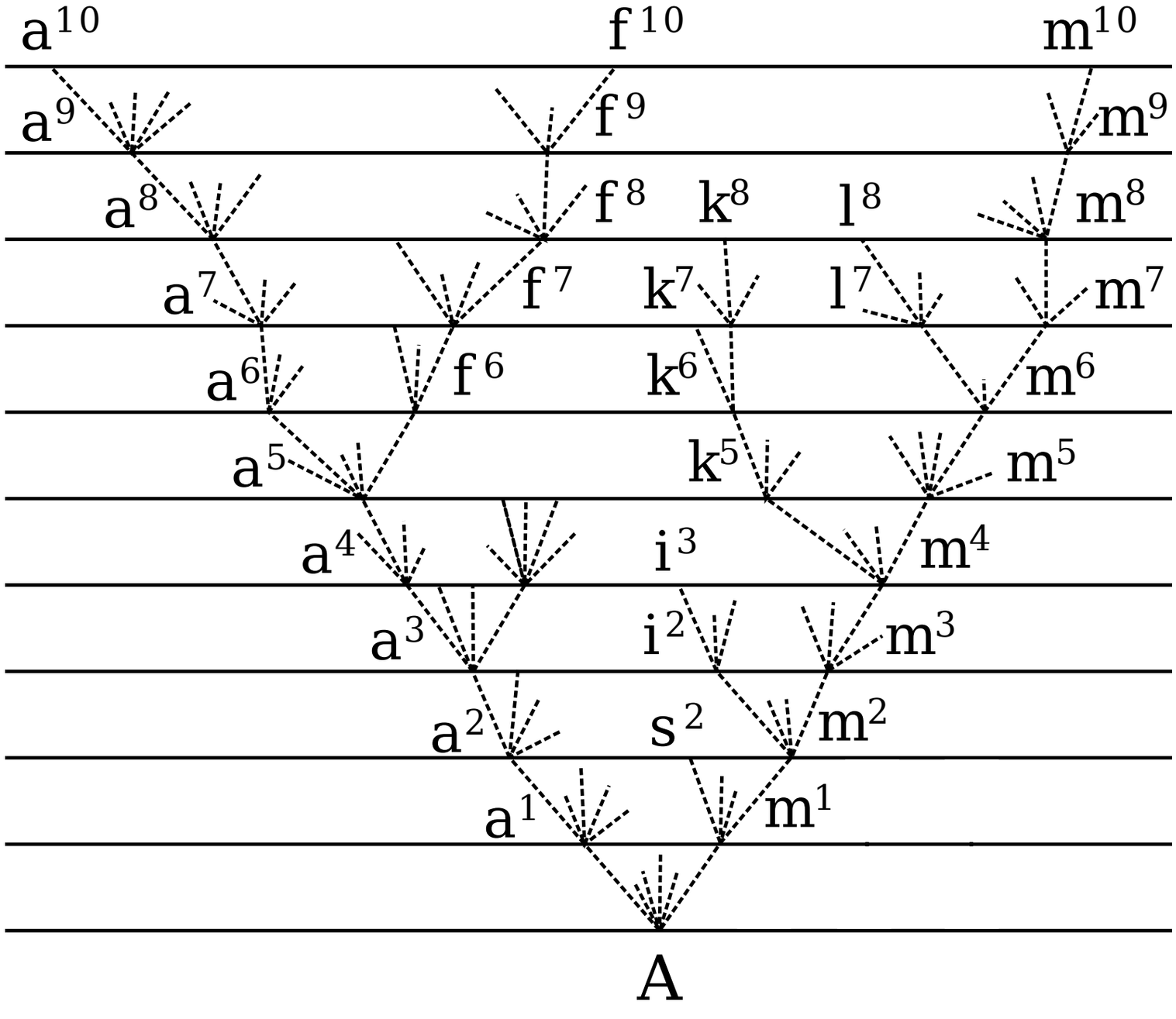
Bayesian phylogenetics

Professor Alexei Drummond

School of Biological Sciences
School of Computer Science
University of Auckland

14th August 2023, Taming the BEAST eh!, Squamish, Canada

Darwin’s Computer



Detail from the only illustration in the *Origin of Species* (Darwin, 1859)

Computational phylogenetics

- **1966-1981 - the early years**
  - Maximum parsimony introduced
  - Least squares
- **1981-1991 - idealogical warfare and the “Dark Ages” for systematics and molecular evolution**
  - Maximum Parsimony and cladistics peaked
  - Maximum likelihood pruning algorithm introduced (1981)
  - Neighbour-joining introduced (1987)
- **1991- 2001 - the statistical phylogenetics revolution and “a reasonably happy ending”**
  - maximum likelihood matures, parametric bootstrap, KH test et cetera.
  - Bayesian phylogenetics introduced (1996)
- **2001-2010 - Bayesian phylogenetics revolution**
  - MrBayes, BEAST, BayesPhylogeny, PhyloBayes, PHYCAS et cetera
- **2010-2019 - Phylogenomics revolution**
  - Multispecies coalescent, Fossilized birth-death models, methods that go beyond trees are maturing
  - BEAST2, RevBayes
- **2020—now - Integrative phylogenomics? Non-MCMC Bayesian methods? Real-time phylogenetics?**

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:
  - All valid reasoning routes give the same result.
  - 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:

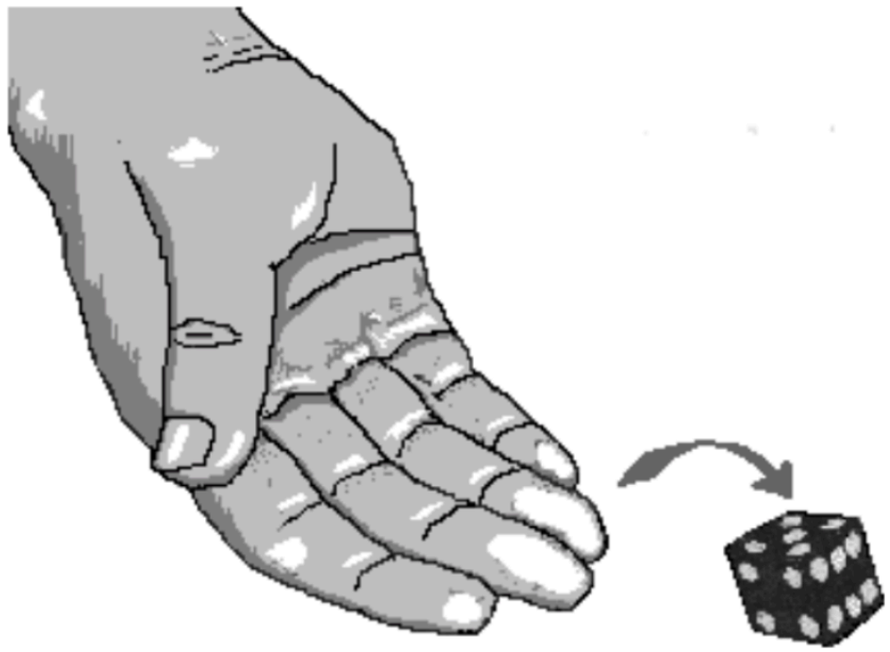
1. Experiments are assumed to be repeatable.
2. Assumes that randomness is a property of the system.
3. Completely ignores ~ 400 years of physics.

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.

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?

- Represents initial beliefs or knowledge about an event or parameter.
- Quantifies our understanding before observing new data.
- Ideally based on:
    - Objective information, when available.
    - Domain expertise or historical data.
- Can be:
    - **Informative**: Specific based on existing knowledge.
    - **Uninformative**: Broad when little is known.
- **In principle**, any two (rational) people with access to the same information should specify exactly the same prior..
- Importance diminishes with more informative data.

Prior probabilities are essential

- **Common Misconception**: Isn't the need for priors a drawback in the Bayesian approach?

  - Answer: NO!

  - Inference is impossible without underlying assumptions.

  - Priors enable the incorporation of past knowledge.

  - It's not just Bayesian: Frequentist and Likelihoodist methods also have underlying priors – they're just less explicit! (e.g. regularisation in “frequentist” regression)

Priors for discrete variables

- Setting priors for discrete variables with finite bounds can be straightforward.

- **Principle of Indifference**:

  - Assign equal probabilities when there's no evidence to prefer one outcome over another.

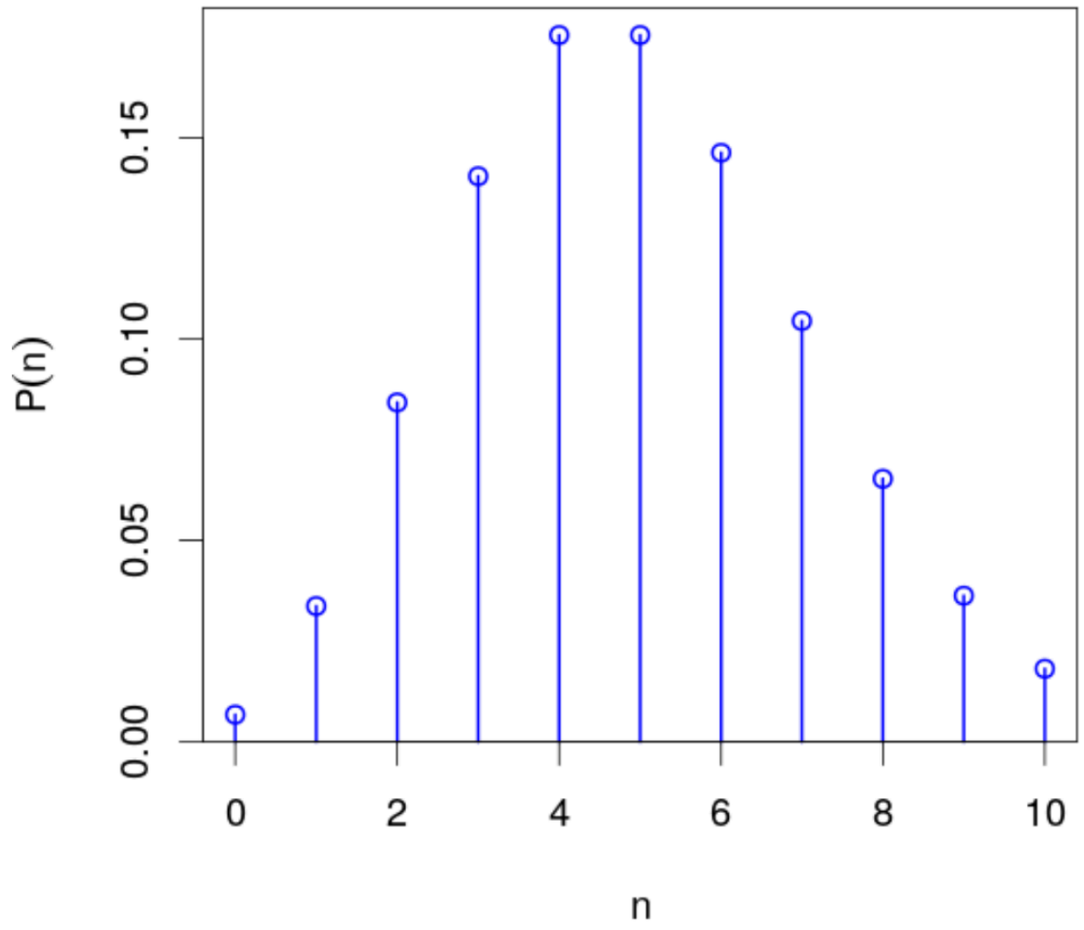
  - Example: For a discrete variable with 4 possible outcomes, each gets a 25% chance.

