Species Tree Estimation

Laura Kubatko
Departments of Statistics and
Evolution, Ecology, and Organismal Biology
Mathematical Biosciences Institute
The Ohio State University

kubatko.2@osu.edu
twitter: Laura_Kubatko

May 27, 2023

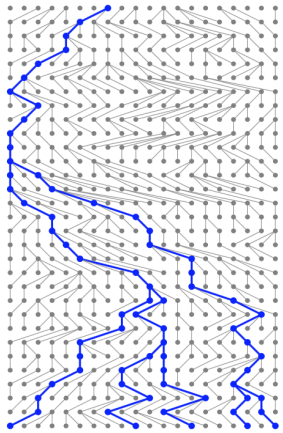
Relationship between population genetics and phylogenetics

- Population genetics: Study of genetic variation within a population
- Phylogenetics: Use genetic variation between taxa (species, populations) to infer evolutionary relationships
- Previously:
  - Each taxon is represented by a single sequence – "exemplar sampling"
  - We have data for a single gene and wish to estimate the evolutionary history for that gene (the **gene tree** or **gene phylogeny**)
- Now:
  - Sample many individuals within each taxon (species, population, etc.)
  - Sequence many genes for all individuals

Relationship between population genetics and phylogenetics

- Need models at two levels:

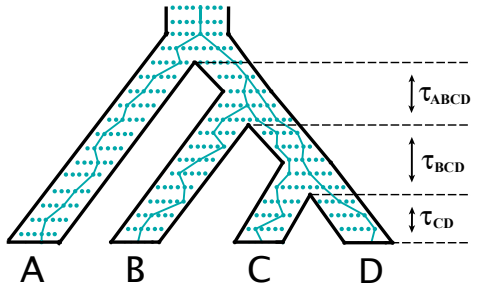
1. Model what happens within each population



$\rightarrow$ *coalescent model*

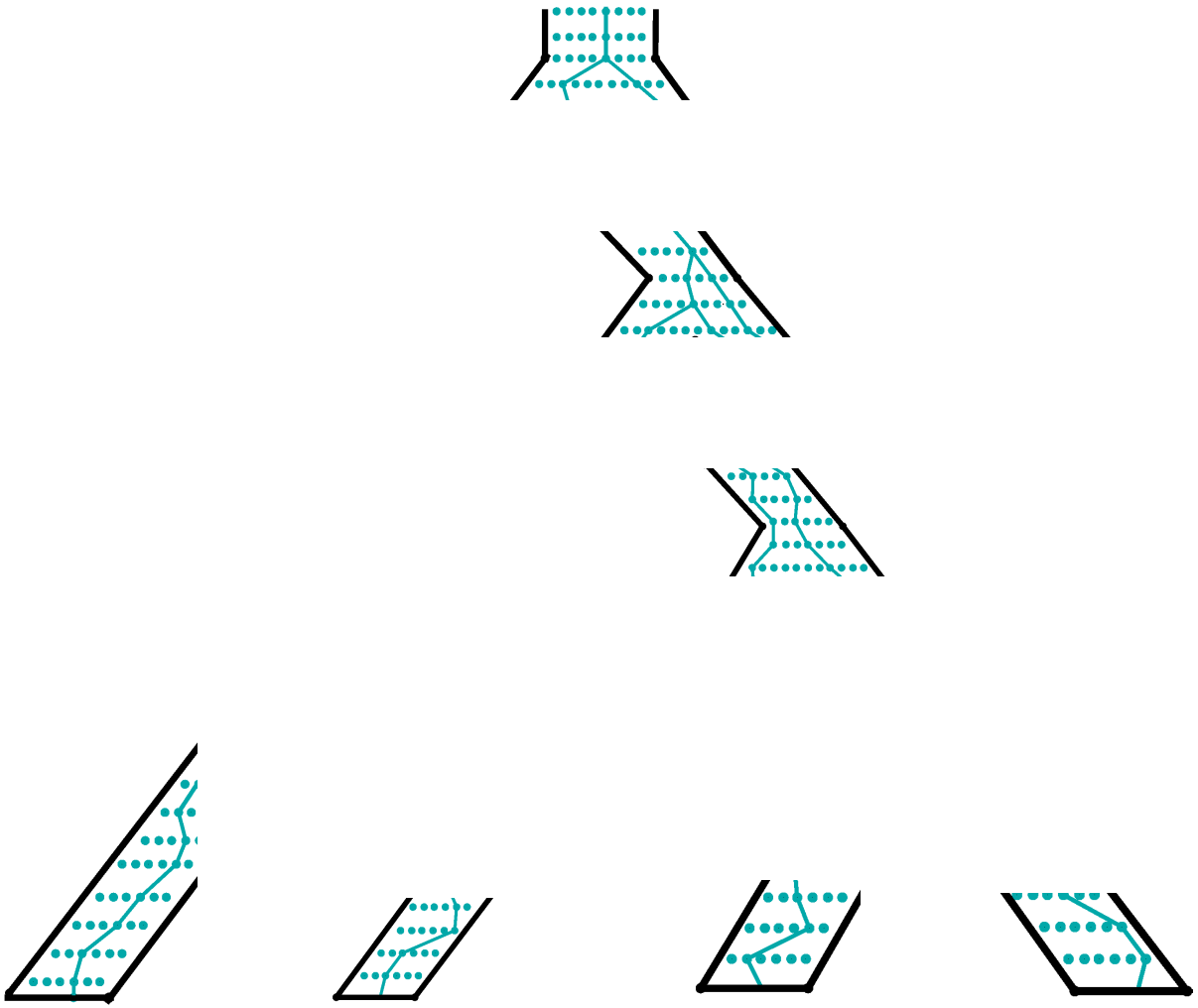
Peter's talk in our first session

2. Link each within-population model on a phylogeny



Relationship between population genetics and phylogenetics

- Build up the species tree from many populations:

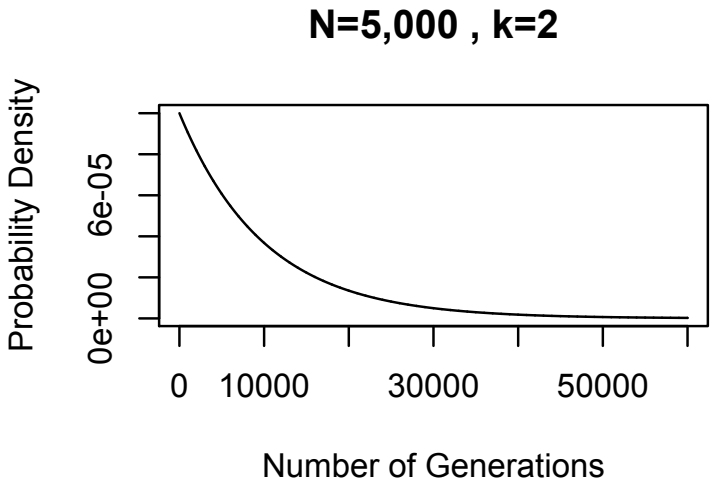


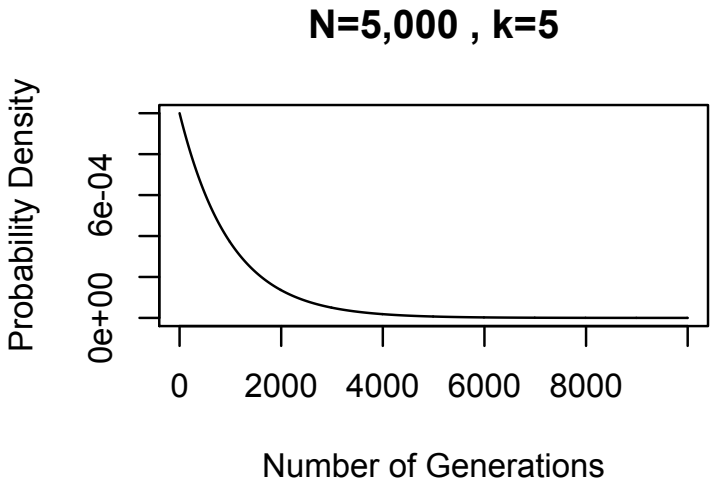
Coalescent review

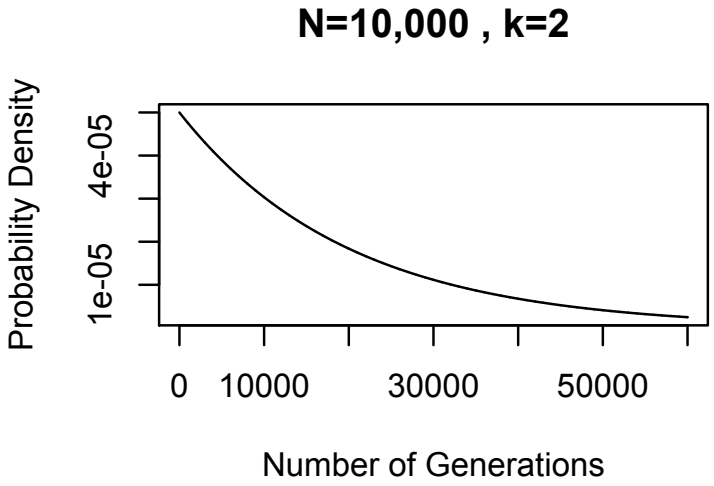
- Recall several important facts from Peter's lecture:
  - Kingman's coalescent: For a sample of k lineages, the distribution of the number of generations until two lineages coalesce is **exponential with rate** $\binom{k}{2}\frac{1}{2N}$
  - k=2: rate $= \frac{1}{2N}$ and mean time to coalescence is $2N$
  - k=5: rate $= \frac{10}{2N}$ and mean time to coalescence is $\frac{2N}{10}$
  - Larger N means that:
  - Larger k means that:
  - 

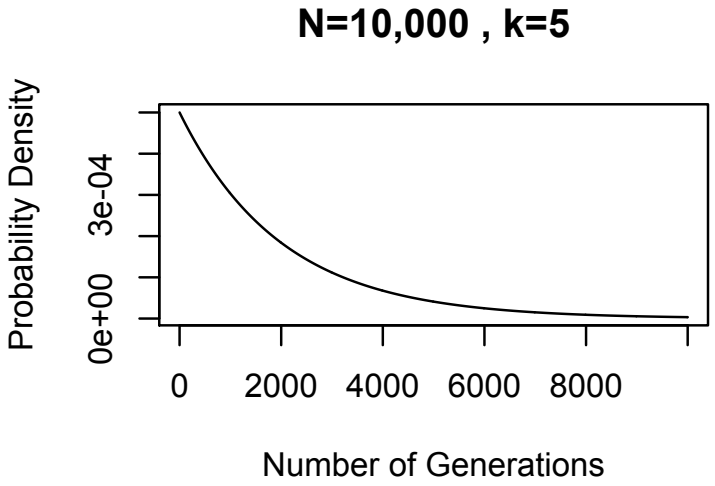
Coalescent review

- What does the exponential distribution look like?







