Inferring Rooted Species Trees from Unrooted Gene Trees

James Degnan
University of New Mexico
Dept. of Mathematics and Statistics

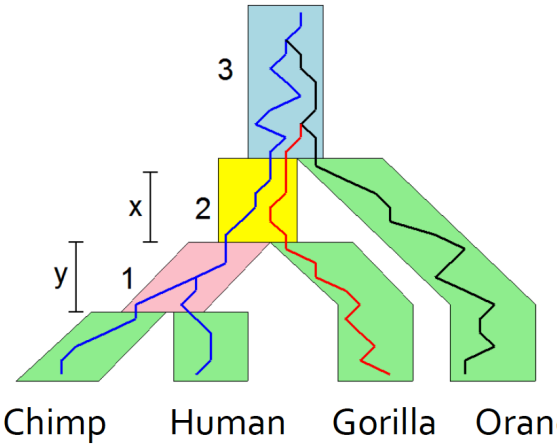
Thanks to Ayed Alanzi, Mathematics Department, University of Majmaah, KSA

NIH R01 GM117590-03

Outline

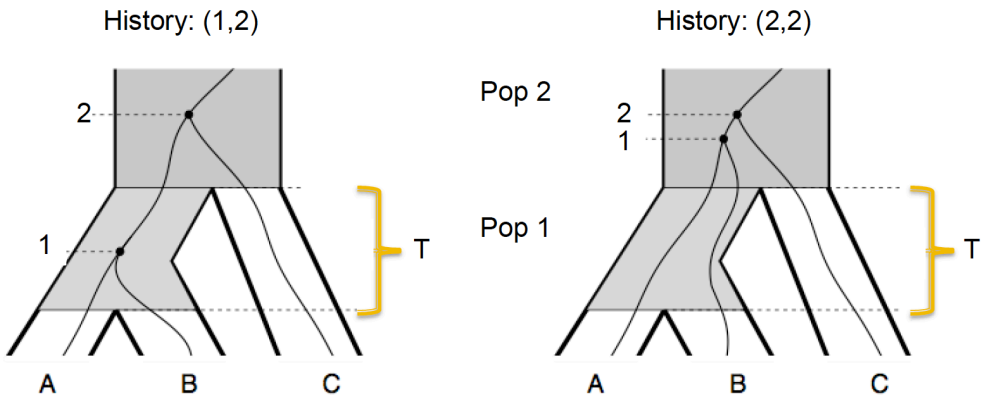
- Motivation
- ABC methods
- ABC simulation study
- Conclusion

Multispecies coalescent



The gene tree is a random variable. The gene tree distribution is parameterized by the species tree topology and internal branch lengths.

Probabilities of gene trees



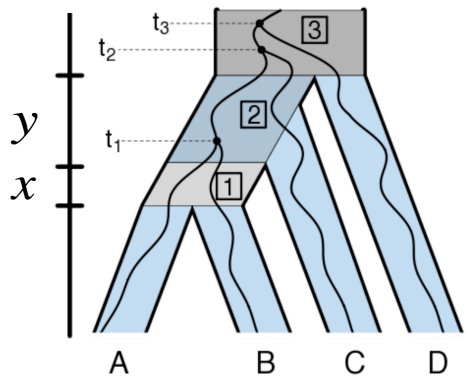
Total probability that the gene tree matches the species tree (Nei 1987):

$$1 - e^{-T} + (1/3)e^{-T} = 1 - (2/3)e^{-T} > 1/3$$

Probabilities of gene trees, use $X = e^{-x}$, $Y = e^{-y}$

History	Probability
h_1: (1,2,3)	$(1-X)(1-Y)$
h_2: (1,3,3)	$\frac{1}{3}(1-X)Y$
h_3: (2,2,3)	$\frac{1}{3}X(1-\frac{3}{2}Y+\frac{1}{2}Y^3)$
h_4: (2,3,3)	$\frac{1}{6}XY-\frac{1}{6}XY^3$
h_5: (3,3,3)	$\frac{1}{18}XY^3$
Total	$1-\frac{2}{3}X-\frac{2}{3}Y+\frac{1}{3}XY+\frac{1}{18}XY^3$

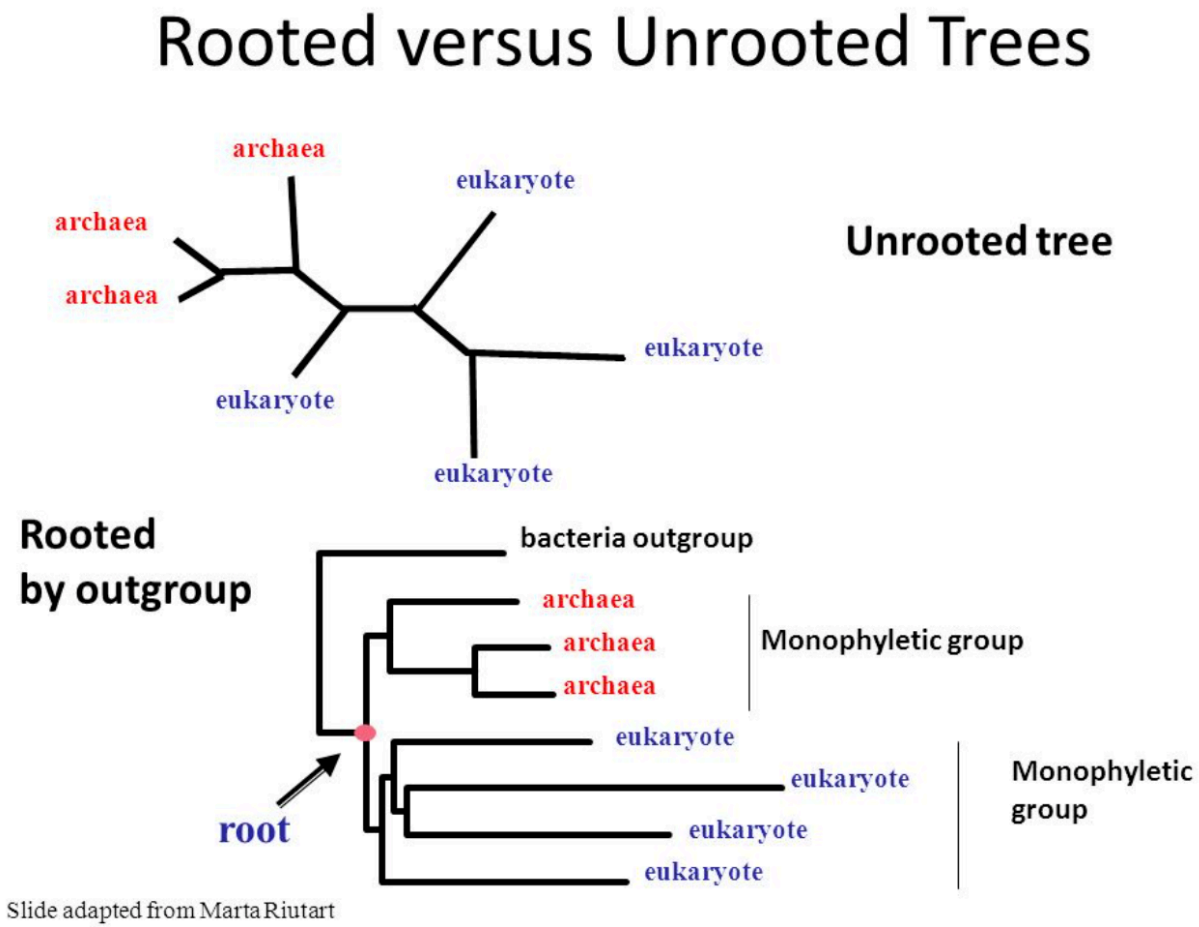
history h_4 = (2,3,3)



$\frac{1}{9}g_{2,2}(x)g_{3,2}(y) = \frac{1}{6}XY-\frac{1}{6}XY^3$

First derived in Pamilo and Nei (1988).

Probabilities of gene trees



Probabilities of gene trees

$$\Pr[\,^{A}_{B}\!\!\diagup\!\!\diagdown\!\!^{C}_{D}\,] = P[(((AB)C)D)] + P[(((AB)D)C)] + P[(((CD)A)B)] + P[(((CD)B)A)] + P[((AB)(CD))]$$

Probabilities of gene trees

```
> p1(x,y);
```

$$(1-e^{-x})(1-e^{-y})+\frac{1}{3}(1-e^{-x})e^{-y}+\frac{1}{3}e^{-x}\left(1-\frac{3}{2}e^{-y}+\frac{1}{2}e^{-3y}\right)+\frac{1}{9}e^{-x}\left(\frac{3}{2}e^{-y}-\frac{3}{2}e^{-3y}\right)+\frac{1}{18}e^{-x}e^{-3y}$$

```
> p2(x,y);
```

$$\frac{1}{3}(1-e^{-x})e^{-y}+\frac{1}{9}e^{-x}\left(\frac{3}{2}e^{-y}-\frac{3}{2}e^{-3y}\right)+\frac{1}{18}e^{-x}e^{-3y}$$

```
> p3(x,y);
```

$$\frac{1}{18}e^{-x}e^{-3y}$$

```
> p4(x,y);
```

$$\frac{1}{18}e^{-x}e^{-3y}$$

```
> p5(x,y);
```

$$\frac{1}{3}(1-e^{-x})e^{-y}+\frac{1}{9}e^{-x}\left(\frac{3}{2}e^{-y}-\frac{3}{2}e^{-3y}\right)+\frac{1}{9}e^{-x}e^{-3y}$$

```
>
```

Probabilities of gene trees

Drumroll....

Probabilities of gene trees

```
> p1(x,y);
```

$$(1-e^{-x})(1-e^{-y})+\frac{1}{3}(1-e^{-x})e^{-y}+\frac{1}{3}e^{-x}\left(1-\frac{3}{2}e^{-y}+\frac{1}{2}e^{-3y}\right)+\frac{1}{9}e^{-x}\left(\frac{3}{2}e^{-y}-\frac{3}{2}e^{-3y}\right)+\frac{1}{18}e^{-x}e^{-3y}$$

```
> p2(x,y);
```

$$\frac{1}{3}(1-e^{-x})e^{-y}+\frac{1}{9}e^{-x}\left(\frac{3}{2}e^{-y}-\frac{3}{2}e^{-3y}\right)+\frac{1}{18}e^{-x}e^{-3y}$$

```
> p3(x,y);
```

$$\frac{1}{18}e^{-x}e^{-3y}$$

```
> p4(x,y);
```

$$\frac{1}{18}e^{-x}e^{-3y}$$

```
> p5(x,y);
```

$$\frac{1}{3}(1-e^{-x})e^{-y}+\frac{1}{9}e^{-x}\left(\frac{3}{2}e^{-y}-\frac{3}{2}e^{-3y}\right)+\frac{1}{9}e^{-x}e^{-3y}$$

```
> simplify(p1(x,y)+p2(x,y)+p3(x,y)+p4(x,y)+p5(x,y));
```

$$1-\frac{2}{3}e^{-x}$$

Identifiability

For four taxa, there is only one parameter in the gene tree distribution. In addition to the branch lengths, the species tree topology (including shape) is not identifiability.

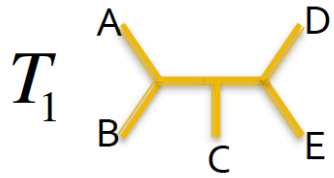
It turns out that for five or more taxa, the species tree, including topology and branch lengths, is identifiabiable using linear invariants and inequalities of the gene tree probabilities.

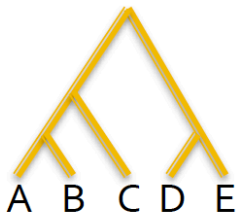
Instead of giving the whole argument, I'll give an example of distinguishing two species trees with the same unrooted topology.

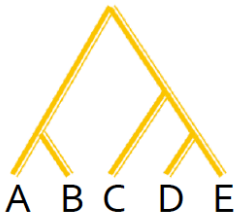
Identifiability for 5 taxa

The labeled, rooted species tree topology can be determined by further considering invariants or inequalities in unrooted gene tree probabilities.

Example: Given the unrooted species tree and given that the species tree is balanced, The rooted species tree is one of:

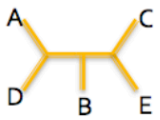


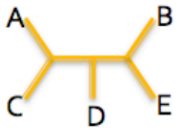




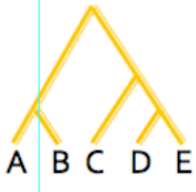
Identifiability for 5 taxa

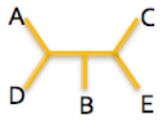
For species tree 

$P($ 

$) < P($ 

$)$

For species tree 

$P($ 

$) > P($ 

$)$

Identifiability for 5 taxa

The general result is

Theorem[Allman, Degnan, Rhodes, 2011] The unrooted gene tree distribution arising from one sample per species identifies the species tree with internal branch lengths for $n = 5$ or more taxa. If $n = 4$, then the gene trees only identify the unrooted species tree with its one internal branch length.

Motivation

A question raised is whether there is a practical method of inferring the root under these conditions?

This would be useful for cases in which a good outgroup is unknown, making it difficult to root an unrooted species tree or the input gene trees.

Motivation

Looking at the history of methods for inferring species trees, most have been implemented in the last 10-12 years, such as the minimize deep coalescence criterion implemented in Maddison and Knowles (2006).

For methods using gene trees as input (summary methods), there has been a shift from methods using rooted gene trees to infer a rooted species tree to methods using unrooted gene trees to infer an unrooted species tree, which is then rooted using an outgroup:

Timeline of summary methods

- Consensus methods (Gadagkar and Kumar, 2005)
- MDC (Maddison and Knowles, 2006)
- BUCKy (Ane et al. 2007)
- Rooted Triple Consensus (Ewing et al., 2008)
- STEAC (Liu et al., 2009)
- STAR (Liu et al., 2009)
- MP-EST (Liu et al., 2010)
- STEM/GLASS/Maximum Tree (Kubatko et al., 2009, Mossell and Roch, 2010, Liu et al., 2010)
- BUCKy with quartets (Larget et al, 2010)
- ST-ABC (Fan and Kubatko, 2011)
- NJ_{st} (Liu and Lu, 2011)
- STELLS (Wu, 2012)
- ASTRAL (Mirarab et al., 2014)
- SNaQ (Solís-Lemus and Ané, 2016)

Shift to unrooted methods

Why has there been this shift?

Because the coalescent is a very rooted idea, it was more natural to think of methods that used rooted gene trees.

In practice, however, gene trees estimated from sequence data tend to be unrooted, especially for faster methods (e.g., PhyML, RAxML), so there is some incentive for thinking of unrooted gene trees as input in applications.

Also in practice, recent unrooted methods such as ASTRAL and NJ_{st} have been outperforming some of the earlier rooted approaches.

Using unrooted gene trees

How to infer the rooted species tree from the unrooted gene trees?

Three possibilities:

- invariants
- maximum likelihood (use the probabilities of the unrooted gene trees)
- Approximate Bayesian Computation

Approximate Bayesian Computation

Approximate Bayesian Computation developed in population genetics for situations in which the likelihood was difficult to write down, but it is relatively easy to simulate from the desired distribution.

1. Start with a data set X
2. Simulate parameters from a prior distribution for θ. Let the simulated parameters be $\theta^{(i)}$
3. Simulate a data set $Y^{(i)}$ from the simulated parameter $\theta^{(i)}$ that is the same size as the original data
4. Compute the distance between the observed and simulated data $D^{(i)} = d(X, Y^{(i)})$
5. If the distance is smaller than some tolerance δ, accept $\theta^{(i)}$
6. Repeat 2–5 until enough simulated parameters have been accepted
7. Use the distribution of the simulated parameters $\{\theta^{(i)}\}$ to approximate the posterior distribution for θ.

Approximate Bayesian Approximation

A variation is the following

1. Start with a data set X
2. Simulate parameters from a prior distribution for θ. Let the simulated parameters be $\theta^{(i)}$
3. Simulate a data set $Y^{(i)}$ from the simulated parameter $\theta^{(i)}$ that is the same size as the original data
4. Compute the distance between the observed and simulated data $D^{(i)} = d(X, Y^{(i)})$
5. Repeat 2–5 J times
6. Accept those simulated parameters that corresponded to the smallest αJ distances. E.g. the best 1%.
7. Use the distribution of the simulated parameters $\{\theta^{(i)}\}$ to approximate the posterior distribution for θ.

ST-ABC

Fan and Kubatko (2011) first applied an ABC-inspired algorithm to infer species trees from rooted gene trees to estimate a rooted species tree.

For four taxa, they used a Yule prior for the species tree topology and uniform priors for the branch lengths. For the distance, they counted the number of times each topology occurred in the gene trees and used

$$D(X, Y^{(i)}) = \sum_{k \text{ topologies}} \frac{(n_{\text{obs},k} - n_{\text{exp},k})^2}{n_{\text{exp},k}}$$

where n_{obs} is a vector of counts for topologies in the observed data, and n_{exp} is the vector of expected counts given the simulated species tree.

ST-ABC

Instead of the expected counts, the usual ABC algorithm would be to simulate a new data set and count the number of times each topology arose.

We can think of the counts as a summary statistic. In this case, the summary counts are **sufficient statistics**, meaning that they retain all of the information in the data relevant for computing the probability of the sample. Note that the vector of counts is 15-dimensional for 4-taxa, and much higher for more taxa.

Approximate Bayesian Approximation

Advantages of ABC:

- You don't have to compute the likehood
- You don't have to worry about convergence (unlike MCMC)
- It is easy to parallelize

Approximate Bayesian Computation

Difficulties of ABC:

- It is not clear how big J needs to be
- There is a variance-bias tradeoff between using large versus small αJ. Smaller cutoffs (similarly a smaller δ) lead to more accurate inference but more variability in estimates
- Sufficient statistics can be hard to find and/or high-dimensional
- A challenge in ABC is often to find lower dimensional summary statistics that are "approximately sufficient"

Approximate Bayesian Computation

For our project, we adapted the ABC approach to inferring rooted species trees from unrooted species trees. The modifications were

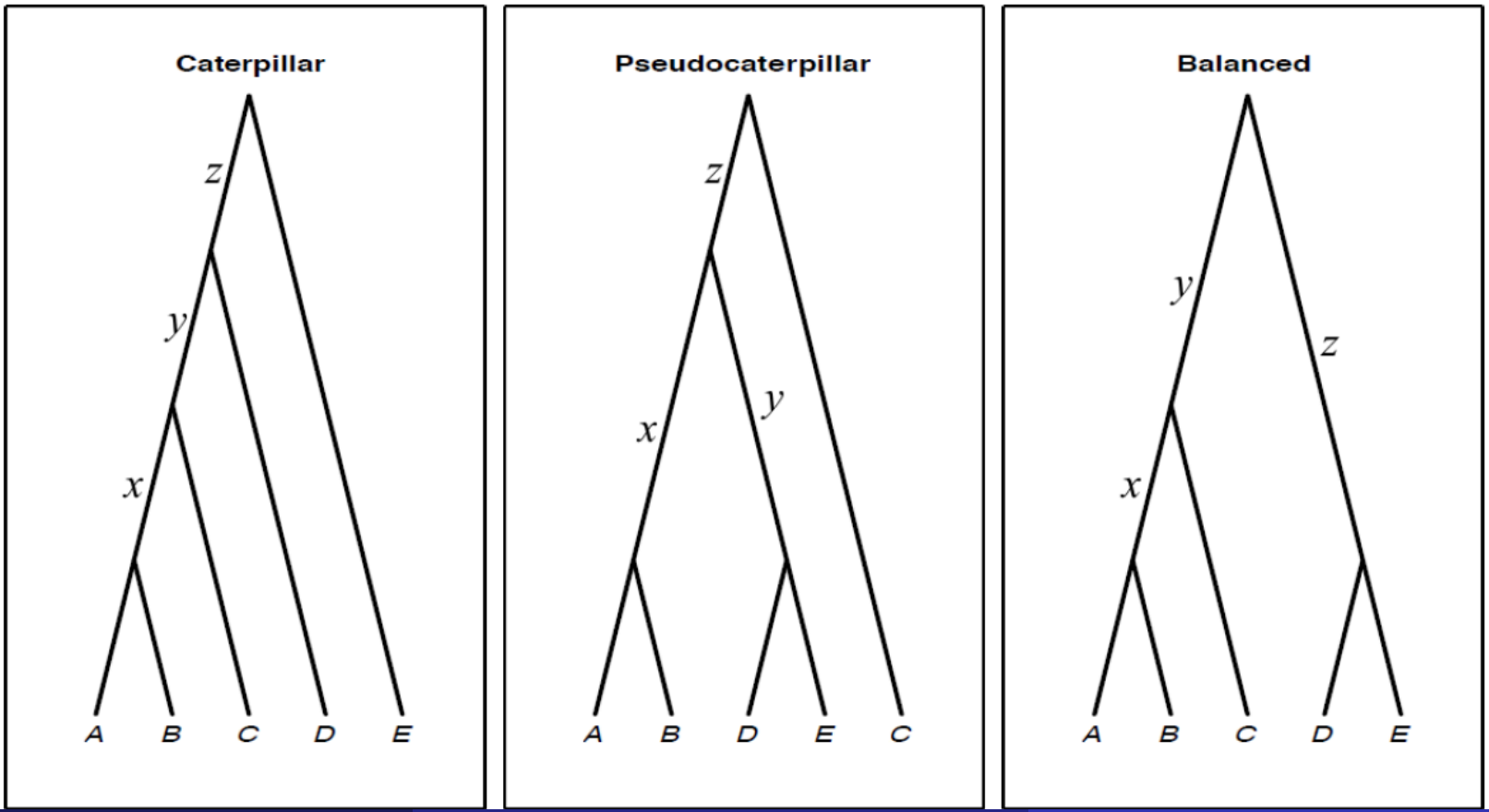
- Have a small number of candidate species trees, perhaps obtained by first estimating an unrooted species tree by another method such as ASTRAL or NJ_{st}. If these programs estimate the unrooted tree correctly, then there are only $n - 3$ possible root locations.
- Simulate unrooted gene trees and use Euclidean distance instead of the goodness-of-fit criterion to avoid division by 0 problems
- Use split counts rather than topology counts as a summary statistic in order to have lower-dimensional summary statistics

ABC Simulation Study

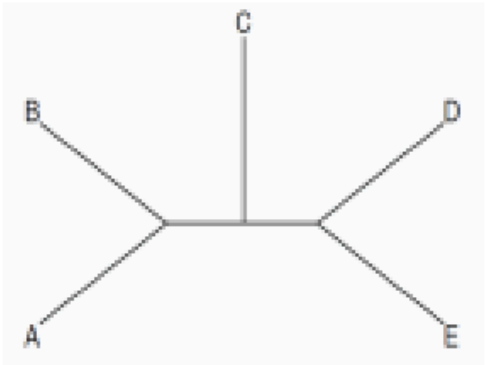
For our simulation study, we used 50 iterations of:

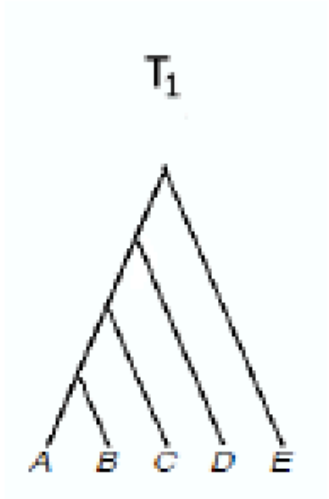
- 5-taxon trees (the smallest for which unrooted gene trees identify the rooted species tree), and 8-taxon trees
- 100 loci
- initially known gene trees, some examples with estimated gene trees
- $J = 50000$, $\alpha = .002$, so retaining the best 100 trees
- Topology counts and split counts
- Euclidean distances based on vectors of counts
- For 5-taxon trees, branch lengths were either 0.1 or 1.0 coalescent units in different combinations
- Uniform prior for topologies (assuming unrooted tree known), and exponential rate 1 prior for branch lengths

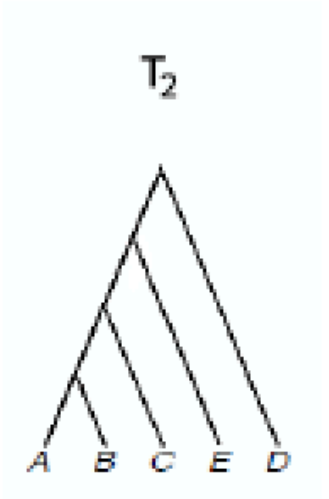
ABC Simulation Study: model species trees, $x, y, z \in \{0.1, 1.0\}$

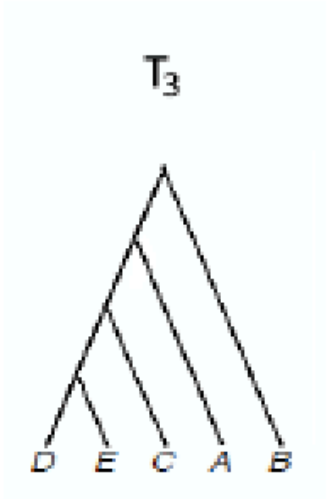


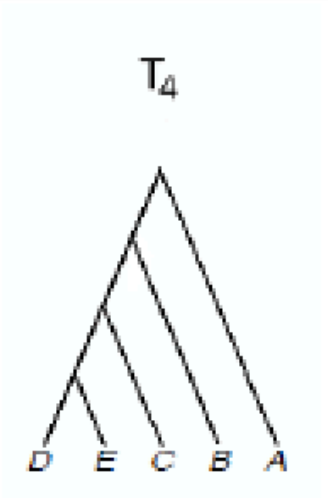
ABC Simulation Study: prior for the topology

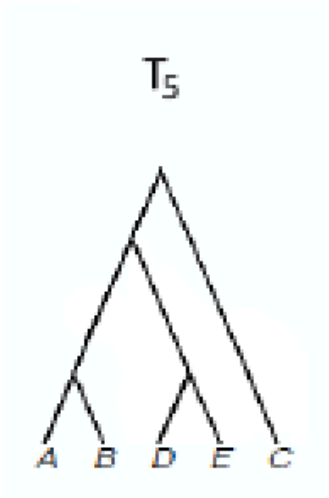


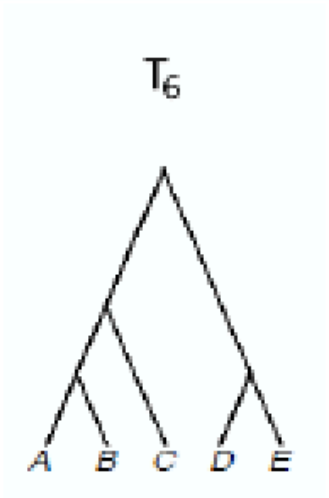


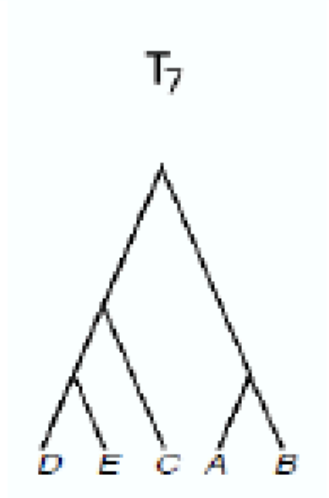












ABC Simulation Study: results using topology counts

Table: Average posterior probabilities using topology counts for five-taxon trees. Average posterior probabilities for the correct species tree topology are in bold.

	Species tree, branch len. (x, y, z)	Prop. Correct (%)	Coverage prob. (%)	Average of the posterior probability (%)							# of trees in 90% CR.
				T_1	T_2	T_3	T_4	T_5	T_6	T_7	
Cat.	(0.1,0.1,0.1)	76	96	**54**	24	2	3	7	7	3	3
	(0.1,0.1,1.0)	94	100	**68**	21	0	2	4	5	1	2
	(1.0,0.1,0.1)	50	96	**40**	34	1	1	16	4	5	3
	(0.1,1.0,0.1)	47	100	**36**	33	0	0	1	29	0	3
	(1.0,1.0,1.0)	36	94	**25**	25	2	3	10	29	7	4
Pseudo-cat.	(0.1, 0.1, 0.1)	92	96	8	9	8	8	**56**	6	5	4
	(1.0, 1.0, 1.0)	9	54	16	17	12	14	**12**	15	14	4
	(0.1, 0.1, 1.0)	94	100	4	4	3	3	**80**	3	3	2
	(0.1, 1.0, 0.1)	52	96	2	3	24	21	**31**	13	5	4
Bal.	(0.1,0.1,0.1)	2	50	30	28	5	5	16	**11**	4	4
	(1.0,1.0,1.0)	65	92	16	15	9	8	13	**29**	11	5
	(0.1,0.1,1.0)	32	84	11	10	12	15	23	**24**	4	4
	(0.1,1.0,0.1)	26	88	35	36	1	0	1	**27**	0	3
	(1.0,0.1,0.1)	0	24	29	31	1	2	22	**7**	9	4

ABC Simultation Study: 5 taxa, known gene trees

- Performance was mixed, with accuracy depending on the combination of topology and branch lengths.
- Caterpillar and Pseudocaterpillar were reasonably accurately inferred when species tree branch lengths were short
- Balanced was very inaccurate when branches were short
- A small number of simulations showed increased accuracy with 5000 loci for the balanced tree, but this case is still very difficult to infer
- Overall, this appears to be a very difficult inference problem
- Patterns for split and topology counts are very similar, suggesting that splits are reasonable summary statistics

ABC Simulation Study: 5 taxa, known gene trees, all branch lengths equal, Caterpillar species tree

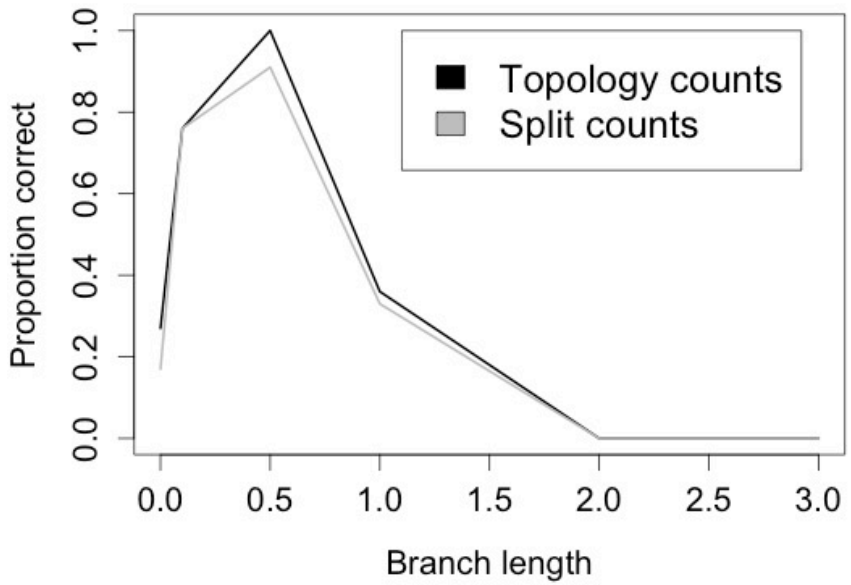


Figure: Proportion of times the species tree has the highest posterior probability when all internal branches are equal, $x = y = z \in \{0, 0.1, 0.5, 1.0, 2.0, 3.0\}$

ABC Simulation Study: understanding long and short branch lengths

- For a star species tree, the caterpillar is inferred as the species tree too often ($> 1/7$). An explanation is that for exponential branch lengths, a caterpillar will predict higher levels of gene tree incongruence than a balanced tree. Consequently, balanced trees are underrepresented in the posterior when the species tree has short branches.
- When branches are long in the true species tree, there is less gene tree incongruence, and the posterior will tend to favor balanced trees as opposed to the caterpillar shape.

Conclusions

- Inferring rooted species trees from unrooted gene trees is possible but difficult
- simulations here were limited to 100 loci — more accurate results would be expected with more loci
- Inferring species trees from unrooted gene trees benefits from gene tree incongruence