Bayesian phylogenetics

(and basic principles of statistical inference)

Professor Alexei Drummond

Director of Centre for Computational Evolution
Department of Computer Science
University of Auckland

12th August 2019, Taming the BEAST Eh!, Squamish, Canada

Inference

Inference is the act of deriving logical conclusions from premises assumed to be true**:**

Premise: **If A is true, then B is true**
Premise: **A is true.**

Inference: **B is true.**

Premise: **All humans are mortal.**
Premise: **Alexei is a human.**

Inference: **Alexei is mortal.**

Statistical inference

Statistical inference generalises this to situations where the premises are not sufficient to draw conclusions without uncertainty.

Premise: **Squamish is a popular destination for rock climbers.**
Premise: **Alexei is visiting Squamish.**

Statistical inference: **Alexei is a rock climber?**

To perform statistical inference we need a theory of plausible reasoning.

Requirements for a theory of plausible reasoning

Cox suggested that a satisfactory theory of plausible reasoning in the face of uncertainty must satisfy the following requirements:

- Degrees of plausibility are represented by real numbers.
- There be a qualitative correspondence with common sense
- Consistency:
    1. All valid reasoning routes give the same result.
    2. Equivalent states of knowledge must have equivalent degrees of plausibility.

Probability: extending logic

These requirements are enough to uniquely identify the essential rules of probability theory [1,2]

- The probability $P(A \mid B)$ is the degree of plausibility of proposition A given that B is true.

- Product rule: $P(A \mid B, C)P(B \mid C) = P(A, B \mid C)$

- Sum rule: $P(A \mid B) + P(\bar{A} \mid B) = 1$

By convention, $P(A) = 0$ indicates A is certainly false while $P(A) = 1$ means A is certainly true.

1. Richard Cox, Am. J. Phys., 1946
2. E. T. Jaynes, Probability Theory: The Logic of Science, Cambridge Uni. Press, 2003

Bayes' rule arises directly from the rules of probability

$$P(A, B) = P(A \mid B)P(B)$$

$$P(A, B) = P(B \mid A)P(A)$$

$$P(B \mid A)P(A) = P(A \mid B)P(B)$$

Bayes' rule:

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$$

Notation

Strictly speaking, probabilities only ever concern propositions (i.e. with true or false values):

- Alexei is a Rock Climber
- $N = 5$

A statement such as $P(N)$ is therefore as meaningless as $P(\text{Alexei})$.

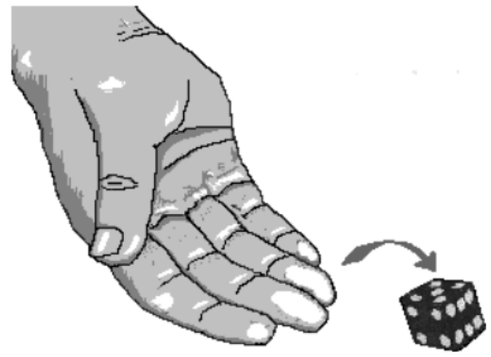
However, where propositions concern the value of a variable like N, we often use $P(n)$ as shorthand for $P(N = n)$.

Abusing notation, this is sometimes written as $P(N)$

Frequentist definition of probability

Traditionally, probability has been defined in terms of relative frequencies of outcomes of repeated random (weakly controlled) "experiments".

- N: Total number of rolls.
- n_5: Total number of 5's rolled.
- $P(d = 5) \equiv n_5/N$ as $N \to \infty$



There are several problems here:

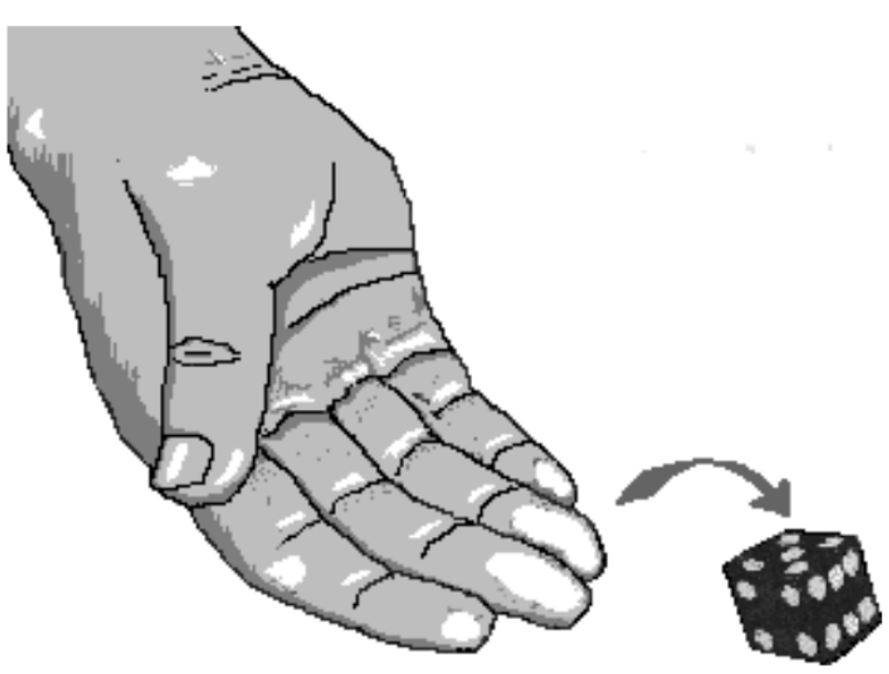
1. Experiments are assumed to be repeatable.
2. Assumes that randomness is a property of the system.
3. Completely ignores ~ 400 years of physics.

Bayesian interpretation of probability

- The Bayesian interpretation treats probability as a measure of the plausibility of propositions conditional on **available information.**

- A single proposition can therefore have multiple probabilities depending on the available information!

$P(d=5|I_{Einstein})=1$



$P(d=5|I_{Homer})=1/6$

It is impossible for a Die, with such determin'd force and direction, not to fall on such determin'd side, only I don't know the force and direction which makes it fall on such determin'd side, and therefore I call it Chance, which is nothing but the want of art… John Arbuthnot, 1692

Continuous hypothesis spaces

Propositions regarding continuous variables require special treatment.*

Suppose X may take any value between 0 and 10.

- The probability $P(X = x)$ will always be zero!
- Instead, define $P(x < X < x + dt) = f(x)dt$

> - $f(x)$ is a probability density.
> - It is normalized: $\int_0^{10} f(x)dx = 1$
> - It is positive: $f(x) \geq 0$
> - At a given point $f(x)$ may be > 1

Often, $f(x)$ follows the standard rules of probability.

*probabilities on continuous and discrete hypothesis spaces can be united by measure theory.

What is “data”

- An urn contains 11 balls: N_r red and $11 - N_r$ blue.

- Suppose we remove a ball (no peeking), record its colour, then replace it.

- Suppose we repeat this 2 more times, obtaining the sequence **R**,**B**,**R**.

How many of the balls in the urn are red? In other words, what is
$P(N_r | d_1 = R, d_2 = B, d_3 = R)$?

Urn example

Given the description of the process, it is more tractable to consider

$$P(d_1, d_2, d_3 \,|\, N_r) = P(d_3 \,|\, N_r, d_1, d_2) P(d_2 \,|\, N_r, d_1) P(d_1 \,|\, N_r)$$

- Knowing nothing about the internal arrangement of the balls in the urn, we must have $P(d_1 \,|\, N_r) = N_r/11$

- In general $P(d_2 \,|\, N_r, d_1)$ depends on the result of the first draw!

- Cheat by assuming urn is shaken between draws and we know nothing of physics, so that

$$P(d_2 \,|\, N_r, d_1) = P(d_2 \,|\, N_r)$$

Urn example

Now we can claim:

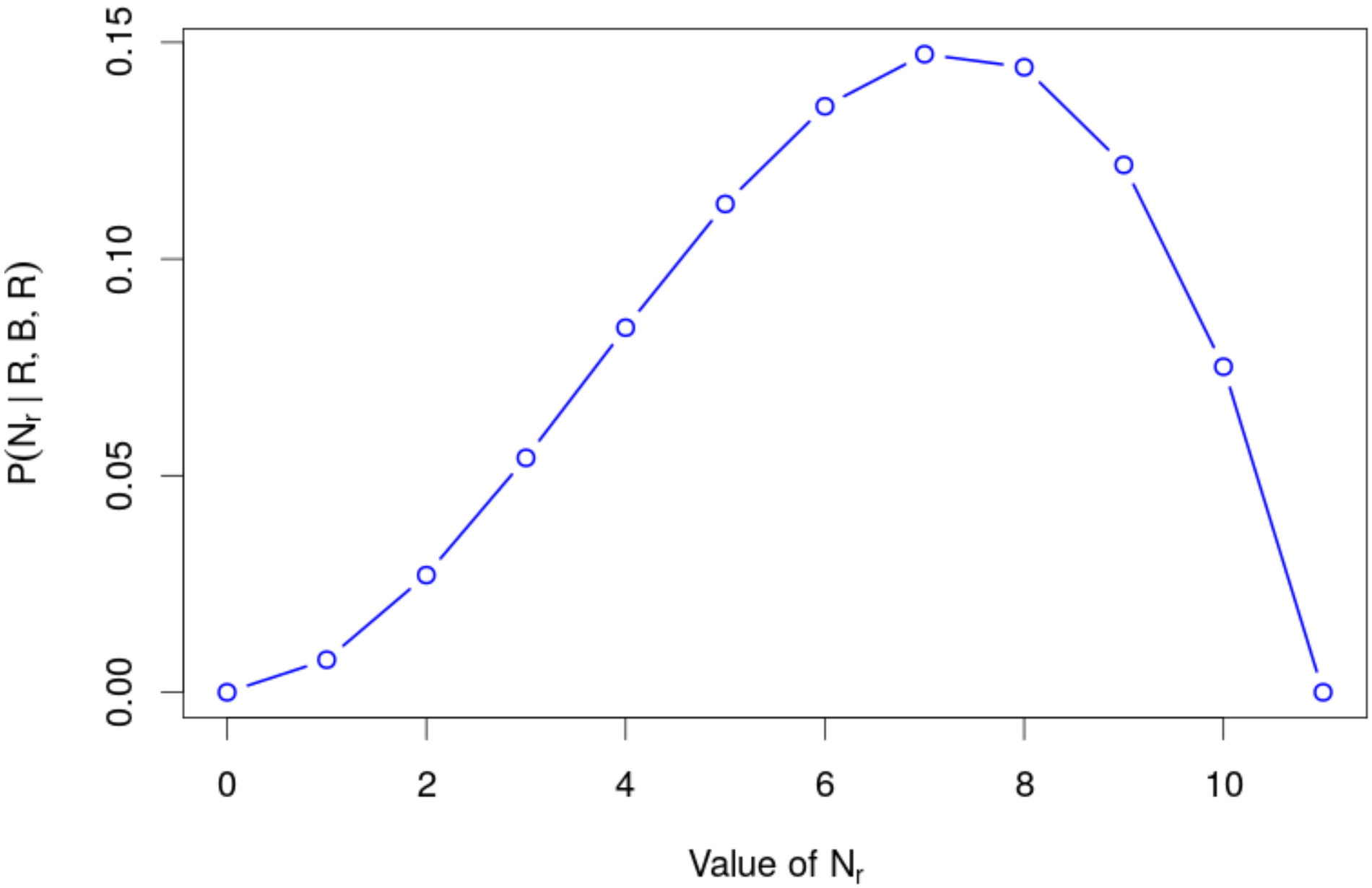
$$P(d_1 = R, d_2 = B, d_3 = R \,|\, N_r) = \frac{N_r}{11} \times \frac{(11 - N_r)}{11} \times \frac{N_r}{11}$$
$$= N_r^2(11 - N_r)/11^3$$

Applying the chain rule a couple of times yields:

$$P(N_r \,|\, R, B, R)P(R, B, R) = P(R, B, R \,|\, N_r)P(N_r)$$
$$P(N_r \,|\, R, B, R) = \frac{1}{P(R, B, R)} P(R, B, R \,|\, N_r)P(N_r)$$

The term $P(R, B, R)$ is a function only of the data: constant. The term $P(N_r)$ specifies the plausibility of each possible value of N_r in the absence of the data.

