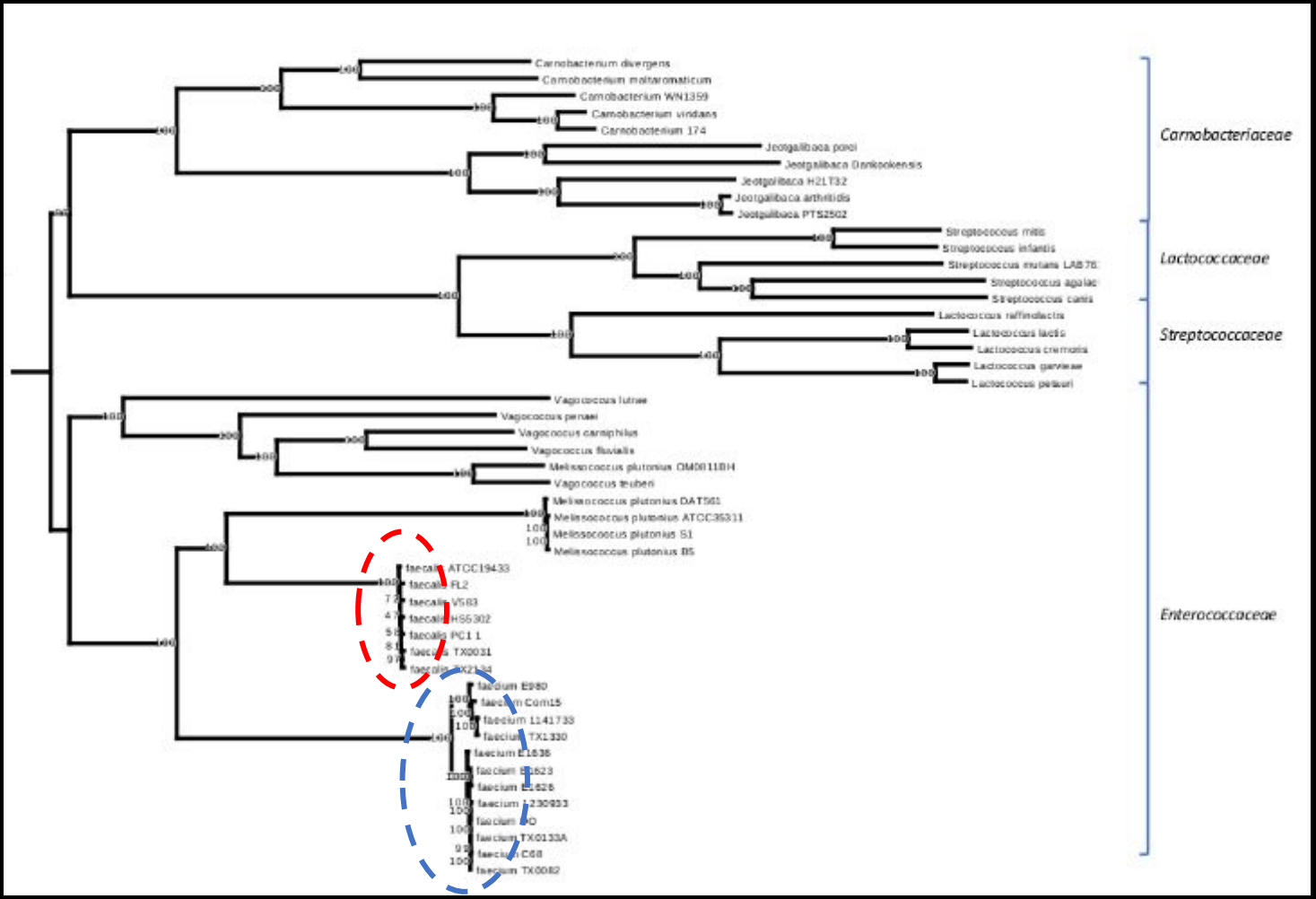
- Resample with replacement from the original population

- Original alignment: *n* columns
- Bootstrapped alignment: still *n* columns

- But some columns will be missing, and some will be present more than once


Alignment
123456
A catcga
B ccgggt
C gcggga
D gaacgt
Resample the alignment
Bootstrapped alignments
515621
364122
615343
414436
Trees
Inferred (true) tree

