Fundamentals of Evolution

Session 6 - 2018

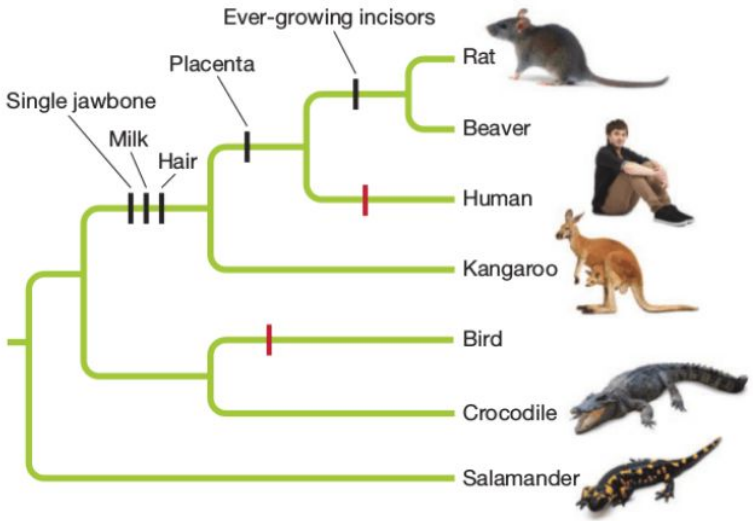
Bayesian phylogenetics & big trees

Recap of last session

- History of systematics and phylogenetics
- Tree thinking
- Character analysis; synapomorphy, homoplasy
- Parsimony methods for phylogenetic inference
- Distance methods for phylogenetic inference
- Likelihood methods for phylogenetic inference

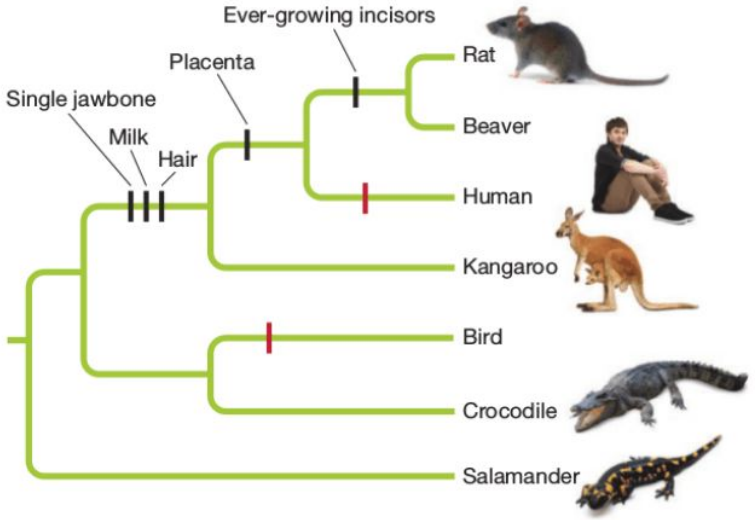
Recap of last session

Phylogenetic relationships are based on shared derived characters (synapomorphies).



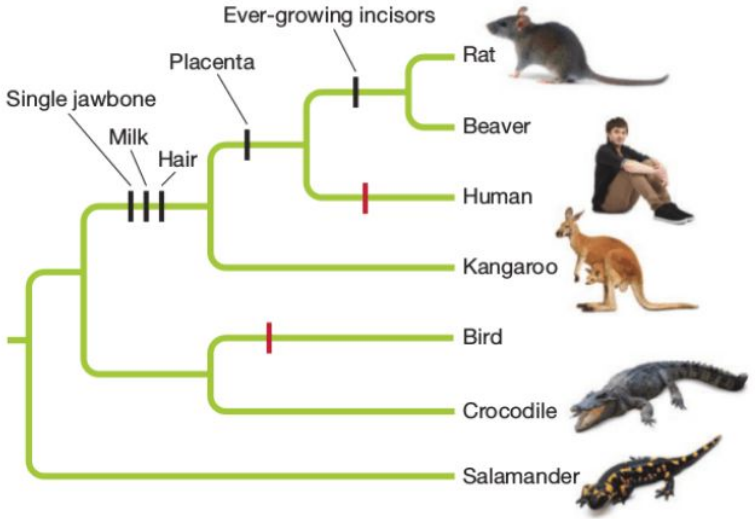
Recap of last session

Most inference methods infer unrooted trees, by counting/estimating changes along branches, and thus do not require us to know which trait is derived vs. ancestral.



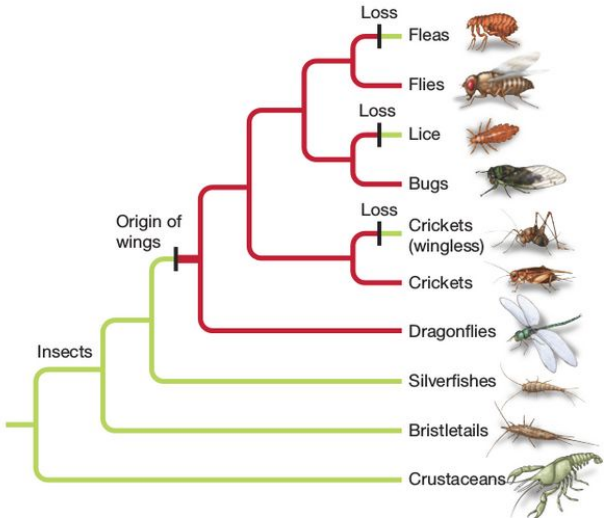
Recap of last session

Based on other knowledge we can then root the tree, which provides *polarization* to the characters, so we know which is derived versus ancestral.



Recap of last session

Homoplasy is a pattern of independent evolution of a character multiple times. It can be caused by *parallel evolution* of homologous characters, or be visualized by mapping convergently evolved characters (non-homologous characters) on the tips of a phylogeny.



Recap of last session

The Likelihood of the data depends on the *topology (branching order), branch lengths, and rate matrix.*

A maximum likelihood optimization finds the *best fitting parameters* of the model (e.g., a substitution matrix) to estimate branch lengths on a given topology. The tree likelihood is the product of all site likelihoods.

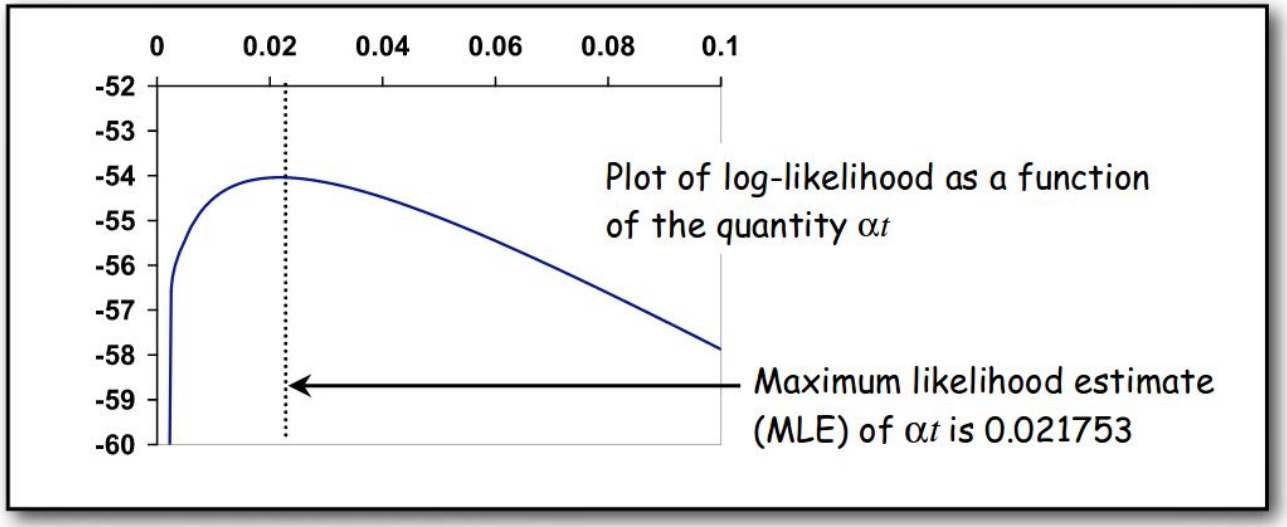
A tree search repeats this process for many or all topologies.

Maximum likelihood estimation

First 32 nucleotides of the ψη-globin gene of gorilla and orangutan:

```
gorilla   GAAGTCCTTGAGAAATAAACTGCACACACTGG
orangutan GGACTCCTTGAGAAATAAACTGCACACACTGG
```

$$L = \left[\left(\tfrac{1}{4}\right)\ \left(\tfrac{1}{4}+\tfrac{3}{4}e^{-4\alpha t}\right)\right]^{30}\left[\left(\tfrac{1}{4}\right)\ \left(\tfrac{1}{4}-\tfrac{1}{4}e^{-4\alpha t}\right)\right]^{2}$$



Recap of last session

Think about the logical steps involved in inferring a phylogeny, and at least one example of each:

- Starting tree (e.g, UPGMA, NJ)
- Optimality criterion (e.g., parsimony, likelihood)
- Heuristic search of tree space (e.g., Hill-climbing)
    - tree rearrangements (e.g, NNI, SPR)

What are pros/cons of using parsimony vs. likelihood?

