

# 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

August 5, 2019

## 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:

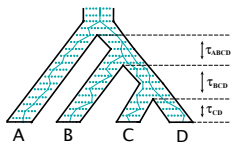
1. Model what happens within each population

→ *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:



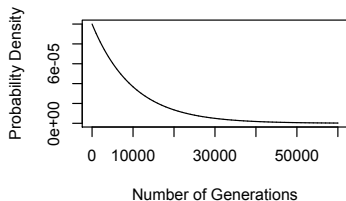
- Recall several important facts from Peter's lecture:
  - ▶ **Kingman's coalescent:** For a sample of  $k$  lineages, the distribution of the number of generations until two lineages coalesce is **exponential with rate**  $\binom{k}{2} \frac{1}{2N}$
  - ▶  $k=2$ : rate =  $\frac{1}{2N}$  and mean time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



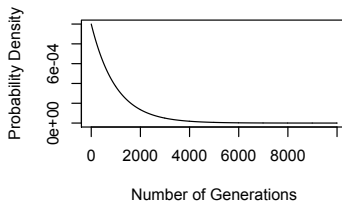
▶

- What does the exponential distribution look like?

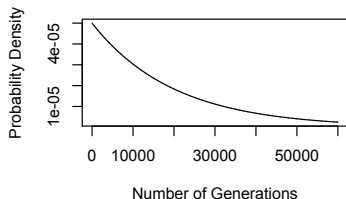
**N=5,000 , k=2**



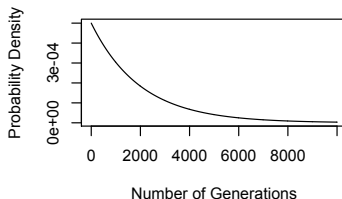
**N=5,000 , k=5**



**N=10,000 , k=2**



**N=10,000 , k=5**

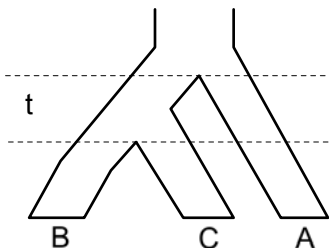


## Coalescent units

- 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

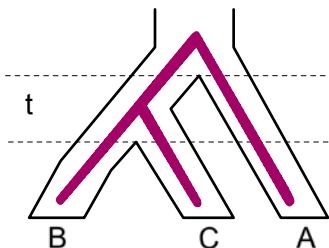
- **Species tree:** phylogeny that displays a sequence of speciation events
- **Gene tree:** phylogenetic history for an individual gene, that evolves “within” the speciation process





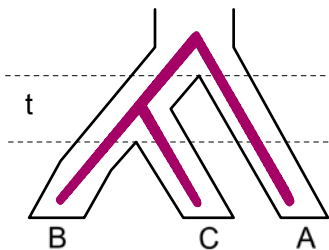
## Phylogenetic coalescent model

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



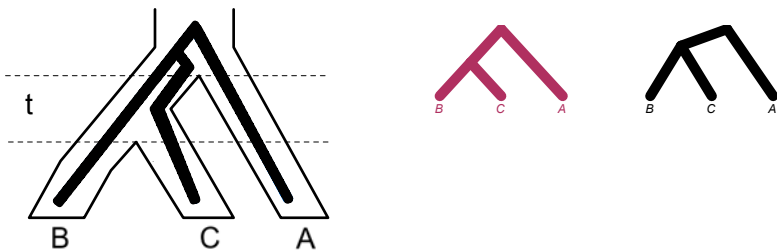
## Phylogenetic coalescent model

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



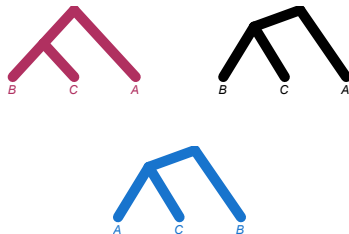
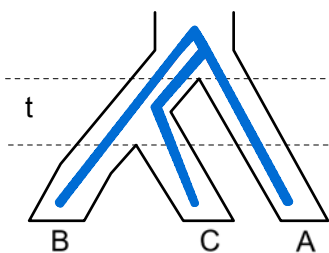
## Phylogenetic coalescent model