Urn example



Bayes' theorem

In answering this question we have accidentally used Bayes theorem:

$$P(\theta_M | D, M) = \frac{P(D | \theta_M, M) P(\theta | M)}{P(D | M)}$$

Here θ_M are parameters of some model M and D is data assumed to be generated by that model.

The components of the equation even have names:

- The posterior of θ: $P(\theta | D)$
- the likelihood of θ: $P(D | \theta)$
- the prior of θ: $P(\theta)$
- the marginal likelihood or evidence for M: $P(D | M)$

What is a prior probability?

A prior probability is:

- A probability!
- The probability of whatever you're interested in but in the absence of possibly relevant data.
- In principle, any two (rational) people with access to the same information should specify exactly the same prior.
- In practice this often isn't true.

Prior probabilities are necessary

- Isn't the need for priors a problem with the Bayesian approach?

  - **NO!**

- It is not possible to do inference without making assumptions.

- Priors allow previous knowledge to be incorporated.

- Frequentist (and Likelihoodist) methods also use priors: it's just not clear what they are!

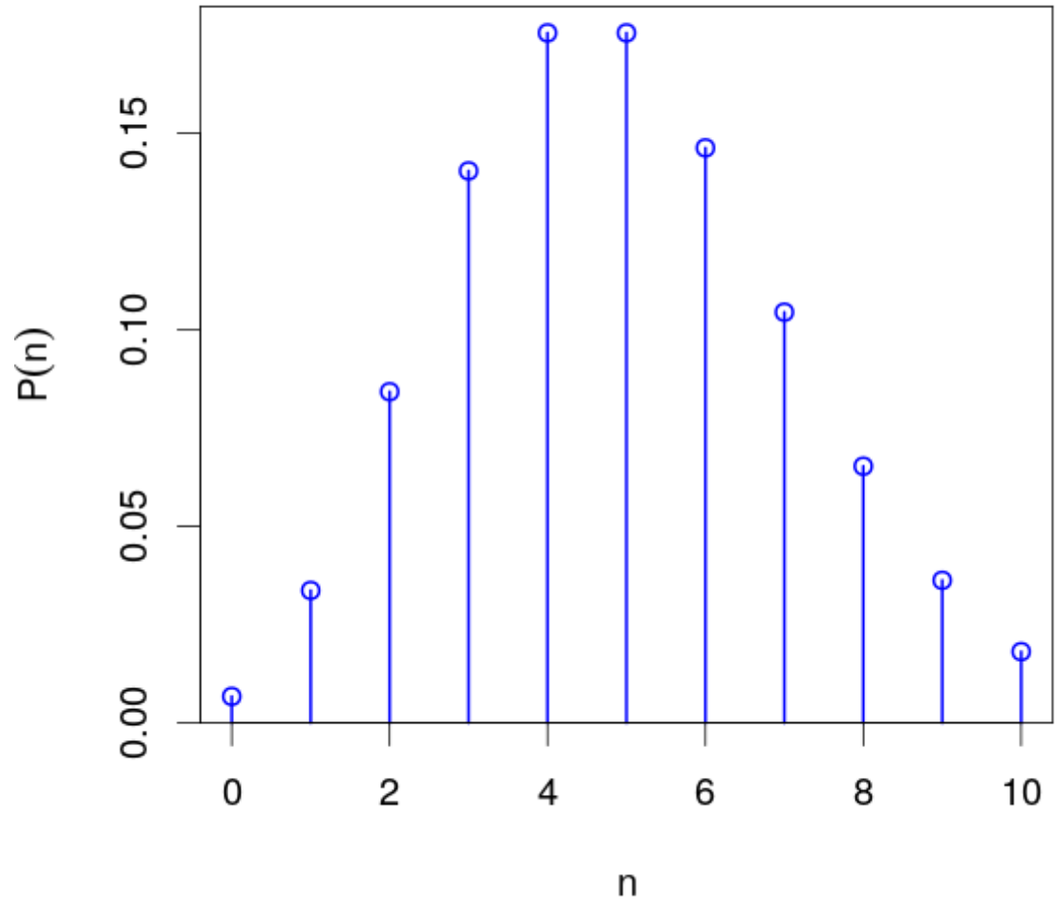
Priors for discrete variables

- Defining priors for discrete variables with finite bounds is often straight-forward.

- **Principle of indifference**.

- E.g. for discrete variables representing the number of events that occur in a given interval, a Poisson distribution may be appropriate.

- The **principle of maximum entropy** works beautifully here for specifying priors under constraints (e.g. Geometric distribution is the maximum entropy distribution for a non-negative random variable with known prior mean)



Priors for continuous variables

For a continuous variable $a < x < b$, sensible priors may include

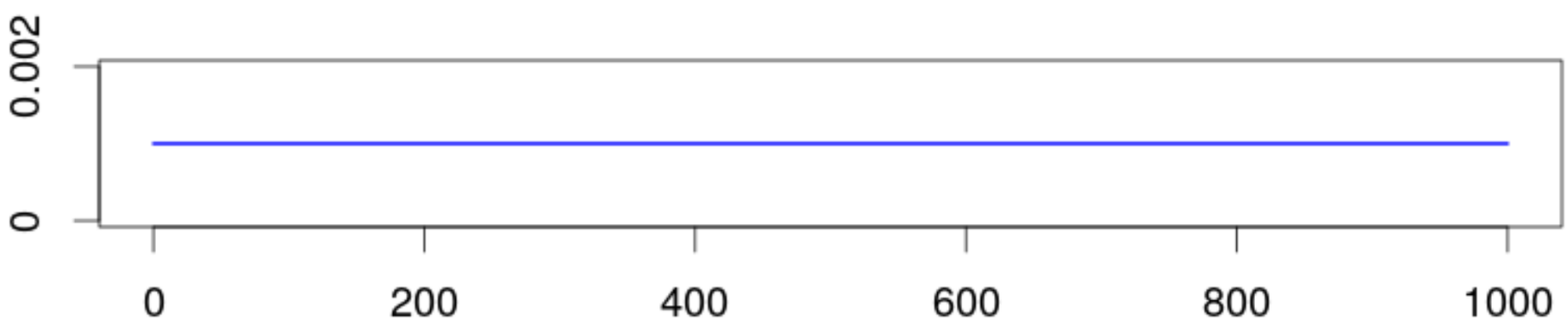
- The uniform distribution $f(x) = 1/(b-a)$
- A Beta distribution, $f(x) \propto (x-a)^{\alpha-1}(b-x)^{\beta-1}$

For a rate variable $\lambda > 0$

- May fix $f(\lambda) = c$ to indicate complete ignorance
  - But this is probably not what you want!
  - Places almost all probability on large values
- $f(\lambda) = 1/\lambda$ is a better choice (uniform in log-space)

Improper priors

Hold on, how can we choose a value of c in $f(\lambda) = c$ so that $f(\lambda)$ is normalized on the domain of λ?

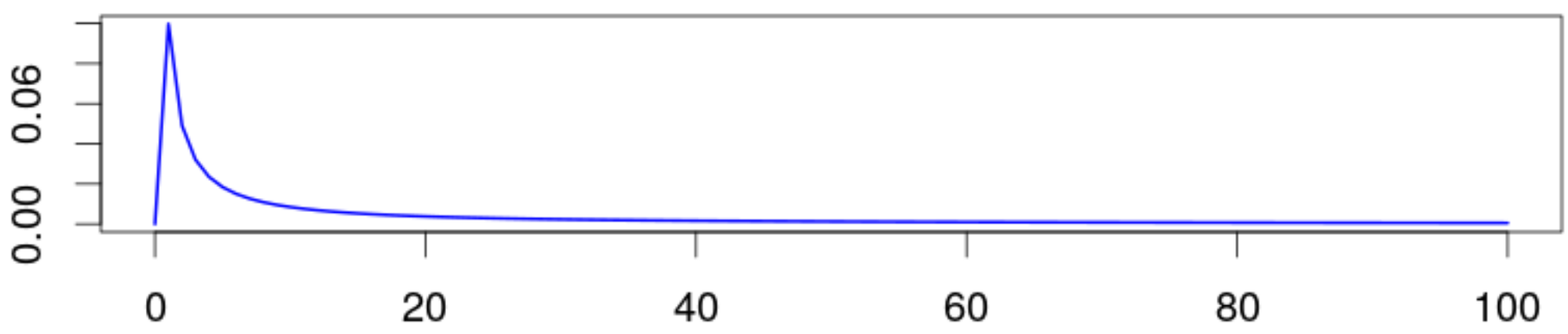


We can't! This $f(\lambda)$ is not a true probability density.

Improper priors

It is important to remember that:

- One almost never knows absolutely nothing.
- Upper and lower bounds can almost always be placed.
- The log-normal prior can be considered a normalizable replacement for the $1/x$ prior.



Which prior is best?

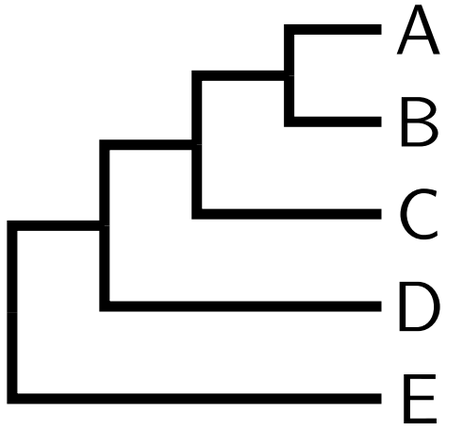
- **Only the person doing the analysis can answer this!**

- **Priors encapsulate expert knowledge (or its absence). This is your opportunity to contribute your hard-won expertise to the analysis.**

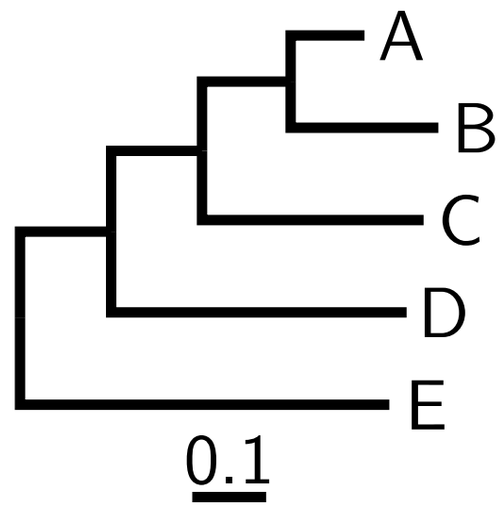
What about phylogenetics?