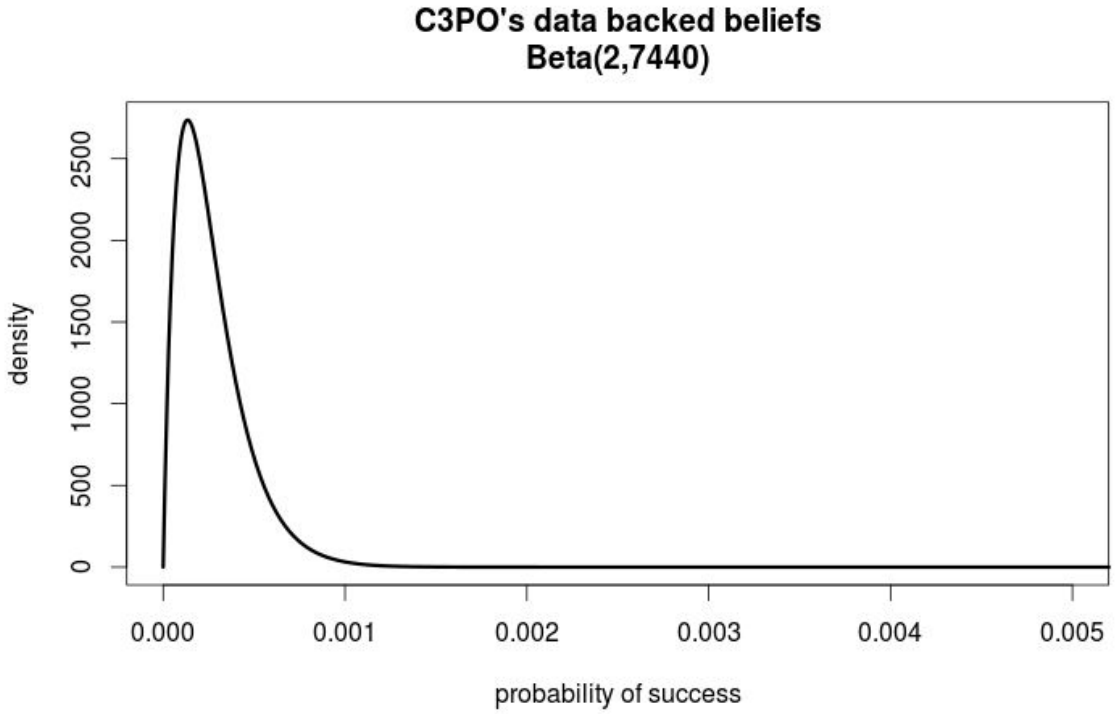
Reconstructing Evolution II

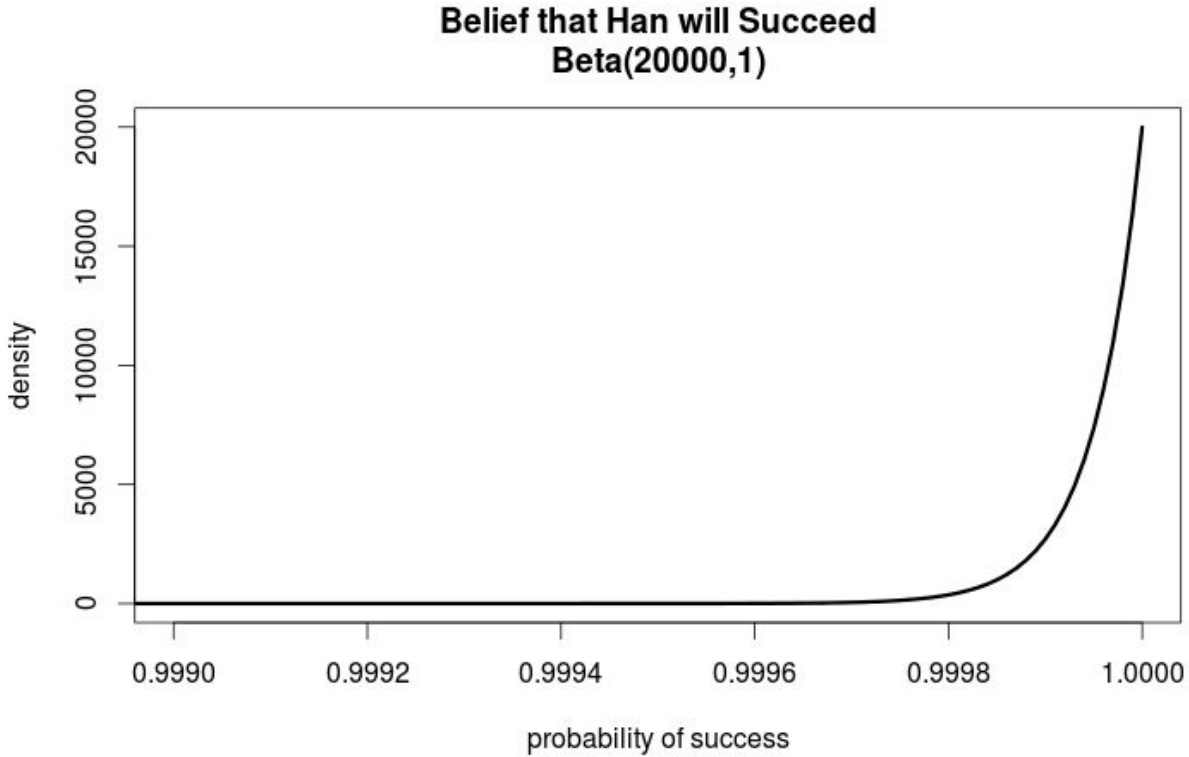
- **Bayesian inference and dated phylogenies**
- **Large-scale phylogenetics: Tree of Life**

Bayesian philosophy

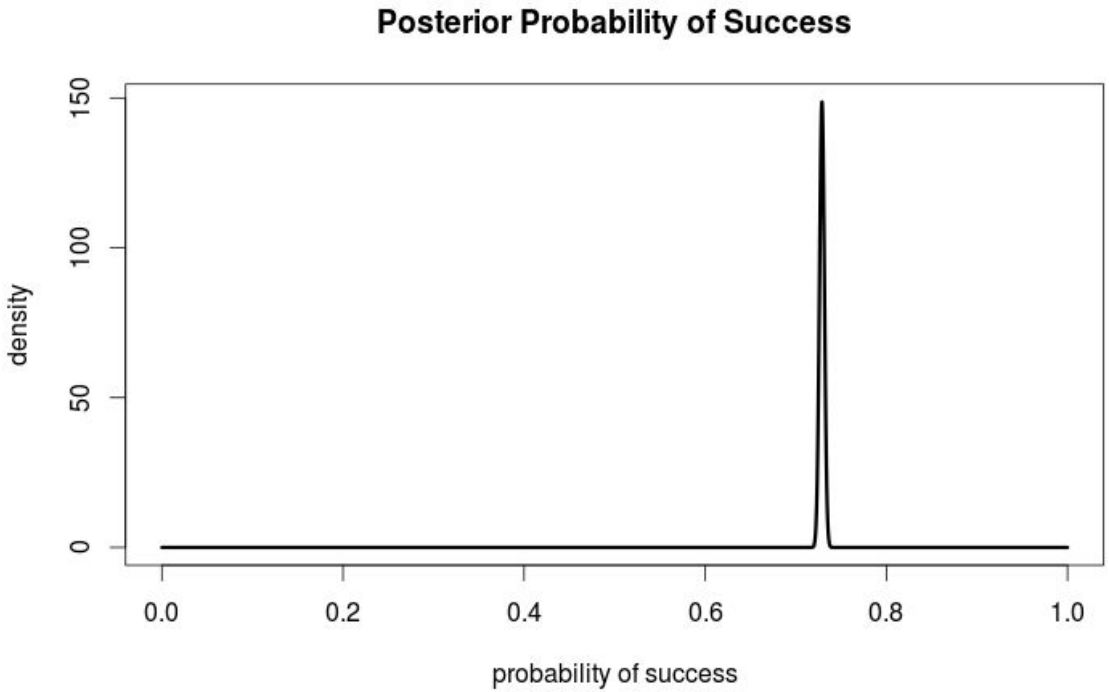
Frequentist (Maximum Likelihood) asks “what is the probability of the data given my hypothesis (model)?”
Bayesian inference asks “What is the probability of my hypothesis (model) given the data?”

Likelihood says, assuming my model is true, what is the probability it generated these data?
Bayesian says, assuming my prior beliefs about this model, how much should I be convinced by new evidence (what is the posterior probability)?


C3PO's data backed beliefs
Beta(2,7440)
density
0
500
1000
1500
2000
2500
0.000
0.001
0.002
0.003
0.004
0.005
probability of success

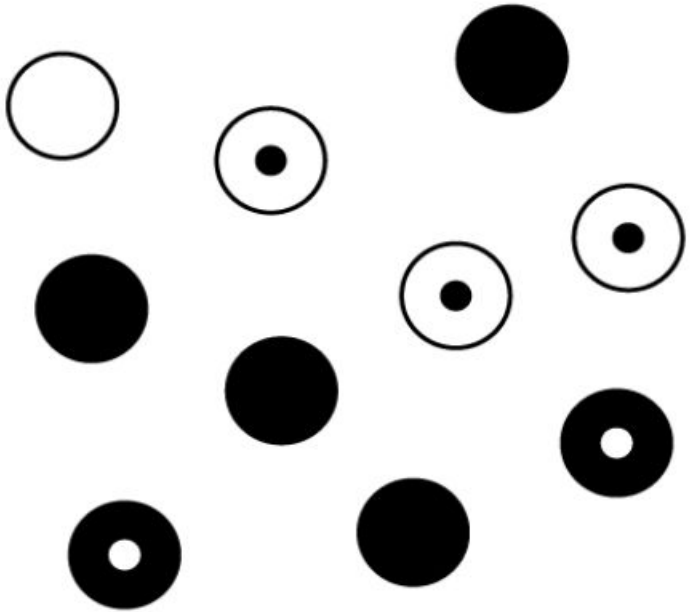

Belief that Han will Succeed
Beta(20000,1)
density
0
5000
10000
15000
20000
0.9990
0.9992
0.9994
0.9996
0.9998
1.0000
probability of success

$$\text{Posterior} = \text{Likelihood} \cdot \text{Prior}$$



Joint probabilities

B = Black S = Solid
W = White D = Dotted



Pr(B) = 0.6 Pr(S) = 0.5

Pr(W) = 0.4 Pr(D) = 0.5

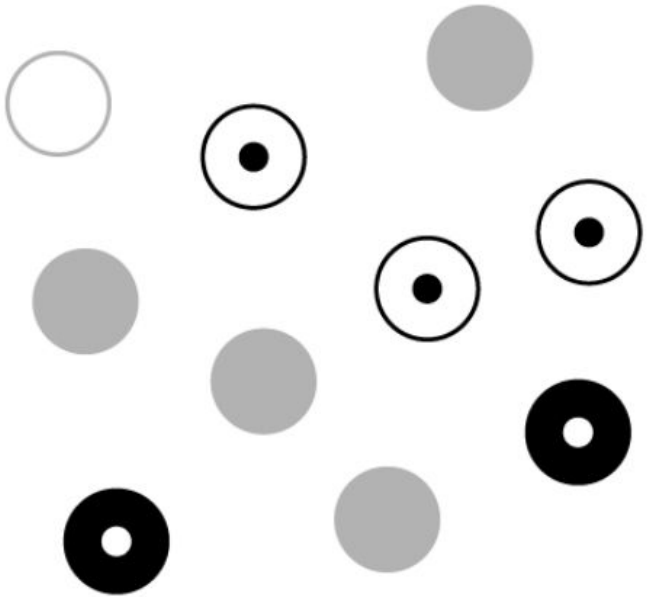
Pr( ) = Pr(B, D) = 0.2

Pr( ) = Pr(B, S) = 0.4

Pr( ) = Pr(W, D) = 0.3

Pr( ) = Pr(W, S) = 0.1

Conditional probabilities

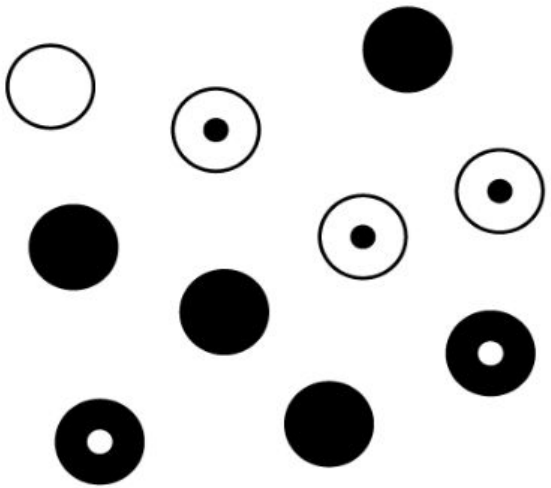


$$\Pr(B|D) = \frac{2}{5} = 0.4$$

Hide all solid marbles
(leaving 5 with dot)

