Workshop in Molecular Evolution

Jan 7 - 11, 2019

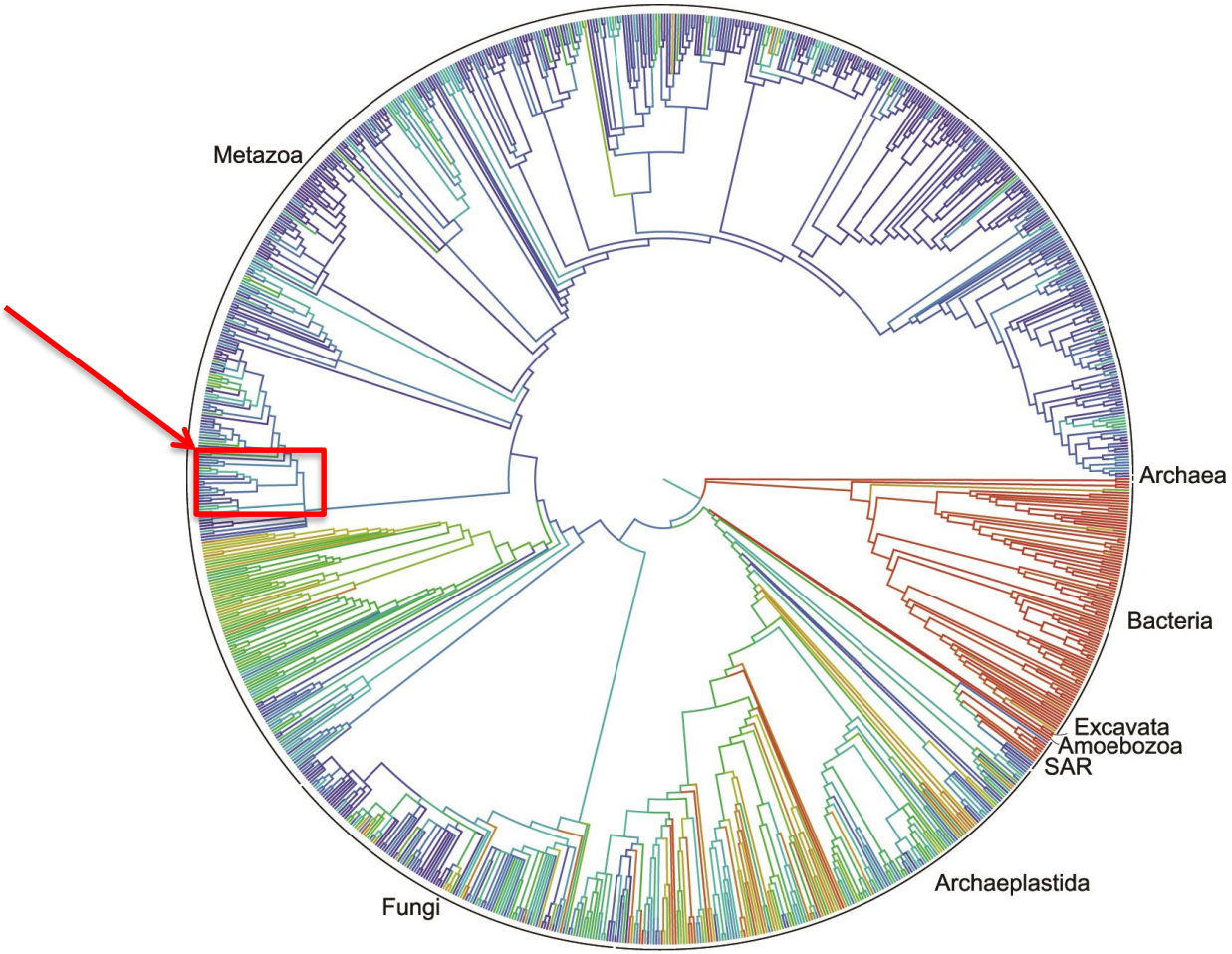
Shanghai

Workshop on Molecular Systematics and Evolution

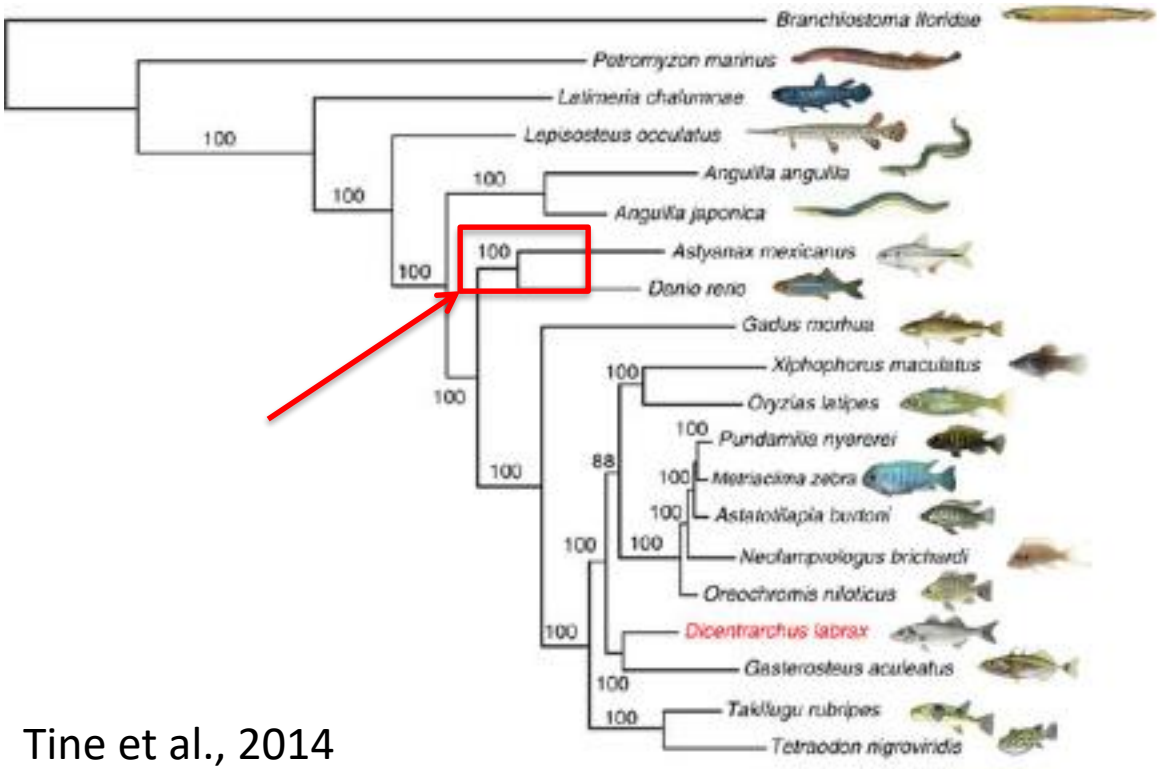
The lab of Molecular Systematics & Ecology
Shanghai Ocean University, Lingang, Shanghai, Jan 7-11, 2019