Support for tree features



Map bootstrap values onto the original tree

The bootstrap for a given grouping of taxa in the tree (supported by an edge) is equal to the frequency that grouping is observed among the bootstrap replicates

70% is often used as a threshold (based on simulation)

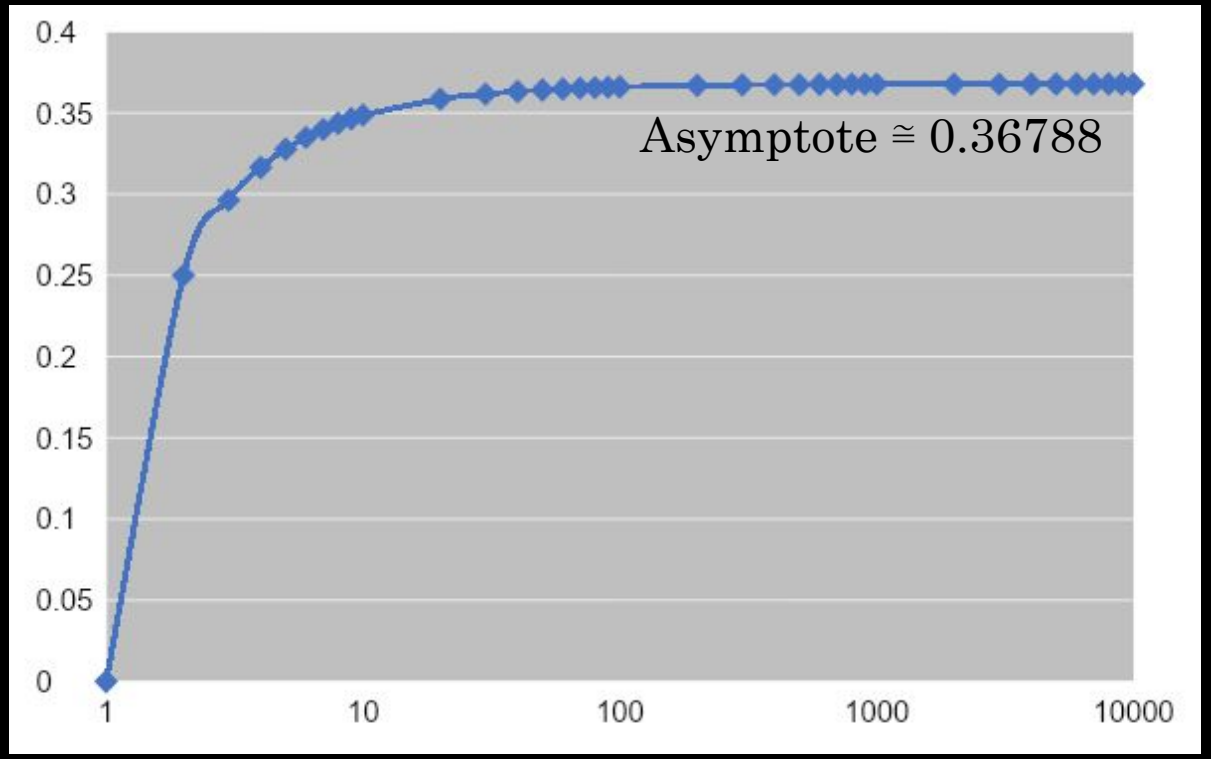
~ 100% (complete support)
~ 50% (much weaker support)

Tree from Alex Manuele

What is the bootstrap doing?

- The bootstrap is randomly reweighting characters in the alignment, and assessing the impact on the phylogeny

- The probability of a given character being excluded (weight = 0) is equal to $(1 - 1/N)^N$



What is the bootstrap doing?

- The goal of the bootstrap is to simulate an infinite population (number of alignment columns) by considering a range of reweightings on the existing data

Limitation of nonparametric methods

- The nonparametric bootstrap method is limited by the availability of reliable data

- This resampling procedure may therefore not cover the range of alternative trees

- The parametric bootstrap simulates data on the proposed tree, and determines how often that tree can be recovered. Cool but rarely used in practice

