Identifying Species Trees from Gene Trees

Elizabeth S. Allman

University of Alaska

IPAM

Los Angeles, CA

November 17, 2011

Workshop III: Evolutionary Genomics

Collaborators

The work in today's talk is joint with

John A. Rhodes
University of Alaska Fairbanks

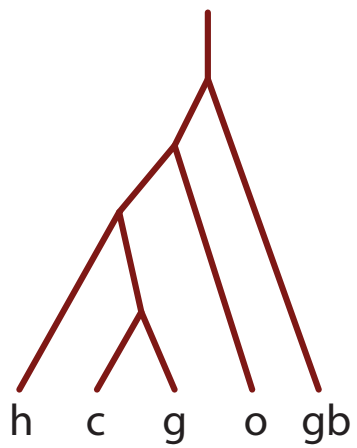
James H. Degnan
University of Canterbury
New Zealand

Gene trees vs. species trees

From genetic data like DNA sequences,

```
Gorilla      AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTTACATCAT...
Orangutan    AAGCTTCACCGGCGCAACCACCCTCATGATTGCCCATGGACTCACATCCT...
Human        AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGGCTTACATCCT...
Chimpanzee   AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTTACATCCT...
Gibbon       AAGCTTTACAGGTGCAACCGTCCTCATAATCGCCCACGGACTAACCTCTT...
```

phylogenetic methods construct **gene trees**.



It is well-known that

gene trees may differ from the underlying species tree.

Sources of conflict

There are many reasons gene trees may differ from species trees

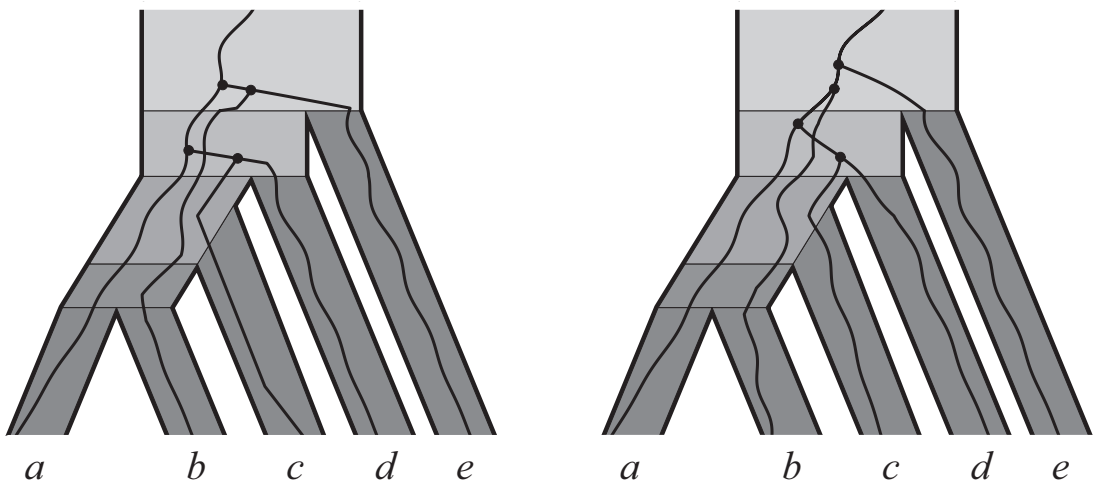
- lateral gene transfer (subtree-prune-and-regraft)
- hybridization (network)
- effects from population genetics

  — incomplete lineage sorting

Gene tree discordance

Different gene trees for the same set of taxa often disagree.

- One cause is incomplete lineage sorting.



Gene trees $(((C, D), A), (B, E))$ and $((((C, D), A), B), E)$

in species tree $((((a, b), c), d), e)$

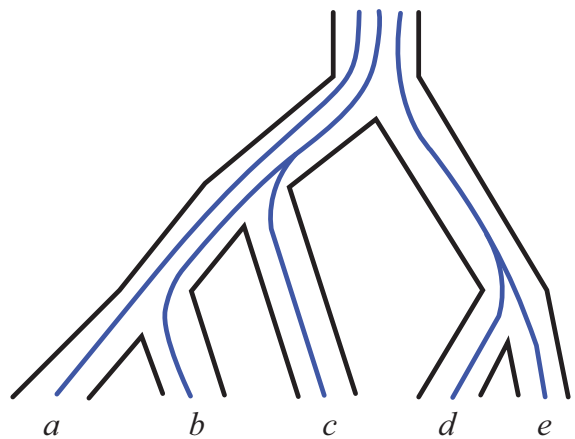
Multispecies coalescent model

Incomplete lineage sorting is modeled by the

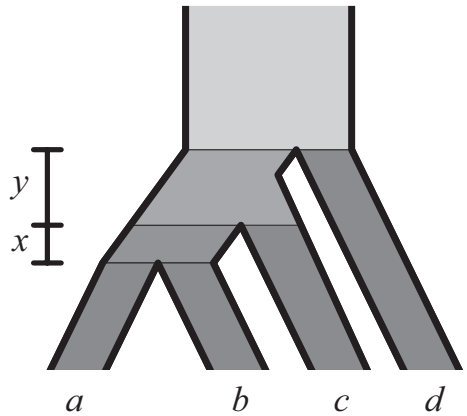
multispecies coalescent model.

- Viewing time backwards (present $\longrightarrow$ past), lineages within a population coalesce, one pair at a time.

The species tree constrains which lineages may coalesce at any given time.