	Jan 7 (Mon.)	Jan 8 (Tue.)	Jan 9 (Wend.)	Jan 10 (Thur.)	Jan 11 (Fri.)
8:30 – 9:00 9:00 – 9:15 *discussion*	EvolMarkers2 *Junman Huang*	Date filtering, partition *Hao Yuang*	Population structure, AMOVA, PCA *Qingwen Xue*	F-dist, Bayescan, Other adaptive methods *Longlong & Suhan*	Intro AI *Liang Lu*
Tea Break					
9:30 – 10:00 10:00 – 10:15 *discussion*	Lib prep & gene cap *Lifang Peng*	Gene tree, species tree *Guoxin Yin*	Spatial structure, population dynamics *Huirui & Ying*	GWAS introduction, improved method *Ziqiang Gong*	Convolutional network *Liang Lu*
Tea Break					
10:30 – 11:00 11:00 – 11:15 *discussion*	Read assembling *Junman Huang*	Time calibration, topology test *Guoxin Yin*	Species delimitation *Lei & Songjun*	Environment GWAS, pedigree deducing *Ziqiang Gong*	GANs *Hao Yuan*
Lunch					
1:30 – 2:00 2:00 – 2:15 *discussion*	Post assembling data processing *Hao Yuan*	Biogeography, character mapping *Yinyi Yang*	ABC *Anirban Sarker*	Transcriptomic analysis *Tao Zhou*	Protein folding *Liang Lu*
Tea Break					
2:30 – 3:00 3:00 – 3:15 *discussion*	Molecular evolution *Chenhong Li*	SNP calling SNPs vs sequences *Qiaoyun Ai*	Fastsimcoal2 *Lifang Peng*	Comparative genomics, EP *Hao Yuan*	Genome prediction *Hao Yuan*
Tea Break					
3:30 – 4:00 4:00 – 4:15 *discussion*	Population genetics *Chenhong Li*	Summary statistics, Arlequin *Qiaoyun Ai*	Land markers *Qiaoyun Ai*	Open: How to identify phenotype associated genes	Open: Ideas applying AI

Two faces of one process: phylogenetics vs. population genetics

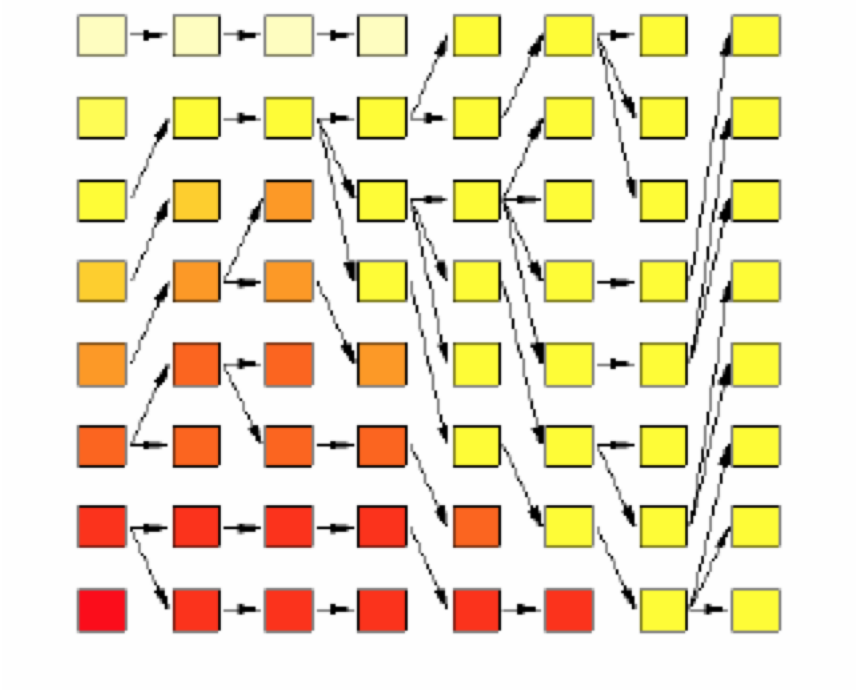


Phylogenetics – model of speciation

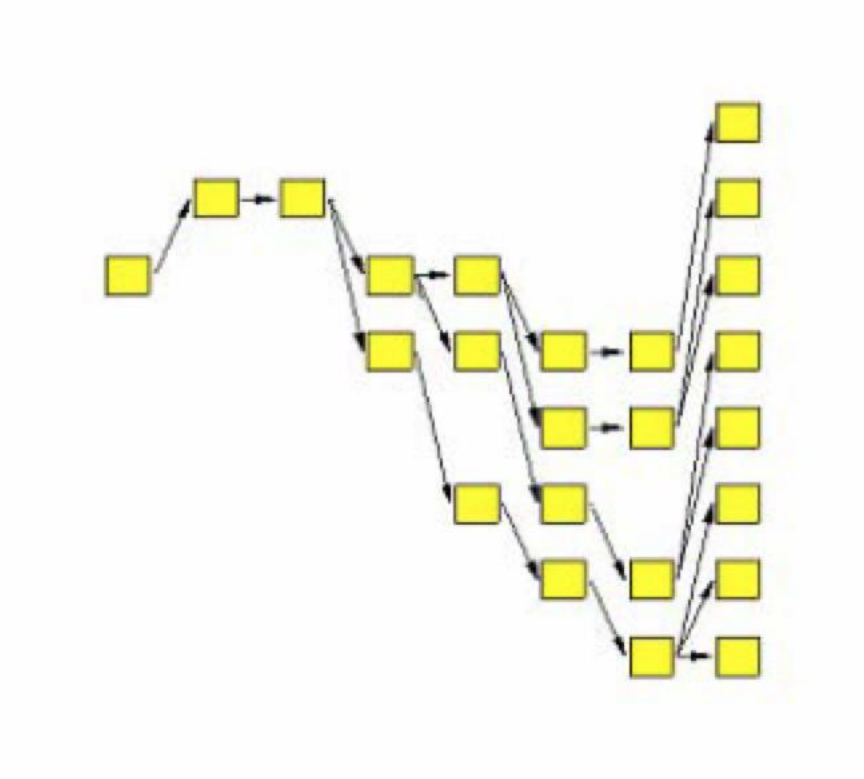


Tine et al., 2014

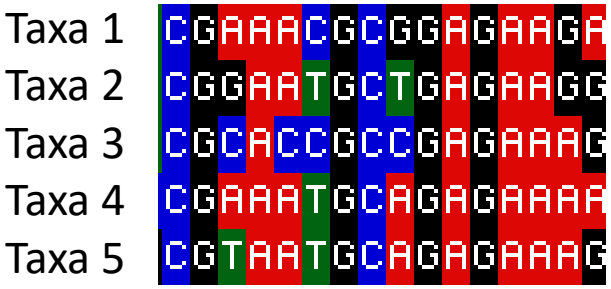
Population genetics – model of coalescence