Phylogenies are complex hypotheses

rooted trees

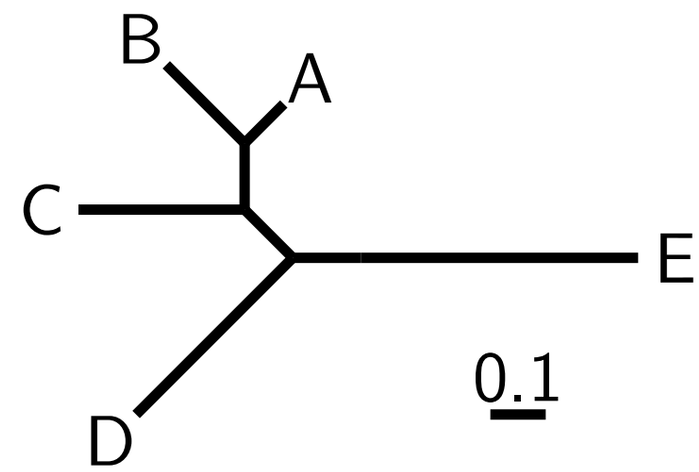


(a) cladogram



(b) phylogram

unrooted tree



(c) unrooted tree

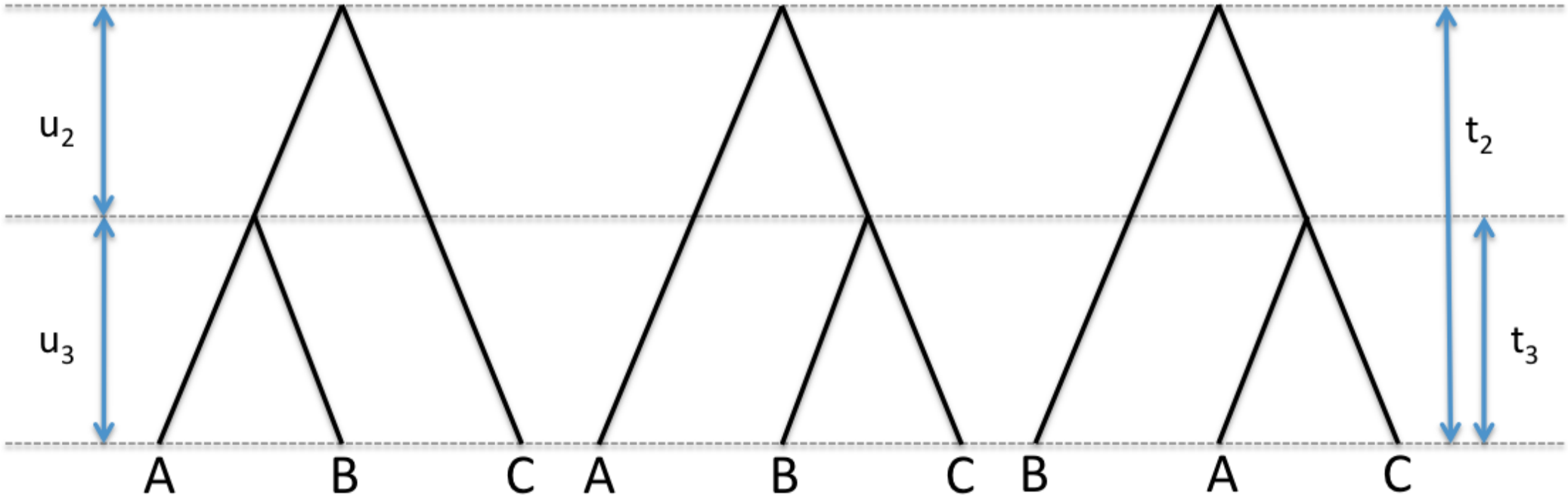
((((A, B), C), D), E);

((((A:0.1, B:0.2):0.12, C:0.3):0.123, D:0.4):0.1234, E:0.5);

branches (edges) and their lengths, nodes, tips (leaves)

The tip-labeled **time-tree**

A tip-labeled time-tree is described by a *tip-labeled ranked topology* of size k and *coalescent times*, $\mathbf{u} = \{u_2, \ldots, u_k\}$.



These time-trees of size 3 can be interpreted as describing the possible alternative evolutionary histories for three species or (uniparental) ancestries of the three individuals represented by the labeled tips.

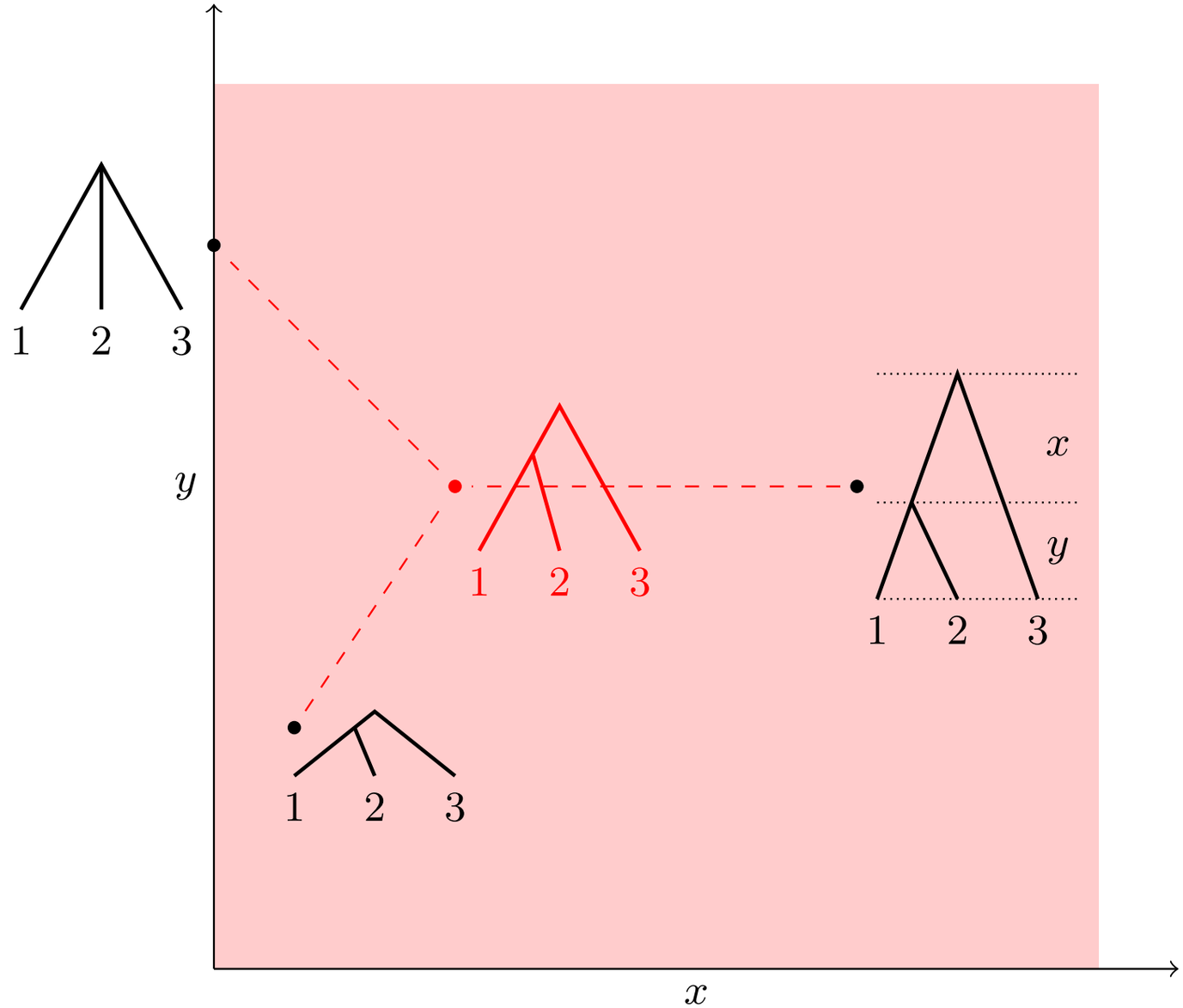
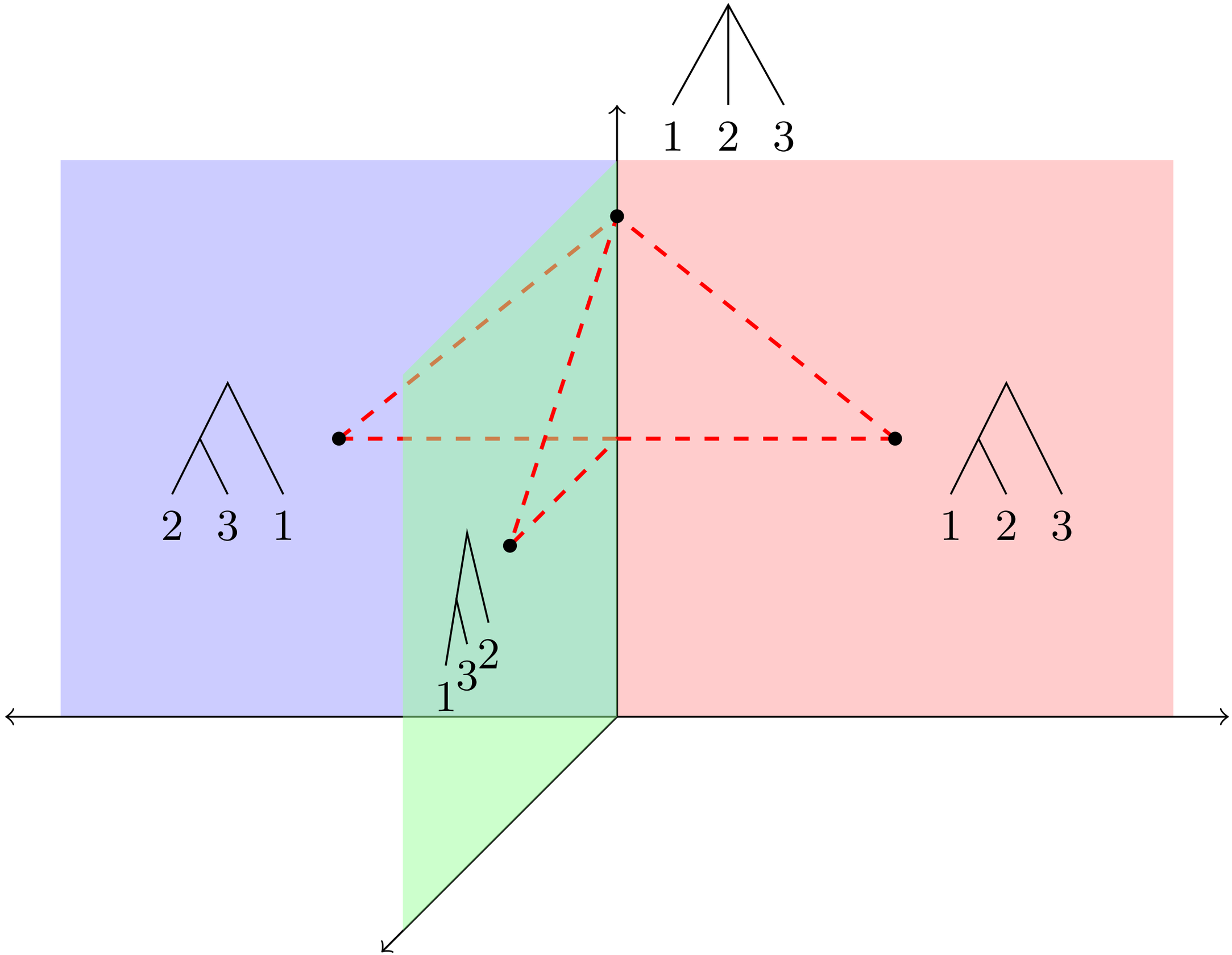


Figure: A Euclidean two-dimensional space representing the space of all possible time-trees for the topology ((1,2),3). There are two parameters, x and y, one for each of the two inter-coalescent intervals, the sum of which is the age of the root ($t_{root} = x + y$). Three trees are displayed, along with their arithmetic mean tree, also called the *centroid*. The dashed lines show the path connecting each of the three trees to the mean tree by the shortest distance (i.e. their deviations from the mean).