  - The Poisson Distribution arises from the assumption that events are rare, independent, and no small interval of time (or space) is privileged over another.

- **Principle of Maximum Entropy**:

  - Ideal for defining priors under constraints.

  - E.g., the Geometric distribution is the max entropy distribution for a non-negative variable with a known prior mean.

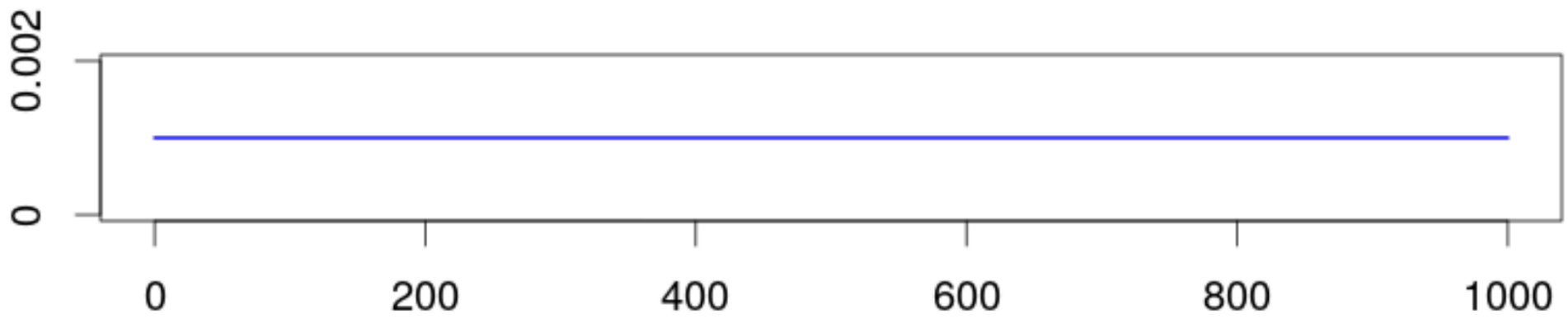


Priors on continuous variables

- Bounded Continuous variable $a < x < b$:
  - Uniform distribution: $f(x) = \dfrac{1}{b-a}$
  - Beta distribution: $f(x) \propto (x-a)^{\alpha-1}(b-x)^{\beta-1}$
- Rate variable $\lambda > 0$:
  - One might think to use $f(\lambda) = c$ to signify complete ignorance.
    - Caution: This might not be ideal!
  - A better option: $f(\lambda) = 1/\lambda$
    - Why? It's uniform in log-space, reducing undue weight on larger values.

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?

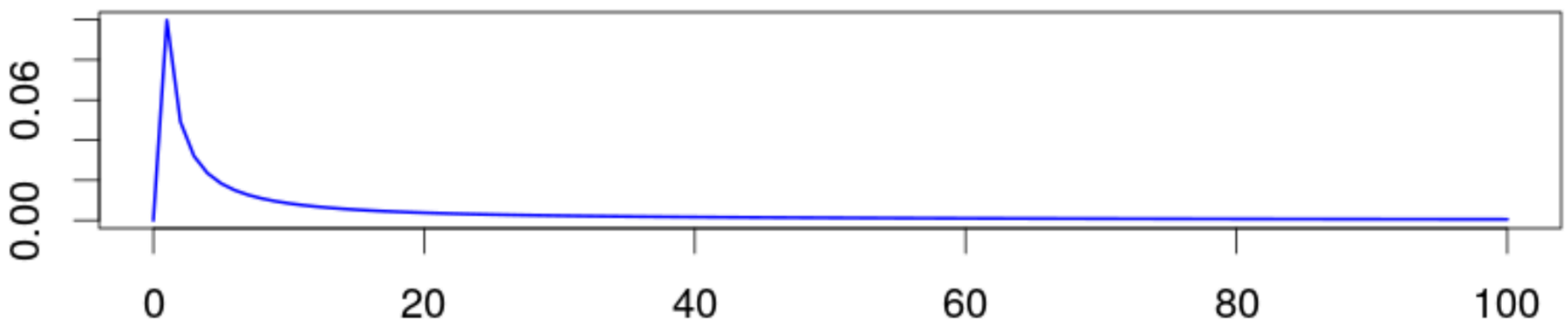


We can't! This $f(\lambda)$ is not a true probability density.

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.



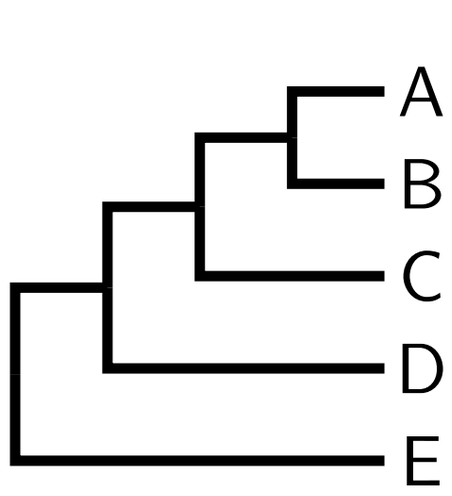
Which prior is best?

- The answer lies with the analyst!
- Role of Priors:
  - Priors capture expert knowledge or acknowledge the lack thereof.
  - Priors provide an avenue to incorporate domain expertise into the analysis.
  - Priors allow for the explicit comparison of different assumptions and testing the sensitivity of your results to them.

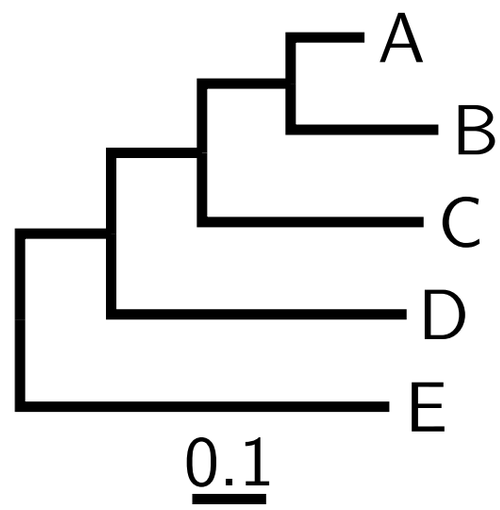
What about phylogenetics?

Types of phylogenies

rooted trees

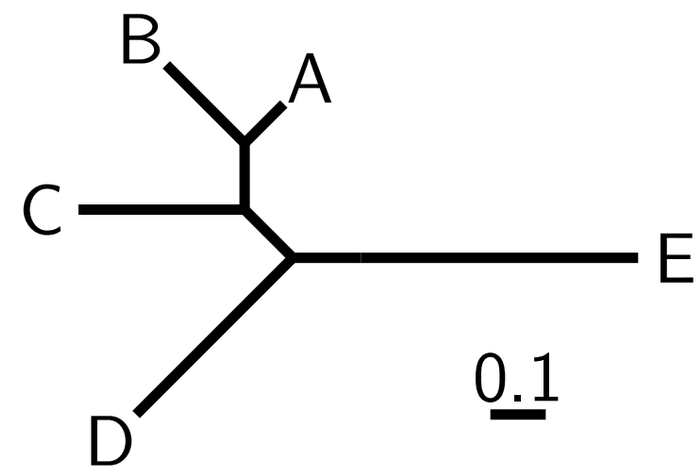


(a) cladogram



(b) phylogram

unrooted tree



(c) unrooted tree

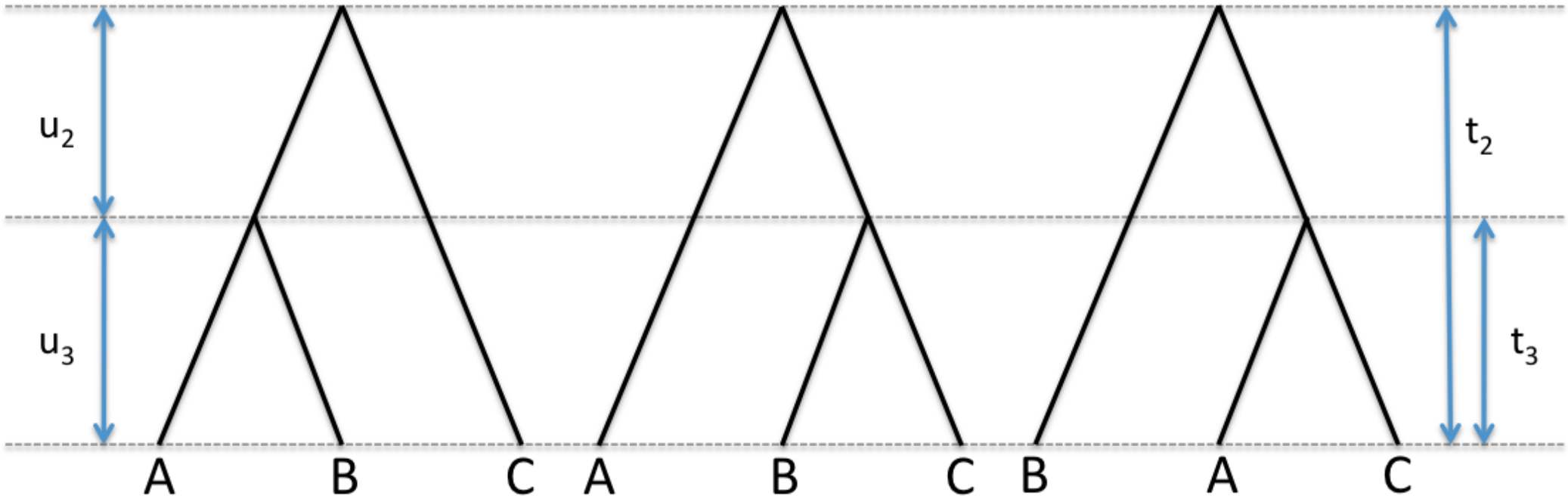
((((A, B), C), D), E);

((((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.