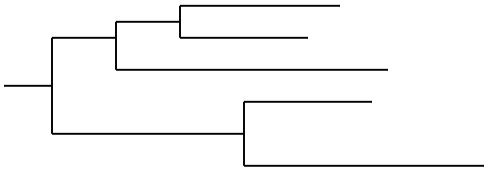
Population genetics



Null model of phylogenetics



- Topology and branch length



- Substitution matrix

  r_{TC} (= r_{CT}), r_{TA} (= r_{AT}), r_{TG} (= r_{GT})
  r_{CA} (= r_{AC}), r_{CG} (= r_{GC})
  r_{AG} (= r_{GA})

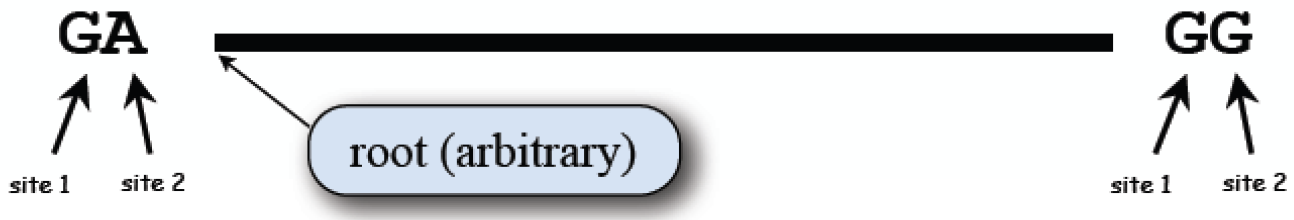
- Stationary base frequencies

  f_T, f_C, f_A, f_G,

Likelihood of the simplest tree

sequence 1 ——— sequence 2

To keep things simple, assume that the sequences are only 2 nucleotides long:



$$
\begin{aligned}
L &= L_1 \, L_2 \\
&= \left[\left(\frac{1}{4}\right) \left(\frac{1}{4} + \frac{3}{4}e^{-4\alpha t}\right)\right] \left[\left(\frac{1}{4}\right) \left(\frac{1}{4} - \frac{1}{4}e^{-4\alpha t}\right)\right]
\end{aligned}
$$

Pr(G) Pr(G|G, αt) Pr(A) Pr(G|A, αt)

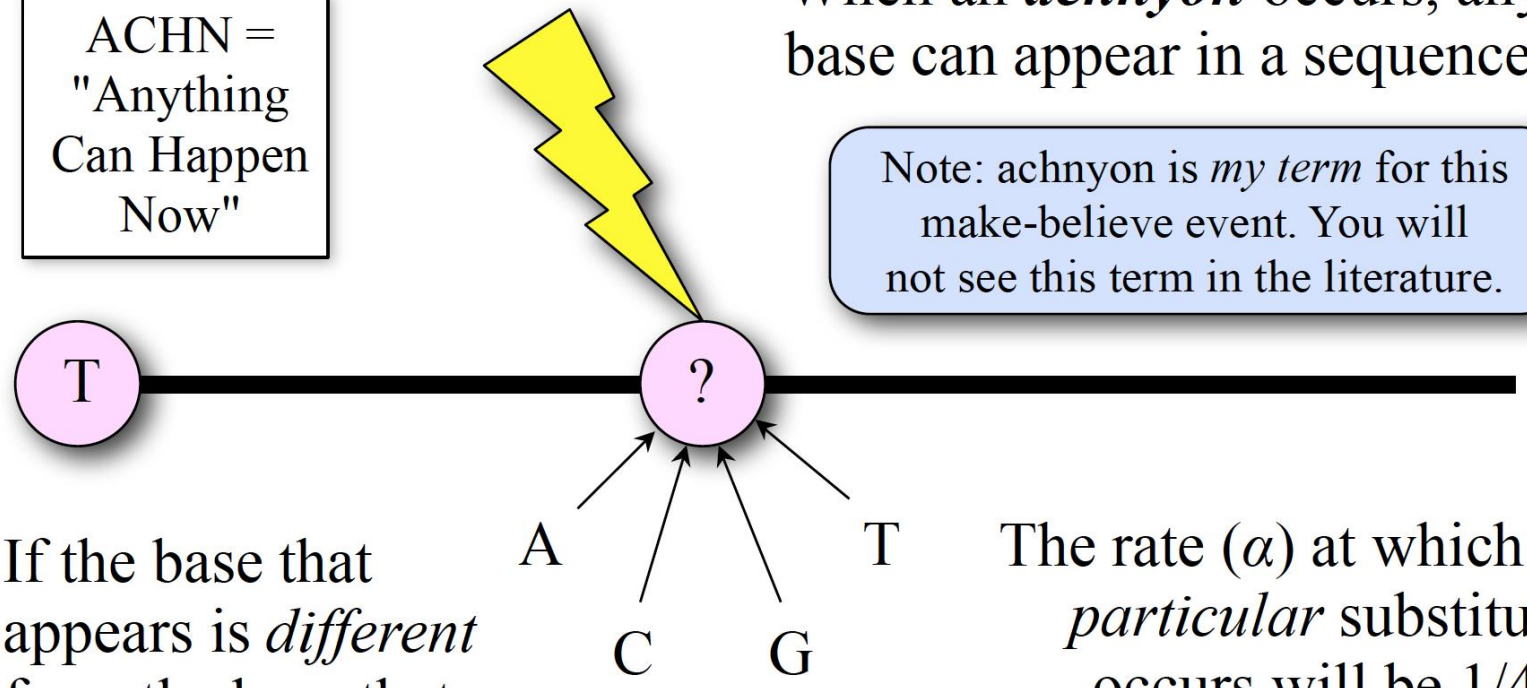
Note that we are NOT assuming independence here

"ACHNyons" vs. substitutions

ACHN = "Anything Can Happen Now"

When an ***achnyon*** occurs, any base can appear in a sequence.

Note: achnyon is *my term* for this make-believe event. You will not see this term in the literature.



If the base that appears is *different* from the base that was already there, then a **substitution** event has occurred.