1 2 3
2 3 1
1 3 2
1 2 3

Tip-labeled time-trees of size 4

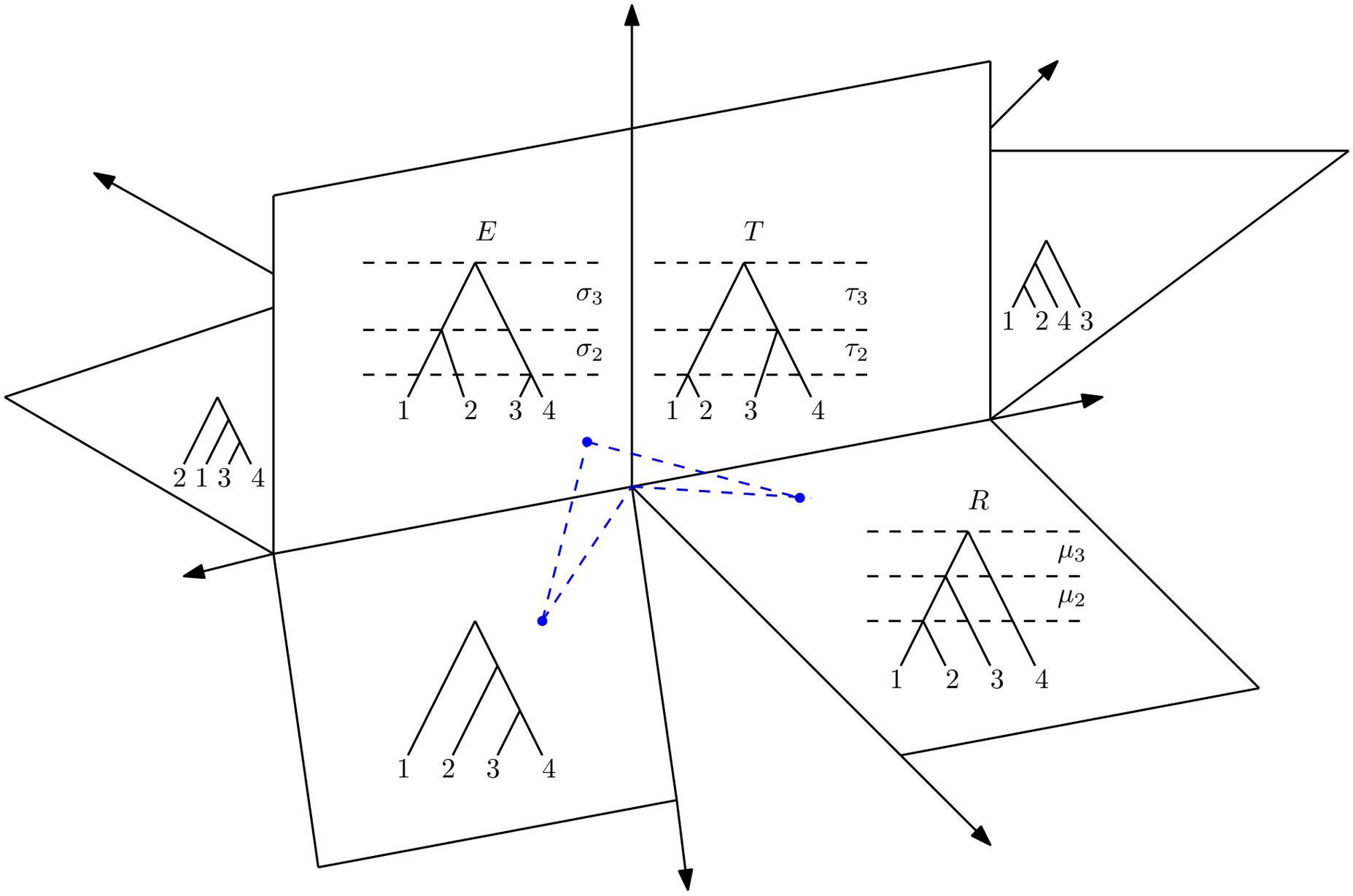
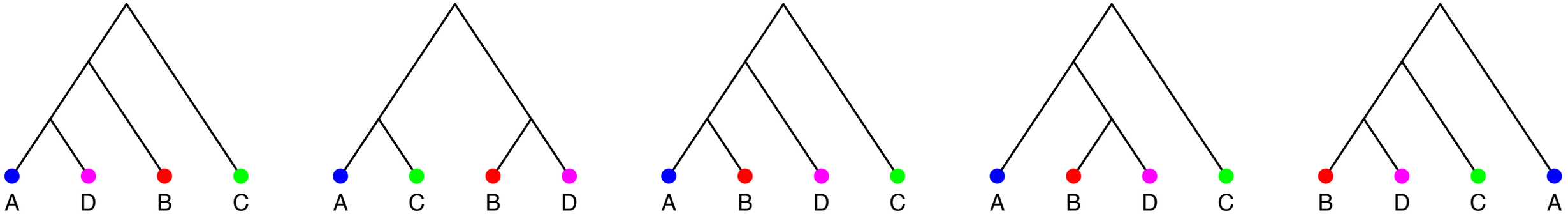


Figure: Three-dimensional projection of 4-dimensional τ-space T_4.

Figure from Gavryushkin and Drummond (2016)



15 possible (unranked) trees of 4 individuals/species

105 possible trees of 5 individuals/species

945 possible trees of 6 individuals/species

Question: How many possible trees are there relating seven taxa?

How many trees are there?

For n species there are

$$T_n = 1 \times 3 \times 5 \times \cdots \times (2n-3) = \frac{(2n-3)!}{(n-2)!2^{n-2}}$$

rooted, tip-labelled binary trees:

n	#trees	
4	15	enumerable by hand
5	105	enumerable by hand on a rainy day
6	945	enumerable by computer
7	10395	still searchable very quickly on computer
8	135135	about the number of hairs on your head
9	2027025	greater than the population of Auckland
10	34459425	$\approx$ upper limit for exhaustive search
20	8.20×10^{21}	$\approx$ upper limit of branch-and-bound searching
48	3.21×10^{70}	$\approx$ the number of particles in the Universe
136	2.11×10^{267}	number of trees to choose from in the "Out of Africa" data (Vigilant *et al*. 1991)

Counting different types of trees

n	#shapes	#trees, $\lvert\mathcal{T}_n\rvert$	#ranked trees	#fully ranked trees
2	1	1	1	1
3	1	3	3	4
4	2	15	18	34
5	3	105	180	496
6	6	945	2700	11056
7	11	10395	56700	349504
8	23	135135	1587600	14873104
9	46	2027025	57153600	819786496
10	98	34459425	2571912000	56814228736

Table: The number of unlabeled rooted tree shapes, the number of labelled rooted trees, the number of labelled ranked trees (on contemporaneous tips), and the number of fully-ranked trees (on distinctly-timed tips) as a function of the number of taxa, n.

Consensus
1. Tarsius_syrichta
2. Lemur_catta
3. Homo_sapiens
4. Pan
5. Gorilla
6. Pongo
7. Hylobates
8. Macaca_fuscata
9. M_mulatta
10. M_fascicularis
11. M_sylvanus
12. Saimiri_sciureus

~10,000 nucleotides

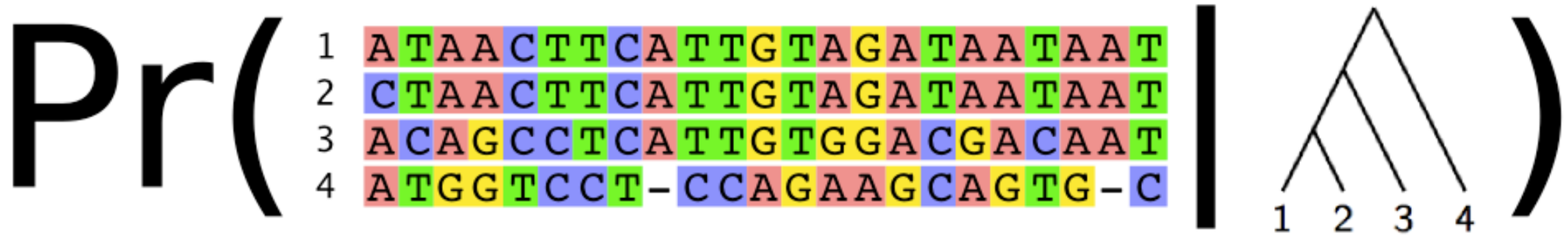
The phylogenetic likelihood

Felsenstein (1981)

Besides coding for function, DNA serves as a record of evolutionary history.

But, in order to reconstruct the phylogenetic tree we need to have a procedure to evaluate each tree in light of the sequence data.

One means of evaluating a tree would be to calculate the **probability of the data** under a statistical model of DNA evolution.



This is known as the **Likelihood** of the tree. One method of reconstructing the evolutionary history is then to find the tree that has the **Maximum Likelihood**.

Bayes theorem

$$P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{P(D)}$$

posterior

likelihood prior

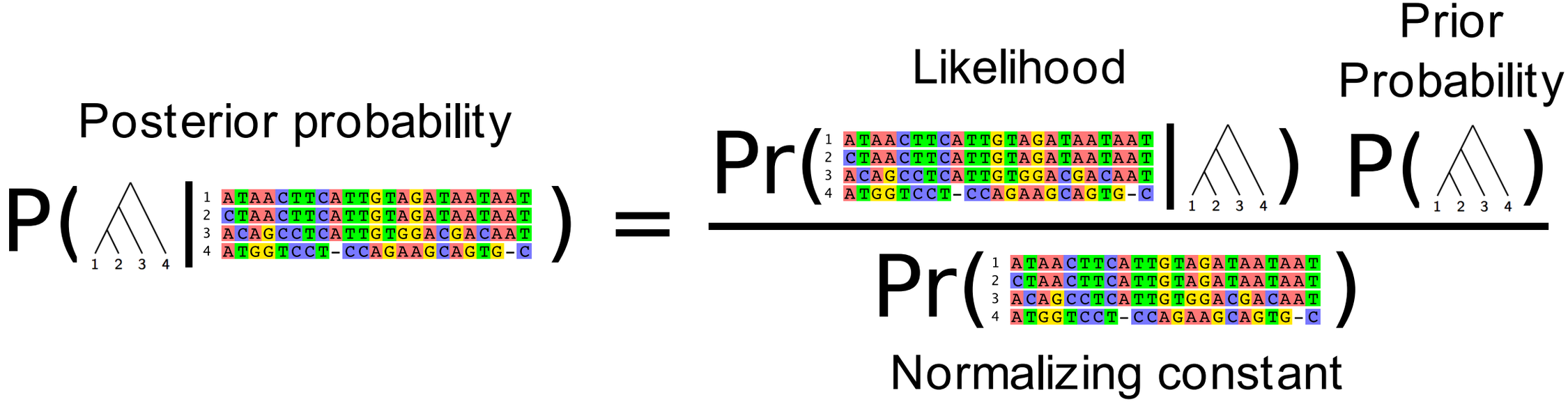
marginal likelihood

Bayesian reconstruction of phylogenetic trees

Yang & Rannala (1997), Mau, Newton & Larget (1998), Wilson and Balding (1998)

In phylogenetics what we want is the **probability of each tree** given the aligned sequence data.