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



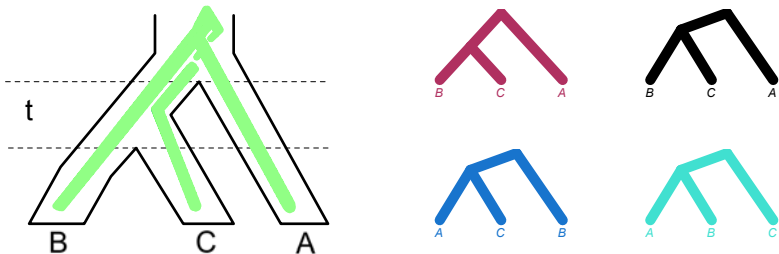
## Phylogenetic coalescent model

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



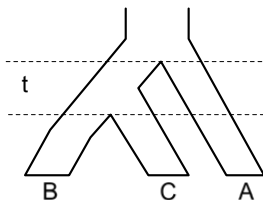
## Phylogenetic coalescent model

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



## Phylogenetic coalescent model

- 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



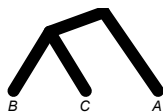
- Example:** 1.2 coalescent units for an organism with population size  $N = 10,000$  and a generation time of 3 years =  $1.2 \times 20,000 \times 3 = 72,000$  years

## Phylogenetic coalescent model

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



$$1 - e^{-t}$$



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



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



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

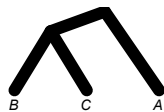
## Phylogenetic coalescent model

$t = \text{length of interval between coalescent events} = 1.0$



$$1 - e^{-t}$$

0.63



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

0.12



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

0.12



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

0.12



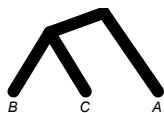
## Phylogenetic coalescent model

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



$$1 - e^{-t}$$

0.63  
0.40



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

0.12  
0.20



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

0.12  
0.20



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

0.12  
0.20

## Phylogenetic coalescent model

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

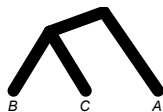


$$1 - e^{-t}$$

0.63

0.40

0.85



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

0.12

0.20

0.05



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

0.12

0.20

0.05



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

0.12

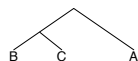
0.20

0.05

## Effect of speciation time

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

(b)



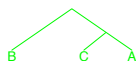
$$\text{prob} = 1 - \exp(-t)$$



$$\text{prob} = (1/3)\exp(-t)$$

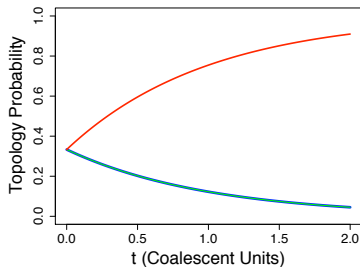


$$\text{prob} = (1/3)\exp(-t)$$



$$\text{prob} = (1/3)\exp(-t)$$

(c)



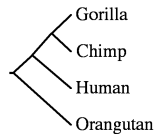
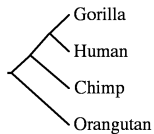
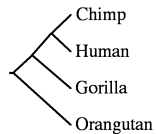
- What did we assume in carrying out these computations?
  - ▶ Events that occur in one population are independent of what happens in other populations within the phylogeny.
  - ▶ More specifically, given the number of lineages entering and leaving a population, coalescent events within populations are 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.

- 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

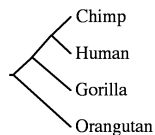
## Application 1: Goodness of fit to empirical data

- **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.

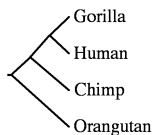
## Application 1: Goodness of fit to empirical data



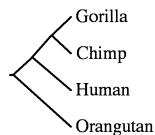
## Application 1: Goodness of fit to empirical data



76.6%



11.4%

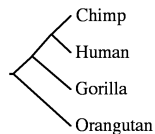


11.5%

Observed proportions of each  
gene tree among ML phylogenies

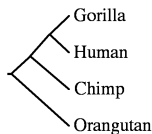


## Application 1: Goodness of fit to empirical data



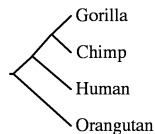
76.6%

79.1%



11.4%

9.9%



11.5%

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 you were given a **sample of gene trees**, i.e.,



70 genes



15 genes

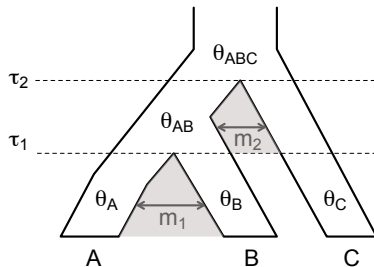


15 genes

- We also know from our earlier work that the probability that the gene tree matches the species tree, say  $p$ , is  $p = 1 - \frac{2}{3}e^{-t}$
- From the data, we estimate that  $p = 0.7$  – use this to estimate  $t$ :

What about gene flow?

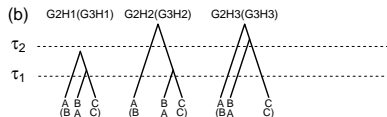
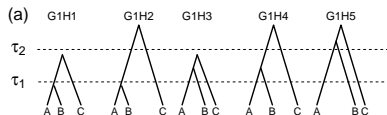
**Question:** What happens to gene tree topology probabilities under a model with gene flow?



Tian and Kubatko, MPE, 2017

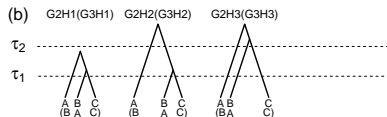
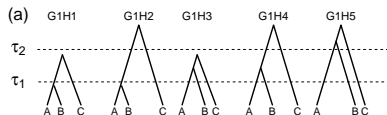
## What about gene flow?

**Complication:** More **histories** are possible, because coalescent events can happen “before” speciation



## What about gene flow?

**Complication:** More **histories** are possible, because coalescent events can happen “before” speciation

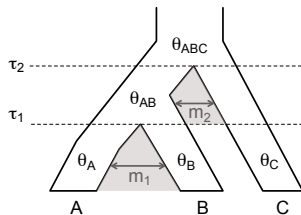


The gene tree that matches the species tree may not have the highest probability!

## Anomalous three-taxon gene trees in the presence of gene flow

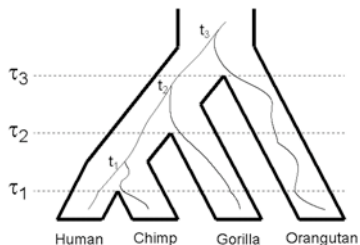
### Recent results (Long and Kubatko, Systematic Biology, 2018):

- When  $\theta_{AB} = \theta_C$ , the gene tree that matches the species tree will have the highest probability (i.e., there are **no anomalous gene trees**)
- When  $\theta_{AB} \neq \theta_C$  and  $m_2 > 0$ , **anomalous gene trees are possible** – the probability of the gene tree matching the species tree could be as low as  $\frac{1}{9}$  (leaving probability  $\frac{4}{9}$  for each of the other two gene trees)
- When  $\theta_{AB} \neq \theta_C$  and there is asymmetric gene flow between populations *AB* and *C*, **anomalous gene trees are possible** – the probability of the gene tree matching the species tree can go to 0 for highly asymmetric rates



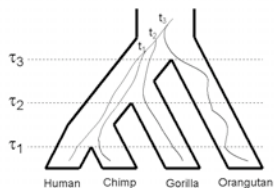
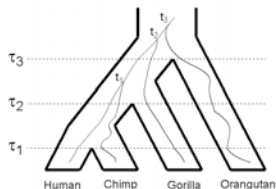
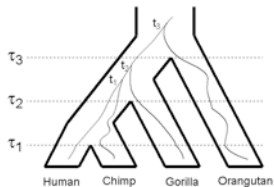
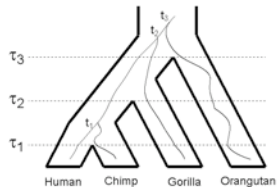
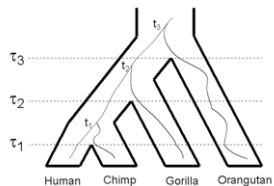
## A slightly larger case – no gene flow

- 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).

Taxa	Number of histories		Number of topologies
	Asymmetric trees	Symmetric trees	
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	$6.190 \times 10^{15}$
20	1,767,263,190	100,360,324	$8.201 \times 10^{21}$

Degnan and Salter, *Evolution*, 2005

- In the general case, we have the following:

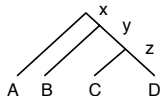
The probability of a gene tree  $g$  given the species tree  $\mathcal{S}$  is given by

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

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

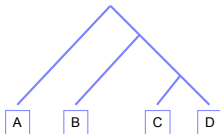
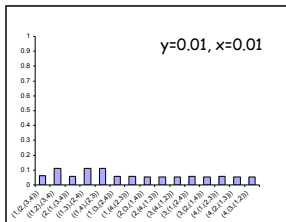
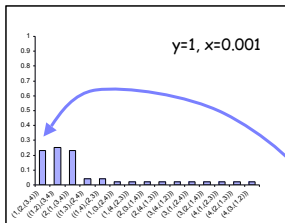
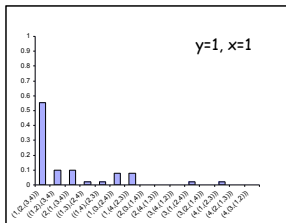
## Gene tree distribution for four taxa

- In the three-taxon case in the absence of gene flow, there are **no anomalous gene trees**
- **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:

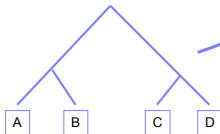
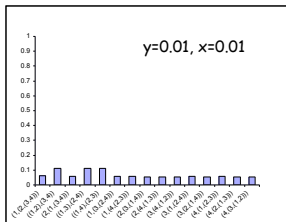
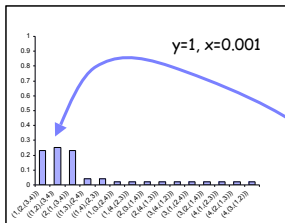
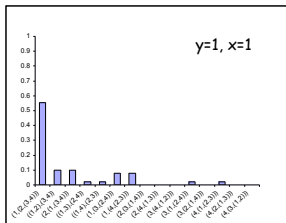


look at probabilities of all 15 gene tree topologies for values of  $x$ ,  $y$ , and  $z$

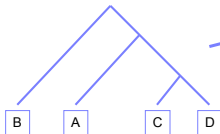
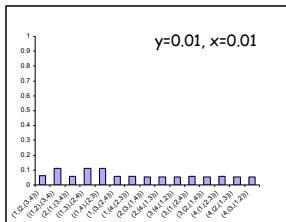
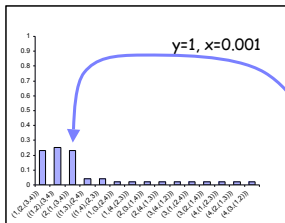
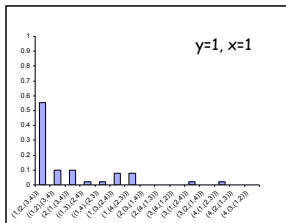
# Gene tree distribution for four taxa



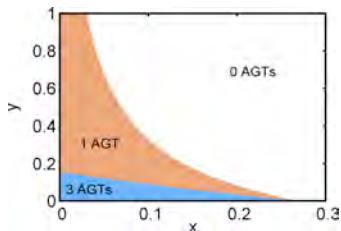
# Gene tree distribution for four taxa



# Gene tree distribution for four taxa



## 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

## 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



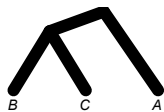
## Phylogenetic coalescent model with mutation

$t = \text{length of interval between coalescent events} = 1.0$



$$1 - e^{-t}$$

0.63



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

0.12



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

0.12



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

0.12

**Example:** Want to compute the probability that taxon  $A$  has nucleotide  $T$ , taxon  $B$  has nucleotide  $G$  and taxon  $C$  has nucleotide  $T$  – call this  $p_{TGT}$

## Phylogenetic coalescent model with mutation

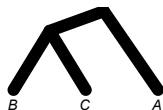
**Example:** Want to compute the probability that taxon  $A$  has nucleotide  $T$ , taxon  $B$  has nucleotide  $G$  and taxon  $C$  has nucleotide  $T$  – call this  $p_{TGT}$



$$1 - e^{-t}$$

$$0.63$$

$$p_{TGT}^{1a} = 0.1$$



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

$$0.12$$

$$p_{TGT}^{1b} = 0.025$$



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

$$0.12$$

$$p_{TGT}^2 = 0.2$$



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

$$0.12$$

$$p_{TGT}^3 = 0.025$$

## Phylogenetic coalescent model with mutation

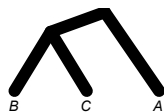
**Example:** Want to compute the probability that taxon  $A$  has nucleotide  $T$ , taxon  $B$  has nucleotide  $G$  and taxon  $C$  has nucleotide  $T$  – call this  $p_{TGT}$



$$1 - e^{-t}$$

$$0.63$$

$$p_{TGT}^{1a} = 0.05$$



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

$$0.12$$

$$p_{TGT}^{1b} = 0.025$$



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

$$0.12$$

$$p_{TGT}^2 = 0.2$$



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

$$0.12$$

$$p_{TGT}^3 = 0.025$$

$$p_{TGT} = 0.63 \times 0.05 + 0.12 \times 0.025 + 0.12 \times 0.2 + 0.12 \times 0.025 = 0.0615$$

## Phylogenetic coalescent model with mutation

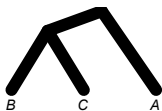
**Example:** Want to compute the probability that taxon  $A$  has nucleotide  $T$ , taxon  $B$  has nucleotide  $G$  and taxon  $C$  has nucleotide  $T$  – call this  $p_{TGT}$



$$1 - e^{-t}$$

$$0.63$$

$$p_{TGT}^{1a} = 0.05$$



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

$$0.12$$

$$p_{TGT}^{1b} = 0.025$$



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

$$0.12$$

$$p_{TGT}^2 = 0.2$$



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

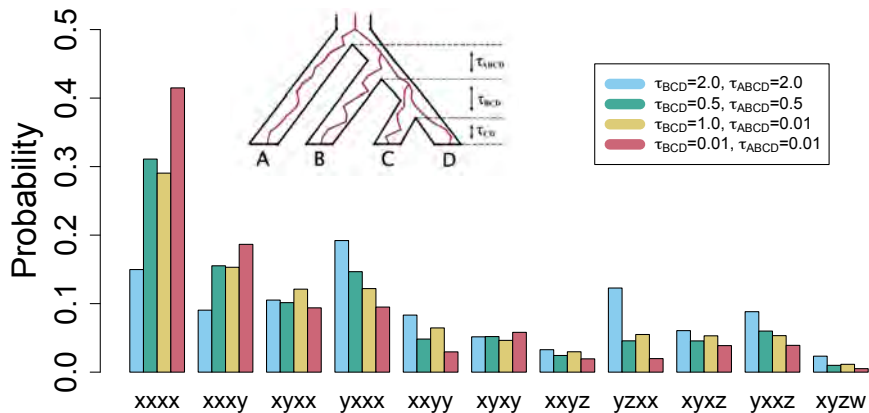
$$s 0.12$$

$$p_{TGT}^3 = 0.025$$

$$p_{TGT} = 0.63 \times 0.05 + 0.12 \times 0.025 + 0.12 \times 0.2 + 0.12 \times 0.025 = 0.0615$$

↑ *For intuition only, not completely correct ...*

# What does the site pattern probability distribution look like?



## What about mutation?

### 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)

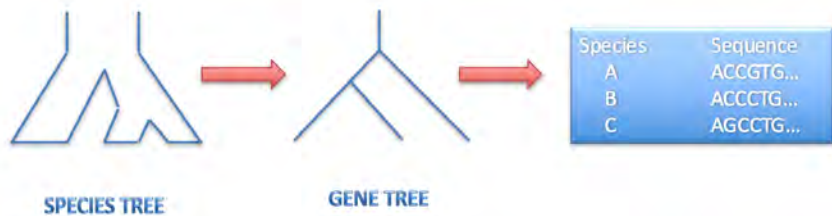
Kubatko and Degnan, 2007; Roch and Steel, 2015

Is there a better way to estimate species phylogenies?

**Explicitly model the coalescent process!**



## Phylogenetic coalescent model with mutation



## Why is this so hard?

### The likelihood function

- Suppose that we have available alignments for  $N$  genes, denoted by  $D_1, D_2, \dots, D_N$
- 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

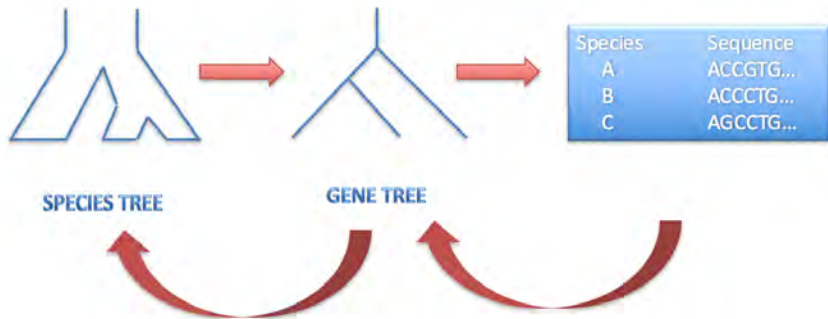
- Recall the **Felsenstein equation** from Peter's lecture, except that now we replace  $\theta$  with  $S$ , the species tree. Use this to form the species tree likelihood for a multi-locus data set:

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

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

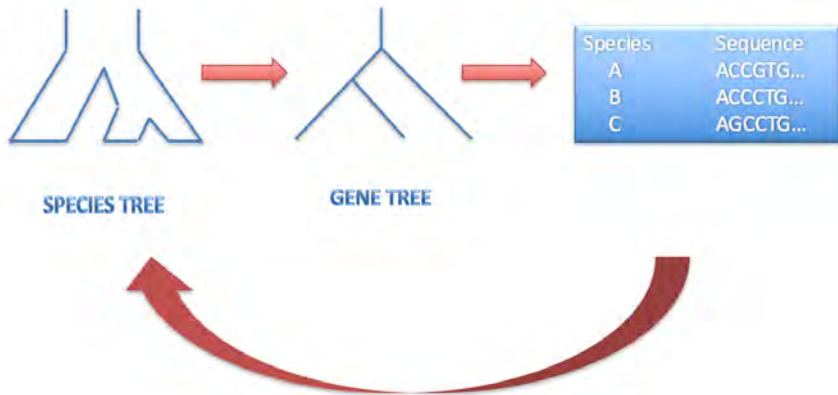
- This likelihood is difficult to evaluate directly, because of the dimension of the inner sum (which is really an integral) **[recall Peter's "galaxy slide"]**

## Inference option 1: Summary statistics methods



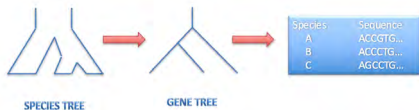
- **Summary statistics methods:** Start with estimated gene trees
  - ▶ Using estimated branch lengths:
    - ★ STEM (Kubatko et al. 2009)
    - ★ STEAC (Liu et al. 2009)
  - ▶ Using topology information only:
    - ★ STAR (Liu et al. 2009)
    - ★ Minimize Deep Coalescences (PhyloNet; Than & Nakhleh 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



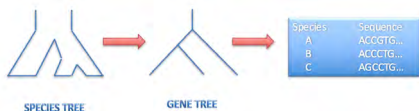
## Full data methods I: BEST, \*BEAST/STARBEAST2, BPP, SNAPP

- 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.

- 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 Thursday'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/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**

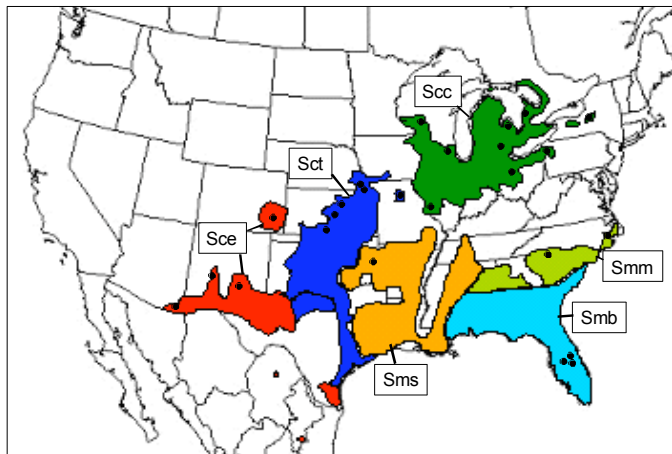
## Example 1: *Sistrurus* rattlesnakes



- 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 Chiuicchi and Brian Fedorko]

## Geographic Distribution of Snake Populations



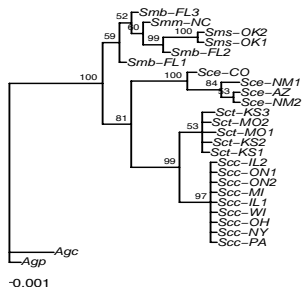
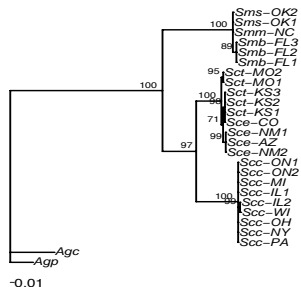


- Data: 7 (sub)species, 26 individuals (52 sequences), 19 genes

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

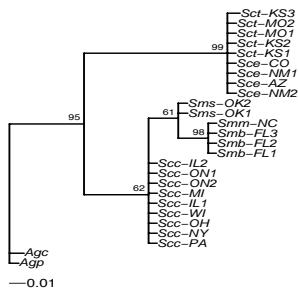
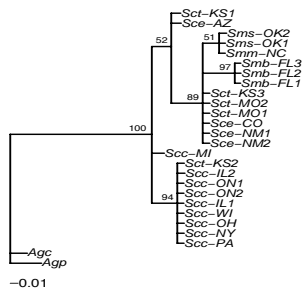
## Individual Gene Tree Estimates

Some are very informative:



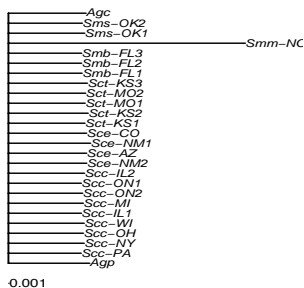
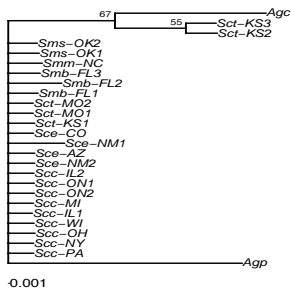
## Individual Gene Tree Estimates

Some are a little informative:



## Individual Gene Tree Estimates

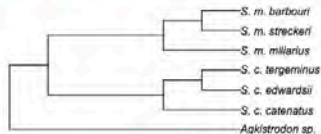
And then there are others .....



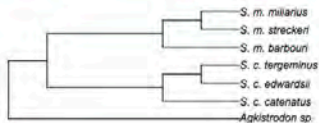


## Example 1: Sistrurus rattlesnakes

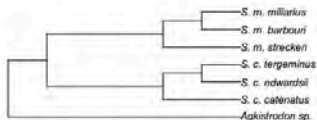
### STEM, STEAC



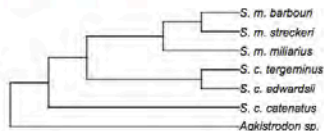
### BEAST (concatenated data), \*BEAST



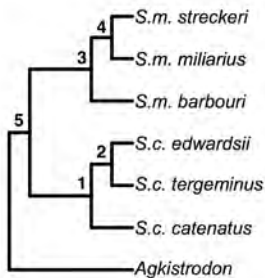
### BEST, Parsimony & MrBayes (concatenated data), Astral



### 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)

## Example 1: *Sistrurus rattlesnakes*

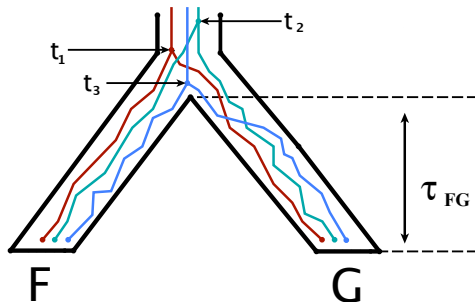
- How does concatenation do?

- ▶ 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) <sup>a</sup>	Divergence estimates from species tree (Ma) <sup>a</sup>	Percent difference <sup>b</sup> (%)
(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!

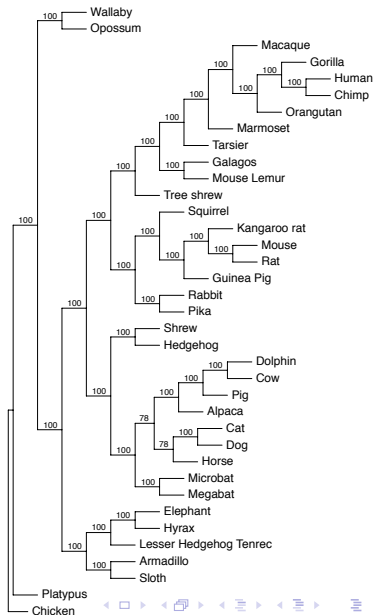


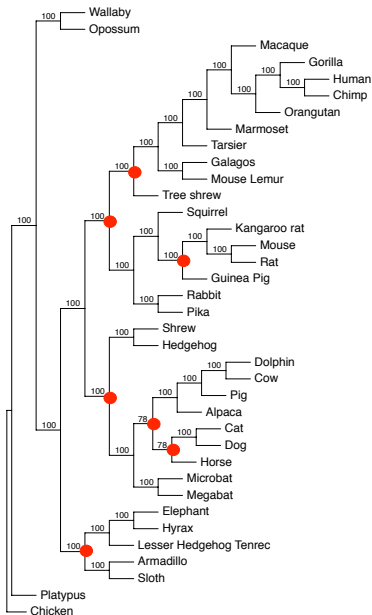
## Multilocus data example 2: Mammals

- **Series of papers** in the literature debating proper phylogenetic relationships among a group of mammals
  - ▶ **Meredith RW, et al. (Science, 2011)** criticized by **Song et al. (PNAS, 2012)**:
    - ★ Amount of data “insufficient” (26 genes, 35,603 bp, 164 mammals)
    - ★ Concatenation not appropriate
  - ▶ Response by **Gatesy and Springer (PNAS, 2013)** criticizing Song et al.:
    - ★ Loci chosen not representative (“concatalescence” – exons ‘pasted’ together)
    - ★ Many nodes still not well supported
    - ★ Subset of 36 species
  - ▶ **Wu et al. (PNAS, 2013)** criticize Gatesy and Springer’s response:
    - ★ Concatenation of all genes is worse than within a few genes
    - ★ The approach of treating exons from a single gene with introns stripped has worked well in other cases
  - ▶ etc. ...

## Example 2: Mammals

- **Dataset:** obtained from Liang Liu, 36 mammal species + outgroup, ~ 1.4 million bp from 447 genes
- **SVDQ** run on 8-year old dual-core linux machine – **27 hours** required to estimate the tree and obtain bootstrap support from 100 replicates





- “Historically problematic nodes” identified by McCormack et al. (Genome Research, 2012) are identified with a red circle
- Overall, the SVDQ analysis agrees with the analysis of Song et al. (2012), who used the coalescence-based method MP-EST
- The SVDQ analysis differs from analyses based on concatenation for some of the difficult nodes, but agrees with concatenation for the two nodes with lower bootstrap support

## 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/ STARBEAST2	multilocus	posterior probability	intermediate; can be run in parallel	coalescent; all reversible substitution models; relaxed clock; variable population sizes
BPP	multilocus	posterior probability	long	coalescent; JC69 model on molecular clock; species delimitation
SVDQ	multilocus; SNP	bootstrap	short	coalescent; all reversible substitution models; non-cl 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	local posterior probability	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.

## 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: