Identifying species trees from clade probabilities

James H. Degnan

University of Canterbury
New Zealand

Phylomania, 4-5 Nov 2010

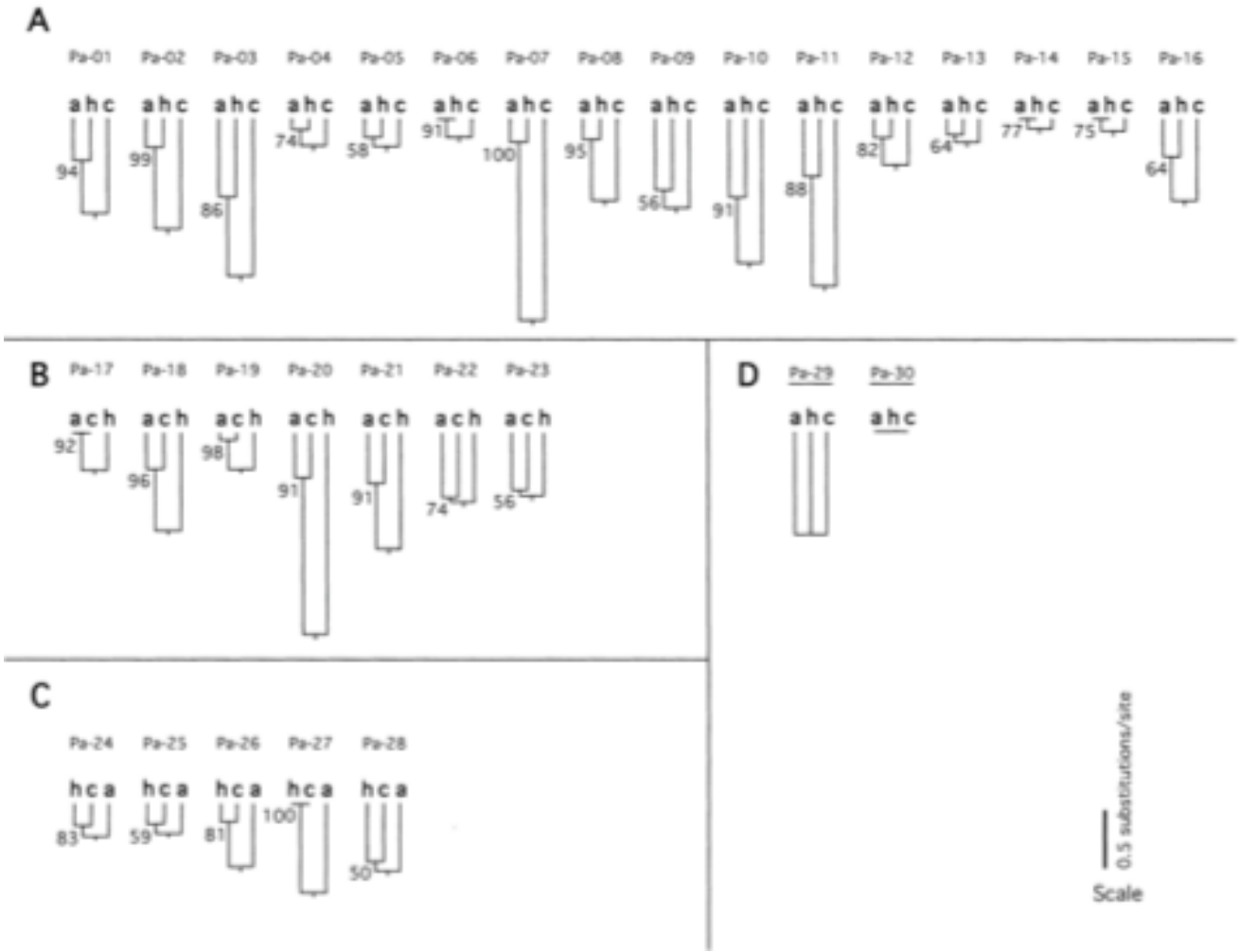
Joint work with
Elizabeth Allman and John Rhodes
University of Alaska Fairbanks

Outline

- Species trees and gene trees
- Inferring species trees from clade support/probabilities (greedy consensus)
- Invariants on clade probabilities
- Conclusion

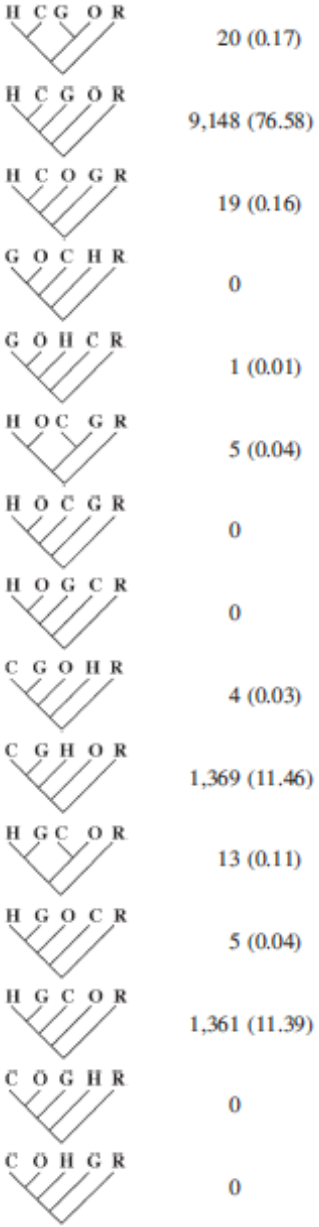
Several loci, several gene trees

From Jennings and Edwards, Evolution (2005), 30 loci for 3 ingroup species of Australian grass finches



Many loci, many gene trees

From Ebersberger et al., 2007,
Mol. Biol. Evol.

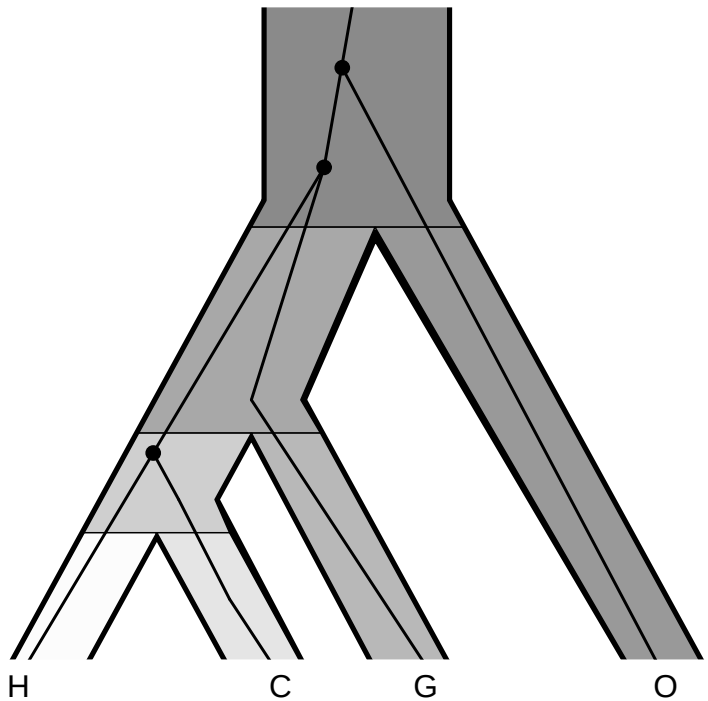


The multispecies coalescent model

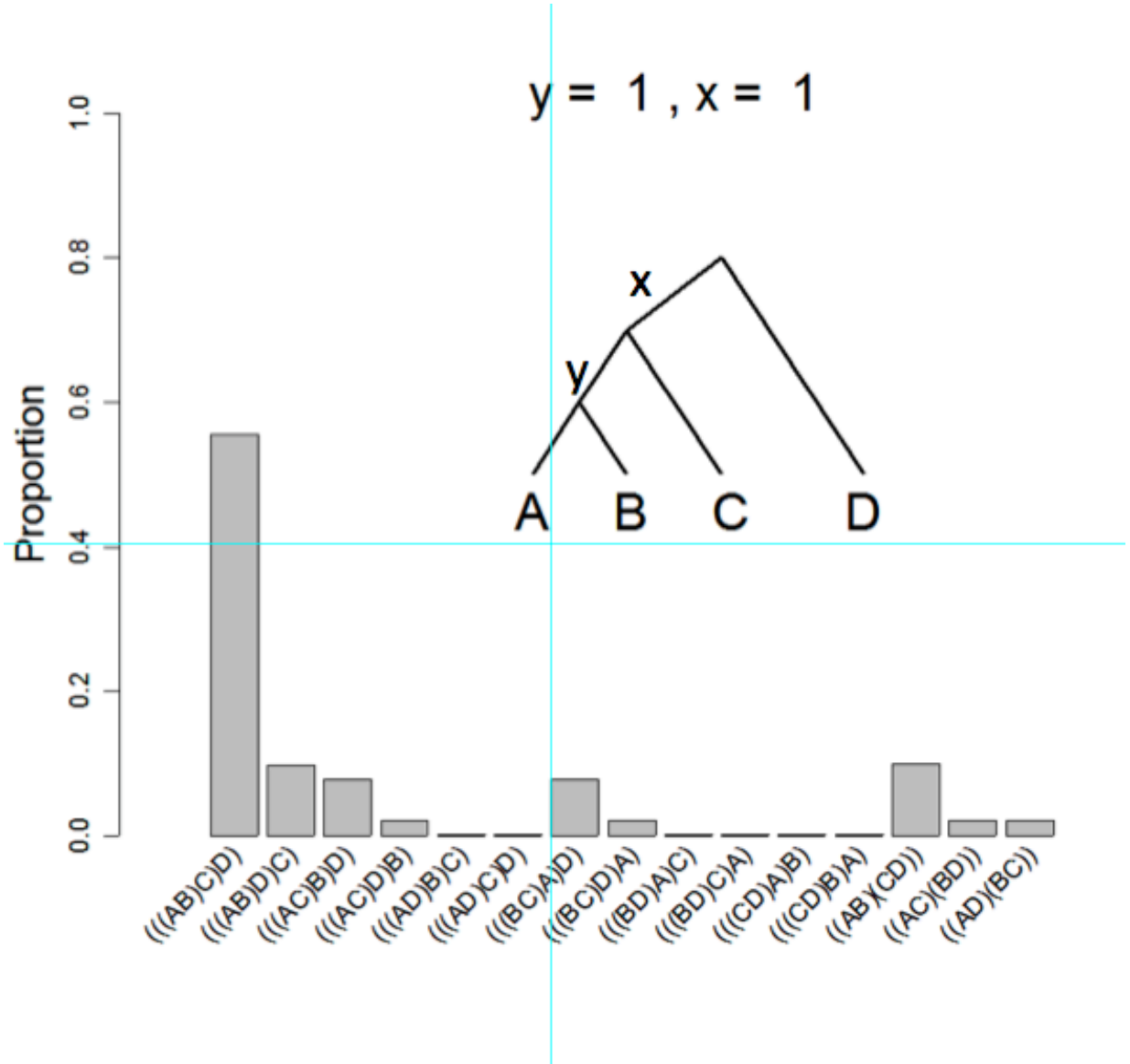
Incomplete lineage sorting is modeled by the