Multispecies coalescent model



Coalescent events occur in 'populations,' or internal branches of species tree σ.

x, y in coalescent units ($\Delta x \propto \Delta t / N(t)$)

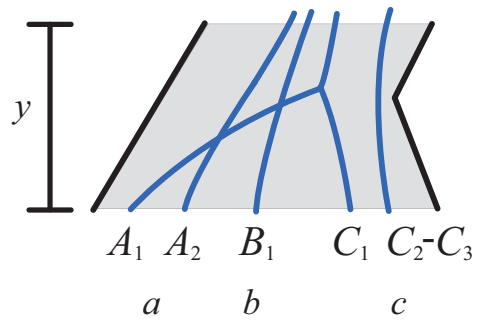
If k lineages enter a population from below, then

- coalescent events follow a Poisson process with rate $\binom{k}{2}$;

  Thus, the waiting time T_k for a coalescent event in any population is modeled by an exponential random variable, $T_k \sim \exp(\binom{k}{2})$.

- When a coalescent event occurs, any two lineages are equally likely to coalesce, with probability $\binom{k}{2}^{-1}$.

Multispecies coalescent model



Coalescent events occur in 'populations,' or internal branches of species tree σ.

x, y in coalescent units ($\Delta x \propto \Delta t / N(t)$)

If k lineages enter a population from below, then

- coalescent events follow a Poisson process with rate $\binom{k}{2}$;

  Thus, the waiting time T_k for a coalescent event in any population is modeled by an exponential random variable, $T_k \sim \exp(\binom{k}{2})$.

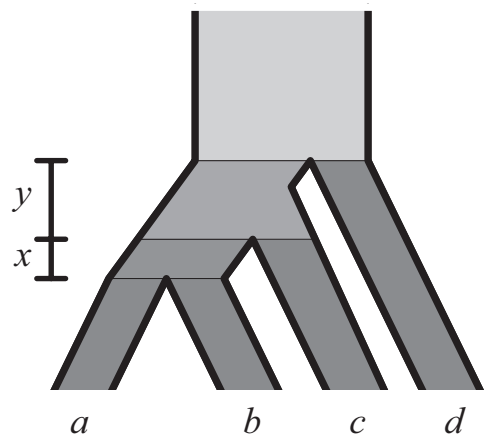
- When a coalescent event occurs, any two lineages are equally likely to coalesce, with probability $\binom{k}{2}^{-1}$.

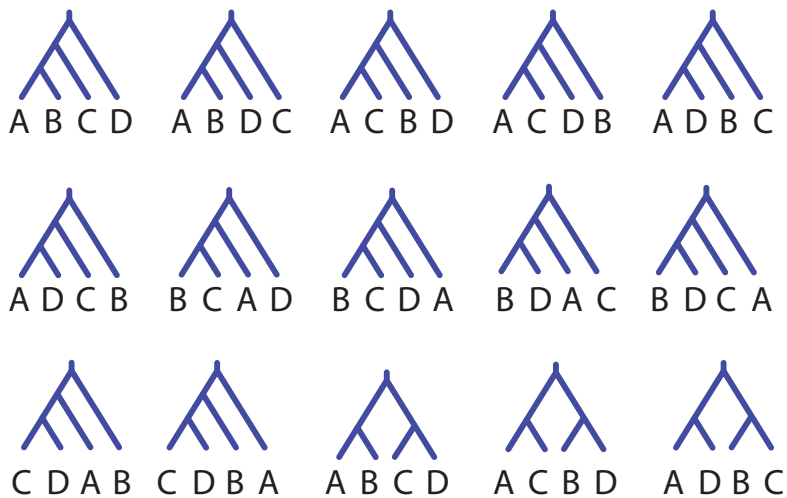
Gene tree distributions

Given a species tree σ on X, one can compute the (rooted)

gene tree distribution.

This distribution is parameterized by the species tree topology ψ and internal branch lengths $\boldsymbol{\lambda}$; $\sigma = (\psi, \boldsymbol{\lambda})$.





Discordance in gene trees can indicate the species tree.

Gene tree distributions

Given a species tree σ on X, one can compute the (rooted)

gene tree distribution.

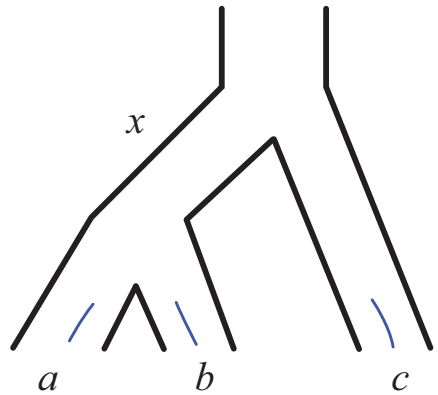
This distribution is parameterized by the species tree topology ψ and internal branch lengths $\boldsymbol{\lambda}$; $\sigma = (\psi, \boldsymbol{\lambda})$.

- Nei 1987: Probs for rooted gene trees with 3 species derived
- Pamilo and Nei 1988: Probs that rooted gene tree matches the species tree topology for 4 and 5 species derived
- Rosenberg 2002: Probs for all 30 species trees/gene tree combinations for 4 species
- Degnan and Salter 2005: General case solved and implemented in COAL

Discordance in gene trees can indicate the species tree.

Ex: Gene tree distribution

3 taxa, species tree
$\sigma = ((a, b){:}x, c)$



The probability that two specific entering lineages have *not* coalesced before x units is $X = e^{-x} = 1 - \int_0^x e^{-\binom{2}{2}t} dt$.