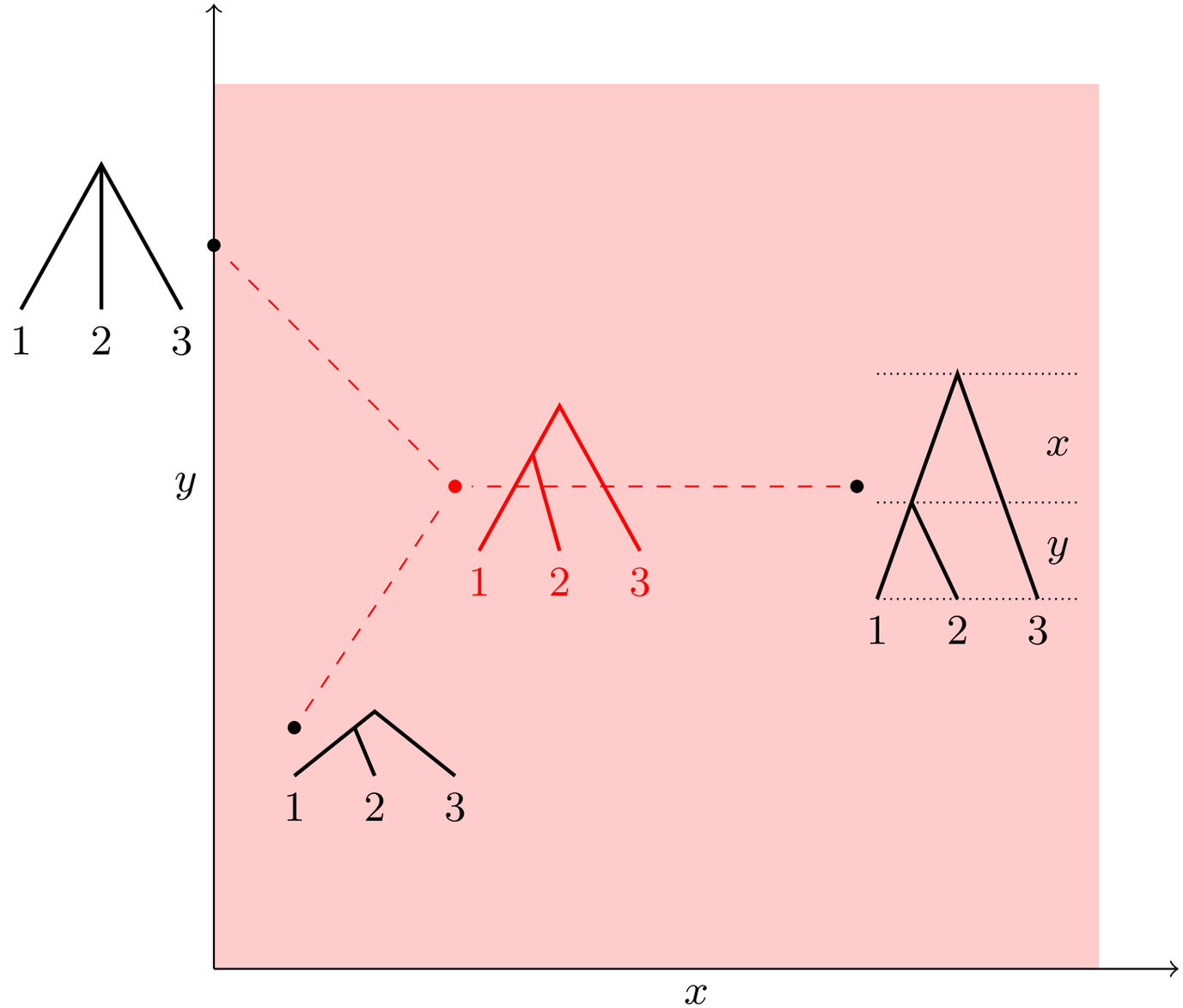


Figure: A Euclidean two-dimensional space representing the space of all possible time-trees for the topology ((1,2),3). There are two parameters, x and y, one for each of the two inter-coalescent intervals, the sum of which is the age of the root ($t_{root} = x + y$). Three trees are displayed, along 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).

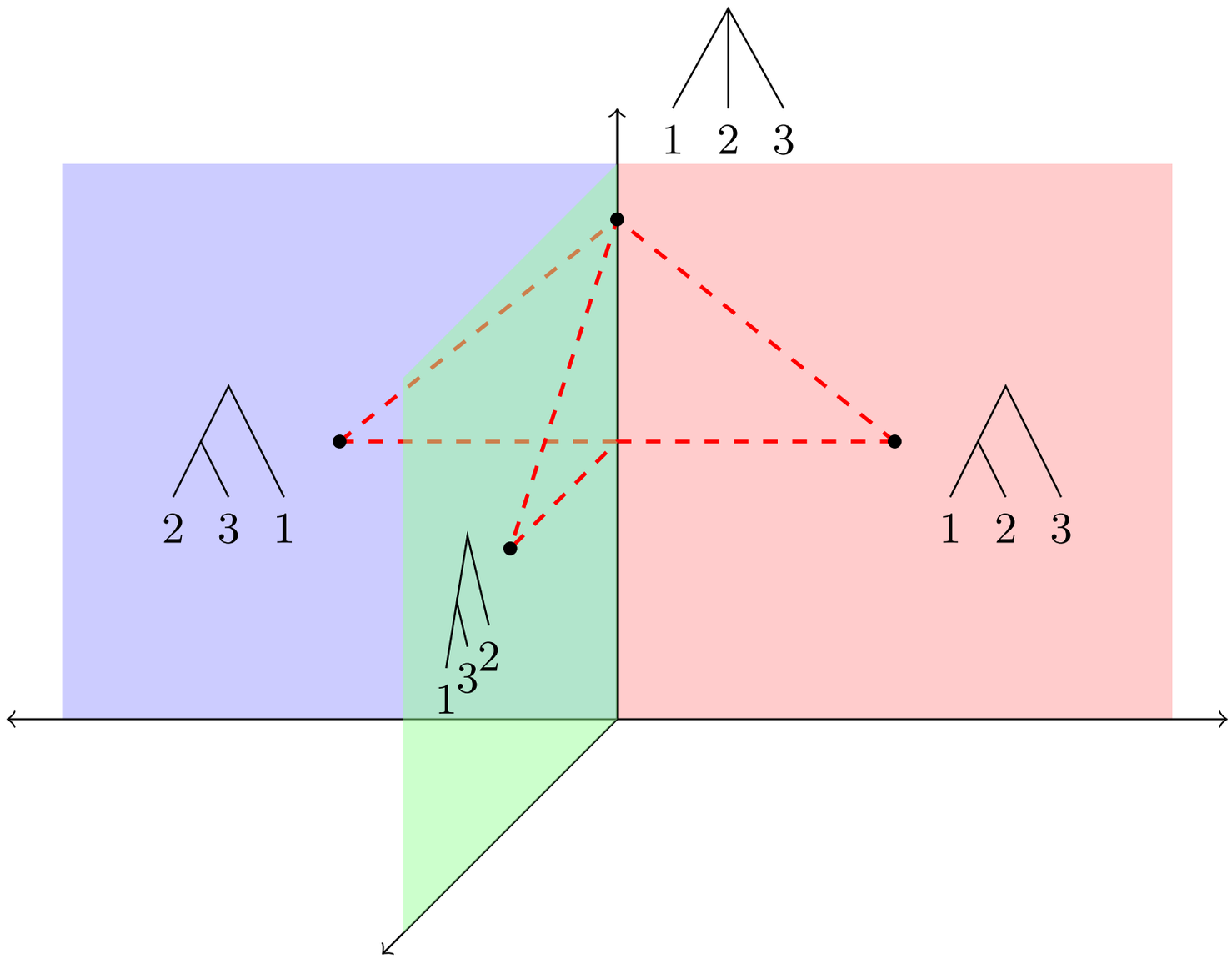


Figure: τ_3, the simplest non-trivial tree space (for time-trees), representing the space of time-trees for $n = 3$ taxa sampled contemporaneously. Each of the three non-degenerate tree topologies is represented by a two-dimensional Euclidean space (as illustrated in Figure 1) and these subspaces meet at a single shared edge representing the star tree, which is a one-dimensional subspace and thus has a single parameter (the age of the root). The dashed lines shows the paths of shortest distance between the four displayed trees.