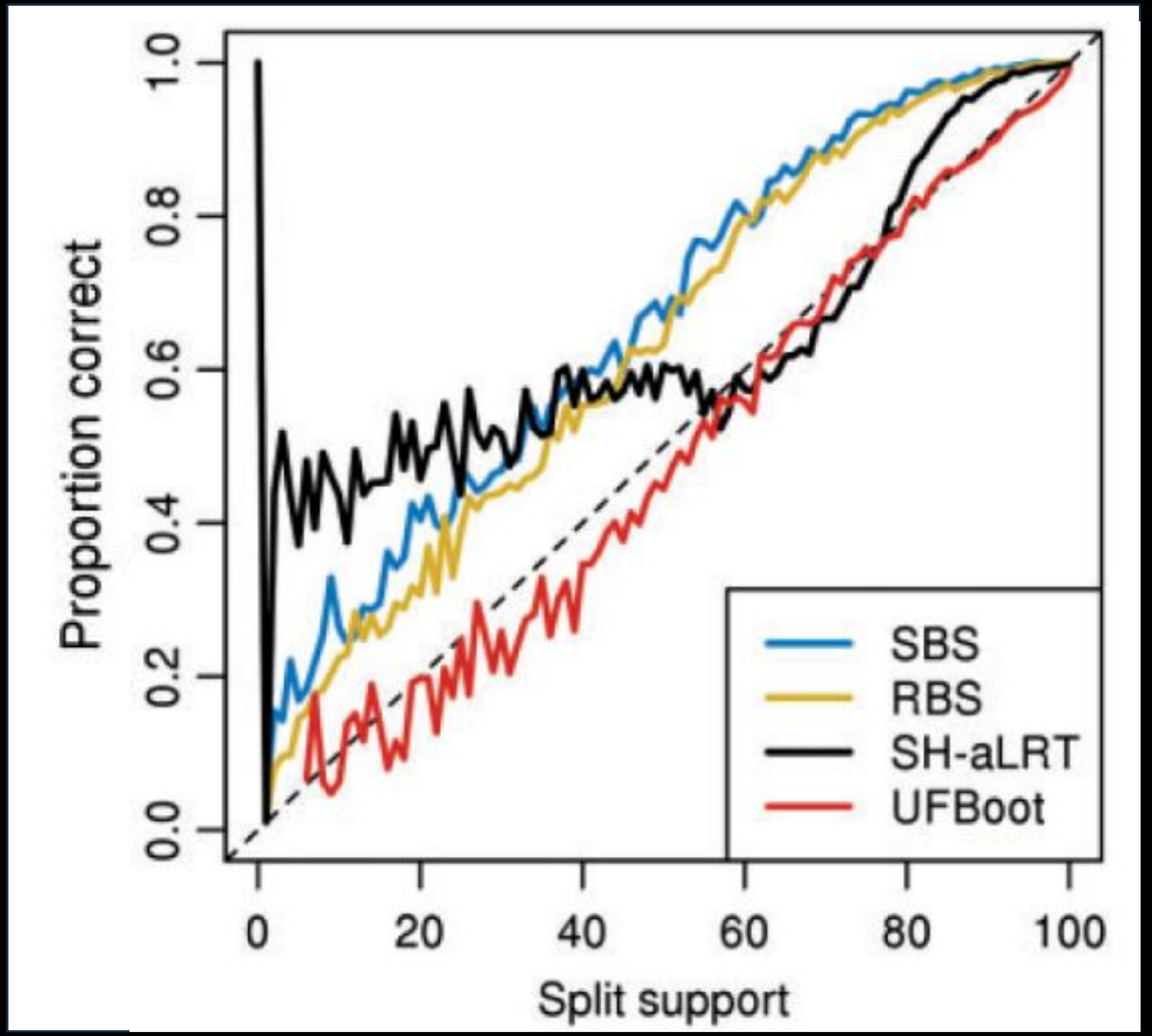
https://en.wikipedia.org/wiki/Snail#/media/File:Snail.jpg

Alternatives to doing the likelihood search 100 or 1000 or 10,000 times

- aLRT: Estimate *local* support using e.g. NNI and re-use likelihoods (since the bootstrap replicates are just the same columns re-weighted)
- SH-aLRT: use simulations to generate a realistic distribution of likelihoods
- Ultrafast bootstrap: Perform the search for all bootstrap replicates *simultaneously*. Keep a record of the best tree for each bootstrap replicate, and update as better trees are found during the search