The rate (α) at which any *particular* substitution occurs will be 1/4 the achnyon rate (μ). That is, $\alpha = \mu/4$ (or $\mu = 4\alpha$)

Deriving a transition probability

Calculate the probability that a site currently T will change to G over time t when the rate of this particular substitution is α:

Pr(zero achnyons) = $e^{-\mu t}$ (Poisson probability of zero events)

Pr(at least 1 achnyon) = $1 - e^{-\mu t}$

Pr(last achnyon results in base G) = $\frac{1}{4}$

Pr(end in G | start in T) = $\frac{1}{4}\left(1 - e^{-\mu t}\right)$

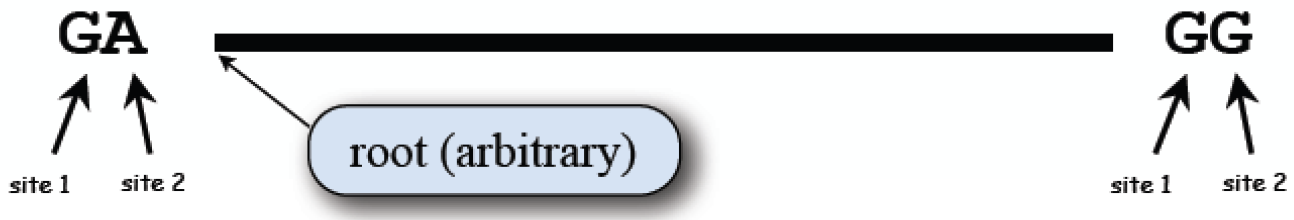
Remember that the rate (α) of any particular substitution is one fourth the achnyon rate (μ):

$$P_{GT}(t) = \frac{1}{4}\left(1 - e^{-4\alpha t}\right)$$

Likelihood of the simplest tree

sequence 1 ——— sequence 2

To keep things simple, assume that the sequences are only 2 nucleotides long:

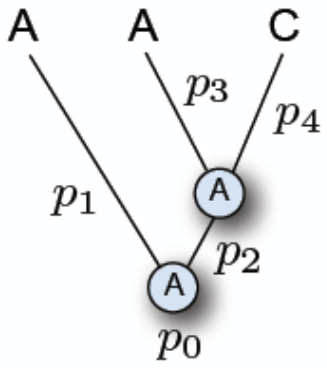


$$
\begin{aligned}
L &= L_1 \, L_2 \\
&= \left[\left(\frac{1}{4}\right) \left(\frac{1}{4} + \frac{3}{4}e^{-4\alpha t}\right)\right] \left[\left(\frac{1}{4}\right) \left(\frac{1}{4} - \frac{1}{4}e^{-4\alpha t}\right)\right]
\end{aligned}
$$

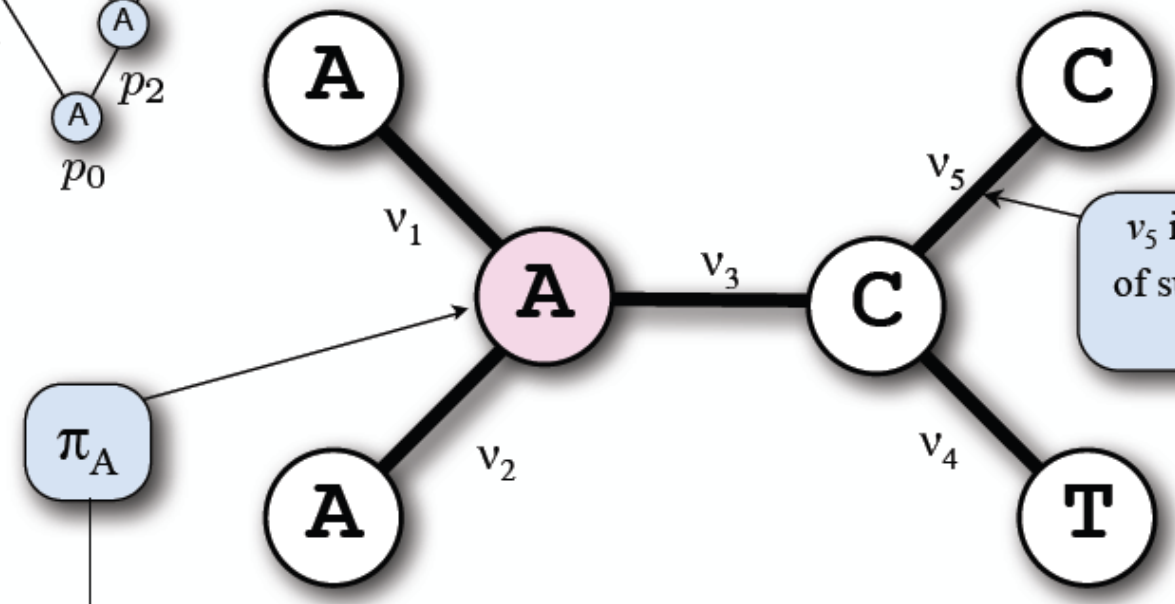
Pr(G) Pr(G|G, αt) Pr(A) Pr(G|A, αt)

Note that we are NOT assuming independence here

From slide 6



Likelihood for site k



v_5 is the expected number of substitutions for just this one branch

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

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

Note use of the AND probability rule

i.i.d. assumption

- Each site evolves independently and according to the identical process, so called "i.i.d." process.

Assumptions in basic models

- Stationarity and time reversibility. Stationarity and time reversibility assure the expected frequencies of the nucleotides or amino acids are constant along the evolutionary pathway.
- The conditional probabilities of nucl. subst. are the same for all sites and do not change over time or among lineages.

- Q..... Are these assumptions reasonable?

INDEPENDENCE?

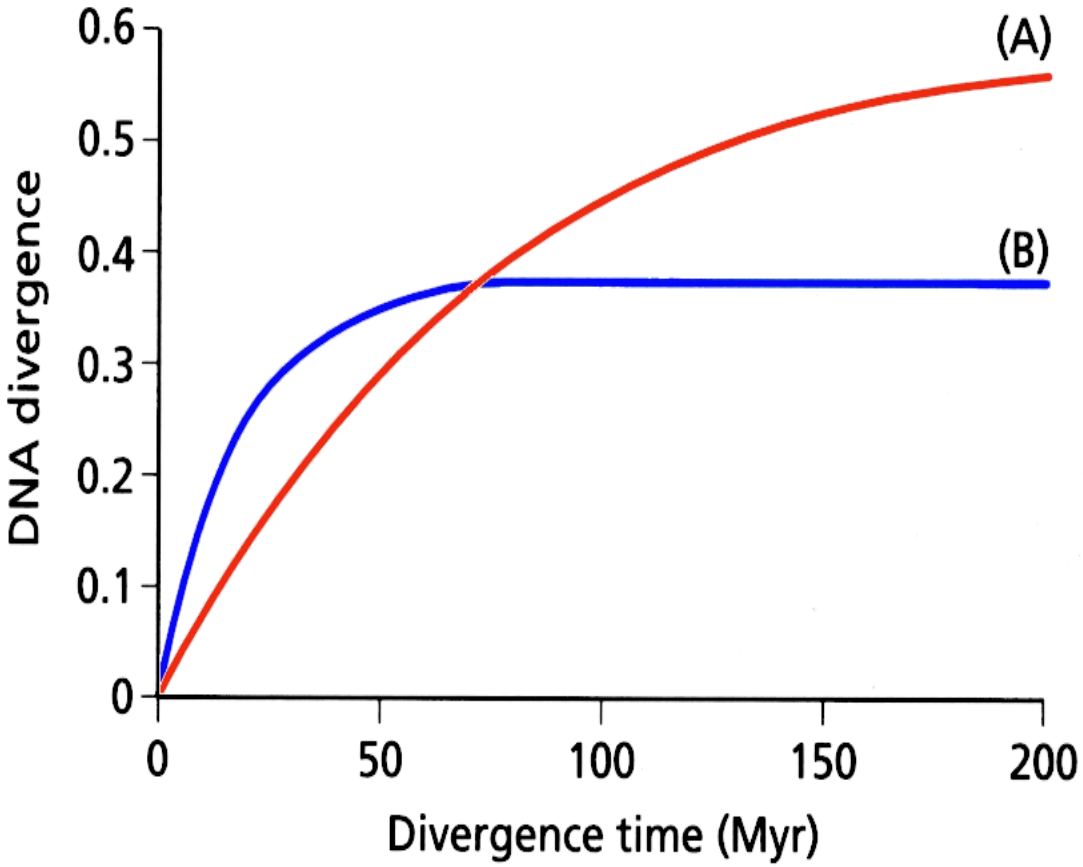
- We assume that change at one site has no effect on other sites. Frequently violated. eg. Ribosomal RNA
- A substitution in a stem region can result in a pair of nucleotides that cannot “Watson-Crick pair” correctly, reducing stability of the structure.
- Often we find that single changes are accompanied by compensatory changes.
- Clearly violates the independence assumption.