Of those left, 2 are black

Bayes' rule



$$\Pr(B, D)$$

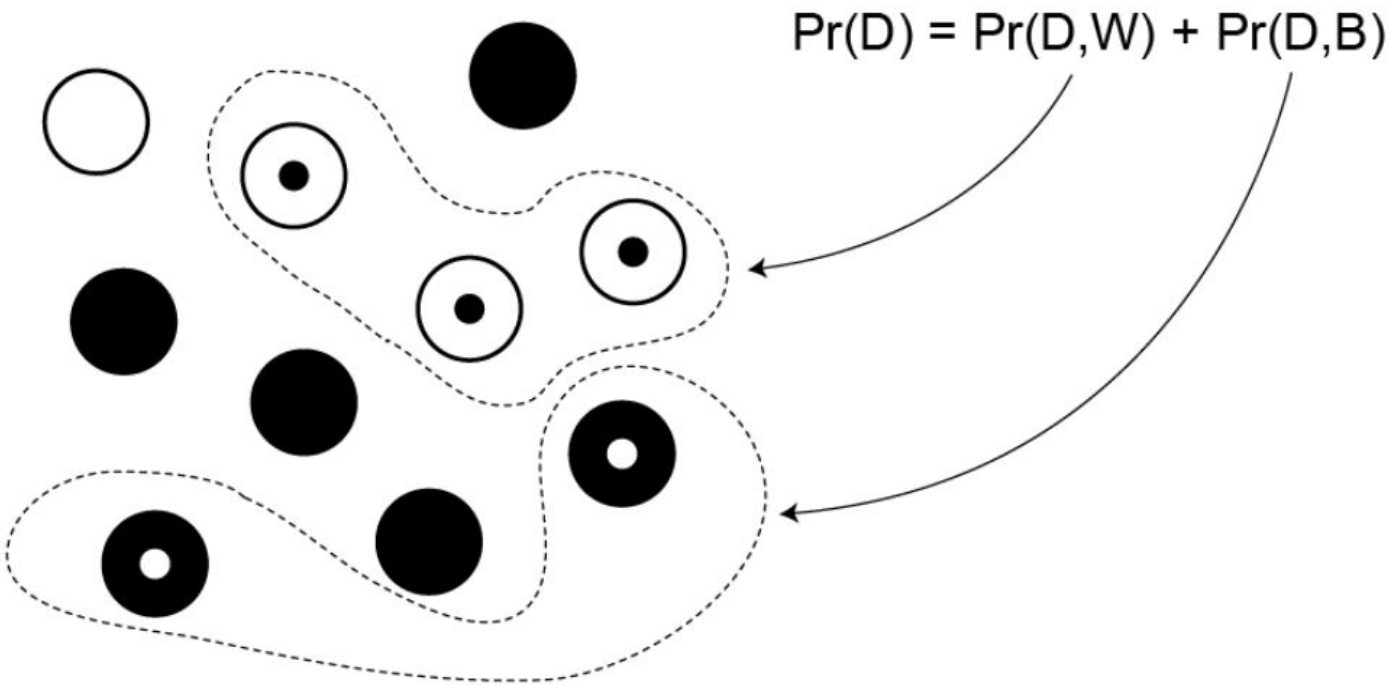
$$\Pr(D)\,\Pr(B|D) = \Pr(B)\,\Pr(D|B)$$

$$\frac{1}{2} \times \frac{2}{5} = \frac{3}{5} \times \frac{1}{3}$$

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

$$= \frac{\frac{3}{5} \times \frac{1}{3}}{\frac{1}{2}} = \frac{2}{5}$$

Probability of "Dotted"



Bayes' rule (cont.)

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

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

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

Bayes' rule (cont.)

It is easy to see that Pr(*D*) serves as a *normalization constant*, ensuring that Pr(*B*|*D*) + Pr(*W*|*D*) = 1.0

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

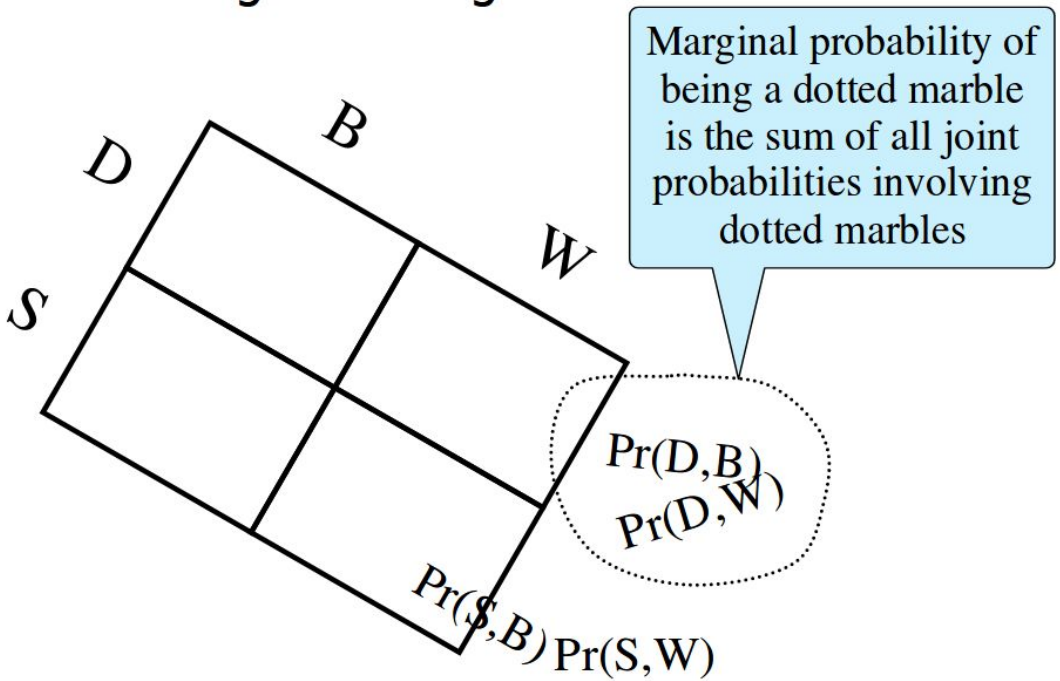
$$\Pr(W|D) = \frac{\Pr(D,W)}{\Pr(D,B) + \Pr(D,W)} \longleftarrow \Pr(D)$$

$$\Pr(B|D) + \Pr(W|D) = \frac{\cancel{\Pr(D,B) + \Pr(D,W)}}{\cancel{\Pr(D,B) + \Pr(D,W)}} = 1$$

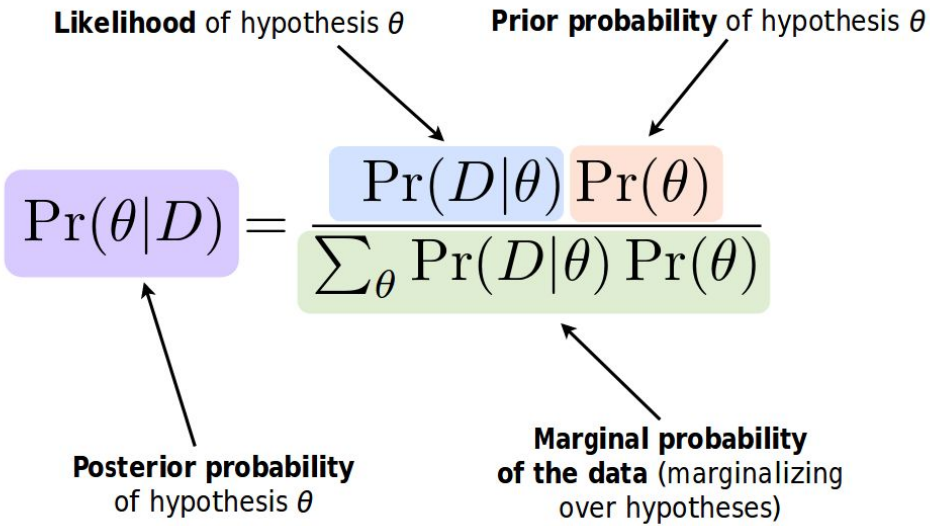
Joint probabilities

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

Marginalizing over colors



Bayes' rule in statistics



Bayes' rule in Statistics

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

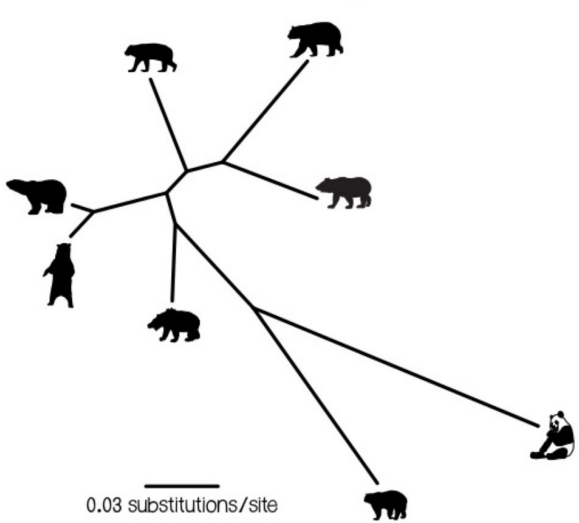
D refers to the "observables" (i.e. the **Data**)

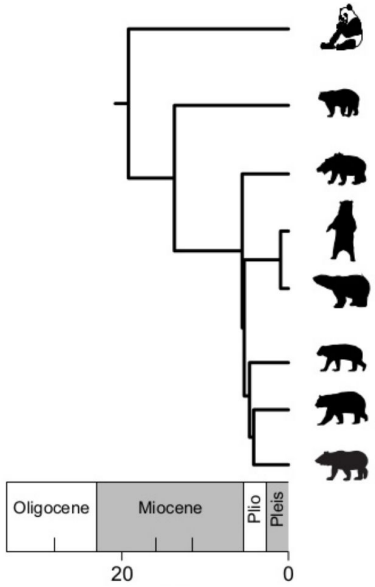
θ refers to one or more "unobservables" (i.e. **parameters** of a model, or the **model itself**):

- *tree model* (i.e. tree topology)
- *substitution model* (e.g. JC, F84, GTR, etc.)
- *parameter* of a substitution model (e.g. a branch length, a base frequency, transition/transversion rate ratio, etc.)
- *hypothesis* (i.e. a special case of a model)
- a *latent variable* (e.g. ancestral state)

Why is Bayesian analysis useful for phylogenetics?

Phylogenies with branch lengths in units of time provide more information than unrooted trees with branch lengths in units of substitutions.

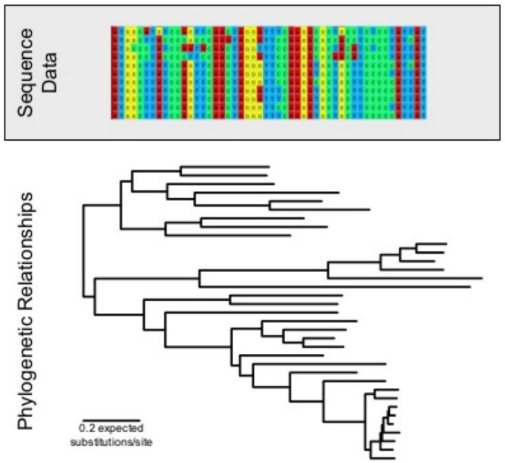




Sequence data provide information about **branch lengths**

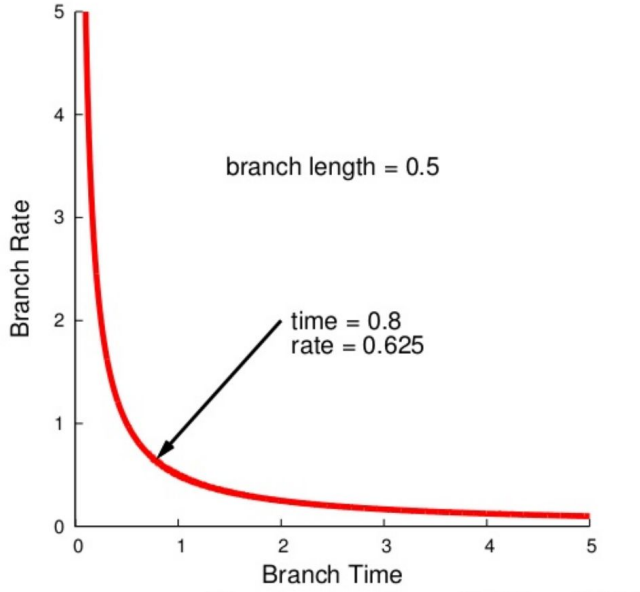
In units of **the expected # of substitutions per site**

branch length = rate × time



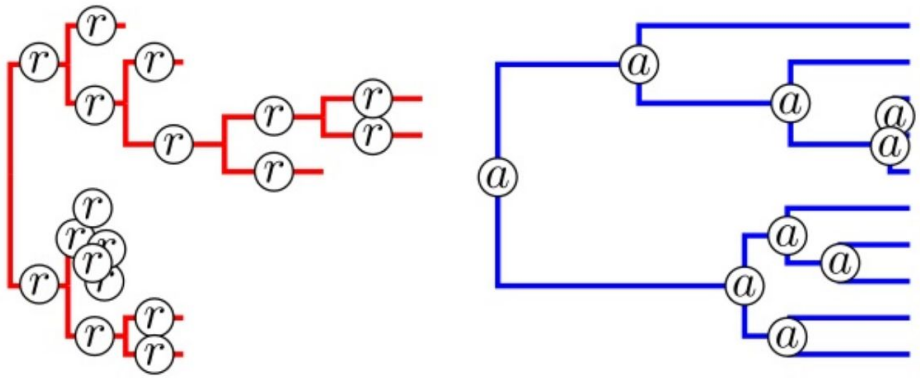
The sequence data provide information about branch length

for any possible rate, there's a time that fits the branch length perfectly



(figure based on Thorne & Kishino, 2005)

Methods for dating species divergences estimate the substitution rate and time separately

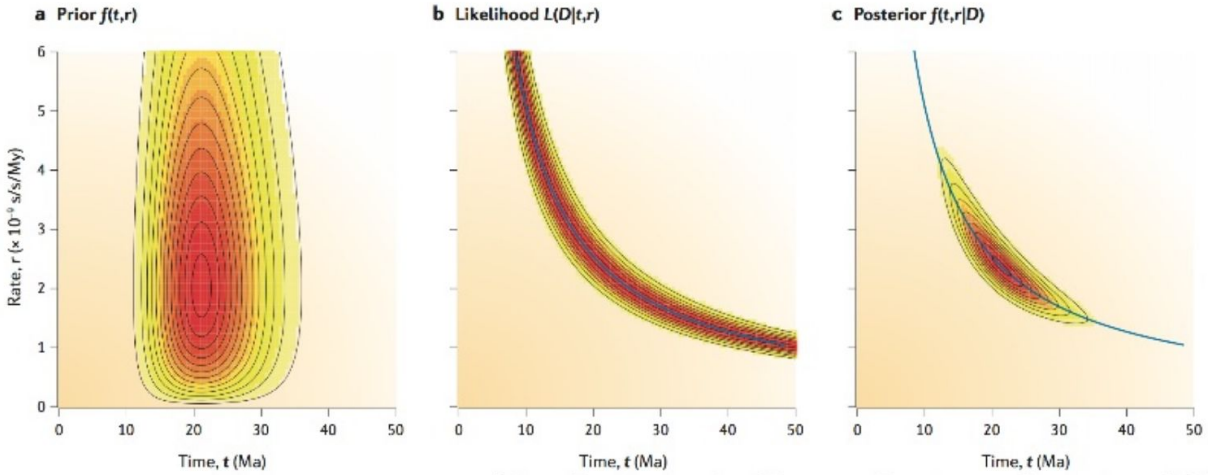


$$\mathcal{R} = (r_1, r_2, r_3, \ldots, r_{2N-2})$$

$$\mathcal{A} = (a_1, a_2, a_3, \ldots, a_{N-1})$$

$$N = \text{number of tips}$$

Methods for dating species divergences estimate the substitution rate and time separately



(dos Reis et al. *Nature Reviews Genetics*, 2016)

Tree-time priors for molecular phylogenies are only informative on a **relative** time scale

Bayes' rule in Statistics

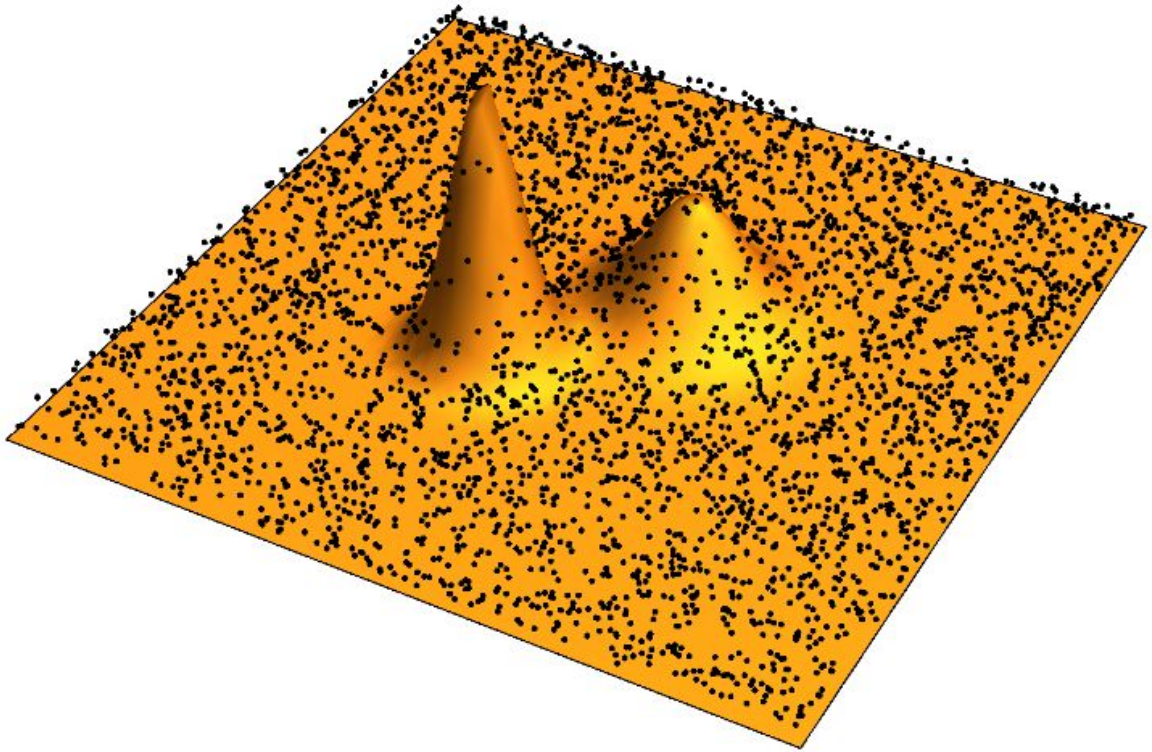
$$\Pr(\theta|D) = \frac{\Pr(D|\theta)\Pr(\theta)}{\sum_\theta \Pr(D|\theta)\Pr(\theta)}$$

D refers to the "observables" (i.e. the **Data**)

θ refers to one or more "unobservables" (i.e. **parameters** of a model, or the **model itself**):

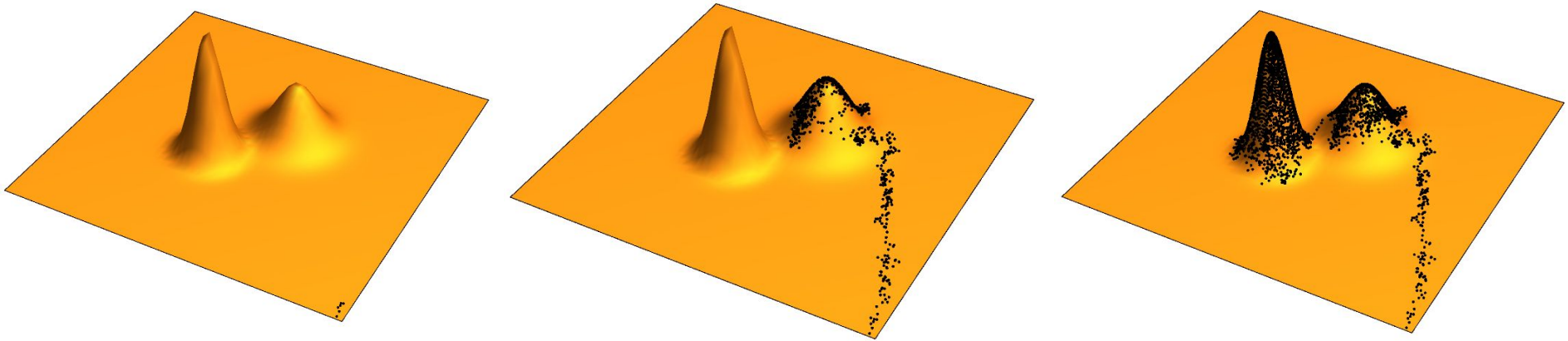
- *tree model* (i.e. tree topology)
- *substitution model* (e.g. JC, F84, GTR, etc.)
- *parameter* of a substitution model (e.g. a branch length, a base frequency, transition/transversion rate ratio, etc.)
- *hypothesis* (i.e. a special case of a model)
- a *latent variable* (e.g. ancestral state)