Tip-labeled time-trees of size 4

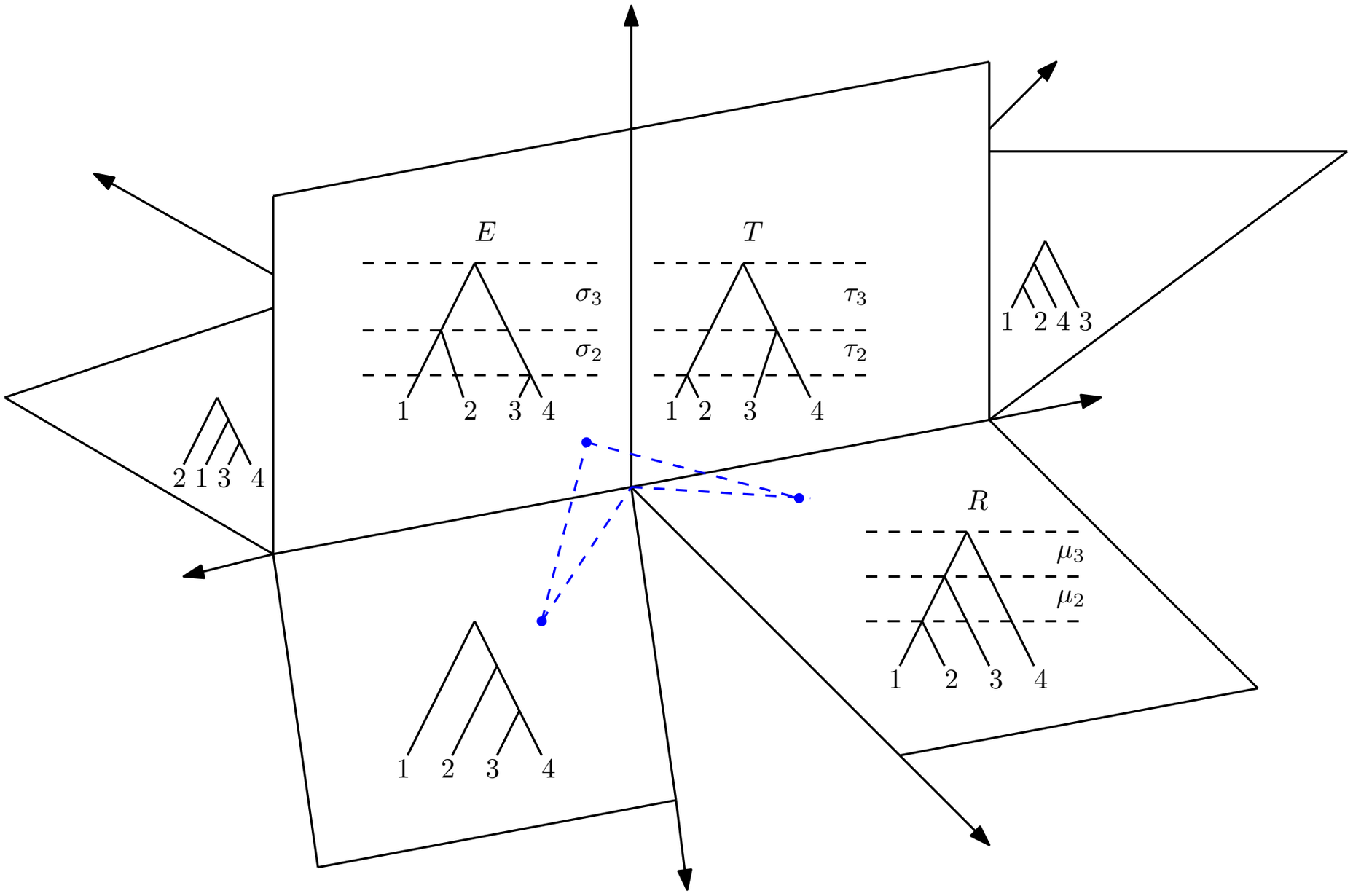
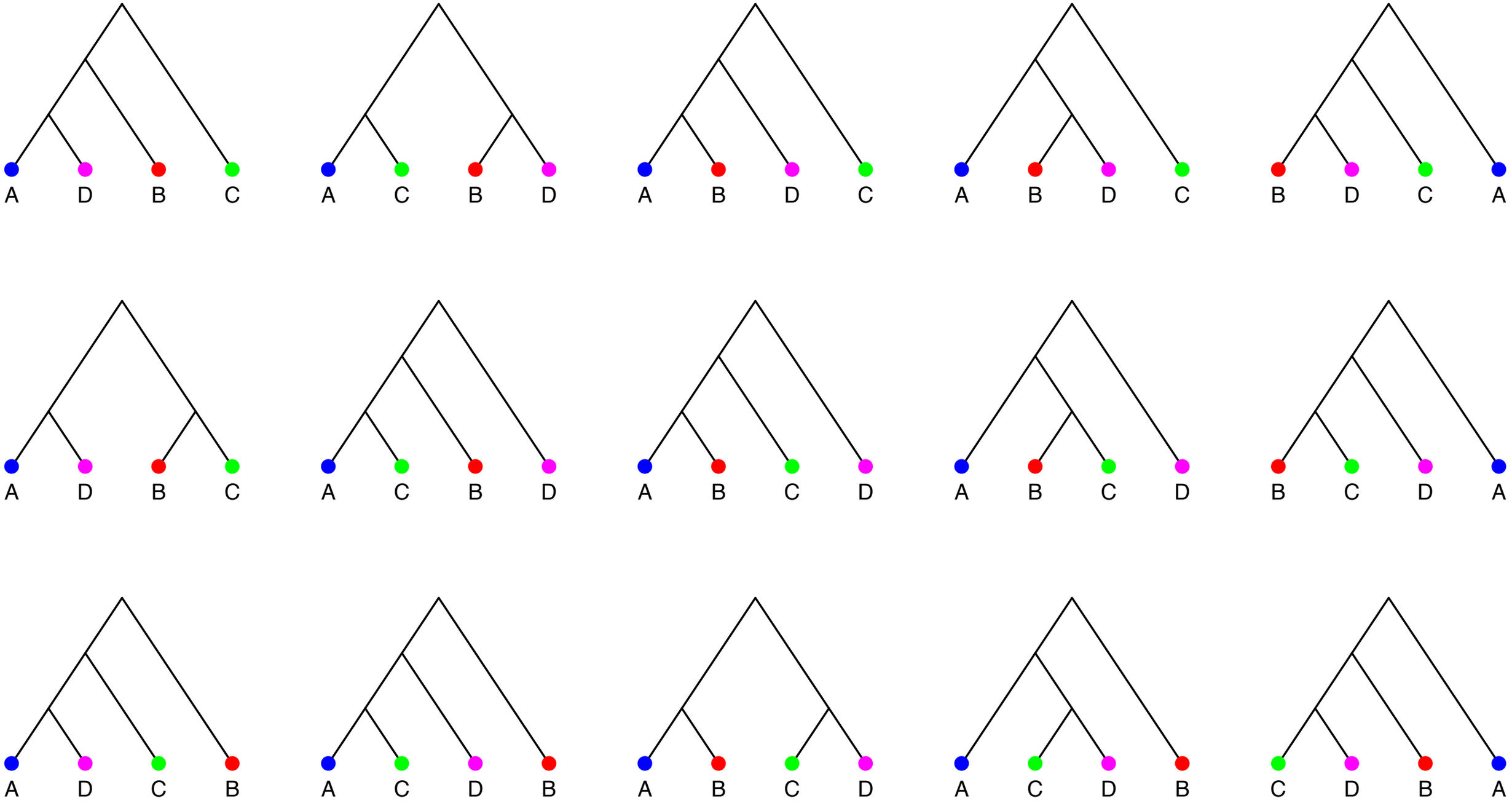


Figure: Three-dimensional projection of 4-dimensional τ-space T_4.

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

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

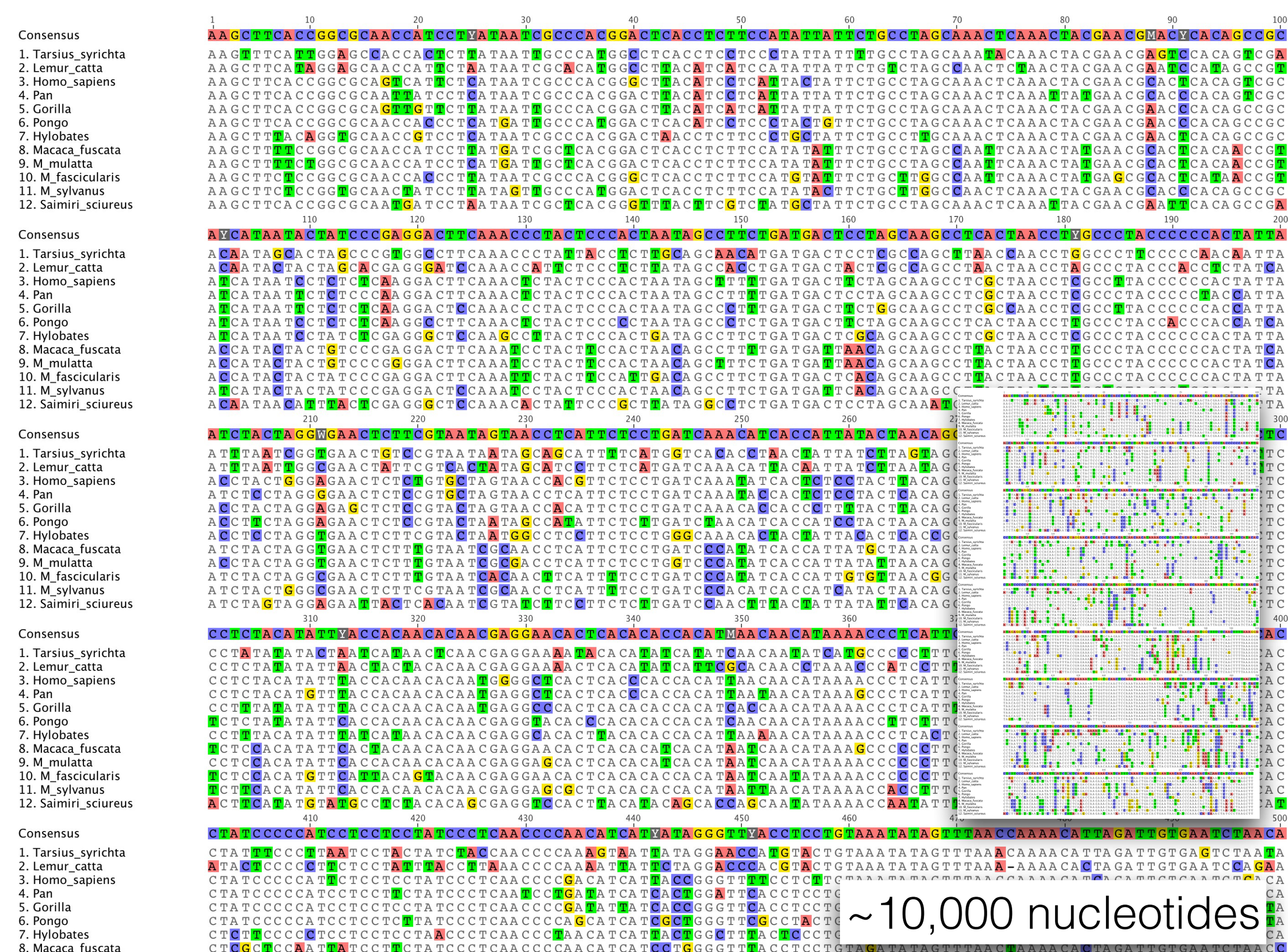
rooted, tip-labelled binary trees:

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.



~10,000 nucleotides

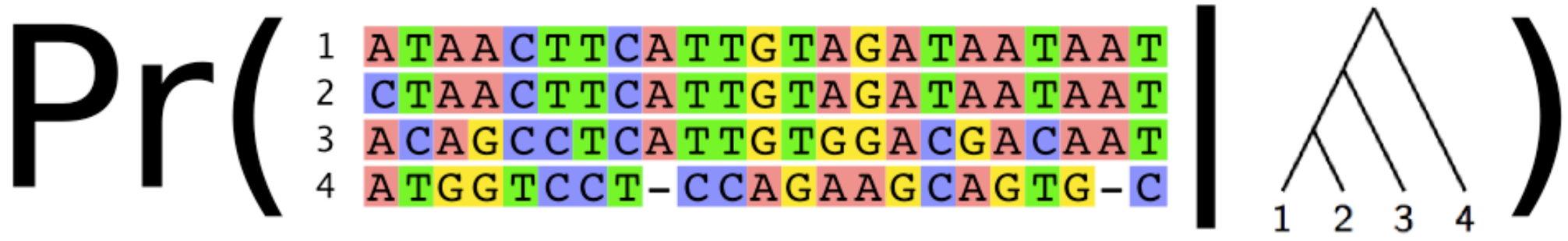
The phylogenetic likelihood

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

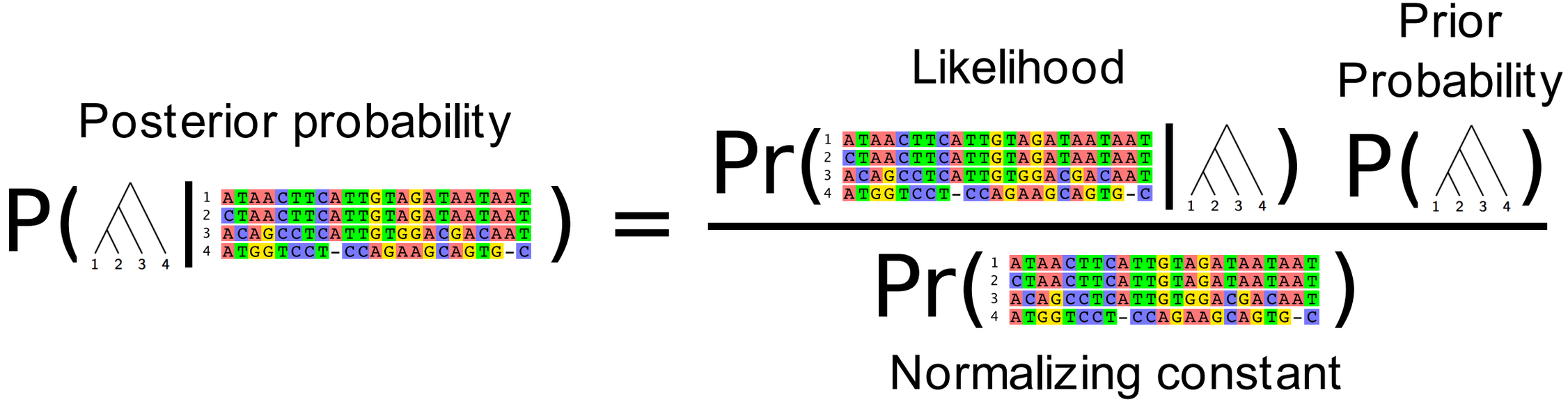
$$\underbrace{P(\theta \mid D)}_{\text{posterior}} = \frac{\overbrace{P(D \mid \theta)}^{\text{likelihood}}\overbrace{P(\theta)}^{\text{prior}}}{\underbrace{P(D)}_{\text{marginal likelihood}}}$$

Bayesian reconstruction of phylogenetic trees

Yang & Rannala (1997), Mau, Newton & Larget (1998)

What we really 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 the posterior probability distribution **without knowing the normalising constant**.

The posterior distribution on Darwin's Finches

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

Posterior entropy = 0.04548418

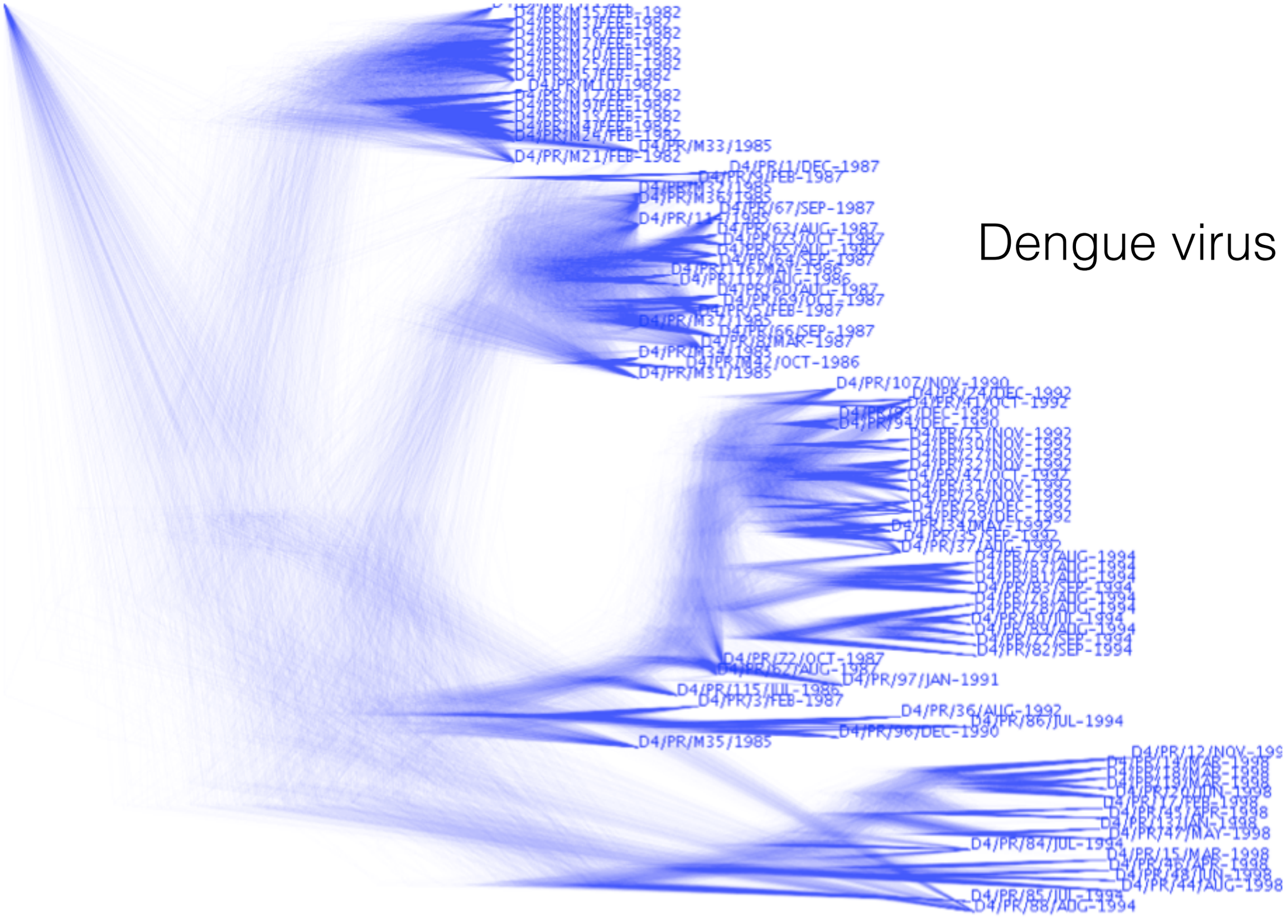
exp(0.04548418) = 1.046534

Total trees in prior: 15

Prior entropy: 14.28661

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

The posterior distribution of larger trees



Dengue virus

Elaborations of the phylogenetic model

Basic model: (posterior proportional to likelihood x prior)

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

Substitution model parameters:

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

Substitution model and tree branching process parameters:

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

The phylogenetic posterior

Standard application of Bayes theorem gives the posterior:

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

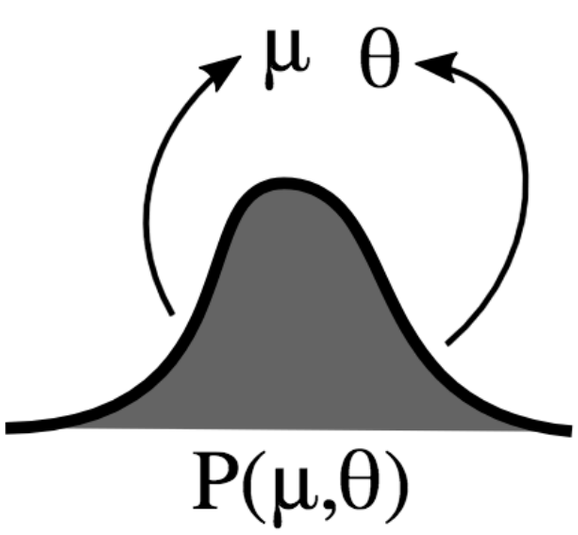
But you will normally see it written like this

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

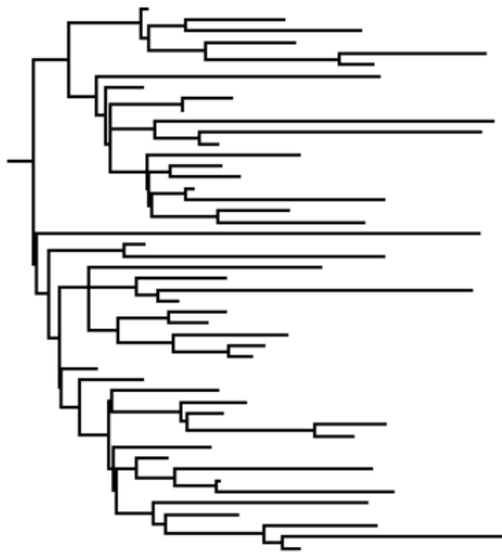
- the probability of the data depends on θ **only through the tree**.
- the prior probability of the tree depends on θ but not on Q.
- the prior probability of θ and the prior probability of Q are independent.

The neutrality assumption

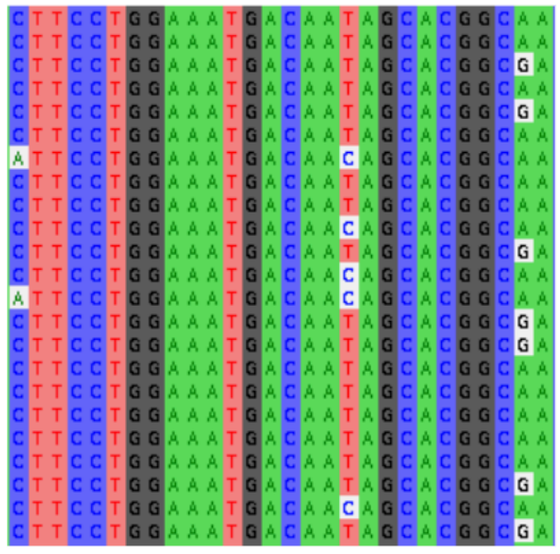
Because of the way we've factorized the joint probability for the data and model parameters, we are implicitly assuming that our alignment could have been produced in the following fashion:



Sample parameters from prior

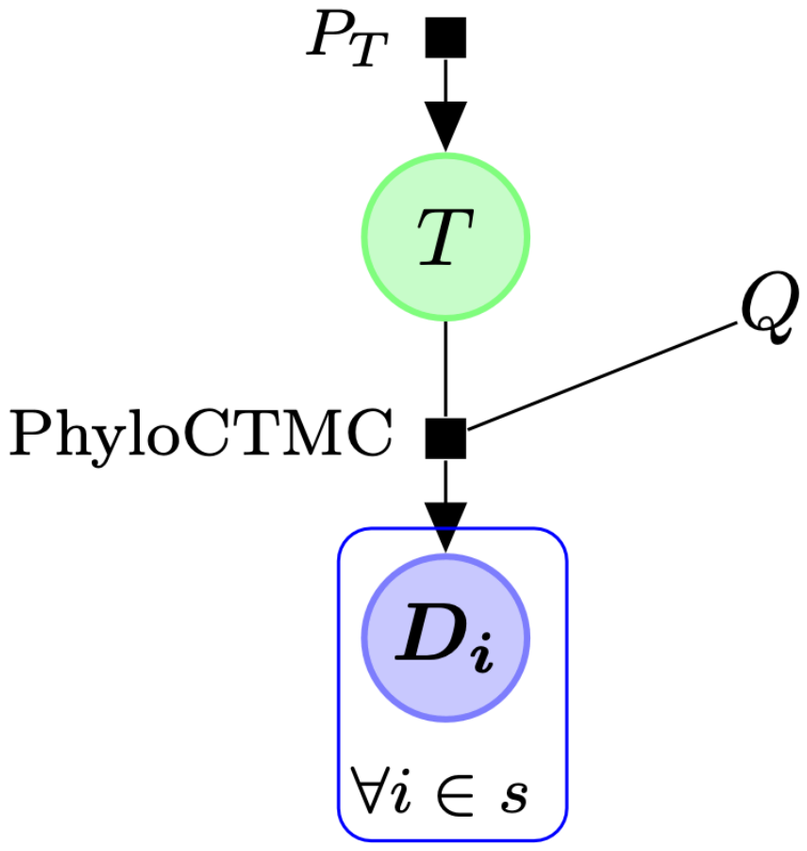


Sample tree from prior

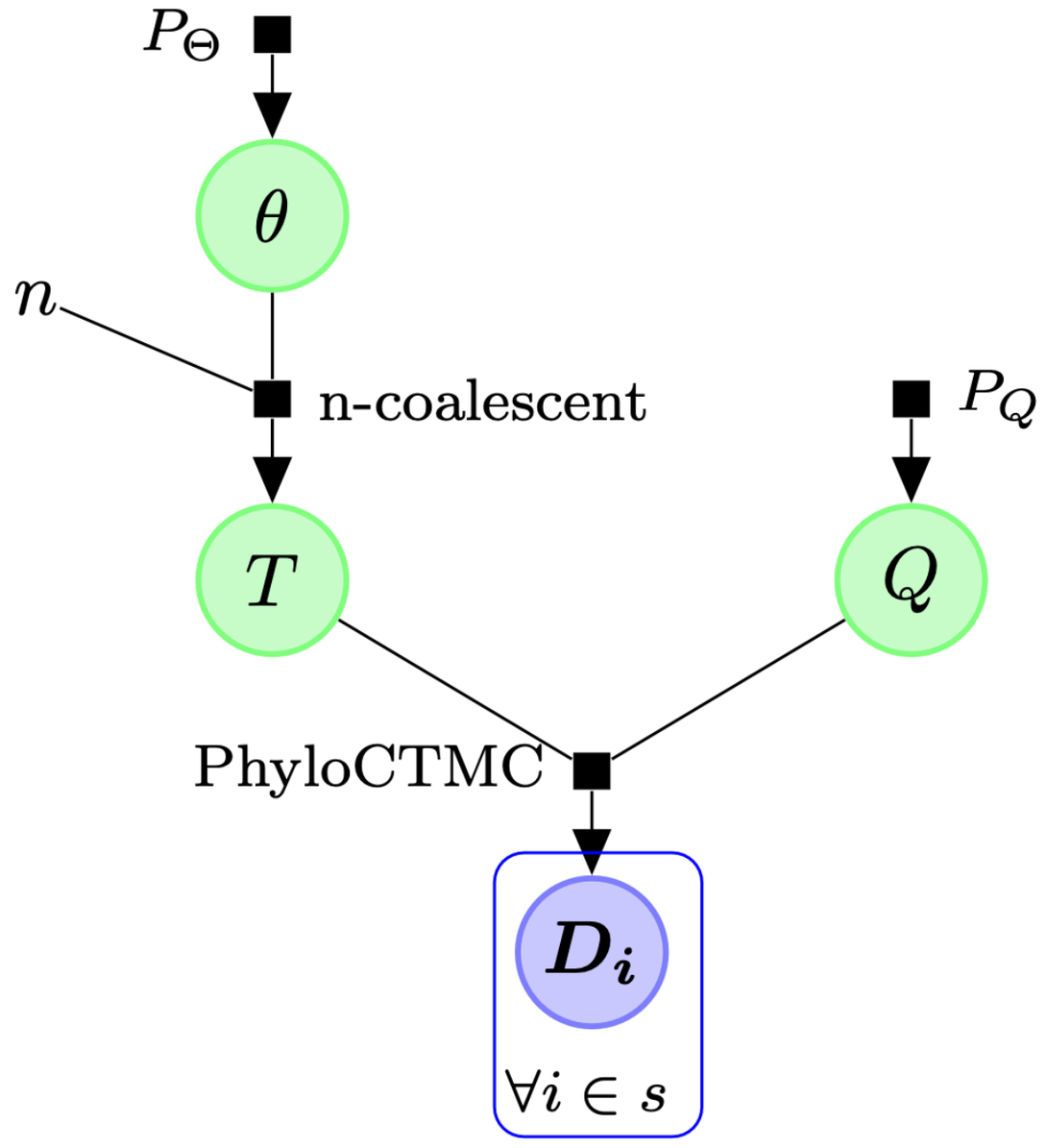


Simulate sequence alignment

Graphical models

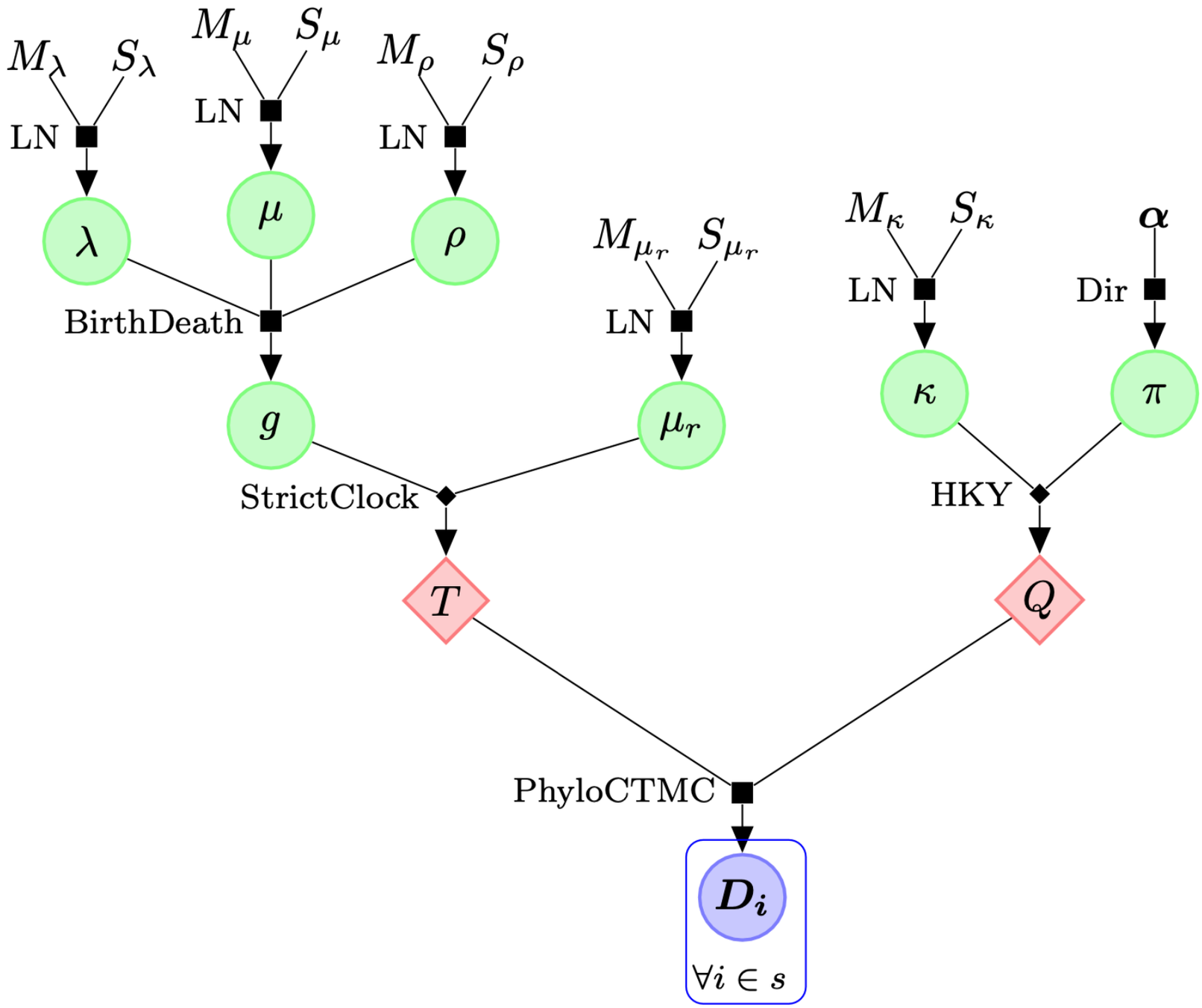


$$P(T|\boldsymbol{D}) \propto \Pr(\boldsymbol{D}|T)P(T)$$