Weight differently for stem and loop sites

Variation in rates of substitution among sites?

- All of the methods presented assume that each site in a sequence is equally likely to undergo substitution.
- If rates of substitution vary, can have considerable influence on sequence divergence (i.e. how much change we estimate to have occurred)
- Consider the case where some sites are free to vary while others are constrained to be invariant

If a large proportion of sites are not free to vary then paradoxically, sequences that evolve at a fast rate can appear to show less sequence divergence than more slowly evolving sequences that have fewer constraints.



(A)

rate of subst. 0.5%/Myr: 80% of sites free to vary

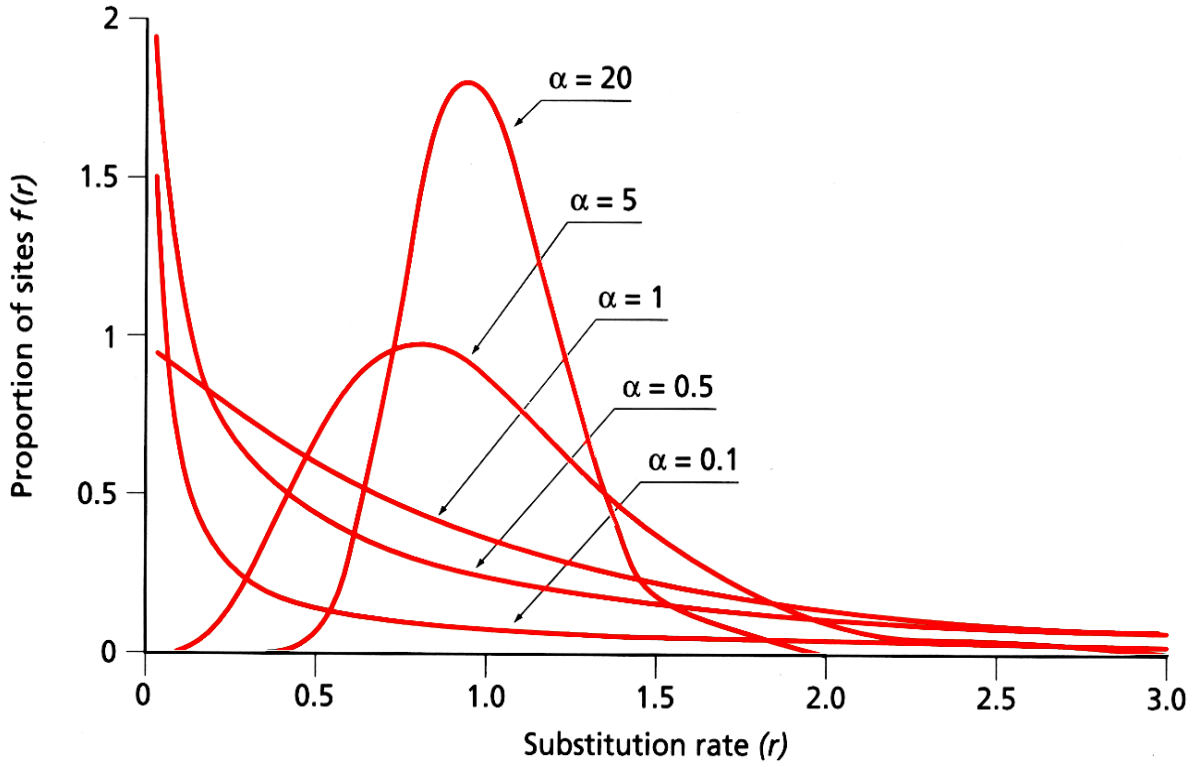
(B)

rate of subst. 2%/Myr: 50% of sites free to vary

In reality sites show a range of probabilities of distribution of rates

- Challenge is to develop a tractable model of the rate variation
- Most widely used approach uses the “gamma distribution”
- Gamma distrib has a shape parameter α that specifies range of rate variation among sites
- small values of α result in L-shaped distrib. larger values smaller range of rates.
- when $\alpha > 1$ distribution is “bell shaped”

Estimates of alpha vary from nuclear and mitochondrial genes vary between 0.16 (12sRNA) - 1.37 (prolactin)



note. Values of α from first & 2nd codon positions tend to be smaller than those from 3rd codon positions

Can modify models of evolutionary change to include the gamma distribution - typically represented by the symbol Γ

HKY + Γ

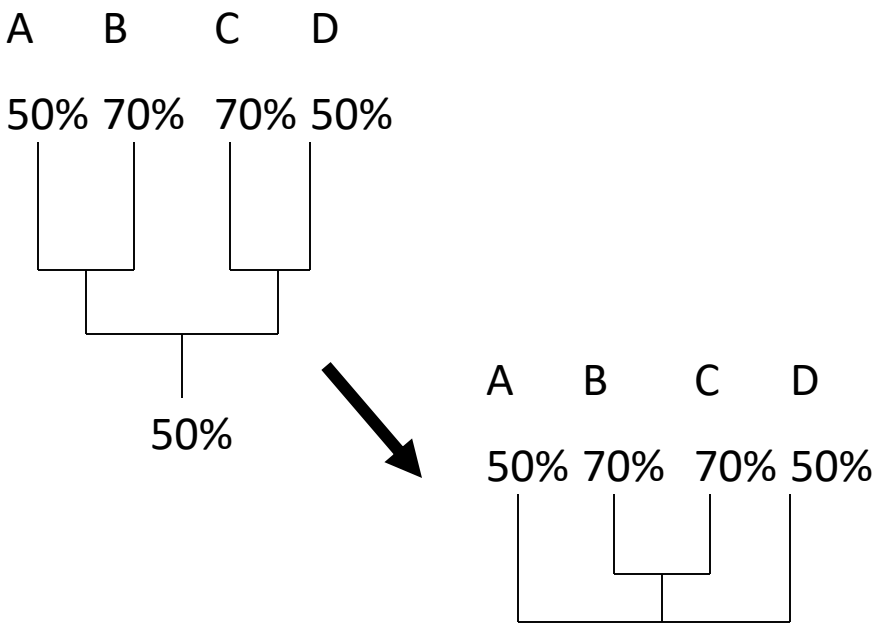
Base Composition Equilibrium?