multispecies coalescent model.

This model gives the *gene tree distribution*, the theoretical distribution of gene trees (with or without branch lengths) arising on a species tree.

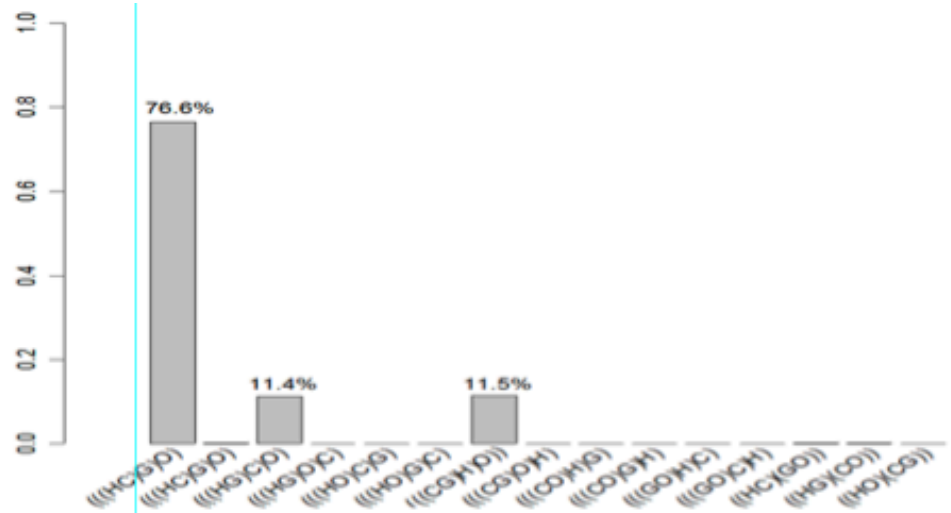


Gene tree distribution

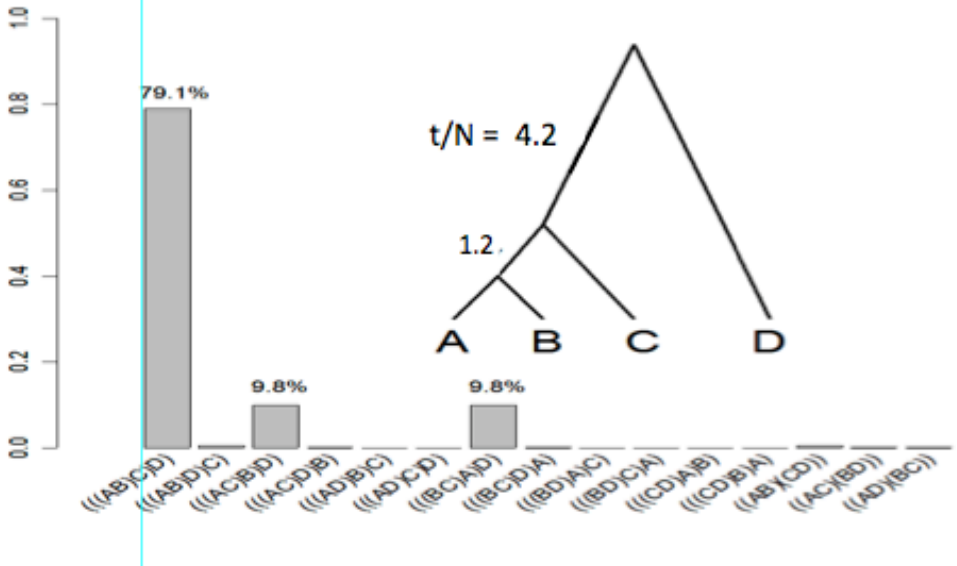


Gene tree distribution: HCG

Data from Ebersberger et al. 2007. Mol. Biol. Evol. 24:2266-2276.