Naive integration approach

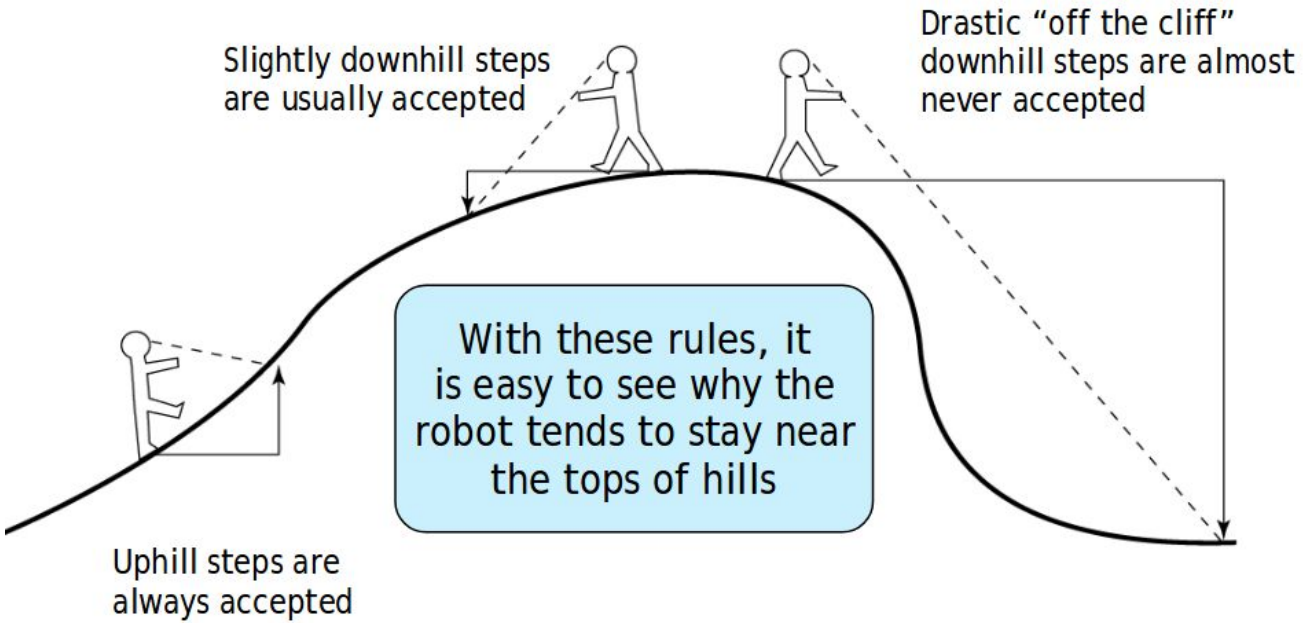


Markov chain Monte-Carlo (MCMC)

Heuristic method of integrating across marginal probabilities. Mechanistic algorithm to search parameter space where the proportion of steps spent in any part of search space reflects the posterior probability support for that parameter. The result is a posterior probability distribution.



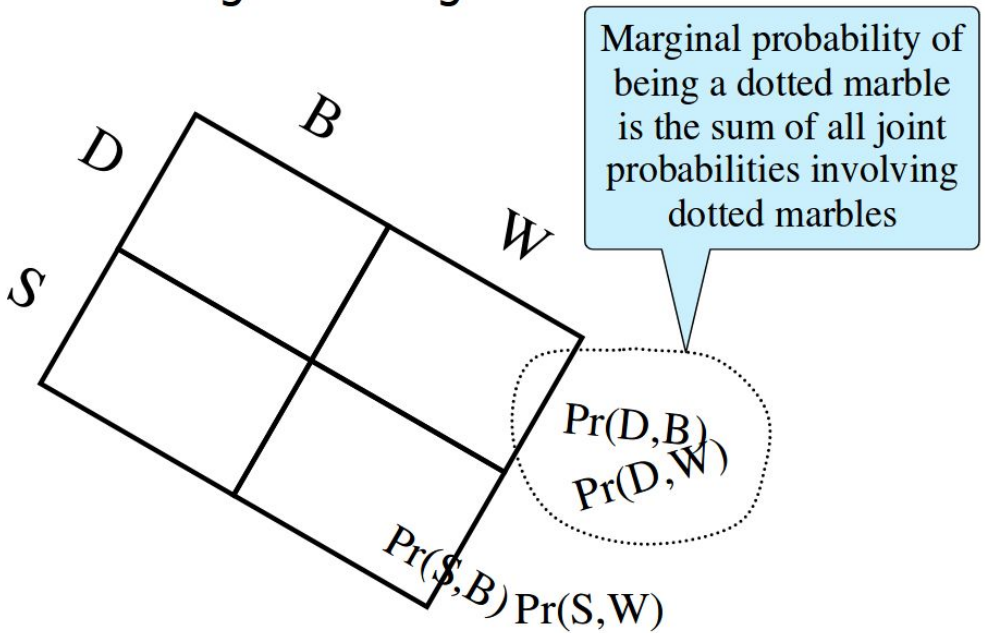
MCMC robot's rules

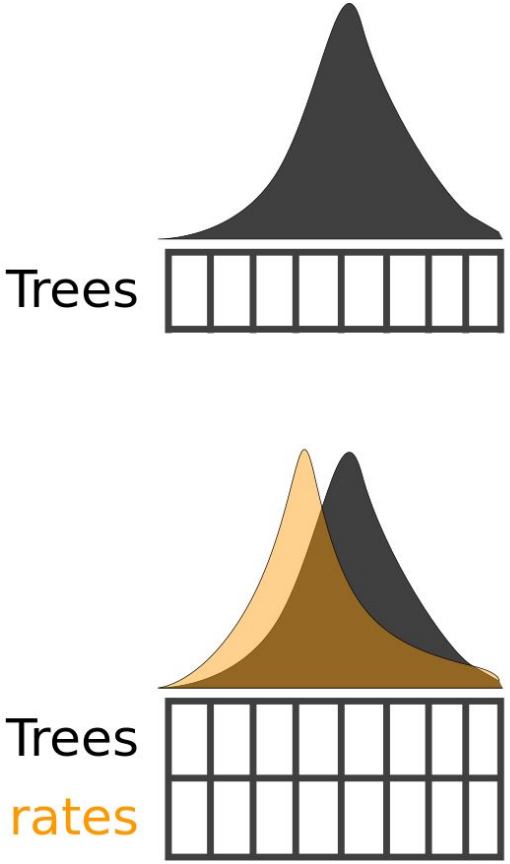


$$f(\mathcal{R}, \mathcal{A}, \theta_{\mathcal{R}}, \theta_{\mathcal{A}}, \theta_s \mid D) = \frac{f(D \mid \mathcal{R}, \mathcal{A}, \theta_s)\, f(\mathcal{R} \mid \theta_{\mathcal{R}})\, f(\mathcal{A} \mid \theta_{\mathcal{A}})\, f(\theta_s)}{f(D)}$$

$f(D \mid \mathcal{R}, \mathcal{A}, \theta_{\mathcal{R}}, \theta_{\mathcal{A}}, \theta_s)$	Likelihood
$f(\mathcal{R} \mid \theta_{\mathcal{R}})$	Prior on rates
$f(\mathcal{A} \mid \theta_{\mathcal{A}})$	Prior on node ages
$f(\theta_s)$	Prior on substitution parameters
$f(D)$	Marginal probability of the data

Marginalizing over colors



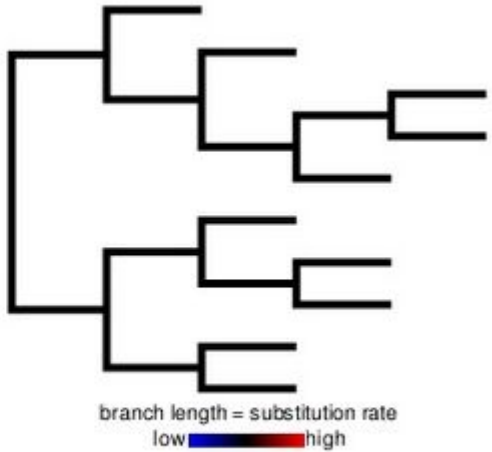

Trees
Trees
rates

- **Global clock** (Zuckerkandl & Pauling, 1962)
- **Local clocks** (Hasegawa, Kishino & Yano 1989; Kishino & Hasegawa 1990; Yoder & Yang 2000; Yang & Yoder 2003, Drummond and Suchard 2010)
- **Punctuated rate change model** (Huelsenbeck, Larget and Swofford 2000)
- **Log-normally distributed autocorrelated rates** (Thorne, Kishino & Painter 1998; Kishino, Thorne & Bruno 2001; Thorne & Kishino 2002)
- **Uncorrelated/independent rates models** (Drummond et al. 2006; Rannala & Yang 2007; Lepage et al. 2007)
- **Mixture models on branch rates** (Heath, Holder, Huelsenbeck 2012)

The substitution rate is constant over time

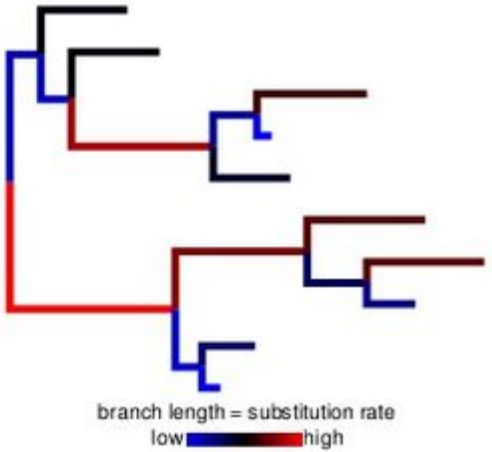
All lineages share the same rate

(Zuckerkandl & Pauling, 1962)