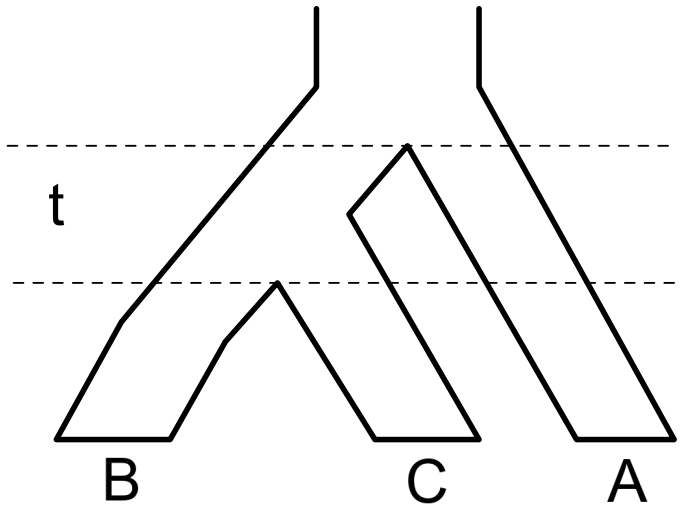


Coalescent units

- Define a common unit of time: coalescent unit, $t = \frac{u}{2N}$
- Examples:
  - $k = 2$ — exponential distribution with rate 1 and mean 1
  - $k = 5$ — exponential distribution with rate 10 and mean 0.1
- t "large" is now relative to population size, but the trends are the same:
  - Longer times lead to a higher probability of coalescence having occurred.
  - Coalescent events happen more quickly when the population size is smaller.
  - Coalescent events happen more quickly when the sample size is larger.
- Now we're ready to think about species trees!

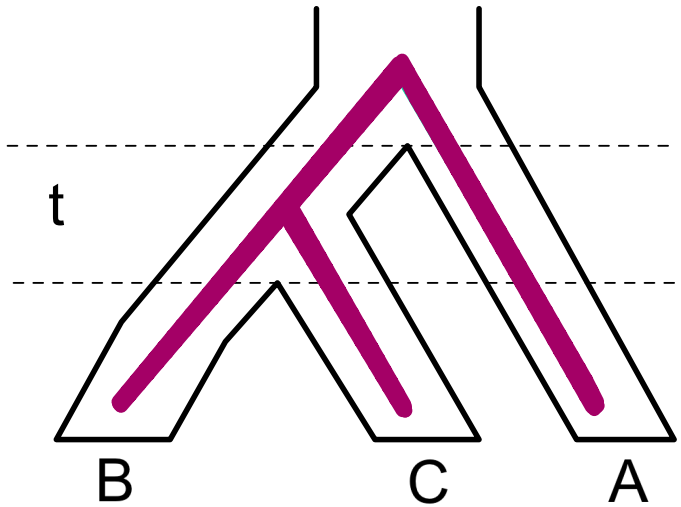
Phylogenetic coalescent model

- Species tree: phylogeny that displays a sequence of speciation events
- Gene tree: phylogenetic history for an individual gene, that evolves "within" the speciation process



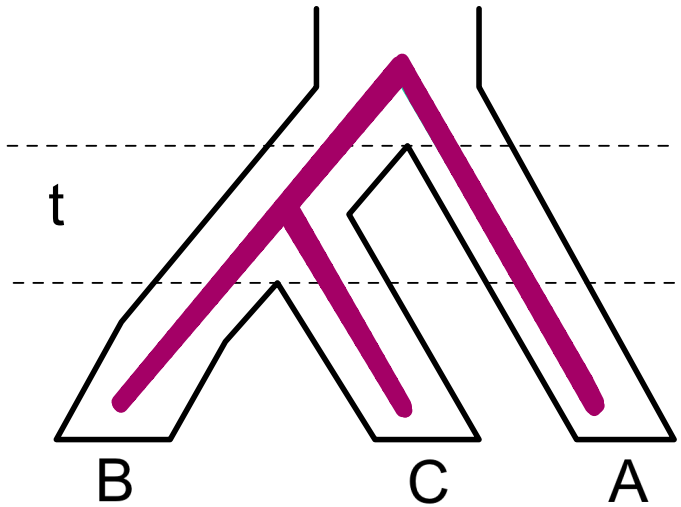
Phylogenetic coalescent model

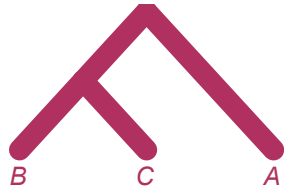
- Species tree: phylogeny that displays a sequence of speciation events
- Gene tree: phylogenetic history for an individual gene, that evolves "within" the speciation process



Phylogenetic coalescent model

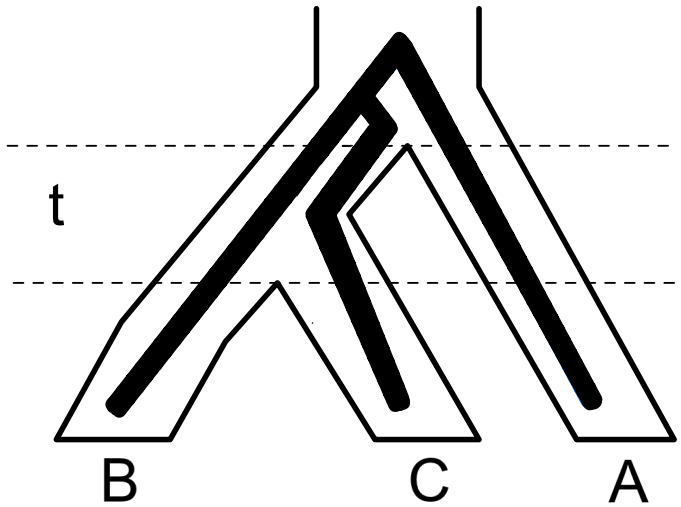
- Species tree: phylogeny that displays a sequence of speciation events
- Gene tree: phylogenetic history for an individual gene, that evolves "within" the speciation process

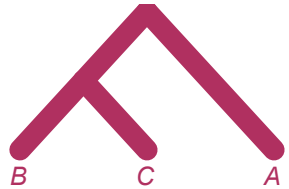


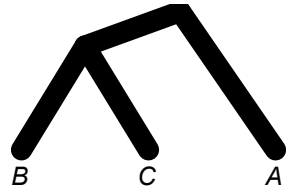


Phylogenetic coalescent model

- Species tree: phylogeny that displays a sequence of speciation events
- Gene tree: phylogenetic history for an individual gene, that evolves "within" the speciation process

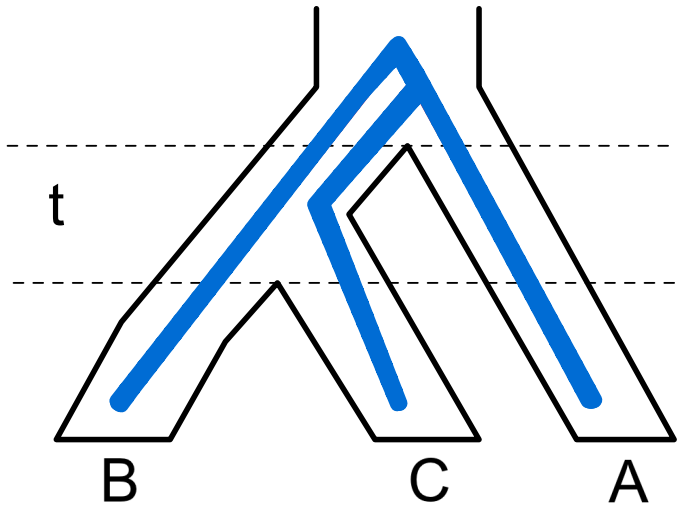


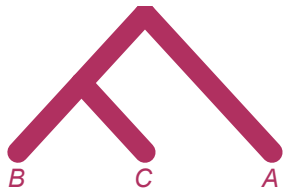


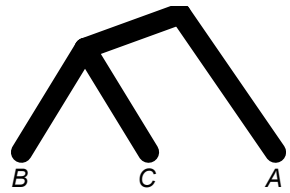


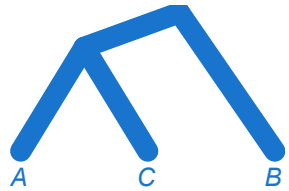
Phylogenetic coalescent model

- Species tree: phylogeny that displays a sequence of speciation events
- Gene tree: phylogenetic history for an individual gene, that evolves "within" the speciation process