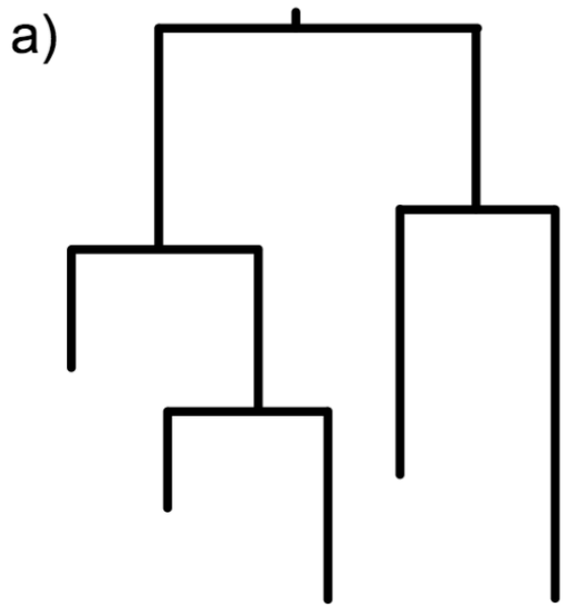
$$P(T, Q, \theta|\boldsymbol{D}) \propto \Pr(\boldsymbol{D}|T, Q)P(T|\theta)P(\theta)P(Q)$$

Graphical models for phylogenomics

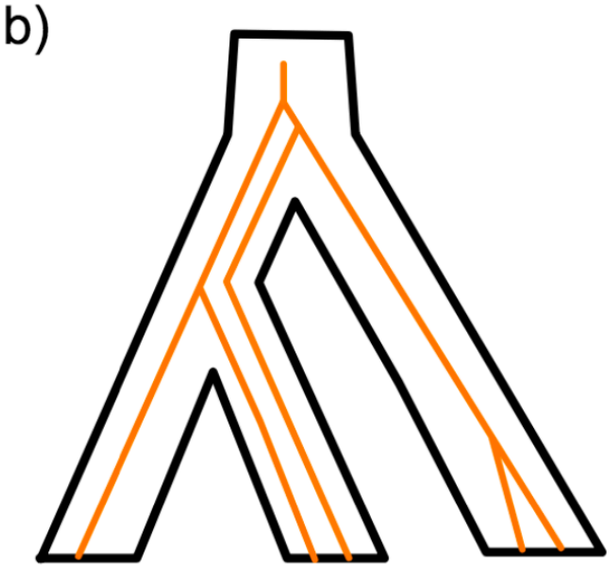


$$P(g, \mu_r, \lambda, \mu, \rho, Q|\boldsymbol{D}) \propto \Pr(\boldsymbol{D}|\mu_r g, Q) P(g|\lambda, \mu, \rho) P(\lambda) P(\mu) P(\rho) P(\mu_r) P(Q)$$

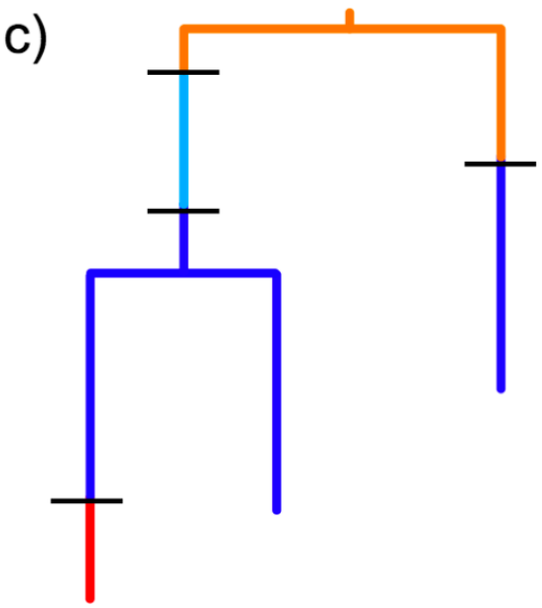
BEAST 2.7 contains many models



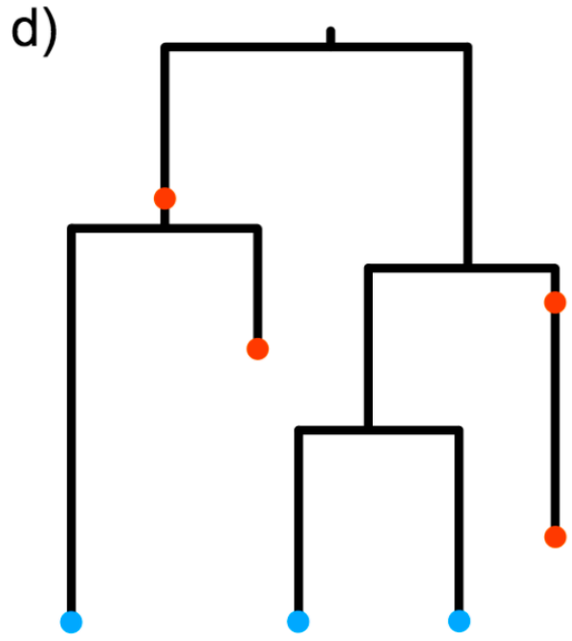
Tip-dated time tree
(leaf times conditioned on)



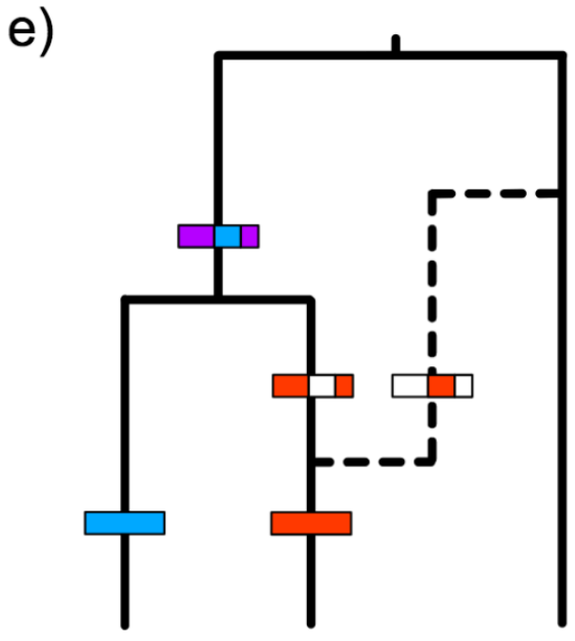
Multi-species coalescent



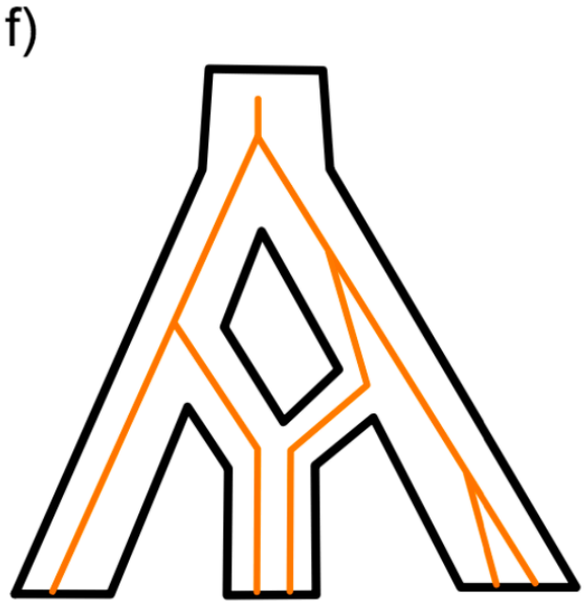
Multi-type time tree



sampled ancestor
time tree

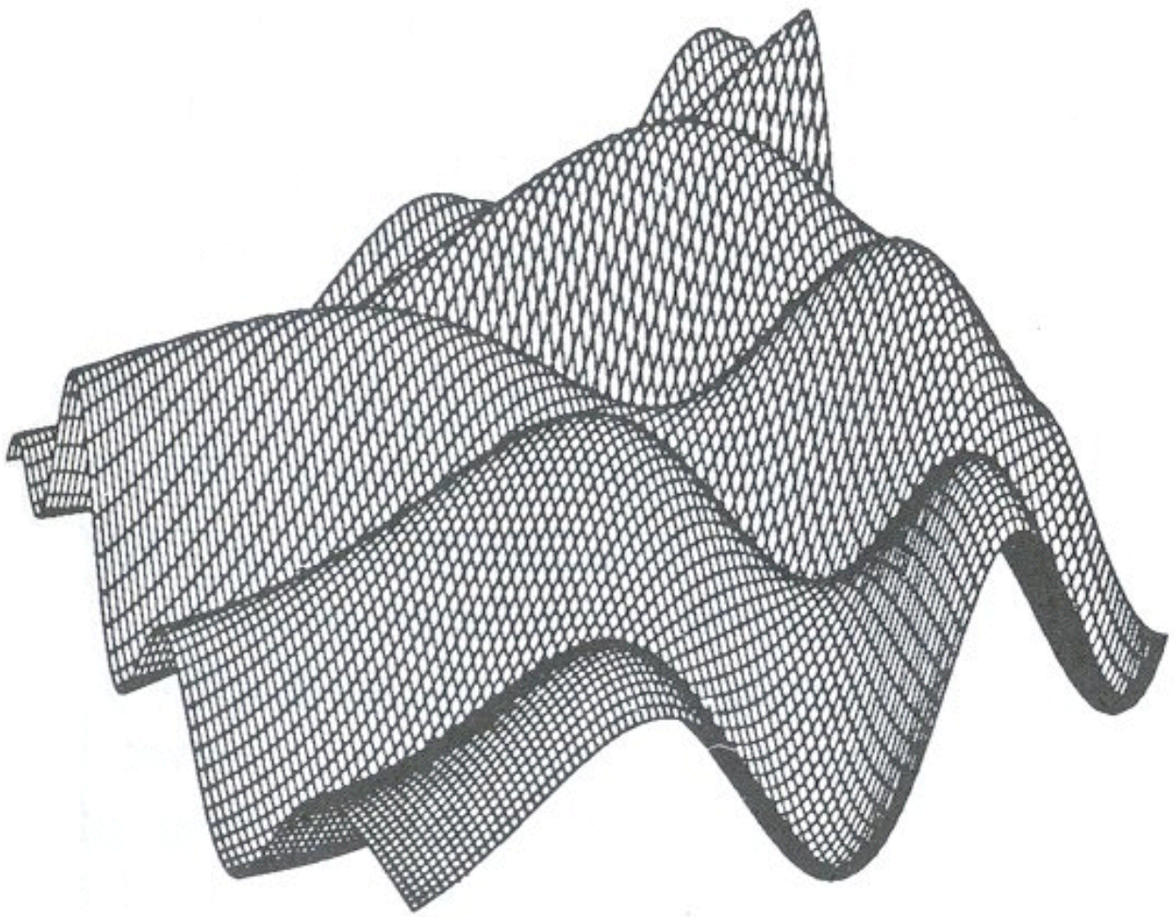


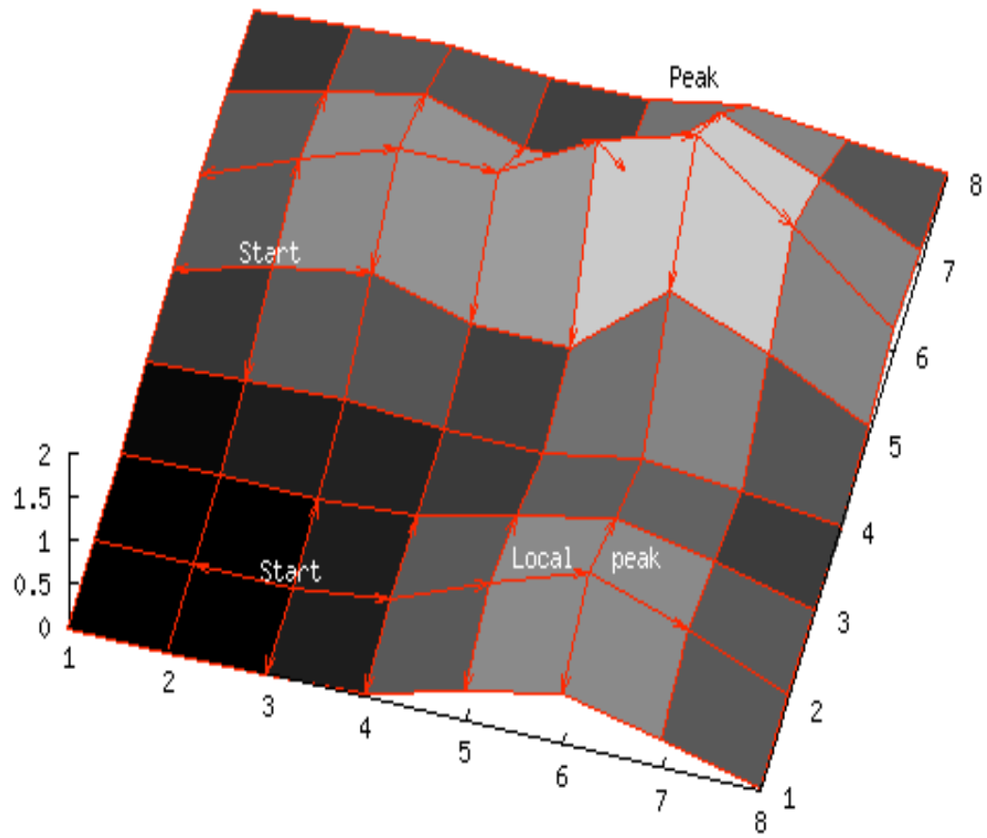
ancestral gene conversion
graph



Species network
with embedded gene tree

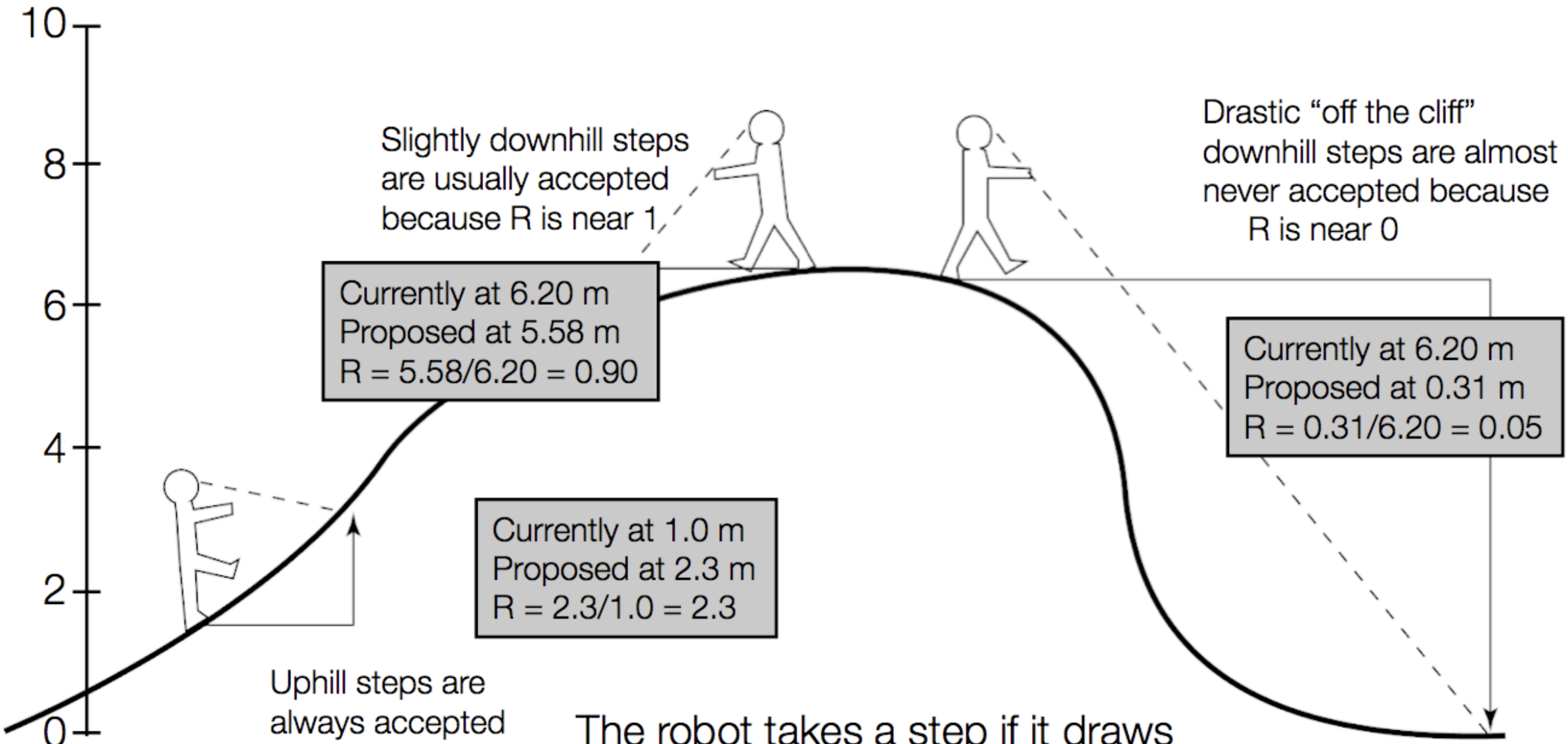
A probability distribution on tree space as a hilly landscape





- 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

Summary

- **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 parameter space of phylogenetic trees (**tree space**) is vast and non-Euclidean and therefore requires special treatment and software.

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

- The **posterior** distribution is approximated using Markov chain Monte Carlo (**MCMC**) algorithm.