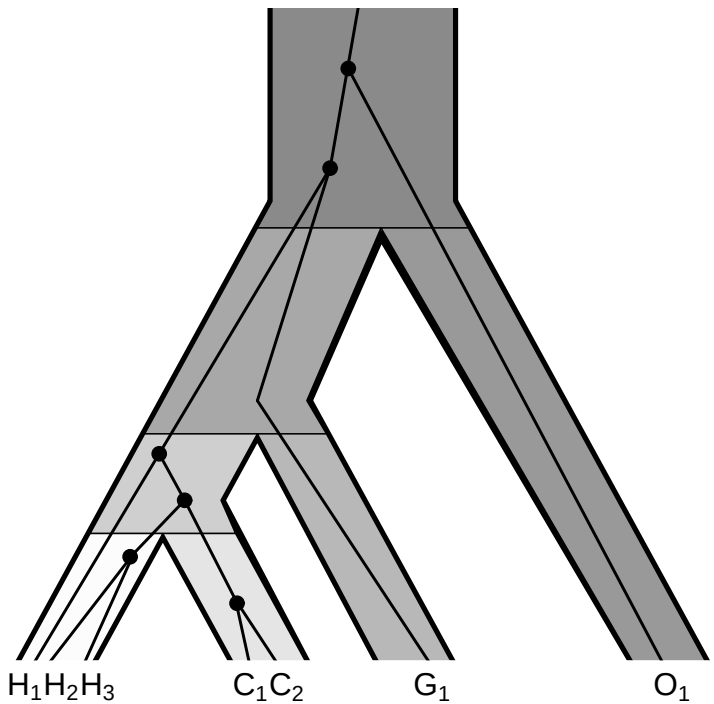
Theoretical distribution based on parameters from Rannala and Yang, 2003. Genetics 164:1645-1656.



The multispecies coalescent model: multiple lineages per species

Incomplete lineage sorting is modeled by the

multispecies coalescent model.



Inferring species trees

Model-Based methods

1. BEST (Liu and Pearl, 2007, *Syst. Biol.*)
2. *BEAST (Heled and Drummod, 2010, *Mol. Biol. Evol*)
3. STEM/GLASS/Maximum Tree (Liu and Pearl, 2010, *J. Math. Biol.*, Kubatko et al., 2009, *Bioinformatics*, Mossel and Roch, 2010, *IEEE Trans. Comp. Biol. Bioinf.*)
4. Pseudo-Maximum Likelihood (Liu et al., 2010, *BMC Evolut. Biol.*)
5. STEW? STEWP? $\int P(X|G)f(G|S)dG$ on SNPS (Bryant et al., 2009, arXiv)

BEST and *BEAST use branch length information in the gene trees but are computationally intensive.

STEM is computationally fast but is sensitive to error in gene tree estimation.

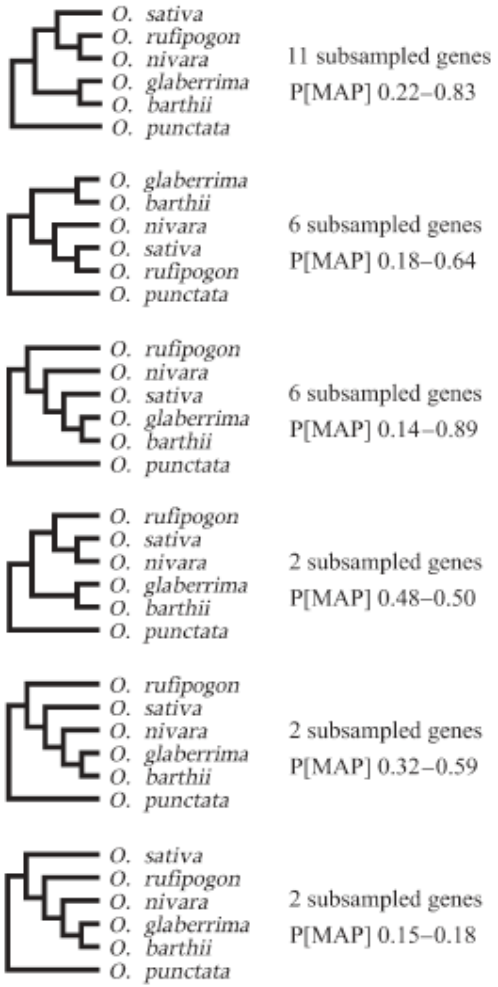
Inferring species trees

Some consensus methods

1. Rooted Triple Consensus (Ewing et al., 2008, *BMC Evolut. Biol.*)
2. R* Consensus (Degnan et al., 2009, *Syst. Biol*)
3. BUCKy, Quartet version (Larget et al., 2010, *Bioinformatics*)
4. BUCKy (Ané et al., 2007, *Mol. Biol. Evol.*)
5. Majority rule and extended majority rule (greedy) consensus (Degnan et al. 2009, *Syst. Biol.*)

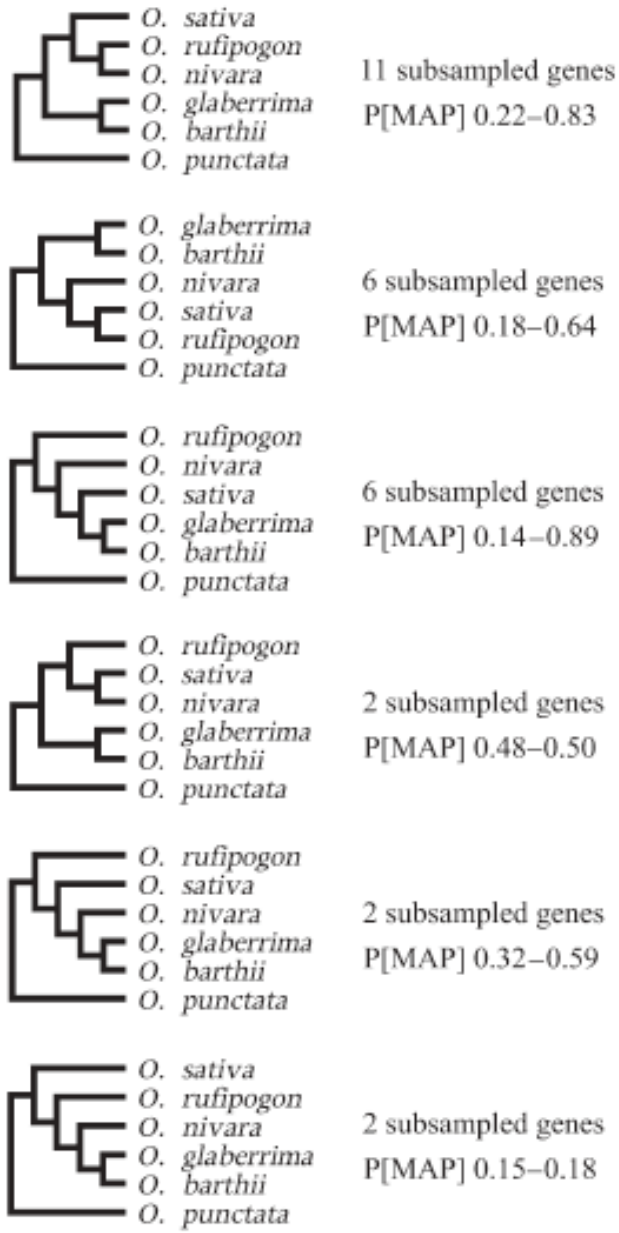
Several loci, several gene trees

From Cranston et al., Systematic Biology (2009), subset of data on 5 rice species



Greedy consensus trees

Clade	frequency (out of 29 genes)
$\{R, N\}$	11
$\{G, B\}$	29
$\{S, R\}$	6
$\{S, N\}$	2
$\{N, R, S\}$	19
$\{B, G, S\}$	6
$\{B, G, N\}$	4



Greedy consensus for rice data

Clade	frequency (out of 29 genes)	
$\{R, N\}$	11	
$\{G, B\}$	29	✓
$\{S, R\}$	6	
$\{S, N\}$	2	
$\{N, R, S\}$	19	
$\{B, G, S\}$	6	
$\{B, G, N\}$	4	

Greedy consensus for rice data

Clade	frequency (out of 29 genes)	
$\{R, N\}$	11	
$\{G, B\}$	29	$\checkmark$
$\{S, R\}$	6	
$\{S, N\}$	2	
$\{N, R, S\}$	19	$\checkmark$
$\{B, G, S\}$	6	$\times$
$\{B, G, N\}$	4	$\times$

Greedy consensus for rice data

Clade	frequency (out of 29 genes)	
$\{R, N\}$	11	$\surd$
$\{G, B\}$	29	$\surd$
$\{S, R\}$	6	$\times$
$\{S, N\}$	2	$\times$
$\{N, R, S\}$	19	$\surd$
$\{B, G, S\}$	6	$\times$
$\{B, G, N\}$	4	$\times$

The greedy consensus tree is $(((R, N), S), (G, B))$, which agrees with BUCKy and BEST.

BUCKy concordance tree

From Cranston et al., Systematic Biology (2009), subset of data on 5 rice species

Clade	Mean CF	95% HPD interval
(Oryza glaberrima, Oryza barthii)	0.707	(0.648, 0.759)
(*Oryza nivara, O. glaberrima*)	0.071	(0.043, 0.111)
(Oryza rufipogon, O. nivara)	0.353	(0.235, 0.451)
(*O. rufipogon, Oryza sativa*)	0.194	(0.105, 0.315)
(O. sativa, O. rufipogon, O. nivara)	0.202	(0.136, 0.272)
(*O. sativa, O. nivara, O. barthii, O. glaberrima*)	0.238	(0.170, 0.302)

Consistency of greedy consensus

Is greedy consensus applied to clade probabilities statistically consistent?

Are you guaranteed to get the correct species tree as the number of gene trees goes to infinity?

Inconsistency of greedy consensus under the multispecies coalescent, species tree$(((a, b):0.05, c):0.05, d)$

gene tree	probability	
(((AB)C)D)	p_1	0.079
(((AB)D)C)	p_2	0.075
(((AC)B)D)	p_3	0.061
(((AC)D)B)	p_4	0.060
(((AD)B)C)	p_5	0.045
(((AD)C)B)	p_6	0.045
(((BC)A)D)	p_7	0.061
(((BC)D)A)	p_8	0.060
(((BD)A)C)	p_9	0.045
(((BD)C)A)	p_{10}	0.045
(((CD)A)B)	p_{11}	0.045
(((CD)B)A)	p_{12}	0.045
((AB)(CD))	p_{13}	0.121
((AC)(BD))	p_{14}	0.105
((AD)(BC))	p_{15}	0.105

Inconsistency of greedy consensus: clade probabilities, species tree $(((a, b){:}0.05, c){:}0.05, d)$

clade	probability	
$\{AB\}$	$c_1 = p_1 + p_2 + p_{13}$	0.275
$\{AC\}$	$c_2 = p_3 + p_4 + p_{14}$	0.226
$\{AD\}$	$c_3 = p_5 + p_6 + p_{15}$	0.196
$\{BC\}$	$c_4 = p_7 + p_8 + p_{15}$	0.226
$\{BD\}$	$c_5 = p_9 + p_{10} + p_{14}$	0.196
$\{CD\}$	$c_6 = p_{11} + p_{12} + p_{13}$	0.212
$\{ABC\}$	$c_7 = p_1 + p_3 + p_7$	0.201
$\{ABD\}$	$c_8 = p_2 + p_5 + p_9$	0.166
$\{ACD\}$	$c_9 = p_4 + p_6 + p_{11}$	0.151
$\{BCD\}$	$c_{10} = p_8 + p_{10} + p_{12}$	0.151

Inconsistency of greedy consensus: clade probabilities, species tree $(((a, b):0.05, c):0.05, d)$

clade	probability		
$\{AB\}$	$c_1 = p_1 + p_2 + p_{13}$	0.275	✓
$\{AC\}$	$c_2 = p_3 + p_4 + p_{14}$	0.226	×
$\{AD\}$	$c_3 = p_5 + p_6 + p_{15}$	0.196	×
$\{BC\}$	$c_4 = p_7 + p_8 + p_{15}$	0.226	×
$\{BD\}$	$c_5 = p_9 + p_{10} + p_{14}$	0.196	×
$\{CD\}$	$c_6 = p_{11} + p_{12} + p_{13}$	0.212	
$\{ABC\}$	$c_7 = p_1 + p_3 + p_7$	0.201	
$\{ABD\}$	$c_8 = p_2 + p_5 + p_9$	0.166	
$\{ACD\}$	$c_9 = p_4 + p_6 + p_{11}$	0.151	×
$\{BCD\}$	$c_{10} = p_8 + p_{10} + p_{12}$	0.151	×

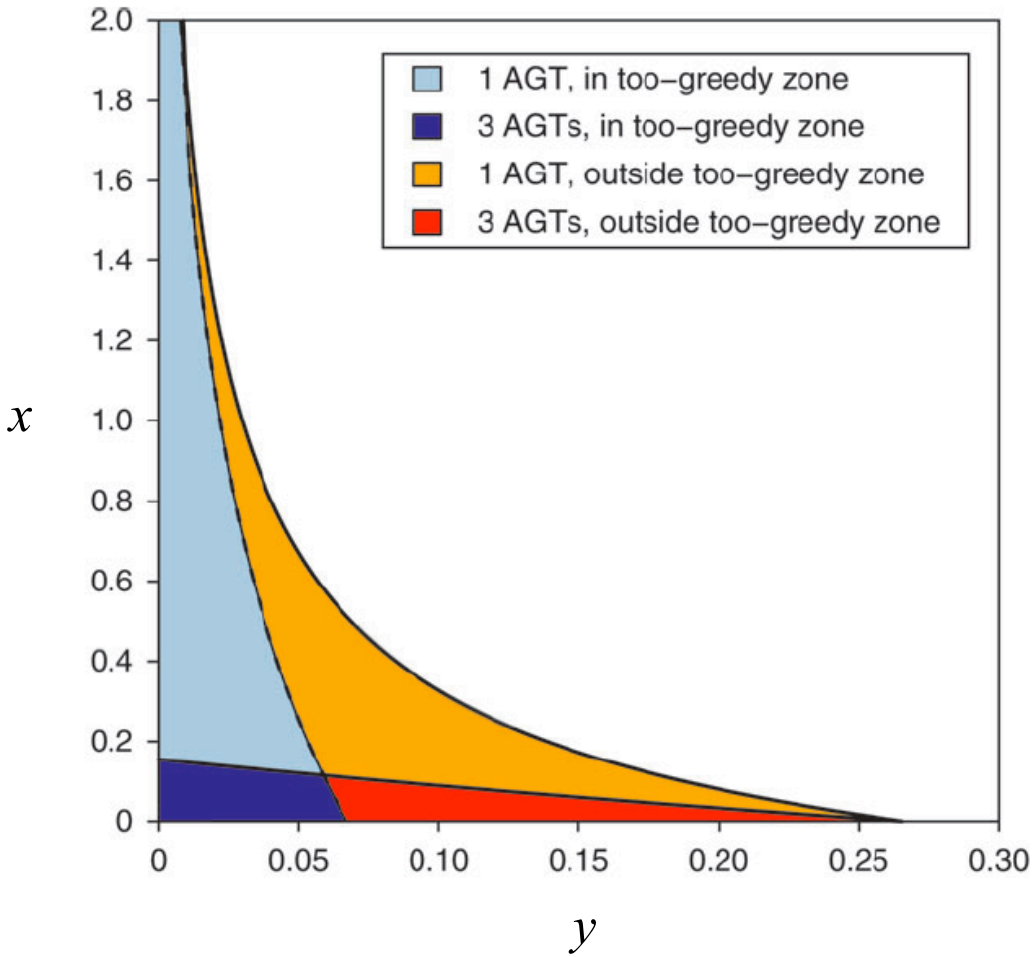
Inconsistency of greedy consensus: clade probabilities, species tree $(((a, b):0.05, c):0.05, d)$

clade	probability		
$\{AB\}$	$c_1 = p_1 + p_2 + p_{13}$	0.275	✓
$\{AC\}$	$c_2 = p_3 + p_4 + p_{14}$	0.226	×
$\{AD\}$	$c_3 = p_5 + p_6 + p_{15}$	0.196	×
$\{BC\}$	$c_4 = p_7 + p_8 + p_{15}$	0.226	×
$\{BD\}$	$c_5 = p_9 + p_{10} + p_{14}$	0.196	×
$\{CD\}$	$c_6 = p_{11} + p_{12} + p_{13}$	0.212	✓
$\{ABC\}$	$c_7 = p_1 + p_3 + p_7$	0.201	×
$\{ABD\}$	$c_8 = p_2 + p_5 + p_9$	0.166	×
$\{ACD\}$	$c_9 = p_4 + p_6 + p_{11}$	0.151	×
$\{BCD\}$	$c_{10} = p_8 + p_{10} + p_{12}$	0.151	×

Greedy consensus tree is $((a, b), (c, d))$, but species tree is $(((a, b), c), d)$.

Inconsistency of greedy consensus

When does this happen? For the species tree $(((a, b):x, c):y, d)$, greedy consensus is misleading in blue regions; the most probable gene tree doesn't match the species tree in all colored regions.



Inconsistency of greedy consensus

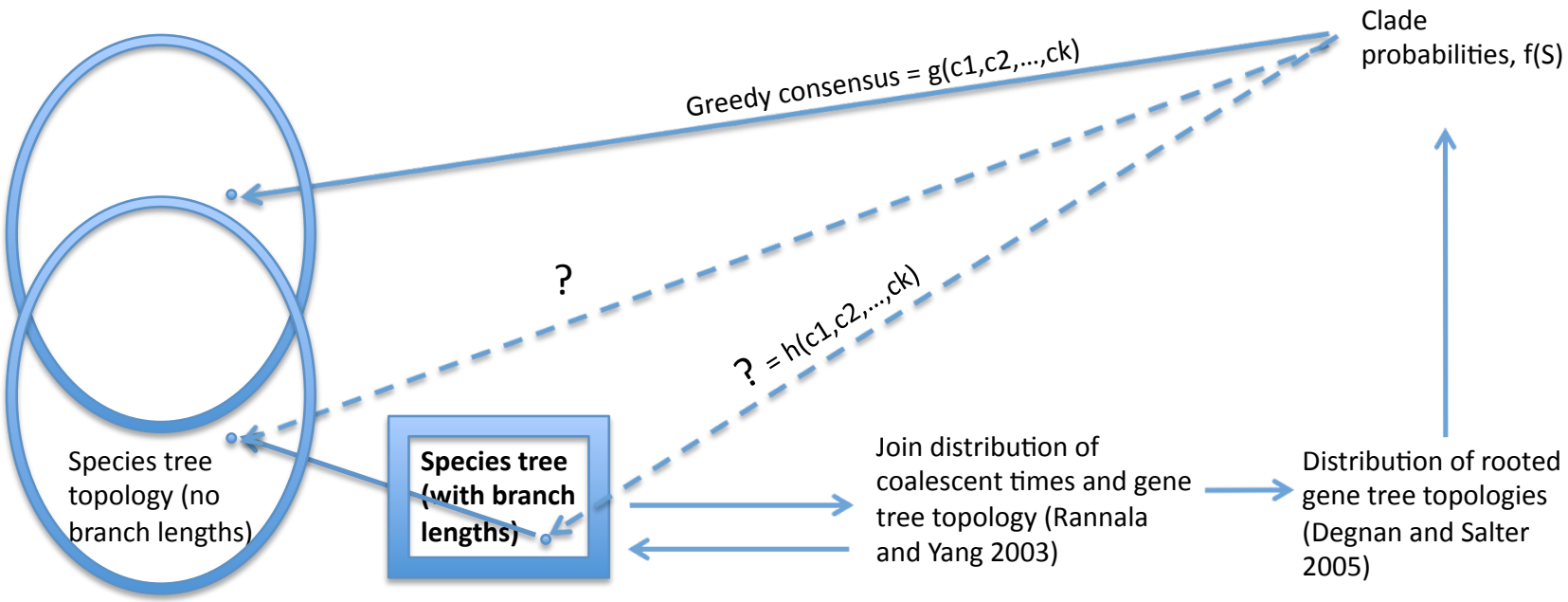
The general result (Degnan, DeGiorgio, Bryant, and Rosenberg, 2009, *Syst. Biol.*):

