James Degnan
University of Canterbury
29 October 2009

(Joint work with Elizabeth Allman and John Rhodes)

# Identifying rooted species trees from unrooted gene tree probabilities

Outline

1. Background
   A. gene trees vs. species trees
   B. coalescence and incomplete lineage sorting

2. Rooted gene tree probabilities as polynomials

3. Unrooted gene tree probabilities

Part 1: The multispecies coalescent
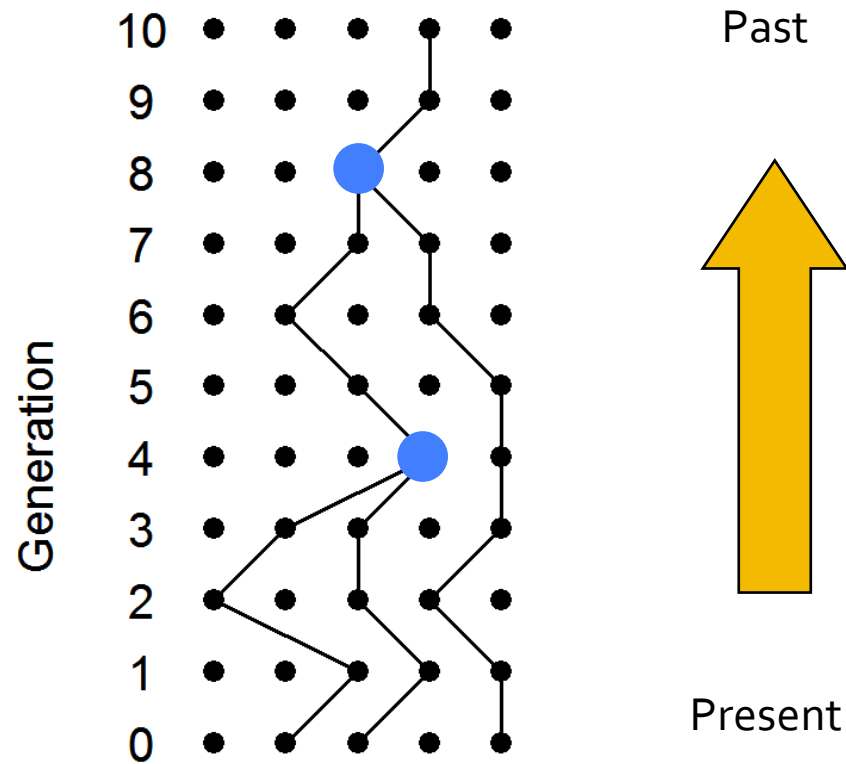
Population Genetics and Phylogenetics

Population genetics: traditionally used to
   analyze single populations.

Phylogenetics: What is the best way to infer
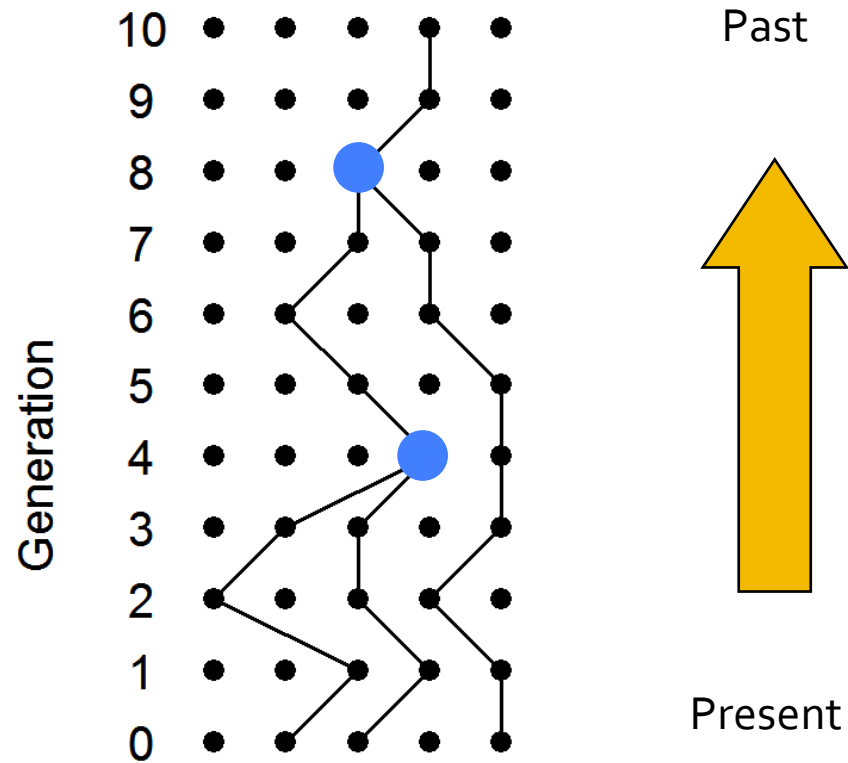relationships between populations/species?
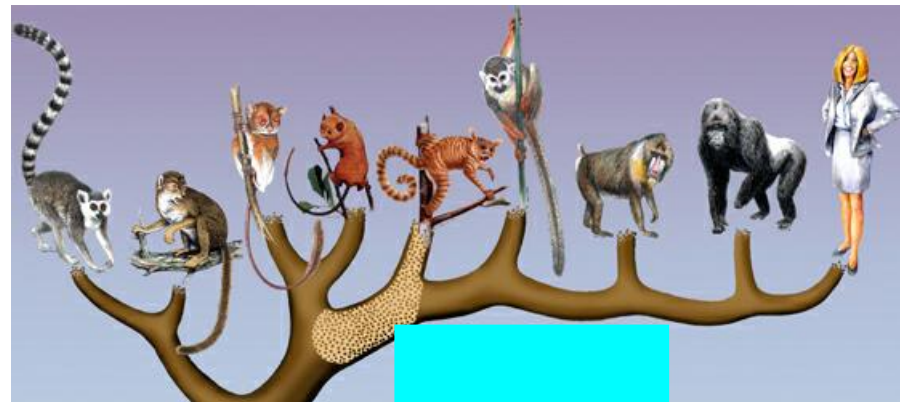
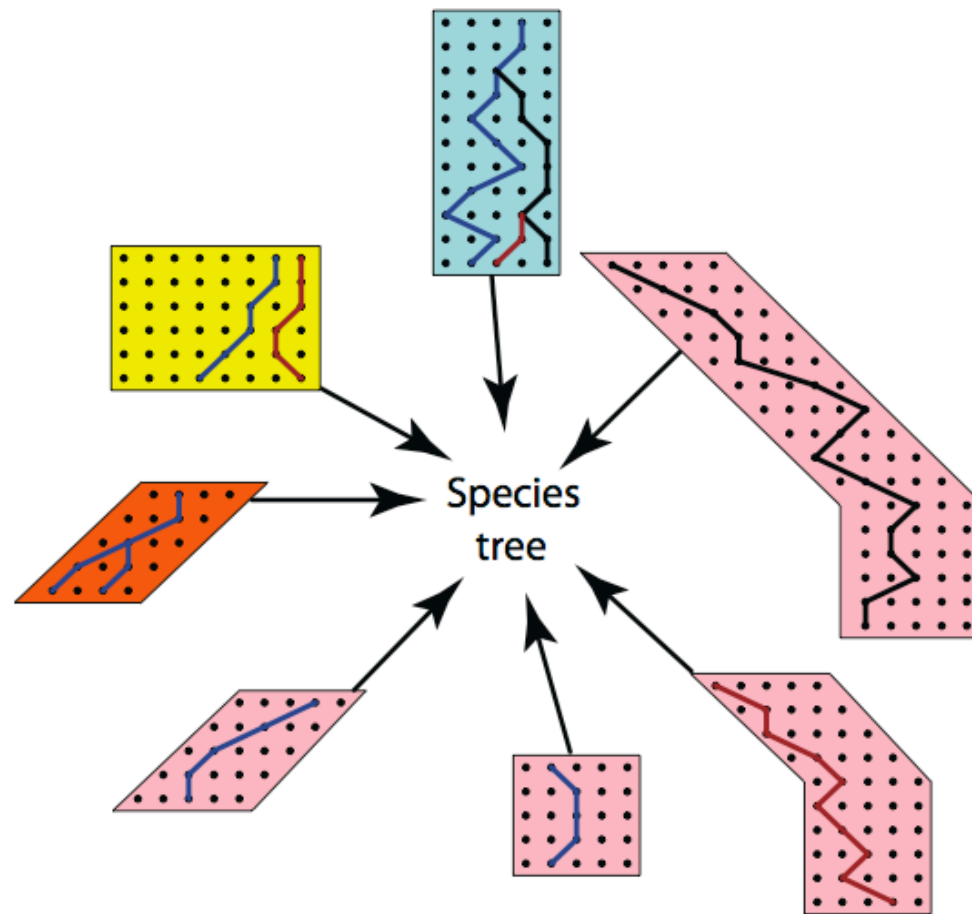Graphic by Mark A. Klinger, Carnegie Museum of Natural History, Pittsburgh
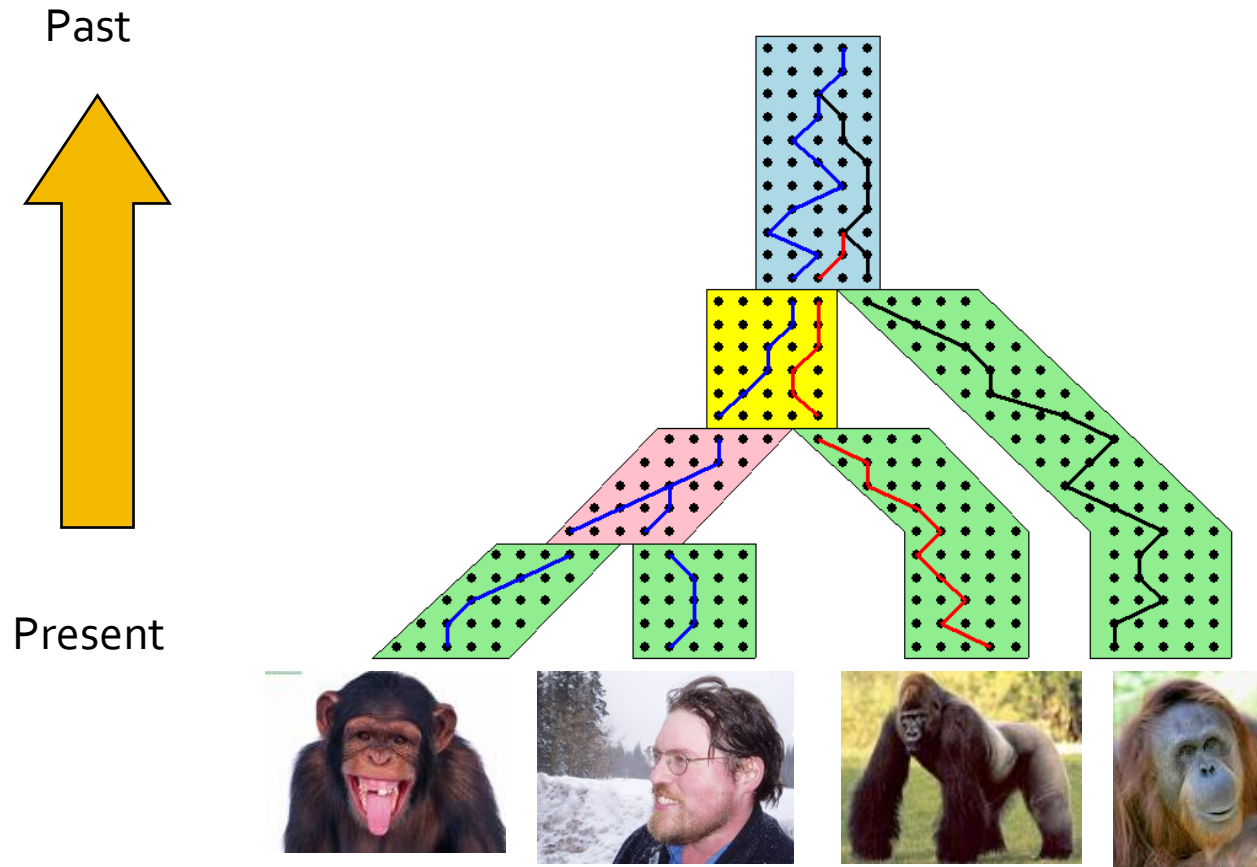
The coalescent process

One population

Model for lineages in populations
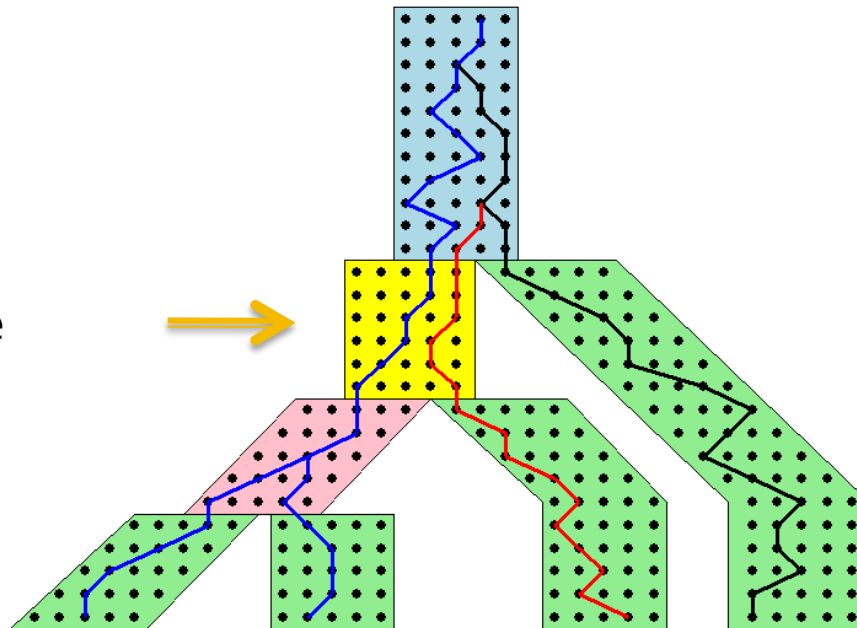
Multiple populations/species

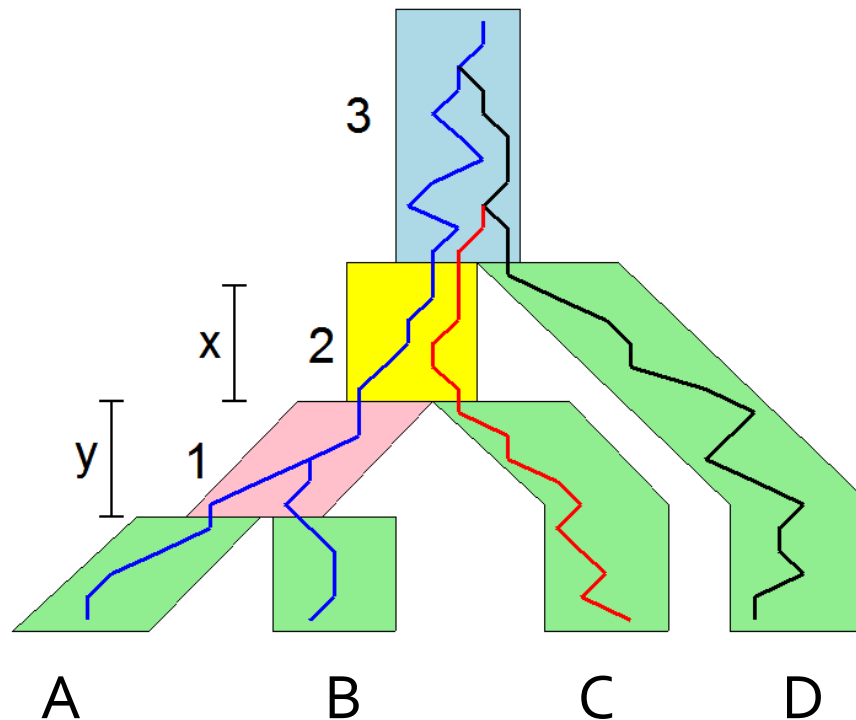# Gene tree in a species tree

Gene tree in a species tree

Incomplete lineage sorting

Model species tree with gene tree

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

Part 2: Rooted gene tree probabilities

How can we compute probabilities of gene trees given species trees?

-Under a coalescent model, probabilities for gene trees with three species were derived by Nei (1987): $1-(2/3)e^{-T}$

-Probabilities for the gene tree to match the species tree topology for 4 and 5 species given by Pamilo and Nei (1988).

-All 30 species tree/gene tree combinations for 4 species given by Rosenberg (2002).

-General case solved by Degnan and Salter (2005) and implemented by program COAL. Also allows $n_i \geq 0$ individuals sampled in species i.

Coalescent histories as cases

Probability that the gene tree matches the species tree for three taxa
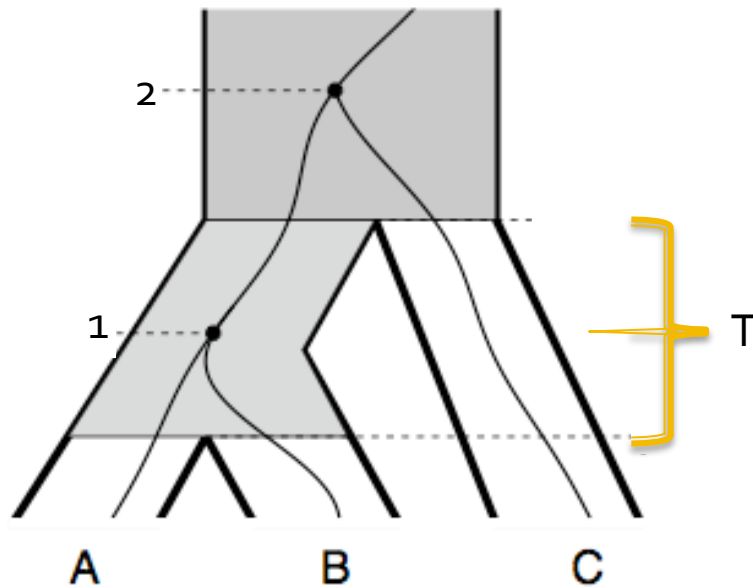
Probability:

$$\Pr[X \le T] = \int_0^T e^{-t}\, dt = 1 - e^{-T}$$
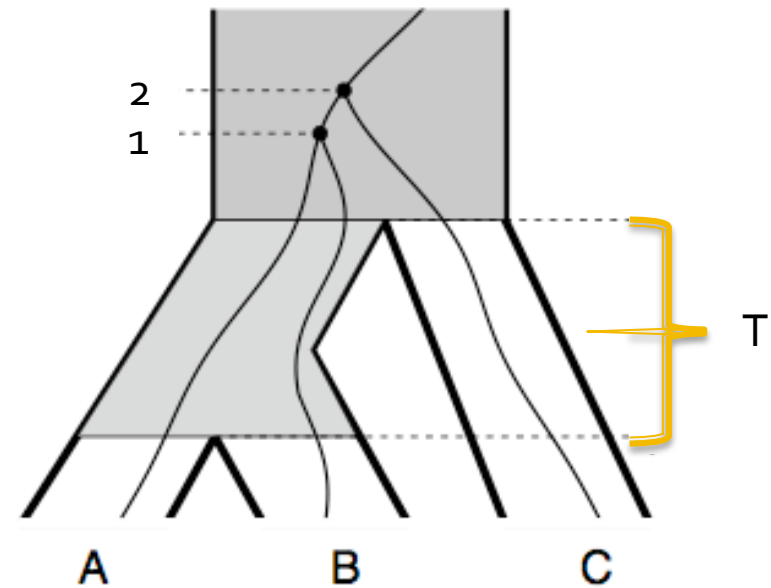
Probability:

$$(1/3)\Pr[X > T] = (1/3)e^{-T}$$

Coalescent histories as cases

Probability that the gene tree matches the species tree for three taxa

History: (1,2)

History: (2,2)

Probability:

$$\Pr[X \le T] = \int_0^T e^{-t}\, dt = 1 - e^{-T}$$
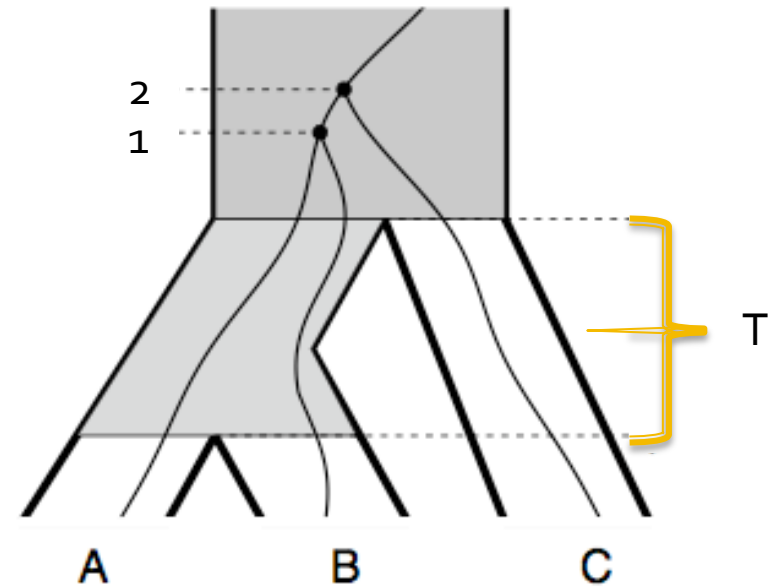
Probability:

$$(1/3)\Pr[X > T] = (1/3)e^{-T}$$

Coalescent histories as cases

Probability that the gene tree matches the species tree for three taxa

History: (1,2)
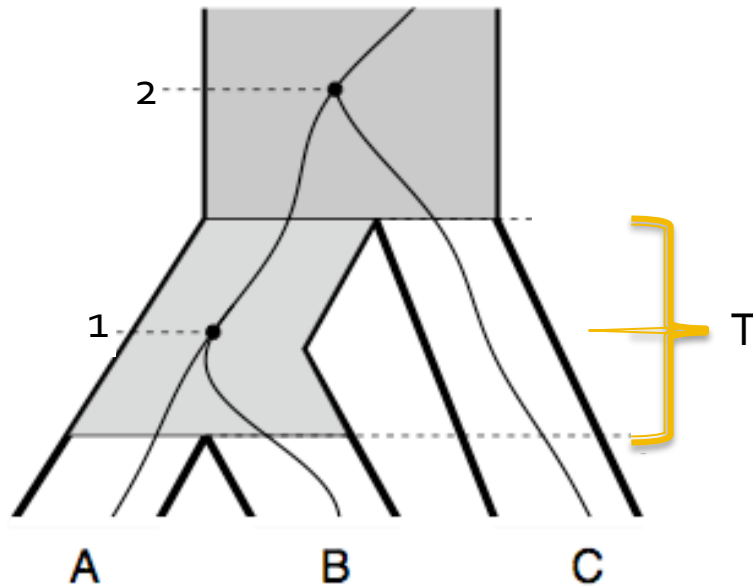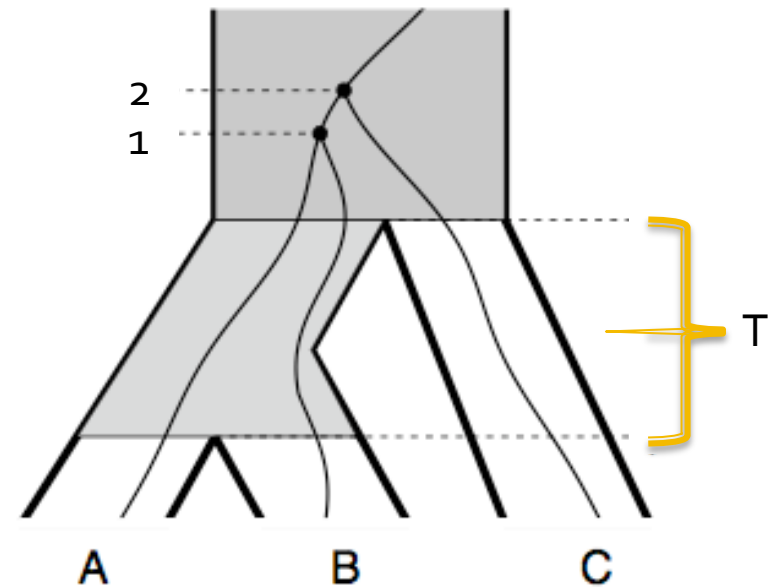
History: (2,2)

Total probability:

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

Coalescent histories as cases

Probability that the gene tree matches the species tree for three taxa
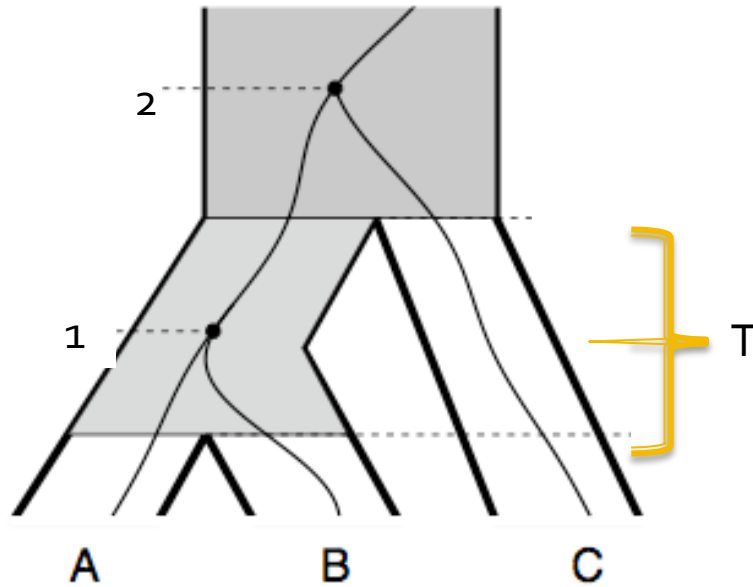
History: (1,2)
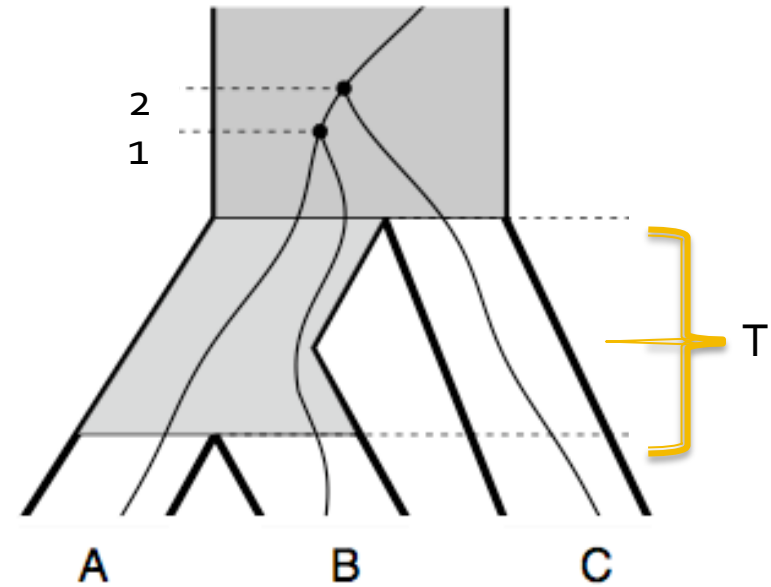
History: (2,2)

Total probability:

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

(1,2)

Coalescent histories as cases

Probability that the gene tree matches the species tree for three taxa

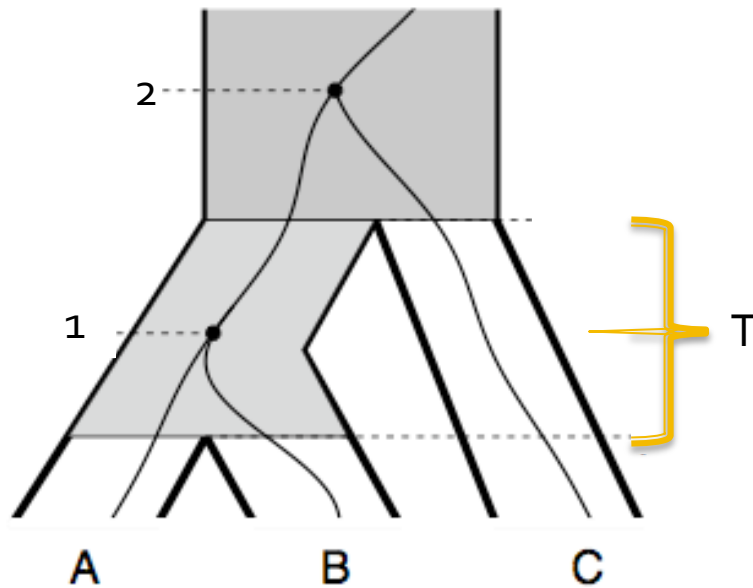History: (1,2)

History: (2,2)

Total probability:

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

(1,2)   (2,2)
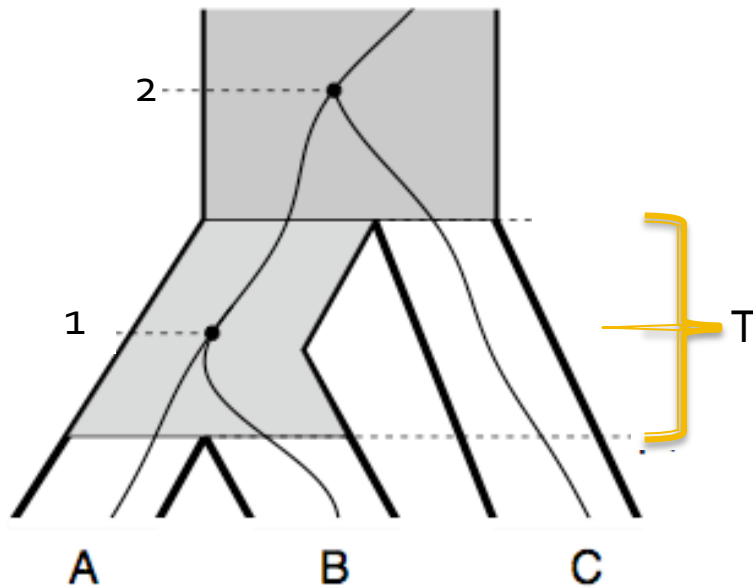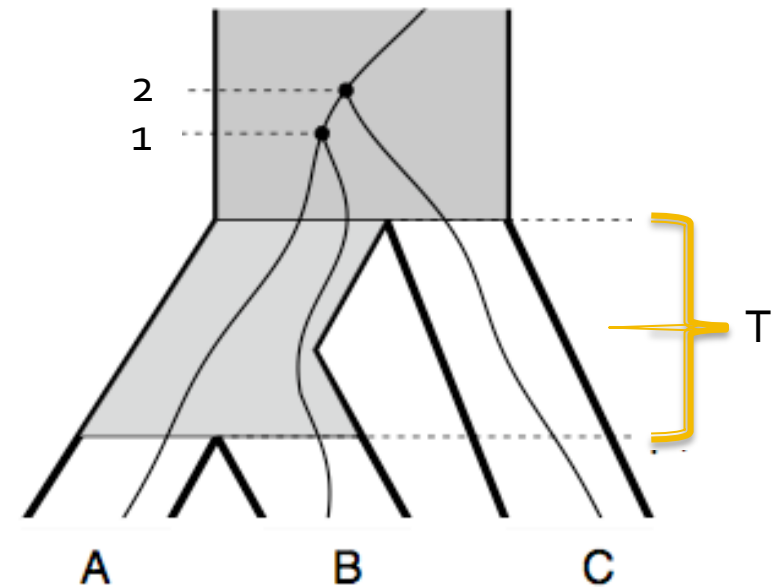
Coalescent history : a list of populations in which coalescent events occur

Species tree
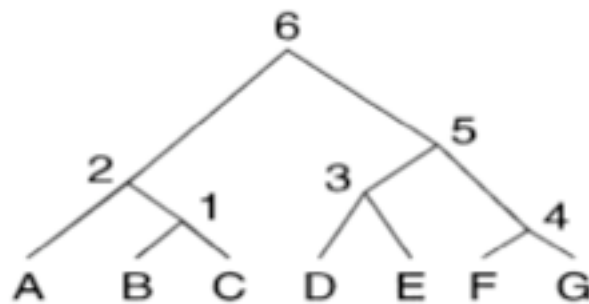
Gene tree

(2,5,5,6,6)
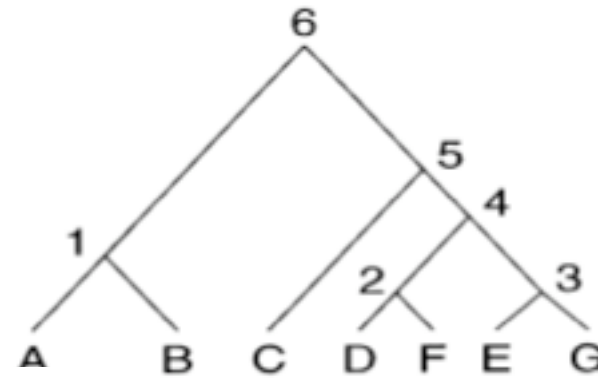
(2,5,6,6,6)

Gene tree probabilities

$$\Pr[G \mid S] = \sum_{histories} \Pr[G, histories \mid S]$$

Gene tree probabilities

$$\Pr[G = g \mid S] = \sum_{histories} \Pr[G = g, histories \mid S]$$

$$= \sum_{histories} \prod_{b} w_b p_{u(b),v(b)}(T_b)$$
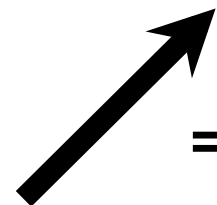
combinatorial enumeration, complexity only known in special cases

internal branches of S

probability coalescences are consistent with g

u coalesce into v

branch length

Gene tree probabilities as linear combinations

Probabilities of coalescent histories are products of functions $p_{i,j}(t_b)$

$$p_{2,1}(t) = 1 - e^{-t}$$

$$p_{2,2}(t) = e^{-t}$$

$$p_{3,1}(t) = 1 - \tfrac{3}{2}e^{-t} + \tfrac{1}{2}e^{-3t}$$

$$p_{3,2}(t) = \tfrac{3}{2}e^{-t} - \tfrac{3}{2}e^{-3t}$$

$$p_{3,3}(t) = e^{-6t}$$

$$p_{4,1}(t) = 1 - \tfrac{9}{5}e^{-t} + 3e^{-3t} - \tfrac{1}{5}e^{-6t}$$

$$p_{4,2}(t) = \tfrac{9}{5}e^{-t} - 9e^{-3t} + \tfrac{6}{5}e^{-6t}$$

$$p_{4,3}(t) = 2e^{-3t} - 2e^{-6t}$$

$$p_{4,4}(t) = e^{-6t}$$

$$p_{5,1}(t) = 1 - 2e^{-t} + \tfrac{10}{7}e^{-3t} - \tfrac{1}{2}e^{-6t} + \tfrac{1}{14}e^{-10t}$$

$$p_{5,2}(t) = 2e^{-t} - \tfrac{30}{7}e^{-3t} + 3e^{-6t} - \tfrac{5}{7}e^{-10t}$$

$$p_{5,3}(t) = \tfrac{20}{7}e^{-3t} - 5e^{-6t} + \tfrac{15}{7}e^{-10t}$$

$$p_{5,4}(t) = \tfrac{5}{2}e^{-6t} - \tfrac{5}{2}e^{-10t}$$

$$p_{5,5}(t) = e^{-10t}$$

Using transformed branch lengths,

$$X_1 = e^{-t_1}, X_2 = e^{-t_2}, \quad etc.$$

gene tree probabilities can be written as linear combinations
of monomials $X_1^{\alpha_1} X_2^{\alpha_2} \cdots X_{n-2}^{\alpha_{n-2}}$

where *n* is the number of tips, and

$$\sum \alpha_i \le \sum_{k=2}^{n-2} \binom{k}{2}$$

Example

history $h_1 = (1,2,3)$
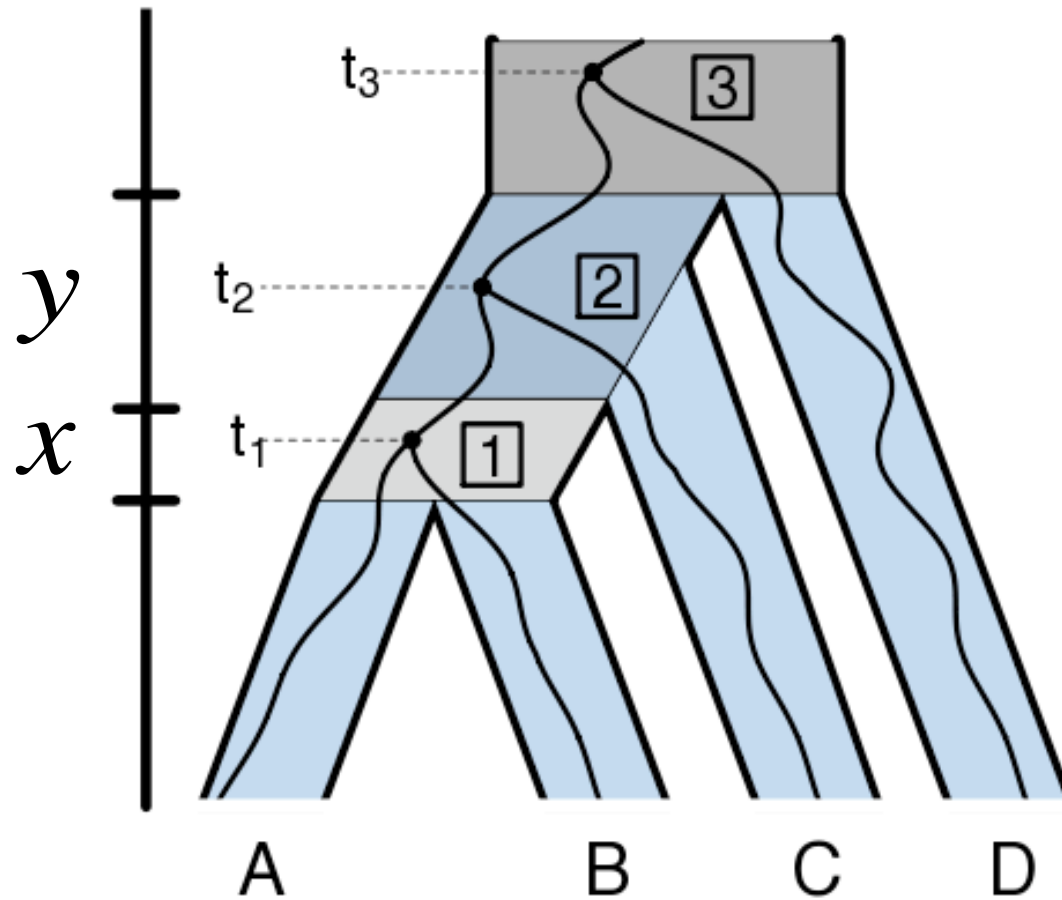
Probability: $p_{2,1}(x)p_{2,1}(y) = (1-X)(1-Y)$

Example

history $h_4 = (2,3,3)$
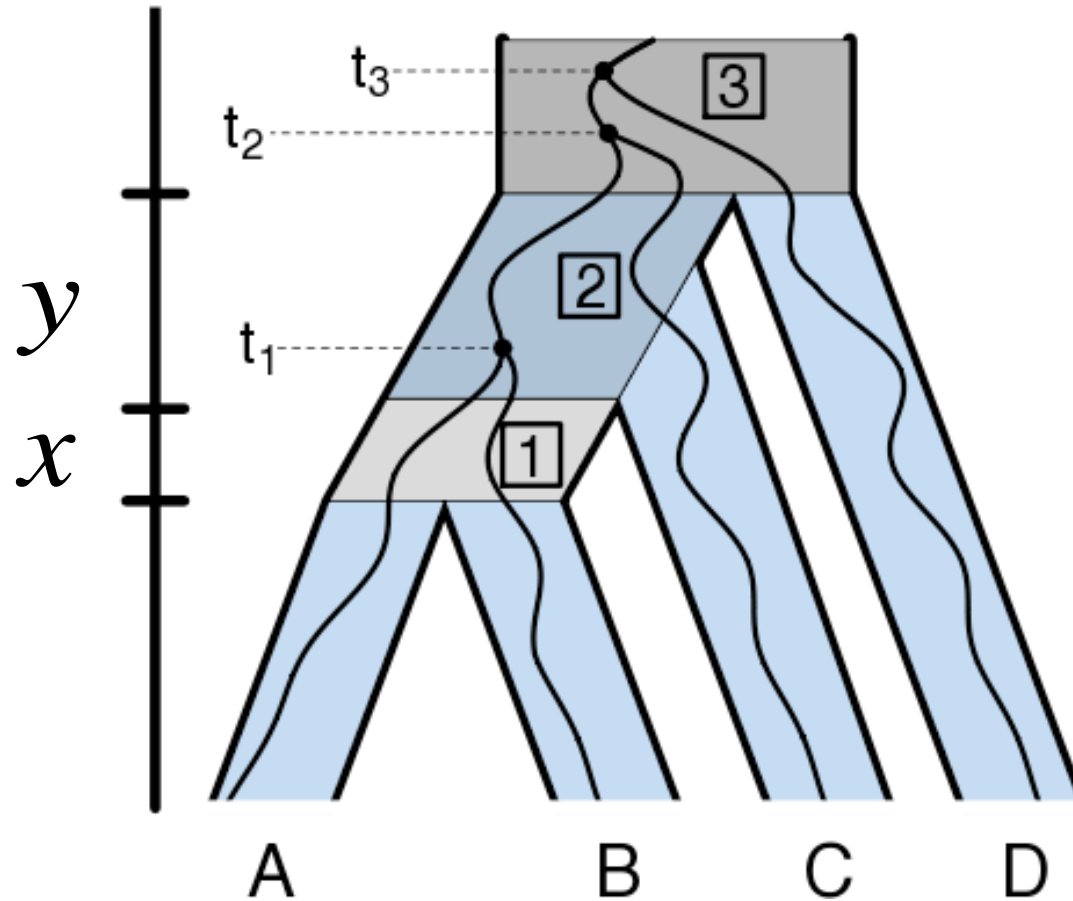
Probability: $\frac{1}{9} p_{2,2}(x) p_{3,2}(y) = \frac{1}{6} XY - \frac{1}{6} XY^3$

Total probability for matching gene tree, species tree $(((a,b){:}x,c){:}y,d)$

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

$$1-\tfrac{2}{3}X-\tfrac{2}{3}Y+\tfrac{1}{3}XY+\tfrac{1}{18}XY^3$$
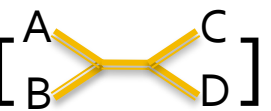
Identifiability of species trees from rooted gene trees

- Given the set of gene tree probabilities, can the species tree be recovered?

- In many cases, the highest probability gene tree has the same topology as the species tree, but not always.

- The most likely triple for any set of three taxa is a rooted triple on the species tree, so the species tree can be recovered by marginalizing gene trees to their rooted triples.

Part 3: Unrooted gene trees

Probabilities of unrooted gene trees

- The probability of an unrooted gene tree is the sum of the probabilities of all gene trees with the same unrooted topology

- $\Pr[\begin{smallmatrix}A \\ B\end{smallmatrix}\!\!\!\times\!\!\!\begin{smallmatrix}C \\ D\end{smallmatrix}]$ = P[((((AB)C)D)]
  $+ \ P[((((AB)D)C)]$
  $+ \ P[((((CD)A)B)]$
  $+ \ P[((((CD)B)A)]$
  $+ \ P[(((AB)(CD))]$

Probabilities of unrooted gene trees

- Probabilities of unrooted gene trees are linear combinations of probabilities of rooted gene trees

- In practice, expressions for probabilities of unrooted gene trees are often simpler than rooted gene tree probabilities for the same number of species.

Species tree (((a,b):x,c):y,d)

Unrooted Gene Trees
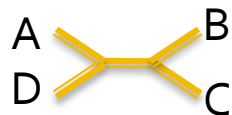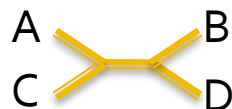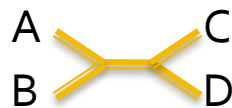
Probability

$$1 - \tfrac{2}{3} X$$

$$\tfrac{1}{3} X$$

$$\tfrac{1}{3} X$$

Species tree $(((a,b):x,(c,d):y)$
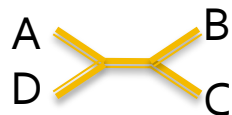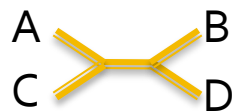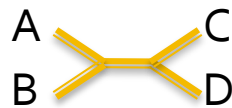
Unrooted Gene Trees	Probability
A ╲╱ C ╱╲ B D	$1 - \frac{2}{3}XY$
A ╲╱ B ╱╲ C D	$\frac{1}{3}XY$
A ╲╱ B ╱╲ D C	$\frac{1}{3}XY$

Results for four taxa

- If the species tree topology is known, the unrooted gene tree distribution only has information about one internal edge (or sum of edges).

- If the species tree is unknown, the unrooted gene tree distribution only identifies the unrooted species tree topology. The following species trees induce the same unrooted gene tree distribution:

$$((a, b):x, c):y_1, d),$$
$$((a, b):x, d):y_2, c),$$
$$((c, d):x, a):y_3, b),$$
$$((c, d):x, b):y_4, a),$$
$$((a, b):z, (c, d):x - z).$$

What happens with five taxa?

- 15 unrooted topologies
- 3 rooted species trees shapes

Caterpilar

Balanced

Pseudocaterpillar

What happens with five taxa?

- 15 unrooted topologies
- 3 rooted species trees shapes
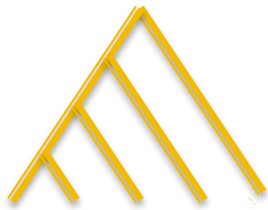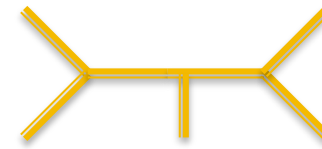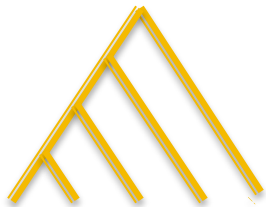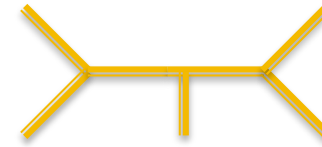
Caterpillar   Balanced   Pseudocaterpillar
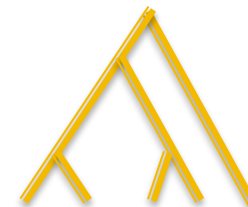
Given a distribution of 15 5-taxon unrooted gene tree probabilities, what information can we recover about the species tree?
(1) Unrooted species tree topology?
(2) Rooted species tree topology?
(3) Branch lengths?

Species tree (((a,b):x,c):y,(d,e):z)

T_1
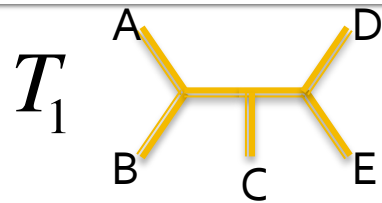
$$1 - \frac{2}{3}X - \frac{2}{3}YZ + \frac{1}{3}XYZ + \frac{1}{15}XY^3Z$$

Species tree (((a,b):x,c):y,(d,e):z)

T_1

A   D
B   E
  C

$$1 - \frac{2}{3}X - \frac{2}{3}YZ + \frac{1}{3}XYZ + \frac{1}{15}XY^3Z$$

T_2

A   C
B   E
  D
A   D

T_3

A   C
B   D
  E
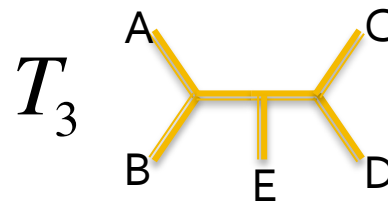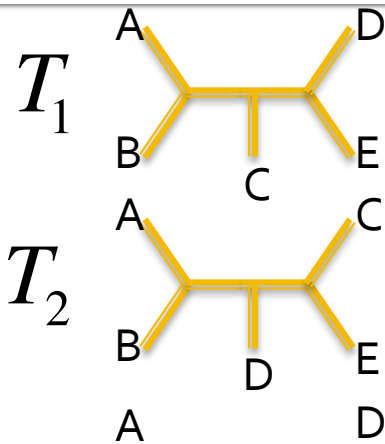
$$\frac{1}{3}YZ - \frac{1}{6}XYZ - \frac{1}{10}XY^3Z$$

Species tree (((a,b):x,c):y,(d,e):z)
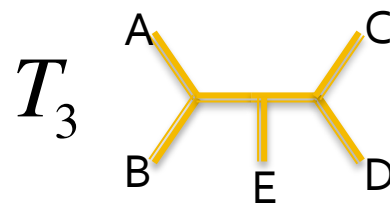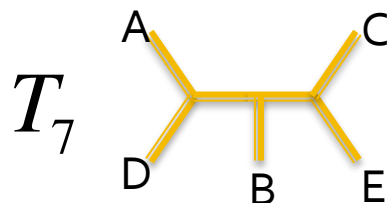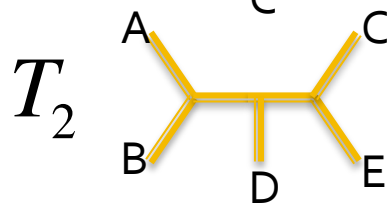
T_1

$1 - \frac{2}{3}X - \frac{2}{3}YZ + \frac{1}{3}XYZ + \frac{1}{15}XY^3Z$

T_2

T_3

$\frac{1}{3}YZ - \frac{1}{6}XYZ - \frac{1}{10}XY^3Z$

T_4

T_{13}

$\frac{1}{3}X - \frac{1}{3}XYZ + \frac{1}{15}XY^3Z$

T_5

$T_6 \quad T_9 \quad T_{12}$

$\frac{1}{6}XYZ - \frac{1}{10}XY^3Z$

T_7

5 others

$\frac{1}{15}XY^3Z$

Observations

(1) Putting unrooted gene trees that are tied in probability into equivalence classes, the size of these classes depends on the unlabeled, rooted species tree topology on five taxa:

Caterpillar class sizes: 1,1,1,2,2,2,6

Balanced class sizes: 1,2,2,4,6

Pseudocaterpillar sizes: 1,2,2,2,8

(2) The unlabeled, rooted species tree topology can therefore be determined from the class sizes.

Observations

(3) From the the four-taxon results, the unrooted species tree topology can be identified by determining the most probable quartet for each subset of four species

Therefore we know (for five taxa):
(i) labeled, unrooted species tree topology
(ii) unlabeled, rooted species tree topology

Observations

(4) 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 T_1 and given that the species tree is balanced, The rooted species tree is one of:
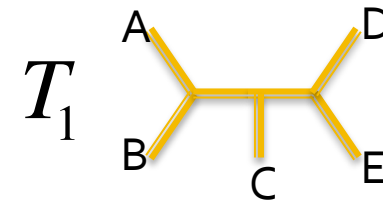
R_1 and R_2 imply different invariants and inequalities for the unrooted gene tree probabilities.

Under R_1, T_7 is in the 6-element class, and

$$\Pr(T_7) > \Pr(T_5)$$

Under R_2, T_7 is in the 4-element class, and

$$\Pr(T_7) < \Pr(T_5)$$

Identifying labeled, rooted topologies

- Similar arguments can be used to identify the caterpillar and pseudocaterpillar. Thus five-taxon rooted species trees are identifiable from five-taxon unrooted gene tree probabilities.

- The results generalize immediately to larger trees: Given a distribution of unrooted trees on more than five taxa, the unrooted gene tree distribution for each subset of five taxa can be obtained by summing over trees that display the five-taxon tree.

- Thus all rooted quintets on the species tree are identifiable. The rooted species tree topology can be constructed from the rooted quintets.

Recovering species tree branch lengths

- All branch lengths on five-taxon species trees can be recovered. Example, (((a,b):x,c):y,(d,e):z)

$$u_1 = 1 - \frac{2}{3}X - \frac{2}{3}YZ + \frac{1}{3}XYZ + \frac{1}{15}XY^3Z$$

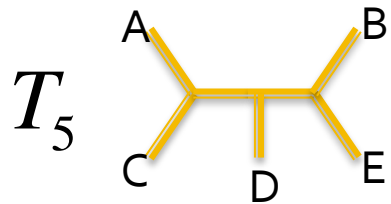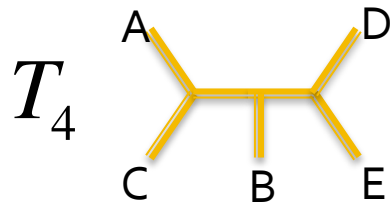$$u_2 = u_3 = \frac{1}{3}YZ - \frac{1}{6}XYZ - \frac{1}{10}XY^3Z$$

$$u_4 = u_{13} = \frac{1}{3}X - \frac{1}{3}XYZ + \frac{1}{15}XY^3Z$$

$$u_5 = u_6 = u_9 = u_{12} = \frac{1}{6}XYZ - \frac{1}{10}XY^3Z$$

$$u_7 = u_8 = u_{10} = u_{11} = u_{14} = u_{15} = \frac{1}{15}XY^3Z$$

$$x = -\log(X) = -\log\left[3u_4 + 6u_5 + 6u_7\right]$$

$$y = -\log(Y) = -\frac{1}{2}\log\left(\frac{5u_7}{2u_5 + 3u_7}\right)$$

$$z = -\log(Z) = -\log\left(\frac{3u_2 + 3u_5 + 9u_7}{Y}\right) = \frac{1}{2}\log\left(\frac{5u_7}{2u_5 + 3u_7}\right) - \log(3u_2 + 3u_5 + 9u_7)$$

Conclusion

Theorem. (i) The unrooted gene tree distribution determines the rooted species tree and branch lengths when there are 5 or more taxa. (ii) The unrooted gene tree distribution given a four-taxon species tree determines the unrooted species tree, but not the rooted species.

Five-taxon Inequalities

- Caterpillar:
$$u_1 > u_2, u_4 > u_5 > u_7$$
$$u_3 > u_2, u_6 > u_5 > u_7.$$

- Balanced:
$$u_1 > u_2, u_4 > u_5 > u_7$$

- Pseudocaterllar:
$$u_1 > u_2, u_4, u_{10} > u_5$$

Linear invariants on gene tree distributions

-- A gene tree distribution has only n-2 parameters (species tree branch lengths)

-- But there are (2n-3)!! gene tree probabilities

-- There are many ties in gene tree probabilities amongst different gene trees

-- There are other linear constraints. For the species tree (((AB)C)D), we have

$$\Pr[((AB)(CD))] - \Pr[(((AB)D)C)] - \Pr[(((AD)B)C)] = 0$$

$$0 = u_2 - u_3 + u_9 - u_{12}$$

-- How many linear constraints are there? How does the number of linear constraints depend on the number of taxa and species tree topology?

-- What are some nonlinear constraints?

Polynomial constraints

Probabilities of coalescent histories, and therefore of gene trees, are polynomials in the transformed branch lengths. An example polynomial constraint:

$$6p_3p_5 + 6p_5^2 + 3p_3p_{13} + 3p_5p_{13} + 3p_3p_{14} + 15p_5p_{14} + 6p_{13}p_{14} + 6p_{14}^2 - 2p_5 - 2p_{14} = 0.$$

Rooted five-taxon gene trees

TABLE 1. The 105 rooted gene trees on five species.

R_1	$((((A,B),C),D),E)$	R_{36}	$((((B,D),E),C),A)$	R_{71}	$(((A,D),(C,E)),B)$
R_2	$((((A,B),C),E),D)$	R_{37}	$((((B,E),A),C),D)$	R_{72}	$(((A,E),(C,D)),B)$
R_3	$((((A,B),D),C),E)$	R_{38}	$((((B,E),A),D),C)$	R_{73}	$(((B,C),(D,E)),A)$
R_4	$((((A,B),D),E),C)$	R_{39}	$((((B,E),C),A),D)$	R_{74}	$(((B,D),(C,E)),A)$
R_5	$((((A,B),E),C),D)$	R_{40}	$((((B,E),C),D),A)$	R_{75}	$(((B,E),(C,D)),A)$
R_6	$((((A,B),E),D),C)$	R_{41}	$((((B,E),D),A),C)$	R_{76}	$(((A,B),C),(D,E))$
R_7	$((((A,C),B),D),E)$	R_{42}	$((((B,E),D),C),A)$	R_{77}	$(((A,C),B),(D,E))$
R_8	$((((A,C),B),E),D)$	R_{43}	$((((C,D),A),B),E)$	R_{78}	$(((B,C),A),(D,E))$
R_9	$((((A,C),D),B),E)$	R_{44}	$((((C,D),A),E),B)$	R_{79}	$(((A,B),D),(C,E))$
R_{10}	$((((A,C),D),E),B)$	R_{45}	$((((C,D),B),A),E)$	R_{80}	$(((A,D),B),(C,E))$
R_{11}	$((((A,C),E),B),D)$	R_{46}	$((((C,D),B),E),A)$	R_{81}	$(((B,D),A),(C,E))$
R_{12}	$((((A,C),E),D),B)$	R_{47}	$((((C,D),E),A),B)$	R_{82}	$(((A,C),D),(B,E))$
R_{13}	$((((A,D),B),C),E)$	R_{48}	$((((C,D),E),B),A)$	R_{83}	$(((A,D),C),(B,E))$
R_{14}	$((((A,D),B),E),C)$	R_{49}	$((((C,E),A),B),D)$	R_{84}	$(((C,D),A),(B,E))$
R_{15}	$((((A,D),C),B),E)$	R_{50}	$((((C,E),A),D),B)$	R_{85}	$(((B,C),D),(A,E))$
R_{16}	$((((A,D),C),E),B)$	R_{51}	$((((C,E),B),A),D)$	R_{86}	$(((B,D),C),(A,E))$
R_{17}	$((((A,D),E),B),C)$	R_{52}	$((((C,E),B),D),A)$	R_{87}	$(((C,D),B),(A,E))$
R_{18}	$((((A,D),E),C),B)$	R_{53}	$((((C,E),D),A),B)$	R_{88}	$(((A,B),E),(C,D))$
R_{19}	$((((A,E),B),C),D)$	R_{54}	$((((C,E),D),B),A)$	R_{89}	$(((A,E),B),(C,D))$
R_{20}	$((((A,E),B),D),C)$	R_{55}	$((((D,E),A),B),C)$	R_{90}	$(((B,E),A),(C,D))$
R_{21}	$((((A,E),C),B),D)$	R_{56}	$((((D,E),A),C),B)$	R_{91}	$(((A,C),E),(B,D))$
R_{22}	$((((A,E),C),D),B)$	R_{57}	$((((D,E),B),A),C)$	R_{92}	$(((A,E),C),(B,D))$
R_{23}	$((((A,E),D),B),C)$	R_{58}	$((((D,E),B),C),A)$	R_{93}	$(((C,E),A),(B,D))$
R_{24}	$((((A,E),D),C),B)$	R_{59}	$((((D,E),C),A),B)$	R_{94}	$(((B,C),E),(A,D))$
R_{25}	$((((B,C),A),D),E)$	R_{60}	$((((D,E),C),B),A)$	R_{95}	$(((B,E),C),(A,D))$
R_{26}	$((((B,C),A),E),D)$	R_{61}	$(((A,B),(C,D)),E)$	R_{96}	$(((C,E),B),(A,D))$
R_{27}	$((((B,C),D),A),E)$	R_{62}	$(((A,C),(B,D)),E)$	R_{97}	$(((A,D),E),(B,C))$
R_{28}	$((((B,C),D),E),A)$	R_{63}	$(((A,D),(B,C)),E)$	R_{98}	$(((A,E),D),(B,C))$
R_{29}	$((((B,C),E),A),D)$	R_{64}	$(((A,B),(C,E)),D)$	R_{99}	$(((D,E),A),(B,C))$
R_{30}	$((((B,C),E),D),A)$	R_{65}	$(((A,C),(B,E)),D)$	R_{100}	$(((B,D),E),(A,C))$
R_{31}	$((((B,D),A),C),E)$	R_{66}	$(((A,E),(B,C)),D)$	R_{101}	$(((B,E),D),(A,C))$
R_{32}	$((((B,D),A),E),C)$	R_{67}	$(((A,B),(D,E)),C)$	R_{102}	$(((D,E),B),(A,C))$
R_{33}	$((((B,D),C),A),E)$	R_{68}	$(((A,D),(B,E)),C)$	R_{103}	$(((C,D),E),(A,B))$
R_{34}	$((((B,D),C),E),A)$	R_{69}	$(((A,E),(B,D)),C)$	R_{104}	$(((C,E),D),(A,B))$
R_{35}	$((((B,D),E),A),C)$	R_{70}	$(((A,C),(D,E)),B)$	R_{105}	$(((D,E),C),(A,B))$

Unrooted gene trees

Tree	Probability	splits
T_1	$u_1 = \mathbb{P}[((A,B),(D,E),C)] = r_1 + r_2 + r_{59} + r_{60} + r_{67} + r_{76} + r_{105}$	$AB|CDE, ABC|DE$
T_2	$u_2 = \mathbb{P}[((A,B),(C,E),D)] = r_3 + r_4 + r_{53} + r_{54} + r_{64} + r_{79} + r_{104}$	$AB|CDE, ABD|CE$
T_3	$u_3 = \mathbb{P}[((A,B),(C,D),E)] = r_5 + r_6 + r_{47} + r_{48} + r_{61} + r_{88} + r_{103}$	$AB|CDE, ABE|CD$
T_4	$u_4 = \mathbb{P}[((A,C),(D,E),B)] = r_7 + r_8 + r_{57} + r_{58} + r_{70} + r_{77} + r_{102}$	$AC|BDE, ABC|DE$
T_5	$u_5 = \mathbb{P}[((A,C),(B,E),D)] = r_9 + r_{10} + r_{41} + r_{42} + r_{65} + r_{82} + r_{101}$	$AC|BDE, ACD|BE$
T_6	$u_6 = \mathbb{P}[((A,C),(B,D),E)] = r_{11} + r_{12} + r_{35} + r_{36} + r_{62} + r_{91} + r_{100}$	$AC|BDE, ACE|BD$
T_7	$u_7 = \mathbb{P}[((A,D),(C,E),B)] = r_{13} + r_{14} + r_{51} + r_{52} + r_{71} + r_{80} + r_{96}$	$AD|BCE, ABD|CE$
T_8	$u_8 = \mathbb{P}[((A,D),(B,E),C)] = r_{15} + r_{16} + r_{39} + r_{40} + r_{68} + r_{83} + r_{95}$	$AD|BCE, ACD|BE$
T_9	$u_9 = \mathbb{P}[((A,D),(B,C),E)] = r_{17} + r_{18} + r_{29} + r_{30} + r_{63} + r_{94} + r_{97}$	$AD|BCE, ADE|BC$
T_{10}	$u_{10} = \mathbb{P}[((A,E),(C,D),B)] = r_{19} + r_{20} + r_{45} + r_{46} + r_{72} + r_{87} + r_{89}$	$AE|BCD, ABE|CD$
T_{11}	$u_{11} = \mathbb{P}[((A,E),(B,D),C)] = r_{21} + r_{22} + r_{33} + r_{34} + r_{69} + r_{86} + r_{92}$	$AE|BCD, ACE|BD$
T_{12}	$u_{12} = \mathbb{P}[((A,E),(B,C),D)] = r_{23} + r_{24} + r_{27} + r_{28} + r_{66} + r_{85} + r_{98}$	$AE|BCD, ADE|BC$
T_{13}	$u_{13} = \mathbb{P}[((B,C),(D,E),A)] = r_{25} + r_{26} + r_{55} + r_{56} + r_{73} + r_{78} + r_{99}$	$BC|ADE, ABC|DE$
T_{14}	$u_{14} = \mathbb{P}[((B,D),(C,E),A)] = r_{31} + r_{32} + r_{49} + r_{50} + r_{74} + r_{81} + r_{93}$	$BD|ACE, ABD|CE$
T_{15}	$u_{15} = \mathbb{P}[((B,E),(C,D),A)] = r_{37} + r_{38} + r_{43} + r_{44} + r_{75} + r_{84} + r_{90}$	$BE|ACD, ABE|CD$