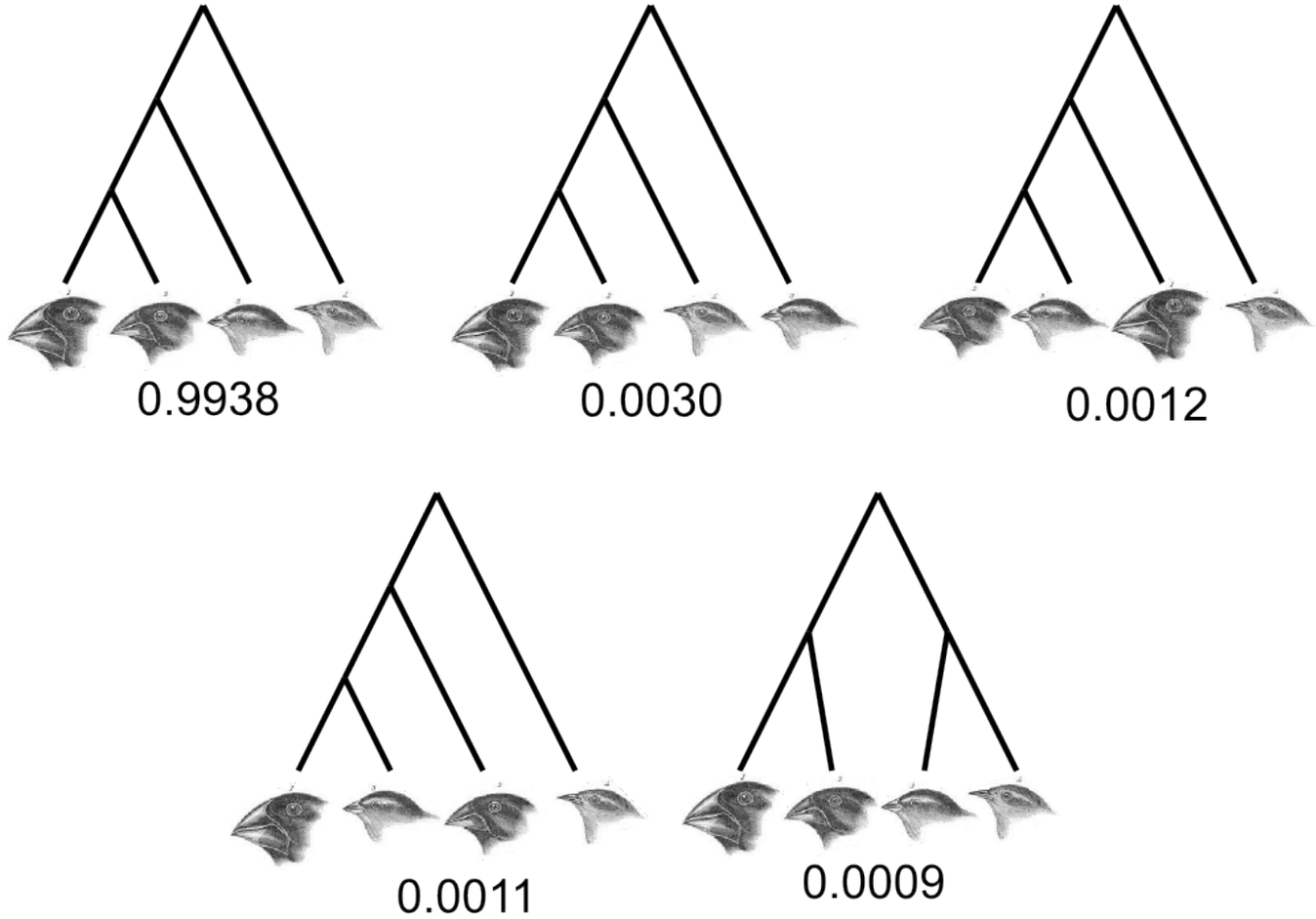
We can compute the probability of a tree using **Bayes Theorem**:



Using the **Markov chain Monte Carlo algorithm** we can produce a sample of trees from this posterior probability distribution **without knowing the marginal likelihood (normalizing constant)**.

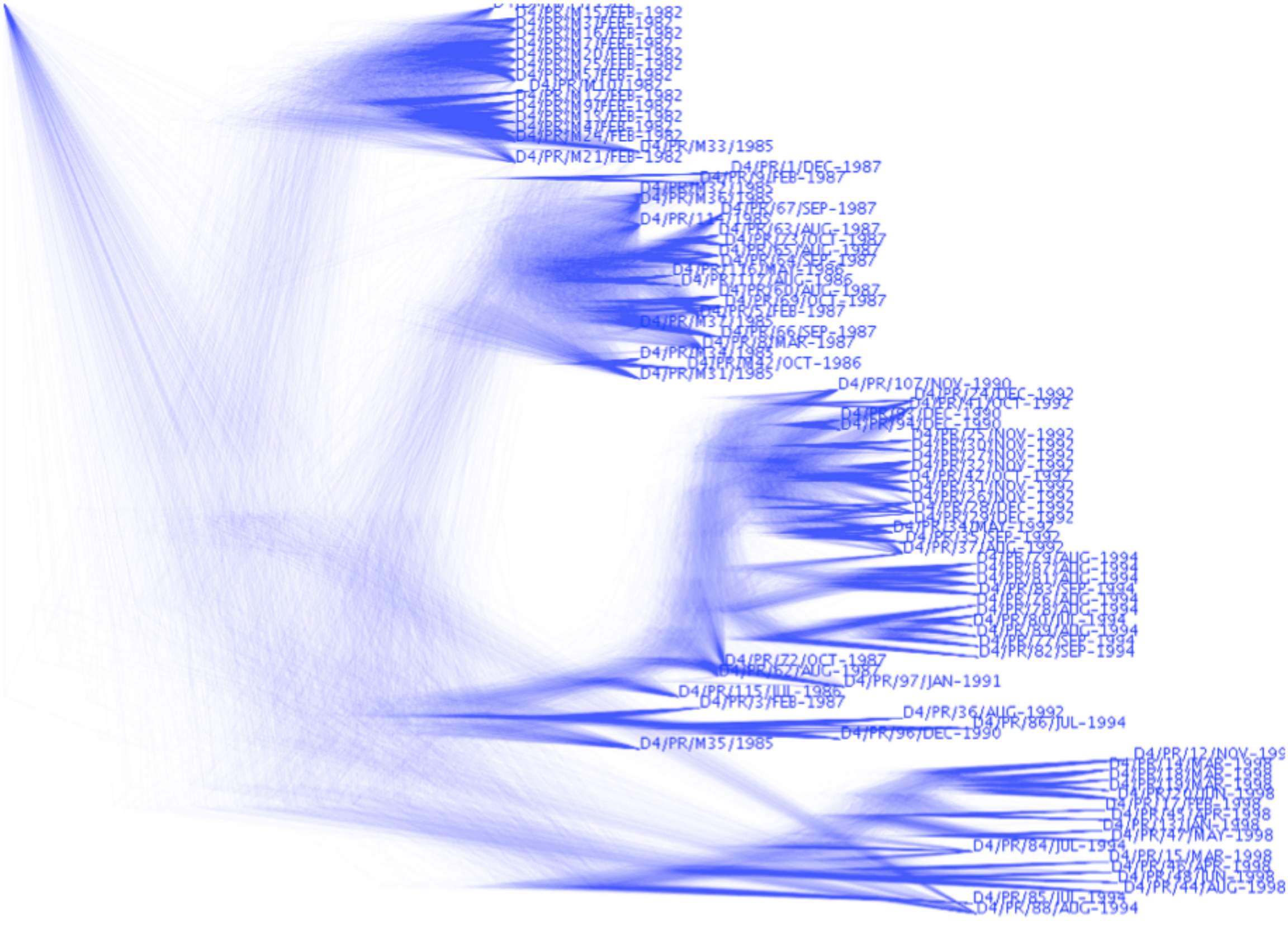
The posterior distribution on Darwin's Finches

```
1  ATAACTTCATTGTAGATAATAAT
2  CTAACTTCATTGTAGATAATAAT
3  ACAGCCTCATTGTGGACGACAAT
4  ATGGTCCT-CCAGAAGCAGTG-C
```



This posterior probability distribution was computed using **Markov chain Monte Carlo** implemented in the BEAST software package.

The posterior distribution of larger trees



Elaborating the model

Basic model: (posterior proportional to likelihood x prior)

$$P(T \mid D) \propto \Pr(D \mid T)P(T)$$

Substitution model parameters:

Assuming independence

$$P(T, Q \mid D) \propto \Pr(D \mid T, Q)P(T)P(Q)$$

Substitution model and tree branching process parameters:

Assuming independence

$$P(T, Q, \theta \mid D) \propto \Pr(D \mid T, Q)P(T \mid \theta)P(\theta)P(Q)$$

The phylogenetic posterior

Standard application of Bayes theorem gives the posterior:

$$P(T, Q, \theta \mid D) = \frac{\Pr(D \mid T, Q, \theta) P(T, Q, \theta)}{\Pr(D)}$$

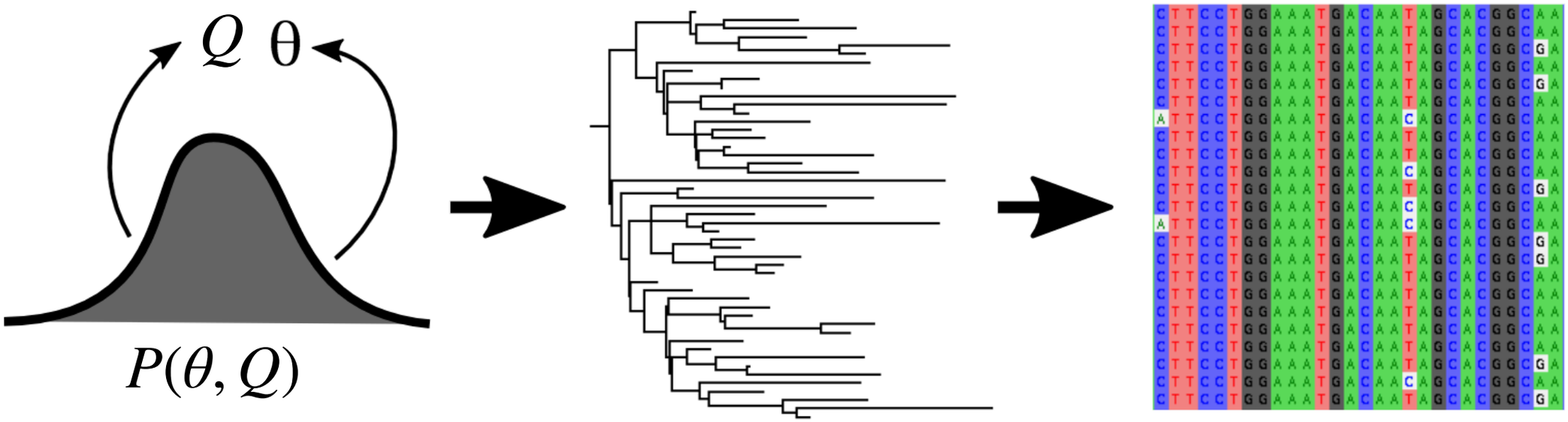
But you will normally see it written like this

$$P(T, Q, \theta \mid D) = \frac{1}{\Pr(D)} \Pr(D \mid T, Q) P(T \mid \theta) P(\theta) P(Q)$$

- the probability of the data doesn't depend on θ **except through the tree**.
- the prior probability of the tree depends on θ but not on Q.
- the prior probability of θ and the prior probability of Q are independent.