- Assumes that base composition is roughly the same over the collection of sequences.
- Deviations from this assumption occur commonly and often lead to misleading inferences.
- When constructing trees there is a tendency to cluster sequences together that have similar base compositional profiles.

Explicitly modeling the non-stationary process

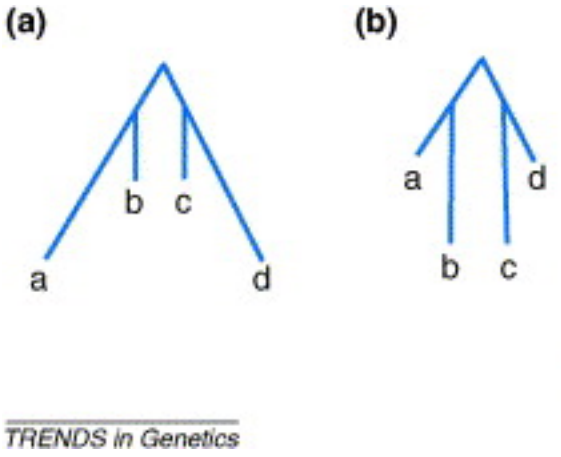
Compositional bias (non-stationary)

- “Compositional bias can result in the artefactual grouping of species with similar nucleotide composition, because most methods assume the homogeneity of the substitution process and the constancy of sequence composition (stationarity) through time ” (Delsuc et al. 2005).



Heterotachy

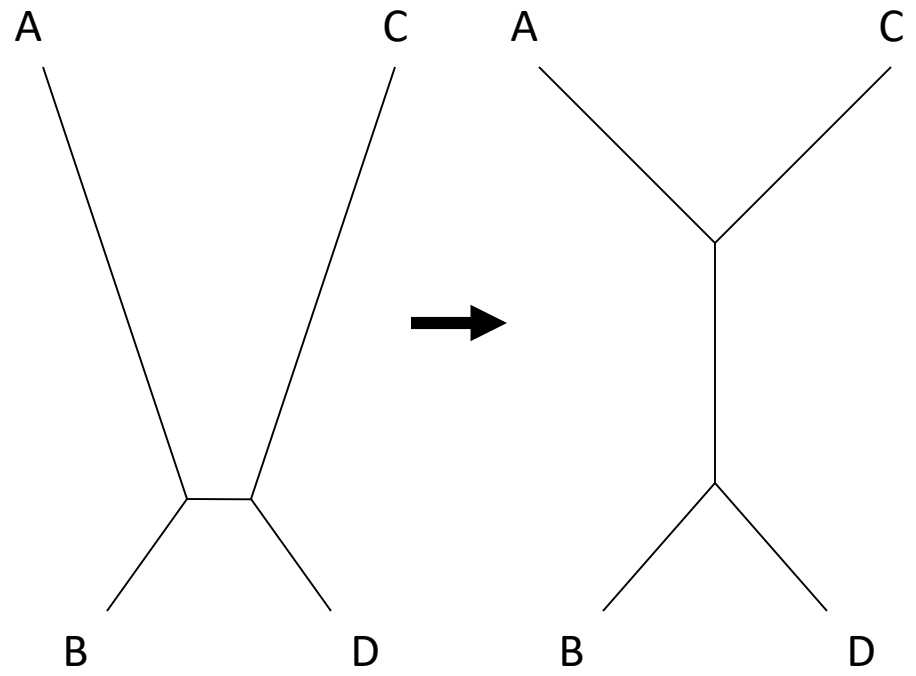
- Heterotachy is the variation of evolutionary rate of a given position of a molecule through time.
- The diagram on the right is a simple scenario used by Kolaczkowski and Thornton (2004).



From Steel, 2005

Long branch attraction

- “Intuitively, with long branches leading to speices A and C, the probability of parallel changes that arrive at the same state becomes greater than the probability of an informative single change in the interior branch of the tree” (Felsenstein, 2004).



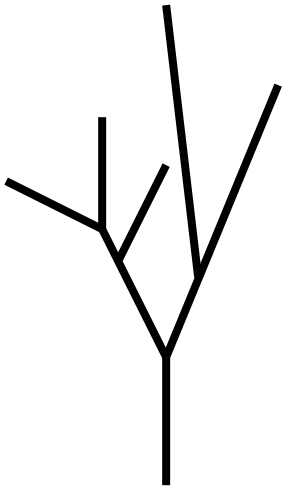
Phylogenomics

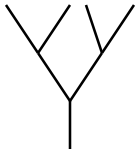
- Prediction of gene function (Eisen, 1998)
- Establishment of evolutionary relationships using genome or genome-scale data

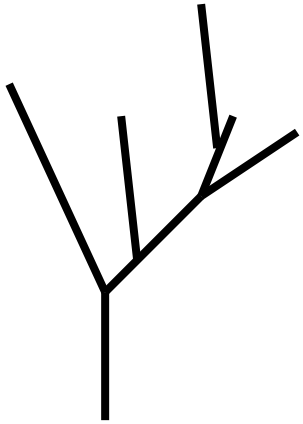
One gene or more genes?

- Single gene or a few genes often result low resolution.
- Single gene or a few genes may even reach to the wrong phylogeny.

Systematic error







Phylogenetic signal

+

+

+ ...

Gene A

Gene B

Gene C

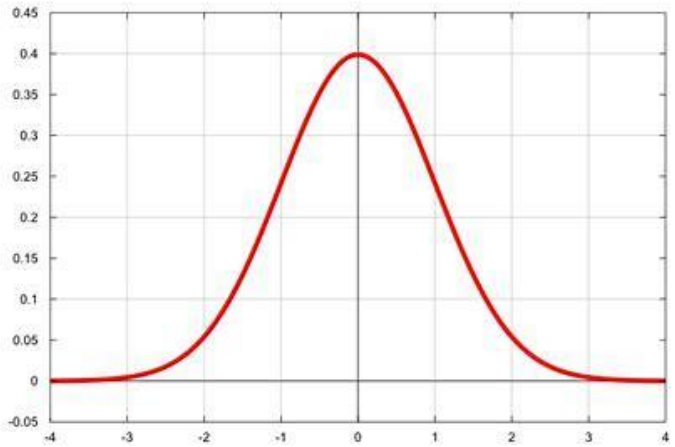
Statistics and concepts

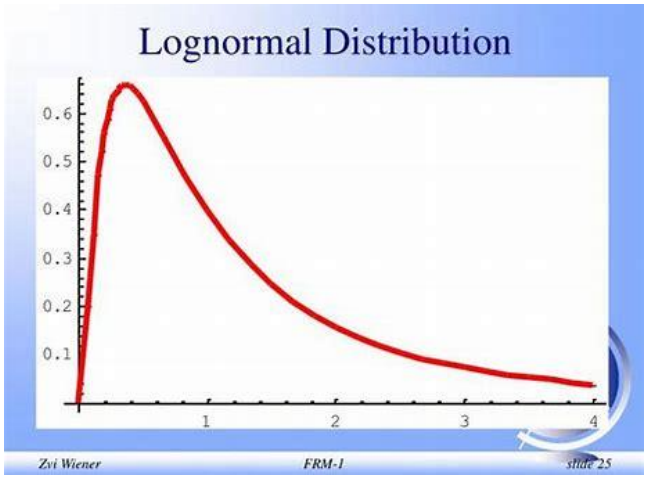
- Likelihood function
- Distribution
- Bayesian approach
- MCMC
- Model selection
- Testing hypothesis