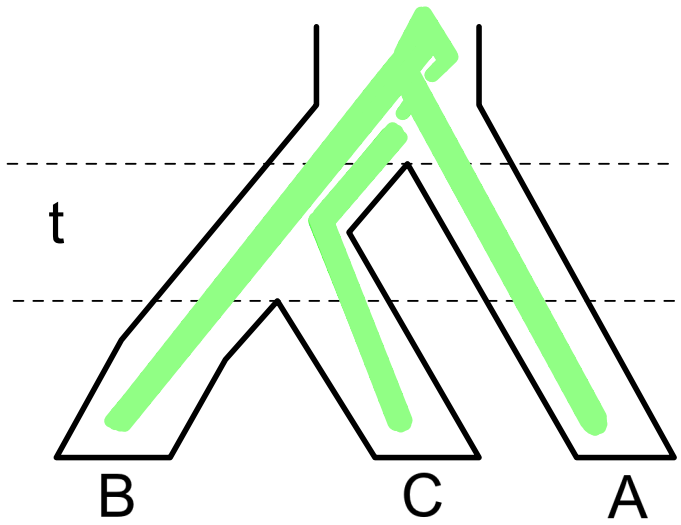


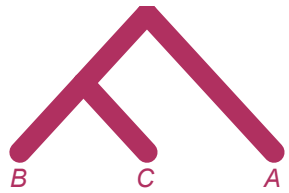


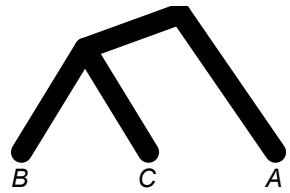


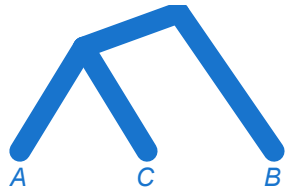
Phylogenetic coalescent model

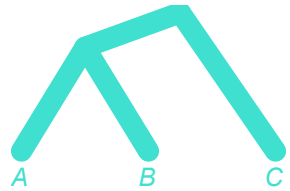
- Species tree: phylogeny that displays a sequence of speciation events
- Gene tree: phylogenetic history for an individual gene, that evolves "within" the speciation process





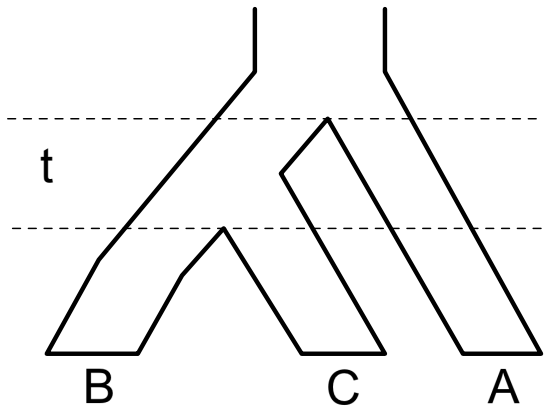






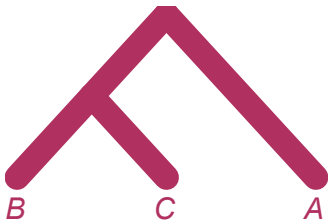
- Let's use what we've learned about the coalescent process to compute some probabilities

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

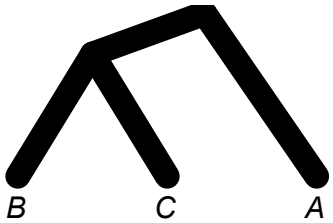


- **Example:** 1.2 coalescent units for an organism with population size $N = 10,000$ and a generation time of 3 years $= 1.2 \times 20,000 \times 3 = 72,000$ years

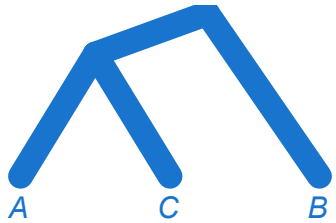
Probabilities of each gene tree history are shown below them
t = length of interval between speciation events



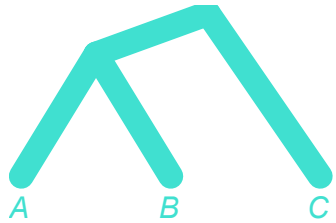
$1 - e^{-t}$



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



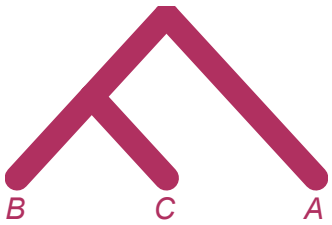
$\frac{1}{3}e^{-t}$



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

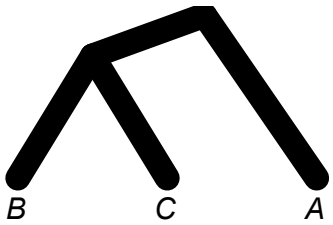
Phylogenetic coalescent model

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



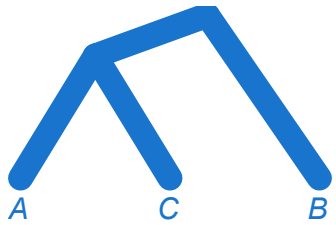
$1 - e^{-t}$

0.63



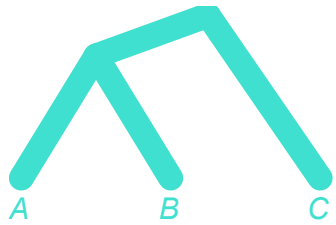
$\frac{1}{3}e^{-t}$

0.12



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

0.12

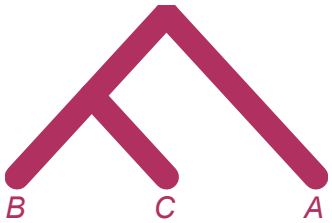


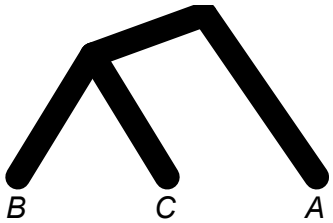
$\frac{1}{3}e^{-t}$

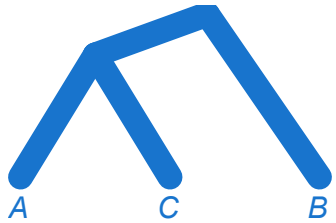
0.12

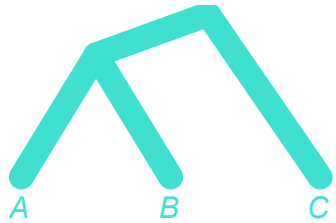
Phylogenetic coalescent model

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





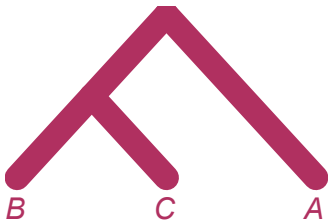


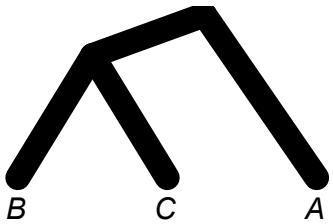


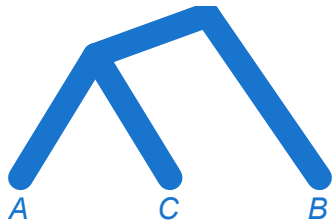
$1 - e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$
0.63	0.12	0.12	0.12
0.40	0.20	0.20	0.20

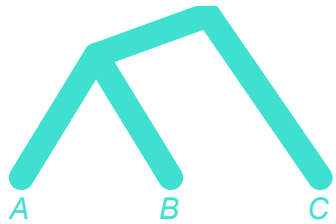
Phylogenetic coalescent model

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





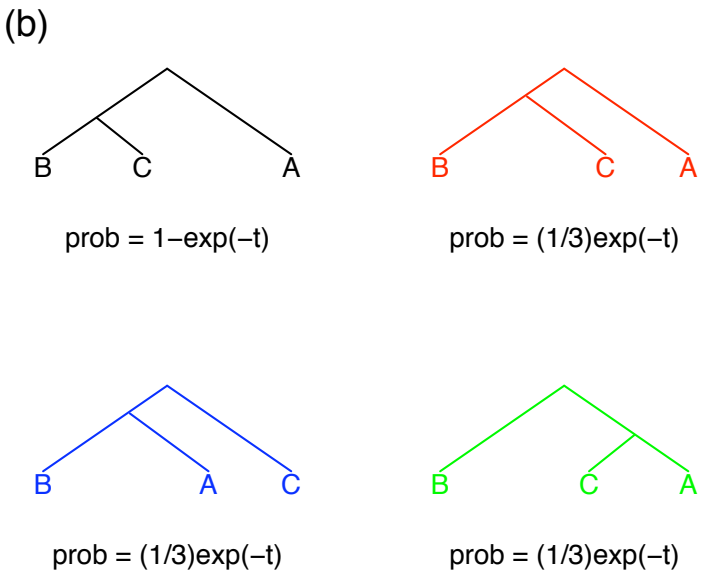


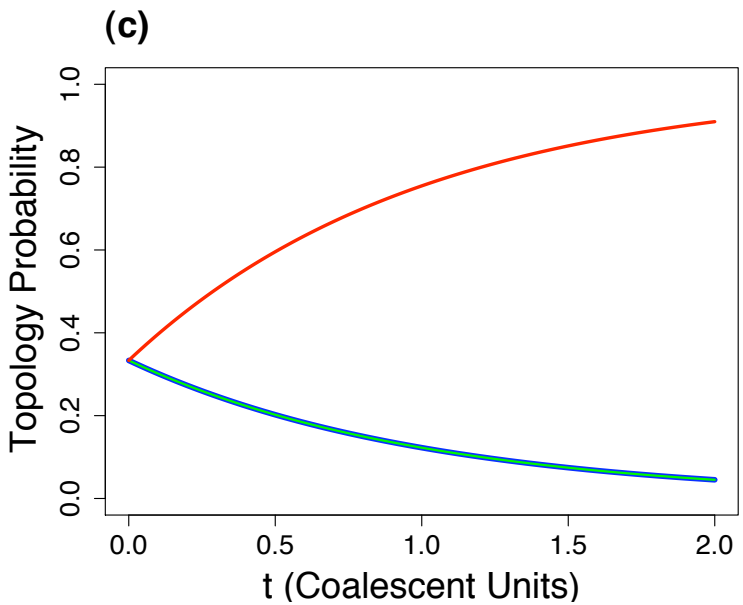


$1 - e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$	$\frac{1}{3}e^{-t}$
0.63	0.12	0.12	0.12
0.40	0.20	0.20	0.20
0.85	0.05	0.05	0.05

Effect of speciation time

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





Assumptions of the phylogenetic coalescent model

- What did we assume in carrying out these computations?
  - Events that occur in one population are independent of what happens in other populations within the phylogeny.
  - More specifically, given the number of lineages entering and leaving a population, coalescent events within populations are independent of other populations.
  - It is also important to recall an assumption we "inherit" from our population genetics model: all pairs of lineages are equally likely to coalesce within a population.
  - No gene flow occurs following speciation.
  - No other evolutionary processes (e.g., horizontal gene flow, duplication, . . .) have led to incongruence between gene trees and the species tree.