Lineage-specific rates are uncorrelated when the rate assigned to each branch is independently drawn from an underlying distribution

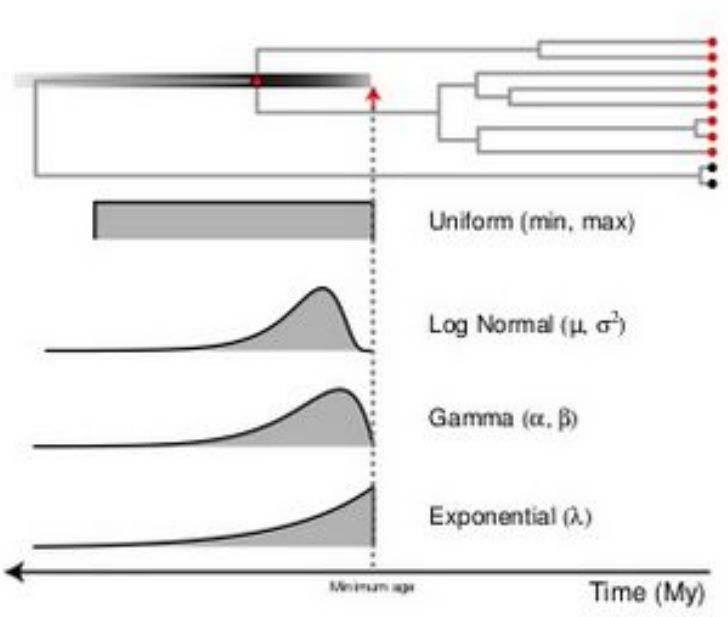
(Drummond et al. 2006; Rannala & Yang 2007; Lepage et al. 2007)



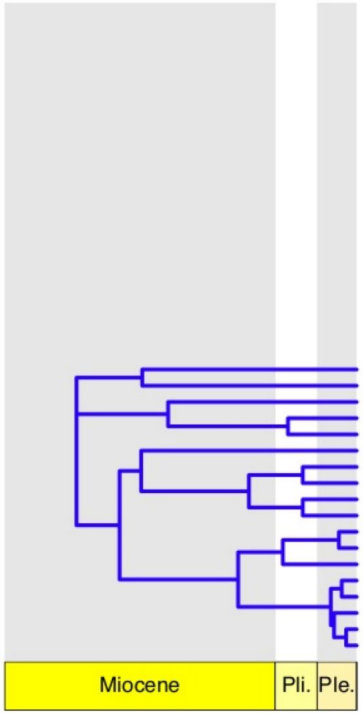
Common practice in Bayesian divergence-time estimation:

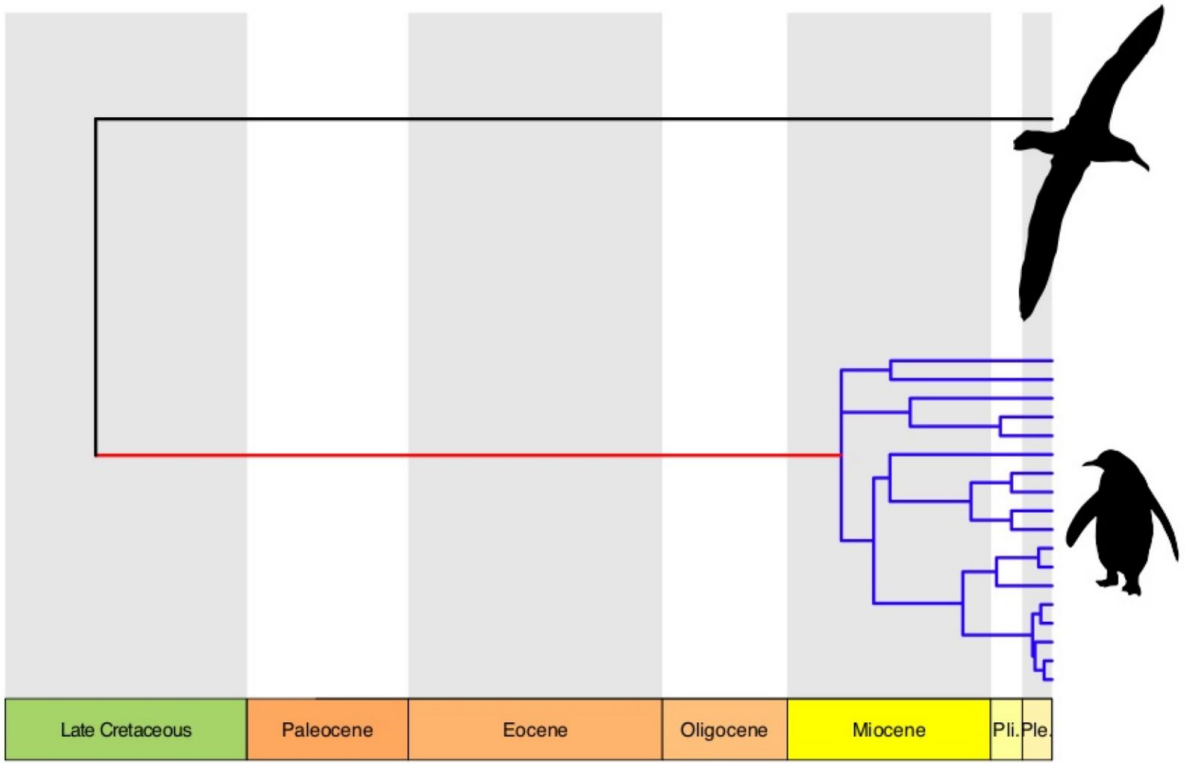
Parametric distributions are typically off-set by the age of the oldest fossil assigned to a clade

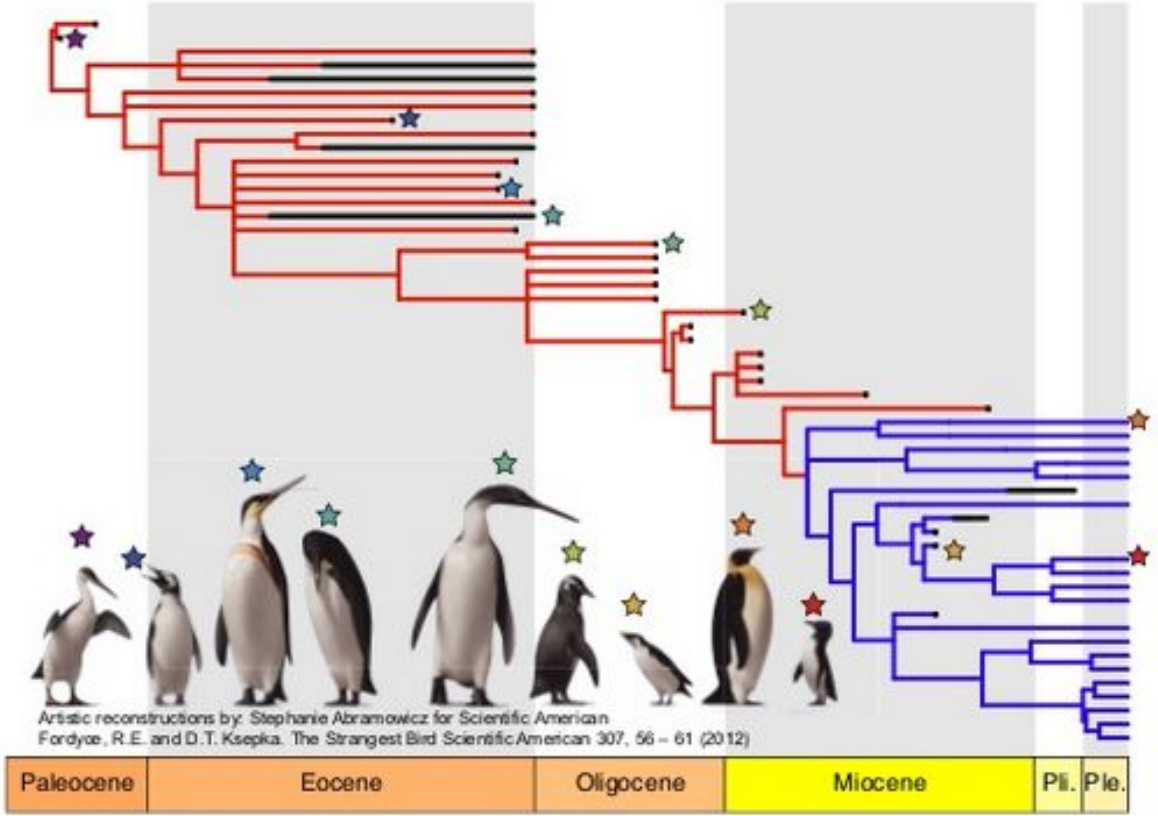
These prior densities do not (necessarily) require specification of maximum bounds



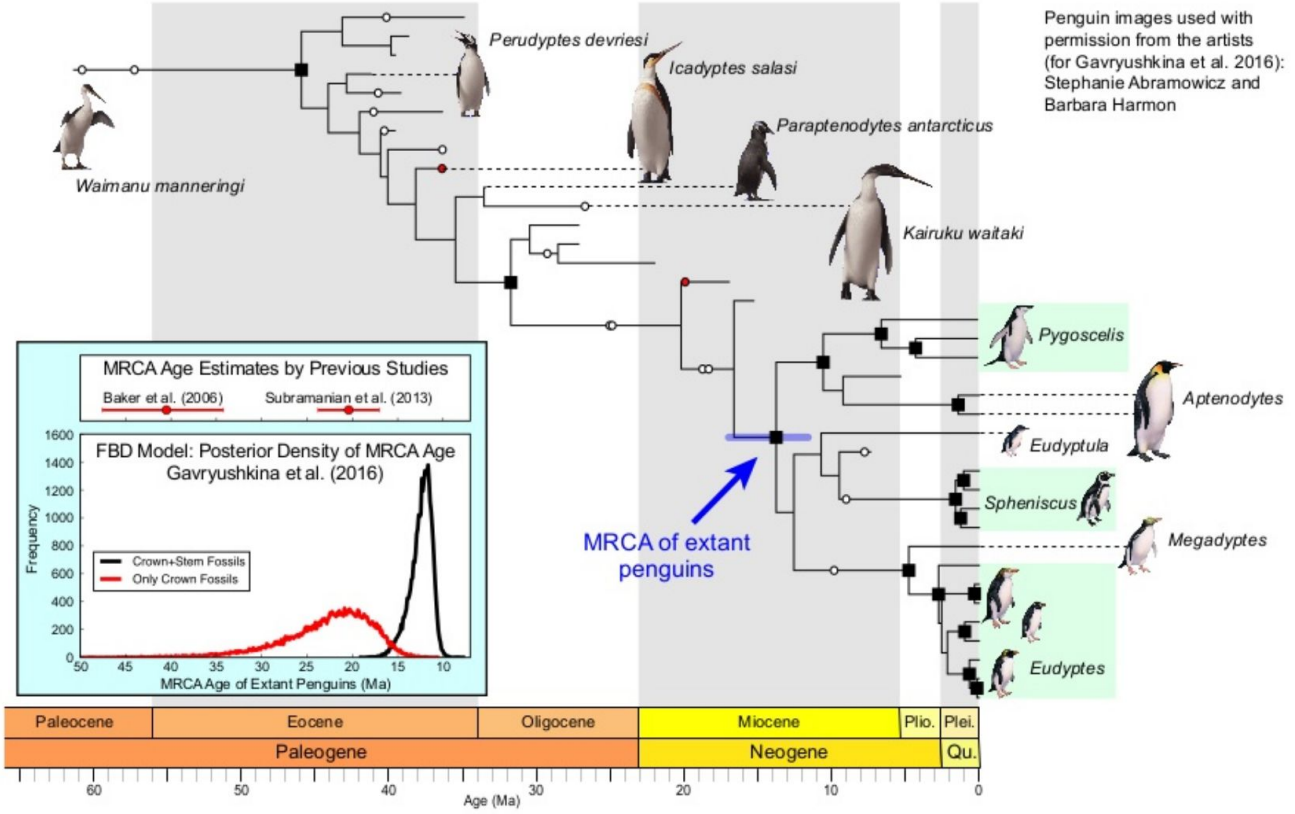
Penguin Diversity



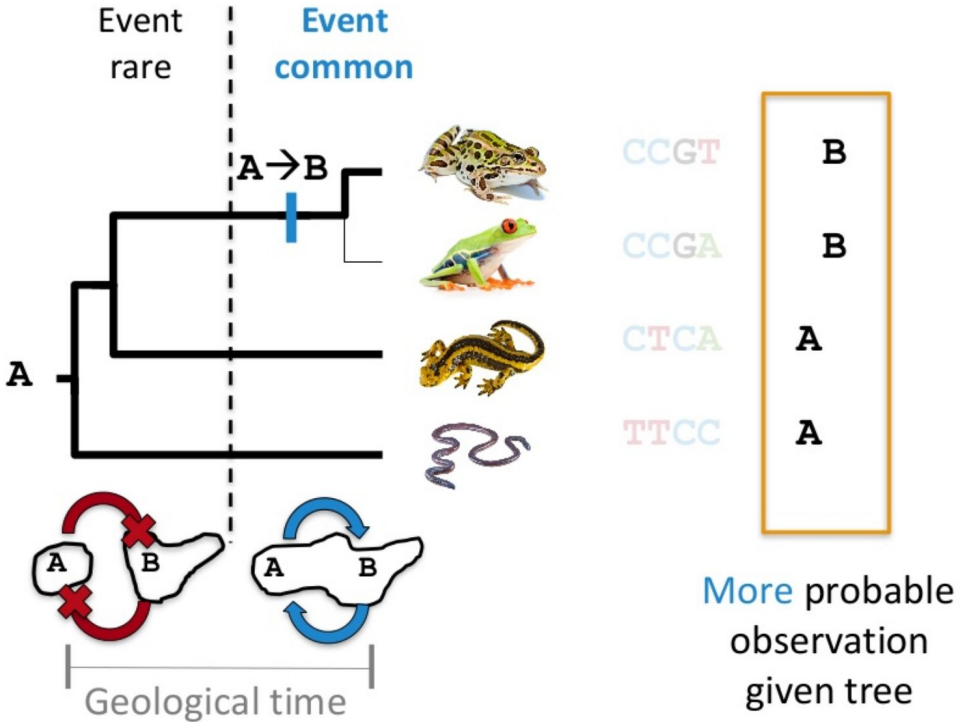

Late Cretaceous
Paleocene
Eocene
Oligocene
Miocene
Pli.
Ple.


Artistic reconstructions by: Stephanie Abramowicz for Scientific American
Fordyce, R.E. and D.T. Ksepka. The Strangest Bird Scientific American 307, 56 – 61 (2012)
Paleocene
Eocene
Oligocene
Miocene
Pli.
Ple.

Incorporating both fossils and DNA sequences, and informed priors on the fossil placements, Gavryushkina et al. (2016) found the crown age of extant penguins is much younger than previously thought.



Even without fossils, time-informed priors



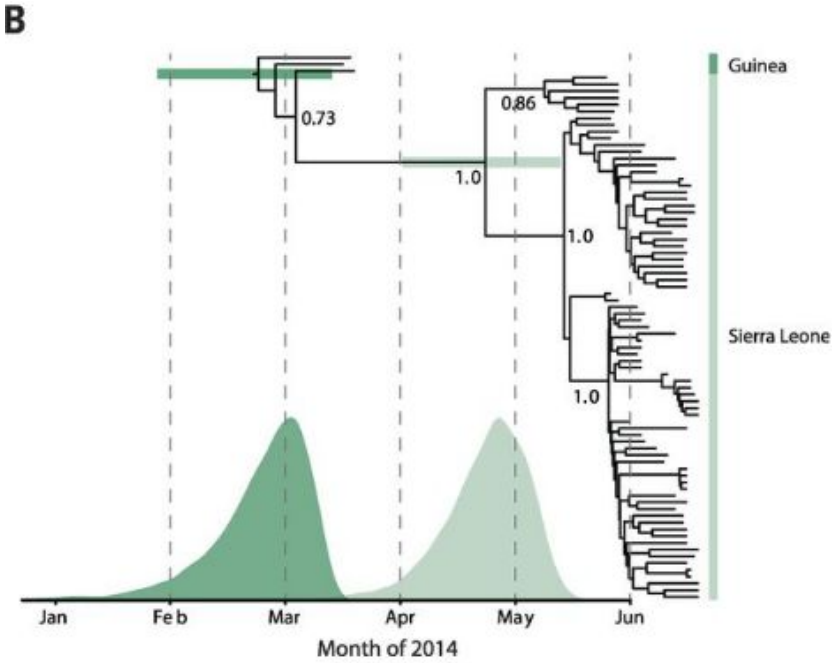
Phylodynamics

- The study of how epidemiological, immunological, and evolutionary processes act and potentially interact to shape viral phylogenies.

- Bayesian phylogenetics is highly important because *rate* varies dramatically during viral outbreaks

Estimating the rate of infection of Ebola

- The 2013 West African Ebola virus epidemic spread primarily through Guinea, Sierra Leone and Liberia and killed over 11,000 people
- Estimated that strain began at a funeral in Guinea December 2013
- Phylogenetic analysis shows MRCA was February 2014 with 2 strains introduced to Sierra Leone.



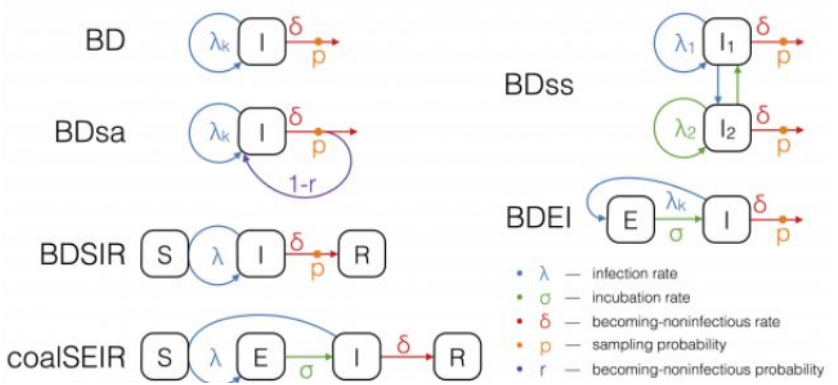
Stephen K. Gire et al. Science 2014;345:1369-1372

Estimating the rate of infection of Ebola

- Multiple birth-death model approaches were used to estimate epidemiological parameters across a Bayesian phylogeny.

- Birth is the rate of transmission, death is recovery or death of host.

- Incubation time: 4.92 days
- Infectious period: 2.58 days
- RO: 2.18 people



Stadler T *et al*. PLOS Currents Outbreaks. 2014

Summary of Bayesian phylogenetics

- Broadly applicable statistical framework that allows one to combine data from many different sources through defining priors.

- In practice, often used for dated phylogenies because with priors on ages or rates you can better differentiate age from rate (which cannot be done in ML)

- However, it can be rather slow (MCMC search)
- And if you define too strict of priors then your results may just return what you put it. Requires careful testing/refining.

Large-scale phylogenetics

- Increasingly, phylogenetic and phylogenomics is a field of informatics, or *data science, and computer science.*

- Data archiving and mining. Researchers focus on specific groups and over time accumulate enough data to span deeper and deeper in time.

- Methods for combining knowledge and minimizing the need to optimization + tree search.

How many species are there?

Article | OPEN

Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life

How many species are there?

- Globally, our best approximation to the total number of species, based on taxonomic expertise, is 3-100 million species (May 2010).

- Many methods are employed to estimate the number of undiscovered/described species: e.g., body-size distribution, species-area relationship, ratios between taxa, time-series relationships (Mora et al. 2011)

Species	Earth			Ocean		
	Catalogued	**Predicted**	**±SE**	**Catalogued**	**Predicted**	**±SE**
Eukaryotes						
Animalia	953,434	7,770,000	958,000	171,082	2,150,000	145,000
Chromista	13,033	27,500	30,500	4,859	7,400	9,640
Fungi	43,271	611,000	297,000	1,097	5,320	11,100
Plantae	215,644	298,000	8,200	8,600	16,600	9,130
Protozoa	8,118	36,400	6,690	8,118	36,400	6,690
Total	1,233,500	8,740,000	1,300,000	193,756	2,210,000	182,000
Prokaryotes						
Archaea	502	455	160	1	1	0
Bacteria	10,358	9,680	3,470	652	1,320	436
Total	10,860	10,100	3,630	653	1,320	436
Grand Total	**1,244,360**	**8,750,000**	**1,300,000**	**194,409**	**2,210,000**	**182,000**

Predictions for prokaryotes represent a lower bound because they do not consider undescribed higher taxa. For protozoa, the ocean database was substantially more complete than the database for the entire Earth so we only used the former to estimate the total number of species in this taxon. All predictions were rounded to three significant digits.

doi:10.1371/journal.pbio.1001127.t002

(Mora et al. 2011)

Large-scale phylogenetics

- Super trees:
- Inferring large trees is difficult and time consuming, it is easier to join together smaller trees. Several techniques.
- This type of method has regained some popularity recently in the study of quartet trees (e.g., SVDquartets)

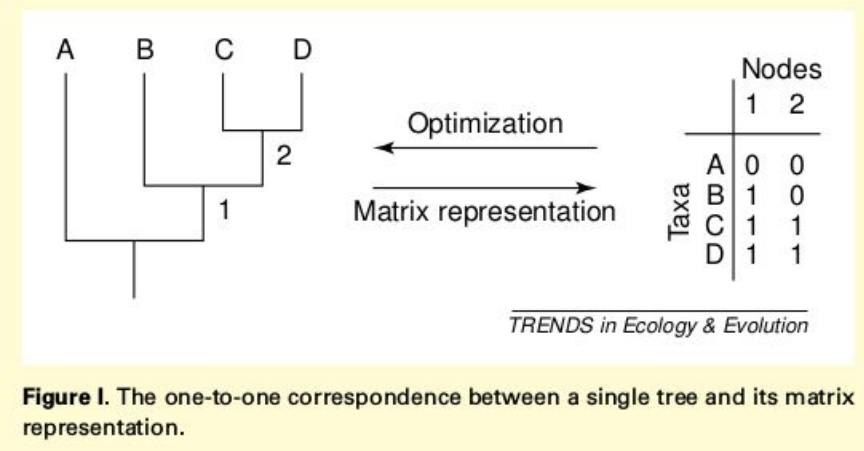


Figure I. The one-to-one correspondence between a single tree and its matrix representation.

Large-scale phylogenetics

- Supermatrices:
- Around the early 2000s common markers were discovered that could be sequenced reliably across many organisms, which made it possible to combine their data into larger analyses. Faster inference methods developed.
- Hundreds of taxa, one or more genes. Sparse matrices.

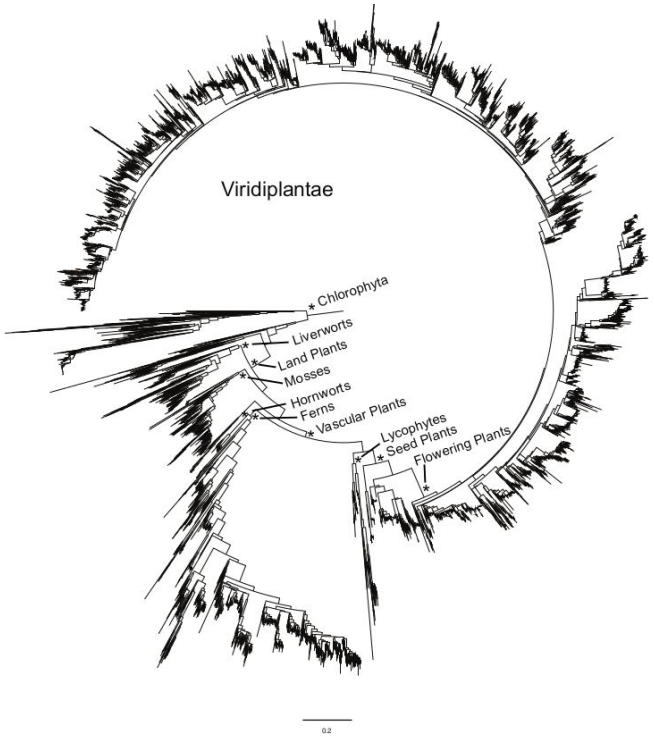


Figure 3
Maximum-likelihood phylogeny for 13,533 species of green plants based on *rbcL* DNA sequences. The data matrix was constructed using the mega-phylogeny method; major clades are labeled and denoted with a star.

Large-scale phylogenetics

- Megaphylogeny pipelines: Automated procedures to build supermatrices by finding sequences in databases and aligning them at multiple hierarchical levels.
- Example: >13K species of plants analyzed for one gene.

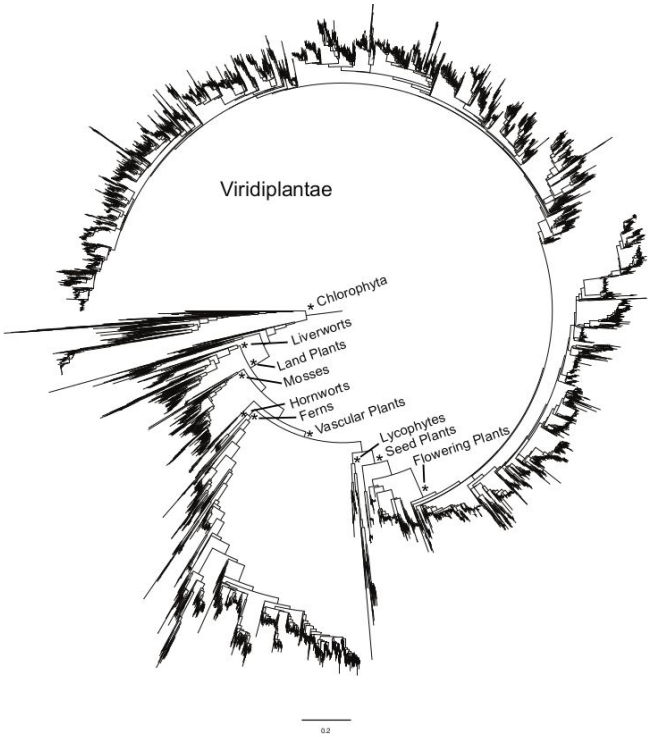
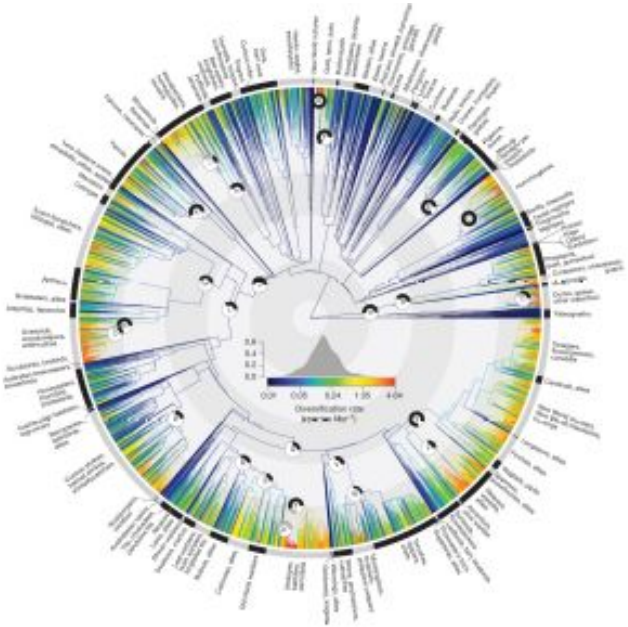


Figure 3
Maximum-likelihood phylogeny for 13,533 species of green plants based on *rbcL* DNA sequences. The data matrix was constructed using the mega-phylogeny method; major clades are labeled and denoted with a star.

Large-scale phylogenetics

- Dated megaphylogenies:
- Bayesian relaxed clock analysis on a reduced set of taxa to infer the backbone.
- Bayesian relaxed clock analyses subclades that are then added to the backbone.



W Jetz *et al. Nature* **000**, 1-5 (2012) doi:10.1038/nature11631

Large-scale phylogenetics

- National Science Foundation initiatives to support Assembling the Tree of Life programs starting in early 2000s.

Assembling the fungal tree of life: progress, classification, and evolution of subcellular traits [1]

Plant scientists plan massive effort to sequence 10,000 genomes

Genome 10K is a project to sequence the genome of at least one individual from each vertebrate genus, approximately 10,000 genomes. It is a key milestone on

Large-scale phylogenetics

- Open Tree of Life.
- Compilation of all published phylogenetic knowledge.
- Uses a taxonomy (groups within groups) to stitch trees together where information is missing.
- Stores conflict among different published studies as a network.

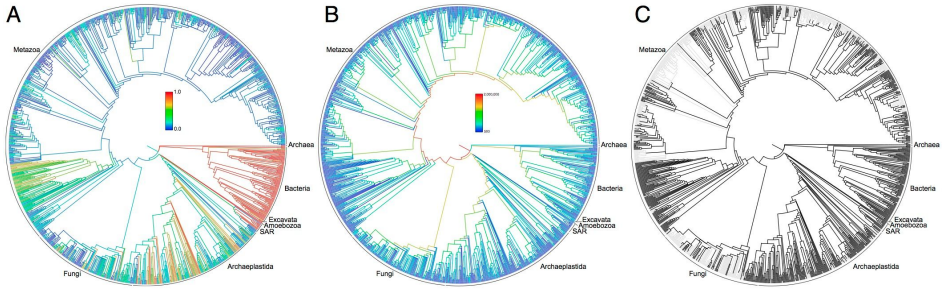


Fig. 1. Phylogenies representing the synthetic tree. The depicted tree is limited to lineages containing at least 500 descendants. (*A*) Colors represent proportion of lineages represented in NCBI databases. (*B*) Colors represent the amount of diversity measured by number of descendant tips. (*C*) Dark lineages have at least one representative in an input source tree.

Large-scale phylogenetics

- However, some groups are difficult to characterize as ‘species’, and therefore to confirm sampling.
- Most data does not end up in databases
- Manual curation and ranking remains necessary.

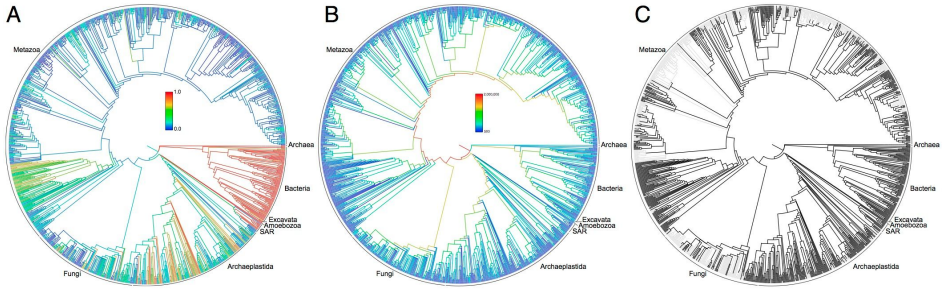


Fig. 1. Phylogenies representing the synthetic tree. The depicted tree is limited to lineages containing at least 500 descendants. (*A*) Colors represent proportion of lineages represented in NCBI databases. (*B*) Colors represent the amount of diversity measured by number of descendant tips. (*C*) Dark lineages have at least one representative in an input source tree.

Summary of large-scale phylogenetics

- Supermatrix approaches combine huge numbers of taxa for few or many genes. Often sparse matrices (missing data). Made possible by algorithmic and computational improvements to likelihood calculations.

- Supertree methods aim to combine information from multiple trees without the need to infer the actual sequence data for all samples at once.

- At the largest scale, both approaches are typically combined to *stitch together* the tree of life with both known (inferred) relationships, and estimated (taxonomy) relationships. *A lot of work remains to be done!*