Species Tree Inference

Laura Kubatko
Departments of Statistics and
Evolution, Ecology, and Organismal Biology
The Ohio State University
kubatko.2@osu.edu

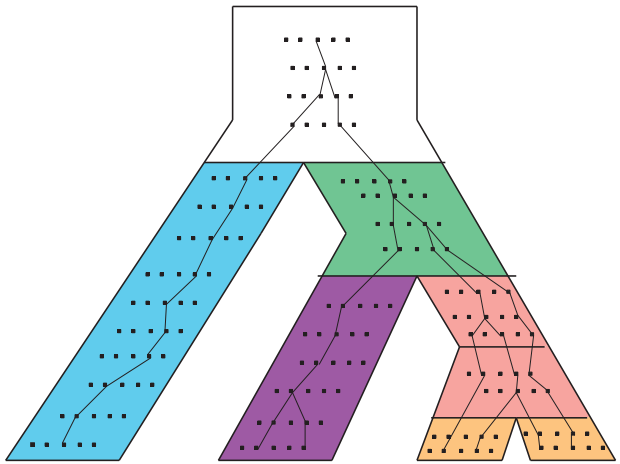
July 30, 2013

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)

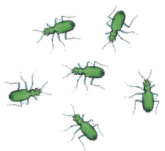
Relationship between population genetics and phylogenetics

- 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]



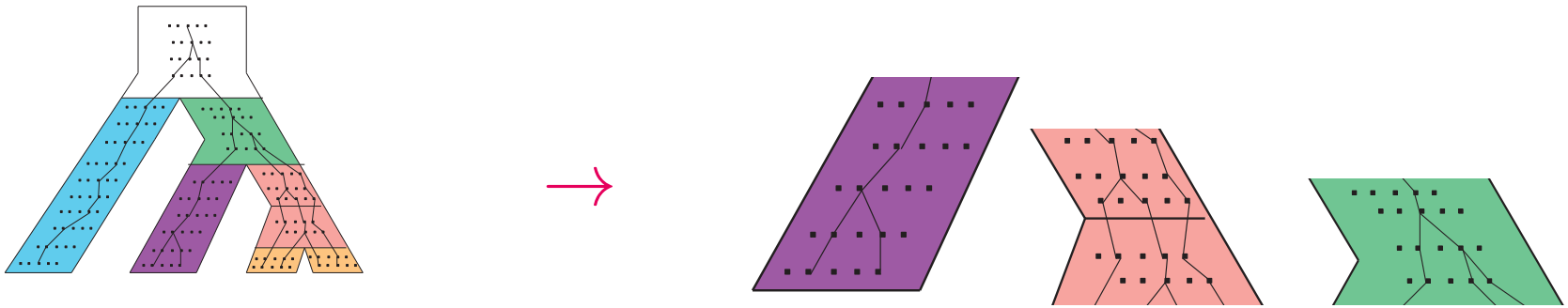
Recall several facts from Peter's lecture

- Under the Wright-Fisher model, the number of generations back into the past until two lineages coalesce $\sim$ Geometric($\frac{1}{2N}$)
- Kingman's approximation: consider continuous time and a sample of k lineages. Then, the time back into the past until two lineages coalesce, U, is exponentially distributed with rate $\binom{k}{2}\frac{1}{2N}$.
  - The probability density function is $g(u) = \binom{k}{2}\frac{1}{2N}e^{-\binom{k}{2}\frac{u}{2N}}$, for $u > 0$.
  - The mean is $\frac{4N}{k(k-1)}$.



- Peter showed us how to use this model to compute the probability density of a "population tree"

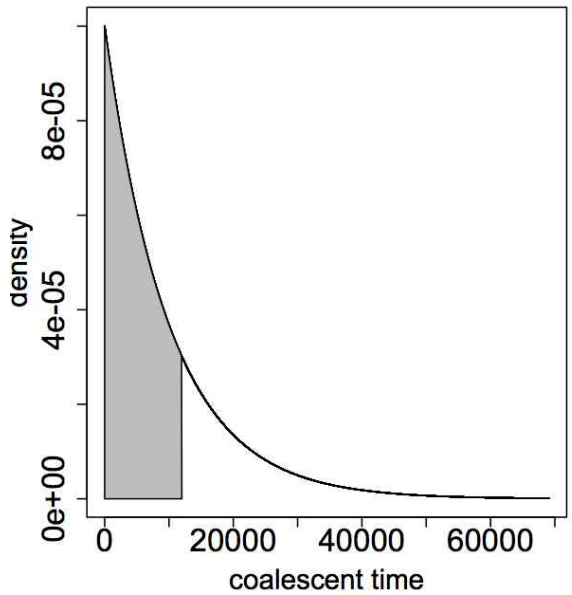
Fitting population trees into a phylogeny

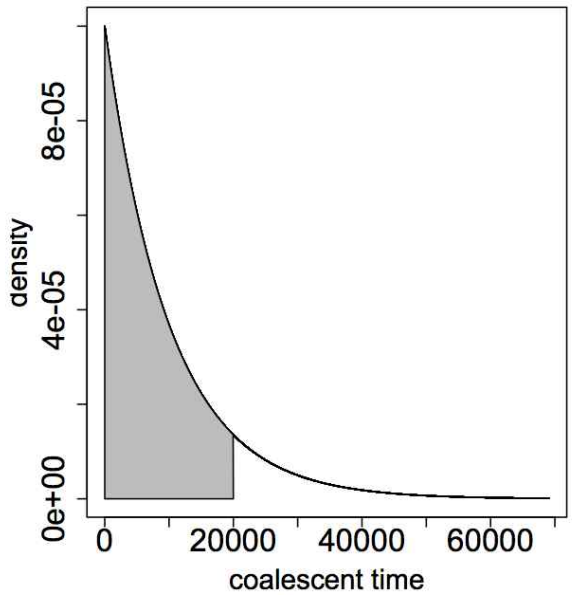


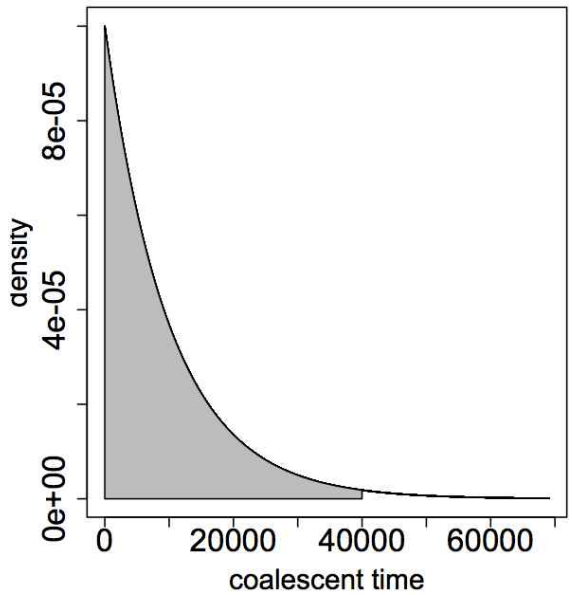
- Focus first on just one speciation interval and a sample of $k = 2$ lineages.
- Then, $\binom{k}{2} = 1$ and we have an exponential distribution with rate $\frac{1}{2N}$ and mean $2N$.
- Suppose $N = 5,000$. Let's find the probability that the two lineages coalesce in an interval of a particular length.

Fitting population trees into a phylogeny

- $N = 5,000$, and consider the times: 12,000, 20,000, and 40,000 generations





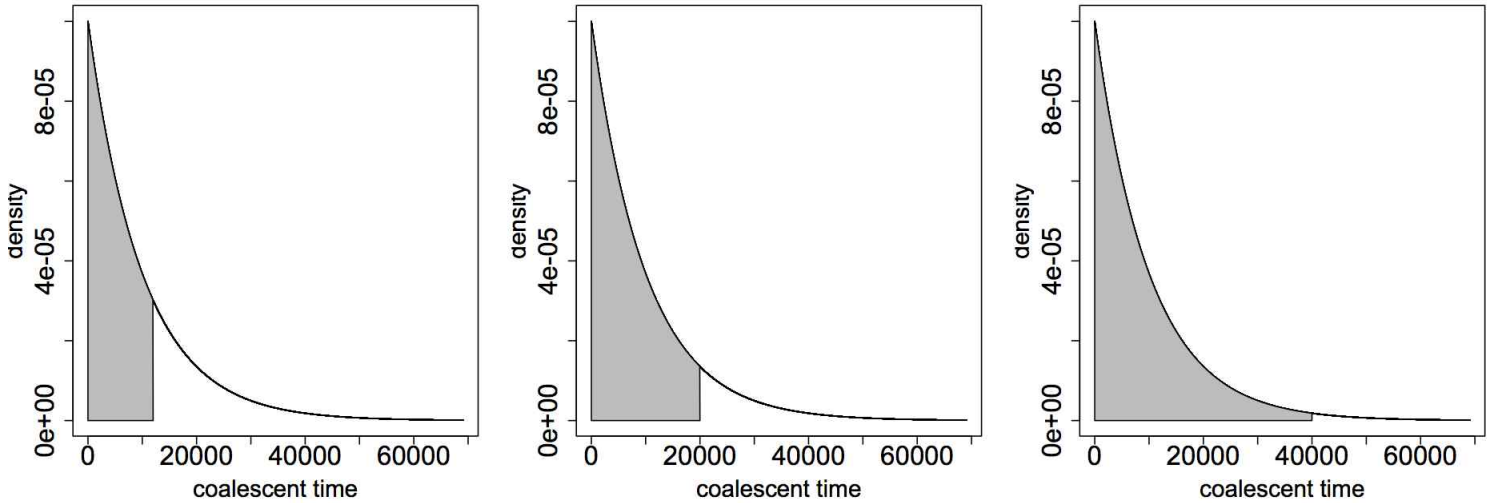


Fitting population trees into a phylogeny

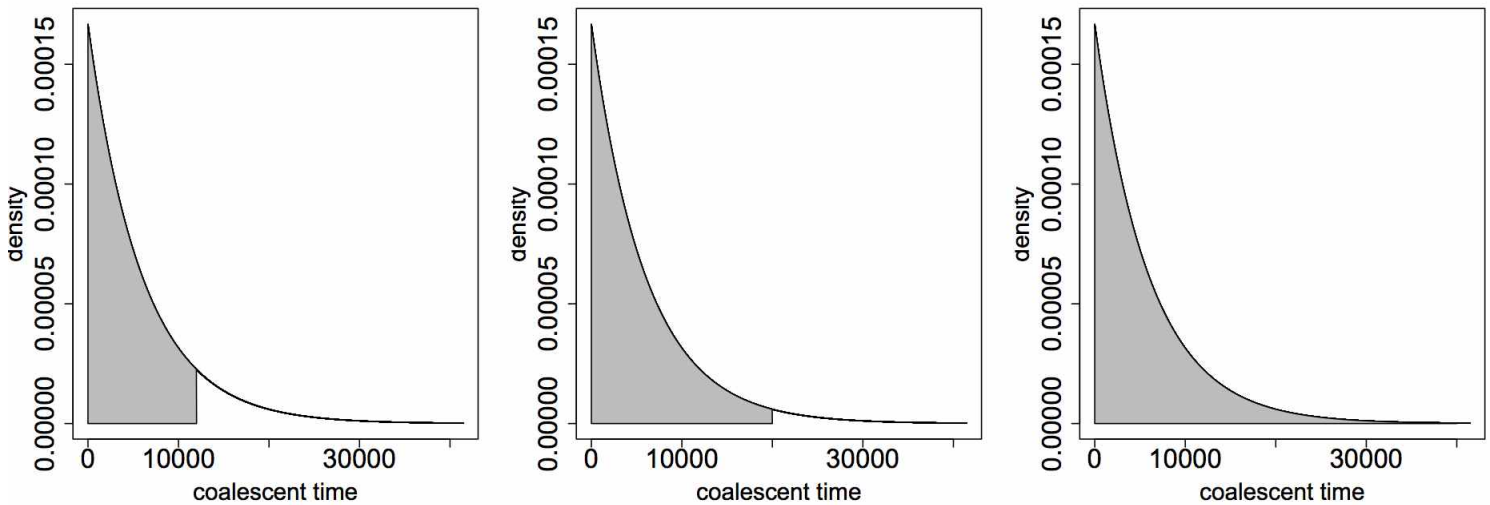
- What happens if we change the population size, N?
- Recall we have an exponential distribution with rate $\frac{1}{2N}$ and mean $2N$.
- Now suppose $N = 3,000$ and look at the same speciation interval lengths

Fitting population trees into a phylogeny

- $N = 5,000$

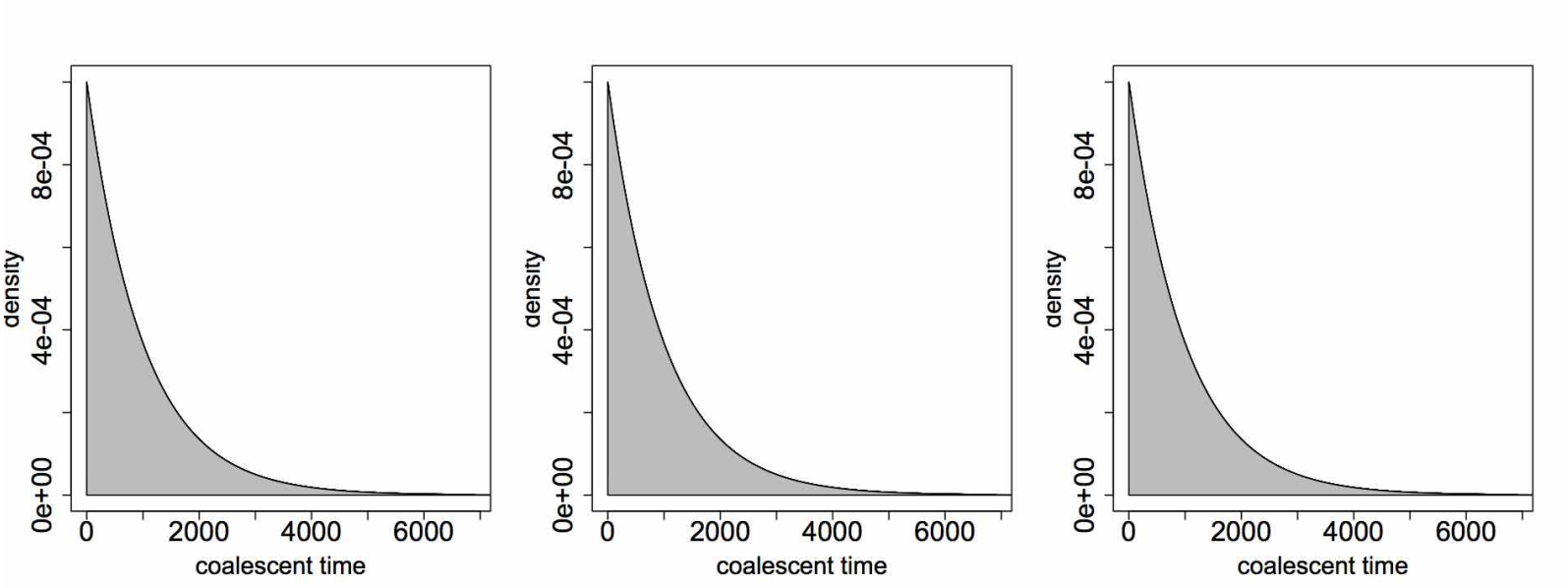


- $N = 3,000$



Fitting population trees into a phylogeny

- What about the effect of sample size, k?
- Consider $N = 5,000$ again, but now use $k = 5$.
  - Rate is $\binom{5}{2}\frac{1}{2N} = \frac{10}{2N}$ (was $\frac{1}{2N}$)
  - Mean is $\frac{4N}{k(k-1)} = \frac{2N}{10}$ (was $2N$)

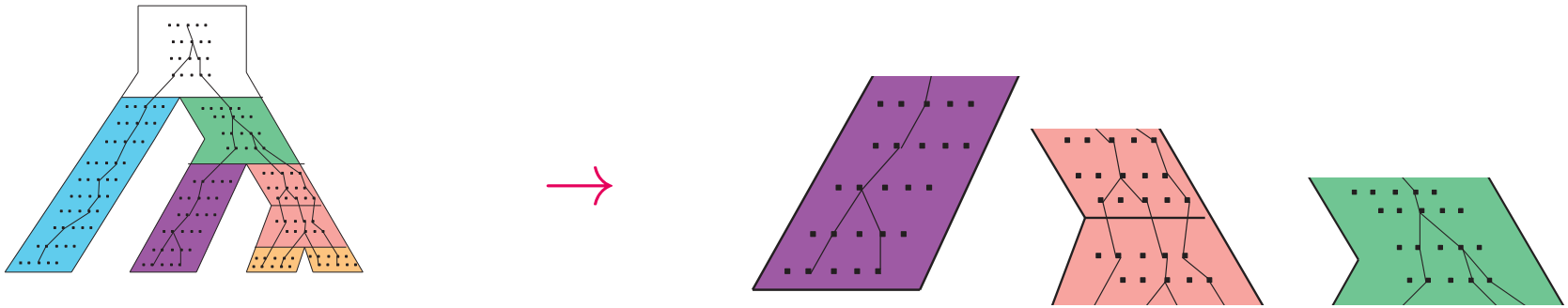


Fitting population trees into a phylogeny

- 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 time intervals lead to a higher probability of coalescence occurring.
  - Coalescent events happen more quickly when the population size is smaller.
  - Coalescent events happen more quickly when the sample size is larger.
- What does this mean for species tree estimation????

Fitting population trees into a phylogeny

- Recall our goal to integrate the population process with the phylogeny:



- Can use our previous results to get the following:
  - The probability that u lineages coalesce into v lineages in time t is given by (Tavare, 1984; Watterson, 1984; Takahata and Nei, 1985; Rosenberg, 2002)

$$P_{uv}(t) = \sum_{j=v}^{u} e^{-j(j-1)t/2} \frac{(2j-1)(-1)^{j-v}}{v!(j-v)!(v+j-1)} \prod_{y=0}^{j-1} \frac{(v+y)(u-y)}{u+y}$$

Fitting population trees into a phylogeny

- When u and v are small, these are easy to compute. For example,

$$\begin{aligned} P_{21}(t) &= \text{probability that 2 lineages coalescence to 1 lineage in time t} \\ &= \text{probability of 1 coalescent event in time t when k=2} \\ &= P(T \leq t), \text{where } T \sim Exp(\mu = 1) \\ &= \int_0^t e^{-x} dx = 1 - e^{-t} \end{aligned}$$

[Note: this is the formula for the gray area in the graphs]

- Similarly,

$$\begin{aligned} P_{22}(t) &= \text{prob. of no coalescence in time t for 2 lineages} \\ &= P(T > t) \\ &= \int_t^\infty e^{-x} dx = e^{-t} \end{aligned}$$

Putting it together . . . the coalescent model along a species tree

- Assumptions:
  - 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.

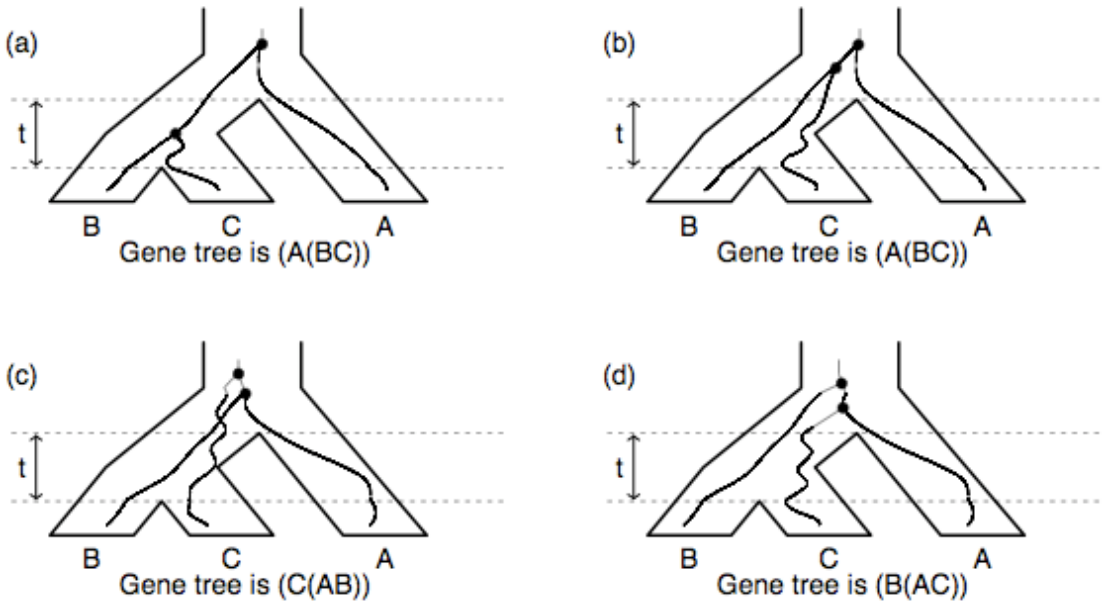
Putting it together . . . the coalescent model along a species tree

- When talking about gene tree distributions, there are two cases of interest:
  - The gene tree topology distribution
  - The joint distribution of topologies and branch lengths
- Start with the simple case of 3 species with 1 lineage sampled in each and look at the gene tree topology distribution

Example: computation of gene tree topology probabilities for the 3-taxon case

Example of gene tree probability computation:

(a) Prob $= 1 - e^{-t}$; (b), (c), (d) Prob $= \frac{1}{3}e^{-t}$



Example: computation of gene tree topology probabilities for the 3-taxon case

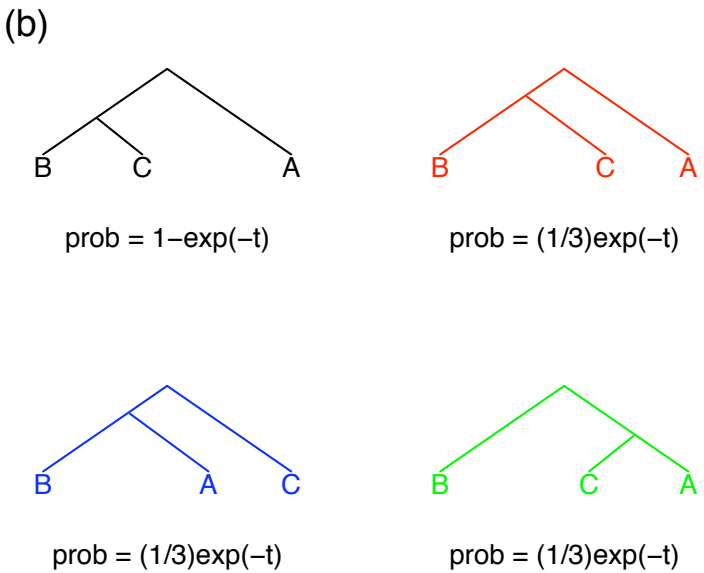
- Thus, we have the following probabilities:
  - Gene tree (A,(B,C)): prob $= 1 - e^{-t} + \frac{1}{3}e^{-t} = 1 - \frac{2}{3}e^{-t}$
  - Gene tree (B,(A,C)): prob $= \frac{1}{3}e^{-t}$
  - Gene tree (C,(A,B)): prob $= \frac{1}{3}e^{-t}$

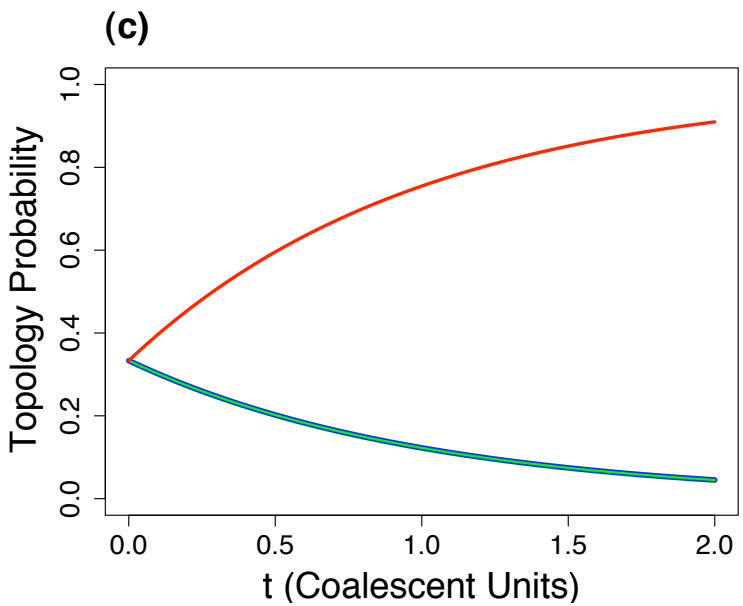
- Note: There are two ways to get the first gene tree. We call these histories.

- The probability associated with a gene tree topology will be the sum over all histories that have that topology.

Example: computation of gene tree topology probabilities for the 3-taxon case

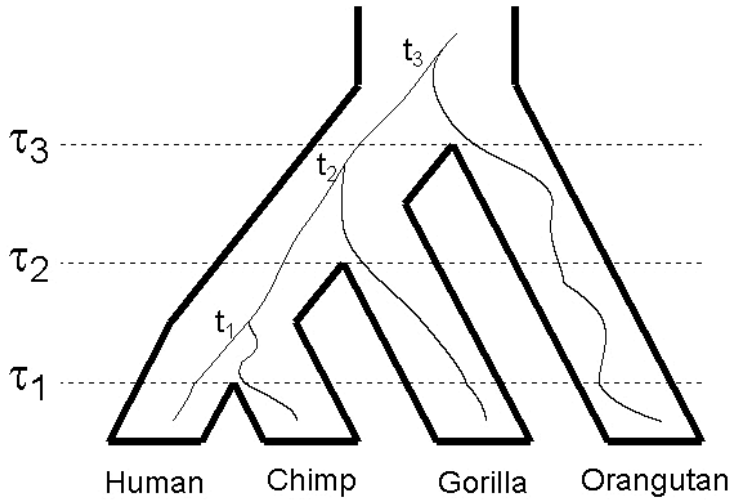
- What are these probabilities like as a function of t, the length of time between speciation events?



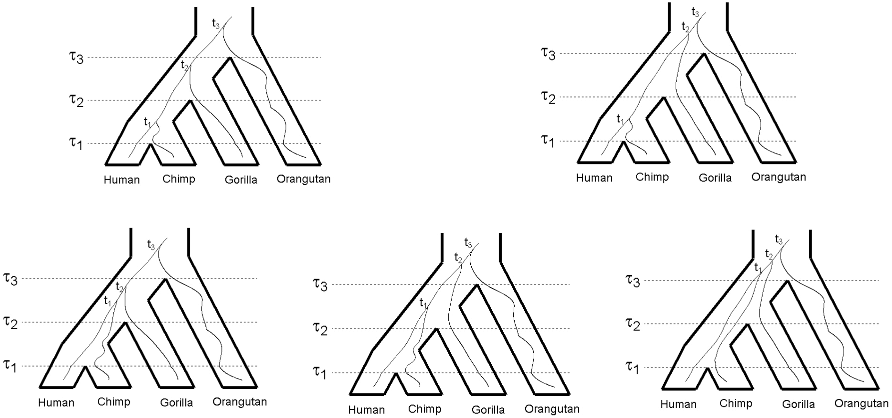


Example: a slightly larger case

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



- There are 5 possibilities for this example:



Computing the topology distribution by enumerating histories

- In the general case, we have the following:

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

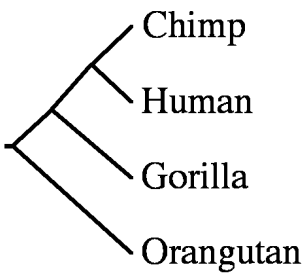
$$P\{G = g|\mathcal{S}\} = \sum_{histories} P\{G = g, history|\mathcal{S}\}$$

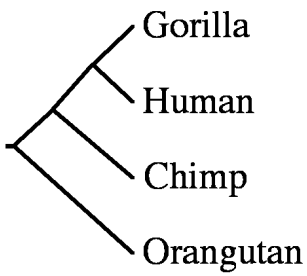
- Implemented in the software COAL (Degnan and Salter, *Evolution*, 2005)
- A more efficient method has been proposed (Wu, *Evolution*, 2012)

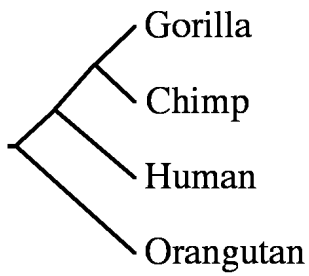
Applications of the topology distribution - example 1

- 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.

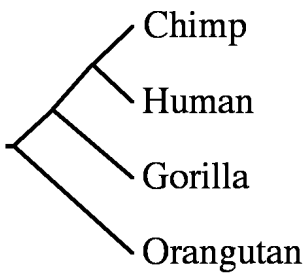
Applications of the topology distribution - example 1



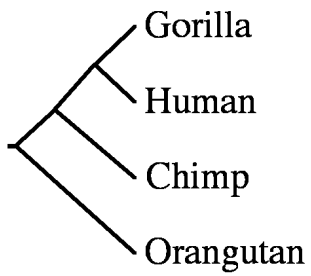




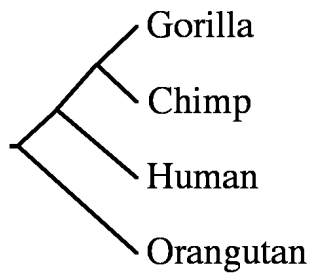
Applications of the topology distribution - example 1



76.6%



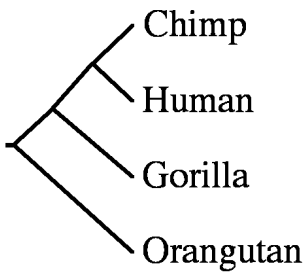
11.4%

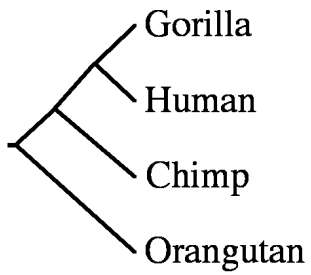


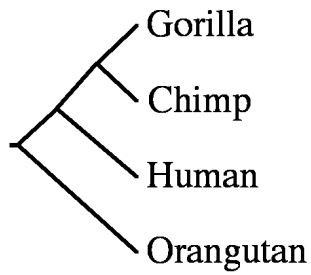
11.5%

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

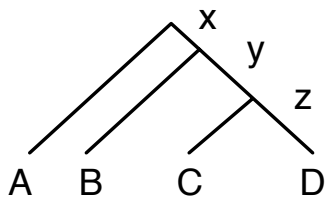
Predicted proportions using parameters from Rannala & Yang, 2003.

Applications of the topology distribution - example 2

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

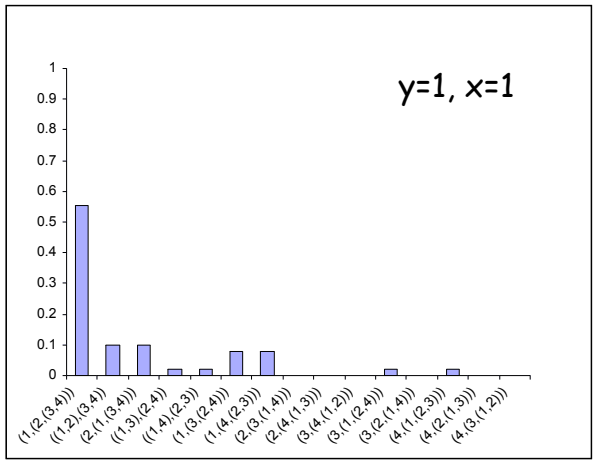
Applications of the topology distribution - example 2

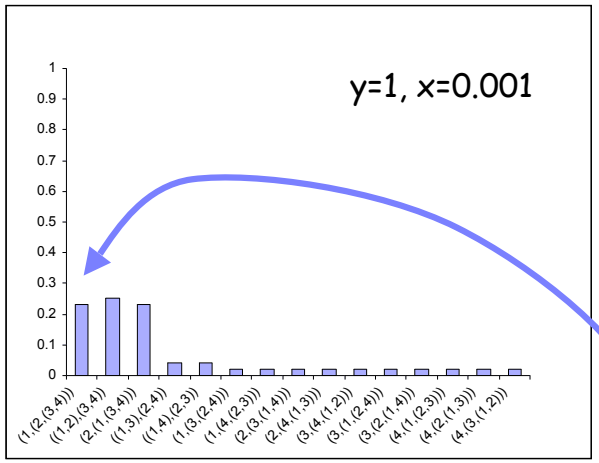
- Consider 4 taxa: A, B, C, and D
- Species tree:

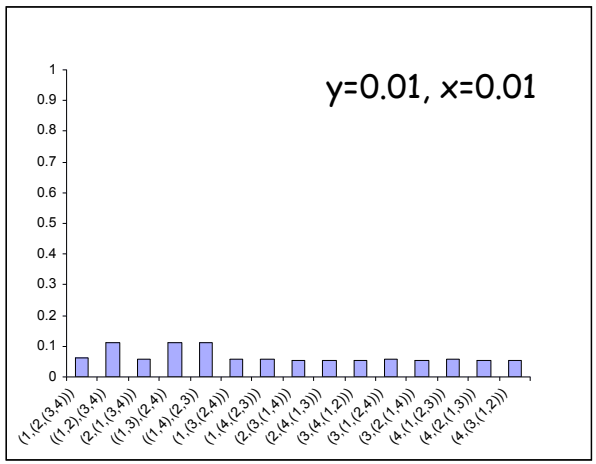


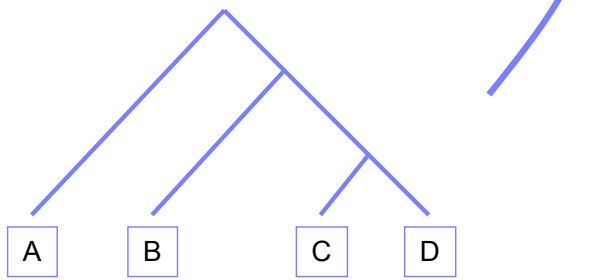
- Look at probabilities of all 15 tree topologies for values of x, y, and z

Applications of the topology distribution - example 2

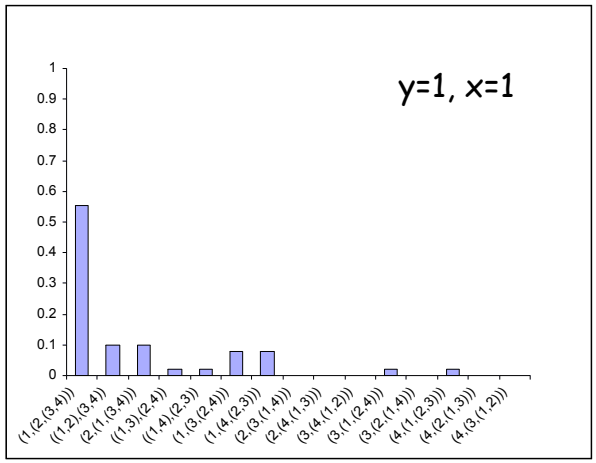


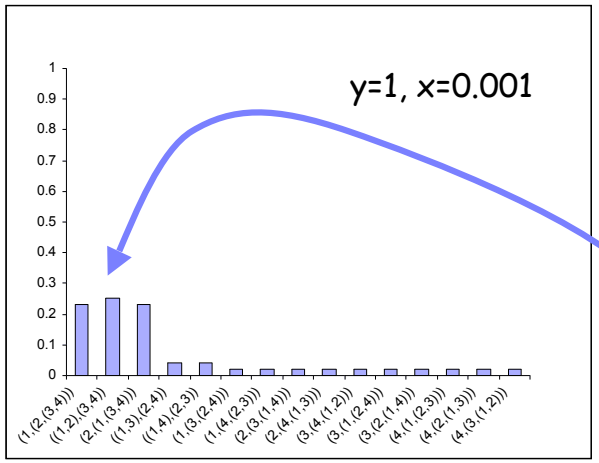


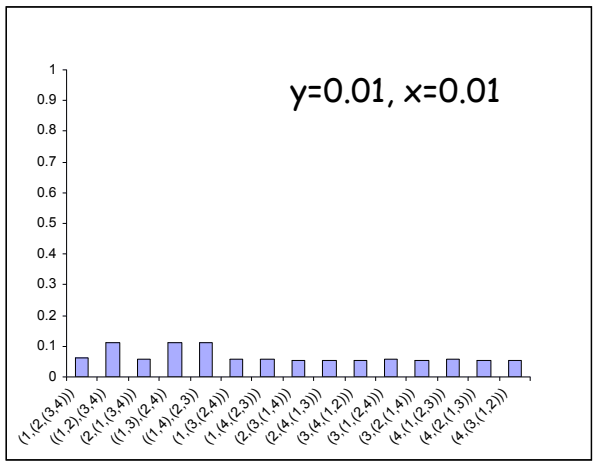


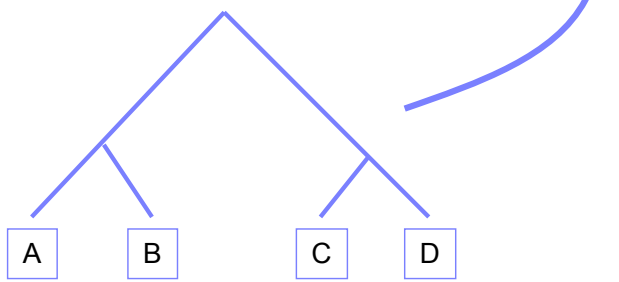


Applications of the topology distribution - example 2

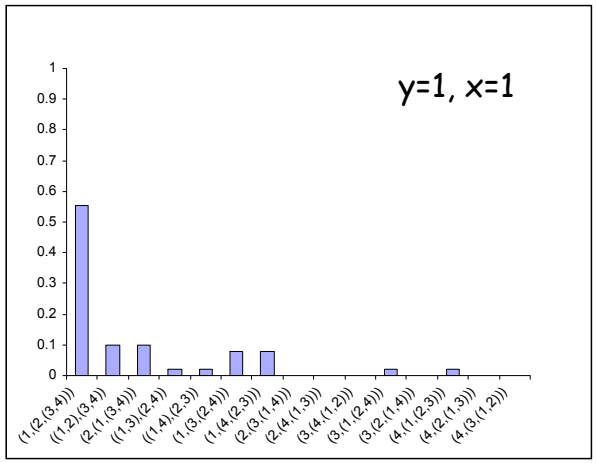


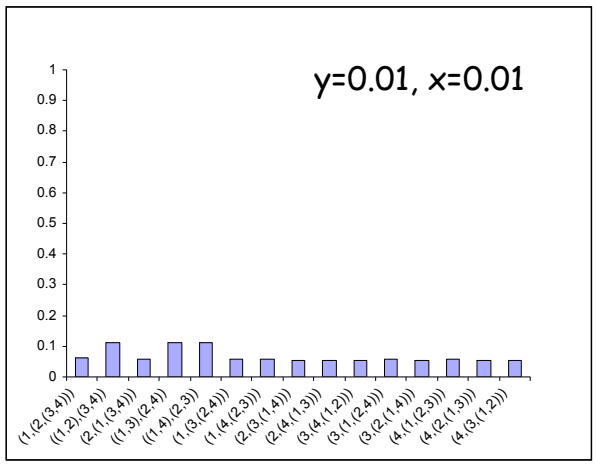


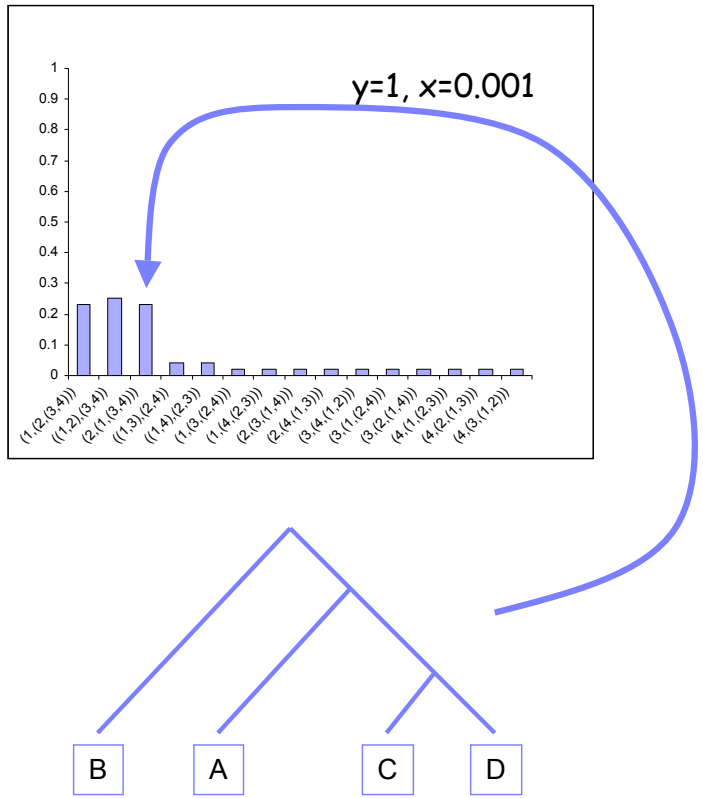




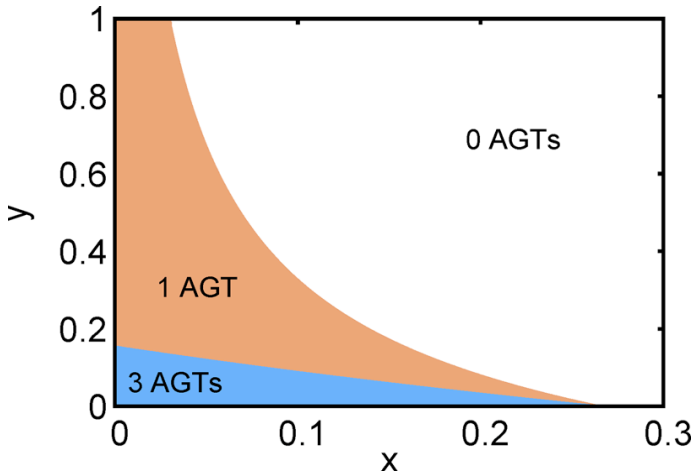
Applications of the topology distribution - example 2







Applications of the topology distribution - example 2



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

Degnan & Rosenberg, *PLoS Genetics*, 2006
Rosenberg & Tao, *Systematic Biology*, 2008

Applications of the topology distribution - example 3

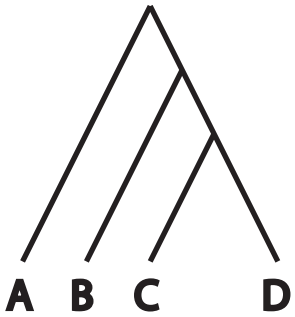
- What about mutation? How does this affect data analysis?
- The coalescent gives a model for determining gene tree probabilities for each gene.
- View DNA sequence data as the result of a two-stage process:
  - Coalescent process generates a gene tree topology.
  - Given this gene tree topology, DNA sequences evolve along the tree.

Applications of the topology distribution - example 3

- Given this model, how should inference be carried out?
- Hypothesis: 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.

Applications of the topology distribution - example 3

Species Tree

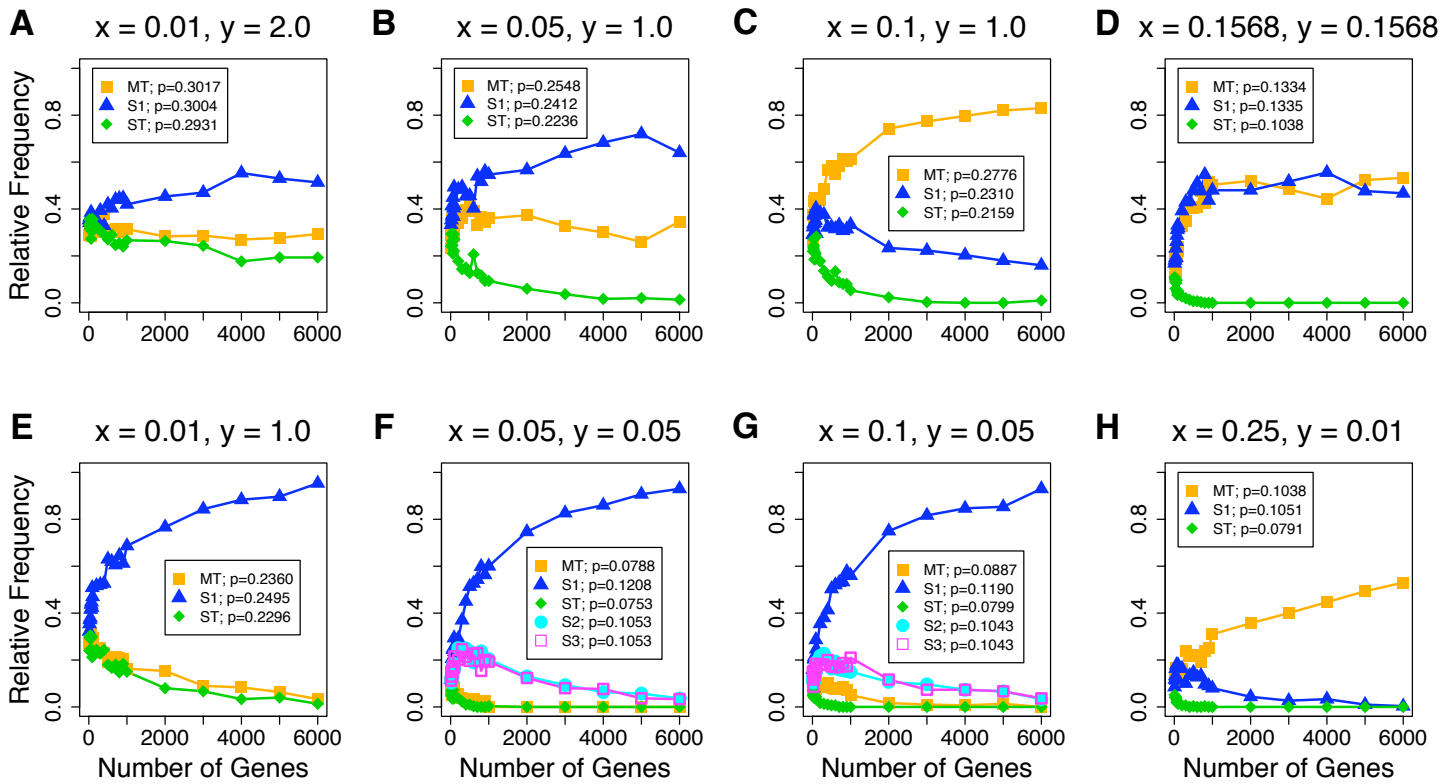


Generate gene trees in COAL → Generate sequence data in Seq-Gen → Estimate tree using concatenation

Repeat 100 times

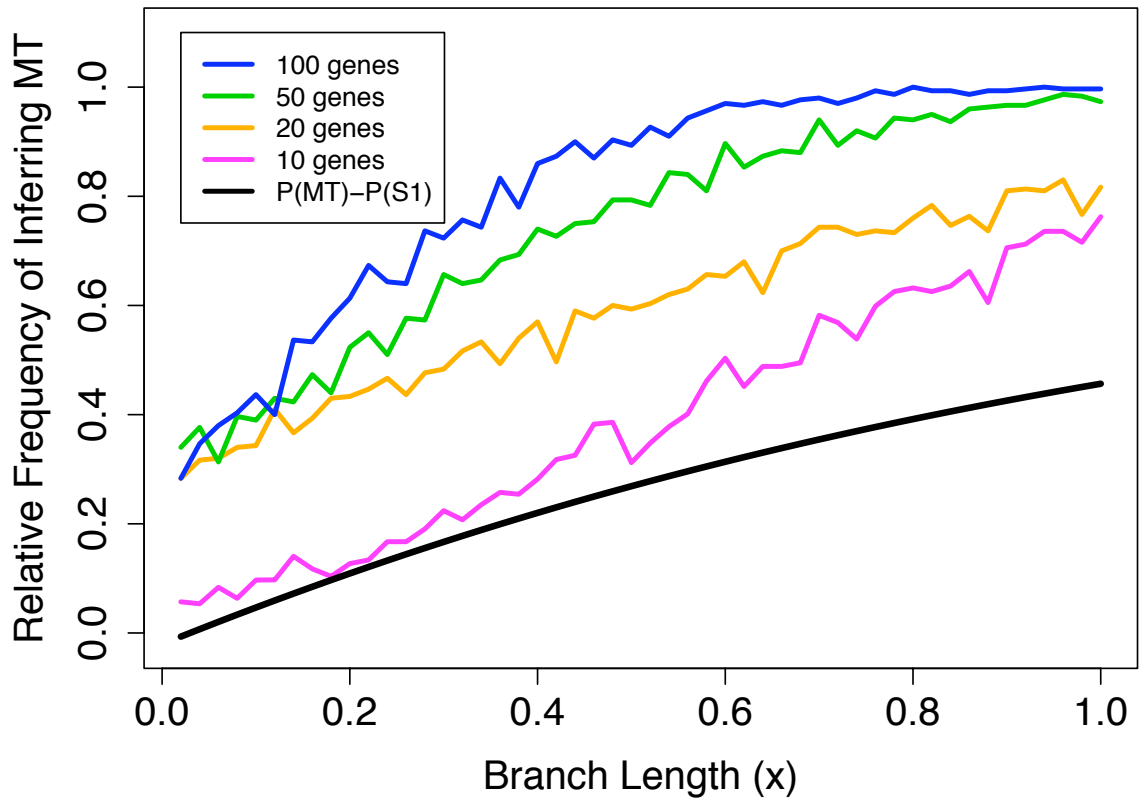
Applications of the topology distribution - example 3

Simulation Study 1



Applications of the topology distribution - example 3

Simulation Study 2

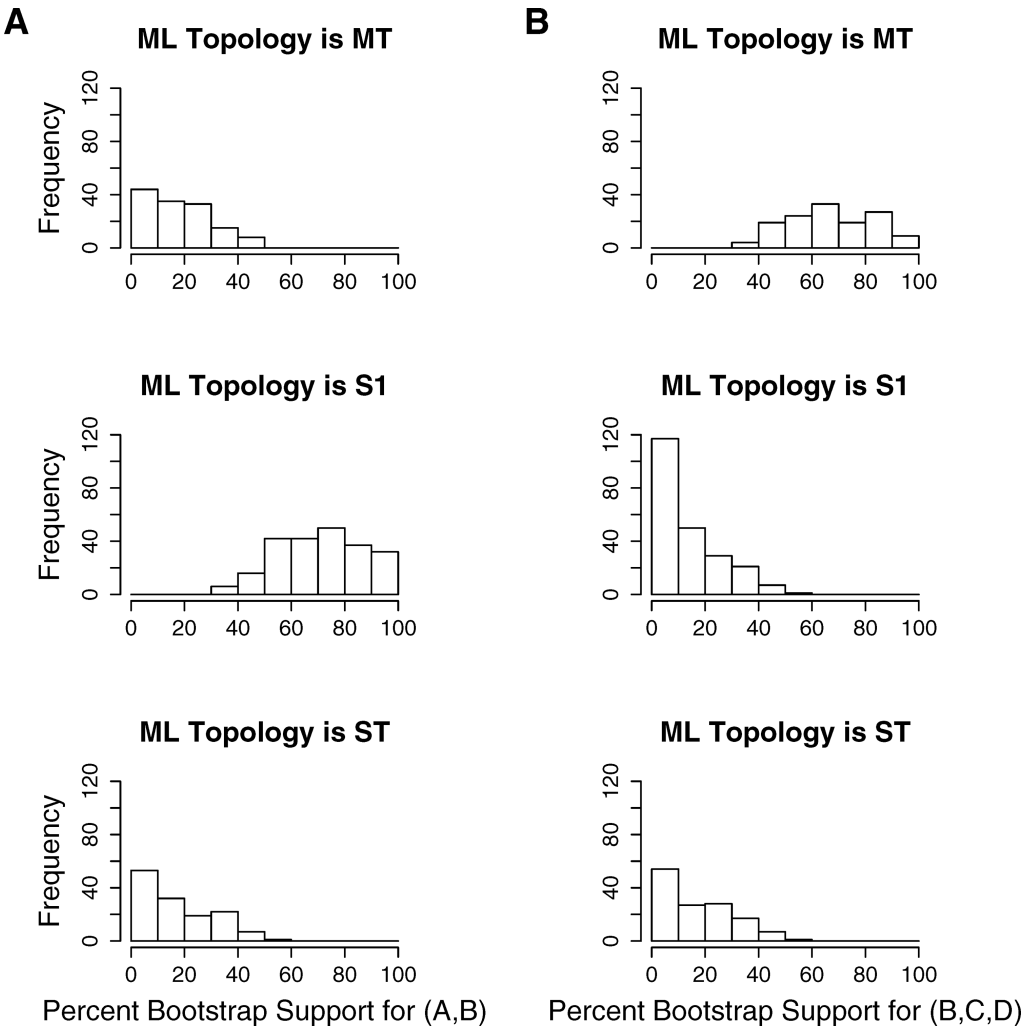


Applications of the topology distribution - example 3

- Performance of the Concatenation Approach:
  - Can be statistically inconsistent when branch lengths in the species phylogeny are sufficiently small
  - May perform poorly even when branch lengths are only moderately short
  - Bootstrap procedure can be positively misled in this situation
- Question: How does the bootstrap perform in these cases?

- Hypothesis: The bootstrap may provide strong support for the incorrect tree when gene trees that are incongruent with the species tree are fairly probable
- Simulation study to examine the performance of the bootstrap:
  - n=100 loci
  - x=0.01, y=1.0
  - $\theta = 0.001$
  - B=200 bootstrap samples per repetition
  - Repeated 500 times

The concatenation approach – performance of the bootstrap



The concatenation approach – performance of the bootstrap

- The bootstrap can 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)
- Question: Is there a better way to estimate species phylogenies?

The concatenation approach – performance of the bootstrap

- The bootstrap can 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)
- Question: Is there a better way to estimate species phylogenies?

Explicitly model the coalescent process!

- Summary statistic 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)

- Methods that utilize the full data: Input is aligned sequences
  - BEST (Liu and Pearl 2007)
  - *BEAST (Heled and Drummond 2010)
  - SNAPP (Bryant et al. 2012)
  - SVDquartets (Chifman and Kubatko 2013)

- Comparison of approaches:
  - Summary statistic methods
    - ★ Advantage: Quick
    - ★ Disadvantage: Ignore information in data
    - ★ Most current implementations do not easily allow for assessment of uncertainty
  - Full data methods
    - ★ Advantage: Fully model-based framework
    - ★ Disadvantage: Computationally intensive, sometimes prohibitively so
    - ★ BEST, *BEAST, and SNAPP utilize a Bayesian framework and involve MCMC
- Give overview of "representative" subset of methods:
  - STEM, BEST, *BEAST, SNAPP, SVDquartets + BUCKy
  - Scott – MP-EST, Phybase

Likelihood function

- Suppose that we have available alignments for N genes, denoted by $D_1, D_2, ...D_N$
- 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 model
  - 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 now we replace θ with S, the species tree. Use this to form the species tree likelihood for a multi-locus data set:

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

  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"]

- To deal with this, either assume gene trees are known (summary statistics), use Bayesian techniques (full data approaches), or think about small problems ☹ .

STEM: The gene tree-species tree likelihood function

- A simpler problem is to suppose that our data consist of a set of gene trees
- Let $g_1, g_2, \ldots g_N$ be a set of N gene trees with branch lengths
- Consider a species tree, S (topology and branch lengths)
- The likelihood function is

$$L(S|g_1, g_2, \ldots g_N) = \prod_{j=1}^{N} f(g_j|S)$$

where $f(g|S)$ is given by Rannala and Yang (2003).

Maximum likelihood estimate of the species tree

- Liu et al. (2009) showed that the ML estimate of the species tree can be computed by sequentially clustering minimum observed divergence times between pairs of species across genes.
- They have shown that when gene trees are known without error, the ML species tree is a consistent estimator.
- A similar result was obtained by Roch & Mossel (2010) – they call their estimator the GLASS tree (an acronym for Global LAteSt Split, based on the algorithm they developed to compute it).
- STEM computes the ML estimate of the species tree this way.

- Five main functions of STEM-hy version 1.0
  - Estimate a species tree given a set of gene trees using maximum likelihood
    run=1
  - Search species tree space for trees of high likelihood
    run=2
  - Compute the likelihood of a user-specified tree
    run=0
  - Carry out a bootstrap analysis (bootstrapping is on sites within genes)
    run=4
  - Assess fit of trees subject to hybridization in the presence of lineage sorting
    run=3

STEM: Strengths and weaknesses

- STEM makes some strong assumptions:
  - Error in estimating gene trees and branch lengths is not incorporated (but the bootstrap helps with this).
  - Information in the sequence data is not used directly; it is only used as summarized by estimated gene divergence times.
  - There is a single value of θ for the entire tree.

- There are trade-offs involved, and STEM does some things well:
  - It is quick (even the tree search does not take long).
  - It can handle missing data easily and intuitively.
  - Simulations demonstrate reasonable performance (unlikely to be misleading; may be uninformative).

Full data methods (1): BEST and *BEAST

- Model the entire process of data generation:

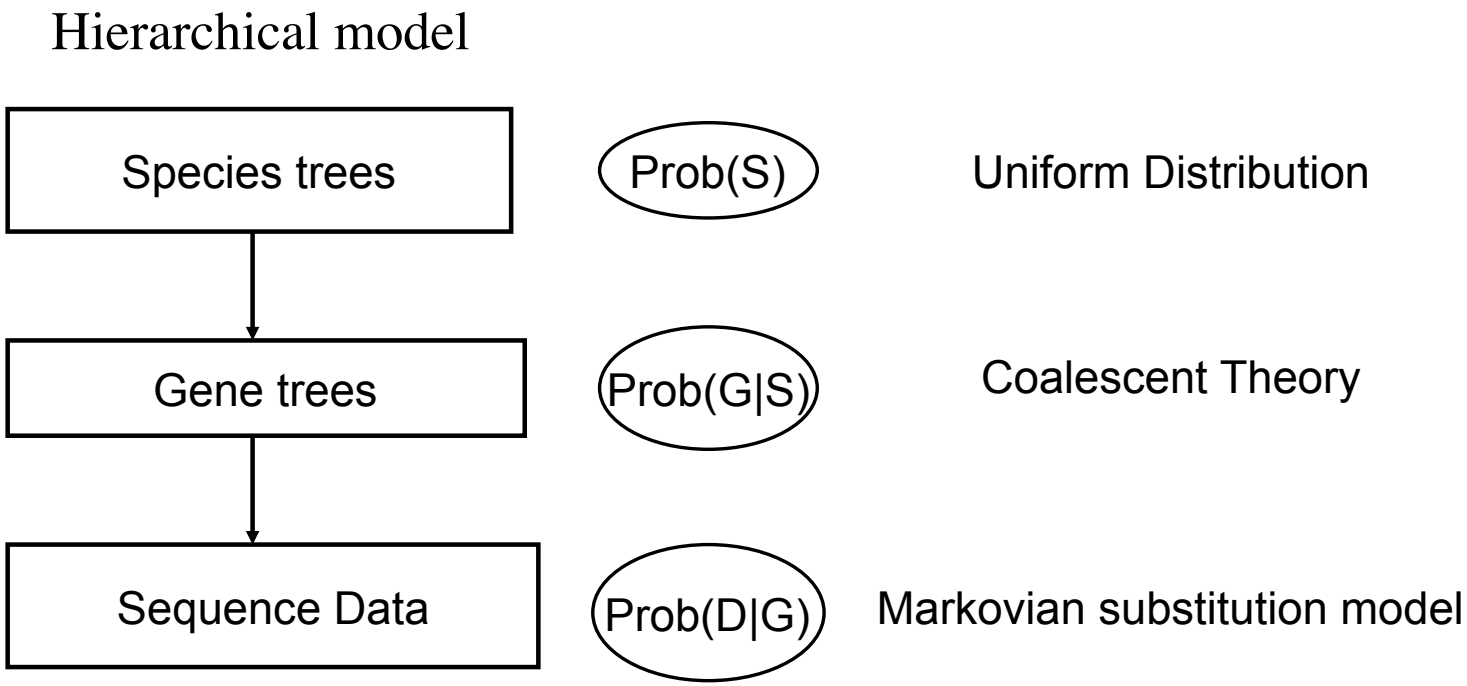
  Species tree $\rightarrow$ gene trees [coalescent process]

  Gene trees $\rightarrow$ sequence data [standard nucleotide substitution models]

- Goal of both methods is to estimate the posterior distribution of the species tree and associated model parameters

- BEST and *BEAST use slightly different algorithms – we will briefly discuss the main ideas of each

BEST: Overview of Method



Source: Lecture 7, Stat 882 Spring 2010, Dennis Pearl

BEST: Overview of Method

- Assumptions:
  - Given the species tree, the gene trees are conditionally independent.
  - Given the gene tree, the DNA sequences are conditionally independent of the species tree.
  - Random mating in each population.
  - No gene flow after species divergence.
  - No recombination within a locus.

Source: Lecture 7, Stat 882 Spring 2010, Dennis Pearl

BEST: Overview of Method

- BEST uses MCMC to sample from the joint posterior distribution of the gene trees and the species tree:

$$f(S, \mathbf{G}|D) = \frac{f(D|\mathbf{G})F(\mathbf{G}|S)f(S)}{f(D)}$$

Source: Lecture 7, Stat 882 Spring 2010, Dennis Pearl

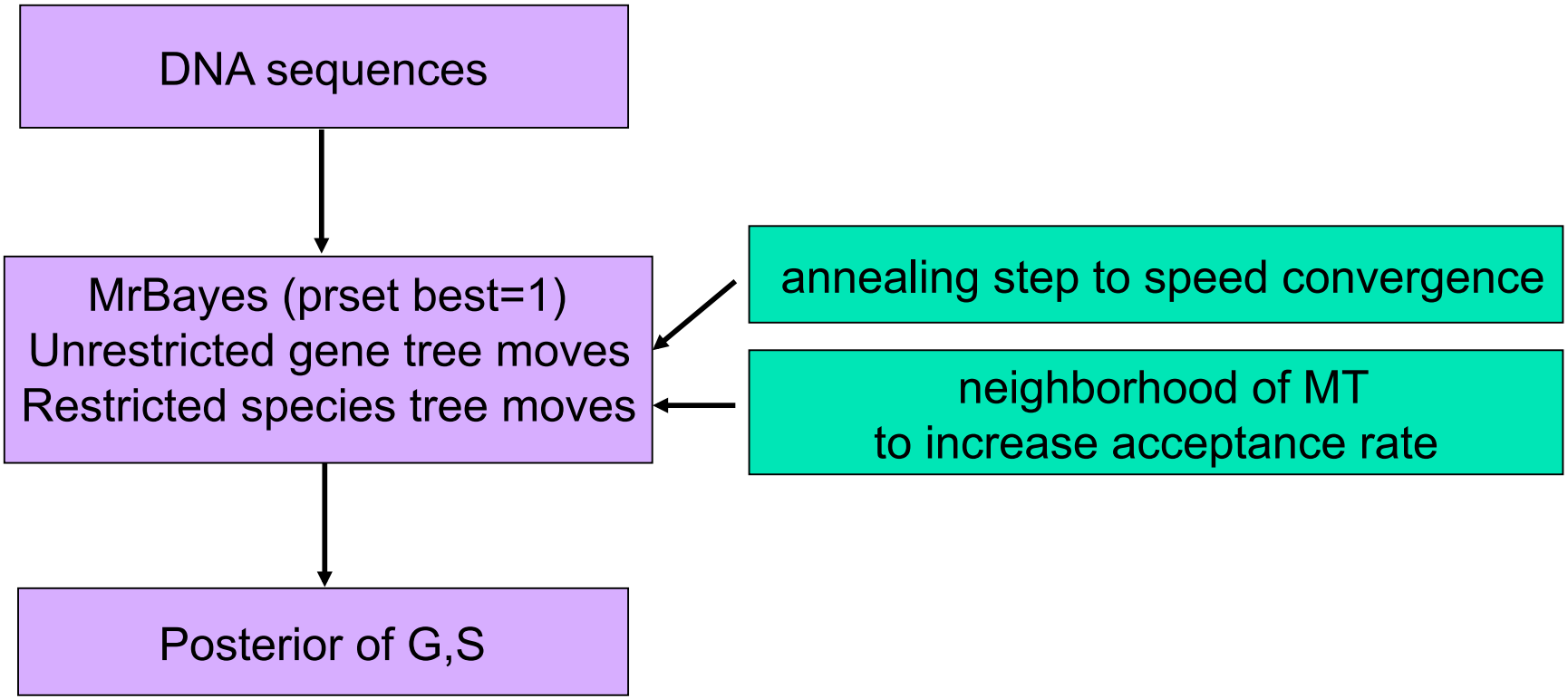
BEST: Overview of Method

- Implementation: MrBayes with BEST
  - Step 1: Use MrBayes to propose vectors of joint gene trees (unlinked).
  - Step 2: Given those gene trees: propose a compatible species tree.
  - Step 3: Implement the chain fully within MrBayes using the usual properties of the MCMC as proposed by the user.

Source: Lecture 7, Stat 882 Spring 2010, Dennis Pearl

BEST: Overview of Method

- BEST algorithm:



Source: Lecture 7, Stat 882 Spring 2010, Dennis Pearl

*BEAST: Overview of Method

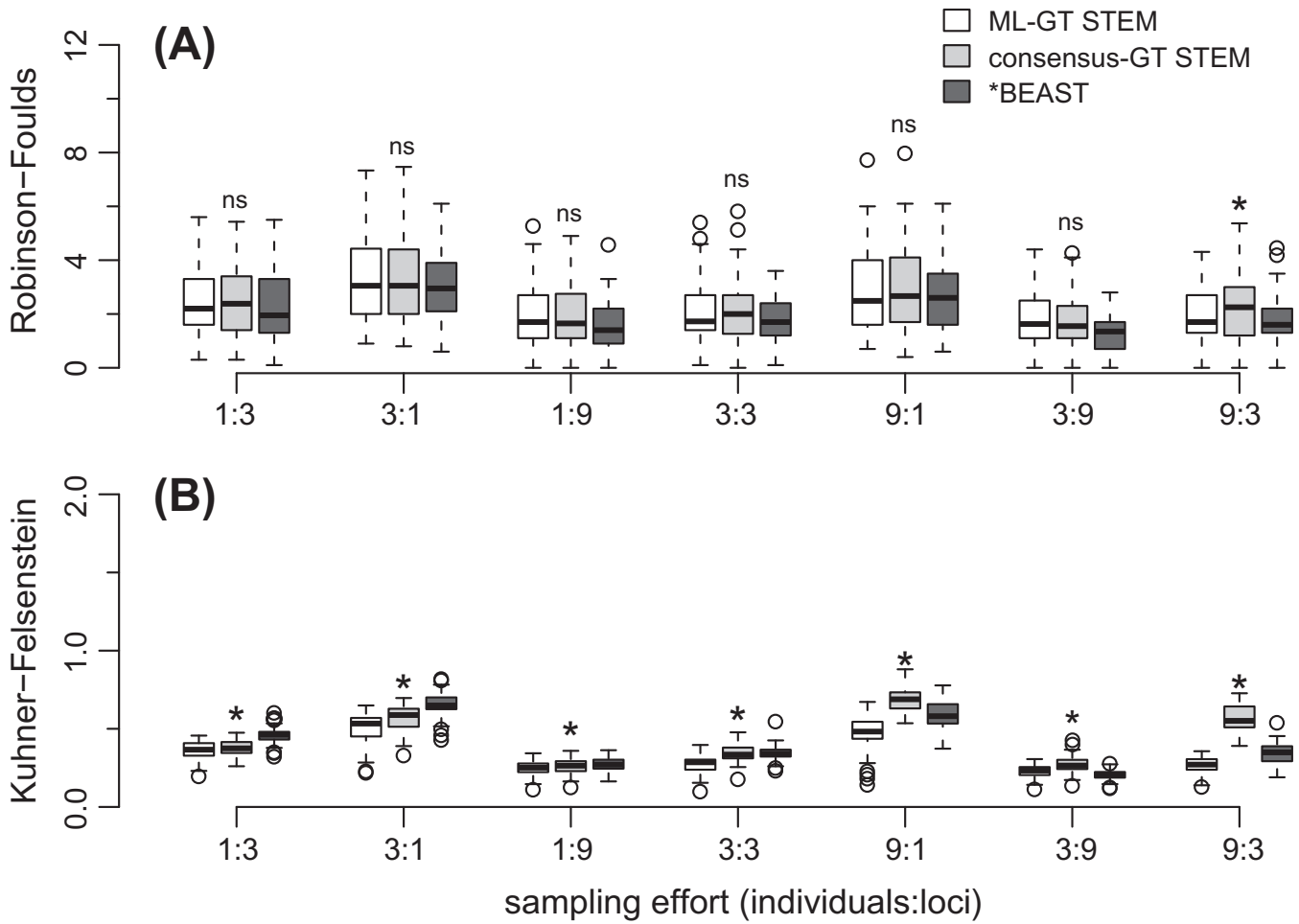
- The MCMC algorithm in *BEAST also samples both gene trees and the species tree at each step of the algorithm $\Longrightarrow$ same goal as BEST (to get posterior distribution of species tree).
- The method of moving in gene trees is one reason that *BEAST works more quickly than BEST.
- Authors have pointed out substantial increases in speed [we'll discuss an empirical example a little later].

Full Data Methods: BEST and *BEAST

- Both methods use MCMC, which means:
  - Need to think carefully about setting prior distributions.
  - Need to check carefully for convergence, in all parameters
  - Need to choose methods of summarizing the estimated posterior distribution carefully, and interpret these summaries correctly
  - etc., ...

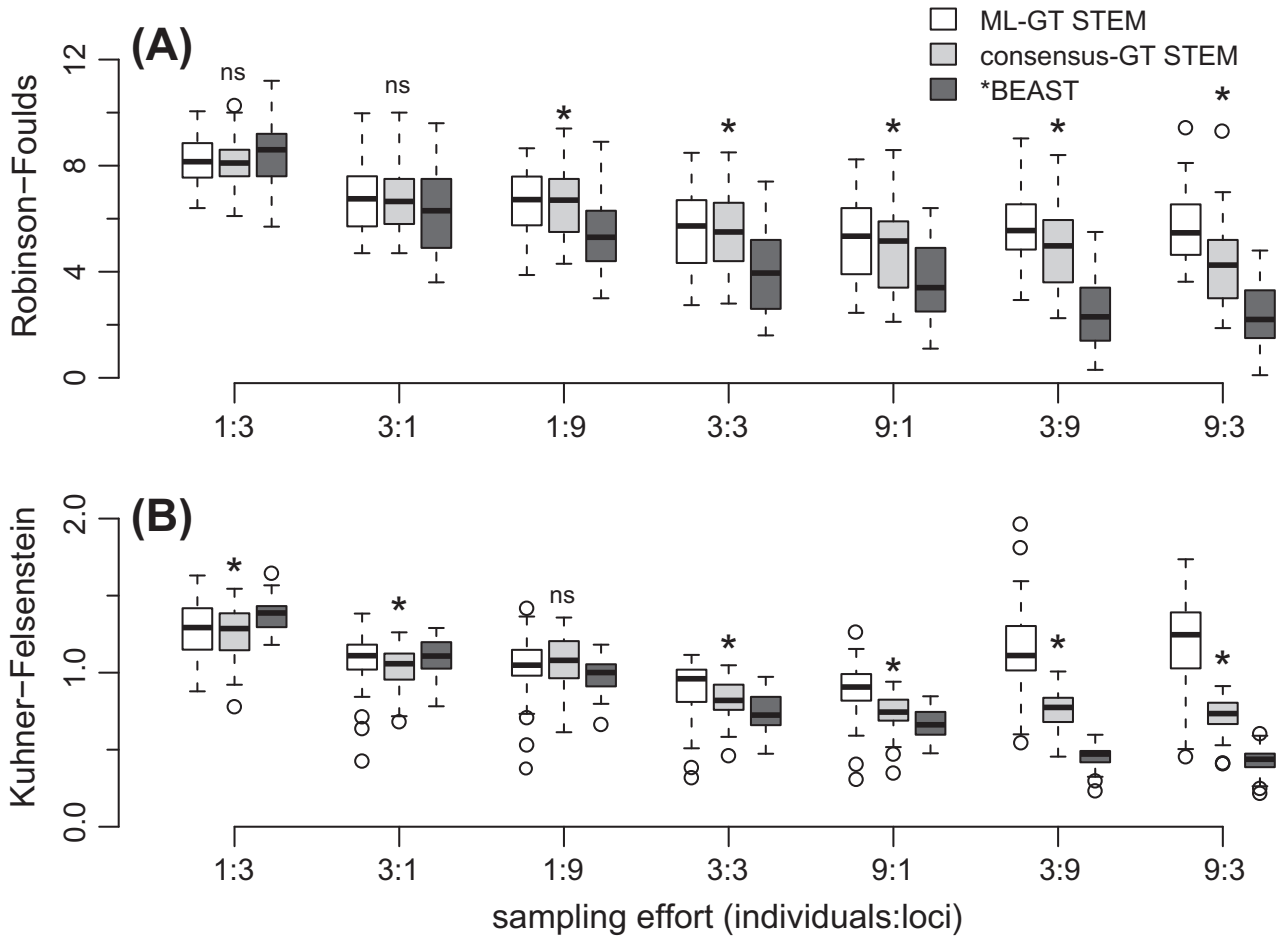
Comparison of STEM with *BEAST (Knowles et al. 2012)

Total tree depth 10N



Comparison of STEM with *BEAST (Knowles et al. 2012)

Total tree depth 1N

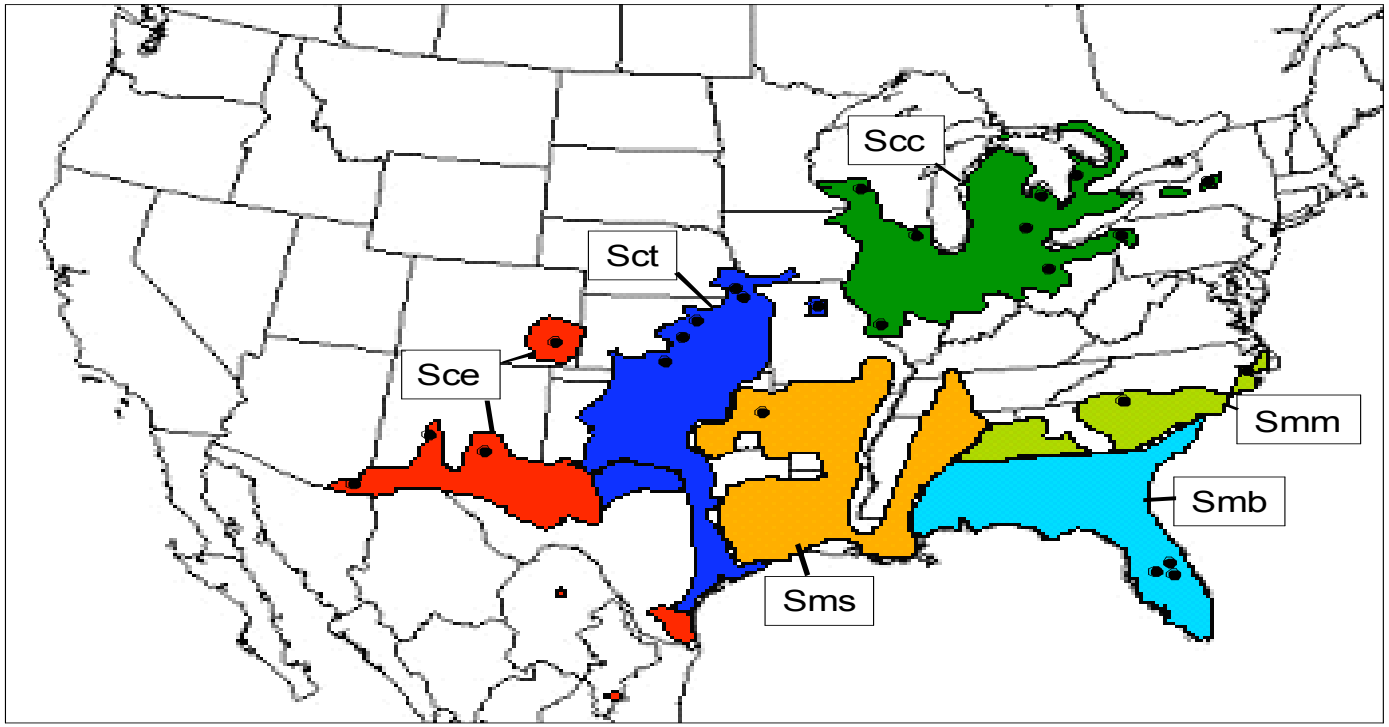


Putting it all together: An empirical example

- 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



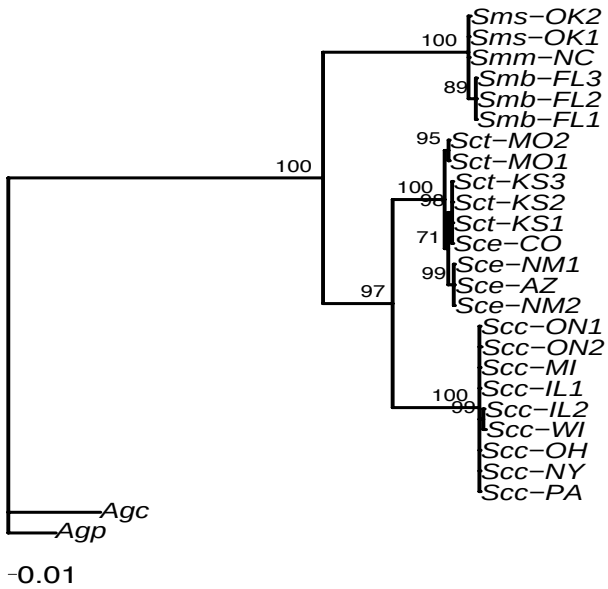
Sistrurus rattlesnakes

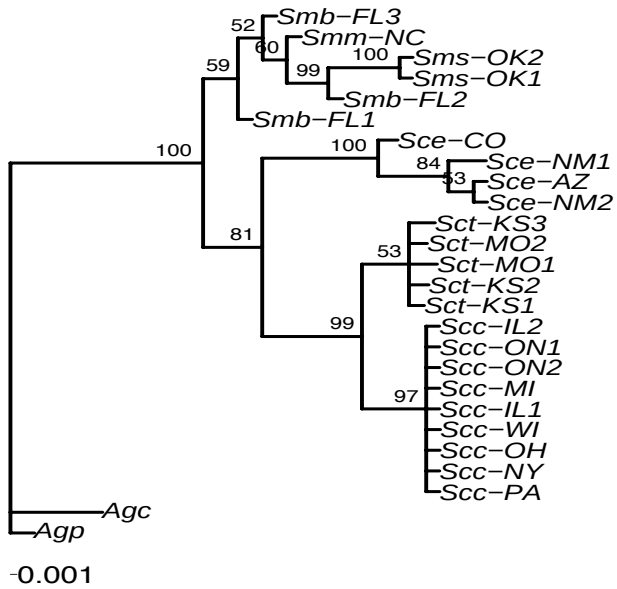
- Data: 7 (sub)species, 26 individuals, 19 genes

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

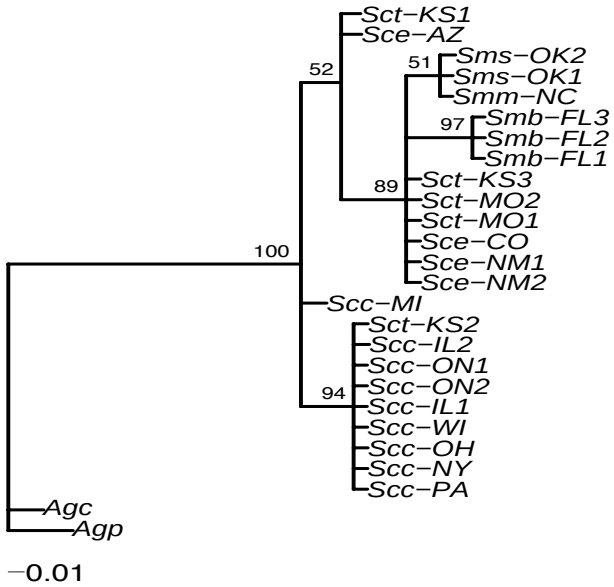
Some are very informative:

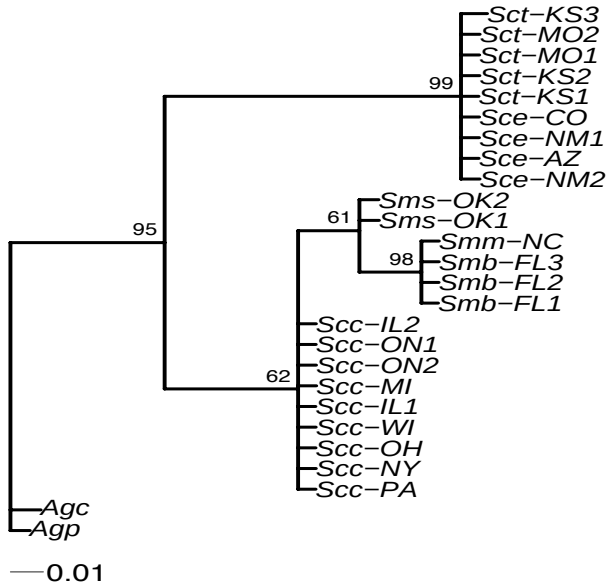




Individual gene tree estimates

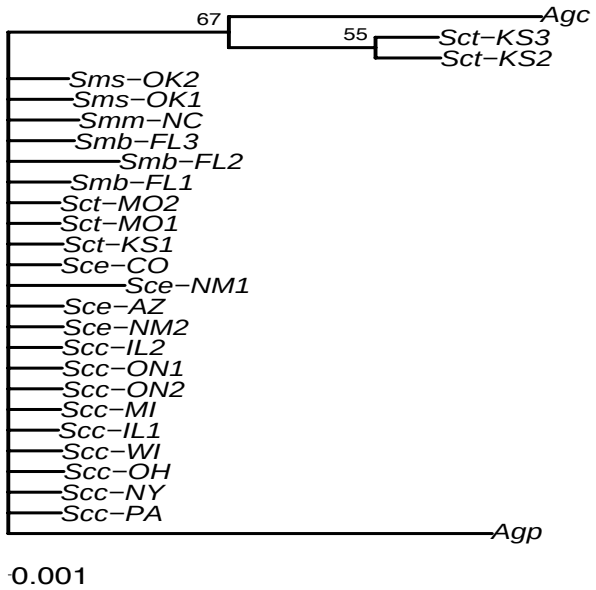
Some are a little informative:

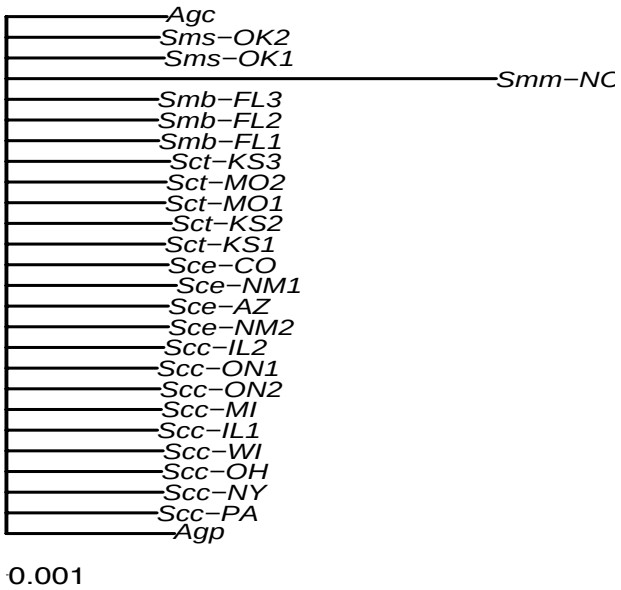




Individual gene tree estimates

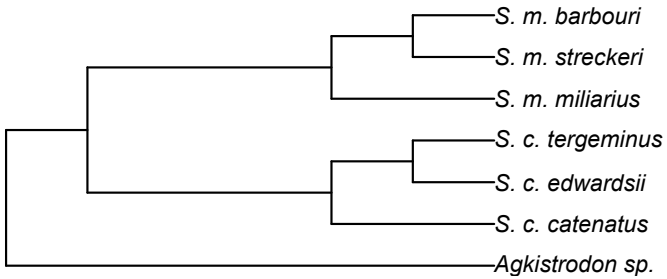
And then there are others .....



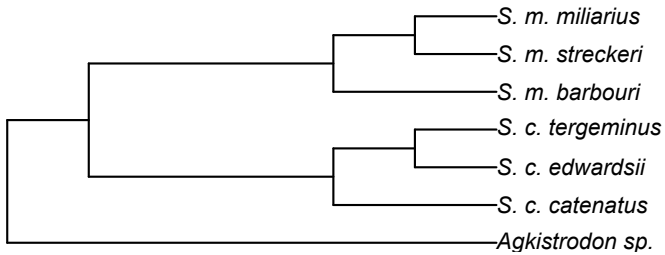


Example: Sistrurus rattlesnakes ... species tree estimation

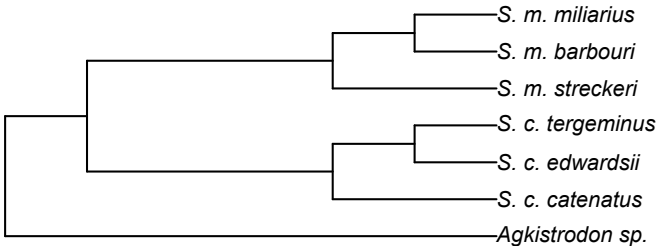
STEM, STEAC



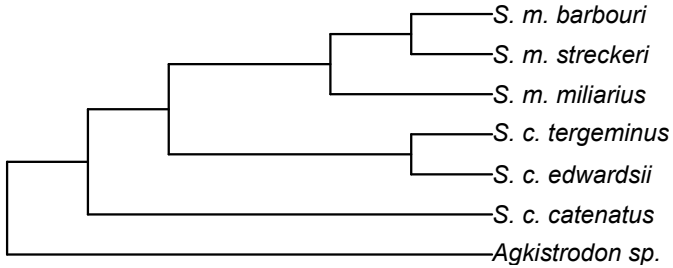
BEAST (concatenated data), *BEAST



BEST, Parsimony & MrBayes (concatenated data)



PhyloNet, STAR



- Some observations:
  - Estimate from PhyloNet places *S. c. catenatus* as sister to the entire clade – it turns out this is due to only two gene trees. If those genes are removed, the estimate agrees with STEM
  - The portion of the tree that differs between STEM, *BEAST, and BEST is the arrangement of the *S. miliarius* subspecies – all three arrangements are observed
  - **Both** BEST and *BEAST have trouble converging: BEST did not converge in the branch length parameters, while *BEAST did not converge in the effective population size parameters, especially for the tip species (same problem?)
  - *BEAST was much faster than BEST (days vs. months for $\sim$ 350 million iterations) – but with an older version of BEST.

Full data methods (2): SNAPP

- SNAPP is a recently-proposed method that bypasses the need to have gene trees explicitly specified at any stage in the algorithm.
- Recall again the Felsenstein equation:

$$L(S|D_1, D_2, ...D_N) = \prod_{i=1}^{N} \sum_{j=1}^{G} P(D_i|g_j) f(g_j|S)$$

- SNAPP uses a clever two-step peeling algorithm to carry out the integration over gene trees.
- Notes:
  - Use MCMC for inference, but gains efficiency due to using only the space of species trees.
  - Currently limited to biallelic SNP data – necessary to make the algorithm for computing the likelihood via integration over gene trees feasible.
  - Can also handle AFLP data.

Full data methods (3): SVDquartets

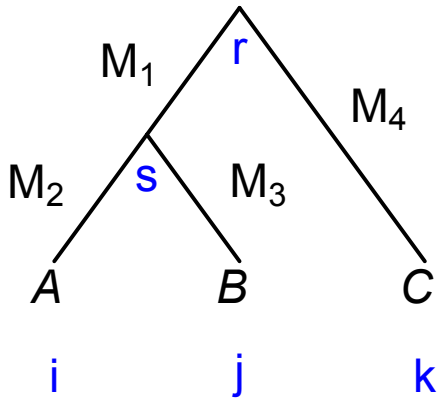
- Motivation: Recall "cartoon time" from Mark's talk yesterday. He noted the connection between trees and the pattern frequency space.
- Data: DNA sequences for gene *i*
- Example:

Taxon	Sequence
(A) Human	GCCGATGCCGATGCCGAA
(B) Chimp	GCCGTTGCCGTTGCCGTT
(C) Gorilla	GCGGAAGCGGAAGCGGAA

- Data: DNA sequences for gene i, D_i
- Example:

Taxon	Sequence		
(A) Human	GCCG	A	TGCCGATGCCGAA
(B) Chimp	GCCG	T	TGCCGTTGCCGTT
(C)Gorilla	GCGG	A	AGCGGAAGCGGAA

- Assume each site in the sequence evolves independently of other sites
- Data are assumed to be an iid sample of sites:
  $(D_i)_j$ = data at the tips of the tree for site j in gene i
- Consider site pattern probabilities – for example, p_{ATA}



- Let T be an n-leaf, rooted, binary tree with distribution of states $\Pi = (\pi_1, \pi_2, \cdots \pi_k)$ at the root
- Edges e of T are labeled by $k \times k$ transition probability matrices M_e that give the probabilities of changes in state from a node to its child
- Let X_H be the state of taxon H

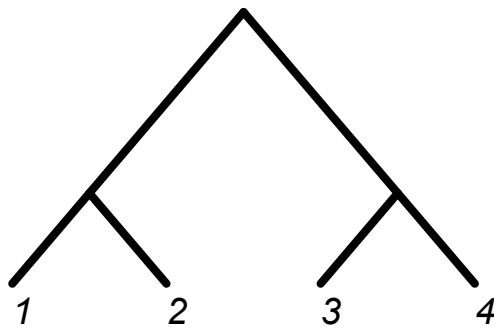
- Together $(T, \{M_e\}, \Pi)$ define the joint distribution at the leaves of tree:

$$\begin{aligned} p_{ijk} &= P(X_A = i, X_B = j, X_C = k) \\ &= \sum_r \sum_s \pi_r M_1(r,s) M_2(s,i) M_3(s,j) M_4(r,k) \end{aligned}$$

- This gives site pattern probability distribution on a gene tree.
- We want the site pattern probability distribution on a species tree under the coalescent model.
- Used Mathematica to get the entire site pattern probability distribution for 4 taxa under the Jukes-Cantor model.

Definition: splits

- **Definition:** A split of a set is a bipartition of the set of taxa into two groups. A split $A|B$ of the leaves of a tree T is valid for T if the induced tree $T|_A$ and $T|_B$ do not intersect.



- Valid: $12|34$
- Not valid: $13|24$

Definition: flattenings

$$p_{ijkl} = P(X_1 = i, X_2 = j, X_3 = k, X_4 = l)$$

$$Flat_{12|34}(P) = \begin{pmatrix} p_{AAAA} & p_{AAAC} & p_{AAAG} & p_{AAAT} & p_{AACA} & \cdots \\ p_{ACAA} & p_{ACAC} & p_{ACAG} & p_{ACAT} & p_{ACCA} & \cdots \\ p_{AGAA} & p_{AGAC} & p_{AGAG} & p_{AGAT} & p_{AGCA} & \cdots \\ p_{ATAA} & p_{ATAC} & p_{ATAG} & p_{ATAT} & p_{ATCA} & \cdots \\ p_{CAAA} & p_{CAAC} & p_{CAAG} & p_{CAAT} & p_{CACA} & \cdots \end{pmatrix}$$

Theorem (Chifman and Kubatko 2012):

- If $A|B$ is a valid split for T, the $rank(Flat_{A|B}(P)) \leq 10$.
- If $C|D$ is not a valid split for T, then generically $rank(Flat_{C|D}(P)) = 16$.

What does this mean?

- When the Flat matrix is constructed for a true tree, some columns will be "duplicates" of the others in some sense.
- When the Flat matrix is constructed for a false tree, all column are independent of the others.
- We do not have the p_{ijkl} directly (they come from the true underlying tree with its branch lengths and substitution model parameters), but we can estimate them with data.
- Idea: Construct an estimate of the Flat matrix, $\hat{Flat}(P)$, and use a measure of whether all columns are independent. We use singular value decomposition (SVD).

- Would like to use these ideas to estimate a species tree when given multi-locus data for L genes, $D_1, D_2, \ldots D_L$
- Issue 1:
  - The model assumes each site has its own gene tree, i.e.,
    $(D_1)_1$ arises from gene tree $(G_1)_1$, $(D_1)_2$ arises from $(G_1)_2$, etc. ...
    $(D_2)_1$ arises from gene tree $(G_2)_1$, $(D_2)_2$ arises from $(G_2)_2$, etc. ...
    ...
    $(D_L)_1$ arises from gene tree $(G_L)_1$, $(D_L)_2$ arises from $(G_L)_2$, etc. ...
  - Multilocus phylogenetics generally assumes that each gene has a single underlying tree, i.e.,
    $(D_1)_1, (D_1)_2, \ldots$ arise as iid observations from G_1
    $(D_2)_1, (D_2)_2, \ldots$ arise as iid observations from G_2
    ...
    $(D_L)_1, (D_L)_2, \ldots$ arise as iid observations from G_L

Goal 2: Species tree inference

- Issue 2:
  - Appropriate for SNPs
  - May need to worry about acertainment – SNP data commonly include only variable sites
  - For our example data:

Taxon	Sequence
(A) Human	GCCGATGCCGATGCCGAA
(B) Chimp	GCCGTTGCCGTTGCCGTT
(C) Gorilla	GCGGAAGCGGAAGCGGAA

this would be

Taxon	Sequence
(A) Human	CATCATCAA
(B) Chimp	CTTCTTCTT
(C) Gorilla	GAAGAAGAA

Simulation study – can we detect the correct split?

Simulate data from the Jukes-Cantor model for a 4-taxon tree and examine split scores

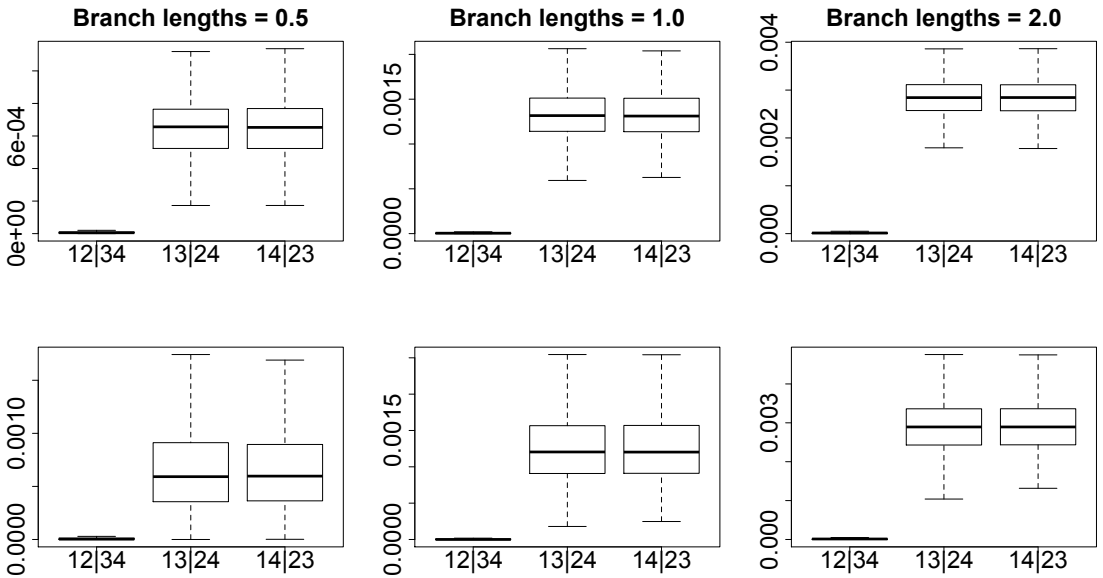


Fig. 2. Simulation results for the JC model. The top row gives results for 5,000 SNP sites and the bottom row gives the results for 10 genes with 500 sites each. The columns correspond to differing branch lengths in the model species tree.

Simulation study – can we detect the correct split?

Simulate data from the GTR+I+Γ model for a 4-taxon tree and examine split scores

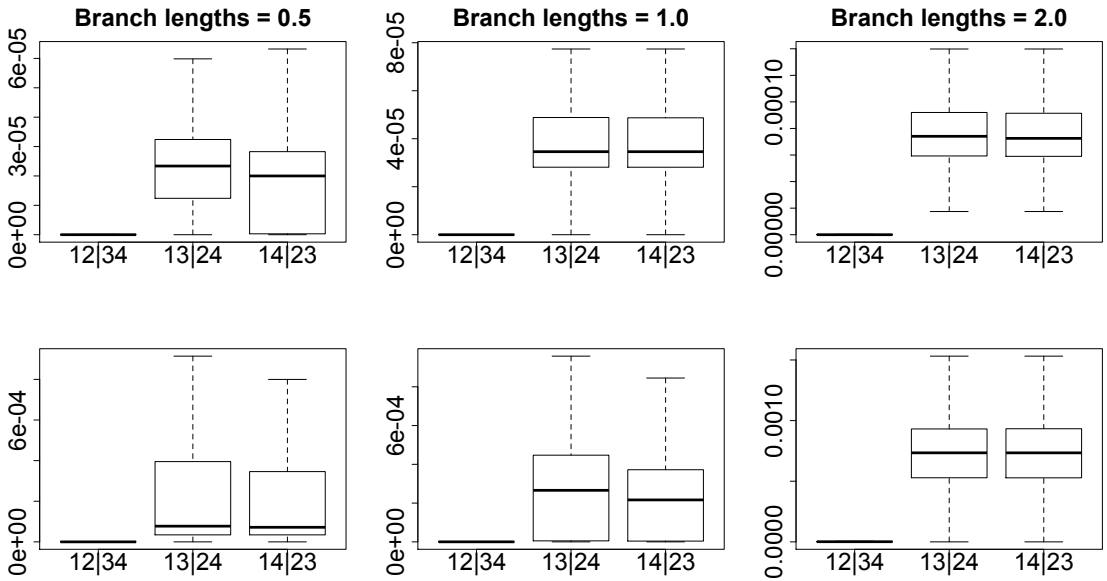
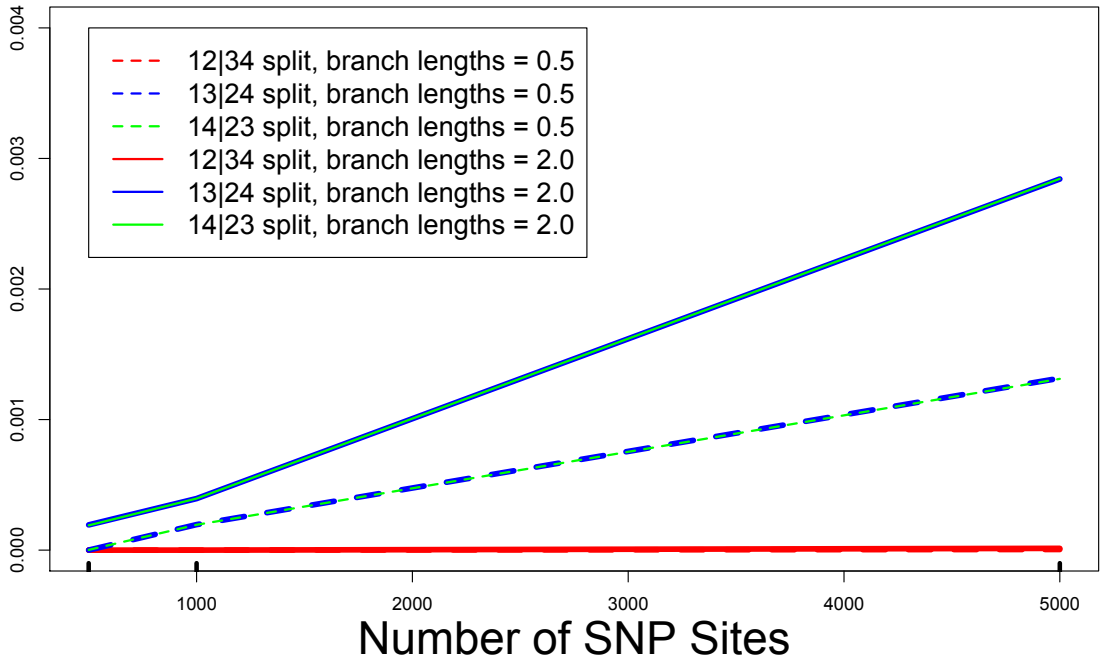


Fig. 3. Simulation results for the GTR+I+Γ model. The top row gives results for 5,000 SNP sites and the bottom row gives the results for 10 genes with 500 sites each. The columns correspond to differing branch lengths in the model species tree.

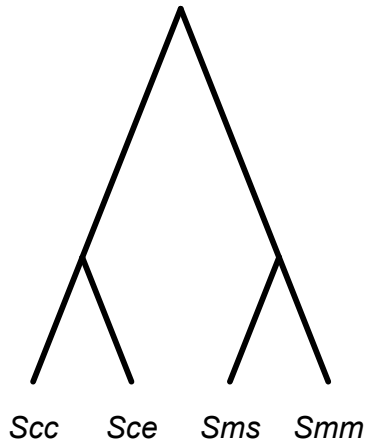
Simulation study – can we detect the correct split?

Change in scores as amount of data increases

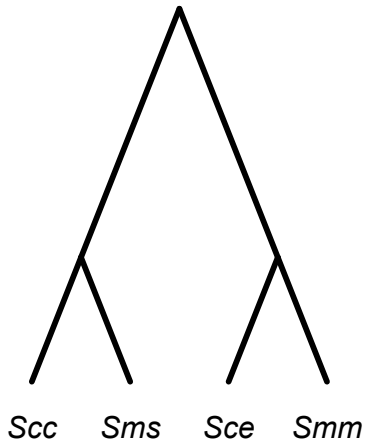


Apply the method to the rattlesnake example

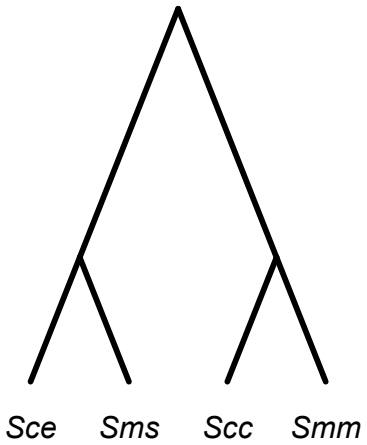
19 genes, ∼8500 bp



SVD score = 3.84 x 10^-14



SVD score = 5.395



SVD score = 5.396

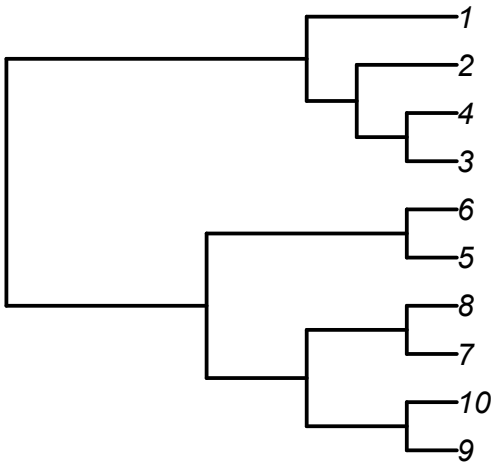
Extending beyond 4 taxa

- The only theoretical results thus far are for 4 taxa, but the use of flattenings in the single gene case extends beyond 4 taxa.
- A straightforward method is to consider all (or a sample of) quartets, and use a quartet-method to reconstruct the species tree.
- Use more simulation studies to see how it works.

Simulation study

- Simulation design:
  1. Generate gene trees from a model species tree [Using COAL]
  2. Generate sequence data from each gene tree under the GTR+I+Γ model [Using Seq-Gen]
  3. Sample K sets of 4 taxa at random, and compute the score based on the flattening matrix for the three possible splits of these 4 taxa for each sample [Easy to code]
  4. Use the inferred quartet relationships to construct the tree [Quartet MaxCut (Snir and Rao, 2010)]
- For a 10-taxon species tree and $K = 500$, it takes $\approx$ 8 seconds to obtain the species tree estimate.

Simulation study results - percent of time correct topology is estimated



black = 500 bp / gene
red = 2,000 bp / gene
blue = No. genes × 500 SNPs

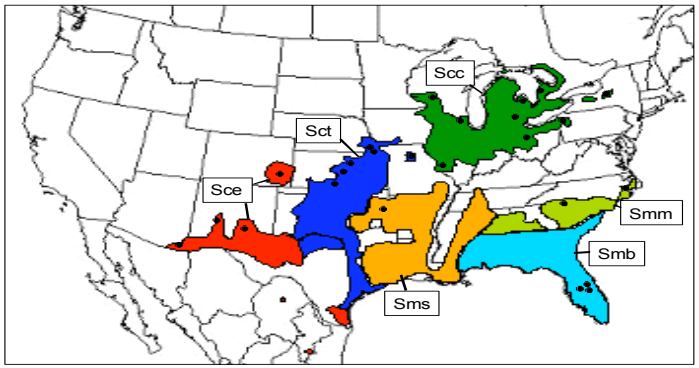
	10 genes	20 genes	50 genes	100 genes
Short (0.5)	13 16 17	16 30 44	60 100 87	81 100 92
Medium (1.0)	42 68 70	69 86 92	91 100 96	92 100 93
Long (2.0)	82 93 86	95 90 91	92 100 94	94 100 94

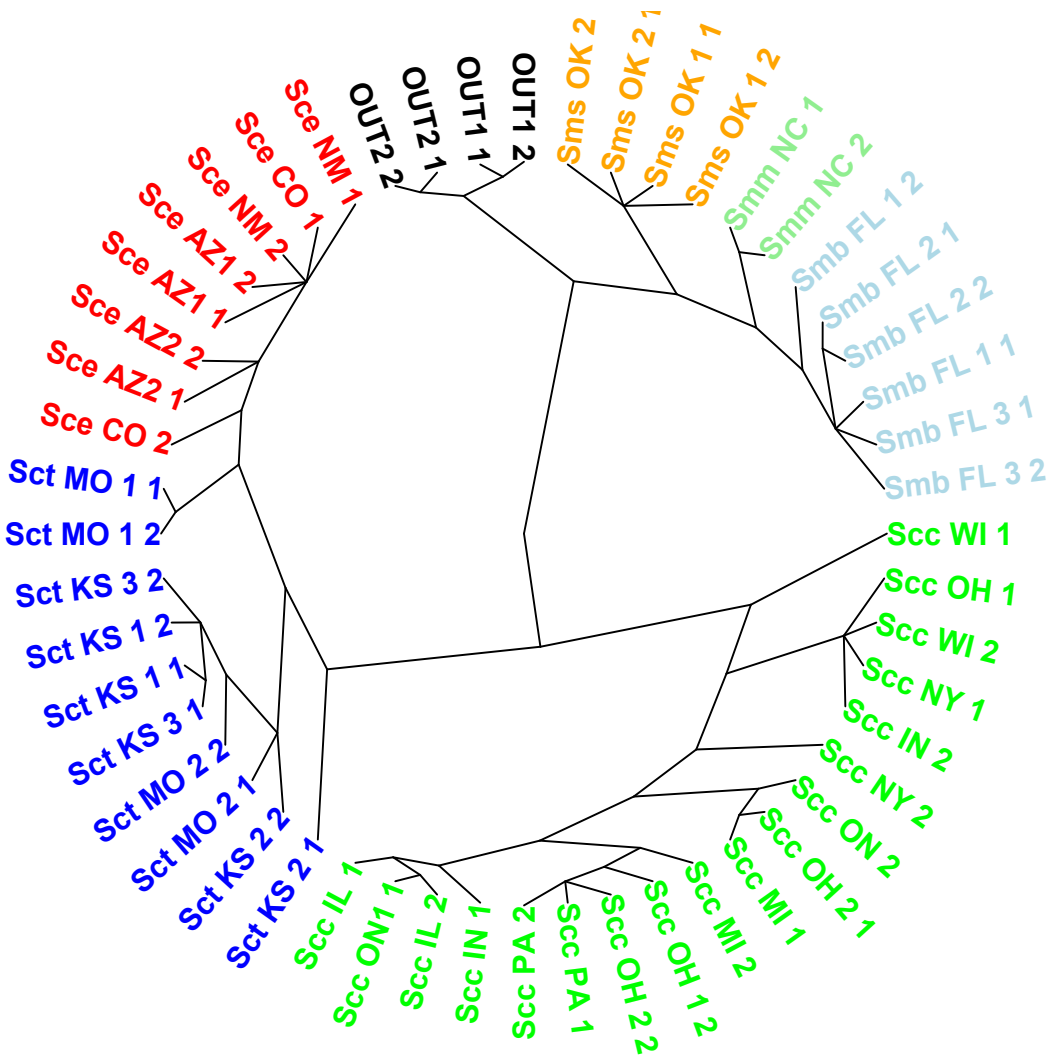
Empirical example: Sistrurus rattlesnakes

Using 3,000 quartets

Quartet MaxCut (Snir and Rao, 2010)

$\approx$ 10 minutes





- Software:
  - Currently, the program to estimate quartet relationships is freely available: http://www.stat.osu.edu/∼lkubatko/software/SVDquartets/
  - Output from this program can be used with any quartet assembly software to produce species trees. We'll try an example in the lab tonight.
- Pros and cons:
  - Quick!!
  - Intuitive way of handling missing data, with possible extensions.
  - Estimates topology only — no branch lengths, population parameters, etc.

BCA: Overview of Method

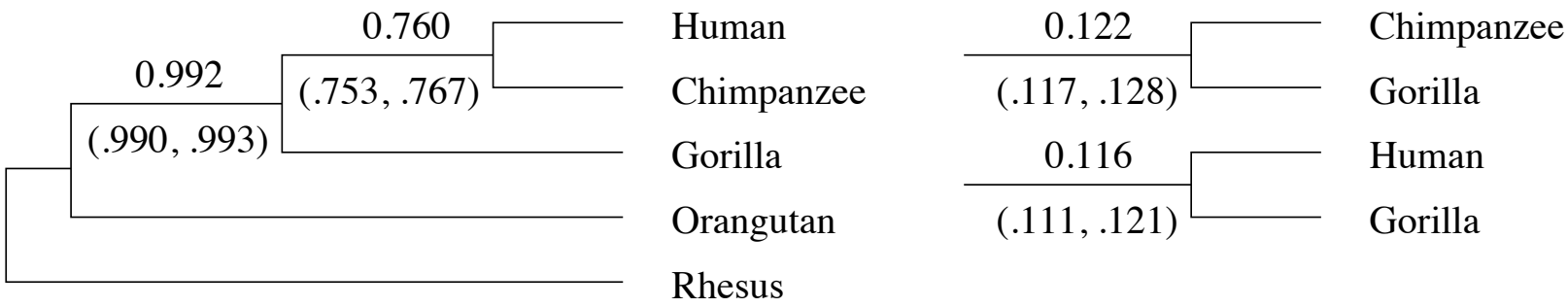
- What about methods that do not assume the coalescent process?
- Bayesian Concordance Analysis (BCA) – implemented in BUCKy (Bayesian Untangling of Concordance Knots)
- Idea: Estimate the proportion of the genome that has a certain clade
- Build a tree that consists of those clades that are inferred to be true for a high proportion of the genome

BCA: Overview of Method

- First step: Run MrBayes to get estimated posterior distribution for each gene.
- Run BCA, a second MCMC step which utilizes the individual gene posterior distributions as input.
- The BCA method clusters genes into a number of groups, so that genes in the same group are assumed to share the same gene tree.

- The user specifies a prior distribution on the number of groups. This controls how much discordance among gene trees is expected.
  - At one extreme, all gene trees are assumed to be the same, and the method mimics concatenation.
  - At the other, all gene trees are assumed to be completely independent, and the method mimics a consensus method.
- Goal is to estimate concordance factors for all possible clades – percent of genes for which that clade is true. These are often displayed in a primary concordance tree.

- Results using the Ebersberger et al. data we looked at earlier:

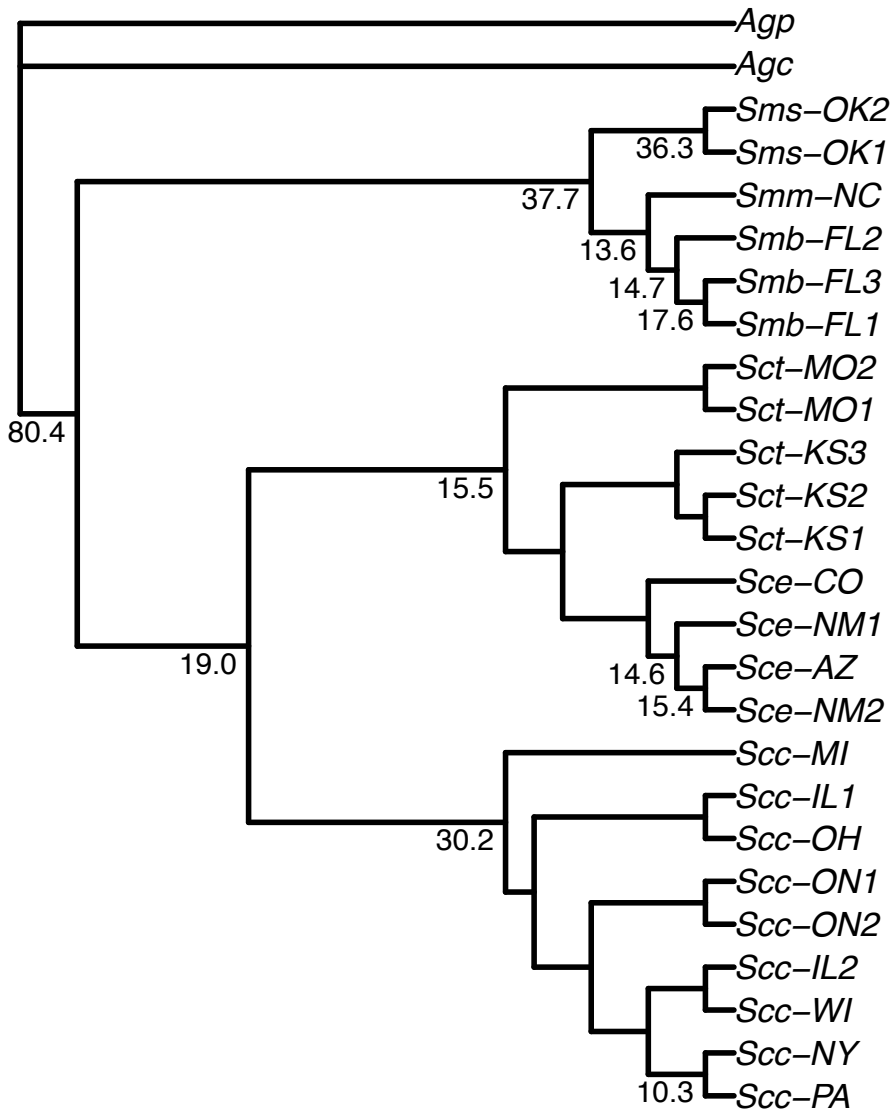


From Ane (2011)

- Numbers above nodes are genome-wide concordance factors
- Intervals below nodes are 95% credibility intervals.

BCA: Application to Rattlesnakes

- The Primary Concordance Tree from BUCKy



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. – get involved and have fun!
- Find me and tell me about your data!

- What next?
- Options:
  - Questions and discussion
  - STEM/STEM-hy demo and lab warm-up

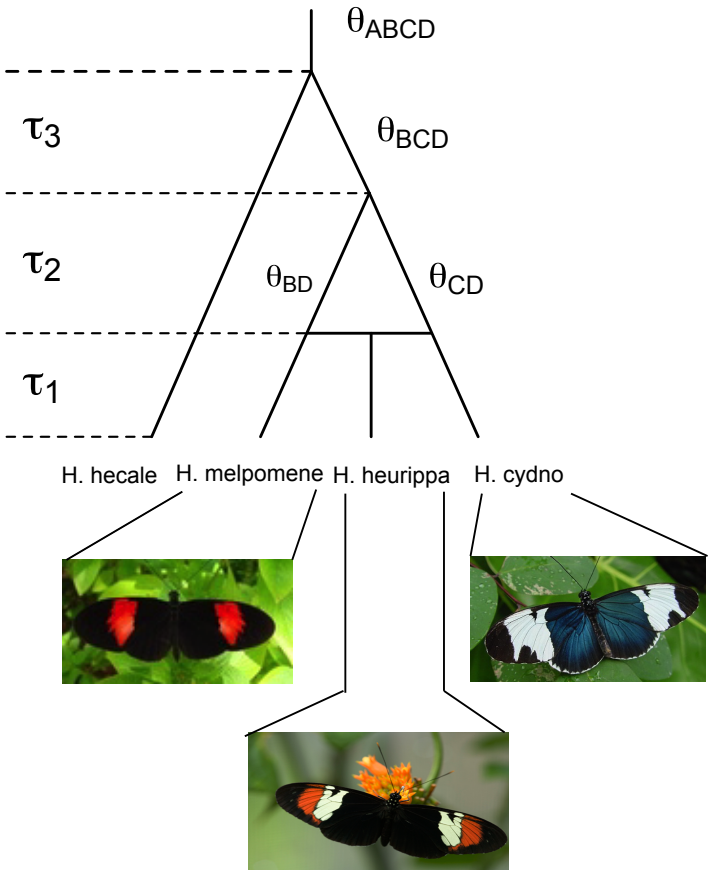
STEM Data Preparation - Gene Trees

- STEM takes as its input one gene tree for each locus.
- Thus, a first step in an analysis using STEM is to estimate gene trees with branch lengths for each locus.
- Any method can be used to do this, but note a couple requirements:
  - Branch lengths are assumed to be in units of expected number of substitutions per site per unit time.
  - Branch lengths must be estimated subject to a molecular clock. This is not checked by the program.
  - Gene trees must be fully resolved; however, polytomies can be included by setting branch lengths to 0 for an arbitrary resolution of the polytomy.

- A value of the parameter $\theta = 4N\mu$ must be provided. Note that this is the "per-site θ", not a "per-locus" value as used by other population genetics programs.
- This will be used to convert gene tree branch lengths to coalescent units by dividing all gene tree branch lengths by θ.
- Estimates of θ could be obtained by standard methods. Typical values of θ will be between 0.001 and 0.1.
- The species tree estimate is returned with branch lengths in coalescent units.

- Each locus can also be given a rate multiplier.
- These can adjust for
  - Variation in mutation rate across loci.
  - Ploidy (e.g., haploid loci – mtDNA – should be given a rate of 0.5).
- At the least, one should estimate rate variation from the data by something like the following:
  - Compute average pairwise sequence divergence of each sequence to the outgroup.
  - Divide all of these values by their overall mean, and assign that number as the rate multiplier for each gene.
  - Adjust specific genes for ploidy, if necessary.

Example: Heliconius butterflies



STEM example

- Example genetrees.tre file:

```
[0.37137]((Hheurippa:0.005989,(Hcydno:0.001322,Hmelpomene:0.001322):0.004667):0.022778,Hhecale:0.028767);
[1.17059]((Hmelpomene:0.049843,(Hcydno:0.000001,Hheurippa:0.000001):0.049843):0.001,Hhecale:0.049943);
[0.11434](((Hcydno:0.021024,Hheurippa:0.021024):0.020051,Hmelpomene:0.041076):0.002610,Hhecale:0.043685);
[1.35454](((Hheurippa:0.010740,Hcydno:0.010740):0.003498,Hmelpomene:0.014238):0.037654,Hhecale:0.051892);
[0.39096](((Hheurippa:0.008764,Hmelpomene:0.008764):0.001686,Hcydno:0.010450):0.003969,Hhecale:0.014419);
[1.22683](((Hheurippa:0.002431,Hcydno:0.002431):0.062919,Hmelpomene:0.065350):0.0000001,Hhecale:0.065351);
```

- Five main functions of STEM-hy version 1.0
  - **Estimate a species tree given a set of gene trees using maximum likelihood**
    run=1
  - Search species tree space for trees of high likelihood
    run=2
  - Compute the likelihood of a user-specified tree
    run=0
  - Carry out a bootstrap analysis (bootstrapping is on sites within genes)
    run=4
  - **Assess fit of trees subject to hybridization in the presence of lineage sorting**
    run=3

STEM example

- Example settings file:

```
properties:
                run: 1
                theta: 0.001
                beta: 0.0005
                burnin: 100
                seed: 3435893
                bound_total_iter: 20
                num_saved_trees: 10
                hybrid_species: H._heurippa
                hybrid_tree: user-heliconius.tre
species:
                H._melpomene: Hmelpomene
                H._hecale: Hhecale
                H._cydno: Hcydno
                H._heurippa: Hheurippa
```

STEM example

- To run STEM-hy:

  java -jar STEM-hy.jar

STEM example

**********Results**********

D_AB Matrix:

[0.00000 1.05113 3.68810 0.35598]
[0.00000 0.00000 3.68810 0.00009]
[0.00000 0.00000 0.00000 3.68810]
[0.00000 0.00000 0.00000 0.00000]

Maximum Likelihood Species Tree (Newick format):

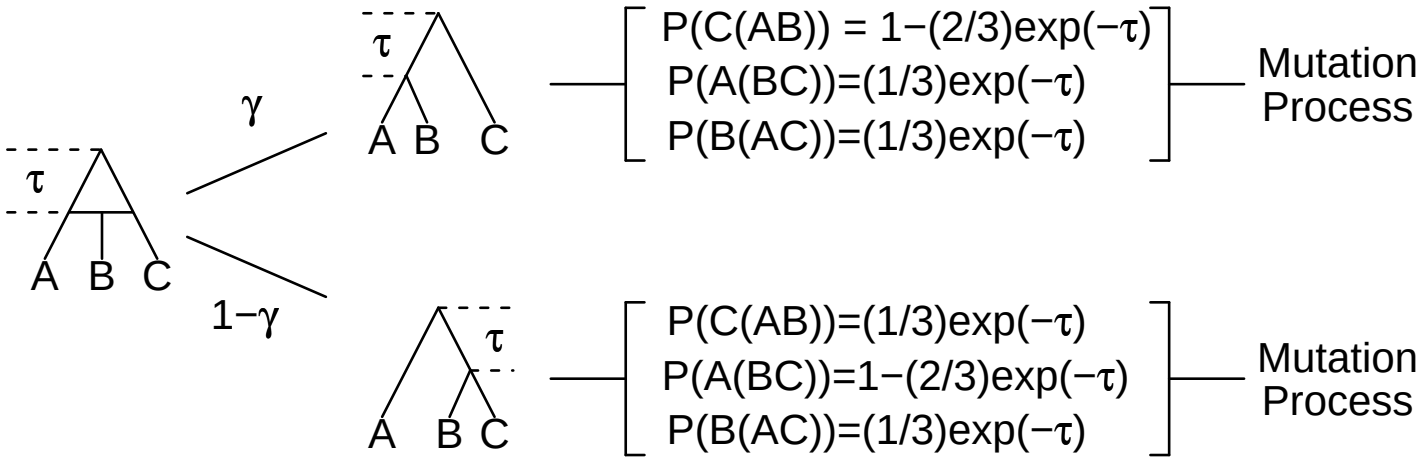
(H._hecale:3.68810,(H._melpomene:0.35598,(H._heurippa:0.00009,H._cydno:0.00009):0.35589):3.33212);

log likelihood for tree: -349.9185707499209

********** Done **********

STEM example

STEM's hybridization model



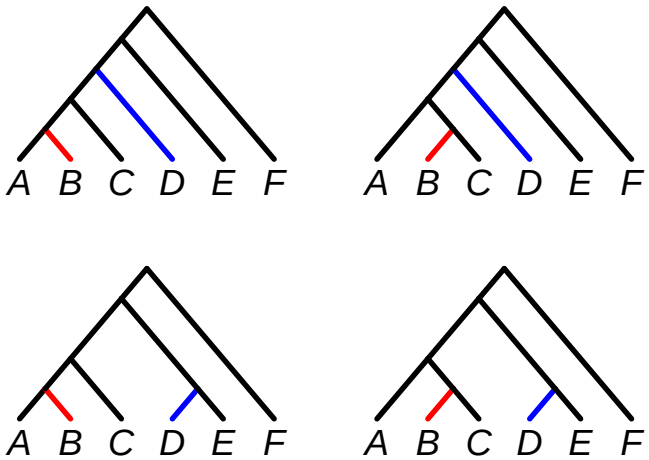
STEM example

Assumptions:

- Hybridization results in a mosaic genome, so that a sampled gene has a probability distribution that its history originated from one of several parental species trees
- Genes in the sample are independent given the species tree
- Hybridization events happen only between sister taxa
- No factors other than coalescence and hybridization lead to incongruence between gene trees and the species tree

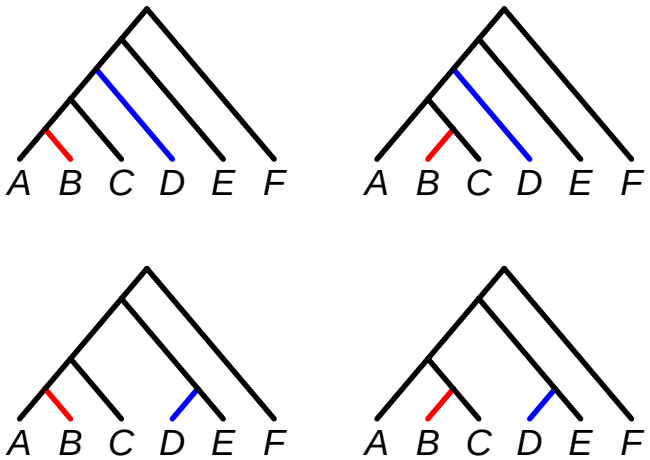
A bigger example

Motivating example:

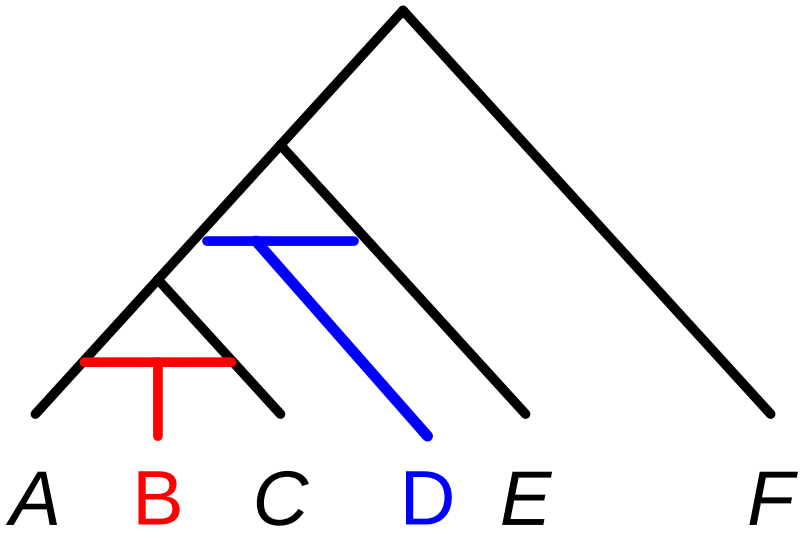


A bigger example

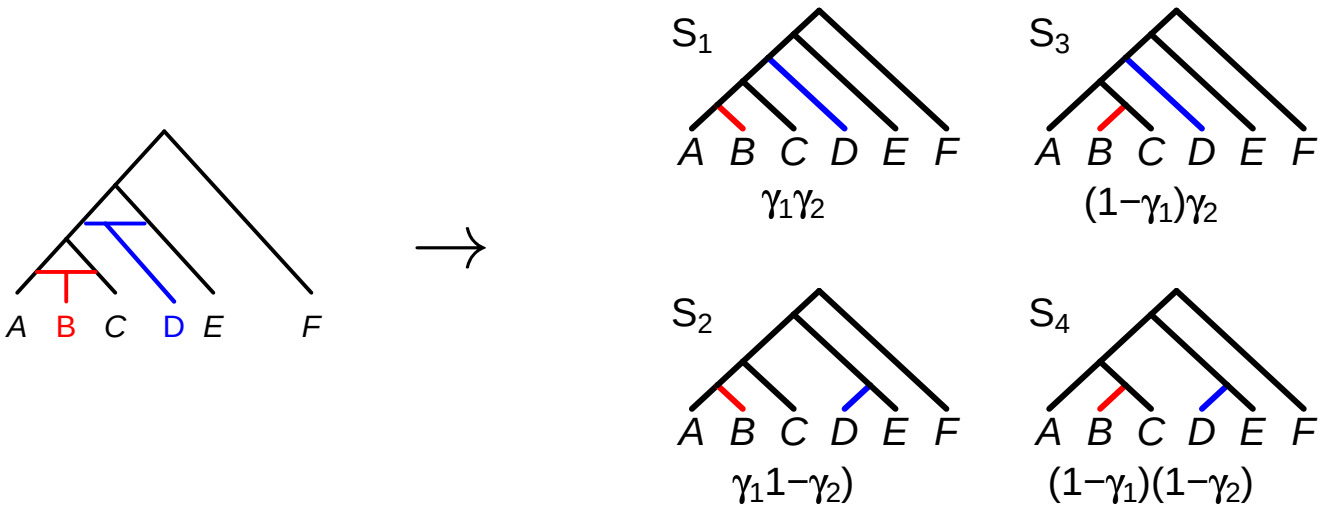
Motivating example:



Consider the hybrid species tree:



The likelihood function



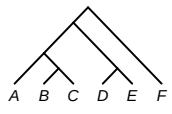
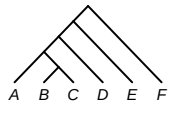
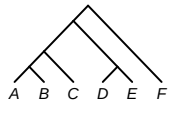
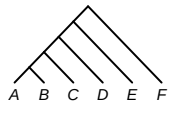
$$\prod_{i=1}^{N} \quad \{\gamma_1\gamma_2 f(g_i|S_1) + \gamma_1(1-\gamma_2) f(g_i|S_2) \\ +(1-\gamma_1)\gamma_2 f(g_i|S_3) + (1-\gamma_1)(1-\gamma_2) f(g_i|S_4)\}$$

Comments on computation

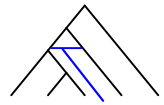
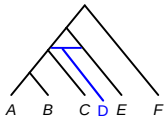
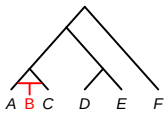
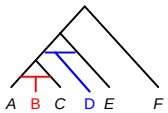
- Parameters in the likelihood function: γ_1, γ_2, branch lengths
- For a given hybrid species tree and sample of gene trees with divergence times, maximum likelihood branch lengths can be analytically determined
- Fitting the likelihood model for a hypothesized hybrid species tree only requires optimization of γ parameters

Selecting the best hybrid species tree

- For the example hybrid species tree, pick the best hybrid model from among possible models using the AIC:

Model	Tree	γ_1	γ_2	Number of Parameters
1		0	0	5
2		0	1	5
3		1	0	5
4		1	1	5

Selecting the best hybrid species tree

Model	Tree	γ_1	γ_2	Number of Parameters
5		0	(0,1)	6
6		1	(0,1)	6
7		(0,1)	0	6
8		(0,1)	1	6
9		(0,1)	(0,1)	7

- Important point: STEM-hy looks for evidence of hybridization in the presence of incomplete lineage sorting.
- By using the model from STEM to compute likelihoods, the coalescent process is incorporated.
- The AIC is used to compare models:
  - $AIC = -2lnL(M|D) + 2k$

  where M is the model and D is the data. $LnL(M|D)$ is the likelihood from STEM-hy for the hybridization model under consideration.

STEM example

- Changes to settings file:

```
properties:
                run: 3
                theta: 0.001
                beta: 0.0005
                burnin: 100
                seed: 3435893
                bound_total_iter: 20
                num_saved_trees: 10
                hybrid_species: H._heurippa
                hybrid_tree: user-heliconius.tre
species:
                H._melpomene: Hmelpomene
                H._hecale: Hhecale
                H._cydno: Hcydno
                H._heurippa: Hheurippa
```

Example: Hybridization in Heliconius

**********Results**********
....

Parental trees:

gamma(H._heurippa) = 1
((H._cydno:0.00009,(H._heurippa:0.00009,H._melpomene:0.00009):0.00000):3.68801,H._hecale:3.68810);
Lik: -357.4325907499209
AIC: 720.8651814998418
k: 3

gamma(H._heurippa) = 0
(((H._heurippa:0.00009,H._cydno:0.00009):0.35589,H._melpomene:0.35598):3.33212,H._hecale:3.68810);
Lik: -349.9185707499209
AIC: 705.8371414998418
k: 3

Hybrid trees:

(((H._heurippa:0.00009,H._cydno:0.00009):0.35589,H._melpomene:0.35598):3.33212,H._hecale:3.68810);
Lik: -349.5409832924012
gamma(H._heurippa): 0.6600000000000004
AIC: 707.0819665848024
k: 4

********** Done **********

What hybrid species can be considered?

- Care must be taken in selecting hybrid species:
  - Both members of a sister group cannot be selected as hybrid taxa in a single analysis. However, two analyses can be run (one with each of the sister group identified as the hybrid) and results will be comparable across runs.
  - The outgroup cannot be selected as a hybrid.
  - Both of these restrictions result from the fact that for now hybridization is only considered between sister taxa.
- More general hybridization relationships can be considered "by hand" using the user-specified tree feature of STEM-hy.

STEM-hy: Strengths and Weaknesses

- STEM-hy makes some fairly strong assumptions:
  - Error in estimating gene trees and branch lengths is not incorporated!!!! But the possibility of carrying out bootstrap analysis helps.
  - Information in the sequence data is not used directly; it is only used as summarized by estimated gene divergence times.
  - There is a single value of θ for the entire tree.

STEM-hy: Strengths and Weaknesses

- STEM-hy makes some fairly strong assumptions:
  - Error in estimating gene trees and branch lengths is not incorporated!!!! But the possibility of carrying out bootstrap analysis helps.
  - Information in the sequence data is not used directly; it is only used as summarized by estimated gene divergence times.
  - There is a single value of θ for the entire tree.
- There are trade-offs involved, and STEM-hy does some things well:
  - It is quick (even the tree search does not take long).
  - It can handle missing data easily and intuitively.
  - Simulations demonstrate reasonable performance (unlikely to be misleading; may be uninformative).

Challenge Datasets

- I've created four datasets under varying conditions:
  - M1 No hybridization, long intervals between speciation events.
  - M2 No hybridization, short intervals between speciation events.
  - M3 Low-levels of hybridization - B is a hybrid of A and C (species tree as in M1 and M2).
  - M4 Extensive hybridization - B is a hybrid of A and C (species tree as in M1 and M2).
- All data sets have 6 species, 2 individuals/species, and 10 loci.
- GOAL: match the data set to the condition listed above
  Solutions are linked to on my course wiki page.