The neutrality assumption

Because of the way we've factorized the joint probability for the data and model parameters, we are implicitly assuming that our alignment could have been produced in the following fashion:

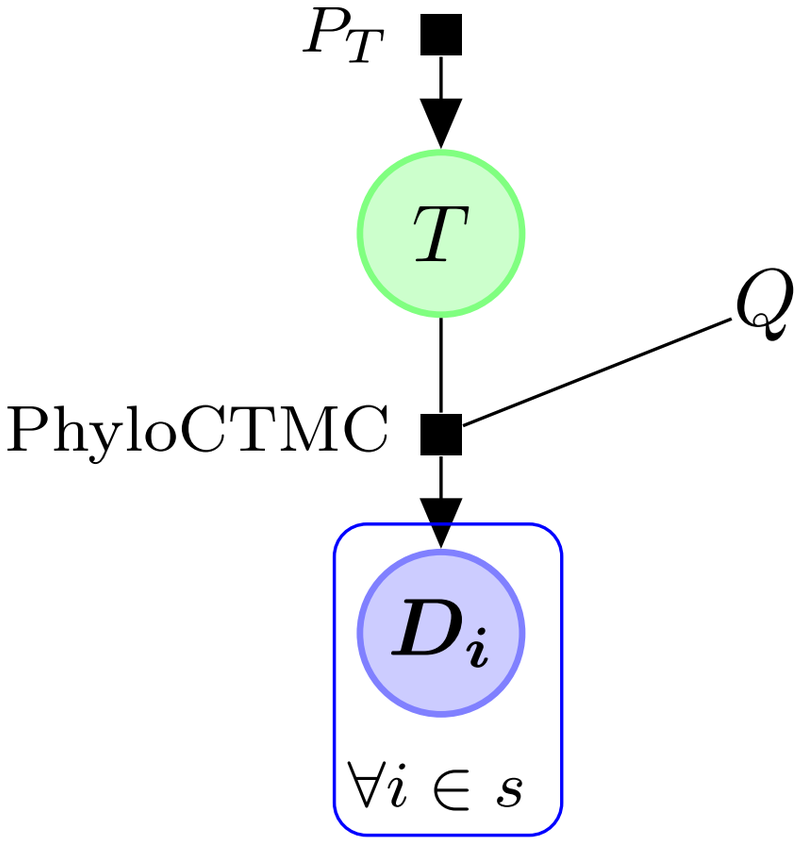


Sample parameters from prior

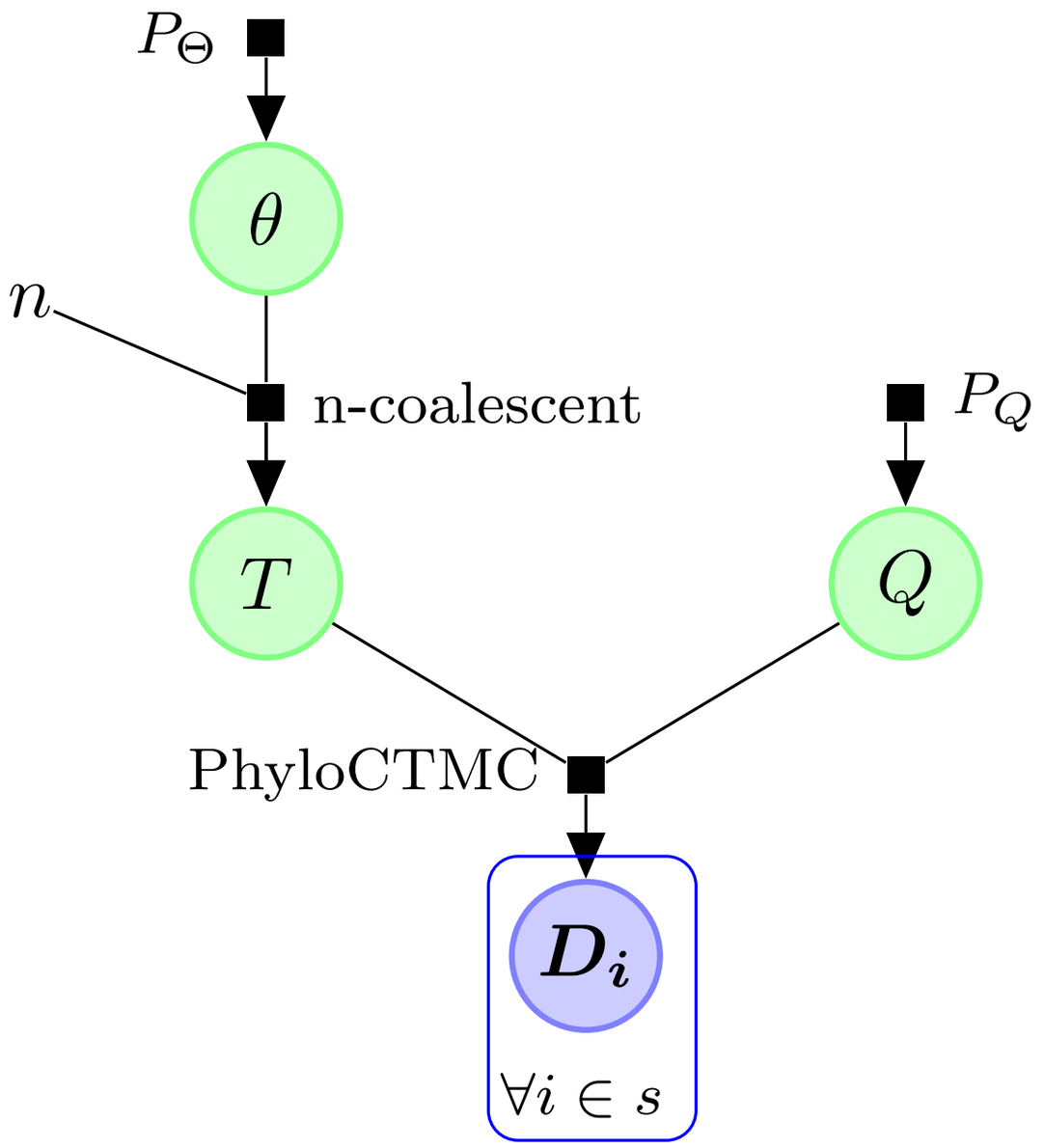
Sample tree from prior

Simulate sequence alignment

Graphical models

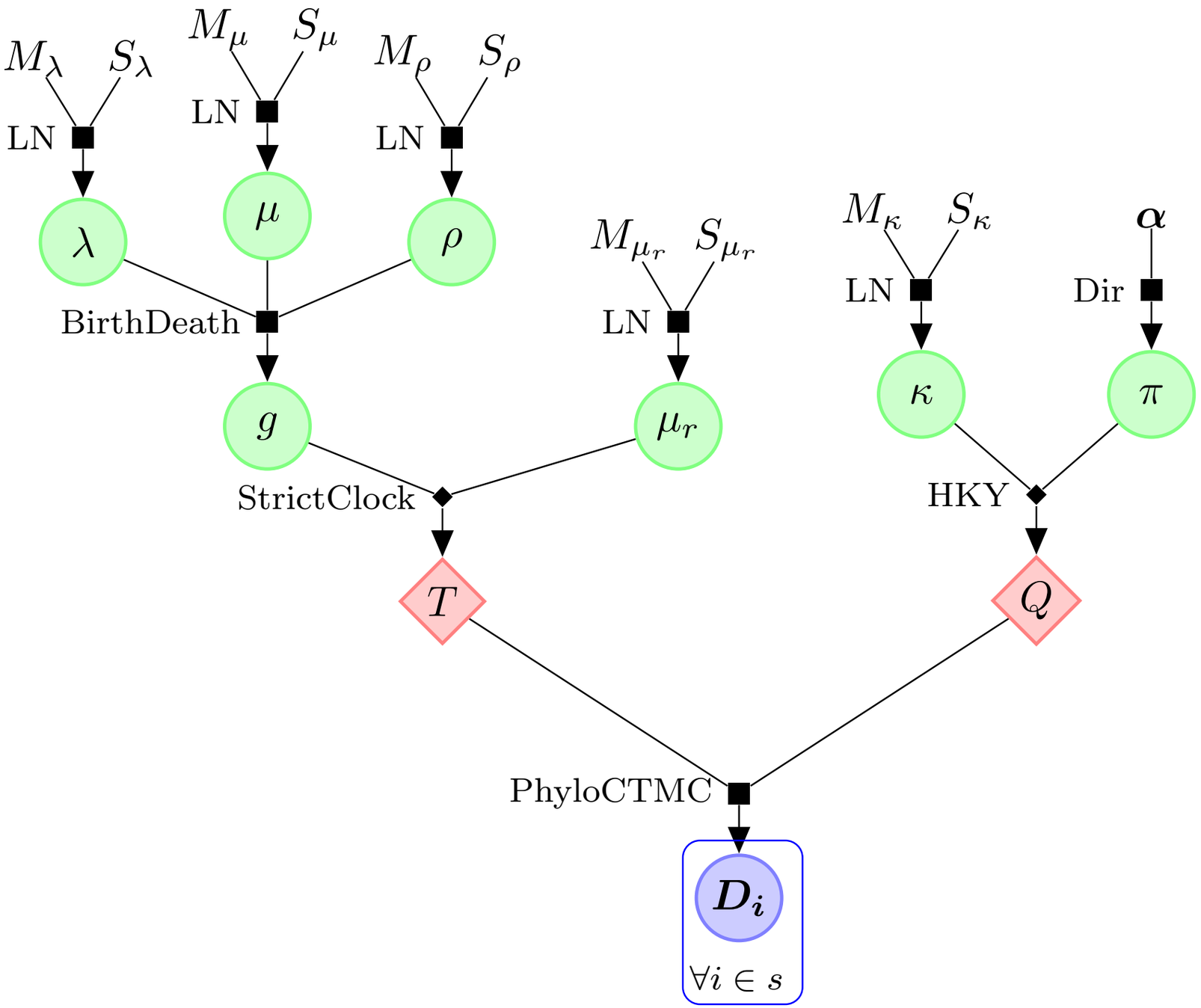


$$P(T|\boldsymbol{D}) \propto \Pr(\boldsymbol{D}|T)P(T)$$



$$P(T, Q, \theta|\boldsymbol{D}) \propto \Pr(\boldsymbol{D}|T, Q)P(T|\theta)P(\theta)P(Q)$$

Graphical models for phylogenomics



$$P(g, \mu_r, \lambda, \mu, \rho, Q|\boldsymbol{D}) \propto \Pr(\boldsymbol{D}|\mu_r g, Q) P(g|\lambda, \mu, \rho) P(\lambda) P(\mu) P(\rho) P(\mu_r) P(Q)$$

plos.org

PLOS COMPUTATIONAL BIOLOGY

BROWSE PUBLISH ABOUT

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis

Remco Bouckaert, Timothy G. Vaughan, Joëlle Barido-Sottani, Sebastián Duchêne, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, Graham Jones, Denise Kühnert, Nicola De Maio, Michael Matschiner, Fábio K. Mendes, Nicola F. Müller, Huw A. Ogilvie, Louis du Plessis, Alex Popinga, Andrew Rambaut, David Rasmussen, Igor Siveroni, Marc A. Suchard, Chieh-Hsi Wu, Dong Xie, Chi Zhang, Tanja Stadler, Alexei J. Drummond [view less]

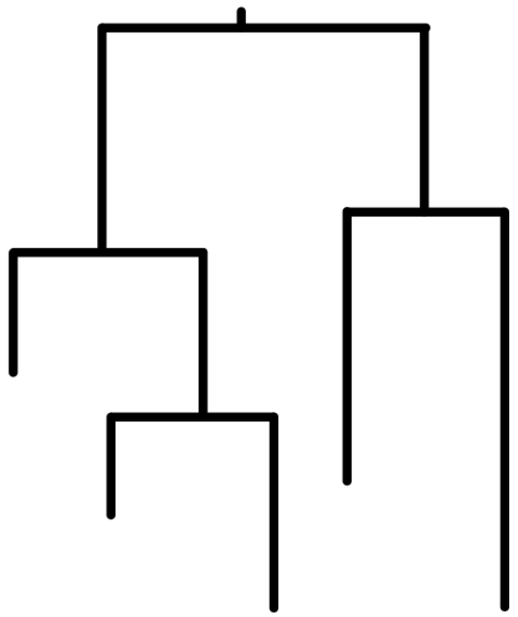
Version 2 Published: April 8, 2019 • https://doi.org/10.1371/journal.pcbi.1006650 • >> See the preprint

BEAST 2.5

- Allows a user to construct a wide range of phylogenetic models to apply to sequence alignments and comparative data using the BEAUti user interface.
- The major components of the phylogenetic models are
  - the time-tree prior (coalescent or birth-death models)
  - the substitution model (nucleotide, codon, trait evolution)
  - the site model (how the substitution model varies among sites/loci)
  - the molecular clock model - strict, relaxed, random local clock
  - How such models are “partitioned” when there are multiple data partitions
- Major differences in models are handled by different top-level templates
- New sub-models and templates can be designed by 3rd party developers

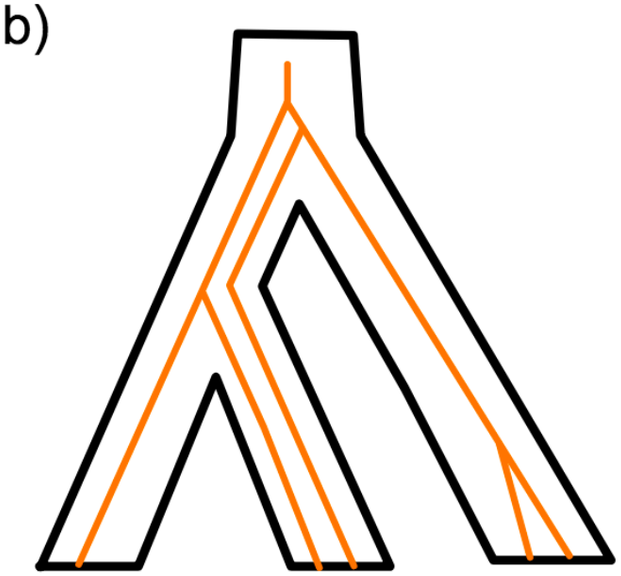
Some BEAST 2.5 tree priors

a)



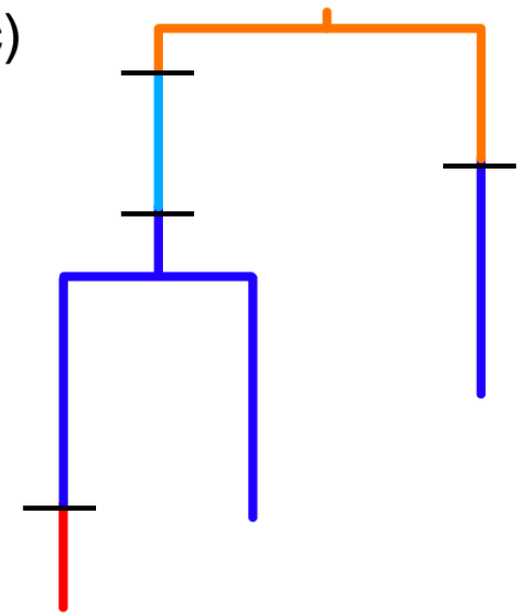
Tip-dated time tree
(leaf times conditioned on)

b)



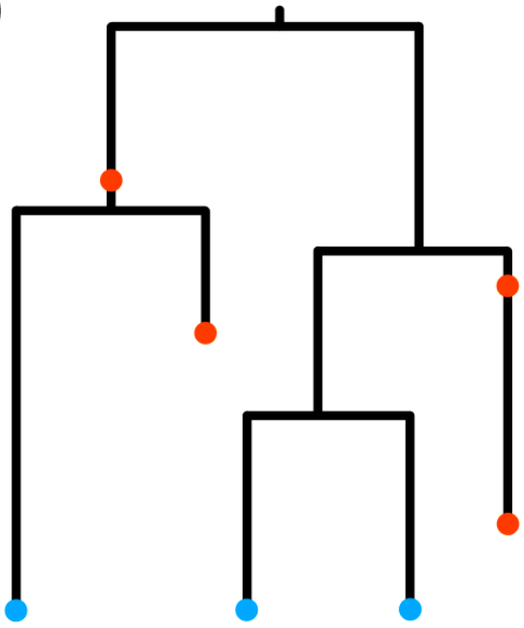
Multi-species coalescent

c)



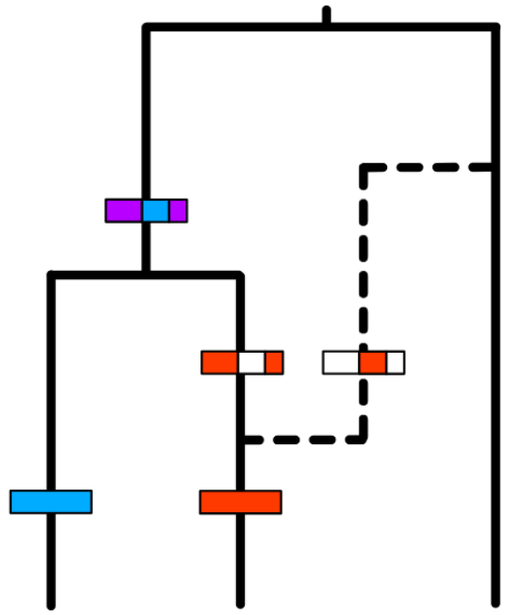
Multi-type time tree

d)



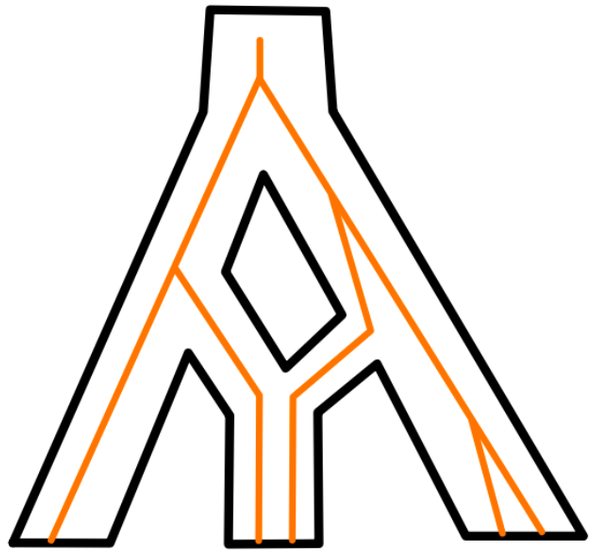
sampled ancestor
time tree

e)



ancestral gene conversion
graph

f)



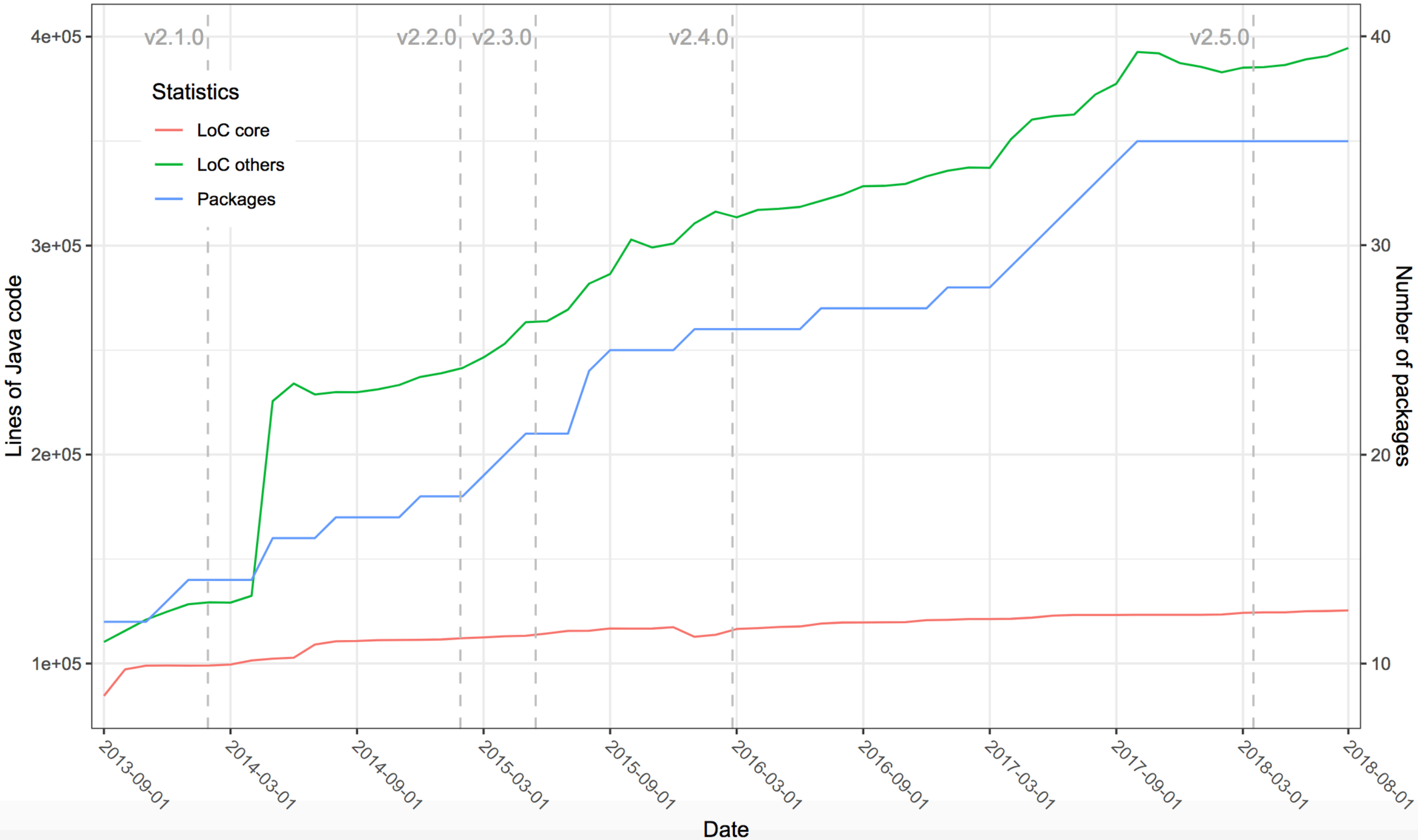
Species network
with embedded gene tree

BEAST 2.5 packages

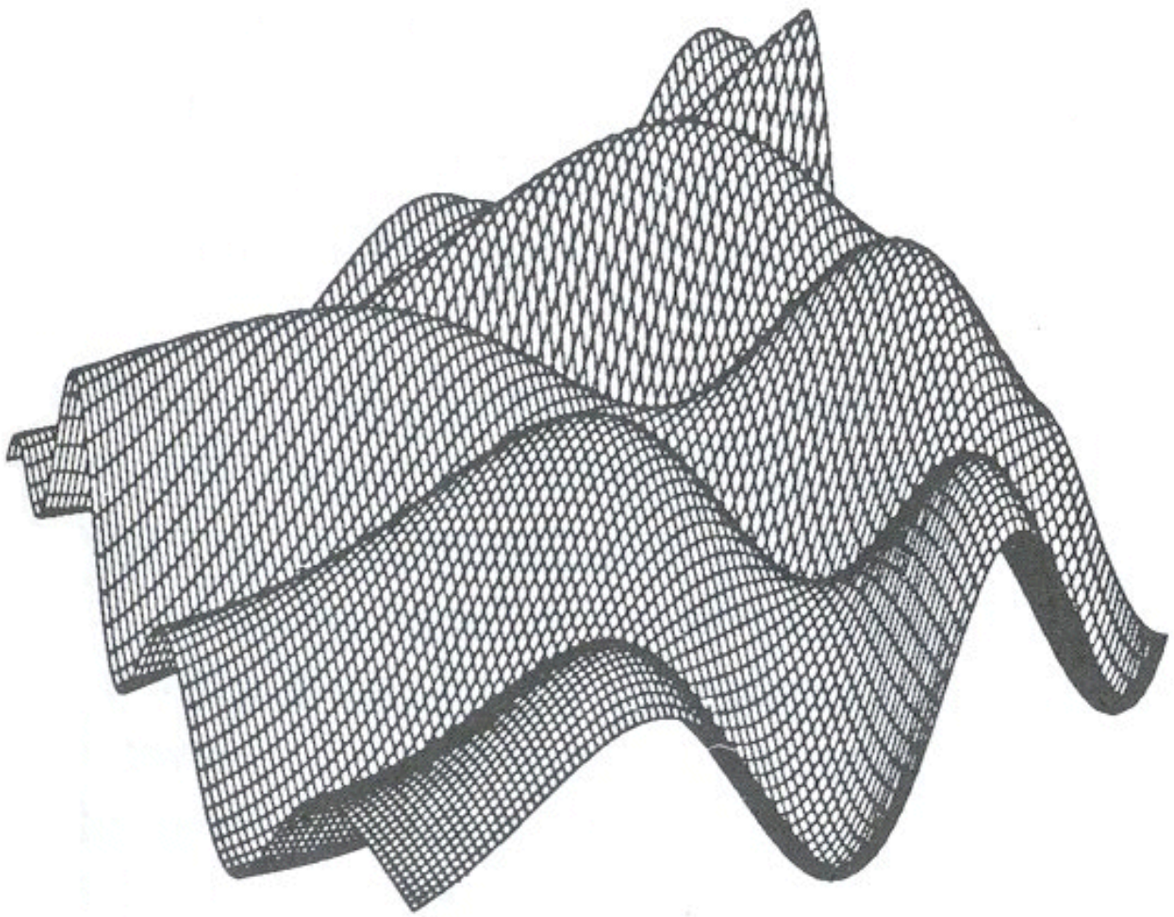
Table 1. BEAST 2 packages

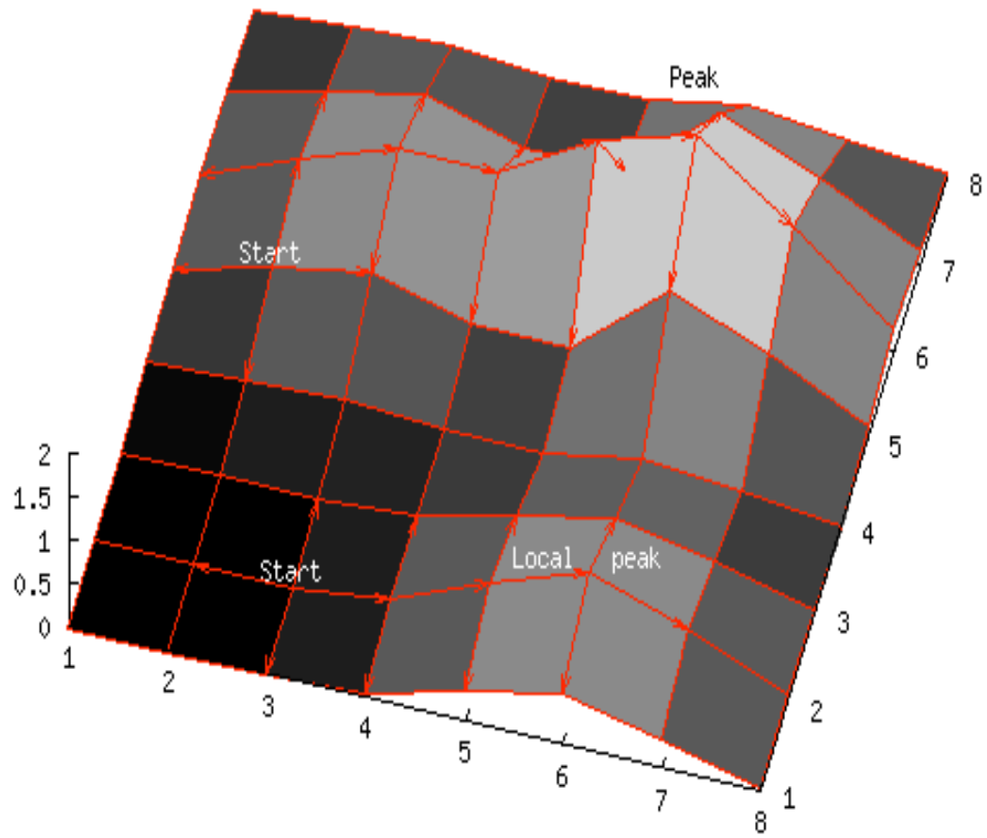
Package	Subspecification	Special Feature	Reference
Substitution models :			
bModelTest	nucleotide subst.[1] model	model averaging, model comparison	[15]
SSM	nucleotide. subst. model	standard named nucleotide models	-
CodonSubstModels	codon subst. model	M0	[16, 17]
MM	morphological model	discrete	[18]
BEASTvntr	microsatellite model	variable number of tandem repeat data	[19, 20]
RBS	subst.[1] model	model averaging for contiguous site partitions	[6]
PoMo	nucleotide subst. model	mutation-selection	[21]
		& species tree	[13]
Site models :			
MGSM	site model	multi-gamma & relaxed gamma	[22]
substBMA	site model	Dirichlet mixture model for site partitions	[8]
Branch model :			
FLC	molecular clock model	strict and relaxed clocks within local clock model	[23]
Tree models :			
SA	unstructured population, non-par.[2]	sampled ancestor* / fossilized BD[3]	[10]
CA	unstructured population, non-par.	calibration density, sampling rate estimate	[24]
BDSKY	unstructured population, non-par.	BD serial skyline*, BD serial sampling	[9]
		BD incomplete sampling (no ψ)	[25]
phylodynamics	unstructured population, par.[2]	deterministic closed SIR, stochastic closed SIR	[26]
		birth-death SIR	[27]
EpiInf	unstructured population, par.	prevalence estimation, particle filtering	[28]
PhyDyn	structured population, par.	define epidemic model by ODEs[4]	[29]
MultiTypeTree	structured population	structured tree	[5]
BadTrIP	structured population	within-host, transmission inference	[14]
BDMM	structured population	multitype BD[3] model	[30]
BASTA	structured population	approx. structured coalescent	[31]
MASCOT	structured population	approx. structured coalescent and time variant GLM's	[32, 33]
SCOTTI	structured population	transmission inference	[34]
BREAK AWAY	geographical model	break-away model of phylogegraphy	[35]
GEO SPHERE	geographical model	whole world phylogeography	[36]
Network models :			
BACTER	network model	clonal frame ancestral recombination graph	[11, 37]
SpeciesNetwork	network model	species networks	[12]
Nested models :			
DENIM	multispecies coalescent	species tree estimation with gene flow	[38]
SNAPP	multispecies coalescent	from independent biallelic markers	[7]
STACEY	multispecies coalescent	species delimitation & species tree estimation	[39]
StarBEAST 2	multispecies coalescent	faster, species tree clocks, FBD-MSC, AIM	[40–43]
Model selection :			
MODEL SELECTION	model selection	path sampling, stepping stone	[44]
NS	model selection	nested sampling	[45]
Simulation tools :			
MASTER	simulation	stochastic population dynamics simulation	[46]

BEAST 2.5 development



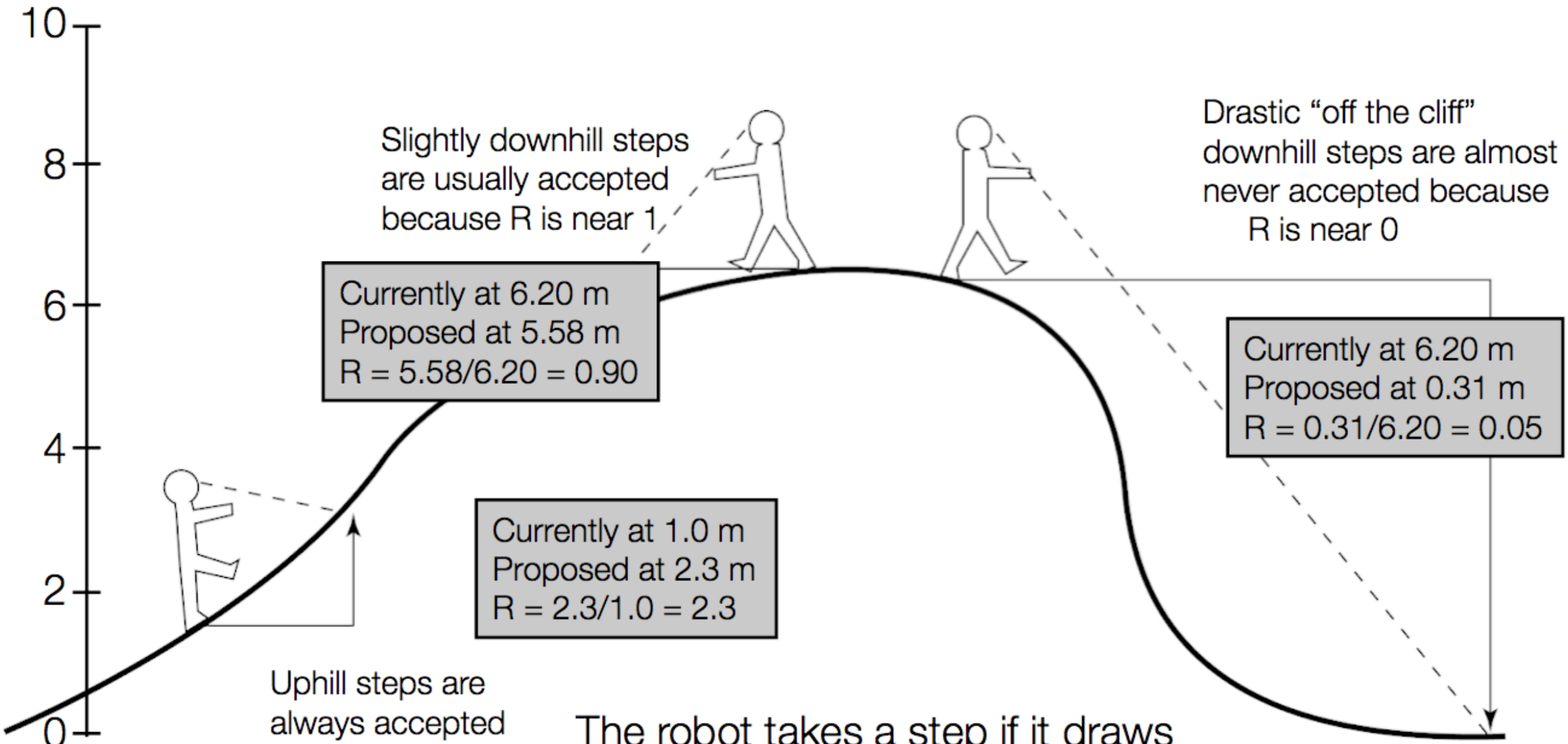
A probability distribution on tree space as a hilly landscape





- This space can be **sampled** in a Bayesian analysis with MCMC
- The peak can be identified by a **search algorithm** in the context of maximum likelihoods

Markov chain Monte Carlo (MCMC) robot



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

Slide courtesy of Paul O Lewis

MCMC animations

Conclusions

- Bayesian statistical inference derives natural from the rules of probability, and is the only inferential method that provides a consistent way to build up knowledge as evidence accumulates, and to bridge differences in prior knowledge.

- The hypothesis space of phylogenetics (tree space + parameters) has a distinctive structure that frustrates attempts to use standard statistical inference software. Thus specialist inference software such as BEAST and MrBayes have been developed.

- Evolutionary biology and phylogenetics is a statistical science, in which mature statistical inference methods are now routinely used.

- Research in this field depends on continuous development and maintenance of large software packages, and this is currently still a challenge for science funding models.