Gene trees, species trees and the coalescent

Emily Jane McTavish

Life and Environmental Sciences
University of California, Merced
ejmctavish@ucmerced.edu, twitter:snacktavish

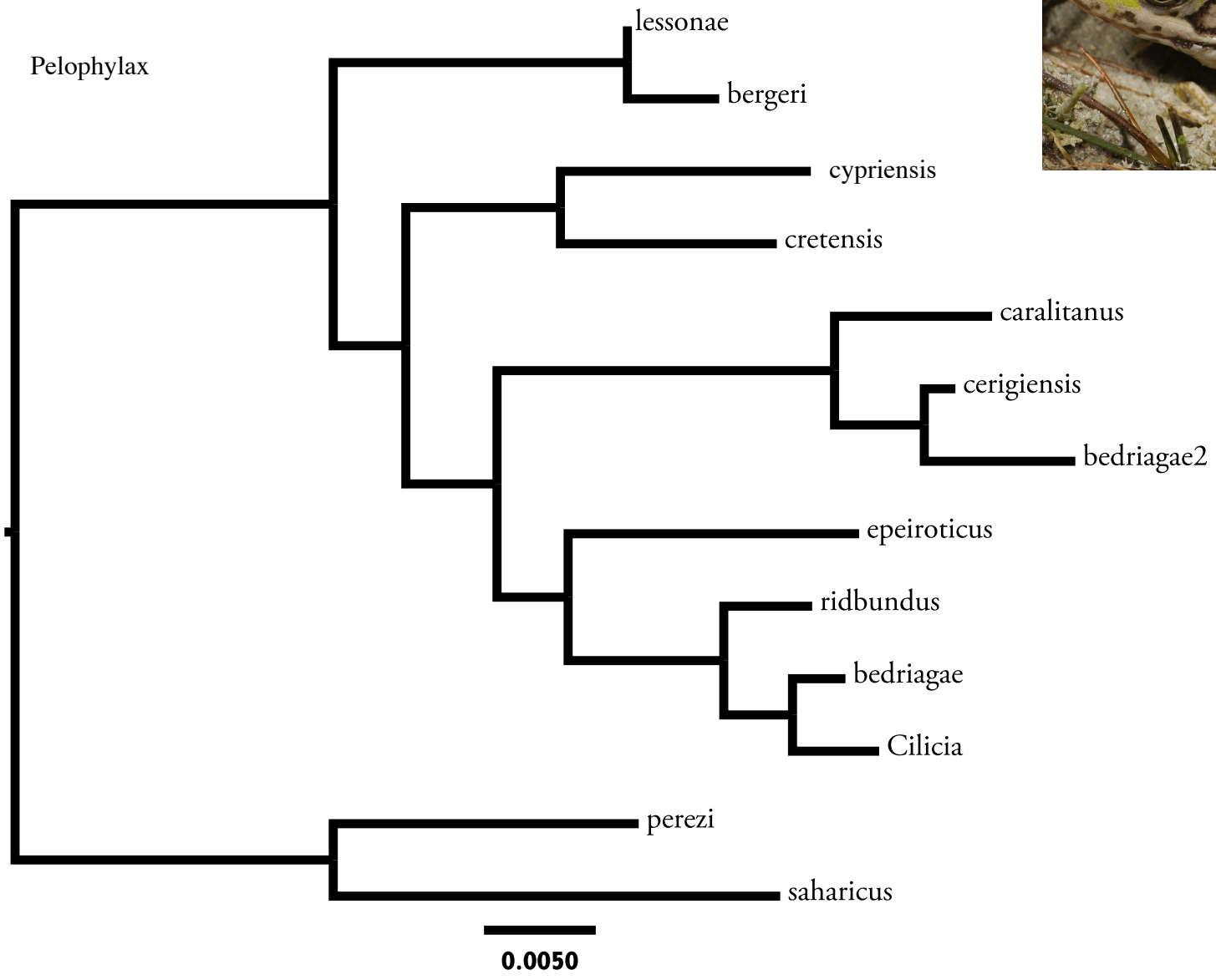
(With thanks to Peter Beerli, Laura Kubatko, Mick Elliot Arne Mooers for slides)

Relationship between phylogenetics and Population Genetics

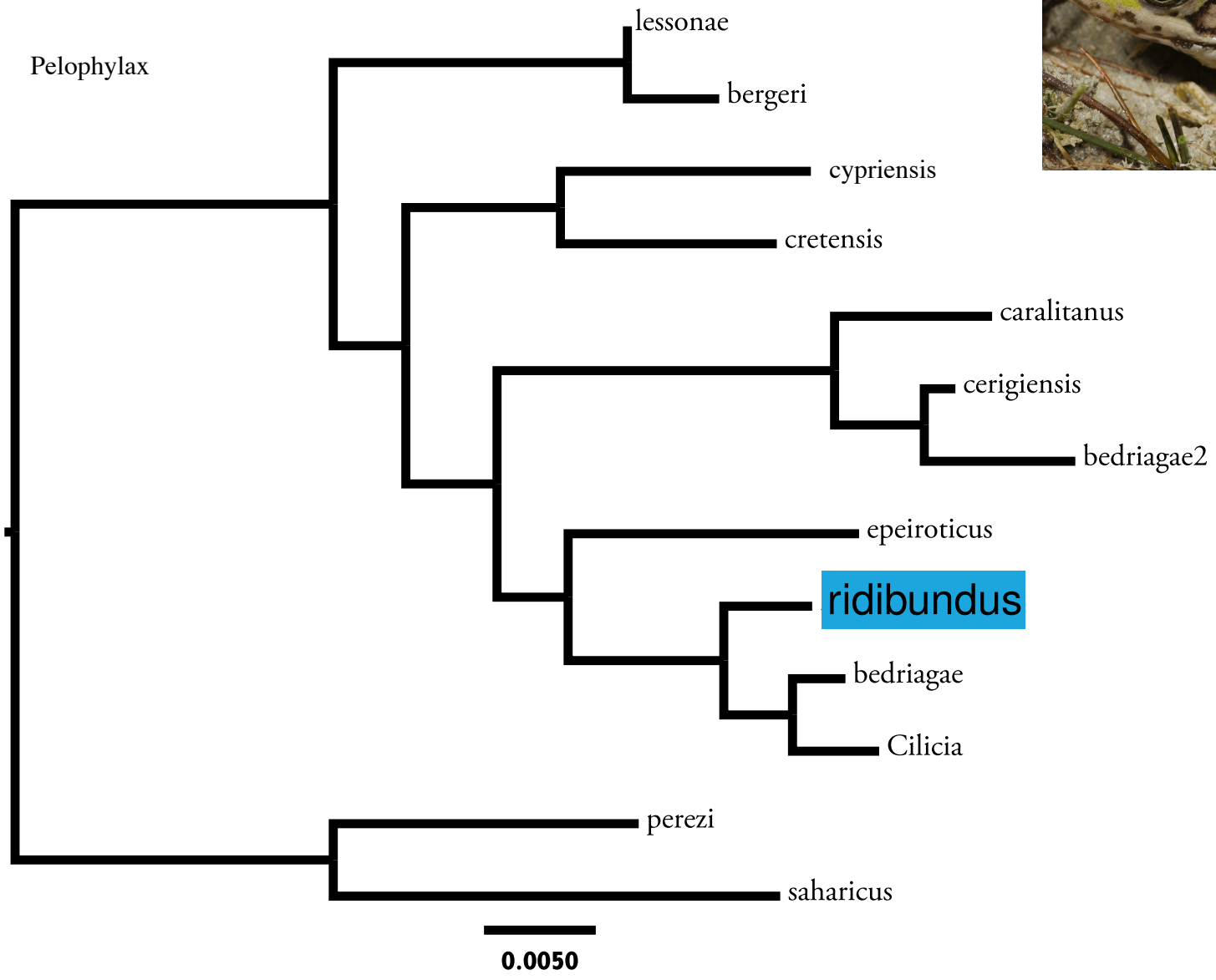
- Population genetics: Study of genetic variation within a population
- Phylogenetics: Use genetic variation between taxa (species, populations) to infer evolutionary relationship

We expected tree like relationships among alleles within species.
The expectations for these trees is described by 'coalescent theory' (Kingman 1982)

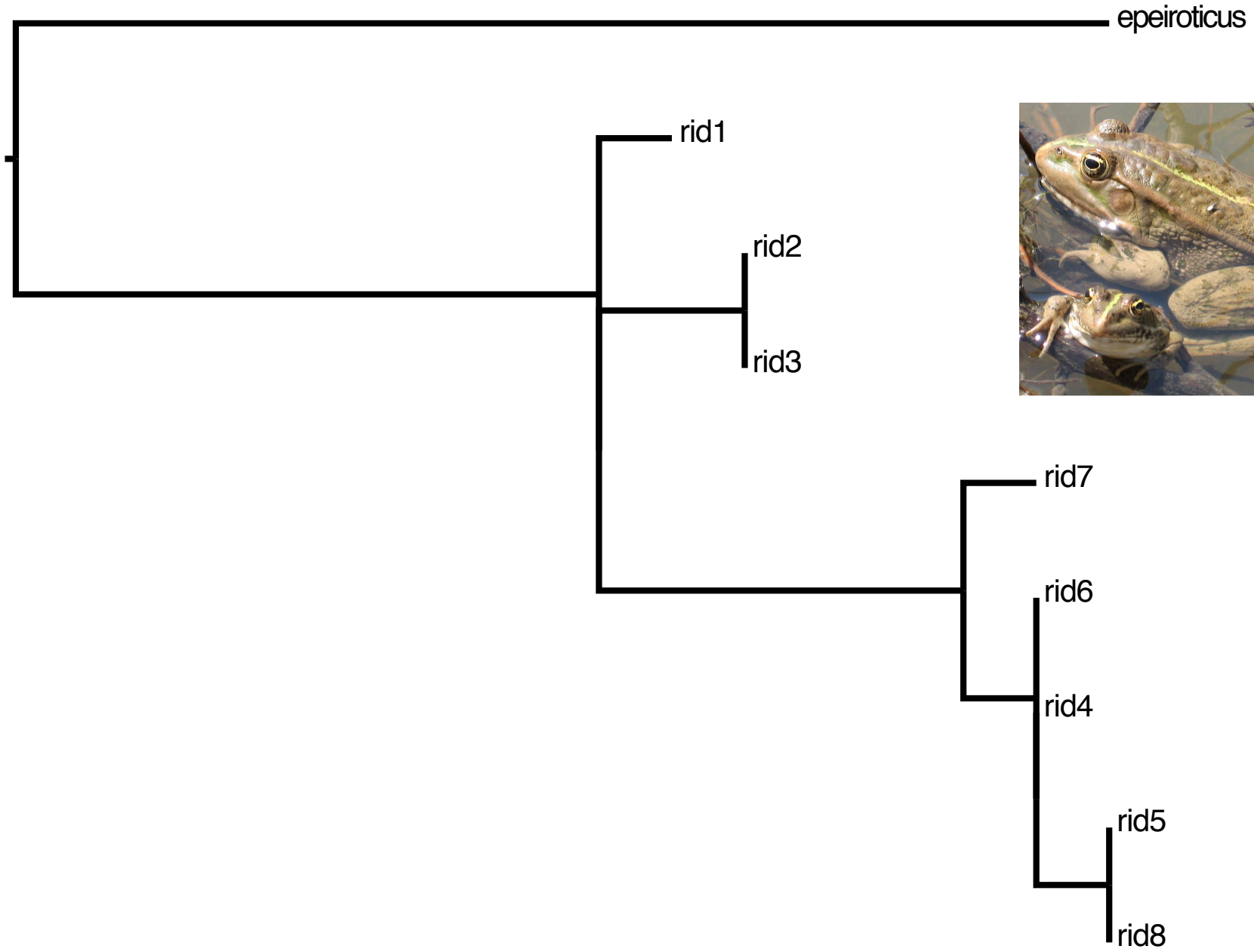
Species trees



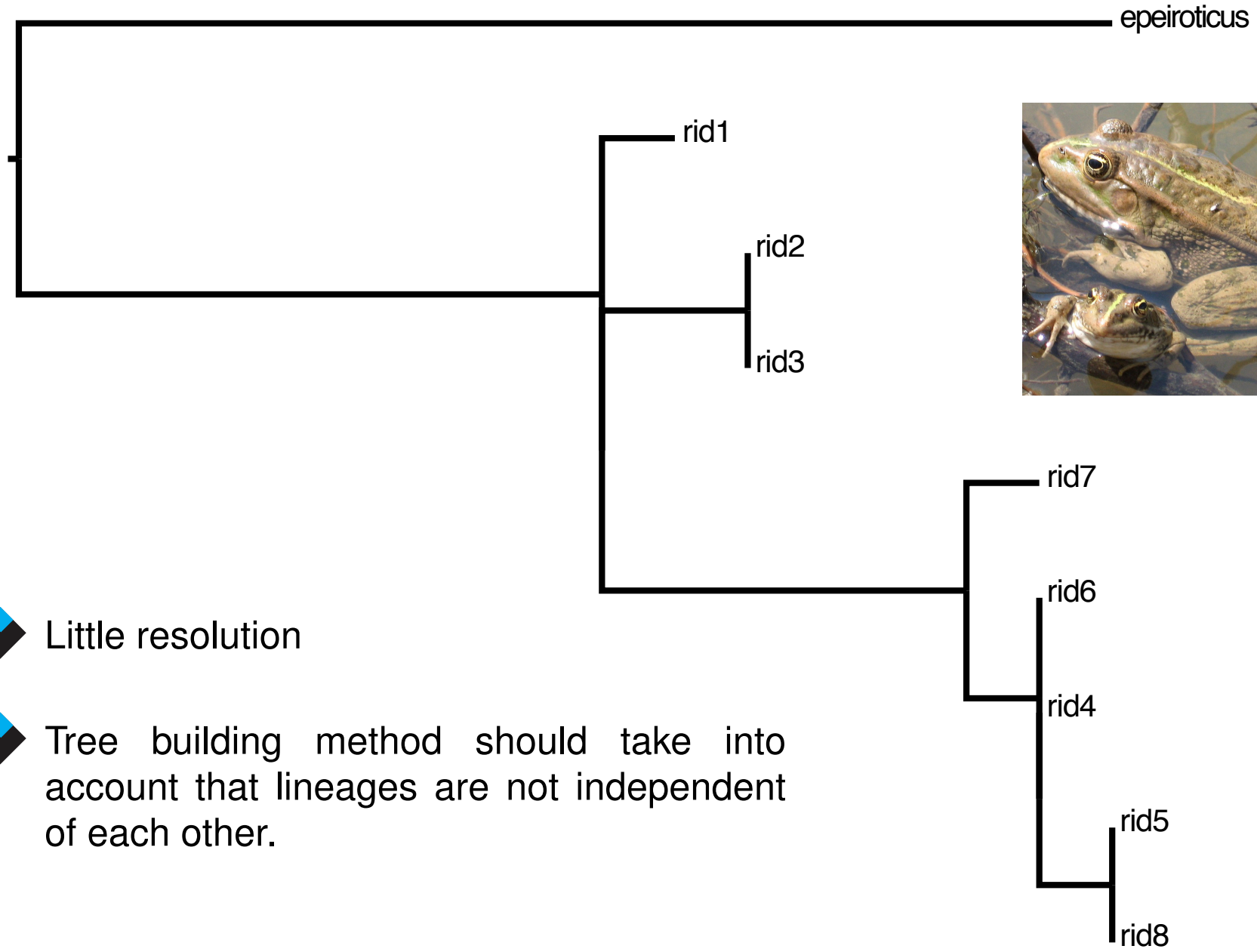
Species trees



Tree of individuals of same species

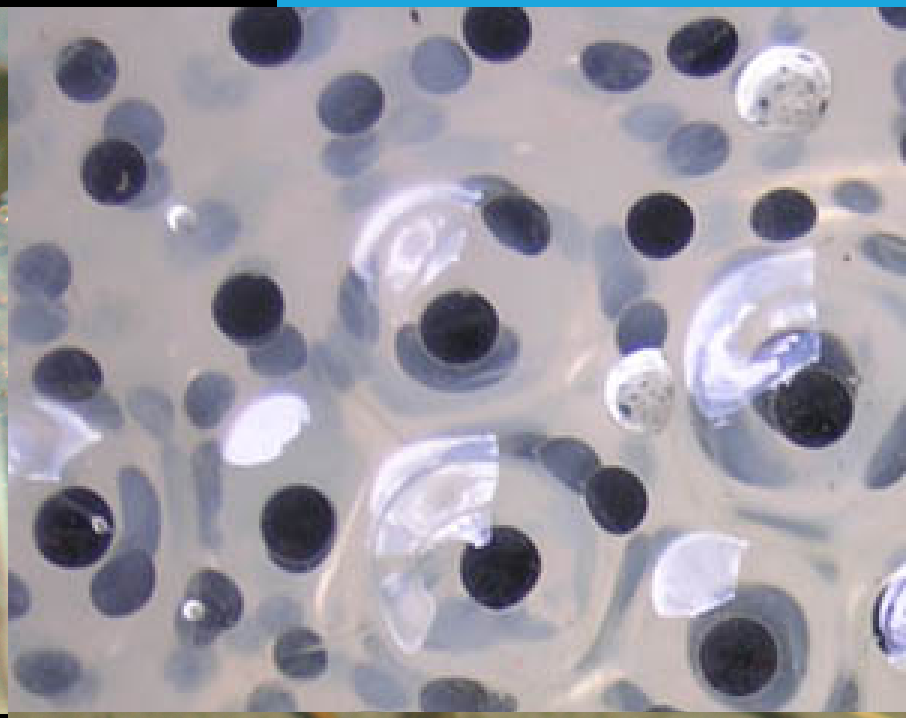


Tree of individuals of same species

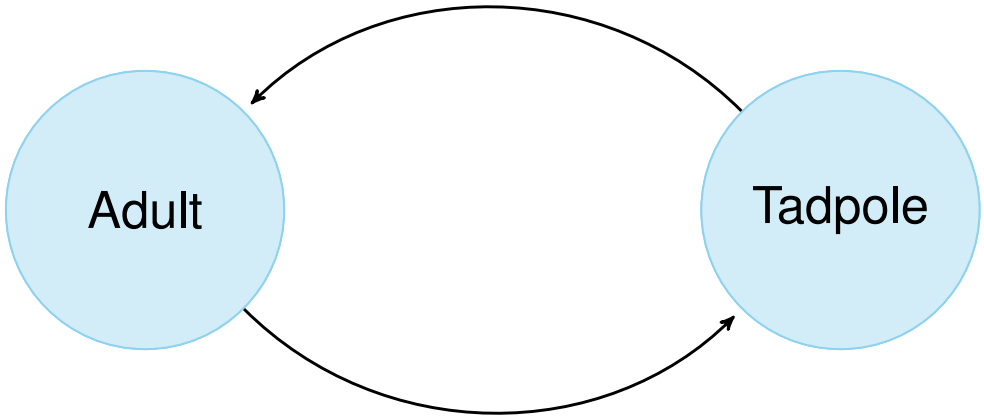


- Little resolution
- Tree building method should take into account that lineages are not independent of each other.

Interaction among individuals



Interaction among individuals



Wright-Fisher population model

- All individuals live one generation and get replaced by their offspring
- All have same chance to reproduce, all are equally fit
- The number of individuals in the population is constant

Population model

Past

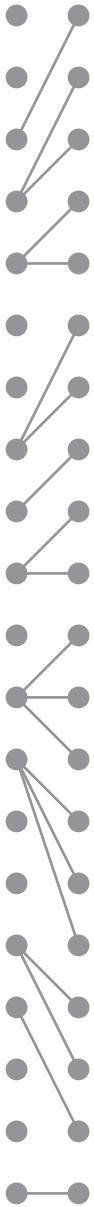
Present

Population model

Past

Present

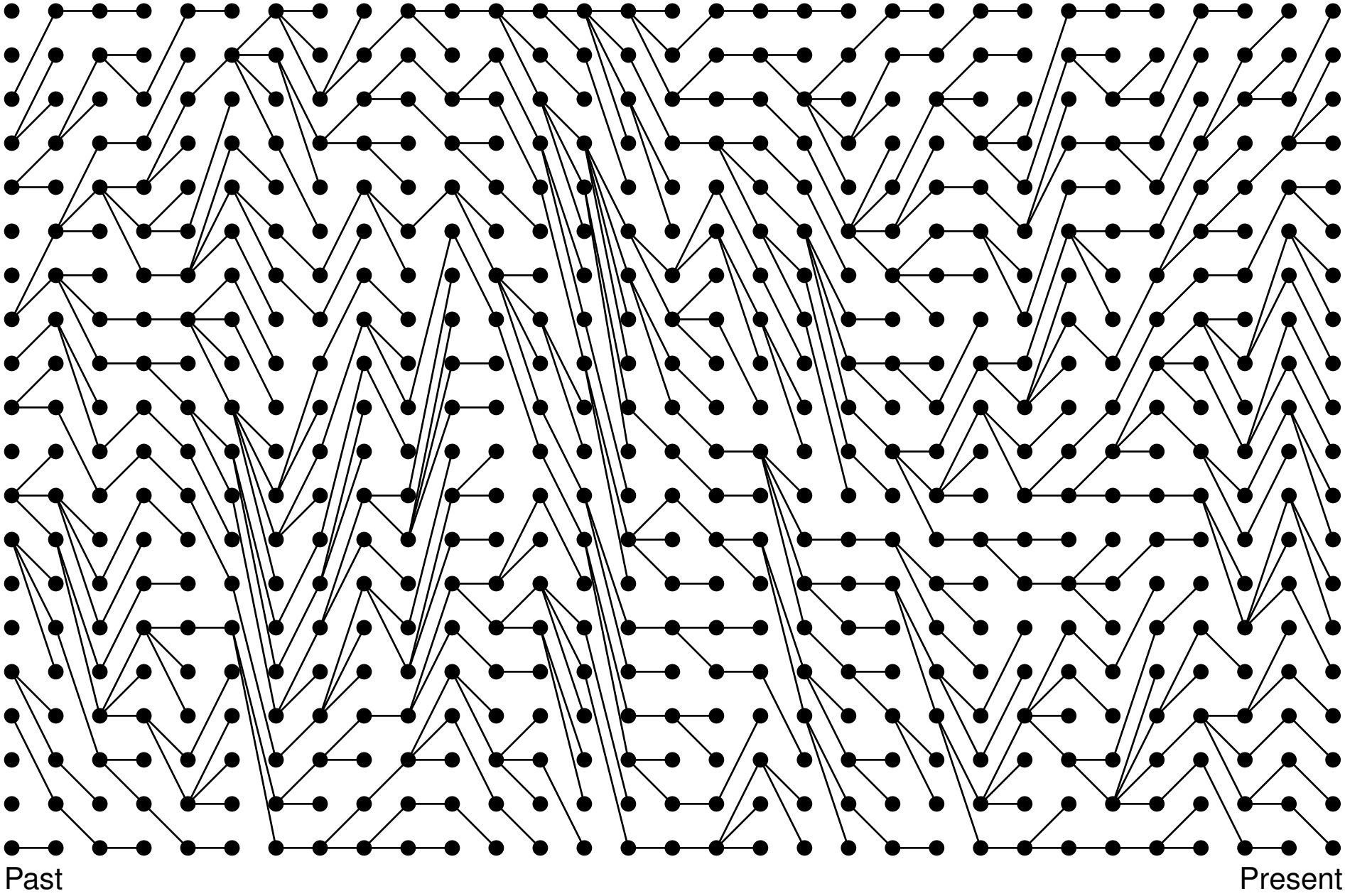
Population model



Past

Present

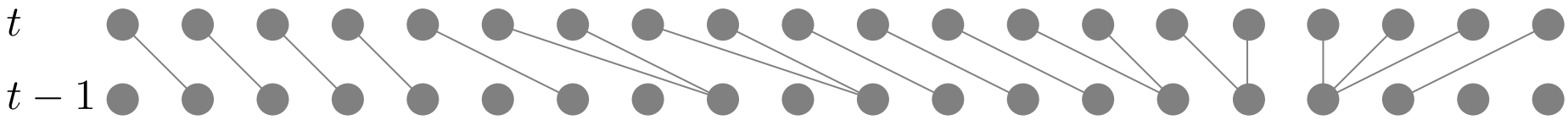
Population model



Population model

Wright

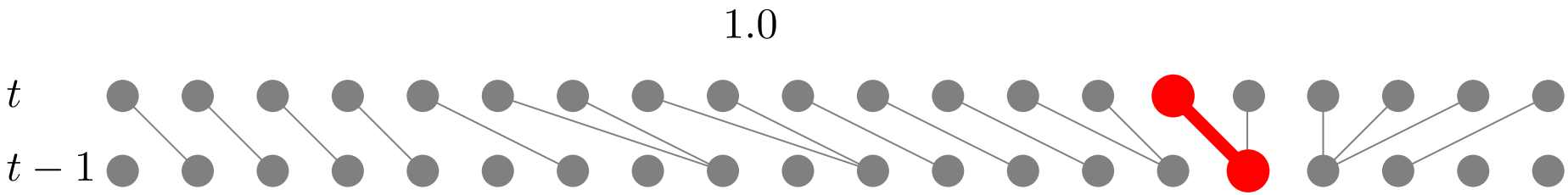
Sewall Wright evaluated the probability that two randomly chosen individuals in generation t have a common ancestor in generation $t-1$. If we assume that there are $2N$ chromosomes then the probability of sharing a common ancestor in the last generation is



Population model

Wright

Sewall Wright evaluated the probability that two randomly chosen individuals in generation t have a common ancestor in generation $t-1$. If we assume that there are $2N$ chromosomes then the probability of sharing a common ancestor in the last generation is

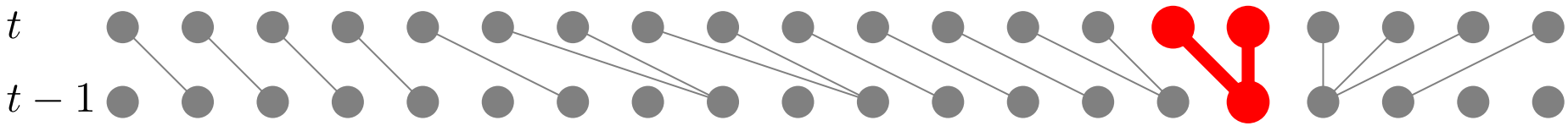


Population model

Wright

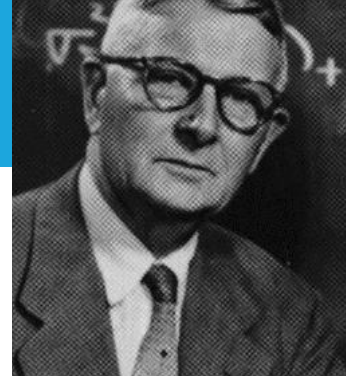
Sewall Wright evaluated the probability that two randomly chosen individuals in generation t have a common ancestor in generation $t-1$. If we assume that there are $2N$ chromosomes then the probability of sharing a common ancestor in the last generation is

$$1.0 \times \frac{1}{2N}$$



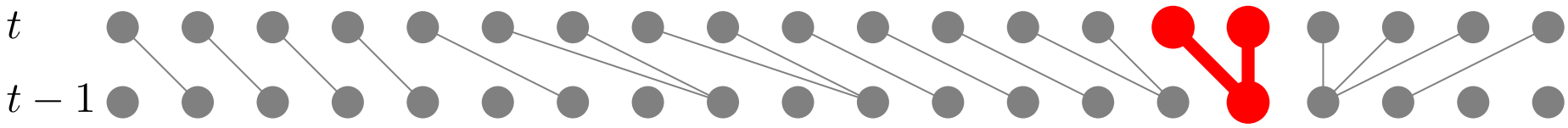
Population model

Wright



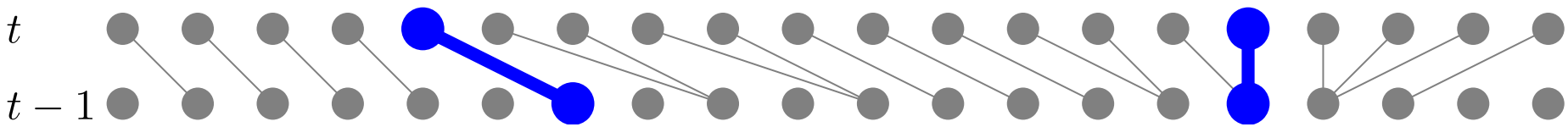
Sewall Wright evaluated the probability that two randomly chosen individuals in generation t have a common ancestor in generation $t-1$. If we assume that there are $2N$ chromosomes then the probability of sharing a common ancestor in last generation is

$$\frac{1}{2N}$$



The probability that two randomly picked chromosome do not have a common ancestor is

$$1 - \frac{1}{2N}$$



Population model

Wright

If we know the genealogy of the two individuals then we can calculate the probability as

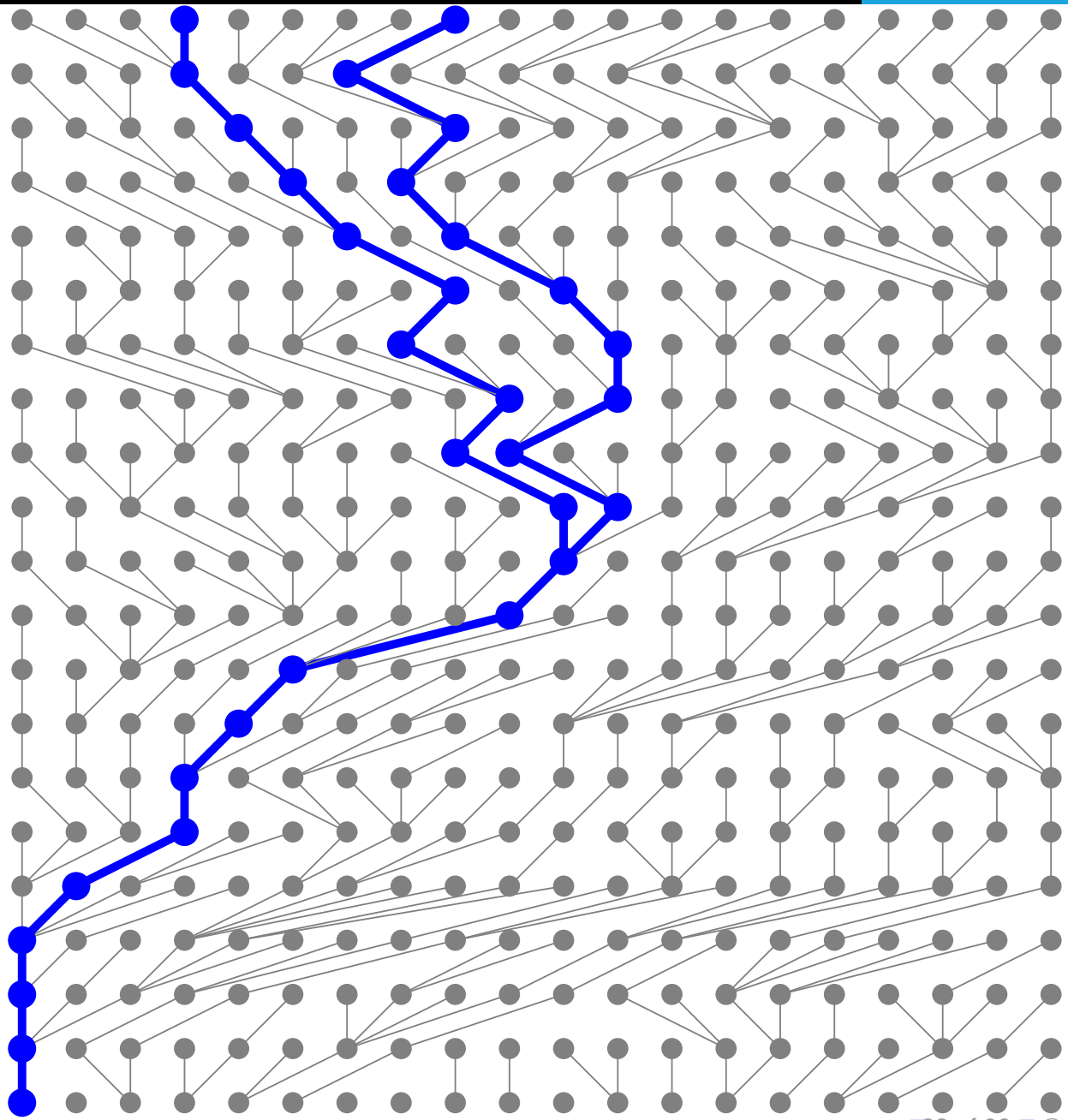
$$\mathrm{P}(\tau|N) = \left(1 - \frac{1}{2N}\right)^{\tau} \left(\frac{1}{2N}\right)$$

where τ is the number of generations with no coalescence. This formula is the Geometric Distribution and we can calculate the expectation of the waiting time until two random individuals coalesce:

$$\mathbb{E}(\tau) = 2N$$

Population model

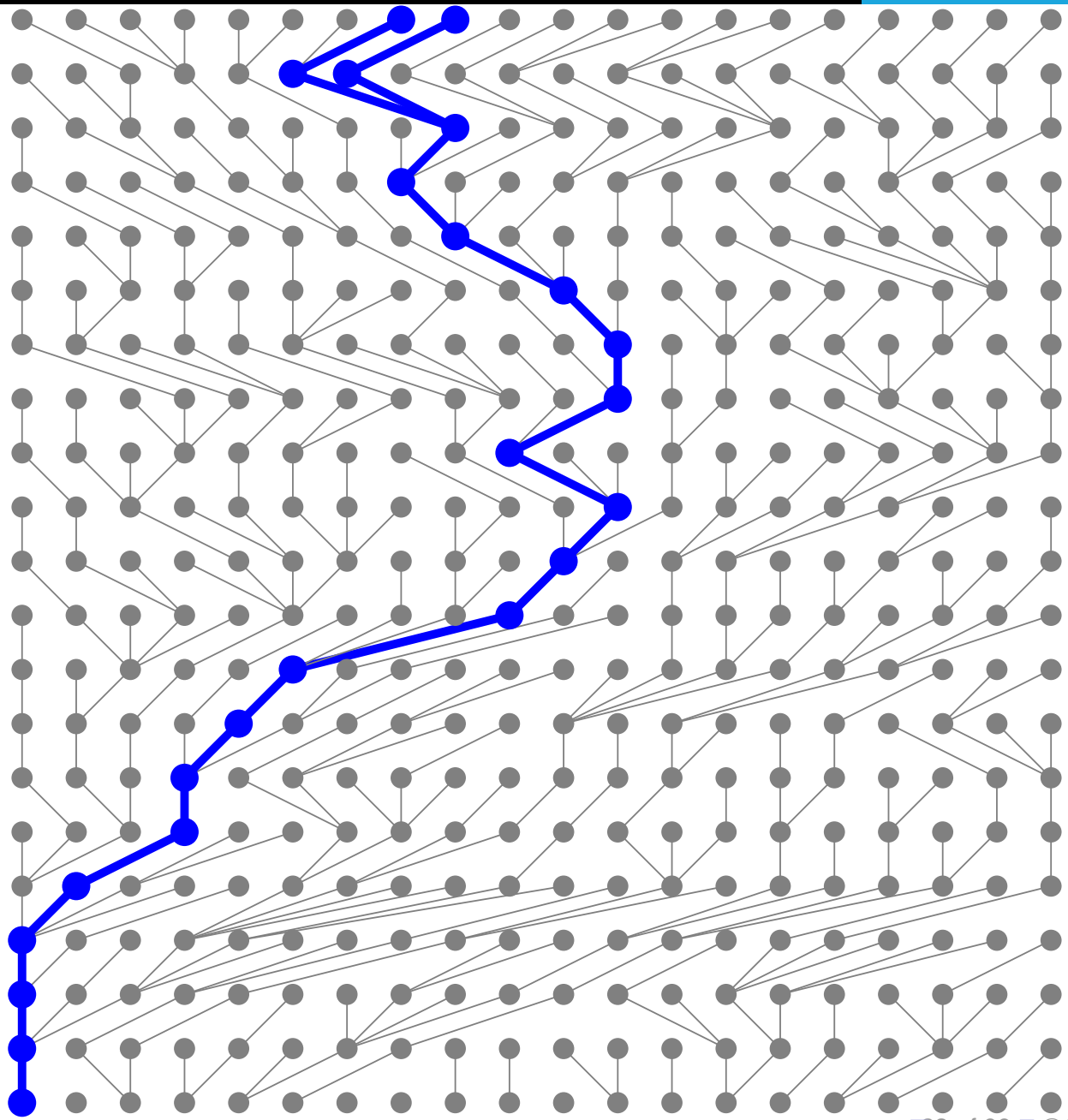
Present



Past

Population model

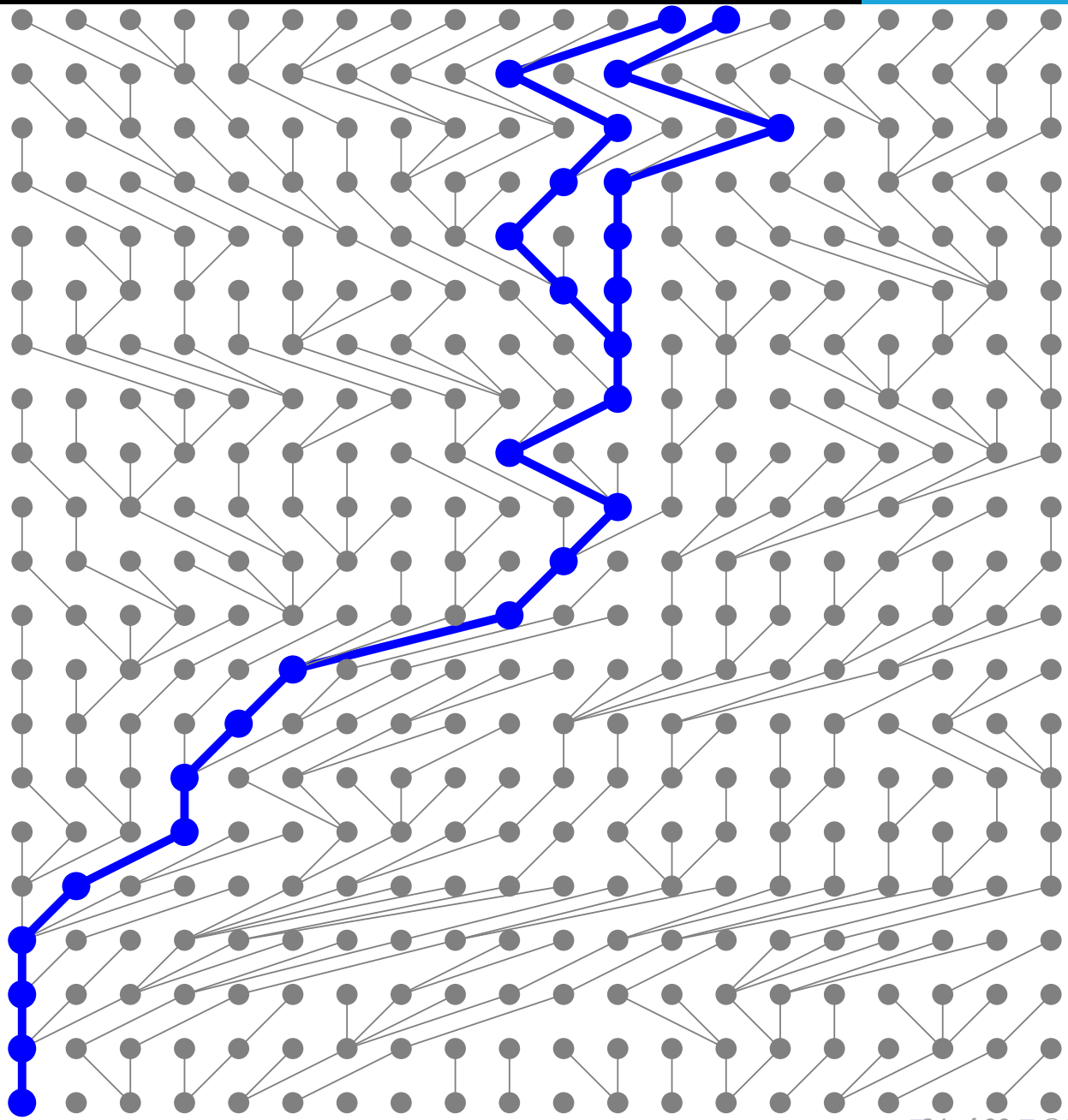
Present



Past

Population model

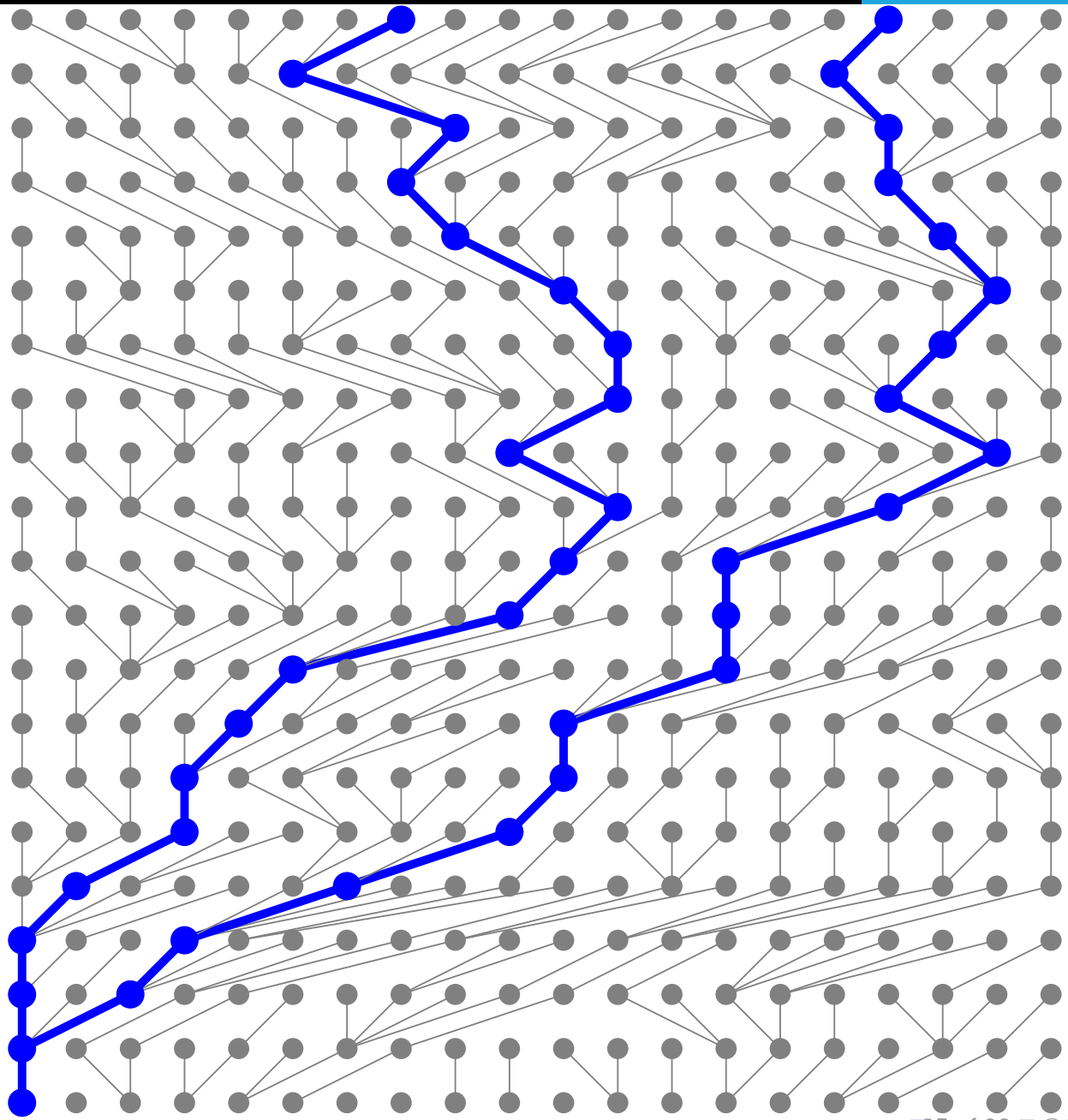
Present



Past

Population model

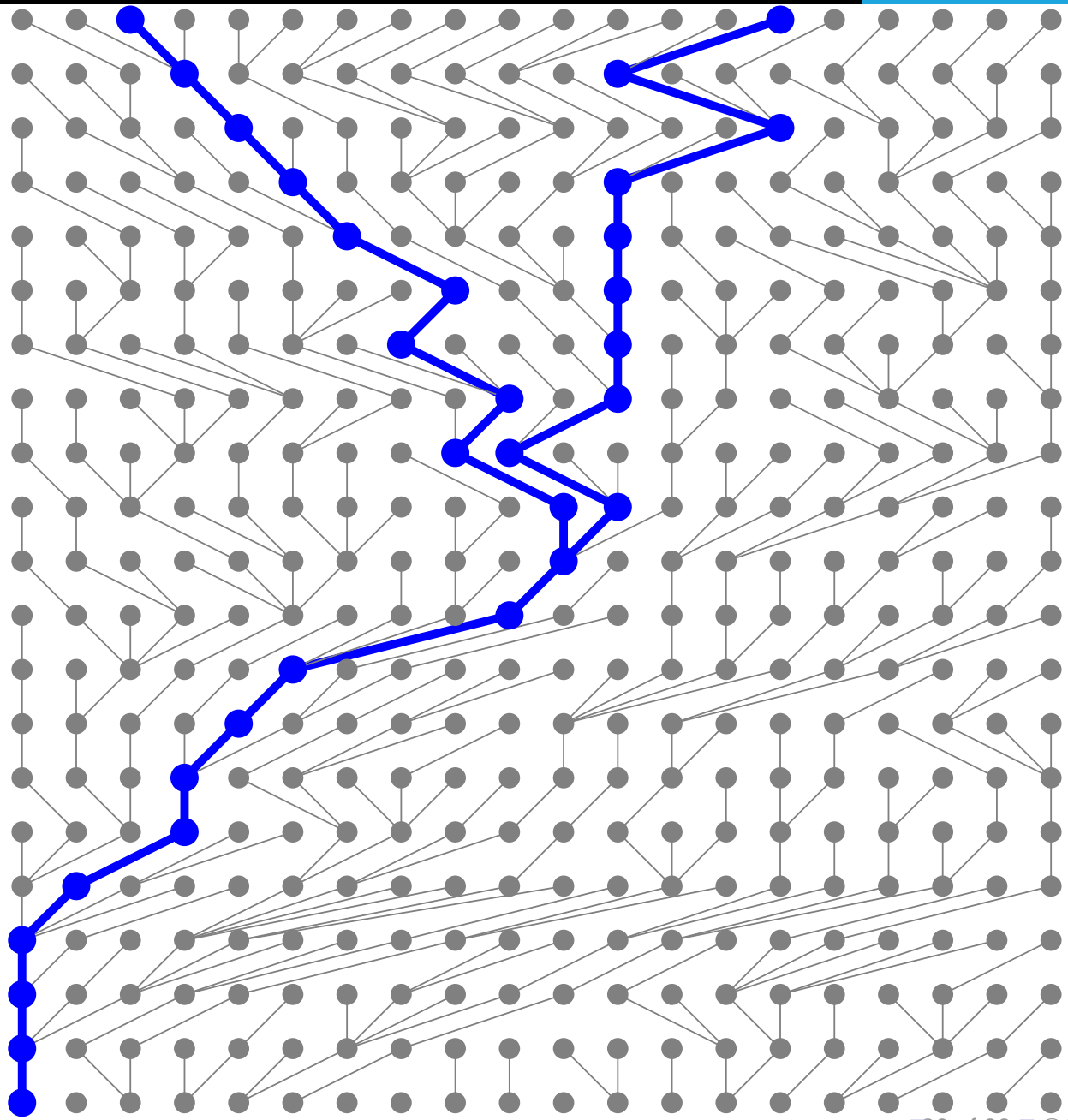
Present



Past

Population model

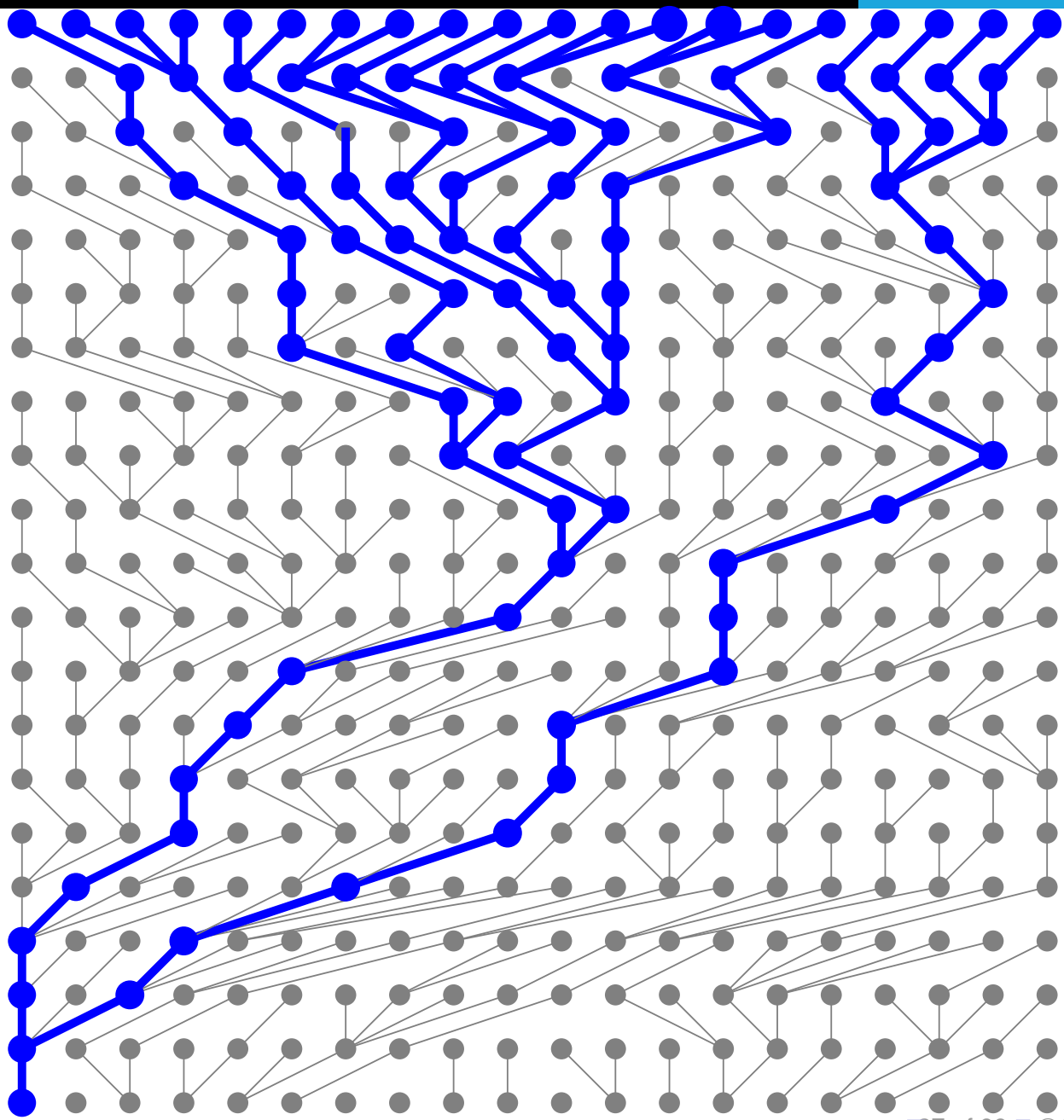
Present



Past

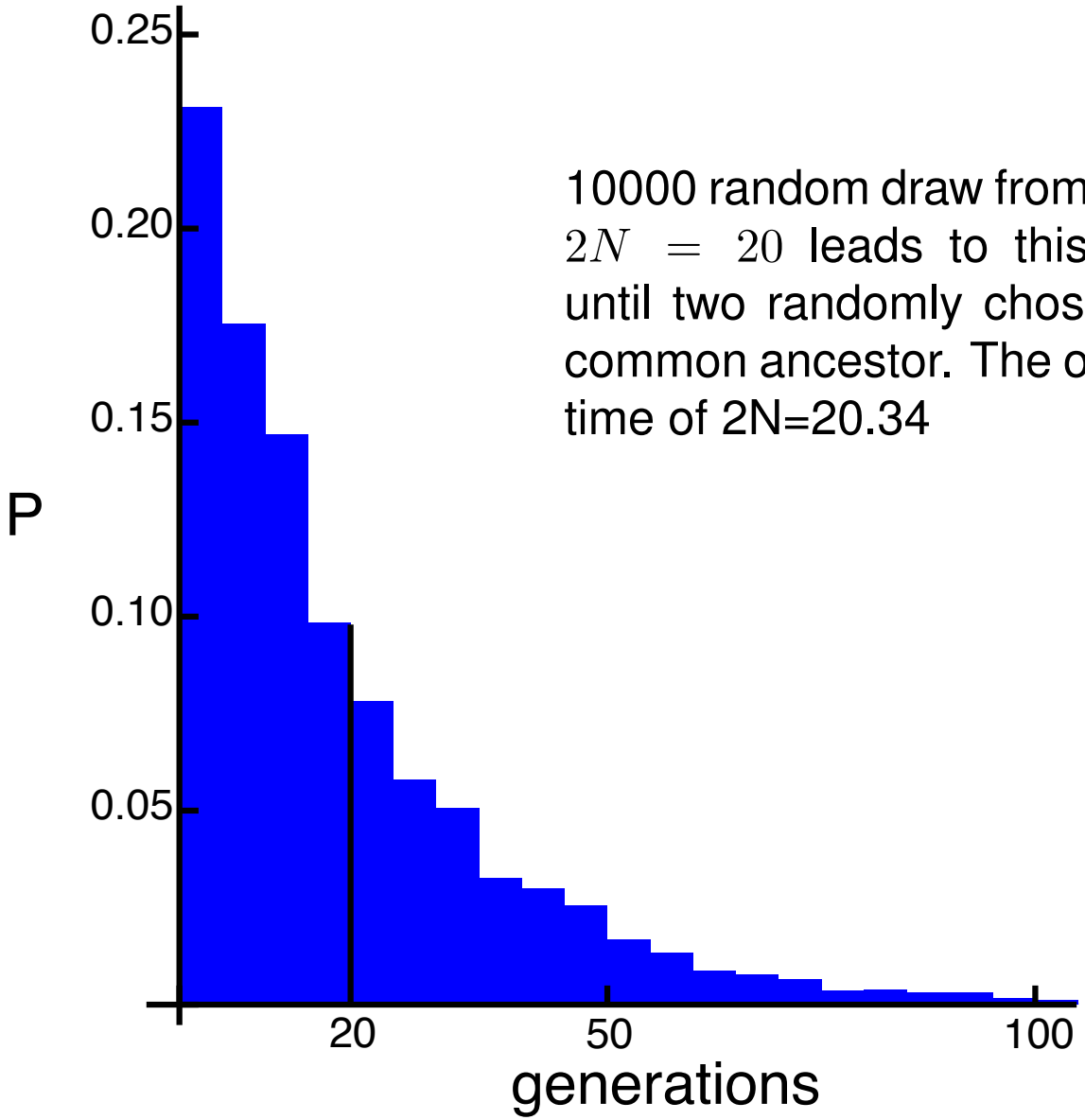
Population model

Present



Past

Probability Distribution



10000 random draw from a population with size $2N = 20$ leads to this distribution of times until two randomly chosen individuals have a common ancestor. The observed mean waiting time of 2N=20.34

Observations

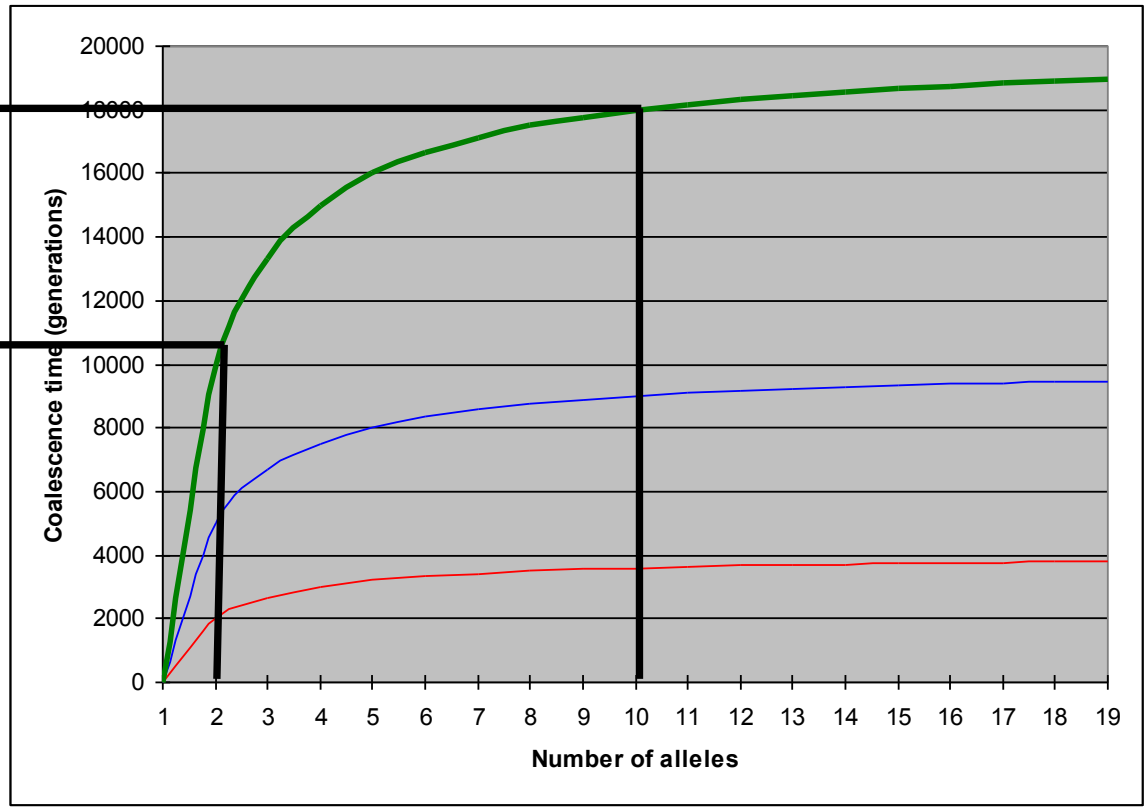
- For the time of coalescence in a sample of TWO , we will wait on average $2N$ generations assuming it is a Wright-Fisher population
- The model assumes that the generations are discrete and non-overlapping
- Real populations do not necessarily behave like a Wright-Fisher (the *'ideal' population)*
- *We assume that calculation using Wright-Fisher populations can be extrapolated to real populations.*

Properties of the Coalescent

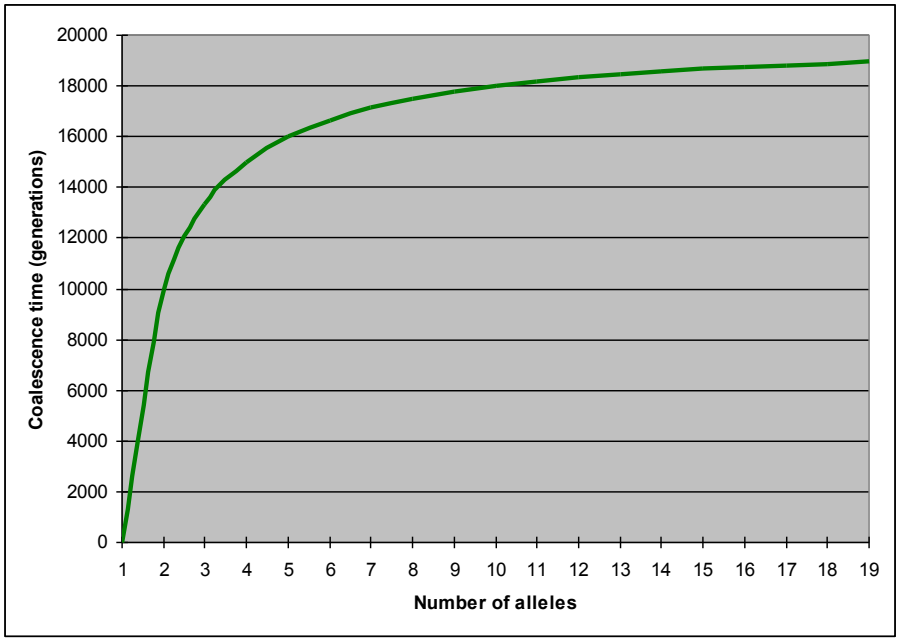
We start with 20 alleles and wait for them to coalesce until we reach the most recent common ancestor of all alleles

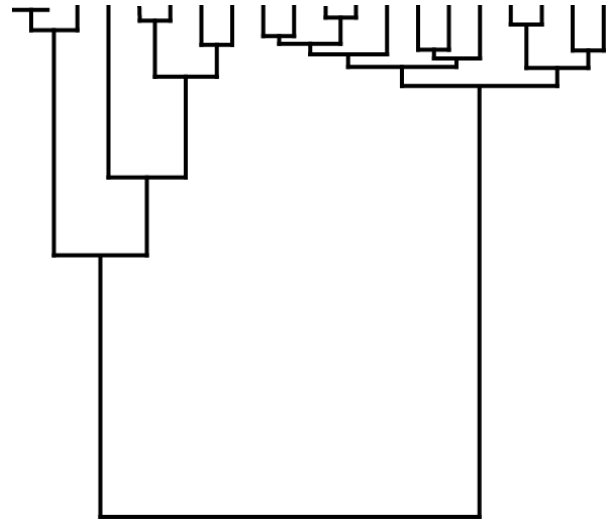
Half the alleles coalesce in the first 10% of time

50% of the total coalescence time is spent waiting for the last pair of alleles to coalesce!



Properties of the Coalescent





This means that coalescent trees are top-heavy!

Properties of the Coalescent

The fact that most branches coalesce at the top of the tree means that deep tree nodes can be inferred from a small number of gene copies

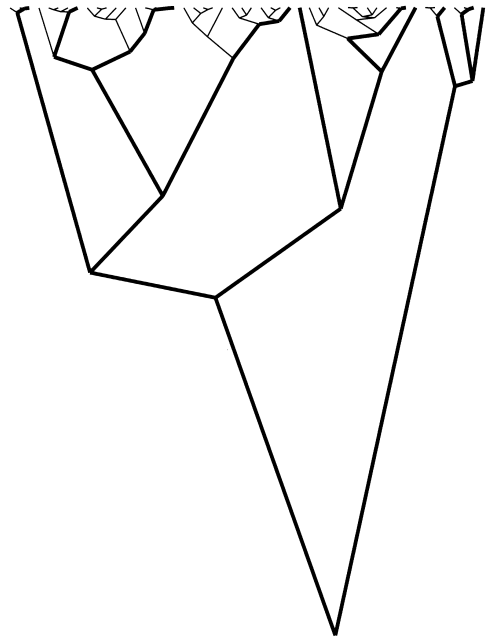
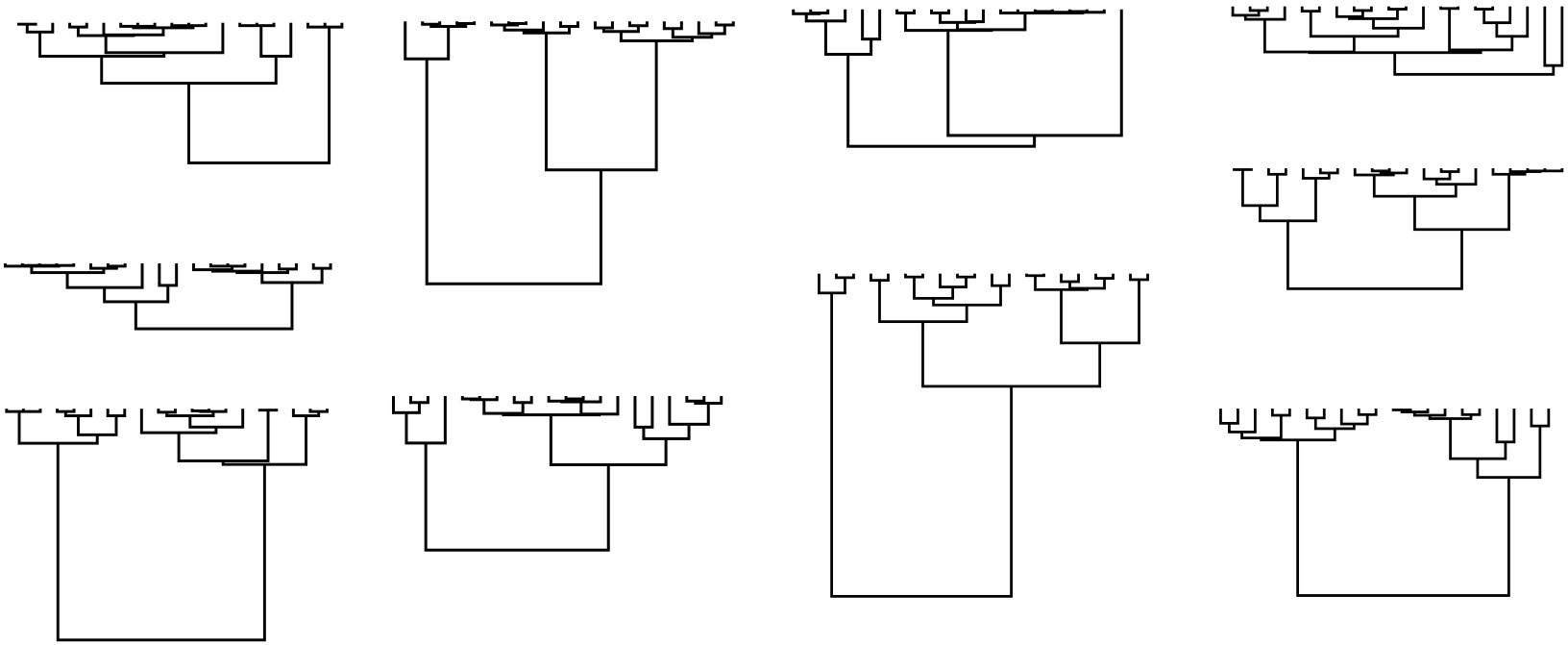


Figure 26.6: A sample genealogy of 50 gene copies, with the ancestry of a random 10 of them indicated by bold lines. Note that adding 40 more gene copies to the sample discloses no new lines in the bottom part of the diagram.

Properties of the Coalescent

The exponential nature of the time between coalescent events makes the coalescent distribution very noisy. These are tree simulated under a stochastic version of the coalescent with an identical N and k.

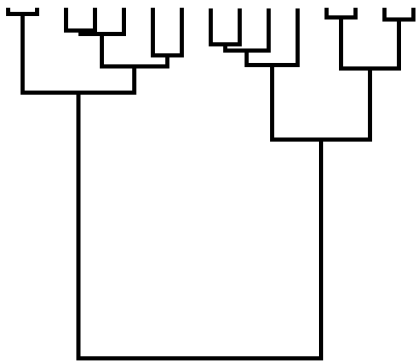


http://www.mesquiteproject.org

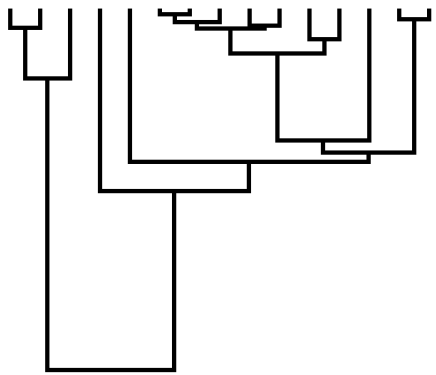
Properties of the Coalescent

The coalescent can be used to simulate a large number of possible genealogies. **Some of these genealogies are more likely than others**.

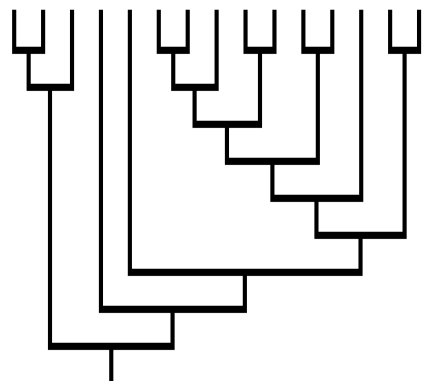
The most likely tree is one in which each coalescence event occurs exactly at the expected time according to the coalescent distribution. The further the topology of the simulated tree is from the expected distribution of the coalescent, the less likely it is to be the REAL history of population coalescence.



High likelihood



medium likelihood



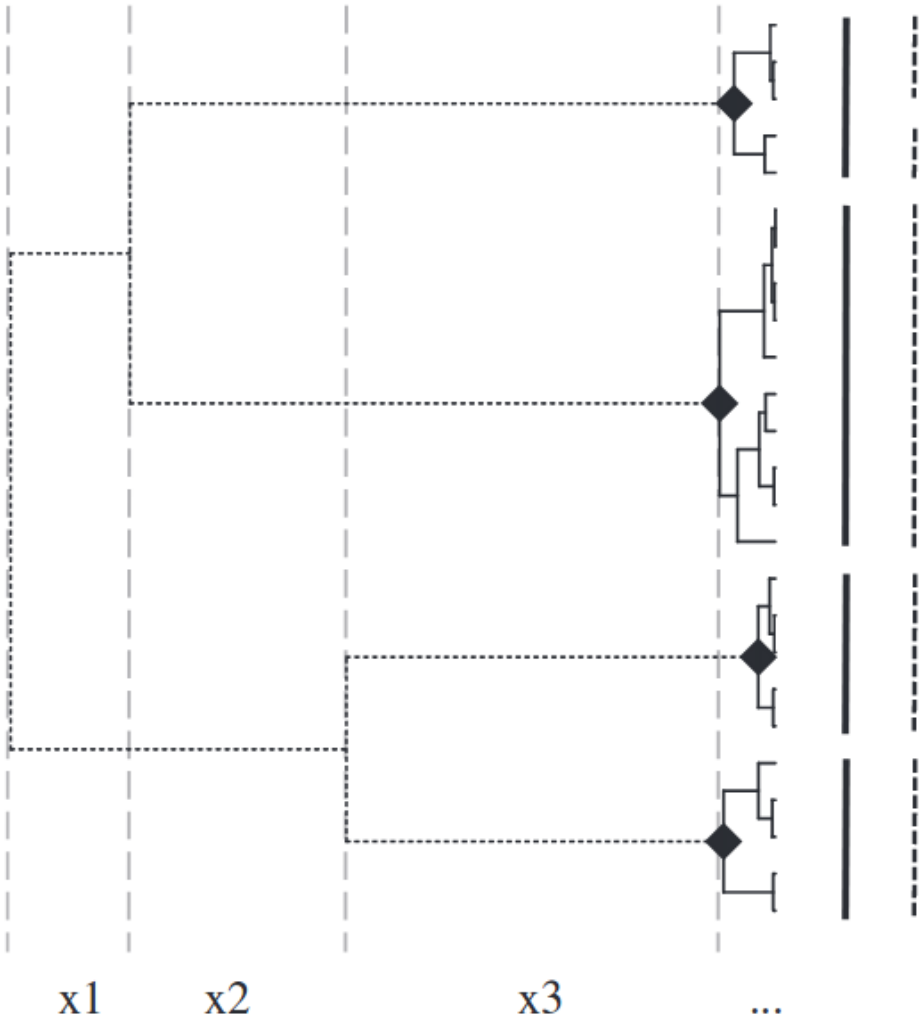
low likelihood

http://www.mesquiteproject.org

- What parameters drive expected timing of branching events of coalescent trees?

- What parameters drive expected timing of branching events of coalescent trees?
- What parameters drive expected timing of branching events of species trees?

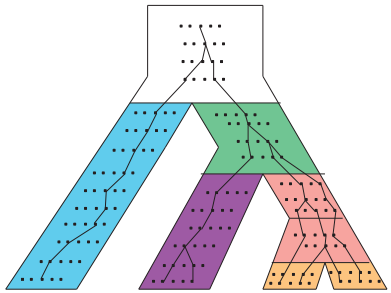
Generalized Mixed Yule Coalescent (GMYC) method is a likelihood method for delimiting species by fitting within- and between-species branching models to reconstructed gene tree



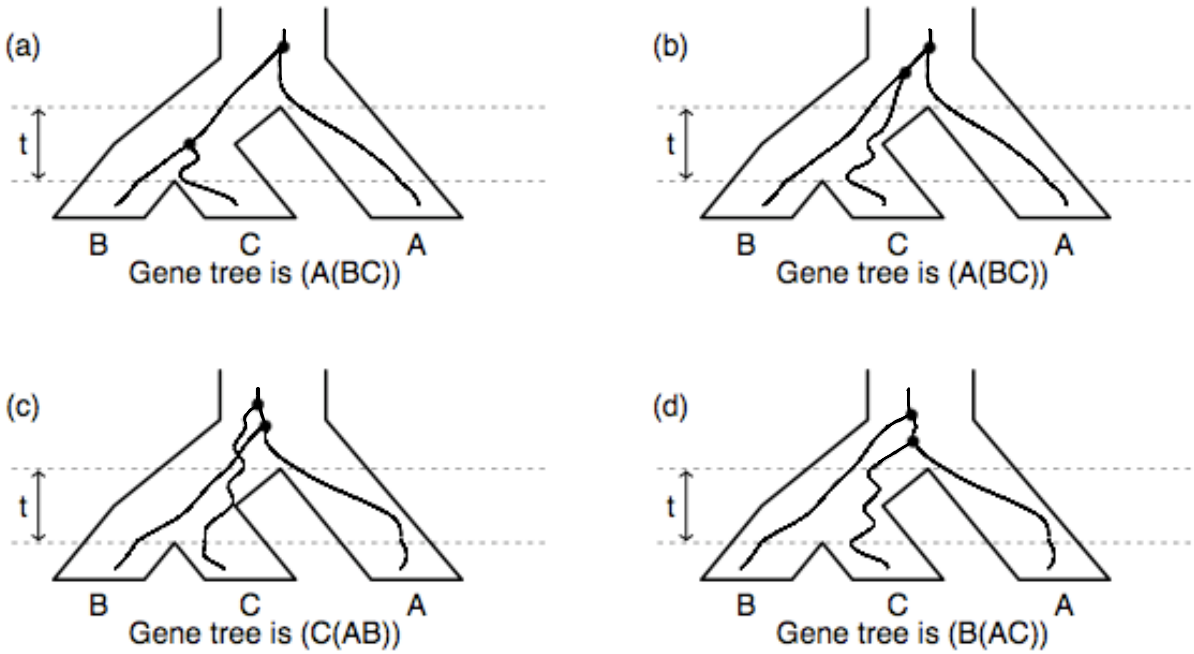
Fujisawa and Barraclough, 2013, Delimiting species using GMYC)

Relationship between population genetics and phylogenetics

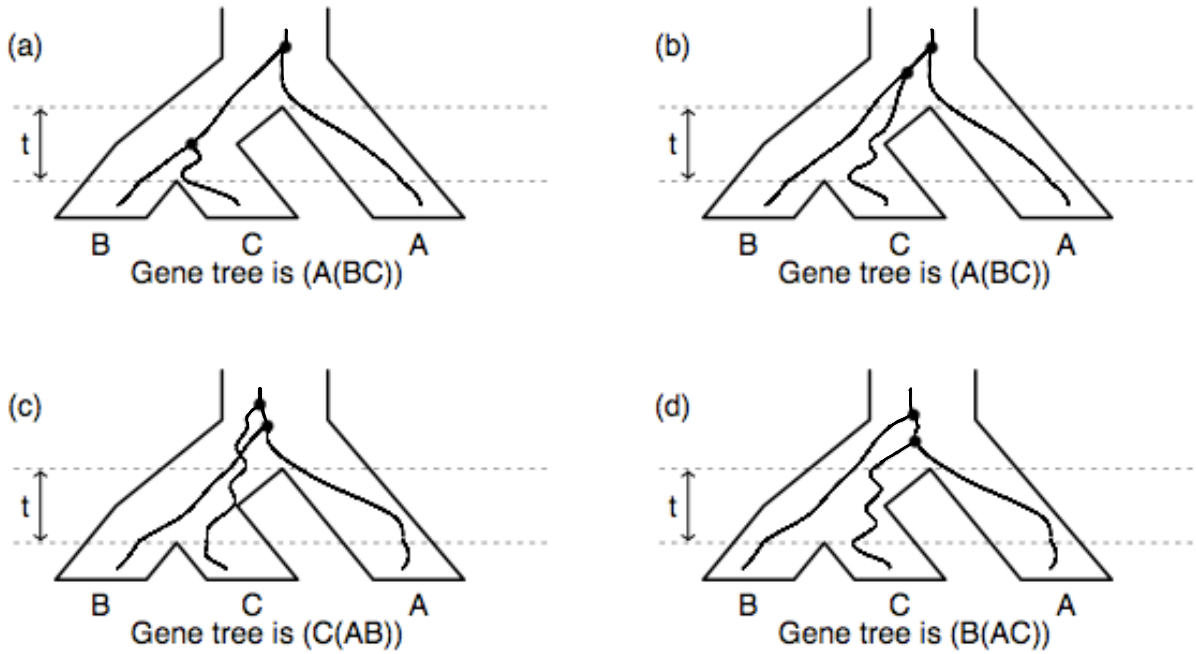
- Given current technology, we can do much more:
  - Sample many individuals within each taxon (species, population, etc.)
  - Sequence many genes for all individuals
- Need models at two levels:
  - Model what happens within each population
    [population genetics – coalescent model]
  - Link each within-population model on a phylogeny
    [phylogenetics]



Phylogenetic coalescent model

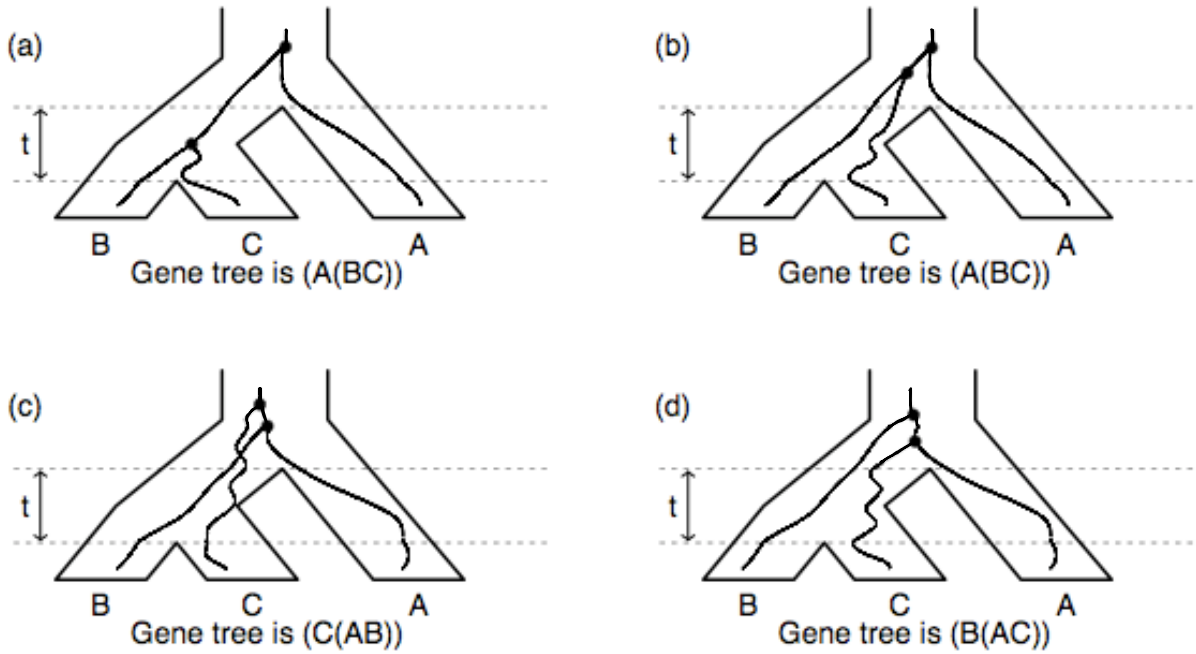


Phylogenetic coalescent model



t = length of interval between speciation events in **coalescent units**
= number of $2N$ generations

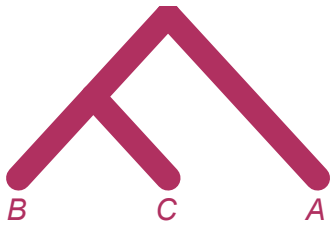
Phylogenetic coalescent model



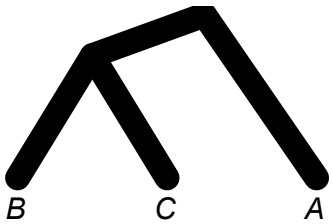
t = length of interval between speciation events in **coalescent units**
= number of $2N$ generations

Example: 1.2 coalescent units for an organisms with population size $N = 10,000$ and a generation time of 3 years $= 1.2 \times 20,000 \times 3 = 72,000$ years

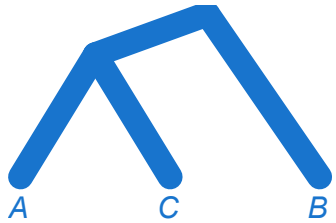
Probabilities of each gene tree history are shown below them
t = length of interval between speciation events



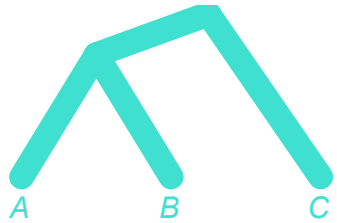
$1 - e^{-t}$



$\frac{1}{3}e^{-t}$



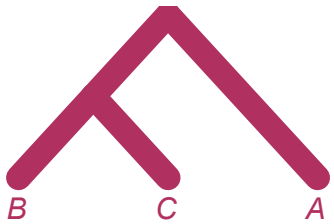
$\frac{1}{3}e^{-t}$



$\frac{1}{3}e^{-t}$

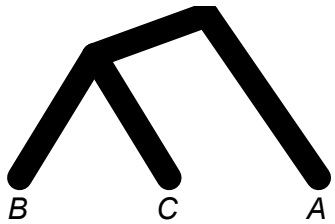
Phylogenetic coalescent model

t = length of interval between coalescent events $= 1.0$



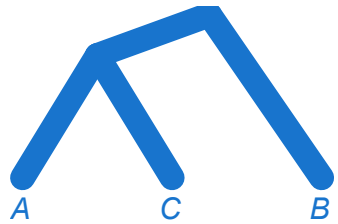
$1 - e^{-t}$

0.63



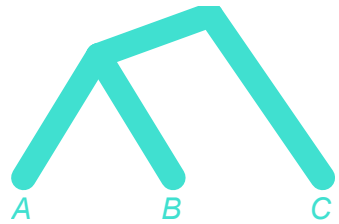
$\frac{1}{3}e^{-t}$

0.12



$\frac{1}{3}e^{-t}$

0.12

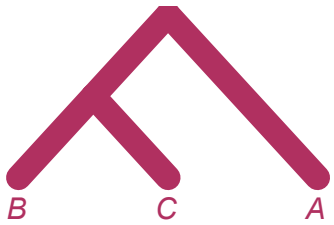


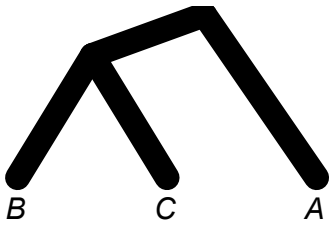
$\frac{1}{3}e^{-t}$

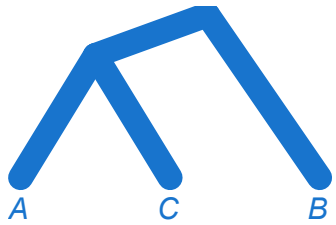
0.12

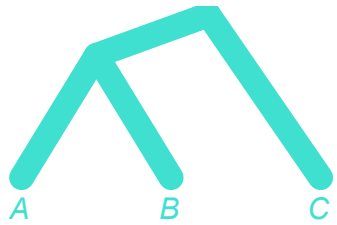
Phylogenetic coalescent model

t = length of interval between coalescent events $= 1.0 = 0.5$





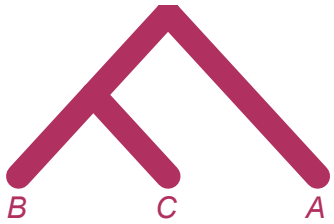


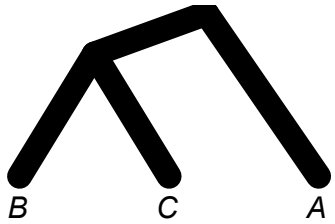


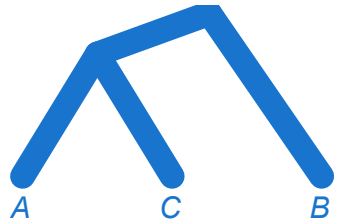
$1 - e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$
0.63	0.12	0.12	0.12
0.40	0.20	0.20	0.20

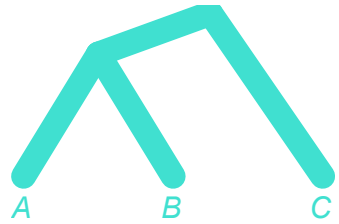
Phylogenetic coalescent model

t = length of interval between coalescent events $= 1.0 = 0.5 = 2.0$





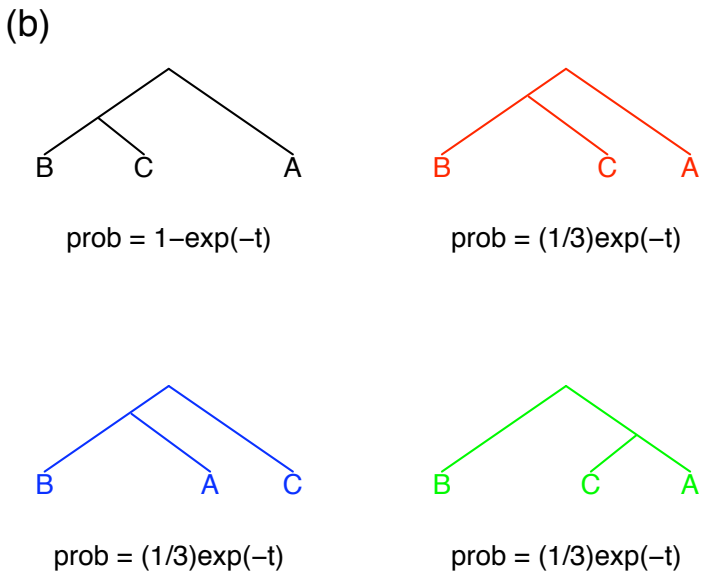


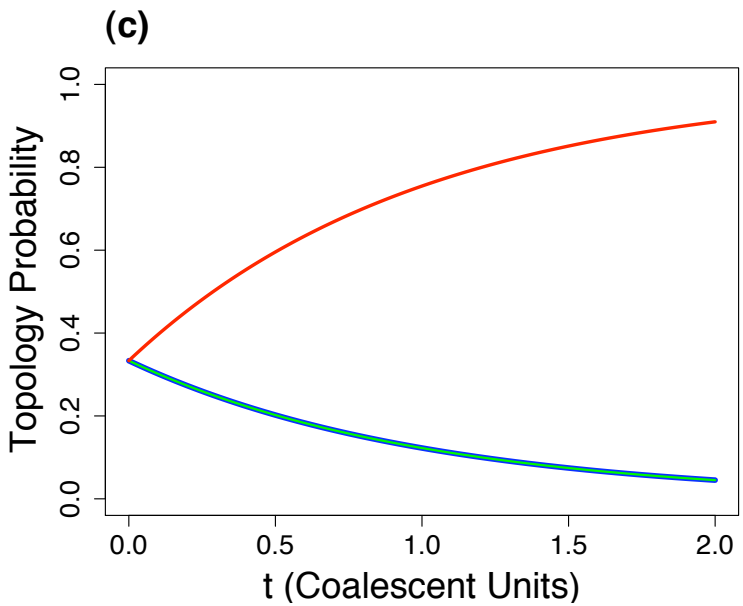


$1 - e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$
0.63	0.12	0.12	0.12
0.40	0.20	0.20	0.20
0.85	0.05	0.05	0.05

Example: Computation of Gene Tree Topology Probabilities for the 3-taxon Case

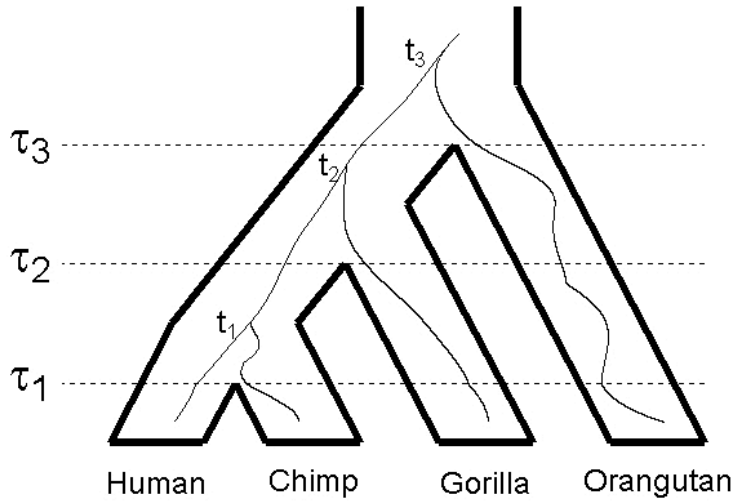
- What are these probabilities like as a function of t, the length of time between speciation events?





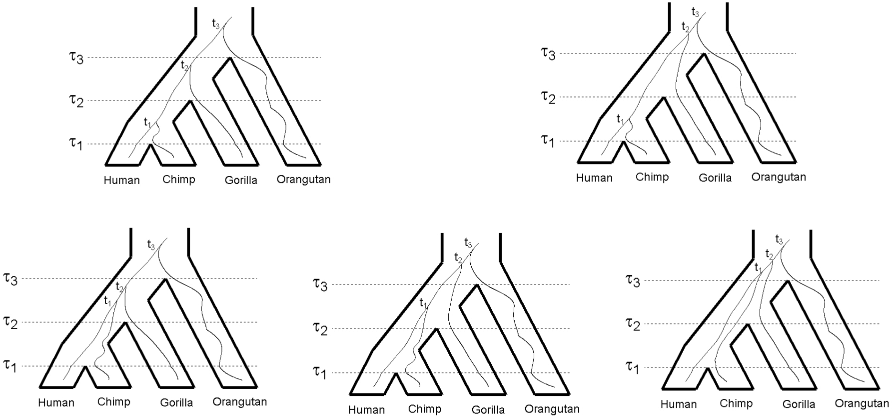
Example: a slightly larger case

- Consider 4 taxa – the human-chimp-gorilla problem

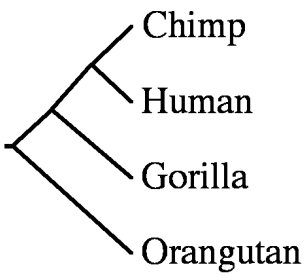


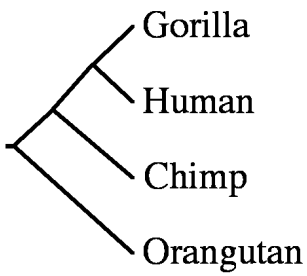
Coalescent histories for the 4-taxon example

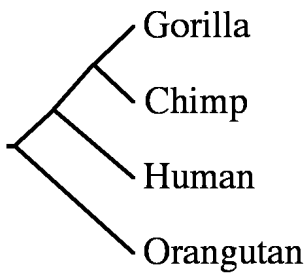
- There are 5 possible histories for this example:



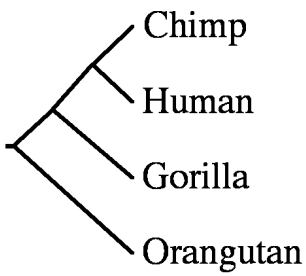
Applications of the topology distribution - example 1



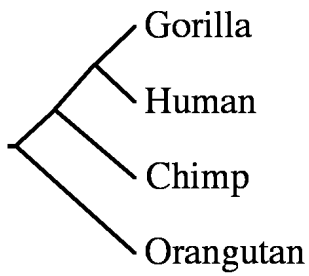




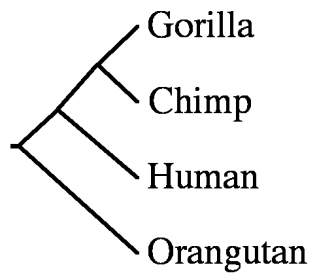
Applications of the topology distribution - example 1



76.6%



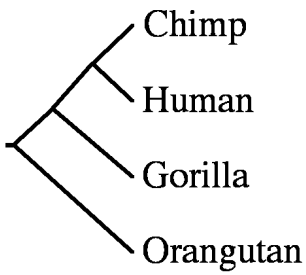
11.4%

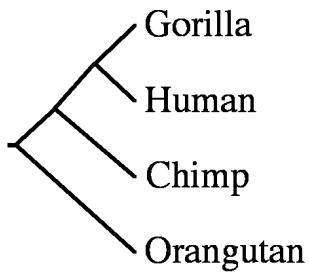


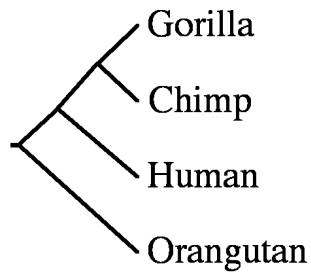
11.5%

Observed proportions of each gene tree among ML phylogenies

Applications of the topology distribution - example 1







76.6% 11.4% 11.5%

79.1% 9.9% 9.9%

Observed proportions of each gene tree among ML phylogenies

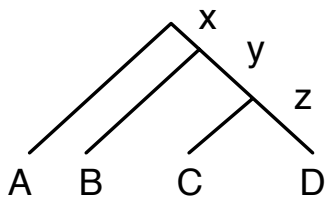
Predicted proportions using parameters from Rannala & Yang, 2003.

Applications of the topology distribution - example 2

- In the previous example, one topology is clear preferred
- Must the distribution always look this way?
- Examine the entire distribution when the number of taxa is small

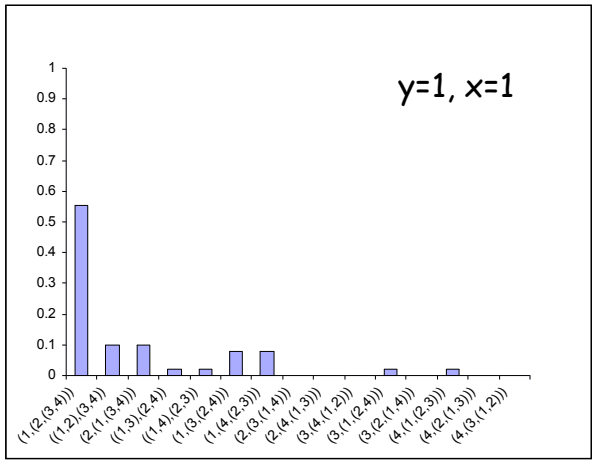
Applications of the topology distribution - example 2

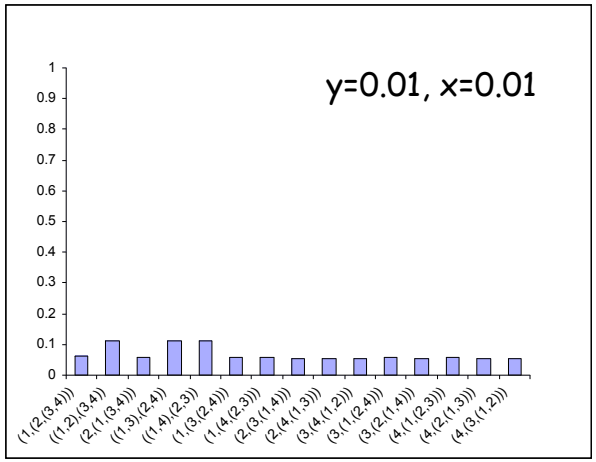
- Consider 4 taxa: A, B, C, and D
- Species tree:

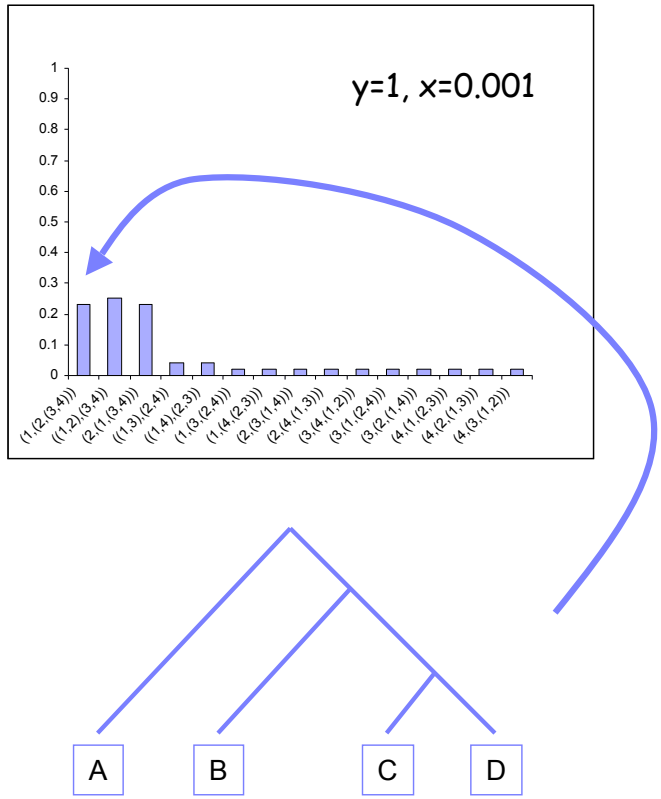


- Look at probabilities of all 15 tree topologies for values of x, y, and z

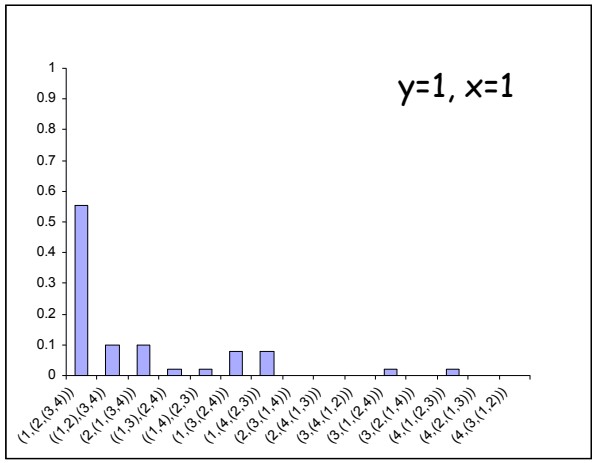
Applications of the topology distribution - example 2

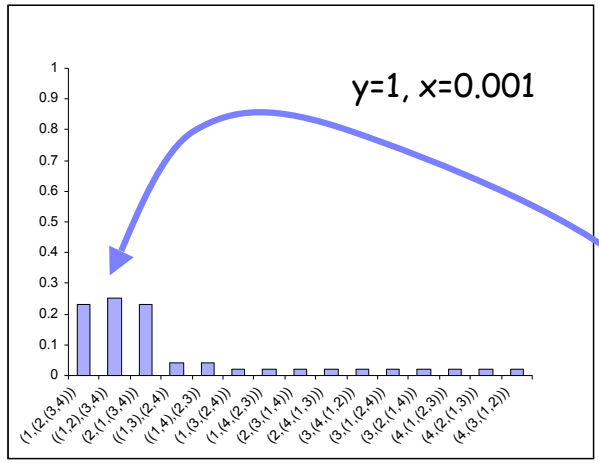


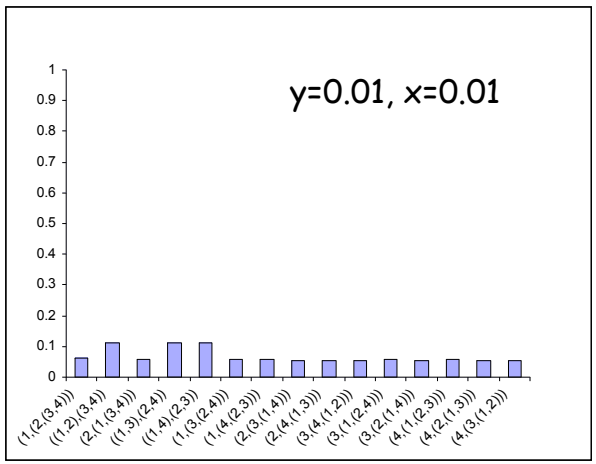


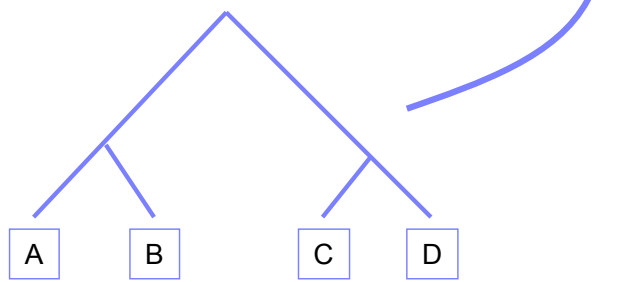


Applications of the topology distribution - example 2

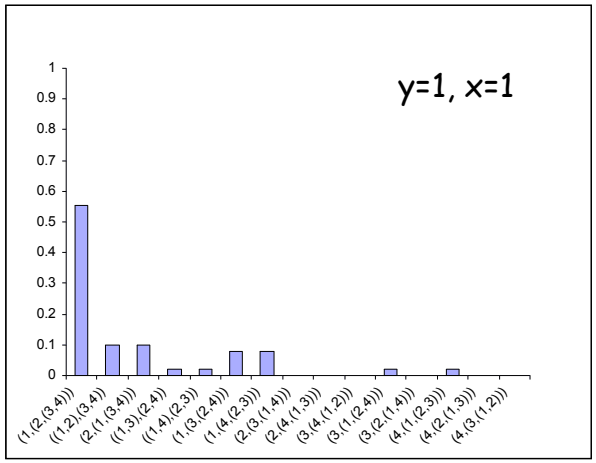


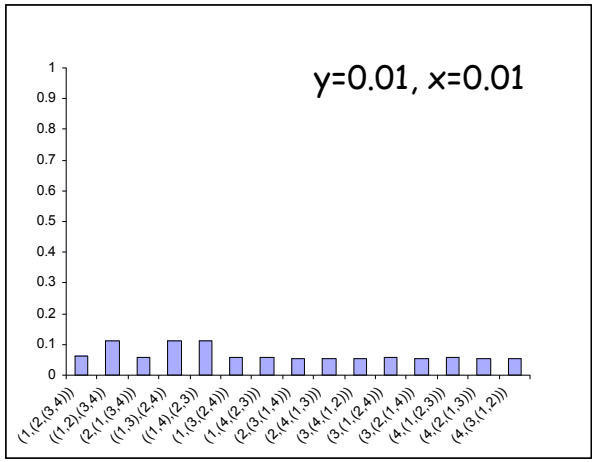


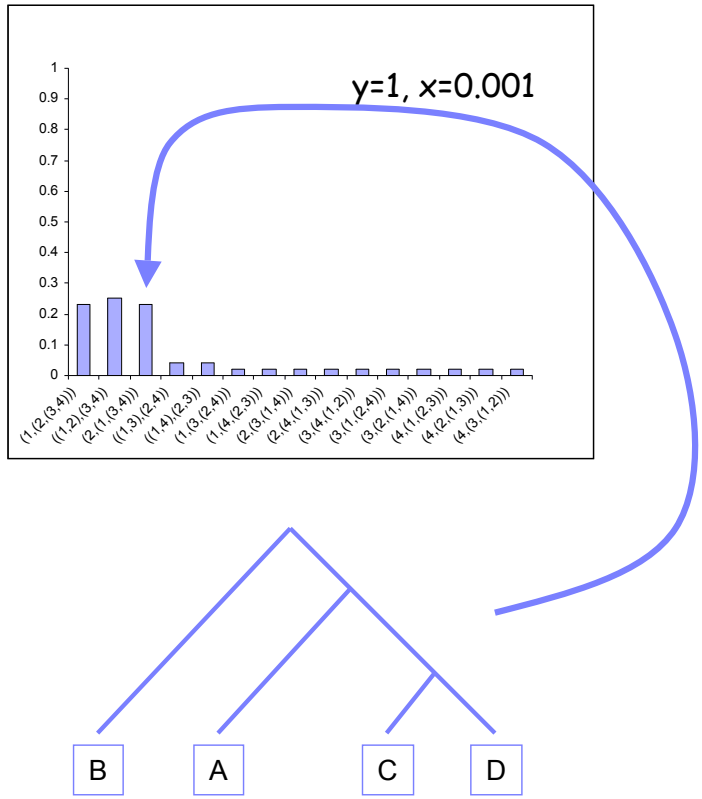




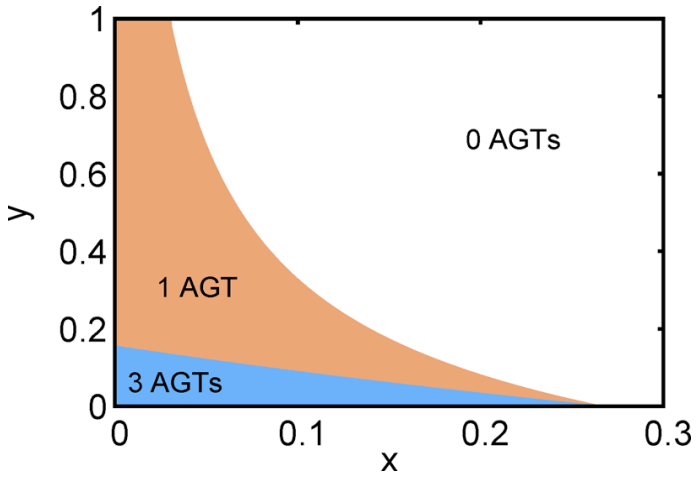
Applications of the topology distribution - example 2







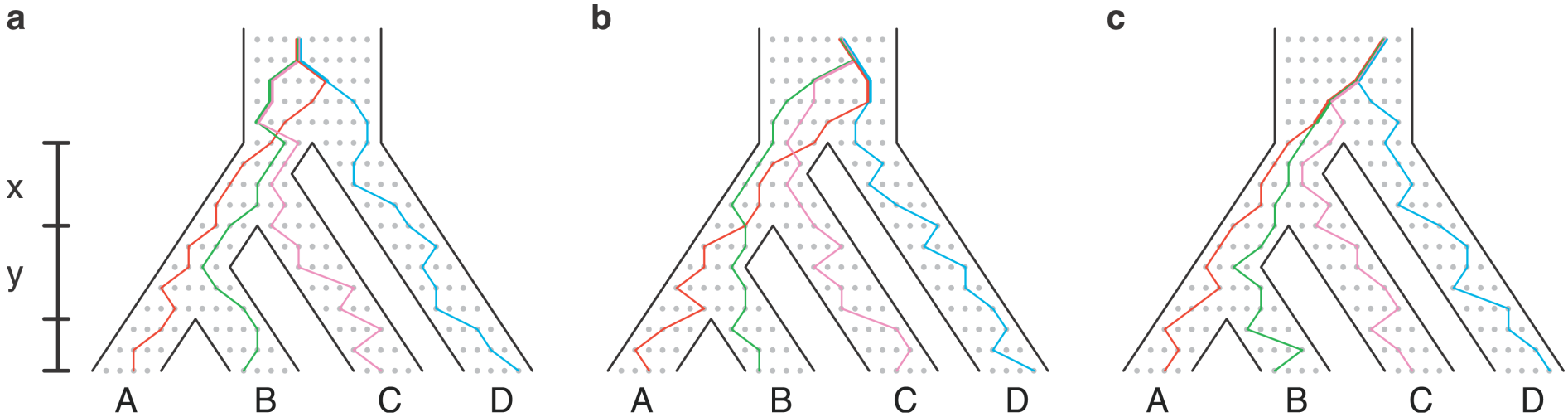
Applications of the topology distribution - example 2



- The existence of anomalous gene trees has implications for the inference of species trees

Degnan and Rosenberg, *PLoS Genetics*, 2006

Rosenberg and Tao, *Systematic Biology*, 2008



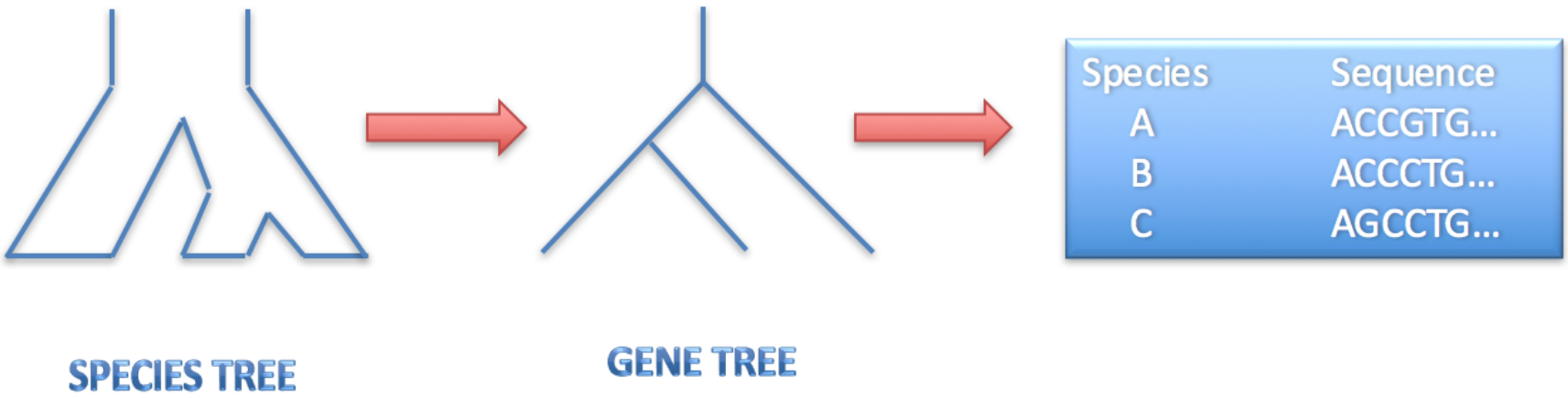
If the internal branches of the species tree—x and y—are short so that coalescences occur deep in the tree, the two sequences of coalescences that produce a given symmetric gene tree topology together have higher probability than the single sequence that produces the topology that matches the species tree. (a) and (b) Two coalescence sequences leading to gene tree topology ((AD)(BC)). In (a), the lineages from B and C coalesce more recently than those from A and D, and in (b), the reverse is true. (c) The single sequence of coalescences leading to gene tree topology (((AB)C)D). (Degnan and Rosenberg, 2006)

How can we deal with incorporate gene tree variation in our species tree estimation?

Is there a better way to estimate species phylogenies?

Explicitly model the coalescent process!

Phylogenetic coalescent model with mutation



The likelihood function

- Suppose that we have available alignments for N genes, denoted by $D_1, D_2, \ldots, D_N$
- We would like to find the likelihood of the species phylogeny given these N alignments, assuming that
  - individual gene trees are randomly generated according to the coalescent
  - evolution of sequences along fixed gene trees occurs following a standard nucleotide-based Markov model
  - the data for the genes are independent given the species tree and associated parameters

Likelihood function

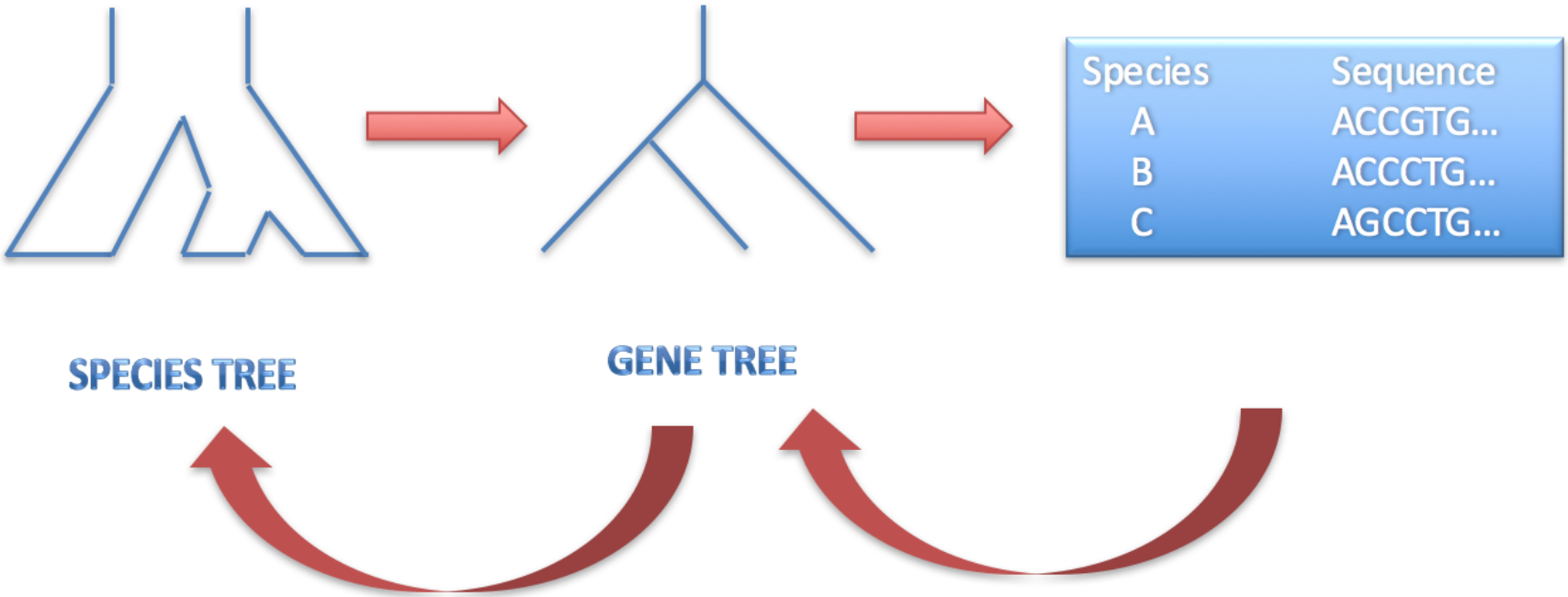
- Recall the Felsenstein equation from Peter's lecture, except that now we replace θ with S, the species tree. Use this to form the species tree likelihood for a multi-locus data set:

$$\begin{aligned} L(S|D_1, D_2, \ldots D_N) &= \prod_{i=1}^{N} P(D_i|S) \text{ [loci conditionally independent]} \\ &= \prod_{i=1}^{N} \sum_{j=1}^{G} P(D_i|g_j) f(g_j|S) \end{aligned}$$

  where S is the species tree (topology and branch lengths) and g_j represents a gene tree.

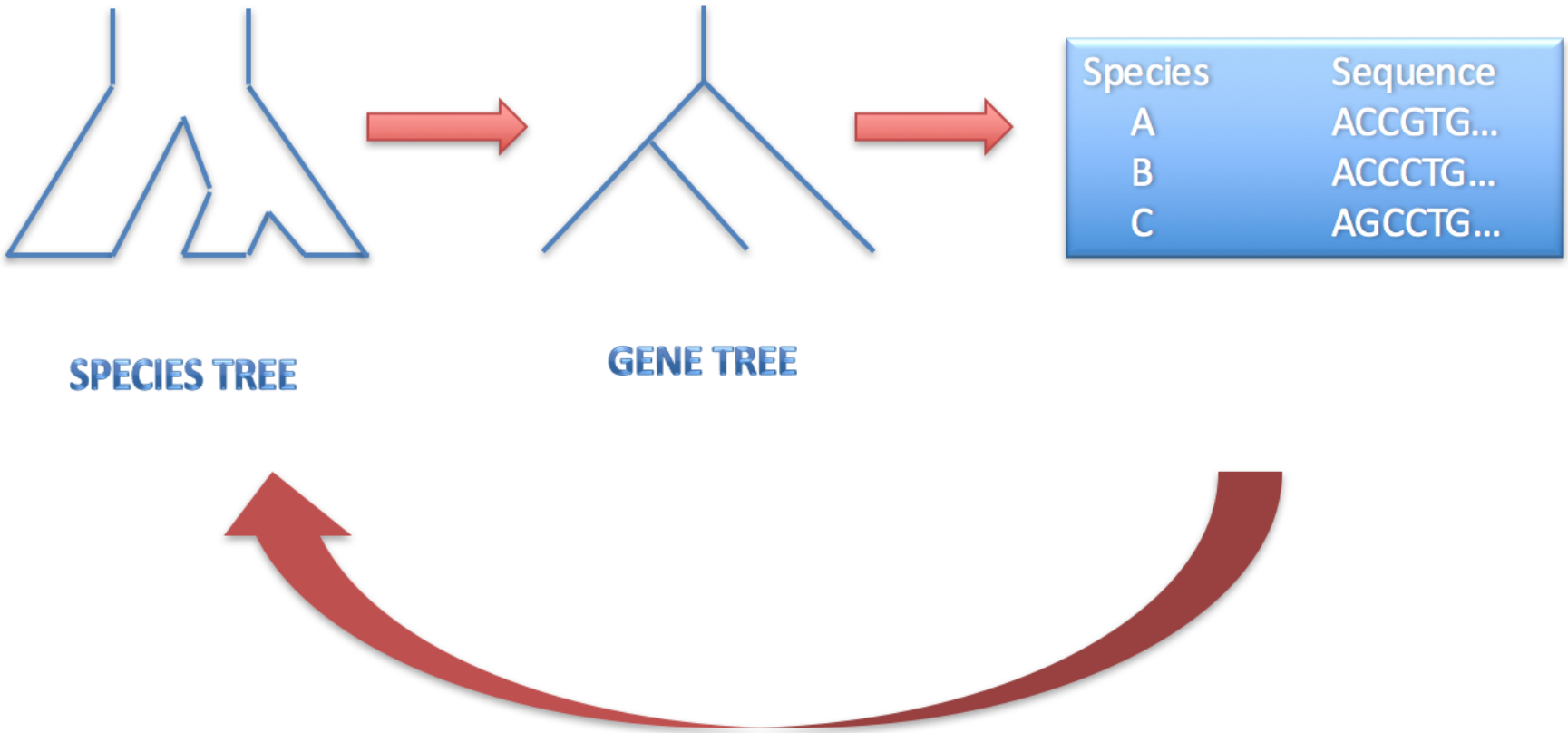
- This likelihood is difficult to evaluate directly, because of the dimension of he inner sum (which is really an integral) [recall Peter's "galaxy slide"]

Inference option 1: Summary statistics methods



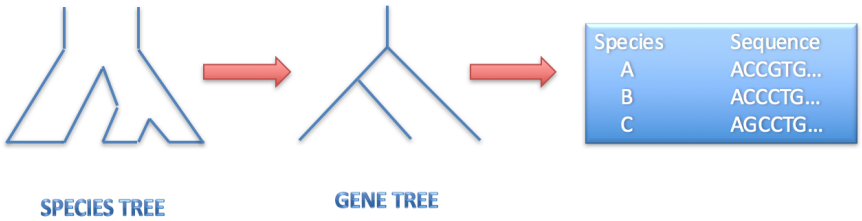
- Summary statistics methods: Start with estimated gene trees
  - Using estimated branch lengths:
    - ⋆ STEM (Kubatko et al. 2009)
    - ⋆ STEAC (Liu et al. 2009)
  - Using topology information only:
    - ⋆ STAR (Liu et al. 2009)
    - ⋆ Minimize Deep Coalescences (PhyloNet; Than & Nakhleh 2009)
    - ⋆ MP-EST (Liu et al. 2010)
    - ⋆ ST-ABC (Fan and Kubatko 2011)
    - ⋆ STELLS (Wu 2011)
    - ⋆ ASTRAL (Mirarab et al. 2014)
    - ⋆ Statistical binning (Bayzid et al. 2014)

Inference option 2: Full data methods



Full data methods I: BEST, *BEAST, BPP, SNAPP

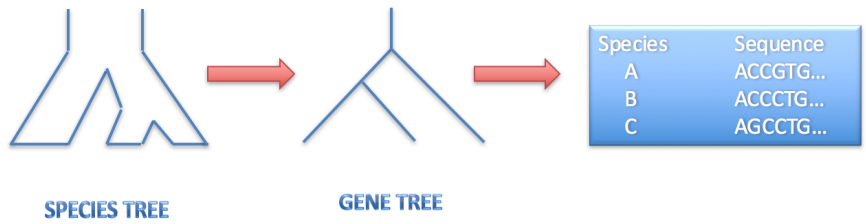
- Model the entire process of data generation
- Goal of these methods is to estimate the posterior distribution of the gene trees and species tree and associated model parameters



- BEST, *BEAST, and BPP use MCMC by considering both gene trees and the species tree, but their implementations are different
- SNAPP uses a clever two-step peeling algorithm to carry out the integration over gene trees, allowing it to consider a reduced space – but currently limited to biallelic data.

Full data methods II: SVDQuartets

- Model the entire process of data generation
- Avoid computing the likelihood by using algebraic structure in the distribution of site pattern probabilities under the model



- SVDQuartets is implemented in PAUP*
- SVDQuartets will be discussed in detail in this afternoon's lab

- Comparison of approaches:
  - Summary statistics methods
    - ⋆ Advantage: Quick
    - ⋆ Disadvantage: Ignore information in the data
    - ⋆ Most current implementations do not easily allow assessment of uncertainty (but bootstrap can be used, at the expense of computational efficiency)
  - Full data methods
    - ⋆ Advantage: Fully model-based framework
    - ⋆ Disadvantage: Computationally intensive, sometimes prohibitively so
    - ⋆ BEST, *BEAST, BPP, and SNAPP utilize a Bayesian framework and involve MCMC

Species Tree Inference Summary – Comparison of Methods

Software	Data Type	Measure of Uncertainty	Computation Time	Models Included
BEST	multilocus	posterior probability	long; can be run in parallel	coalescent; all reversible substitution models
*BEAST	multilocus	posterior probability	intermediate; can be run in parallel	coalesent; all reversible substitution models; relaxed clock; variable population sizes
BPP	multilocus	posterior probability	long	coalescent; JC69 model only; molecular clock; species delimitation
SVDQ	multilocus; SNP	bootstrap	short	coalescent; all reversible substitution models; non-clock; gene flow; parameter estimation ?
SNAPP	biallelic SNP; AFLP	posterior probability	long; can be run in parallel	coalescent; two-state substitution model; Bayes factor delimitation
ASTRAL	unrooted gene trees	bootstrap	short given gene trees	no specific model assumed
MP-EST	rooted gene trees	bootstrap	short given gene trees	coalescent model

Species Tree Inference Summary

- Failure to incorporate the coalescent model in estimation of the species tree can lead to statistical inconsistency, even when a method that is statistically consistent is applied.
- Many new methods for inferring species trees are being developed – each has its advantages and disadvantages.
- In addition, we should continue to think about other ways of using multi-locus data to its full advantage .... and we should be thinking beyond estimation of the species tree.
- Lots of areas emerging: species delimitation, incorporating horizontal events along the phylogeny, etc.

Next week: computer lab on using Astral and SVD quartets to build species trees.