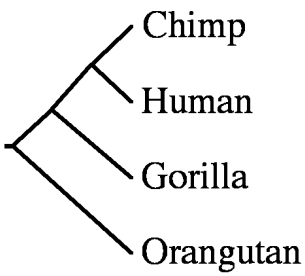
Summary of the three-taxon case

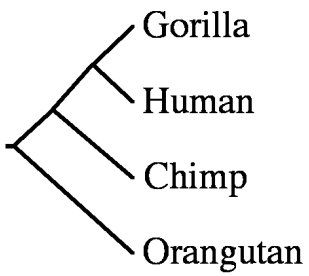
- What have we learned from considering 3 taxa?
  - Gene tree with topology that matches the species tree occurs with probability at least as large as the other two trees
  - The other two trees are expected to occur in equal frequency
  - Shorter intervals between speciation events lead to more disagreement between gene trees and species trees

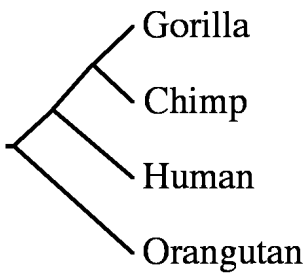
Application 1: Goodness of fit to empirical data

- Motivation: Paper by Ebersberger et al. 2007. *Mol. Biol. Evol.* 24:2266-2276
- Examined 23,210 distinct alignments for 5 primate taxa: Human, Chimp, Gorilla, Orangutan, Rhesus
- Looked at distribution of gene trees among these taxa - observed strongly supported incongruence only among the Human-Chimp-Gorilla clade.

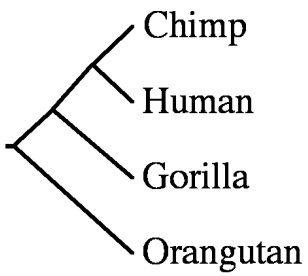
Application 1: Goodness of fit to empirical data



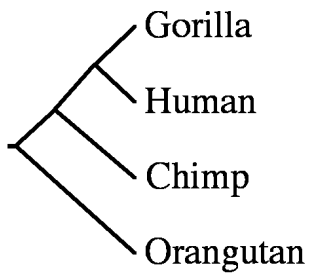




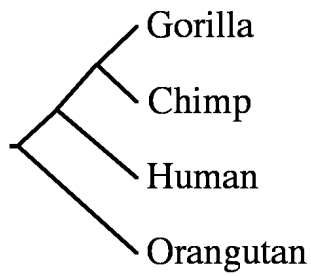
Application 1: Goodness of fit to empirical data



76.6%



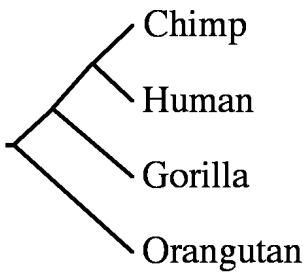
11.4%

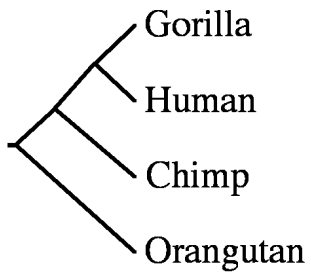


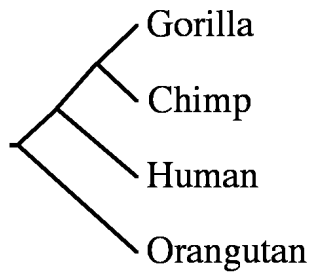
11.5%

Observed proportions of each gene tree among ML phylogenies

Application 1: Goodness of fit to empirical data







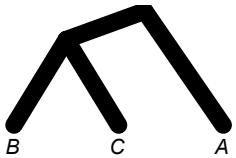
76.6%	11.4%	11.5%
79.1%	9.9%	9.9%

Observed proportions of each gene tree among ML phylogenies

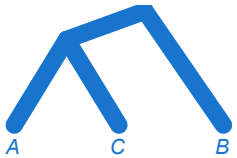
Predicted proportions using parameters from Rannala & Yang, 2003.

Application 2: Branch length estimation

- Suppose we are given a sample of gene trees, i.e.,



70 genes



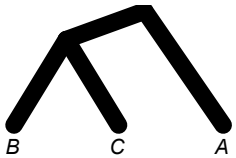
15 genes

15 genes

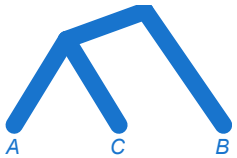
- What do the gene trees tell us?

Application 2: Branch length estimation

- Suppose we are given a sample of gene trees, i.e.,



70 genes

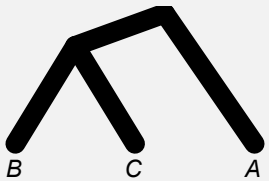


15 genes

15 genes

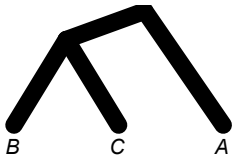
- What do the gene trees tell us?

The species tree

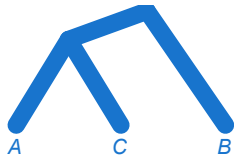


Application 2: Branch length estimation

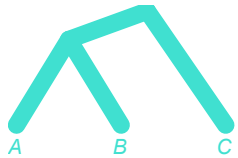
- Suppose we are given a sample of gene trees, i.e.,



70 genes



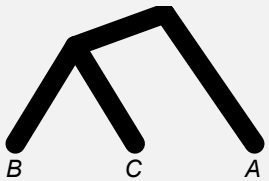
15 genes



15 genes

- What do the gene trees tell us?

The species tree



The branch length t:

Set $0.7 = 1 - \frac{2}{3}e^{-t}$
and solve for t

$t = 0.7985$

How general is this result?



J. Math. Biol. (2011) 62:833–862
DOI 10.1007/s00285-010-0355-7

Mathematical Biology

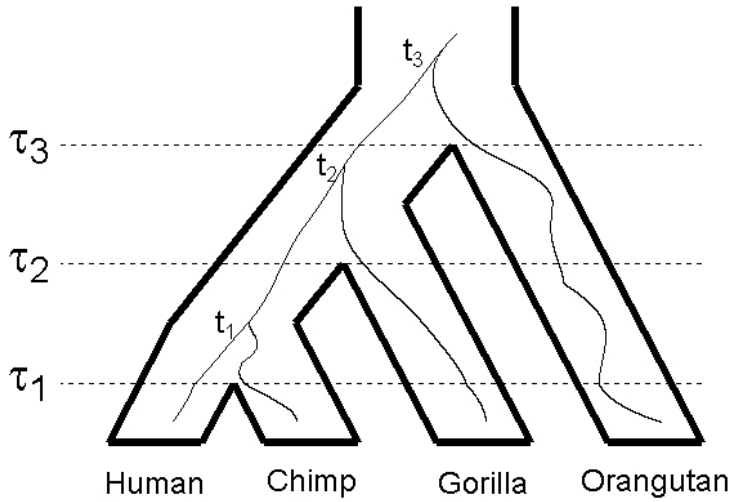
Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent

Elizabeth S. Allman · James H. Degnan · John A. Rhodes

- Four taxa: the distribution of unrooted gene trees determines the unrooted species tree and branch lengths
- Five or more taxa: the distribution of unrooted gene trees determines the rooted species tree and branch lengths.

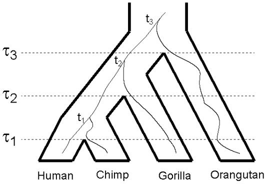
A slightly larger case

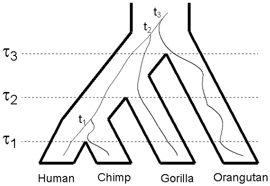
- Consider 4 taxa – the human-chimp-gorilla problem

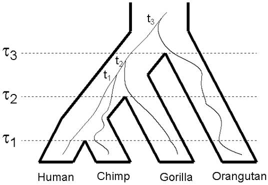


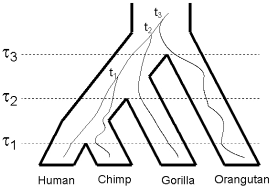
Coalescent histories for the 4-taxon example

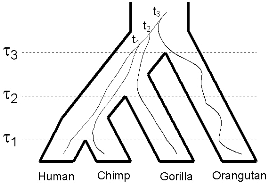
- There are 5 possible histories for this example:











Enumerating Histories

TABLE 3. The number of valid coalescent histories when the gene tree and species tree have the same topology. The number of histories is also the number of terms in the outer sum in equation (12).

	Number of histories		
Taxa	Asymmetric trees	Symmetric trees	Number of topologies
4	5	4	15
5	14	10	105
6	42	25	945
7	132	65	10,395
8	429	169	135,135
9	1430	481	2,027,025
10	4862	1369	34,459,425
12	58,786	11,236	13,749,310,575
16	9,694,845	1,020,100	6.190×10^{15}
20	1,767,263,190	100,360,324	8.201×10^{21}

Degnan and Salter, *Evolution*, 2005

Computing the Topology Distribution by Enumerating Histories

- In the general case, we have the following:

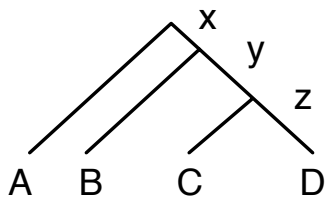
  The probability of a gene tree g given the species tree $\mathcal{S}$ is given by

$$P\{G = g|\mathcal{S}\} = \sum_{histories} P\{G = g, history|\mathcal{S}\}$$

- Implemented in the software COAL (Degnan and Salter, *Evolution*, 2005)
- A more efficient method has been proposed (Wu, *Evolution*, 2012)

Gene tree distribution for four taxa

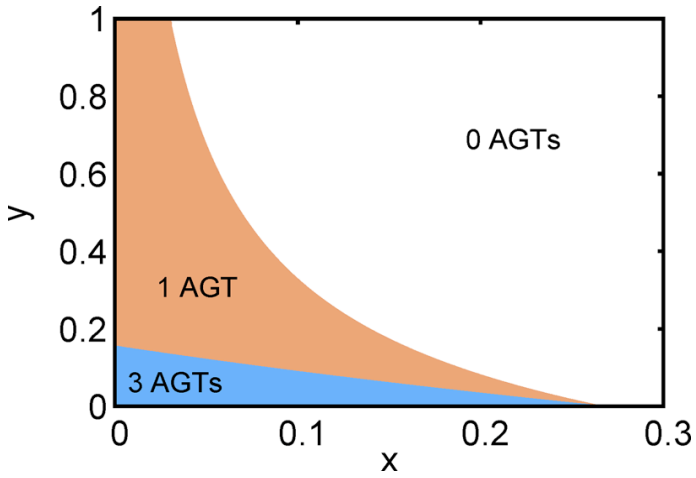
- In the three-taxon case, the gene tree with the highest probability has the same topology as the species tree
- Question: Must the distribution always look this way?
- Examine the entire distribution for four taxa – only 15 gene trees are possible
- For the species tree:



look at probabilities of all 15 gene tree topologies for values of x, y, and z

- https://lkubatko.shinyapps.io/GeneTreeProbs/

Gene tree distribution for four taxa



- 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

Can we use gene trees to estimate the species trees?

- Two problems with using gene trees directly for inference:
- We don't observe gene trees directly

  Rather, we observe sequence data for each gene and need to estimate the gene trees

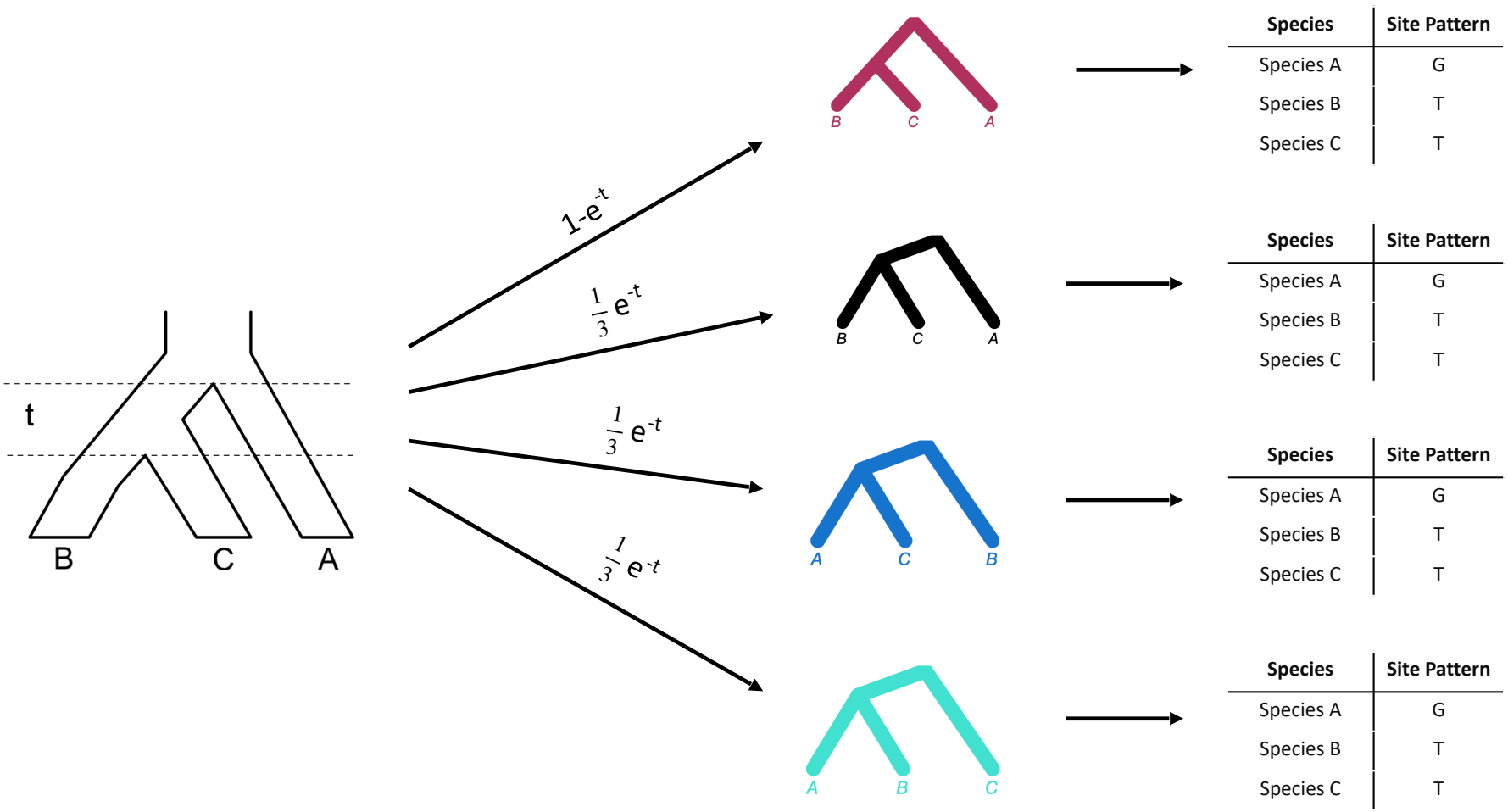
- Sampling error in the gene tree proportions would complicate inference

  For example, if the branch length t is long enough, we would only observe gene trees that matched the species tree ... and then how would we estimate t?

What about mutation?

- What about mutation? How does this affect data analysis?
- The coalescent gives a model for determining gene tree probabilities for each gene.
- View DNA sequence data as the results of a two-stage process:
  - Coalescent process generates a gene tree topology.
  - Given this gene tree topology, DNA sequences evolve along the tree.
- Go back to our three-taxon example to get some intuition about the model

Sequence data



Species	Site Pattern
Species A	G
Species B	T
Species C	T

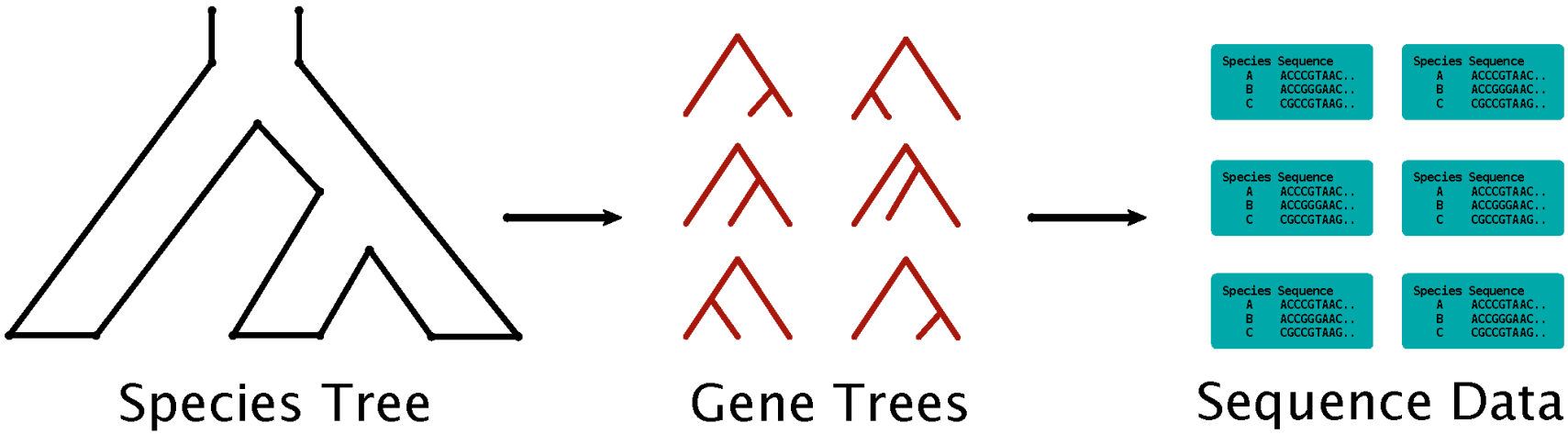
Species	Site Pattern
Species A	G
Species B	T
Species C	T

Species	Site Pattern
Species A	G
Species B	T
Species C	T

Species	Site Pattern
Species A	G
Species B	T
Species C	T

https://lkubatko.shinyapps.io/SitePatternsProbs/

The multispecies coalescent (MSC) model



Question: How do we estimate a species tree under this model that accommodates variation in gene trees?

Given this model, how should inference be carried out?

- As more data (genes) are added, the process of estimating species trees from concatenated data can be statistically inconsistent
- May fail to converge to any single tree topology if there are many equally likely trees.
- May converge to the wrong tree when a gene tree that is topologically incongruent with the species tree has the highest probability.
- The bootstrap may be positively misleading – show strong support for an incorrect clade

  Important note: This is NOT a failing of the bootstrap methodology; the observed "poor" performance is due to the use of an incorrect model (concatenation)

Kubatko and Degnan, 2007; Roch and Steel, 2015

Is there a better way to estimate species phylogenies?

Explicitly model the coalescent process!

BUT, this is hard! Why?

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

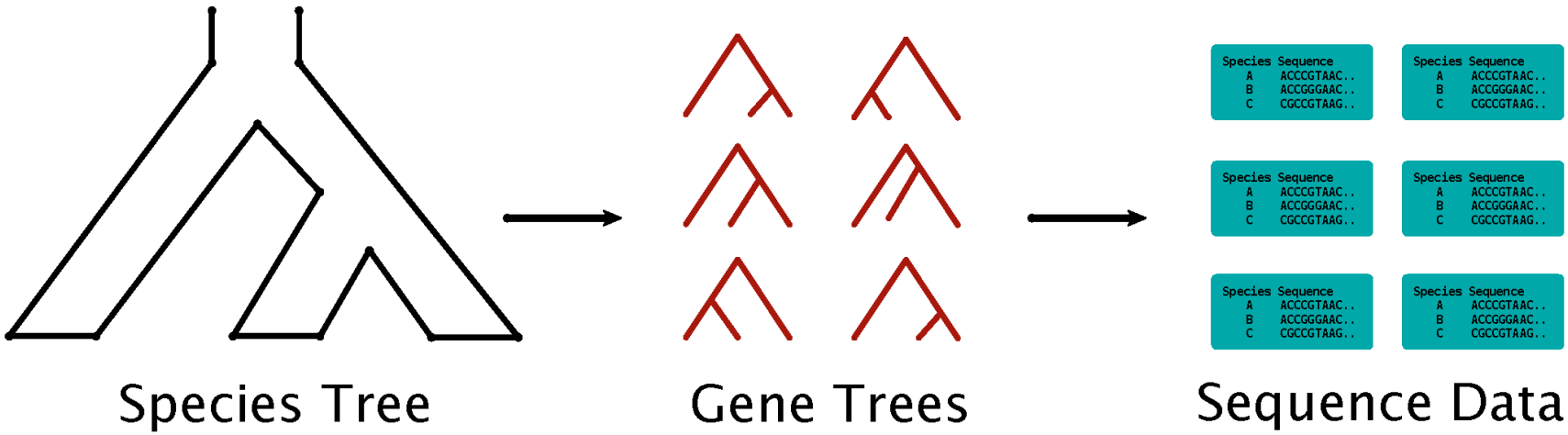
- Recall the Felsenstein equation from Peter's lecture, except that now we replace θ with S, the species tree. Use this to form the species tree likelihood for a multi-locus data set:

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

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

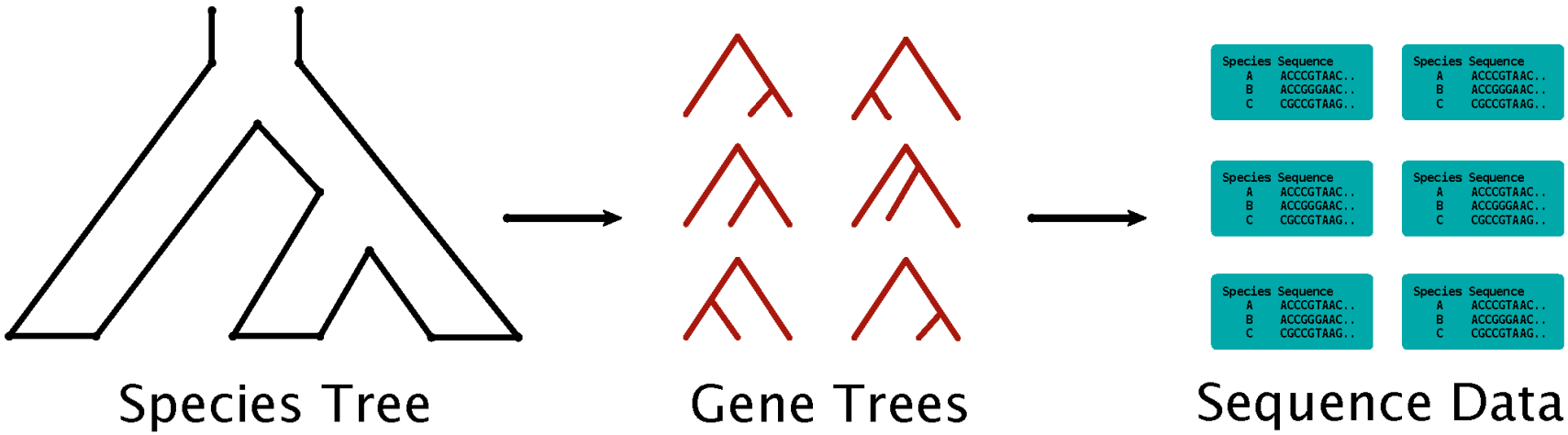
- This likelihood is difficult to evaluate directly, because of the dimension of the inner sum (which is really an integral) [recall Peter's "galaxy slide"]

Inference option 1: Summary statistics methods



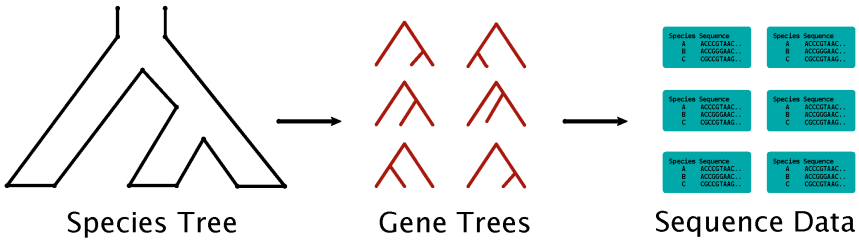
- Summary statistics methods: Start with estimated gene trees
  - Using estimated branch lengths:
    - ⋆ STEM (Kubatko et al. 2009)
    - ⋆ STEAC (Liu et al. 2009)
  - Using topology information only:
    - ⋆ STAR (Liu et al. 2009)
    - ⋆ Minimize Deep Coalescences (PhyloNet; Than & Nakhleh 2009)
    - ⋆ 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/STARBEAST2, BPP, SNAPP

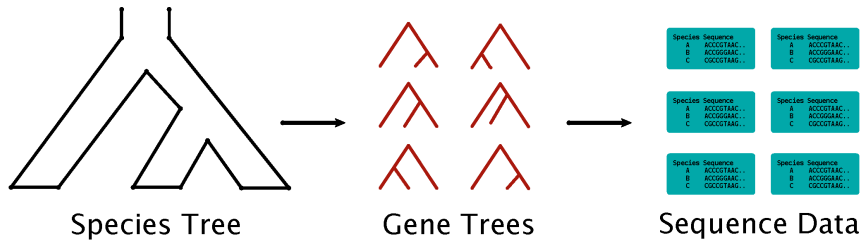
- Model the entire process of data generation
- Goal of these methods is to estimate the posterior distribution of the gene trees and species tree and associated model parameters



- BEST, *BEAST/STARBEAST2, 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

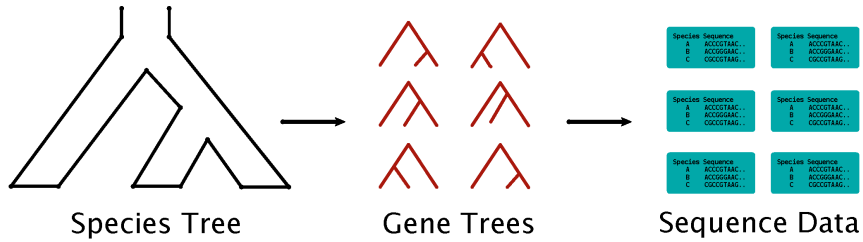
- Model the entire process of data generation
- Avoid computing the likelihood by using algebraic structure in the distribution of site pattern probabilities under the model



- SVDQuartets is implemented in PAUP*
- SVDQuartets will be discussed in detail in lab

Full data methods III: Composite likelihood

- Model the entire process of data generation
- Approximate the likelihood by multiplying likelihoods of 3-tip or 4-tip trees



- For branch length estimation on any fixed species tree: qAge, implemented in PAUP* (lab)
- For tree estimation: part of Kevin's package PhyNEST for estimating phylogenetic networks (`https://github.com/sungsik-kong/PhyNEST.jl`)

- 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/STARBEAST2, BPP, and SNAPP utilize a Bayesian framework and involve MCMC

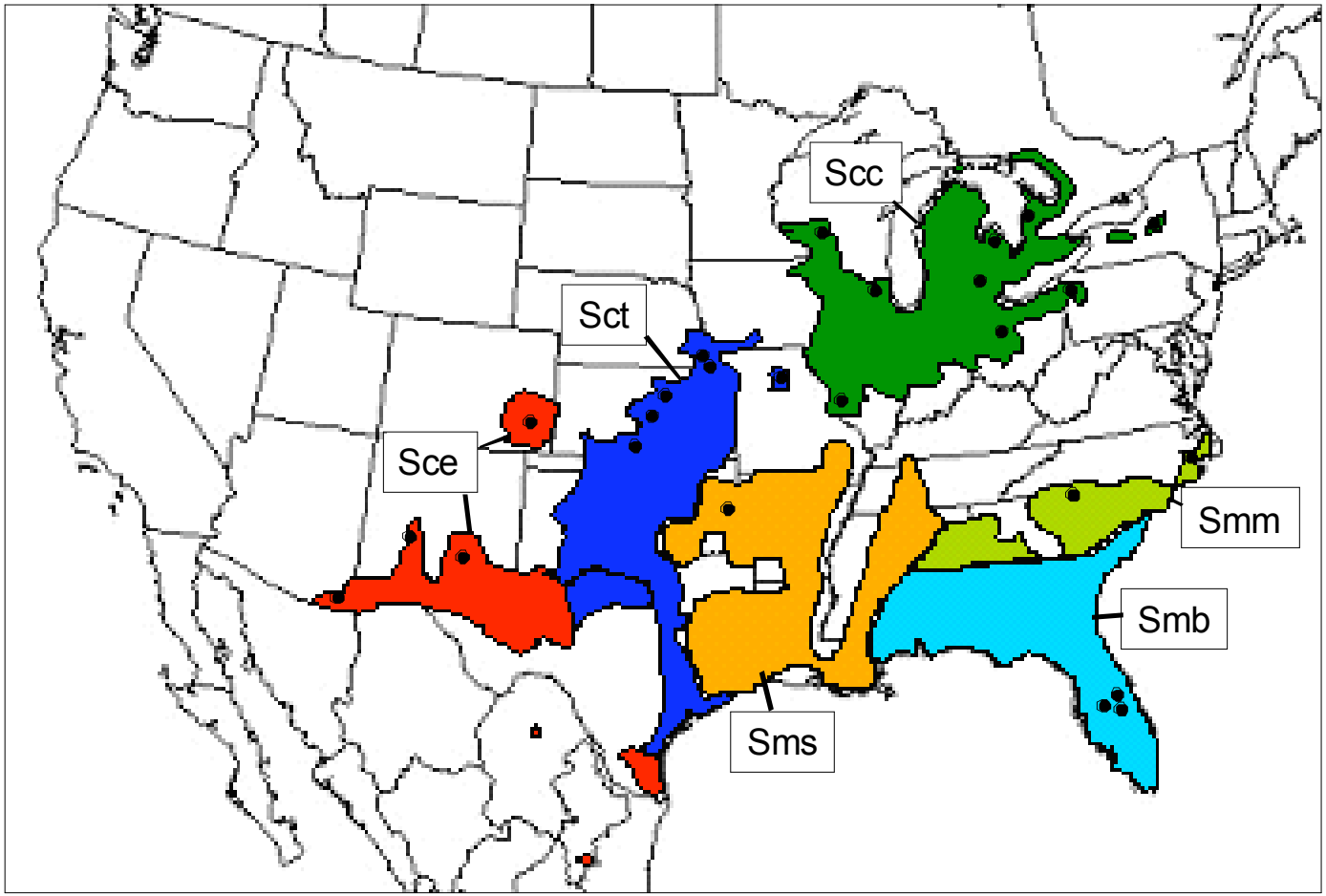
- Comparison of approaches:
  - Summary statistics methods
    - ⋆ Advantage: Quick
    - ⋆ Disadvantage: Ignore information in the data
    - ⋆ Most current implementations do not easily allow assessment of uncertainty (but bootstrap can be used, at the expense of computational efficiency)
  - Full data methods
    - ⋆ Advantage: Fully model-based framework
    - ⋆ Disadvantage: Computationally intensive, sometimes prohibitively so
    - ⋆ BEST, *BEAST/STARBEAST2, BPP, and SNAPP utilize a Bayesian framework and involve MCMC
- Ugh! Do we really need the coalescent? Why not just **concatenate**????
  - Well, the model is incorrect, and alternatives are available with a little effort
  - Also: the model matters for quantification of uncertainty and branch length estimation

Example 1: Sistrurus rattlesnakes

- North American Rattlesnakes - Joint work with Dr. Lisle Gibbs (EEOB at OSU)
- Of interest evolutionarily because of the diversity of venoms present in the various species and subspecies.
- Of conservation interest because population sizes in the eastern subspecies are very small.