Minh et al. (2013) *Mol Biol Evol*

Accuracy matters: bootstrap correctness for simulated data sets

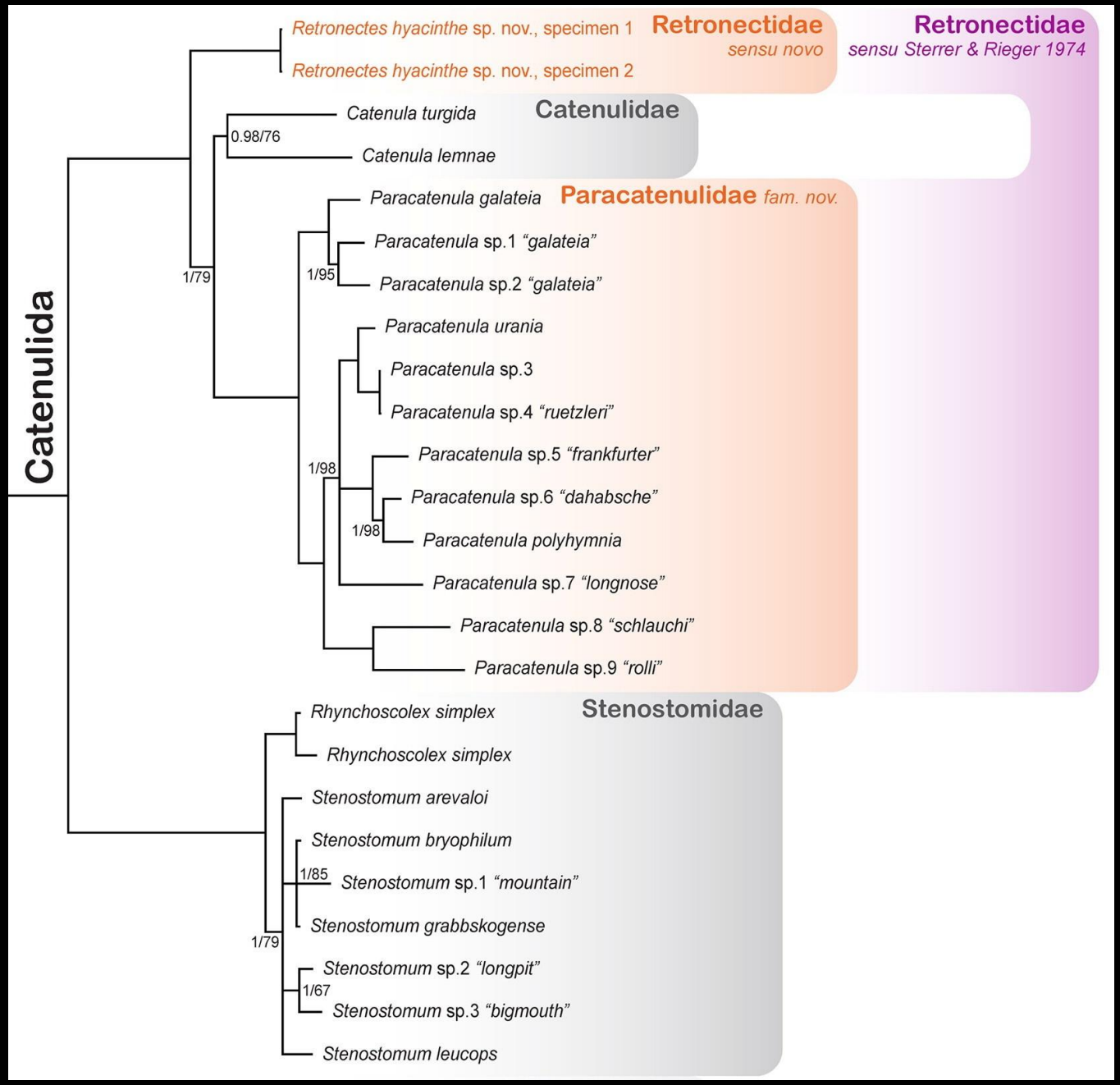


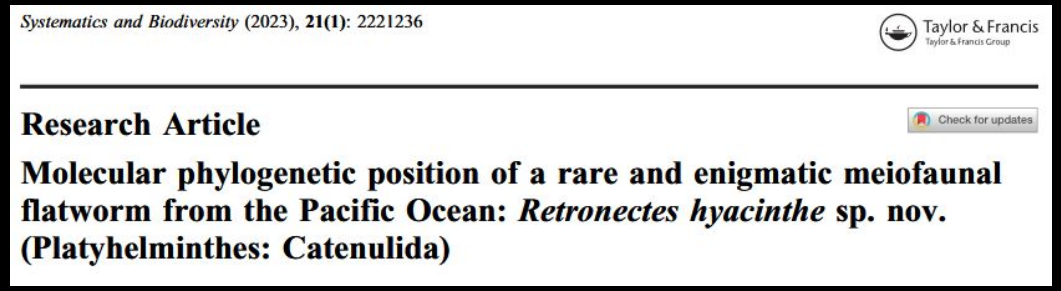
Minh et al. (2013) *Mol Biol Evol*

Bayesian Posteriors vs Bootstraps

For example:

