Molecular Phylogenetics Course

Statistical frameworks for modelling in phylogenetics

Niklas Wahlberg
Jadranka Rota

Some slides from Paul Lewis (University of Connecticut, USA)

Maximum likelihood

Maximum Likelihood Estimation (MLE)

- Statistical method for estimating parameters of a model (e.g. mean and variance of a normal distribution)
- Originally developed by R. A. Fisher in the 1920s
- Adapted for phylogenetics by Joe Felsenstein

Joe Felsenstein

Joe Felsenstein

Born	Joseph Felsenstein May 9, 1942 (age 79)
Alma mater	University of Chicago
Known for	PHYLIP Felsenstein's tree-pruning algorithm

Wikipedia

EVOLUTIONARY TREES FROM DNA-SEQUENCES - A MAXIMUM-LIKELIHOOD APPROACH

FELSENSTEIN, J

1981 | JOURNAL OF MOLECULAR EVOLUTION 17 (6) , pp.368-376

Full Text Finder Full Text at Publisher •••

11,479 Citations

26 References

Web of Science 2024-05-14

3

Likelihood of a hypothesis

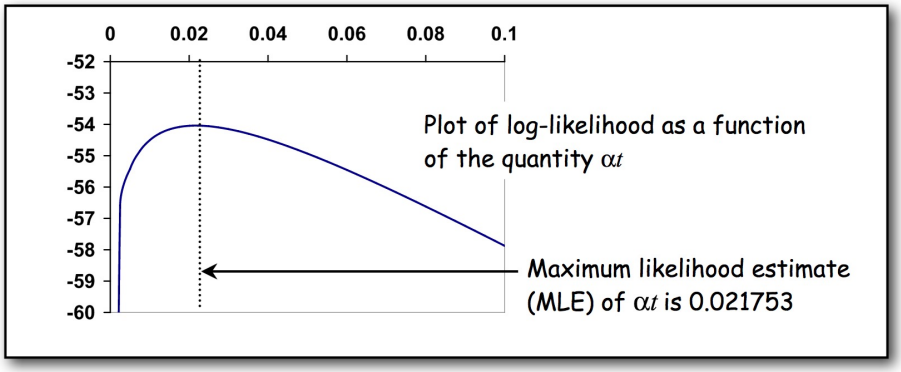
- Likelihood (L) is proportional to the probability (Pr) of observing the data (D) given a model (M) – *conditional probability*

$$L(M) = Pr(D \mid M)$$

- We can examine this likelihood function to find where it is highest and identify the parameters of the model at this point -> Maximum Likelihood Estimates

4

Likelihood function



Paul O. Lewis (2014 Woods Hole Workshop in Molecular Evolution) 22

5

Likelihood of a hypothesis

- Likelihood (L) is proportional to the probability (P) of observing the data (D) given a model (M) – *conditional probability*

  – L(M) = Pr (D | M)

- We can examine this likelihood function to find where it is highest and identify the parameters of the model at this point -> Maximum Likelihood Estimates
- **In molecular phylogenetics, likelihood is the probability of observing the sequences given our model (e.g. GTR+G and our tree topology)**

6

Maximum Likelihood

- For reconstructing phylogenies

Model

  - which tree topology (τ), branch lengths, and parameters of DNA evolution model (θ) (e.g. transition/transversion ratio, base frequencies, …) are maximizing the probability of observing the sequences at hand?

Data

$$L(\tau,\theta) = Pr(Data \mid \tau,\theta)$$

ML analysis in short

- Tree topology is obtained
- Branch lengths and parameters of the DNA substitution model are optimized
- Different topologies (with branch lengths and DNA substitution model parameters optimized) are compared based on their likelihood as the optimality criterion
- The topology with the highest likelihood needs to be found

Likelihood of a single sequence

First 32 nucleotides of the $\psi\eta$-globin gene of gorilla:

GAAGTCCTTGAGAAATAAACTGCACACACTGG

$$L = \pi_G\pi_A\pi_A\pi_G\pi_T\pi_C\pi_C\pi_T\pi_T\pi_G\pi_A\pi_G\pi_A\pi_A\pi_A\pi_T\pi_A\pi_A\pi_A\pi_C\pi_T\pi_G\pi_C\pi_A\pi_C\pi_A\pi_C\pi_A\pi_C\pi_T\pi_G\pi_G$$
$$= \pi_A^{12}\pi_C^{7}\pi_G^{7}\pi_T^{6}$$

$$\ln L = 12\ln(\pi_A) + 7\ln(\pi_C) + 7\ln(\pi_G) + 6\ln(\pi_T)$$

We can already see by eye-balling this that the F81 model will fit better than the JC69 model because there are about twice as many As as there are Cs, Gs and Ts.

Likelihood ratio test

Find lnL under F81 model:

$$\ln L = 12\ln(\pi_A) + 7\ln(\pi_C) + 7\ln(\pi_G) + 6\ln(\pi_T)$$
$$= 12\ln(0.375) + 7\ln(0.21875) + 7\ln(0.21875) + 6\ln(0.1875)$$
$$= -43.1$$

Find lnL under JC69 model:

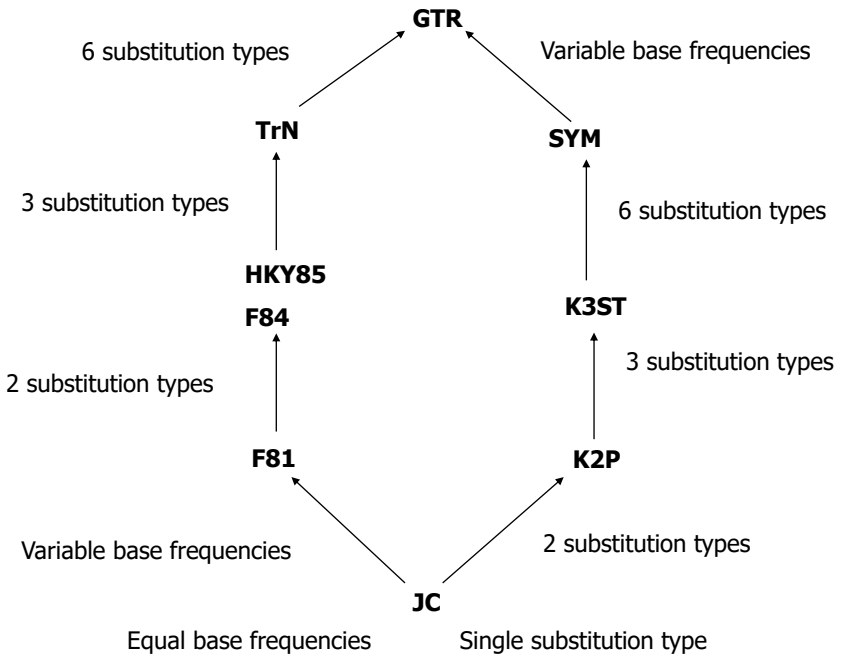
$$\ln L = 12\ln(\pi_A) + 7\ln(\pi_C) + 7\ln(\pi_G) + 6\ln(\pi_T)$$
$$= 12\ln(0.25) + 7\ln(0.25) + 7\ln(0.25) + 6\ln(0.25)$$
$$= -44.4$$

F81 does fit better (-43.1 > -44.4), but not significantly better (P = 0.457, chi-squared with 3 d.f.*)

Find likelihood ratio test statistic:

$$LR = -2\left(\ln L_{JC69} - \ln L_{F81}\right) = -2\left[-44.4 - \left(-43.1\right)\right] = 2.6$$

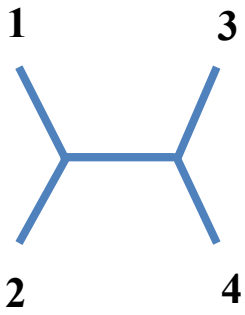
*The number of degrees of freedom equals the difference between the two models in the number of parameters. In this case, F81 has 3 parameters and JC69 has 0, so d.f. = 3 - 0 = 3



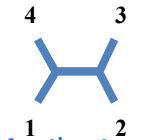
11

Maximum Likelihood tree reconstruction

1 AGAGAC
2 AGCGAC
3 CGATTC
4 TGATAG



Tree A



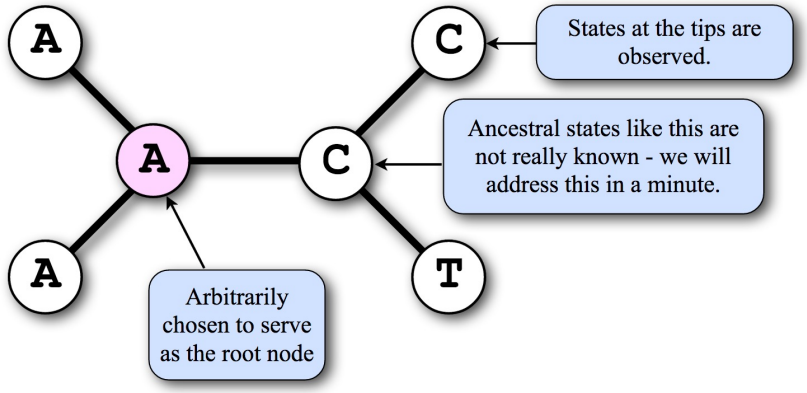
Another tree

What is the likelihood that Tree A (rather than another tree) could have generated the sequence alignment?

12

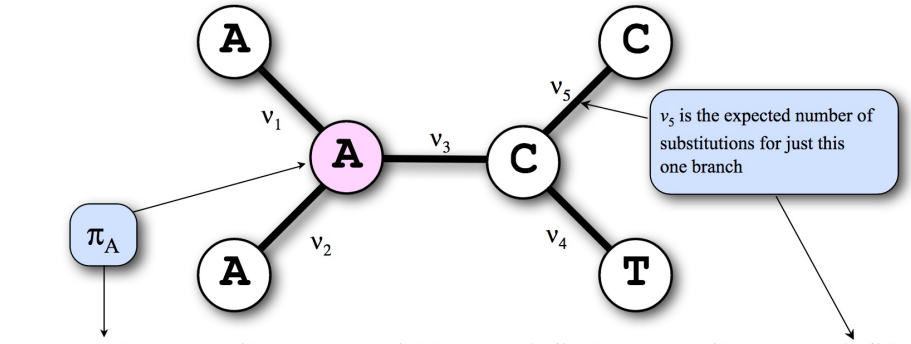
Likelihood of an unrooted tree

(data shown for only one site)



13

Likelihood for site *k*



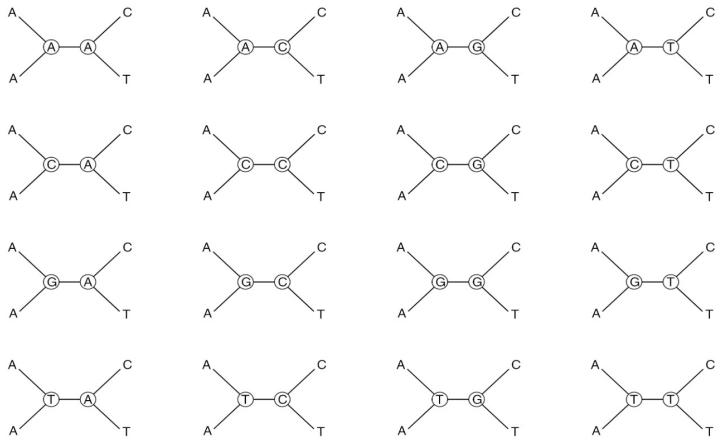
$$L_k = \tfrac{1}{4}\left[\tfrac{1}{4}+\tfrac{3}{4}e^{-4\nu_1/3}\right]\left[\tfrac{1}{4}+\tfrac{3}{4}e^{-4\nu_2/3}\right]\left[\tfrac{1}{4}-\tfrac{1}{4}e^{-4\nu_3/3}\right]\left[\tfrac{1}{4}-\tfrac{1}{4}e^{-4\nu_4/3}\right]\left[\tfrac{1}{4}+\tfrac{3}{4}e^{-4\nu_5/3}\right]$$

$P_{AA}(v_1)$ $P_{AA}(v_2)$ $P_{AC}(v_3)$ $P_{CT}(v_4)$ $P_{CC}(v_5)$

Note use of the AND probability rule

14

Brute force approach would be to calculate L_k for all 16 combinations of ancestral states and sum them

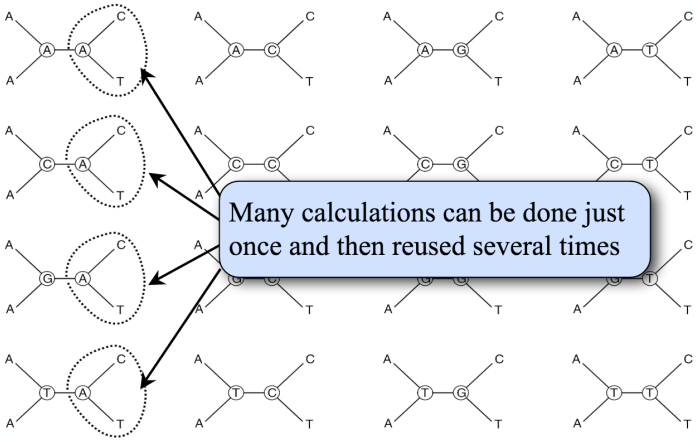


Note use of the OR probability rule

15

Pruning algorithm

(same result, less time)

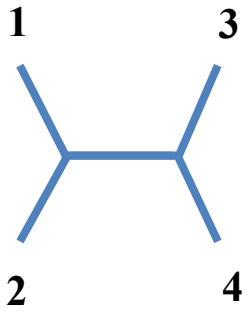


Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**:368-376

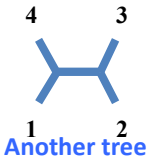
16

Maximum Likelihood tree reconstruction

```
1  AGAGAC
2  AGCGAC
3  CGATTC
4  TGATAG
```



Tree A



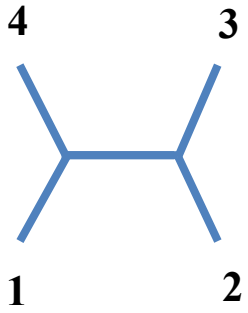
Another tree

What is the likelihood that Tree A (rather than another tree) could have generated the sequence alignment?

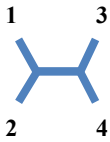
17

Maximum Likelihood tree reconstruction

```
1  AGAGAC
2  AGCGAC
3  CGATTC
4  TGATAG
```



Tree B



Tree A

What is the likelihood that Tree B (rather than another tree) could have generated the sequence alignment?

18

ML analysis in summary

- The likelihood for each site in the alignment for a given topology, branch lengths and the DNA substitution model is calculated
- The probability of the single site is the sum of probabilities of each scenario, taking into account all of the possible nucleotides that may have existed as states at the internal nodes
- The likelihood for a given tree topology for the whole alignment is the product of the likelihoods for each site

Finding the maximum likelihood of a tree

- Problem: the number of possible trees (e.g. for 10 taxa, 2 million unrooted trees possible; for 60 taxa, more possible trees than atoms in the universe!)
  - for each tree topology we need to identify the maximum likelihood estimate for evolutionary parameters and branch lengths
  - then compare the likelihood among all the trees
    - This is simply computationally not feasible
- Solution:
  - Currently no method guarantees finding the best tree
  - Starting tree made usually using MP or NJ
  - Heuristic approaches are used:
    - e.g. NNI = Nearest Neighbour Interchange, SPR = subtree pruning and regrafting, TBR = tree-bisection and reconnection

Typical assumptions of ML substitution models

- The probability of any change is independent of the prior history of the site (a Markov Model)
- Relative frequencies of A, G, C, and T are at equilibrium (stationarity) - S
- Change is time reversible e.g. the rate of change of A to T is the same as T to A - R
- Substitution probabilities do not change with time or over the tree (a homogeneous Markov process) – H
- SRH – we assume that sequence evolution is stationary, reversible, and homogeneous or SRH

How often are the SRH assumptions broken?

The Prevalence and Impact of Model Violations in Phylogenetic Analysis
Suha Naser-Khdour ✉, Bui Quang Minh, Wenqi Zhang, Eric A Stone, Robert Lanfear

Genome Biology and Evolution, Volume 11, Issue 12, December 2019, Pages 3341–3352, https://doi.org/10.1093/gbe/evz193
Published: 19 September 2019 **Article history** ▾

- Review of 35 published phylogenetic datasets
- 23% of them reject the SRH assumptions
- The authors partitioned the data into its various subsets (e.g. 1st, 2nd, 3rd codon positions, nuclear, mitochondrial, introns, exons, etc.)
- SRH tests done on each partition
- In 25% of datasets, tree topology different between the partitions that do and do not violate these assumptions

The Prevalence and Impact of Model Violations in Phylogenetic Analysis

Suha Naser-Khdour ✉, Bui Quang Minh, Wenqi Zhang, Eric A Stone, Robert Lanfear

Genome Biology and Evolution, Volume 11, Issue 12, December 2019, Pages 3341–3352, https://doi.org/10.1093/gbe/evz193
Published: 19 September 2019 **Article history** ▾

Table 2

The Proportion of Partitions That Failed At Least One of the Three Tests—MaxSymTest, $MaxSymTest_{mar}$, and $MaxSymTest_{int}$

Type/Genome	Nuclear	Mitochondrial	Plastid	Virus
First codon positions	20.2%	27.6%	33.3%	25.0%
Second codon positions	21.0%	7.4%	0.0%	25.0%
Third codon positions	76.6%	44.8%	0.0%	75.0%
Other (e.g., intron)	27.8%	100.0%	0.0%	
rRNA	30.0%	25.0%		
UCE	22.5%			
tRNA		0.0%		

- **Take-home message: test your partitions for model violations!**

23

Maximum Likelihood should be seen as a tree estimation procedure instead of a tree reconstruction

"we are making a *best estimate* of an evolutionary history based on incomplete information" Swofford, 1990

24

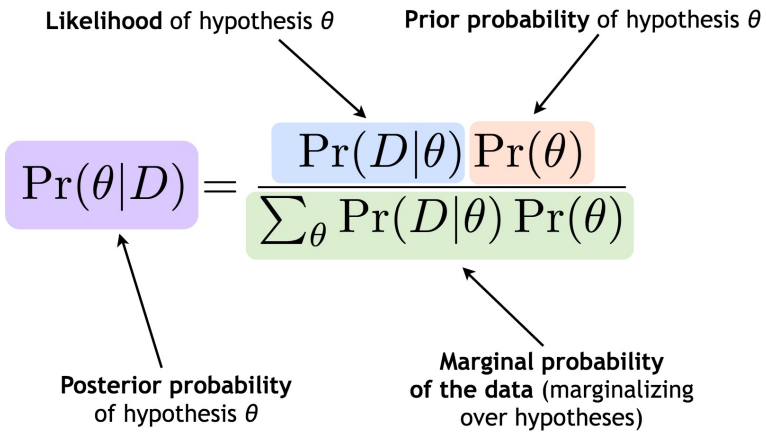
A Bayesian Approach to Phylogenetics

A Bayesian approach compared to ML

- The likelihood is the probability of observing the data given a hypothesis
  - $L = Pr(D \mid \theta)$.
- In ML we search for the parameter values of the model that maximize the likelihood function

- In a Bayesian analysis, we get the probability of a hypothesis given the data (probability of the tree given the sequences)
  - We combine the likelihood of a given hypothesis with a prior expectation for this hypothesis to obtain a posterior probability of the hypothesis

Bayes' rule in statistics

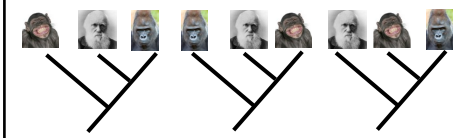


27

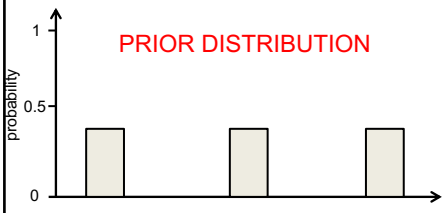
Bayesian inference in general

- D stands for data
- Θ (Gr. theta) means any one of a number of things:
  - a discrete hypothesis
  - a distinct model (e.g. JC, HKY, GTR, etc.)
  - a tree topology
  - one of an infinite number of continuous model parameter values (e.g. ts:tv rate ratio)
- Prior vs. posterior probability
- Posterior probability can be calculated by multiplying the prior probability of a tree (and model parameters) and the likelihood of the observed data (given a tree and its parameters) divided by a normalizing constant

28

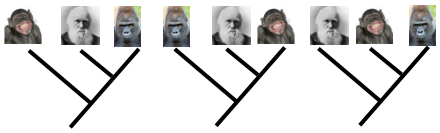


PRIOR DISTRIBUTION



- So-called flat priors for tree topology are used
  - All topologies have the same prior probability
  - Differences in posterior probability result from differences in the likelihood

OBSERVATIONS



POSTERIOR DISTRIBUTION

Slide by Nicolas Chazot (SLU, Uppsala)

29

Major difference between ML and BI

- In ML joint estimation of parameters – likelihood for all parameters estimated at once
  - Likelihood of each parameter depends on likelihood estimation of every other parameter
- In BI marginal estimation – posterior probability of any one parameter is calculated independently
- So even if using flat priors and the same model of DNA evolution, ML and BI could infer different trees because of differences between joint and marginal likelihood estimation

30

Bayes' rule: continuous case

Likelihood

Prior probability ***density***

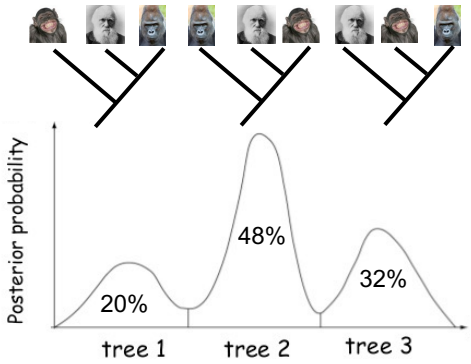
$$f(\theta|D) = \frac{f(D|\theta)f(\theta)}{\int f(D|\theta)f(\theta)d\theta}$$

Posterior probability ***density***

Marginal probability of the data

Paul O. Lewis (2014 Woods Hole Molecular Evolution Workshop) 23

31



Topologies

Branch length

	t1	t2	t3	
vA	0.10	0.07	0.12	0.29
vB	0.05	0.22	0.06	0.33
vC	0.05	0.19	0.14	0.38
	0.20	0.48	0.32	

Joint probabilities

Marginal probabilities

Problem: it is impossible, in most cases,
to derive the posterior probability analytically

or even estimate it by drawing random samples from it

We want something that will "walk" across this parameter space and actively search for the highest point in the parameter "landscape"

Slide by Nicolas Chazot (SLU, Uppsala)

32

Markov chain Monte Carlo (MCMC)

Larget & Simon 1999

33

How does MCMC work? 1/2

A simple summary ☺

- A Markov chain generates a series of random variables
- The probability distribution of future states depends only on the current state (Markov property)
- Phylogenetic inference starts with a randomly generated tree with branch lengths
- Next step is to generate a new tree, based on the previous tree e.g. using tree rearrangements (NNI, SPR, TBR) or changing branch lengths -> this is a proposal
- The proposal is accepted or rejected given a probability based on the Metropolis-Hastings algorithm
  - In practice, it's accepted if it has a better likelihood
- If it is accepted, it becomes the new current state and a new proposal is made

34

How does MCMC work? 2/2

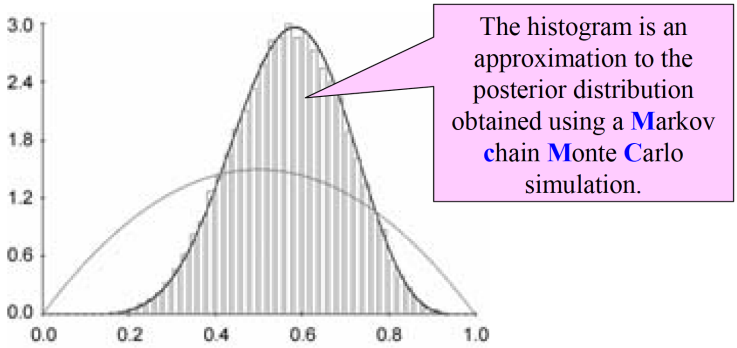
A simple summary ☺

- **Running a Markov chain relatively quickly finds better trees**
- **After a while no better trees can be found and all sampled trees are close to the optimum – "stationary distribution"**
- **The number of times the tree is visited by the chain – interpreted as posterior probability of that tree**

Adapted from Bleidorn (2017) Phylogenomics: An Introduction

35

Markov chain Monte Carlo (MCMC)



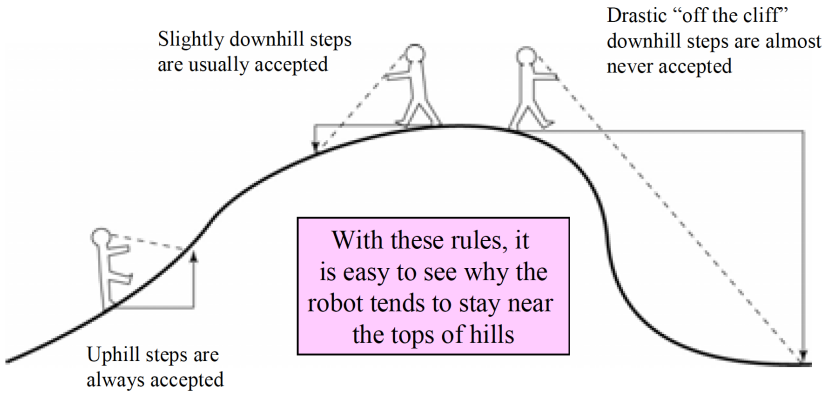
For more complex problems, we might settle for a

good approximation

to the posterior distribution

36

MCMC robot's rules



37

How to decide?

θ = initial position in the parameter space
θ* = new position proposed randomly

$$f(\theta|D) = \frac{f(D|\theta)f(\theta)}{\int f(D|\theta)f(\theta)d\theta}$$

- **Ratio of posterior probabilities R:**

$$\frac{f(\theta^*|D)}{f(\theta|D)} = \frac{\frac{f(D|\theta^*)f(\theta^*)}{\cancel{f(D)}}}{\frac{f(D|\theta)f(\theta)}{\cancel{f(D)}}} = \frac{f(D|\theta^*)f(\theta^*)}{f(D|\theta)f(\theta)} = \mathrm{R}$$

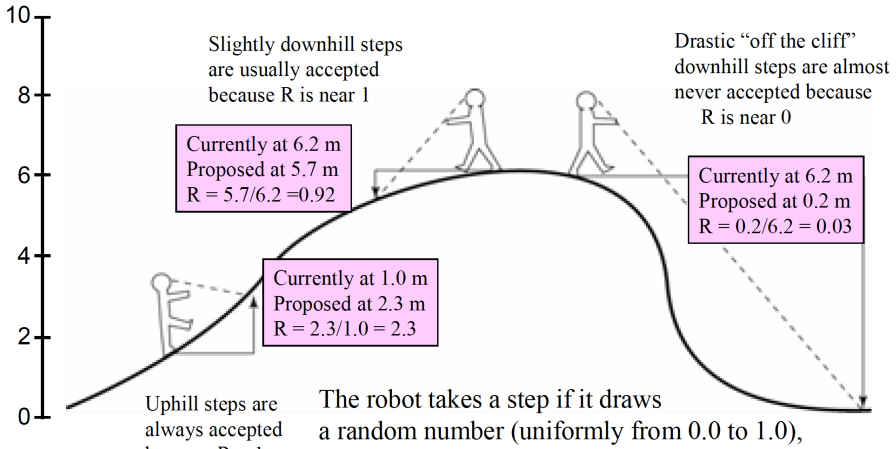
- **Random number between [0,1]:** ***n***

if n ≤ R => new position accepted

if n > R => new position rejected

38

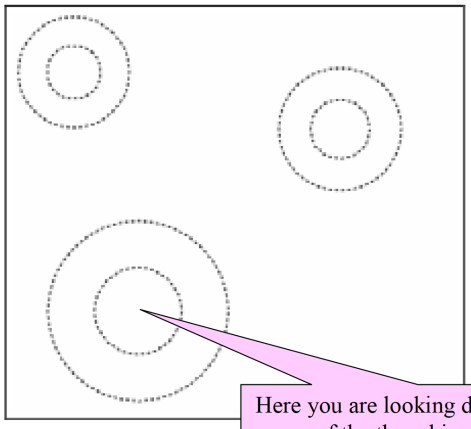
(Actual) MCMC robot rules



The robot takes a step if it draws a random number (uniformly from 0.0 to 1.0), and that number is less than or equal to R

39

What MCRobot can teach us about Markov chain Monte Carlo

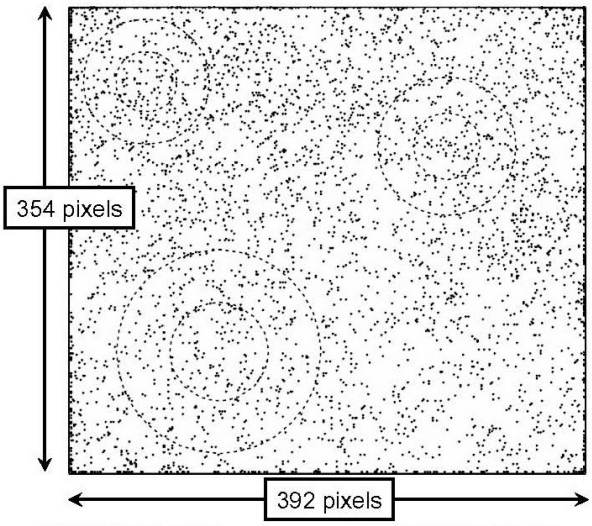


Posterior distribution:
- equal mixture of 3 bivariate normal "hills"
- inner contours: 50%
- outer contours: 95%

Proposal scheme:
- random direction
- gamma-distributed step length
- reflection at edges

40

Pure random walk



Proposal scheme:
- random direction
- gamma-distributed step length (mean 45 pixels, s.d. 40 pixels)
- reflection at edges

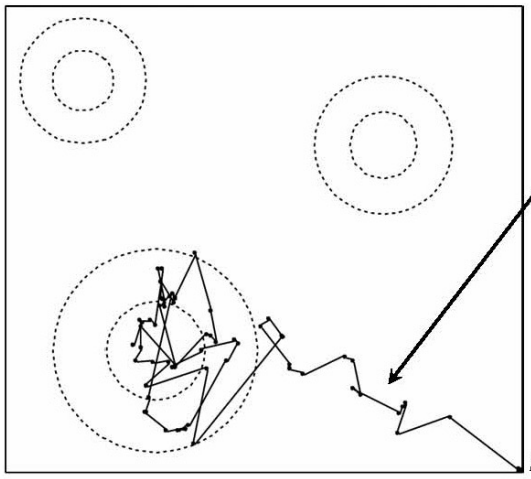
Target distribution:
- equal mixture of 3 bivariate normal "hills"
- inner contours: 50%
- outer contours: 95%

In this case, the robot is accepting every step

5000 steps shown

41

Burn-in



Robot is now following the rules and thus quickly finds one of the three hills.

Note that first few steps are not at all representative of the distribution.

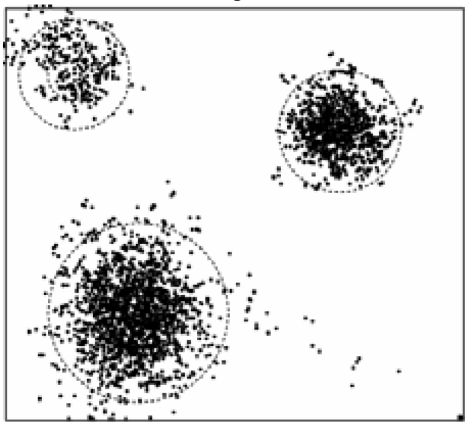
100 steps taken

Starting point

42

Target distribution approximation

5000 steps taken



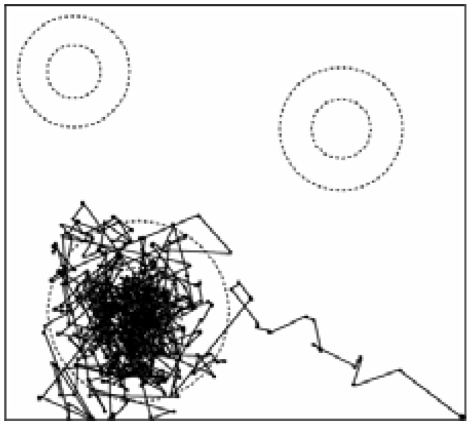
How good is the MCMC approximation?

- 51.2% of points are inside inner contours (cf. 50% actual)
- 93.6% of points are inside outer contours (cf. 95% actual)

Approximation gets better the longer the chain is allowed to run.

Just how long is a long run?

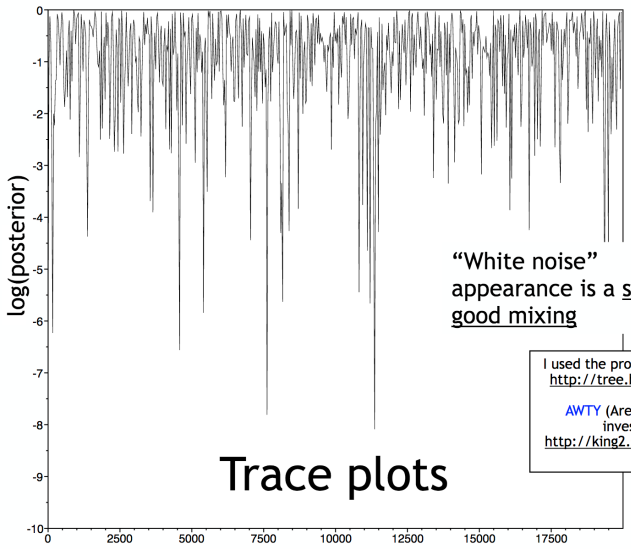
1000 steps taken



What would you conclude about the target distribution had you stopped the robot at this point?

The way to avoid this mistake is to perform several runs, each one beginning from a different randomly-chosen starting point.

Results different among runs? Probably none of them were long enough!

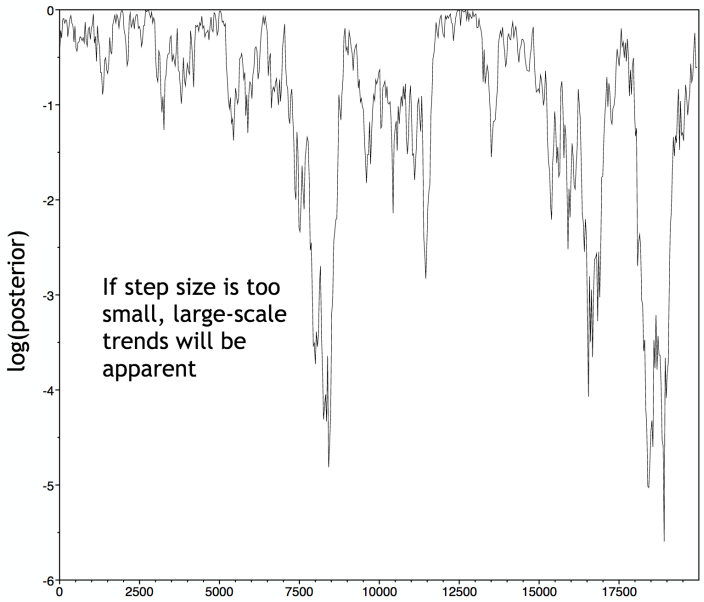


"White noise" appearance is a sign of good mixing

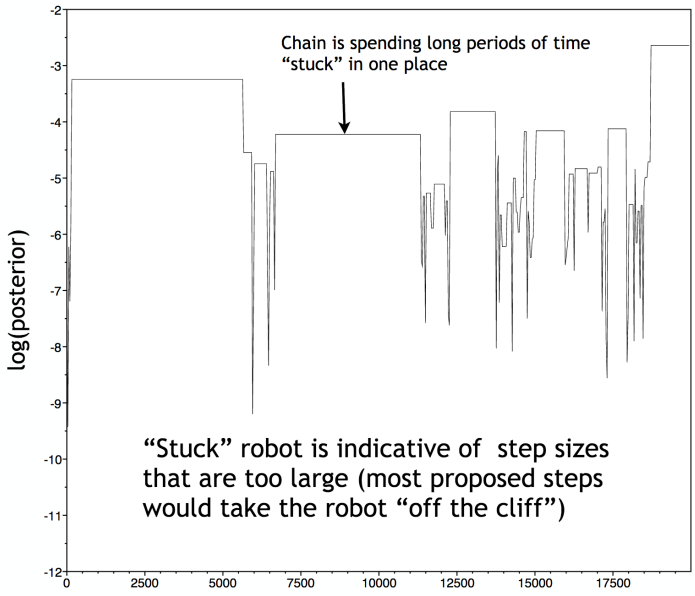
I used the program Tracer to create this plot: http://tree.bio.ed.ac.uk/software/tracer/

AWTY (Are We There Yet?) is useful for investigating convergence: http://king2.scs.fsu.edu/CEBProjects/awty/awty_start.php

Trace plots

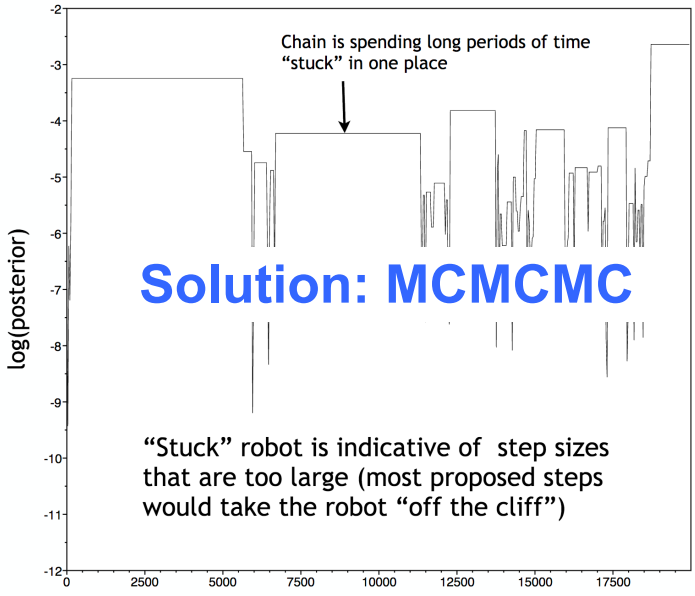


If step size is too small, large-scale trends will be apparent



"Stuck" robot is indicative of step sizes that are too large (most proposed steps would take the robot "off the cliff")

47

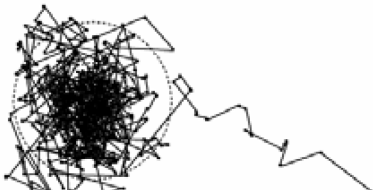


Solution: MCMCMC

"Stuck" robot is indicative of step sizes that are too large (most proposed steps would take the robot "off the cliff")

48

Metropolis-coupled Markov chain Monte Carlo (MCMCMC, or MC^3)



- MC^3 involves running several chains simultaneously (one "cold" and several "heated")
- The cold chain is the one that counts, the heated chains are "scouts"
- Chain is heated by raising densities to a power less than 1.0 (values closer to 0.0 are warmer)

49

Cold vs. heated landscapes

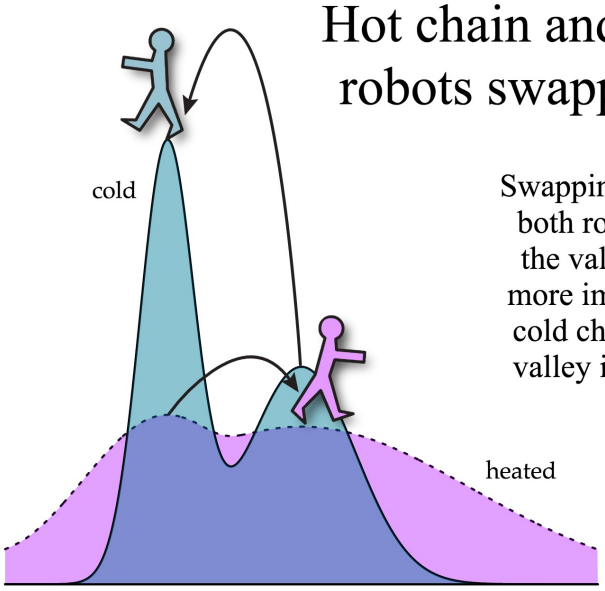
Cold landscape: note peaks separated by deep valleys

Heated landscape: note shallow (easy to cross) valleys

 Bayesian Phylogenetics 36

50

Hot chain and cold chain robots swapping places



Swapping places means both robots can cross the valley, but this is more important for the cold chain because its valley is much deeper

51

MCMC, in short:

- Start with random tree and arbitrary initial values for branch lengths and model parameters
- Each generation consists of one of these (chosen at random):
  - Propose a new tree (e.g. NNI) and either accept or reject the move
  - Propose (and either accept or reject) a new model parameter value
- Every k generations save tree topology, branch lengths and all model parameters (i.e. sample the chain)
- After n generations, summarize sample using histograms, means, credible intervals, etc.

52

How does MCMC work? 1/2

A simple summary ☺

- A Markov chain generates a series of random variables
- The probability distribution of future states depends only on the current state (Markov property)
- Phylogenetic inference starts with a randomly generated tree with branch lengths
- Next step is to generate a new tree, based on the previous tree e.g. using tree rearrangements (NNI, SPR, TBR) or changing branch lengths -> this is a proposal
- The proposal is accepted or rejected given a probability based on the Metropolis-Hastings algorithm
  - In practice, it's accepted if it has a better likelihood
- If it is accepted, it becomes the new current state and a new proposal is made

How does MCMC work? 2/2

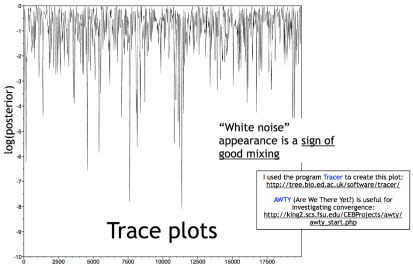
A simple summary ☺

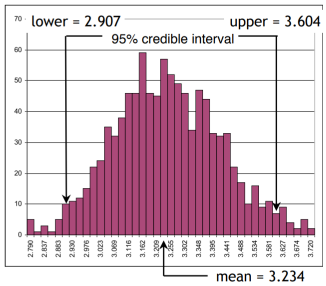
- Running a Markov chain relatively quickly finds better trees
- After a while no better trees can be found and all sampled trees are close to the optimum – "stationary distribution"
- The number of times the tree is visited by the chain – interpreted as posterior probability of that tree

Adapted from Bleidorn (2017) Phylogenomics: An Introduction

Looking at the results

- Graphically





- **Statistically, by computing Effective Sample Size (ESS, minimum should be 200 or more)**
- **Checking convergence of each run, and all runs together**

55

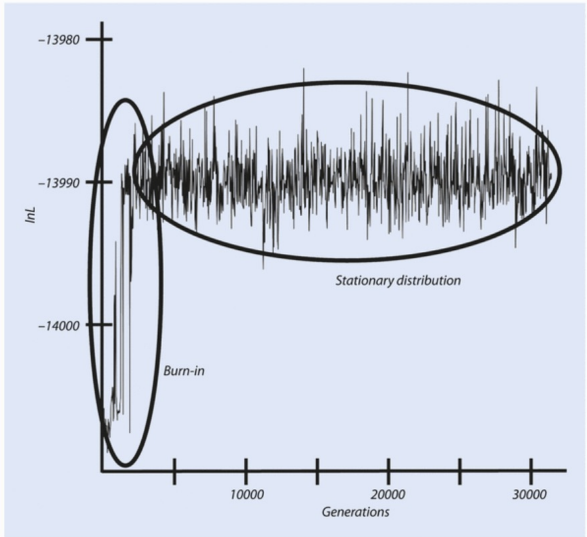
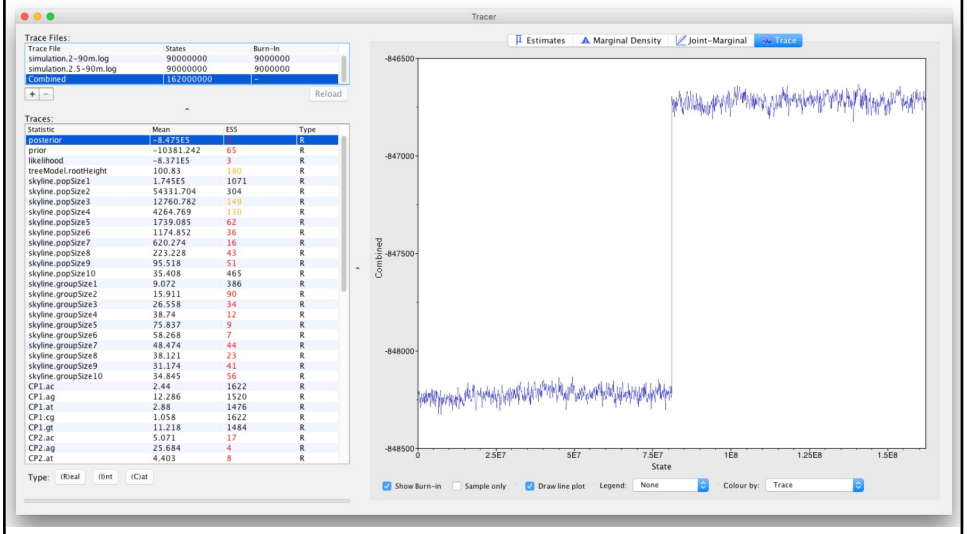


Fig. 8.12 Likelihood scores of a MCMC run plotted against generations. Once stationarity is achieved, trees from this distribution are sampled by discarding all other trees as burn-in. A majority-rule consensus of sampled trees will provide posterior probabilities for every node

Bleidorn (2017) Phylogenomics: An Introduction

56

Lack of convergence



https://beast.community/tracer_convergence accessed on Nov 17. 2021

Summarizing the results

- Autocorrelation: between values that are sampled one after another, so need to sample values at a lower frequency – e.g. every 1000 steps
- MCMC easily runs over millions of states

=> Synthesize/summarize the parameters we are interested in

By computing the marginal posterior distribution of these parameters

- mean, median or variance
- 95% credibility interval

By identifying one or more "best" topologies

e.g. the splits most frequently identified

The number of times a clade in the tree is accepted during the MCMC defines the posterior probability of the clade, and therefore indicates the support for the node

Maximum Likelihood and Bayesian methods: summary

- Both methods are very popular in molecular systematics
- Maximum likelihood is the most important method in phylogenomics
- Bayesian methods are able to take into account uncertainty in parameter estimates
- Bayesian methods can relax the assumption of a homogenous Markov model for rates of change in a tree