[Pictures by Jimmy Chiucchi and Brian Fedorko]

Geographic Distribution of Snake Populations

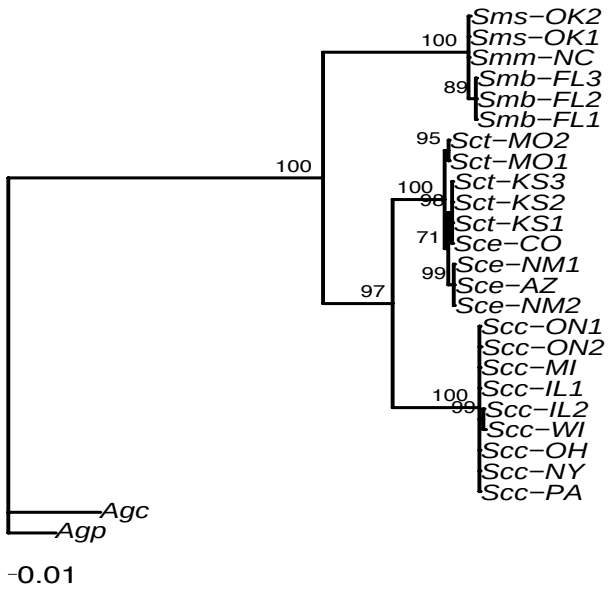


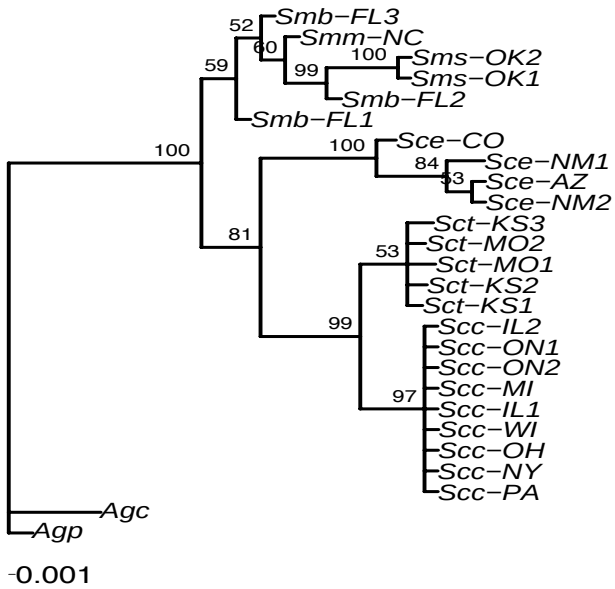
- Data: 7 (sub)species, 26 individuals (52 sequences), 19 genes

Species	Location	No. of individuals per gene
S. catenatus catenatus	Eastern U.S. and Canada	9
S. c. edwardsii	Western U.S.	4
S. c. tergeminus	Western and Central U.S.	5
S. miliarius miliarius	Southeastern U.S.	1
S. m. barbouri	Southeastern U.S.	3
S. m. streckerii	Southeastern U.S.	2
Agkistrodon sp. (outgroup)	U.S.	2

Individual Gene Tree Estimates

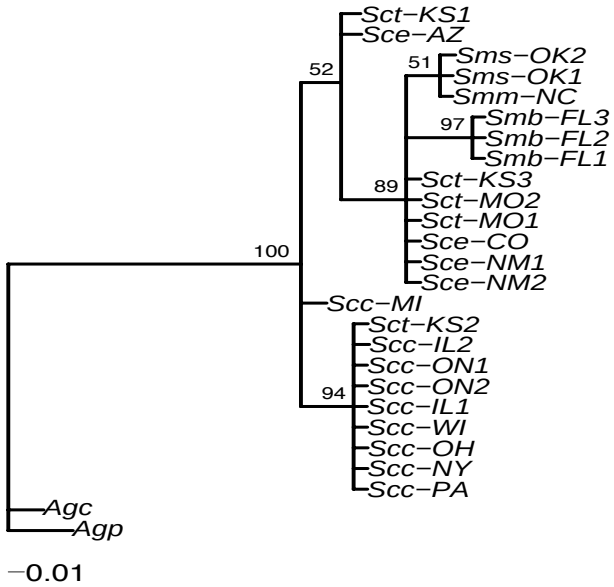
Some are very informative:

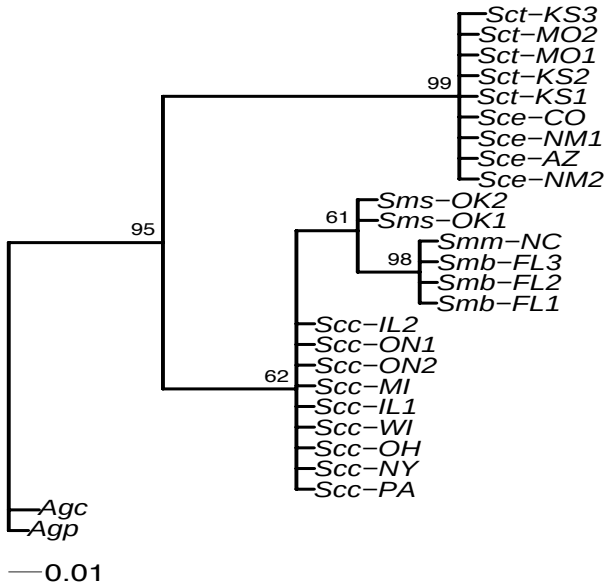




Individual Gene Tree Estimates

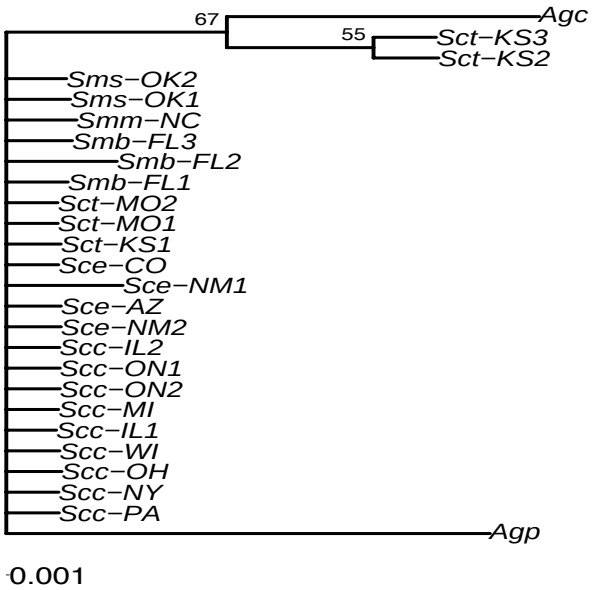
Some are a little informative:

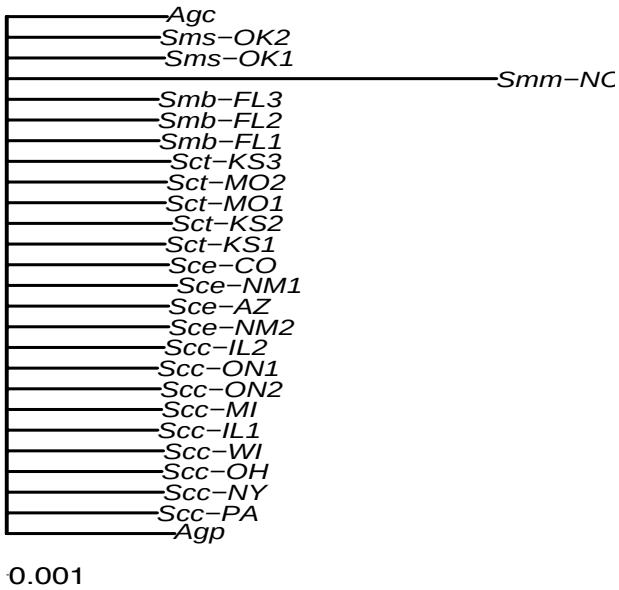




Individual Gene Tree Estimates

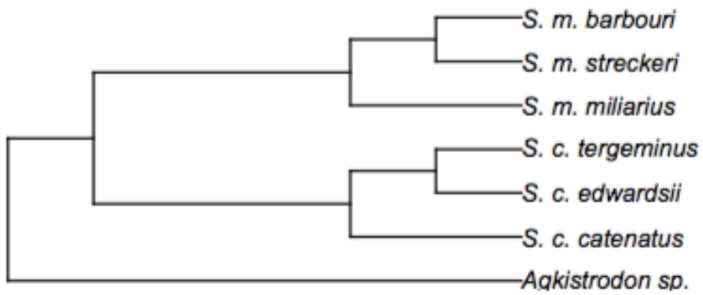
And then there are others .....





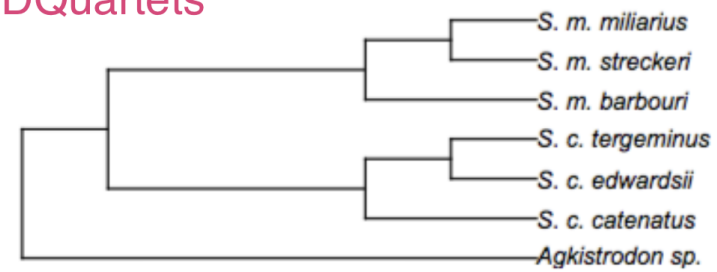
Example 1: Sistrurus rattlesnakes

STEM, STEAC

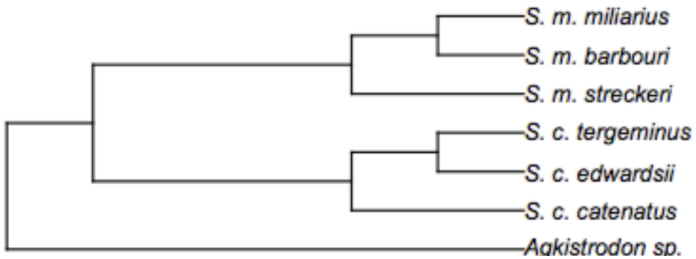


BEAST (concatenated data), *BEAST

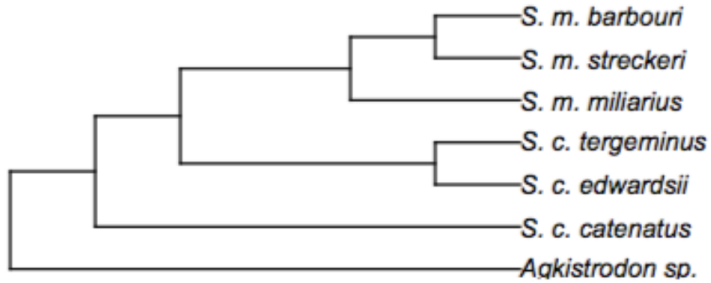
SVDQuartets



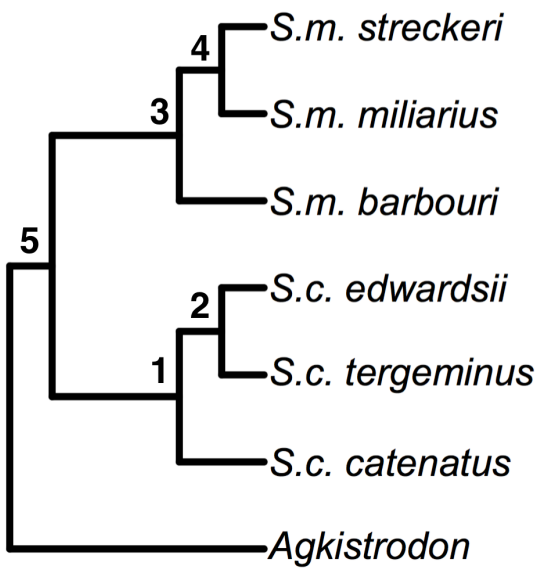
BEST, Parsimony & MrBayes (concatenated data), Astral



PhyloNet, STAR



Example 1: Sistrurus rattlesnakes



Node	1	2	3	4	5
*BEAST	100	100	100	46*	100
BPP	100	99	100	33*	100
SVDQ	93	100	100	46	100

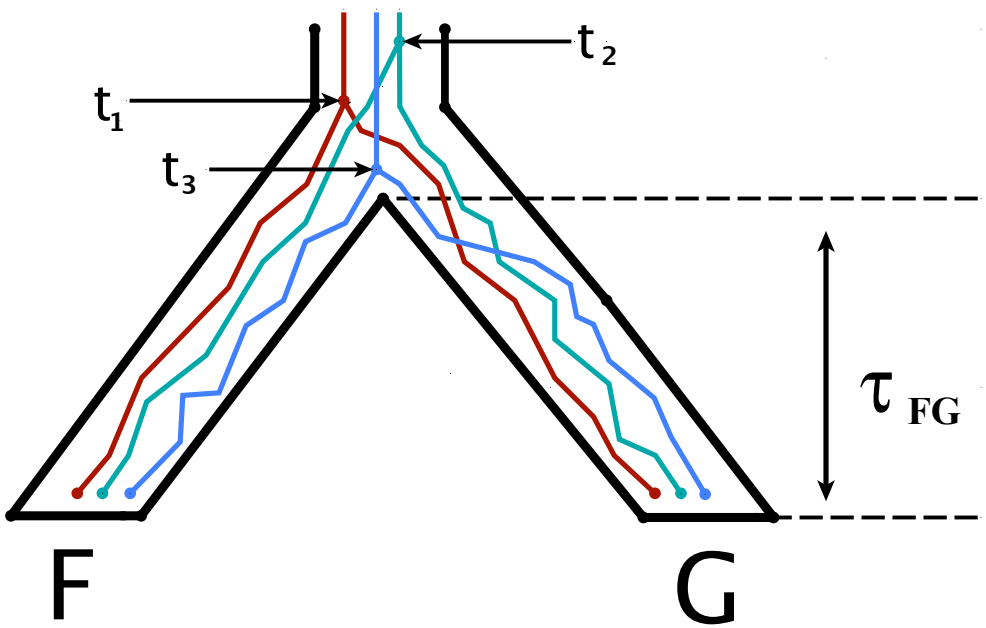
* = This clade was not in the maximum clade credibility (*S. m. miliarius* and *S. m. barbouri* received 48.78% posterior probability with *BEAST and 59% posterior probability with BPP)

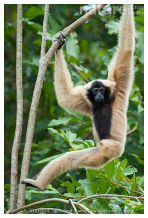
- How does concatenation do?
  - Tree agrees with estimated species tree (both with BEAST and with ML in PAUP*)
    - BEAST: posterior probability on *miliarius* clade: 73%
  - Speciation time estimates are severely biased:

Dated node	Divergence estimates from concatenated gene tree (Ma)[a]	Divergence estimates from species tree (Ma)[a]	Percent difference[b] (%)
(Scc (Sce,Sct)) vs. (Sms(Smb, Smm))	9.45 (9.14, 10.24)	10.04 (9.25, 12.97)	+6
Scc vs. (Sce, Sct)	6.06 (5.22, 7.02)	2.92 (1.58,4.90)	−52
Sce vs. Sct	2.41 (2.01, 2.88)	0.47 (0.24, 0.86)	−79
Smb vs. (Smb, Sms)	1.98 (1.60, 2.47)	0.77 (0.44,1.31)	−62
Sms vs. Smm	1.60 (1.23, 2.06)	0.49 (0.25, 0.92)	−69

Example 1: Sistrurus rattlesnakes

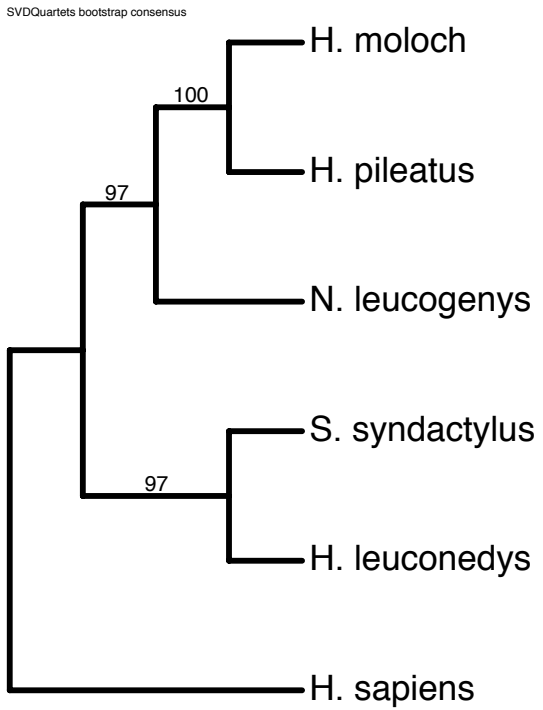
- Why are speciation times biased?
  - We estimate different quantities when using a gene tree vs. species tree analysis!



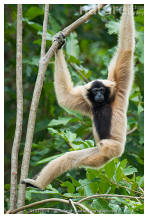


Application to empirical data for gibbons

- Data for 5 species of gibbons (Carbone et al., 2014; Veeramah et al., 2015; Shi and Yang, 2018)
- 11,323 coding loci with 200 sites each = 2,264,400 sites
- 2-4 individuals sampled per species
- 752 total quartets possible with one member from each species



- Computation time for SVDQuartets: 14.46 seconds



Application to empirical data for gibbons

- Performance for branch length estimation (coalescent units)

Table S3: Comparison of the node age estimates for BPP and MAP_{CL} in mutation units.

Age of MRCA of ...	BPP mean	BPP 95% HPD interval	MAP_{CL} mean	MAP_{CL} 95% CI
HmHp	0.00088	(0.00084, 0.00093)	0.00085	(0.00081, 0.00090)
BS	0.00164	(0.00154, 0.00175)	0.00274	(0.00270, 0.00279)
NBS	0.00264	(0.00248, 0.00280)	0.00279	(0.00275, 0.00284)
NBSHmHp	0.00306	(0.00302, 0.00311)	0.00290	(0.00286, 0.00294)
ONBSHmHp	0.01148	(0.01127, 0.01169)	0.01428	(0.01417, 0.01438)

Species Tree Inference Summary – Comparison of Methods

Software	Data Type	Measure of Uncertainty	Computation Time	Models Included
*BEAST/ STARBEAST2	multilocus	posterior probability	intermediate; can be run in parallel	coalesent; all reversible substitution models; relaxed clock
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 4 sister taxa
SNAPP	biallelic SNP; AFLP	posterior probability	long; can be run in parallel	coalescent; two-state substitution model; Bayes factor delimitation
ASTRAL	unrooted gene trees	local posterior probability	short given gene trees	no specific model assumed
MP-EST	rooted gene trees	bootstrap	short given gene trees	coalescent model
PhyNEST	multilocus; SNP	bootstrap	intermediate	coalescent; JC69 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.

Key points to take away ....

- Gene trees and species trees are different – both conceptually and physically
- The coalescent model predicts a distribution of gene trees for a given species tree
  - 3 taxa:
  - Empirical data often fit this predicted distribution
- Three reasons a species tree analysis is preferred over concatenation: