Molecular Evolution & Phylogenetics

Heuristics based on tree alterations, maximum likelihood, Bayesian methods, statistical confidence measures

Jean-Baka Domelevo Entfellner

2nd EANBiT residential training
KEMRI-Wellcome Trust, Kilifi, 11 July 2019

biosciences eastern and central africa

EANBiT Eastern Africa Network for Bioinformatics Training

ILRI INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE

Learning Objectives

- know basic tree rearrangements widely used in the literature and in inference programs
- know what is the likelihood of a tree
- understand Maximum Likelihood methods
- understand Bayesian methods
- know about the bootstrap procedures and other techniques to assess the statistical significance of branches in a tree

Learning Outcomes

- be able to run Maximum Likelihood analyses, understanding how it works
- be able to understand the basic parameters of Bayesian inference methods
- be able to interpret the supports on branches output by phylogenetic inference programs

Strategies in the quest for "the best" phylogenetic tree: browsing the space of topologies

Necessity for guided tree transformations

- Tree inference problem is essentially an **optimization problem**: find the tree that maximizes/minimizes a certain criterion
- Remember: (2*n*-5)!! unrooted binary tree topologies with *n* taxa.

  ⇒ looking for “the best tree”, one cannot just try all of them and calculate e.g. the number of parsimony steps for each one.

  ⇒ necessity to **guide the search** with a certain criterion or set of criteria, and to develop **heuristics** to decide which tree to try next

Pseudo-random walk

⇒ phylogenetic inference software implement **pseudo-random walks** in the space of tree topologies, **trying different topologies**

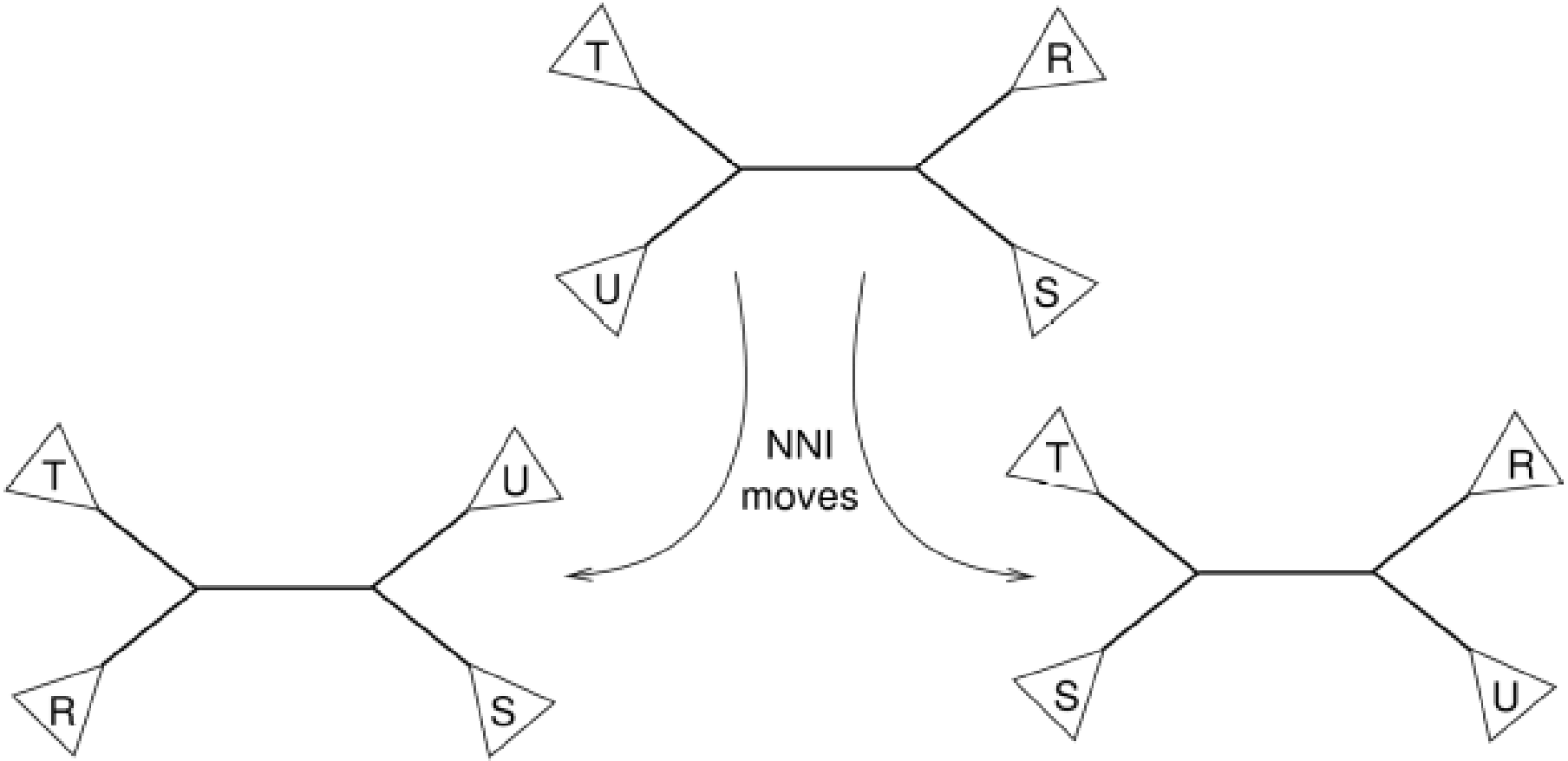
⇒ we go from one tree to the next with **elementary tree alterations**: NNI, SPR or TBR moves

⇒ **iterative trial and error process**: we try one tree, calculate the corresponding parsimony cost, then try improve it with an elementary tree alteration (e.g. picking the move leading to largest improvement), calculate new cost, etc

⇒ **pseudo-random process**: try random alterations, **conserve** alteration if it is an improvement, otherwise **drop** it (backtracking) before attempting another alteration

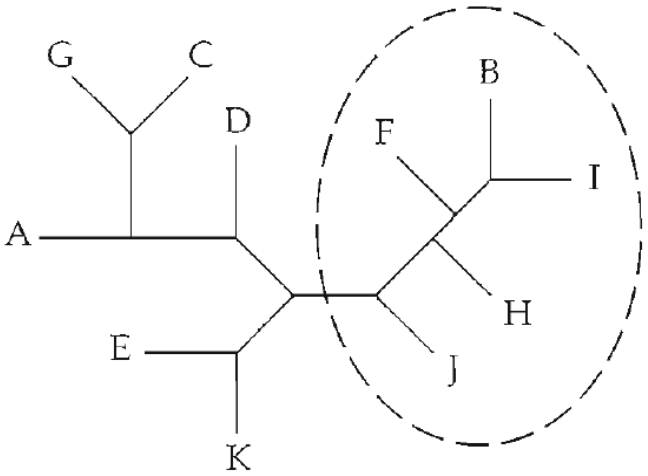
NNI: Nearest Neighbour Interchange

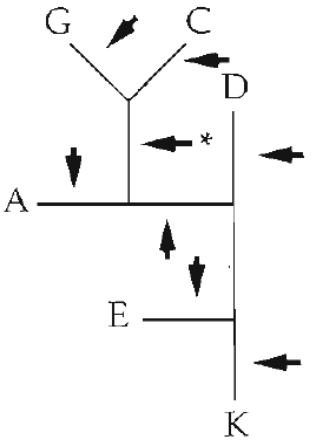
- NNI is a local rearrangement **swapping two** of the four **subtrees** connected to a given internal branch.

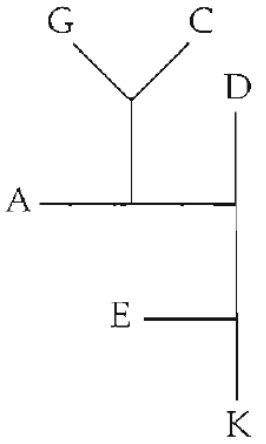


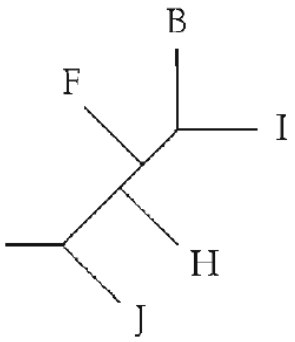
SPR: Subtree Pruning and Regrafting

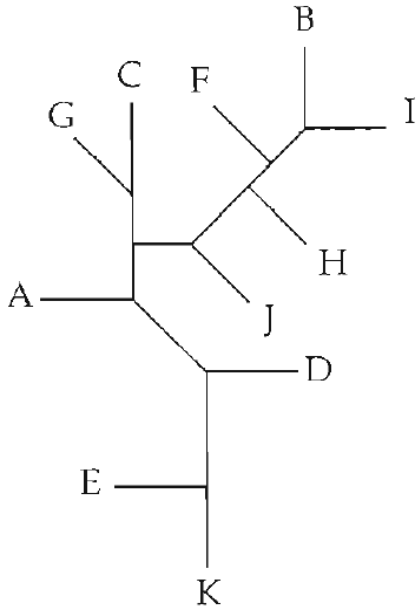
SPR is a “less local” rearrangement **pruning** a subtree and **regrafting** it onto any of the branches of the tree (here, the edge marked with a star).







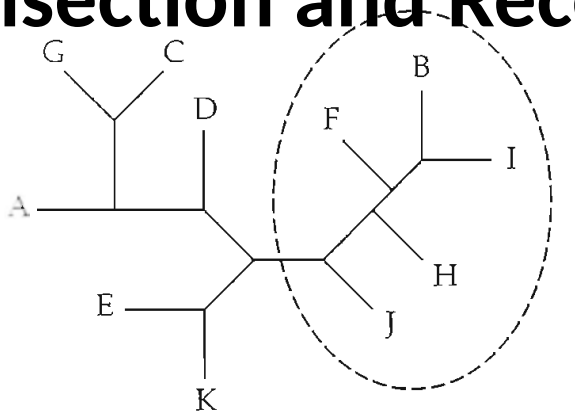


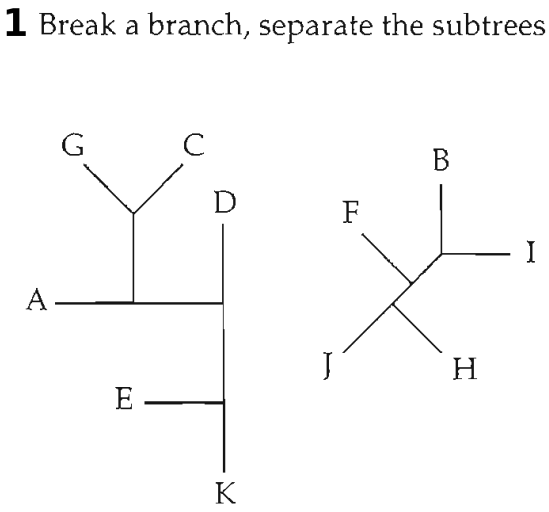


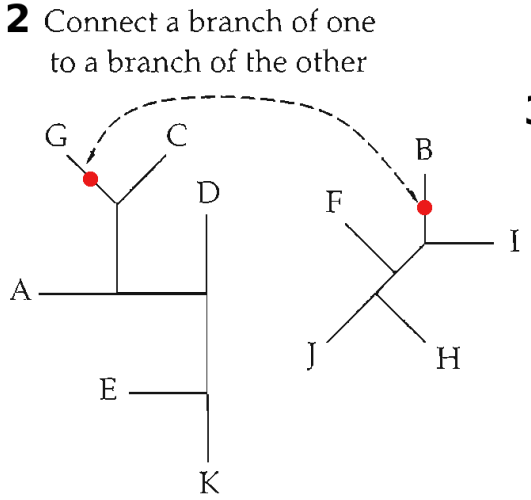
TBR: Tree Bisection and Reconnection

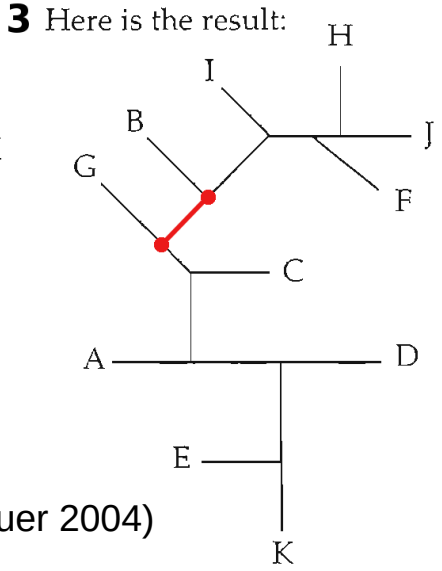
TBR is more involved a rearrangement **bisecting** a tree into two subtrees and **reconnecting** them by joining together **any** branch of the one tree with **any** branch of the other tree (here, marked in red).

⇒ SPR moves form a subset of TBR moves









source: *Inferring Phylogenies*, J. Felsenstein (Sinauer 2004)

The likelihood of a tree

Likelihood: definition

Given some **observed data D**, the likelihood (FR: “vraisemblance”) of a **model M** is the probability that the observations originate from that (generative) model:

$$Lk(M) = Pr(D|M)$$

In phylogenetics, the observations D are the traits corresponding to the taxa on the leaves (e.g. alignment of nucl. or a.a. residues) and the **model M** encapsulates:

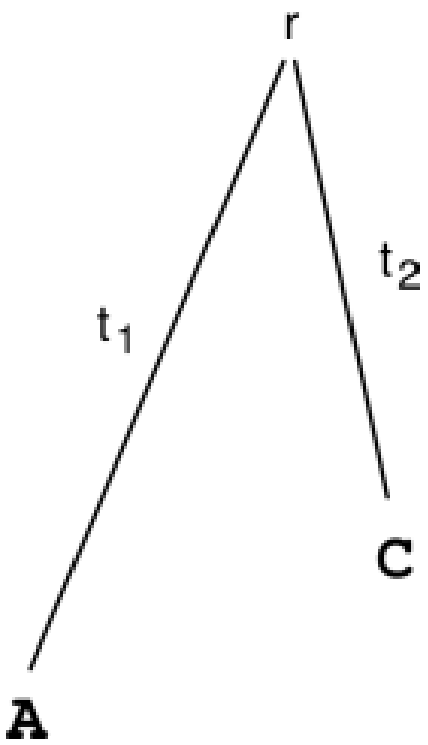
- a **tree topology**;
- the corresponding **branch lengths**;
- an evolutionary **substitution model** (matrix Q of instantaneous substitution rates) believed to be acting on the branches of the tree.

Likelihood calculations

$$Lk(T) = \sum_{i \in \{A,C,G,T\}} Pr(r=i)\ Pr\left(i \xrightarrow{t_1} A\right)\ Pr\left(i \xrightarrow{t_2} C\right)$$

$$= \sum_{i \in \{A,C,G,T\}} \pi_i \left[e^{Qt_1}\right]_{(i,A)} \left[e^{Qt_2}\right]_{(i,C)}$$

$$\ll 1$$



Based on this, a **recursive procedure** enables us to calculate the likelihood for more complex trees, based on **partial likelihood vectors** calculated on the subtrees (**Felsenstein's pruning algorithm**).

Total likelihood value: sites seen as independent

$$Lk(T) = \prod_{site\ s} Lk_s(T) = \prod_{site\ s} Pr(s|T)$$

and because likelihood values are usually very small, computer programs use their logarithms:

$$logLk(T) = \sum_{site\ s} logLk_s(T)$$

Typical log likelihood values for reasonably sized ML trees:
-3000 (small tree), -12000, -63000, etc.

Maximum likelihood

Maximum Likelihood (ML) framework

Under the Maximum Likelihood framework, one tries to find the tree with highest likelihood (optimisation problem), i.e. the model M* such that:

$$Lk(M^*) = \max_M \left(Lk(M) \right) = \max_M \left(Pr(D|M) \right)$$

This implies to try several trees, with heuristics including tree alterations (NNI, SPR, TBR).

⇒ **No guarantee** to find **the best** tree!

⇒ Some popular ML software: PAML (Ziheng Yang), **PhyML** (Guindon/Gascuel), **RAxML** (Stamatakis)

Extra sugar for Maximum Likelihood methods: modeling varying rates among sites, testing models

Modeling rate variation across sites

- Naive models: constant rate of evolution across sites (1 change per unit branch length)
- Not consistent with biology (codon positions, sites under evolutionary pressure…)
- If evolutionary speeds vary across sites, we want to average the relative rates to 1
- E.g. relative rates {2.0, 1.0, 3.0, 2.4, 0.8} become {1.09, 0.54, 1.63, 1.30, 0.43} ⇒ average rate = 1
- Site evolving at rate *r* along branch of length *t* same as evolving at rate 1 along branch of length *rt*.

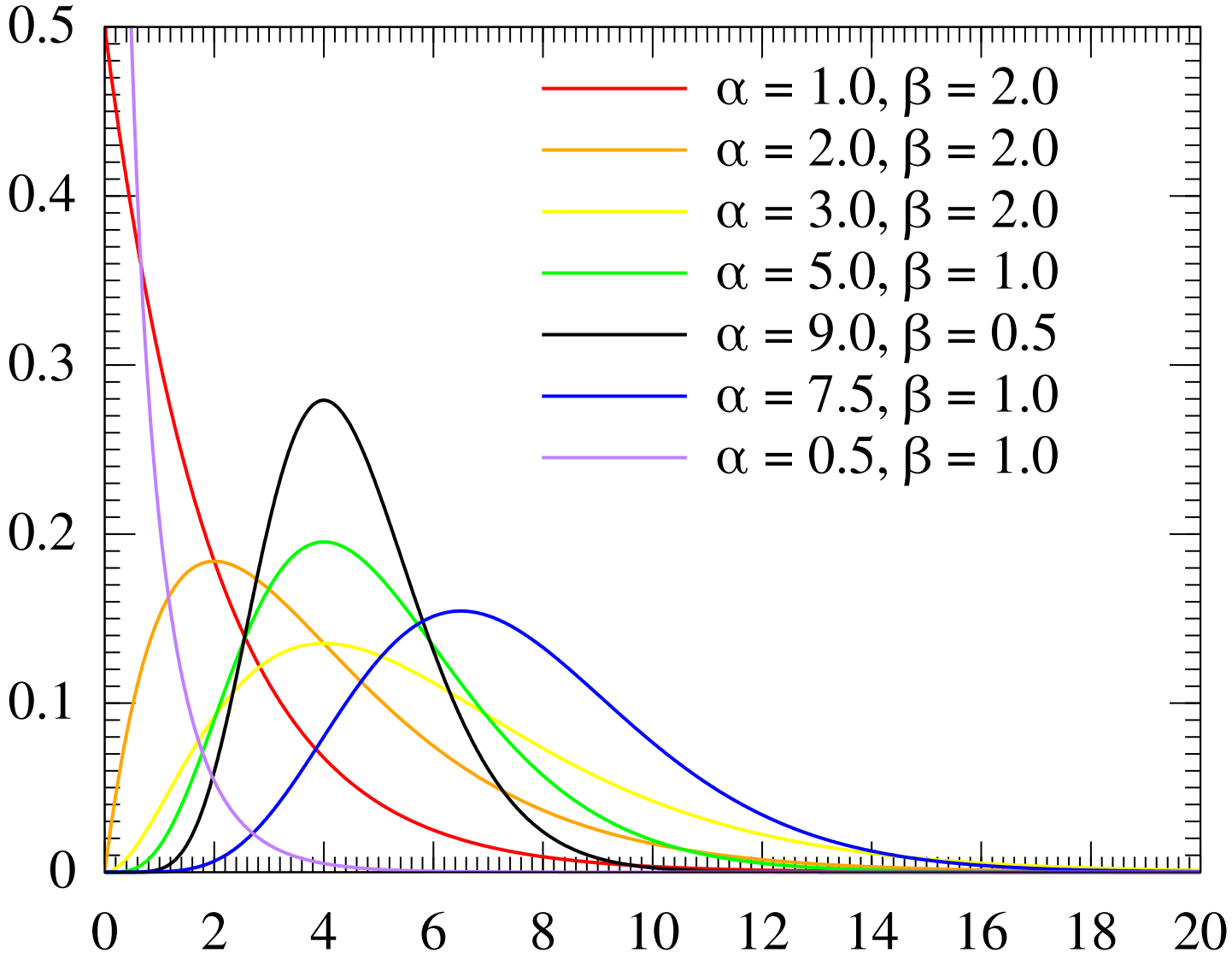
Estimating the rate for every site

- New formula for the elementary likelihood:

$$Lk\left(A \overset{t}{-} C\right) = \int_0^{\infty} \pi_A \, Pr\left(\text{rate} = r\right) Pr\left(A \overset{rt}{\rightarrow} C\right) dr$$

- Difficult to perform full integration, so we approximate with discrete rates r_i drawn from a Gamma distribution (mathematical convenience)
- Density of a Gamma distribution with parameters α and β (distribution mean is αβ): $f(r) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} r^{\alpha-1} e^{-r/\beta}$
- After choosing the number of rate categories, automatic way to set the rates. αβ=1 ⇒ only 1 extra param. (α)

Gamma distributions with different parameters



Comparing models

- More parameters imply more flexibility in the model
- More flexibility ⇒ a better fit to the data (higher likelihood for the model)
- BUT too many parameters estimated from the data is **overfitting**!
- When comparing models, one should assess whether the more complex model really brings significant improvement (likelihood increase), with the need to compensate for the number of free parameters (penalty)

Basic Likelihood Ratio Test (LRT)

- Standard procedure to compare two models
- H_0: simpler model (q parameters)
- H_1: more involved model ($p>q$ parameters)
- **Nested** models: H_0 must be a specialization of H_1
- LRT statistic:

$$2\Delta l = 2\log\left(\frac{Lk(H_1)}{Lk(H_0)}\right) = 2\left(logLk(H_1) - logLk(H_0)\right)$$

- Under H_0 and if large sample (data) size, $2\Delta l \sim \chi^2_{p-q}$

Akaike Information Criterion (AIC)

- Standard procedure to assess the "amount of accurate information" in a model (Akaike 1974)
- Let *M* be a model with *p* free parameters
- Let logLk(*M*) be the optimum log likelihood of *M*.
- AIC statistic: $\mathrm{AIC}(M) = -2\log\mathrm{Lk}(M) + 2p$
- Models with lower AIC are preferred (extra param. worth it if it improves the logLk by one unit).
-

Akaike Information Criterion, corrected

- Was seen that the AIC doesn't penalize enough for the increased number of parameters
- Corrected AIC (Sugiura 1978, Hurvich and Tsai 1989):

$$\mathrm{AIC_c}(M) = -2\log\mathrm{Lk}(M) + \frac{2np}{n-p-1} = AIC(M) + \frac{2p(p+1)}{n-p-1}$$

- One chooses the model with lowest AIC_c

Bayesian Information Criterion

- Even more severe penalization of parameter-rich models (Schwarz 1978)

$$\mathrm{BIC}(M) = -2\,\mathrm{logLk}(M) + p\log(n)$$

- The lower the BIC, the better.
- *n* is the *sample size*: depends mostly on the number of unique sites in the alignment and how much they are correlated

Bayesian inference

Bayes' theorem

Linking conditional, marginal and joint probabilities:

$$Pr(A,B) = Pr(A|B)Pr(B) = Pr(B|A)Pr(A)$$

so:

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

and applied to phylogenies:

likelihood

prior probability of the model

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

posterior probability

Bayes' theorem: denominator

Pr(D) (probability of the data) is a sum of joint probabilities over all models:

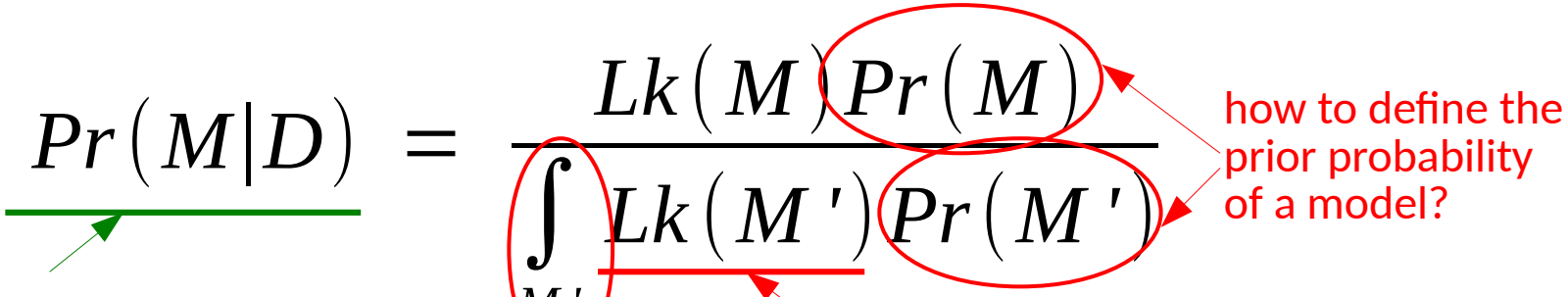
$$\begin{aligned} Pr(D) &= Pr(D, M_1) + Pr(D, M_2) + Pr(D, M_3) + \ldots \\ &= \sum_{M'} Pr(D, M') \\ &= \sum_{M'} Pr(D|M') Pr(M') \end{aligned}$$