So topological gene trees have probabilities

$$\begin{aligned} \Pr((A, B), C) &= 1 - 2X(1/3) && \geq 1/3, \\ \Pr((A, C), B) &= X\,(1/3) && \leq 1/3, \\ \Pr((B, C), A) &= X\,(1/3) && \leq 1/3. \end{aligned}$$

Ex: Gene tree distribution

In particular, for species tree $\sigma = ((a, b){:}x, c)$,

$$\Pr((A, B), C) > \Pr((A, C), B) = \Pr((B, C), A),$$

and gap in inequality identifies branch length x.

That is, the 3-taxon species tree $\sigma = (\psi, \boldsymbol{\lambda})$ is **identifiable** from gene tree probabilities:

- the most probable gene tree matches the species tree;
- if m is the value of the smallest gene tree probability, solve $m = \frac{1}{3}X$ for $X = \exp(-x)$ to obtain the branch length x.

By marginalization to 3-taxon subsets, this shows for $n \geq 3$ taxa, metric rooted species trees can be identified from n-taxon rooted topological gene tree probabilities.

Empirical examples

Ebersberger et al., 2007:

For multiple loci on $\{H, C, G\}$, applying a rooted triple method gives further evidence for ((human,chimp),gorilla) tree, as well as dating divergences in it.

Jennings and Edwards, 2005; Wakeley, 2008:

A rooted triple method can give evidence for evolutionary relationships between a triple of Australian grass finches.

Motivating Questions

How can species trees be inferred?

How can we summarize the gene tree dist and still retain enough information to recover the species tree σ?

In today's talk,

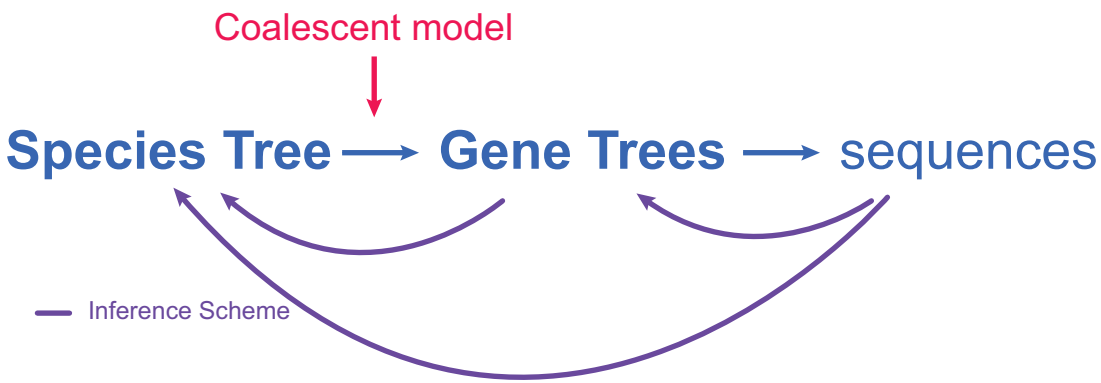
focus on **identifying** species trees from summary statistics for topological gene trees (not practical inference)

Motivating Questions

How can species trees be inferred?

How can we summarize the gene tree dist and still retain enough information to recover the species tree σ?

In today's talk,



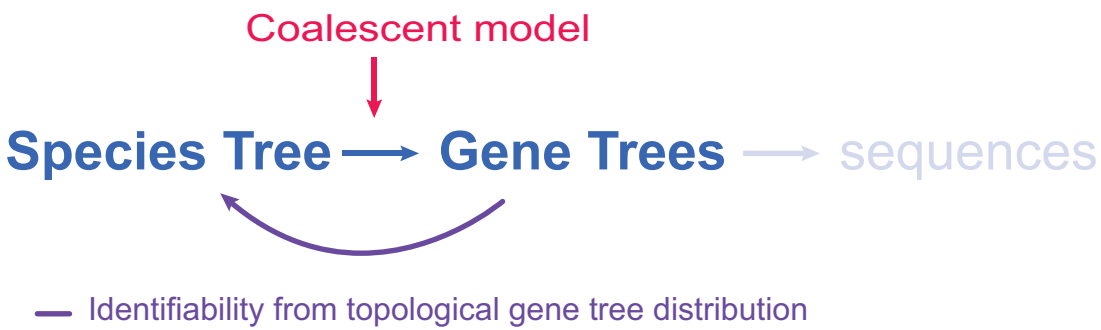
focus on **identifying** species trees from summary statistics for topological gene trees (not practical inference)

Motivating Questions

How can species trees be inferred?

How can we summarize the gene tree dist and still retain enough information to recover the species tree σ?

In today's talk,



focus on **identifying** species trees from summary statistics for topological gene trees (not practical inference)

Main Results

In (ADRa, ADRb, 2011), we prove that

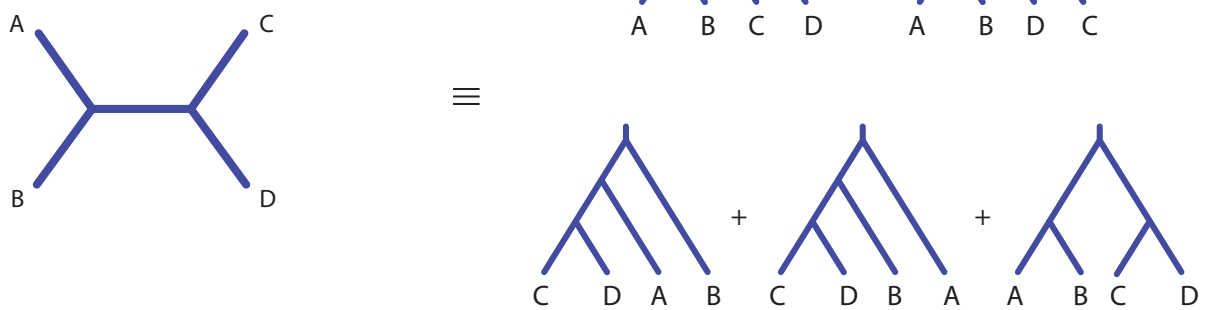
1. Species trees topologies ψ and branch lengths $\boldsymbol{\lambda}$ are identifiable from the distribution of
   unrooted topological gene tree probabilities.