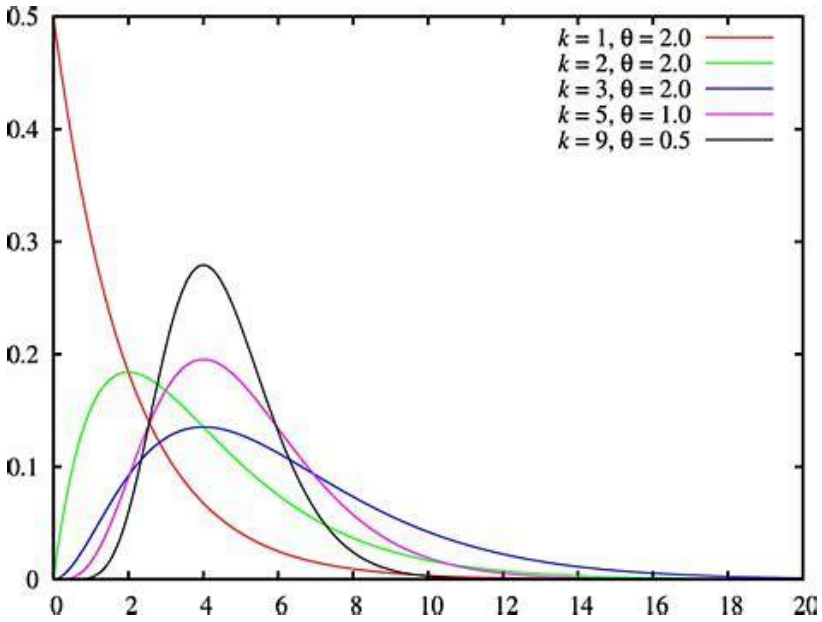
Likelihood function

$$L_D = \Pr(D \mid H)$$

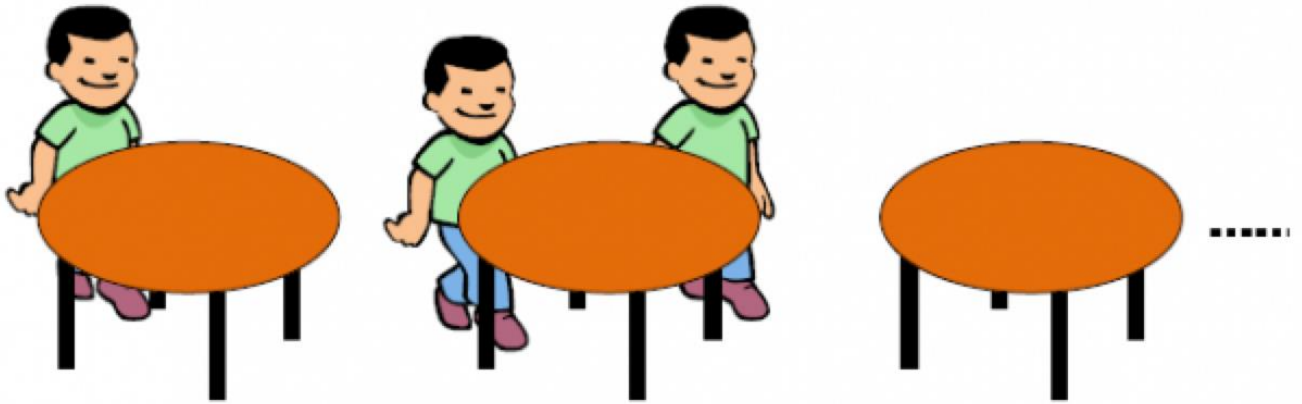
Distribution







The Dirichlet Process
the Chinese Restaurant Process



$P(z_i = c \mid c_{-i}) =$

1	0	0
$\frac{1}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	0
$\frac{1}{2+\alpha}$	$\frac{1}{2+\alpha}$	$\frac{\alpha}{2+\alpha}$
$\frac{1}{3+\alpha}$	$\frac{2}{3+\alpha}$	$\frac{\alpha}{3+\alpha}$

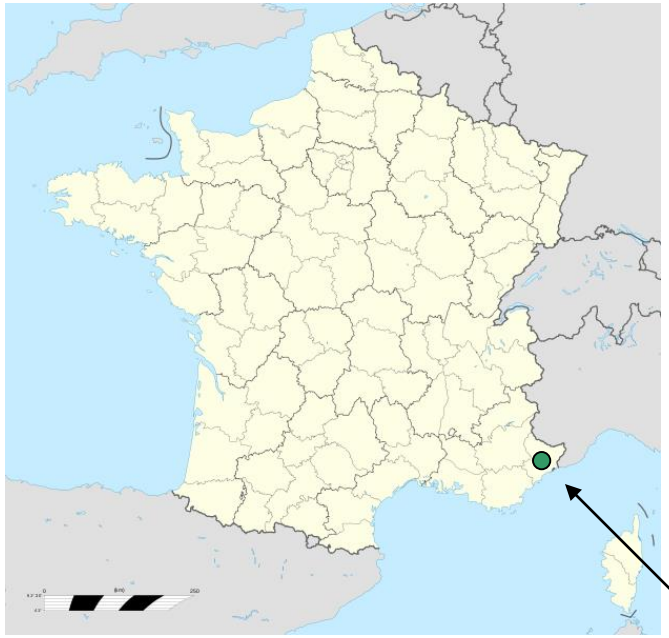
Markov chain

A Markov chain is a model in which changes in states follow transition probabilities.

- It is a stochastic system, i.e. random process
- The probability of the next state depends on the current state, but can also have a chain with memory
- The probability of moving to another state follows a probability distribution
- But it can stay in the same locality, where locality may be in space or time

Markov chain Monte Carlo

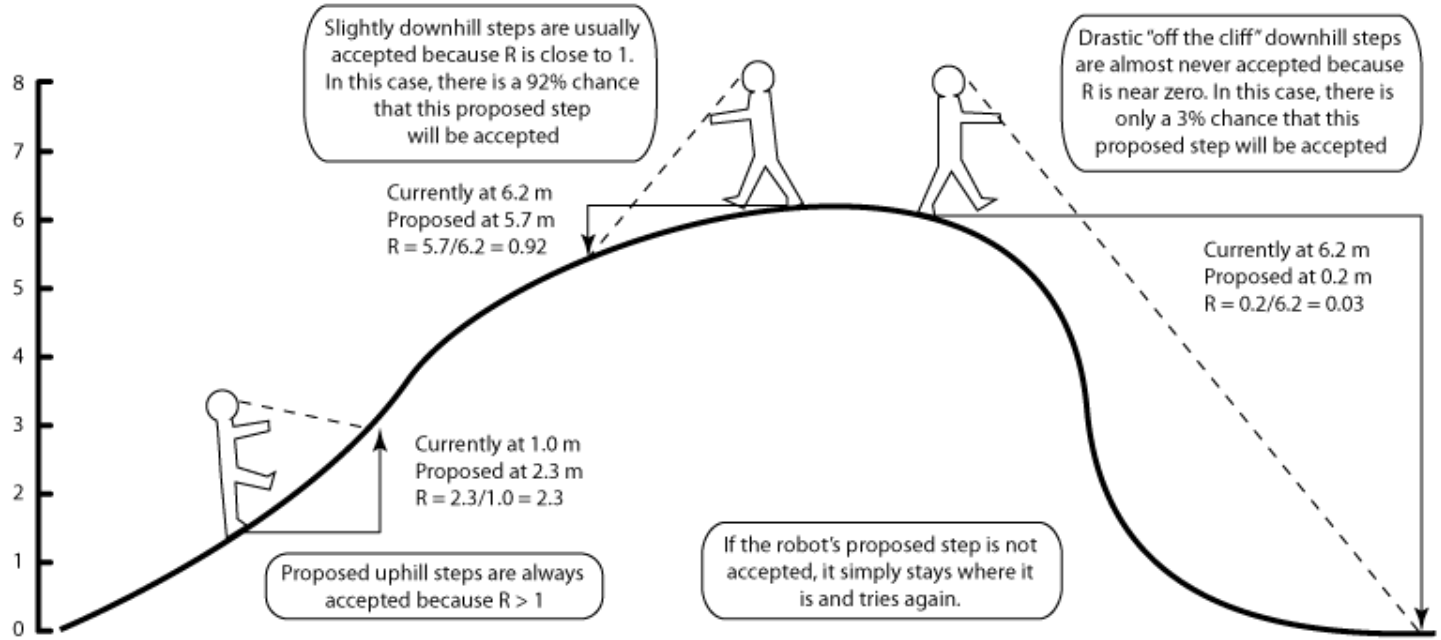
Monte Carlo: town in Monaco famous for its casino (including the European Poker Tour and World Backgammon Championship)



Relevance ?

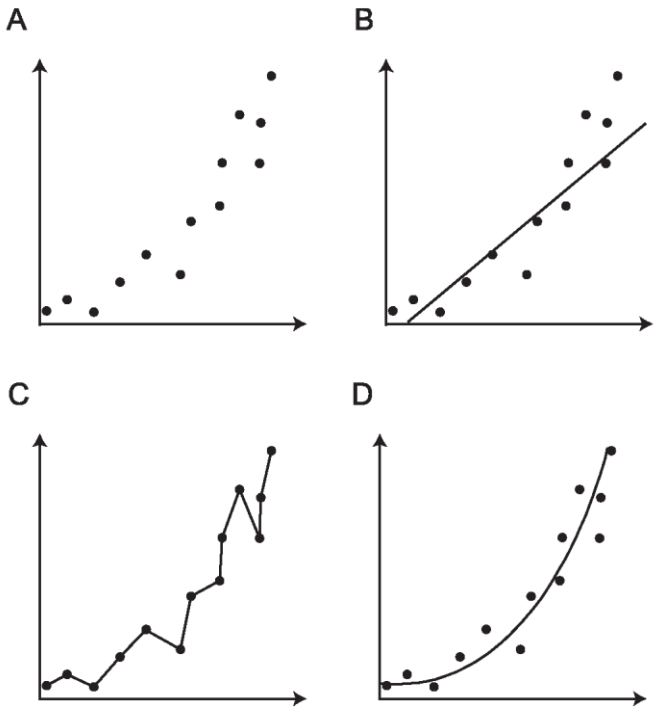
both operate on random processes

MCMC Robot



http://marple.eeb.uconn.edu/mcmcrobot/?page_id=24

Model selection



Likelihood ratio test

$$\delta = 2(\ln L_1 - \ln L_0),$$

where $\ln L_1$ is the likelihood score of the more complex model. The test statistic is then typically evaluated under the assumption of asymptotic convergence to a χ^2 distribution; the degrees of freedom are the difference in number of free parameters in the two models.

Akaike Information Criterion

The Akaike information criterion (AIC) (Akaike 1973) is a simple measure witl a complex derivation. The AIC for model i (AIC_i) is calculated as follows:

$$\mathrm{AIC}_i = -2 \ln L_i + 2k_i,$$

where $\ln L_i$ is the maximum log-likelihood of the model (i.e., with joint ML estimates across parameters) and k_i is the number of parameters in model i. In

quantifying uncertainty in model selection). Burnham & Anderson (2002, 2004) provide the following benchmarks for discerning the relative support for alternative models: $\Delta_i \leq 2$ indicates substantial support, $4 \leq \Delta_i \leq 10$ indicates weak support, and $\Delta_i \geq 10$ indicates no support. Furthermore, these Δ_i values can be

BAYES FACTORS In Bayesian comparison of two models, the Bayes factor permits direct evaluation of the support in the data for one model versus another (Kass & Raftery 1995). This support is calculated as by $B_{12} = \text{pr}(D|M_1)/\text{pr}(D|M_2)$, and it can be multiplied by the ratio of the prior probabilities of each model to give

hLRTs. As with the Δ_i under the AIC, benchmarks are provided by Raftery (1996) to interpret relative support on the basis of the magnitude of the Bayes factor. When $B_{12} > 20$, support for M_1 is strong; when $3 \leq B_{12} \leq 20$, M_1 is slightly favored; and when $1 \leq B_{ij} < 3$, the two models are supported roughly equally by the data. Suchard et al. (2002) used Bayes factors to examine a nested subset of

Annu. Rev. Ecol. Evol. Syst. 2005. 36:445–66
doi: 10.1146/annurev.ecolsys.36.102003.152633

First published online as a Review in Advance on September 16, 2005

MODEL SELECTION IN PHYLOGENETICS

Jack Sullivan[1,2] and Paul Joyce[2,3]
[1]*Department of Biological Sciences, University Idaho, Moscow, Idaho 83844-3051; email: jacks@uidaho.edu*
[2]*Initiative in Bioinformatics and Evolutionary Studies (IBEST), University of Idaho, Moscow, Idaho 83844*
[3]*Department of Mathematics, University of Idaho, Moscow, Idaho 83844-1103; email: joyce@uidaho.edu*

Thank you!