But the space of all models (trees) is continuous:

$$Pr(D) = \int_{M'} Pr(D|M') Pr(M')$$

Bayes' theorem: challenges arising

The final Bayes' formula for phylogenetic inference is:



how to define the prior probability of a model?

posterior probability.
what we want:
to find the model
with highest
posterior probability

how to integrate
over an infinite space
of phylogenetic models?

huge number of likelihood calculations involved in the calculation of a single posterior probability: very computationally intensive

Priors on trees

Several strategies can be used to define prior probabilities on phylogenetic trees:

- flat ("uninformative") priors: all trees have same prior (uniform distribution)
- birth-death markovian process of speciation
- prior using an arbitrary distribution on branch lengths, etc

The frequentist/ML viewpoint: no prior is fully satisfactory.

The Bayesian viewpoint: priors don't matter that much.

Sampling the space of trees

Remember the denominator $\int_{M'} Lk(M')Pr(M')$?

Bayesian methods require to sample **intensively** the space of all phylogenetic models and calculate the corresponding likelihoods.

Idea of the Markov Chain Monte Carlo (MCMC) methods: wander at random and long enough in the likelihood landscape, covering as best as possible the areas of "good-ish" likelihood, to get ultimately a reasonable image of the whole likelihood landscape.

An MCMC example: Metropolis-Hastings

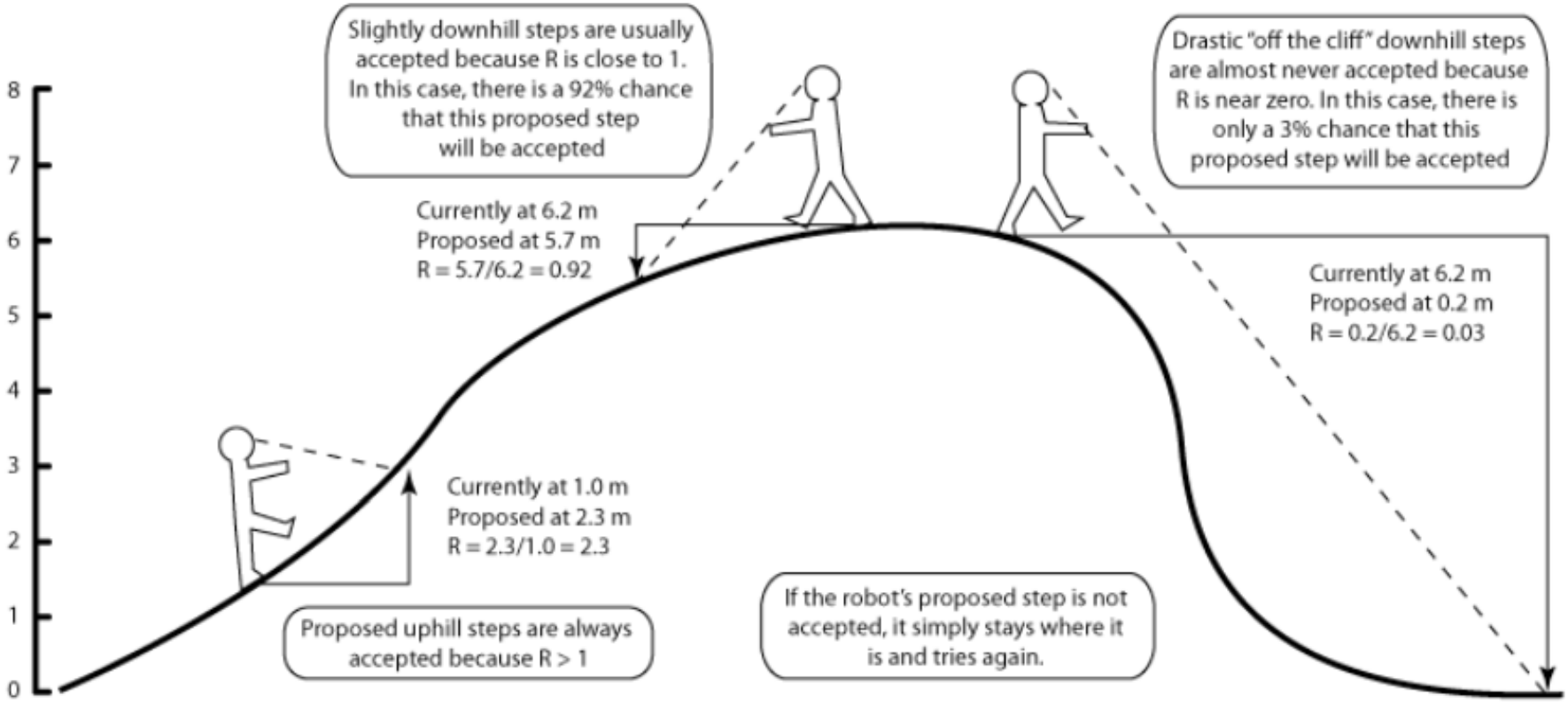
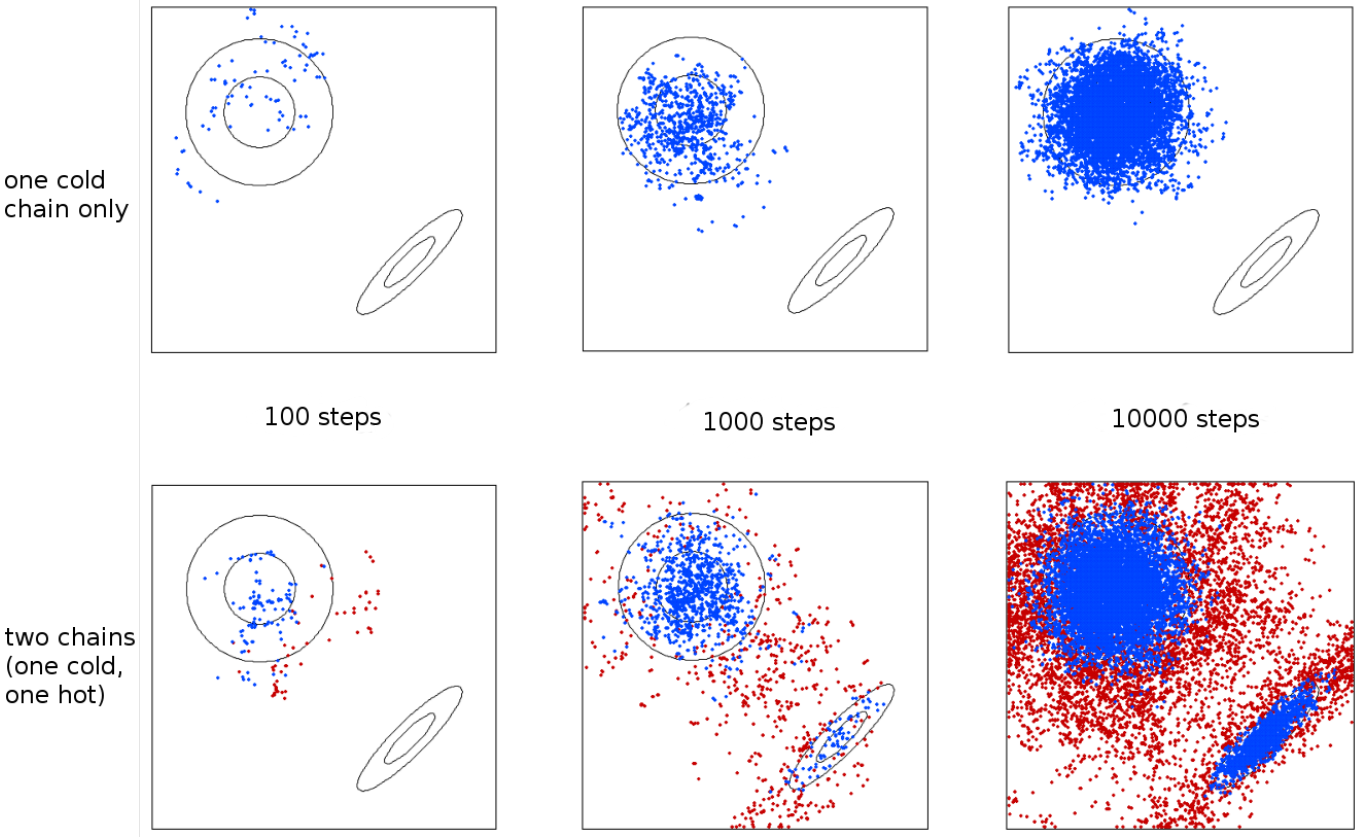


Illustration of MCMC method process (Lewis, 2011)

Some MCMC refinements

- multiple chains with different starting points (initial tree)
- multiple chains with different expected leap length (cold/hot chain)
- swapping a cold chain with a hot chain at certain random times (explore more thoroughly the whole space)
- burnout sequence of *n* initial trees discarded



Bayesian methods: summary

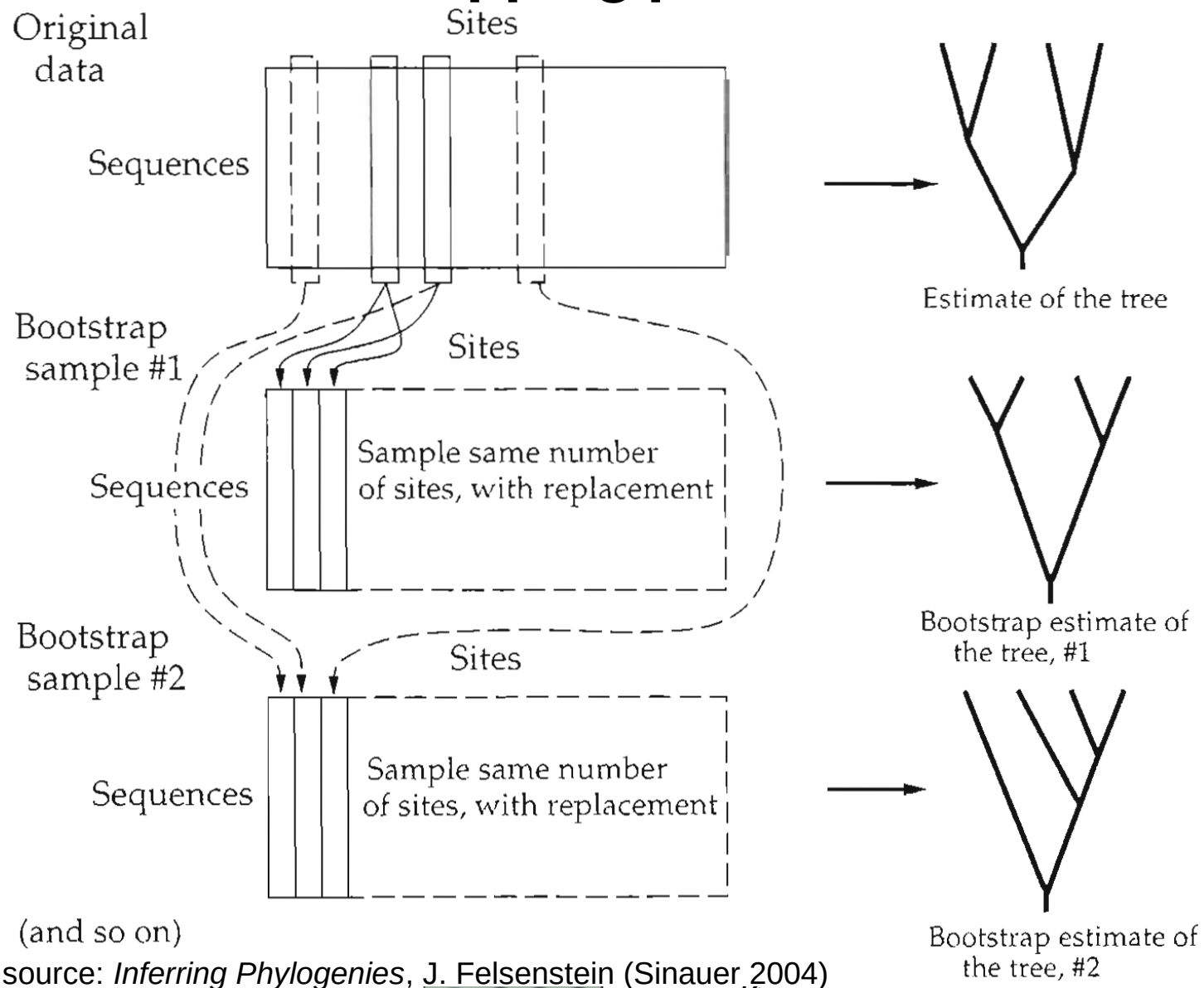
- Bayesian methods represent arguably the most elaborate way to infer phylogenies
- they aim at maximizing the **posterior probability** of a model rather than its likelihood
- they are very **computationally intensive** (common to have jobs running for weeks on relatively large datasets)
- some famous software for Bayesian inference: PAML (Yang/Rannala), **MrBayes** (Huelsenbeck/Ronquist), PhyloBayes (Lartillot/Rodrigue), **BEAST** (Drummond/Rambaut)

Assessing confidence on the branches: bootstrap et al.

Bootstrapping procedure

- Idea: the specific “best tree” we got is a function of the alignment we fed the inference process with.
- different sequence alignments on the same taxa would they lead to alternate trees? Probably…
- we can **resample with replacement from the set of sites** composing the input alignment and infer the “best tree” corresponding to that resampling
- perform this iteratively and then compare all the **bootstrap trees** you got with your original best tree.
- the **statistical support** of a branch is the **proportion of bootstrap trees** containing that split (split = branch)

Bootstrapping procedure



source: *Inferring Phylogenies*, J. Felsenstein (Sinauer 2004)

Alternative statistical supports: likelihood ratios

- Other idea: a branch AB|CD is “certain” if the tree containing it has a much better likelihood than the trees obtained by NNI on that branch to include either AC|BD or AD|BC

- alternatively: likelihood ratio between the phylogeny including the branch and the (non binary) phylogeny having a multifurcation there (branch of length 0)
  → aLRT statistics (Anisimova & Gascuel 2006)