2. Species tree topologies ψ are identifiable from
   clade probabilities.

Unrooted gene tree distribution

The ugt distribution is obtained by summing the probabilities of all rooted gene trees with the same unrooted topology.

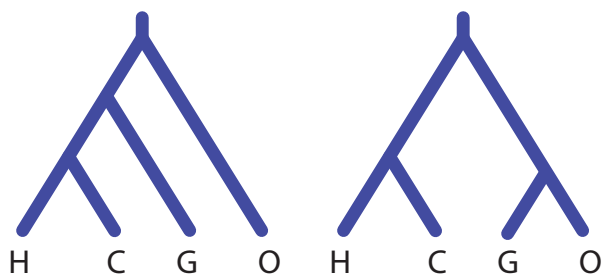
$$P[AB \mid CD] = P[(((AB)C)D)] + P[(((AB)D)C)] + P[(((CD)A)B)] + P[(((CD)B)A)] + P[((AB)(CD))]$$



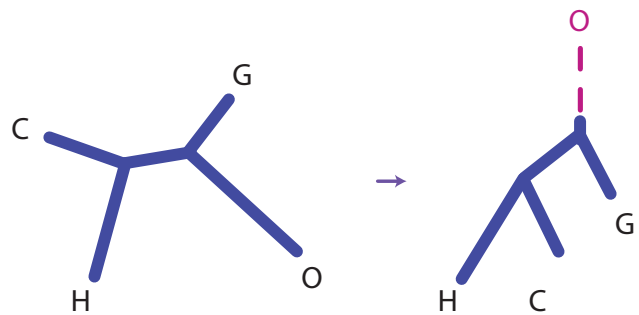
Motivation

Why look at unrooted topological gene trees?

- Many species tree methods using the coalescent currently assume or co-estimate rooted metric gene trees (BEST, *BEAST, STEM, ...)
- To infer roots of gene trees, you need
  - a molecular clock hypothesis (unrealistic?)
  - or an outgroup (non-statistical, perhaps biased placement)



MC ≡ each tree has constant root to tip distance



Outgroup is used to root tree, then deleted.

Motivation

Why look at unrooted topological gene trees?

- Many species tree methods using the coalescent currently assume or co-estimate rooted metric gene trees (BEST, *BEAST, STEM, ...)
- To infer roots of gene trees, you need
  - a molecular clock hypothesis (unrealistic?)
  - or an outgroup (non-statistical, perhaps biased placement)
- Inferred branch lengths in metric gene trees can be sensitive to model used in inference;

  gene tree topology is more robust to model choice

Results (ugt)

Theorem (ADR). Unrooted topological gene tree probabilities determine the metric species tree.

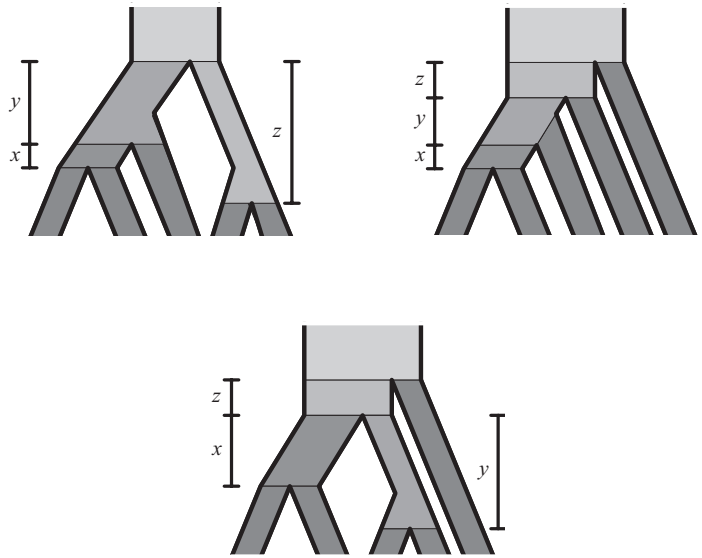
- For 4 taxa, the unrooted gene tree distribution determines the unrooted species tree, but not the rooted species tree.
- The unrooted gene tree distribution determines the rooted species tree and branch lengths when there are 5 or more taxa.
- These results remain valid for species trees with polytomies (non-binary trees).

This theorem justifies the use of ML or Bayesian methods to infer species trees from unrooted topological gene tree probabilities.

Method of proof (Part I)

Let u_i denote the probability of unrooted gene tree T_i.

Tree	Splits	Prob.
T_1	$AB\|CDE$, $ABC\|DE$	u_1
T_2	$AB\|CDE$, $ABD\|CE$	u_2
T_3	$AB\|CDE$, $ABE\|CD$	u_3
T_4	$AC\|BDE$, $ABC\|DE$	u_4
T_5	$AC\|BDE$, $ACD\|BE$	u_5
T_6	$AC\|BDE$, $ACE\|BD$	u_6
T_7	$AD\|BCE$, $ABD\|CE$	u_7
T_8	$AD\|BCE$, $ACD\|BE$	u_8
T_9	$AD\|BCE$, $ADE\|BC$	u_9
T_{10}	$AE\|BCD$, $ABE\|CD$	u_{10}
T_{11}	$AE\|BCD$, $ACE\|BD$	u_{11}
T_{12}	$AE\|BCD$, $ADE\|BC$	u_{12}
T_{13}	$BC\|ADE$, $ABC\|DE$	u_{13}
T_{14}	$BD\|ACE$, $ABD\|CE$	u_{14}
T_{15}	$BE\|ACD$, $ABE\|CD$	u_{15}



- For 5-taxon species trees σ, we show that different topologies satisfy different equations and inequalities in the u_i.
- Once the topology is known, internal branch lengths can be computed.
- Generalize from 5 to n.

Note:

ADR theorem is an *identifiability* result.

Proof method not intended as practical inference method of rooted, metric species trees.

Liu and Yu in *Sys. Biol.* 2011

"Estimating species trees from unrooted gene trees"

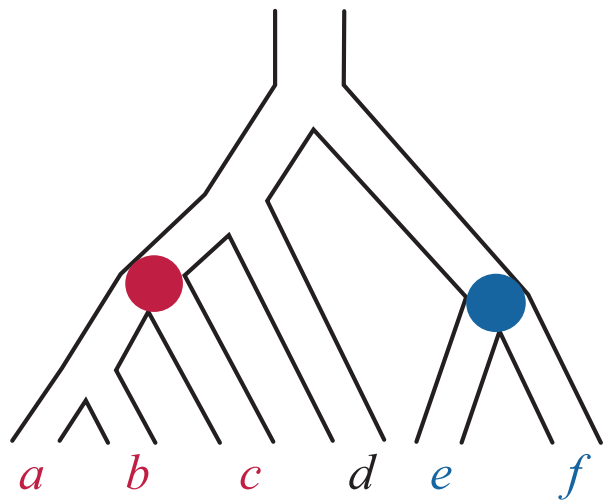
- give consistent algorithm to estimate *unrooted topological* species trees from unrooted topological gene trees

  Method closely related to the STAR algorithm of Liu, Yu, Pearl, and Edwards. More to come on this ....

- Kreidl 2011, arXiv: rigorous proof of consistency of algorithm

Identifiability from clade probabilities

Defn: A *clade* $\mathcal{C}$ on a species tree σ is the set of taxa all descended from a node of σ.



Clades $\mathcal{C}_1 = \{a, b, c\}$ and $\mathcal{C}_2 = \{e, f\}$ depicted at left.

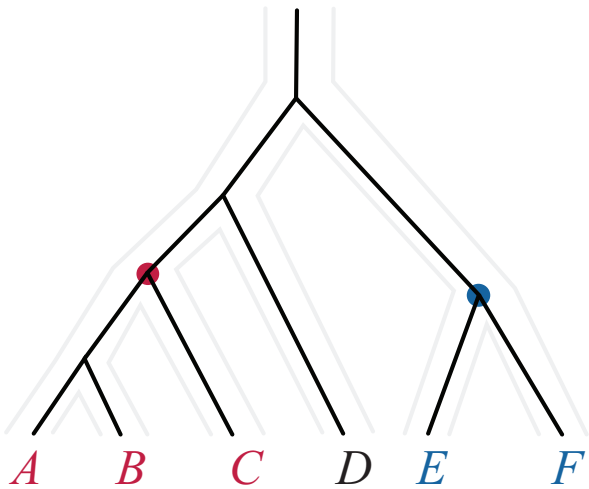
A clade on a gene tree T is defined similarly.

Question:

Do clade probabilities on gene trees determine the species tree?

Identifiability from clade probabilities

Defn: A *clade* $\mathcal{C}$ on a species tree σ is the set of taxa all descended from a node of σ.



Clades $\mathcal{C}_1 = \{a, b, c\}$ and $\mathcal{C}_2 = \{e, f\}$ depicted at left.

A clade on a gene tree T is defined similarly.

Question:

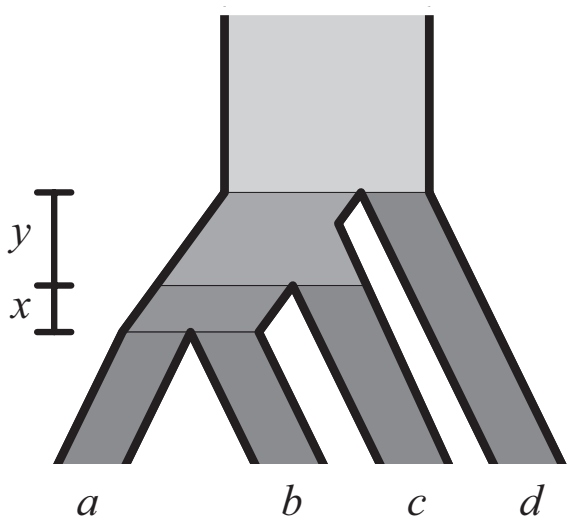
Do clade probabilities on gene trees determine the species tree?

Why clades?

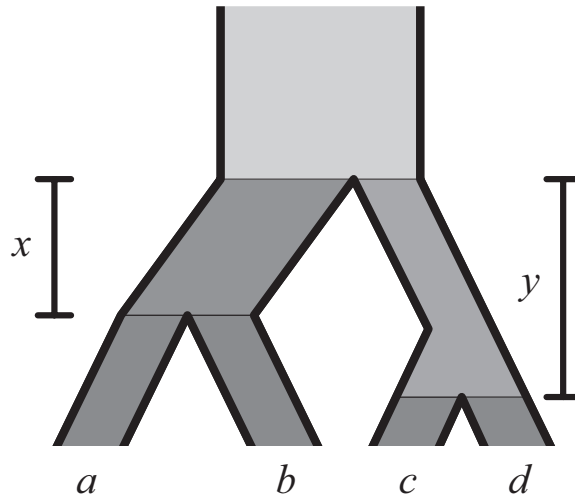
Clades on gene trees are natural to consider since they

- are inherent to the coalescent;
- summarize a collection of gene trees;
- don't depend on metric information on gene trees; (addresses lack of robustness)
- are already being used to deal with the gene tree/species trees differences. (BUCKy)

Eg. Clade probabilities on 4-taxon trees



Caterpillar tree



Balanced tree

Clade probabilities are functions of parameters $\sigma = (\psi, \boldsymbol{\lambda})$.

They do *not* form a distribution.

Eg. Clade probabilities on 4-taxon trees

Probabilities of clades for 4-taxon species trees under the coalescent. $X = \exp(-x)$, $Y = \exp(-y)$.

	probability under species tree	
clade	$(((a,b):x,c):y,d)$	$((a,b):x,(c,d):y)$
$c_1 = \Pr(AB)$	$1 - \frac{2}{3}X - \frac{1}{9}XY^3$	$1 - \frac{2}{3}X - \frac{1}{9}XY$
$c_2 = \Pr(AC)$	$\frac{1}{3}X - \frac{1}{9}XY^3$	$\frac{2}{9}XY$
$c_3 = \Pr(AD)$	$\frac{1}{6}XY + \frac{1}{18}XY^3$	$\frac{2}{9}XY$
$c_4 = \Pr(BC)$	$\frac{1}{3}X - \frac{1}{9}XY^3$	$\frac{2}{9}XY$
$c_5 = \Pr(BD)$	$\frac{1}{6}XY + \frac{1}{18}XY^3$	$\frac{2}{9}XY$
$c_6 = \Pr(CD)$	$\frac{1}{3}Y - \frac{1}{6}XY + \frac{1}{18}XY^3$	$1 - \frac{2}{3}Y - \frac{1}{9}XY$
$c_7 = \Pr(ABC)$	$1 - \frac{2}{3}Y - \frac{1}{3}XY + \frac{1}{6}XY^3$	$\frac{1}{3}Y - \frac{1}{6}XY$
$c_8 = \Pr(ABD)$	$\frac{1}{3}Y - \frac{1}{6}XY$	$\frac{1}{3}Y - \frac{1}{6}XY$
$c_9 = \Pr(ACD)$	$\frac{1}{6}XY$	$\frac{1}{3}X - \frac{1}{6}XY$
$c_{10} = \Pr(BCD)$	$\frac{1}{6}XY$	$\frac{1}{3}X - \frac{1}{6}XY$

† These parameterizations give rise to algebraic varieties V_σ.

High probability clades

Theorem. Suppose $\sigma = (\psi, \boldsymbol{\lambda})$ is a species tree (not necessarily binary), and that $\mathcal{C}$ is a gene tree clade with $\Pr_\sigma(\mathcal{C}) > 1/3$. Then $\mathcal{C}$ is a clade on σ.

- Partial analog for rooted triple gene tree probability.
- But a clade $\mathcal{C}$ can be on σ, with $\Pr(\mathcal{C}) \leq 1/3$.

Marginalization fails

For a proof of identifiability of the species tree topology from clade probabilities, it was not possible to argue from 'small' to 'large' trees.

For example, suppose

- σ is a species tree on $X = \{a, b, c, d\}$, and
- υ is the induced species tree on the 3-taxon subset $\{a, b, c\}$.

Then,

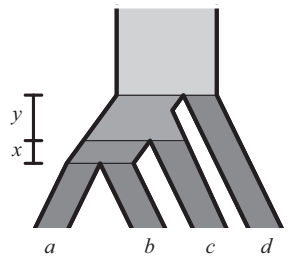
> there does *NOT* exist any linear combination of the clade probabilities on σ that gives the clade probabilities on υ that is valid for all species tree topologies.

i.e. We can't relate clade probabilities on trees and their subtrees..

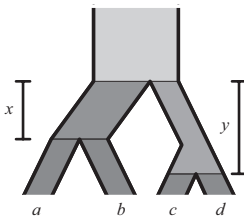
Algebraic proofs of identifiability

Clade probabilities for a particular species tree topology satisfy equations specific to that topology.

Eg. Identifiability of 4-taxon σ:



Caterpillar tree



Balanced tree

Probabilities of clades for 4-taxon species trees under the coalescent.
$X = \exp(-x)$, $Y = \exp(-y)$.

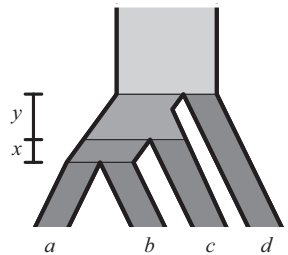
	probability under species tree	
clade	$(((a,b):x,c):y,d)$	$((a,b):x,(c,d):y)$
$c_2 = \Pr(AC)$	$\frac{1}{3}X - \frac{1}{9}XY^3$	$\frac{2}{9}XY$
$c_3 = \Pr(AD)$	$\frac{1}{6}XY + \frac{1}{18}XY^3$	$\frac{2}{9}XY$

For example, $c_2 = c_3$ for clade probabilities arising from the balanced tree, but $c_2 \neq c_3$ for the caterpillar tree.

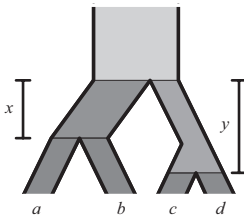
Algebraic proofs of identifiability

Clade probabilities for a particular species tree topology satisfy equations specific to that topology.

Eg. Identifiability of 4-taxon σ:



Caterpillar tree



Balanced tree

Conclusion:

4-taxon species tree topologies are identifiable from $\{c_i\}$.

For example, $c_2 = c_3$ for clade probabilities arising from the balanced tree, but $c_2 \neq c_3$ for the caterpillar tree.

Identifiability of small σ

For small species trees, we were able to find polynomial equations that exactly determined the species tree topology.

(used computational algebra package `Singular` ...)

This proved identifiability for small species tree topologies.

But,

4-, 5-taxon species tree topology identifiability

$\not\Rightarrow$ n-taxon tree identifiability result

For a general theorem,

understanding of relationships between clade probabilites *without computation* was needed.

Polynomials relations

'Cherry swapping' invariants

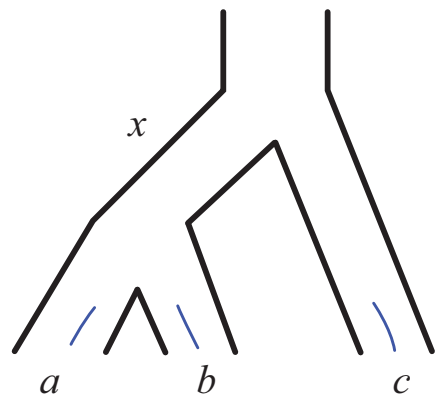
Proposition. If (a, b) is a 2-clade on the species tree, then for any non-empty set $\mathcal{D}$ of taxa excluding A, B,

$$\Pr(\mathcal{D} \cup \{A\}) = \Pr(\mathcal{D} \cup \{B\}).$$

For example,

$$\Pr(\{AC\}) = \Pr(\{BC\}).$$

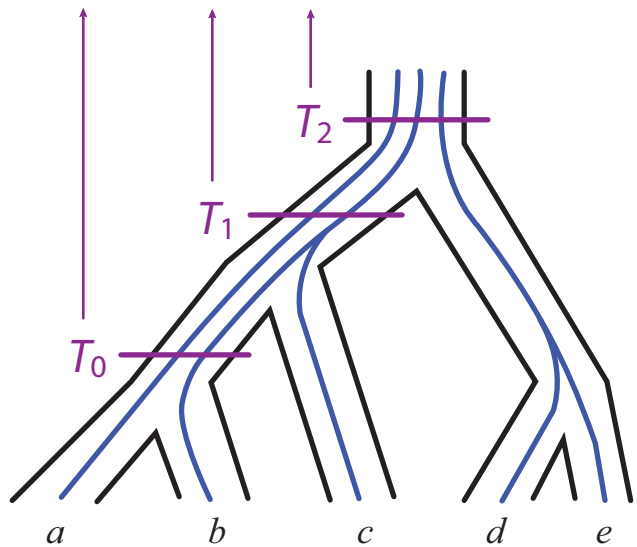
(Equiv. $\Pr(\{AC\}) - \Pr(\{BC\})$ is an invariant for σ.)



Proof: Under the coalescent model, lineages for A, B are exchangeable (i.e., behave identically).

Exchangeability of gene lineages

If two lineages are present in the same population at a particular point in time on σ, then above that point both lineages behave exactly the same way.



Lineages A, B behave exactly the same into the past, for $T < T_0$.

Similarly, lineages A, B-C behave the same for $T < T_1$, and lineages A, B-C, D-E for $T < T_2$.

More polynomial relations

Since no other polynomial relationships for clade probabilities were obvious, from `Singular` we obtain polynomials invariants for small trees.

Ex: Species tree $(((a,b):x,c):y,d)$: $(c_i = \text{probability of clade } i)$

$$stdJHcat[1] = c9 - c10$$
$$stdJHcat[2] = c5 - c6 + c8 - c10$$
$$stdJHcat[3] = c3 - c6 + c8 - c10$$
$$stdJHcat[4] = c2 - c4$$
$$stdJHcat[5] = c1 + 2 * c4 + 9 * c6 - c7 - 11 * c8 - 4 * c10$$
$$stdJHcat[6] = 3 * c4 * c8 + 6 * c6 * c8 - 6 * c8^2 + 3 * c4 * c10 + 12 * c6 * c10 - 2 * c7 * c10 - \ldots$$
$$stdJHcat[7] = 9 * c6^3 - 6 * c6^2 * c7 + c6 * c7^2 - 39 * c6^2 * c8 + 16 * c6 * c7 * c8 - \ldots$$
$$stdJHcat[8] = 9 * c4 * c6^2 - 3 * c4 * c6 * c7 + 6 * c6^2 * c7 - 2 * c6 * c7^2 + 60 * c6^2 * c8 - \ldots$$
$$stdJHcat[9] = 9 * c4^2 * c6 + 12 * c4 * c6 * c7 + 4 * c6 * c7^2 - 84 * c6^2 * c8 + \ldots$$

Insight from computations

For some 5-taxon species trees, elimination calculation did not terminate, but we do obtain some polynomials of low degree.

Such calculations ...

and much staring to discern a pattern ...

led to formulating the following theorem.

Polynomials relations determine σ

Theorem. Let $\mathcal{C}$ be a subset of taxa with at least two elements, and $\mathcal{D}$ a non-empty set of taxa disjoint from $\mathcal{C}$.

For distinct $a, b \in \mathcal{C}$, let $\mathcal{C}' = \mathcal{C} \smallsetminus \{a, b\}$.

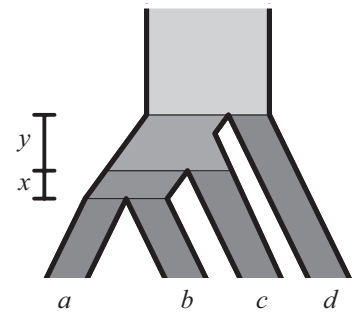
Then if $\mathcal{C}$ is a clade on the species tree σ, the polynomial $f =$

$$\left(\sum_{\mathcal{S} \subseteq \mathcal{C}'} \Pr(\mathcal{S} \cup \{a\} \cup \mathcal{D})\right) - \left(\sum_{\mathcal{S} \subseteq \mathcal{C}'} \Pr(\mathcal{S} \cup \{b\} \cup \mathcal{D})\right) = 0.$$

Moreover, if $\mathcal{C}$ is not a clade on the species tree, then for generic edge lengths the above equality does not hold.

Example

Caterpillar species tree
$\sigma = (((a, b){:}x, c){:}y, d)$



Clade $\mathcal{C} = \{a, b, c\}$; $\mathcal{C}' = \mathcal{C} \setminus \{a, c\} = \{b\}$; $\mathcal{D} = \{d\}$

$$\left(\sum_{\mathcal{S} \subseteq \mathcal{C}'} \Pr(\mathcal{S} \cup \{a\} \cup \mathcal{D})\right) - \left(\sum_{\mathcal{S} \subseteq \mathcal{C}'} \Pr(\mathcal{S} \cup \{c\} \cup \mathcal{D})\right) = 0.$$

or

$$\Big(\Pr\{a, d\} + \Pr\{a, b, d\}\Big) - \Big(\Pr\{c, d\} + \Pr\{b, c, d\}\Big) = 0.$$

Results (clade probs)

This leads to

Theorem (ADR).

> Clade probabilities determine the species tree topology for generic edge lengths.

For small trees, it is also possible to recover branch lengths, but no general method is known.

STAR and clades

Liu, Yu, Pearl, and Edwards (*Sys. Biol.* 2009):

introduce the STAR algorithm for estimating species tree topologies from gene trees.

(2012? ADR give a rigorous proof that STAR consistently estimates ψ.)

STAR and clades

Outline:

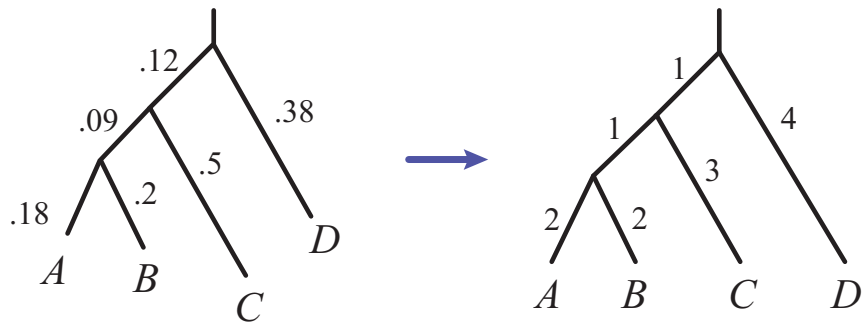
Given a collection of rooted (possibly metric) gene trees on n taxa,

1. For each gene tree, make all internal edges length 1, terminal edges lengths as needed so root to leaf distance is n.
2. Compute pairwise distances on each modified gene tree.
3. To find the distance $d_{STAR}(a, b)$ between species a and b, average distances over all gene trees.
4. Use a distance method to reconstruct the species tree topology, ψ.

... extends to multiple individuals sampled per species

STAR distances

For a rooted n-taxon gene tree, assign 1 to each internal edge and assign weights to terminal edges so that root to tip distance is n.



Original gene tree $\longrightarrow$ Modified gene tree.

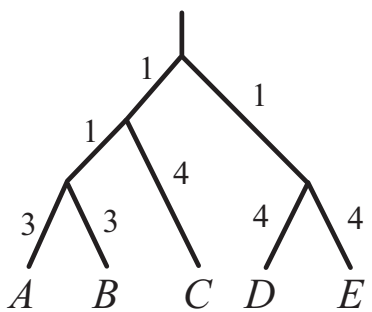
On the modified gene tree, $d(A, C) = 6$ for example.

Connection: STAR and $\{\Pr(c_i)\}$

Theorem (ADR). Suppose $\sigma = (\psi, \boldsymbol{\lambda})$ is an n-taxon rooted binary metric species tree. Then the clade probabilities $\{\Pr(\mathcal{C})\}$ determine the expected STAR distances. (Empirical version too.)

(And both determine ψ.)

Idea:



On this gene tree, the number of non-trivial clades containing A, B is 2;

That is, the distance from the root to $MRCA(A, B)$ is 2.

$$d_{STAR}(A, B) = 2n - 2(\text{number of clades containing } \{A, B\}).$$

Connection: STAR and $\{\Pr(c_i)\}$

Theorem (ADR). Suppose $\sigma = (\psi, \boldsymbol{\lambda})$ is an n-taxon rooted binary metric species tree. Then the clade probabilities $\{\Pr(\mathcal{C})\}$ determine the expected STAR distances. (Empirical version too.)

This idea underlies

$$d_{STAR}(a,b) = 2n - 2 \sum_{\substack{\text{non-trivial} \\ \text{clades } \mathcal{C}}} \Pr(\mathcal{C}) I_{A,B}(\mathcal{C}).$$

Remark: This gives an algorithm for computing the species tree topology from $\{\Pr(c_i)\}$, and explains the underpinnings of STAR.

The first taxa joined in STAR are those which appear in the most observed clades.....

Thank you.