- Posterior < 0.95 or bs < 0.7: nodes collapsed
- Posterior = 1.0 and bs = 100: values not shown
- All other values shown





Systematics and Biodiversity (2023), **21(1)**: 2221236

Taylor & Francis

Taylor & Francis Group

Research Article

Check for updates

Molecular phylogenetic position of a rare and enigmatic meiofaunal flatworm from the Pacific Ocean: *Retronectes hyacinthe* sp. nov. (Platyhelminthes: Catenulida)

Van Steenkiste et al. (2023) *Systematics and Biodiversity*

Problems with resampling in general

- Limited to asking the question, “to what extent do the data support the tree”?

- Questions not addressed:
  - How good are alternative trees?
  - Did we choose the right model?
  - Do the data support a tree representation?

Another Question: How complex should my model be?

Remember: alternative models (plus rate variation, etc…)

$$Q_{\text{JC69}} = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & -3\alpha & \alpha & \alpha & \alpha \\ C & \alpha & -3\alpha & \alpha & \alpha \\ G & \alpha & \alpha & -3\alpha & \alpha \\ T & \alpha & \alpha & \alpha & -3\alpha \end{array}$$

Jukes-Cantor: all substitution rates (α) and nucleotide frequencies are the same

$$Q_{\text{GTR}} = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & -q_A & r_{AC}\pi_C & r_{AG}\pi_G & r_{AT}\pi_T \\ C & r_{AC}\pi_A & -q_C & r_{CG}\pi_G & r_{CT}\pi_T \\ G & r_{AG}\pi_A & r_{CG}\pi_C & -q_G & r_{GT}\pi_T \\ T & r_{AT}\pi_A & r_{CT}\pi_C & r_{GT}\pi_G & -q_T \end{array}$$

General Time Reversible (GTR): different rates of change, and nucleotide frequencies

Too few parameters = cannot realistically model evolution

Too many parameters = too much flexibility, likelihood is inflated

How do we choose?

The problem

- More parameters will nearly always give a higher likelihood: it's easier to fit the pattern, but it's also easier to fit the noise

- We need a way to balance complexity and model fit

Likelihood ratio test: Nested models

If one model is a special case of another, then we can take the negative logarithm of the ratio of the simpler model M_0 to the more-complex model M_1:

$$\lambda = -2ln\left(\frac{P(D|JC)}{P(D|GTR)}\right)$$

Jukes-Cantor (M_0):
1 free parameter

$$Q_{JC69} = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & -3\alpha & \alpha & \alpha & \alpha \\ C & \alpha & -3\alpha & \alpha & \alpha \\ G & \alpha & \alpha & -3\alpha & \alpha \\ T & \alpha & \alpha & \alpha & -3\alpha \end{array}$$

GTR (M_1):
9 free parameters

$$Q_{GTR} = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & -q_A & r_{AC}\pi_C & r_{AG}\pi_G & r_{AT}\pi_T \\ C & r_{AC}\pi_A & -q_C & r_{CG}\pi_G & r_{CT}\pi_T \\ G & r_{AG}\pi_A & r_{CG}\pi_C & -q_G & r_{GT}\pi_T \\ T & r_{AT}\pi_A & r_{CT}\pi_C & r_{GT}\pi_G & -q_T \end{array}$$

λ follows a Chi-squared distribution with degrees of freedom ($M_1 - M_0$) = 8 in this case

Can accept or reject based on p-value

Akaike information criterion (AIC): Non-nested models

- If neither model is a special case of another, we can calculate the AIC for each model and compare them:

$$AIC(M) = -2ln(P(D|M)) + 2k$$

Free parameters in M

Kimura 80:
2 free parameters
(transition / transversion rates)

$$Q_{\text{K80}} = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & -q_A & \beta & \alpha & \beta \\ C & \beta & -q_C & \beta & \alpha \\ G & \alpha & \beta & -q_G & \beta \\ T & \beta & \alpha & \beta & -q_T \end{array}$$

Felsenstein 84:
3 free parameters
(nucleotide frequencies)

$$Q_{\text{F84}} = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & -q_A & \pi_C & \kappa\pi_G & \pi_T \\ C & \pi_A & -q_C & \pi_G & \kappa\pi_T \\ G & \kappa\pi_A & \pi_C & -q_G & \pi_T \\ T & \pi_A & \kappa\pi_C & \pi_G & -q_T \end{array}$$

Favour the model with the smaller AIC

Choosing a Model

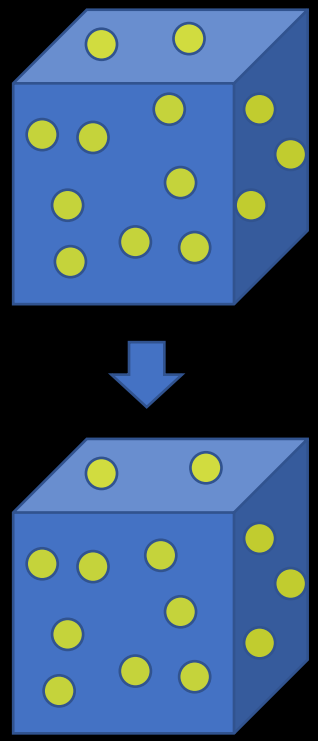
- MODELTEST – propose and evaluate more-complex models hierarchically and stop when AIC stops improving
- ModelFinder (in IQ-TREE) – try them all, choose based on AIC or related statistics
- How aggressively should we optimize our model?

Posada and Crandall (1998) *Bioinformatics*
Kalyaanamoorthy et al (2017) *Nature Methods*

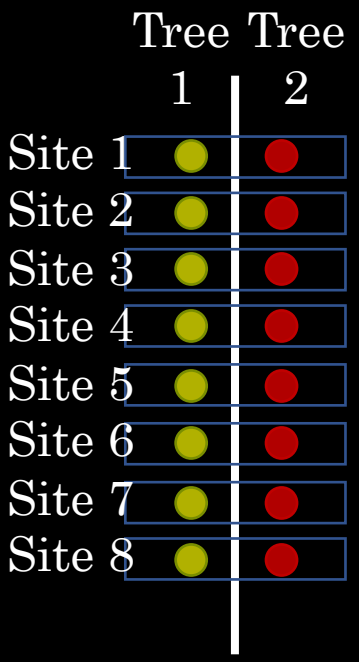
Best tree?

Is the best tree better than some other tree?

We need to approach the data differently



So far – reshuffle data, but only infer results from complete data sets

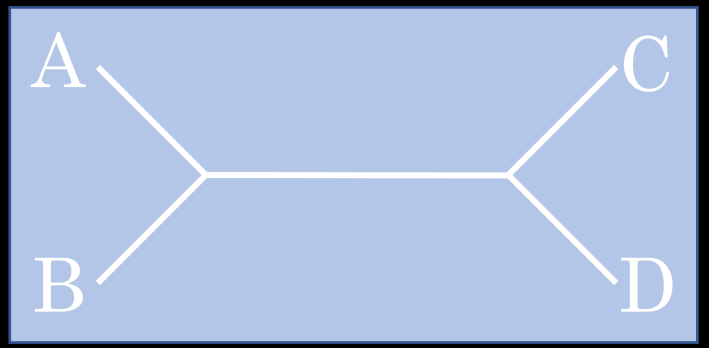


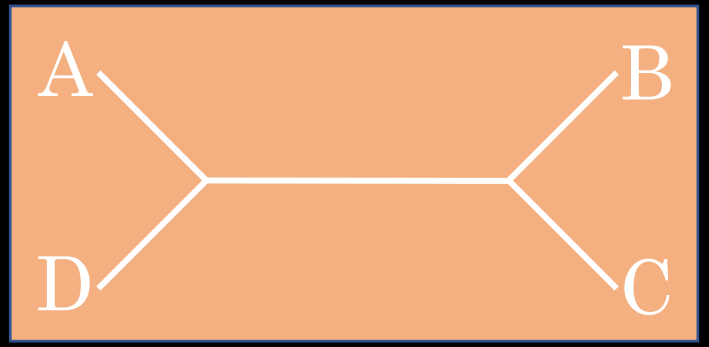
Now – compare individual sites to come up with an overall conclusion of significance

Basic principles

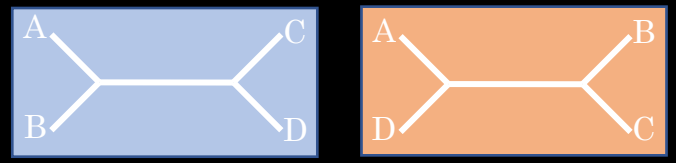
For two trees, compare the fit at each alignment site either quantitatively or qualitatively

	1	2	3	4	5	6
A	c	a	t	c	c	t
B	c	c	g	g	g	t
C	g	c	g	g	g	a
D	g	a	a	c	g	t
Favours?						
By how much?	5.2	3.1	0.9	6.6	0.3	0.2





The winning sites test: "An up-or-down vote"



4 sites favour the red tree
2 favour the blue tree

	1	2	3	4	5	6
A	c	a	t	c	c	t
B	c	c	g	g	g	t
C	g	c	g	g	g	a
D	g	a	a	c	g	t
Favours?						
By how much?	5.2	3.1	0.9	6.6	0.3	0.2

Use the binomal distribution to assess the significance of this difference

$$\binom{n}{k} p^k (1-p)^{n-k}$$

What is the probability that 4 or greater coin tosses will come up with the same result?

p depends on alignment length

4 out of 6: p = 0.6875 (not significant)

40 out of 60: p = 0.0124
(significant at threshold of 0.05)

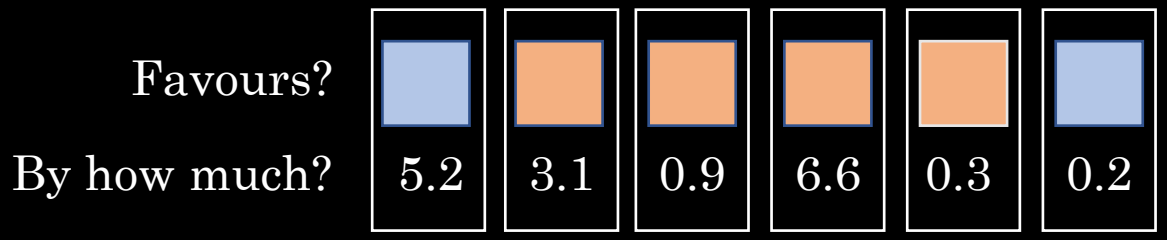
$$\binom{n}{k} p^k (1-p)^{n-k}$$

400 out of 600: p = 2.3×10^{-16}

Paired *t* test: A quantitative approach

	1	2	3	4	5	6
A	c	a	t	c	c	t
B	c	c	g	g	g	t
C	g	c	g	g	g	a
D	g	a	a	c	g	t
Favours?						
By how much?	5.2	3.1	0.9	6.6	0.3	0.2

Consider not the number of sites, but the mean and variance of differences across all sites



Mean of differences: (-5.2 + 3.1 + 0.9 + 6.6 + 0.3 – 0.2) / 6 = 0.916

Variance: 15.22

We compute a t statistic using the following formula:

$$t = \frac{\bar{x}}{\text{var}}\sqrt{N} = 0.148$$

Compare to the t distribution for 5 degrees of freedom: $p = 0.888$

Nobody uses these tests

Why not?

These tests are very biased

- Statistical tests generally assume a random sample
- The (distributions of) trees we want to test are most definitely not!
- Less-biased tests often depend on more-sophisticated comparisons and (again) simulation

What do people actually do?

- Dedicated tests such as Kishino-Hasegawa (KH), Shimodaira-Hasegawa (SH), and Approximately Unbiased (AU) which try to alleviate the bias in tree selection in various ways

- Key concepts include bootstrapping, simulation, and weighted corrections for multiple comparisons

Kishino and Hasegawa (1989) *Journal of Molecular Evolution*
Shimodaira and Hasegawa (1999) *Molecular Biology and Evolution*
Shimodaira (2002) *Systematic Biology*

Summary

- Your trees may look great but be unsupported by the data
- Bootstrap tests: How strong is the support for my tree?
- Statistical comparisons of trees: How much better is this tree than that tree?