Theorem

(i) For 3-taxon species tree topologies and for 4- taxon symmetric species tree topologies, the greedy asymptotic consensus tree (GACT) matches the species tree; (ii) for the asymmetric topology with $n = 4$ taxa and for every species tree topology with $n \geq 5$ taxa, there exist branch lengths such that the GACT does not match the species tree.

Identifiability of species trees from clade probabilities

Is there a function from clade probabilities to species trees that always recovers the species tree?



Identifiability of species trees from clade probabilities

Practical question: Is there a method that is "better" than greedy consensus?

Theoretical question: Does the function from species tree parameters to clade probabilities have an inverse?

Invariants of gene tree distributions

Invariants can be useful for answering questions of identifiability.

A gene tree distribution on 15 trees has probabilities $p = (p_1, \ldots, p_{15}) \in (0, 1)^{15}$.

A **gene tree invariant** is a polynomial in the gene tree probabilities is equal to 0 for all possible branch lengths for a particular species tree topology.

For the asymmetric species tree, the trivial invariant is $p_1 + \cdots + p_{15} - 1 = 0$.

Examples

In addition, many gene tree probabilities are tied.

For example, for the species tree $(((a, b){:}x, c){:}y, d)$, the gene trees $(((A, C), B), D)$ and $(((B, C), A), D)$ are equally likely by symmetry. These are simple examples of invariants. For example,

$$p_2 - p_3 = 0.$$

Fancier examples:

$$p_2 + p_5 - p_{13} = 0$$

$$\begin{aligned} &6p_3p_5 + 6p_5^2 + 3p_3p_{13} + 3p_5p_{13} + 3p_3p_{14} \\ &+15p_5p_{14} + 6p_{13}p_{14} + 6p_{14}^2 - 2p_5 - 2p_{14} = 0 \end{aligned}$$

Clade probabilities

Probabilities of clades can be determined by summing probabilities of gene trees.

The clade probabilities for a species tree form a k-tuple, $c = (c_1, \ldots, c_k) \in (0,1)^k$, $k = 2^n - n - 1$. They do not add up to 1 because clades are not mutually exclusive. However,

$$\sum_k c_k = n - 2 \text{ (trivial invariant)}$$

because each tree contributes $n - 2$ clades.

$$\begin{aligned}\sum_{j=1}^{k} c_j &= \sum_{j=1}^{k} \sum_i p_i I(\text{clade } j \text{ is on tree } i) \\ &= \sum_i p_i \sum_j I(\text{clade } j \text{ is on tree } i) \\ &= \sum_i p_i (n-2) = n - 2\end{aligned}$$

Clade invariants

The trivial invariant for clades holds for all branch lengths and all species tree topologies. A nontrivial invariant holds for all branch lengths for a particular topology (or subset of topologies).

Similar to gene tree invariants, some clade invariants are the result of symmetries in the species tree. For example

$$\mathbb{P}[\{A, C\}] - \mathbb{P}[\{B, C\}] = 0$$

for any species tree with clade $\{a, b\}$.

Using similar observations (and probabilities of clades under the coalescent), one can show, for example that 4- and 5-taxon species trees are identifiable from clade probabilities.

Clade invariants, arbitrary n

How do we generalize arguments to larger trees?

Clade probabilities on smaller trees cannot be written as linear combinations of clade probabilities on larger trees, making inductive arguments difficult.

If the species tree has leafset $\mathcal{X}$, we'd like to find an invariant associated with any nontrivial clade $\mathcal{A} \subset \mathcal{X}$ such that the invariants holds if and only if $\mathcal{A}$ is a clade on the species tree, regardless of $|\mathcal{X}|$.

Clade invariant idea

An idea for identifying clades is that if two lineages are present in a population, each lineage is equally likely to be *excluded* from a clade by exchangeability of lineages.

Notation

1. $\mathcal{X}$ the leafset of the species tree
2. $\mathcal{A}$ a subset of leaves of $\mathcal{X}$ with $1 < |\mathcal{A}| < |\mathcal{X}|$
3. $\pi(\mathcal{A}) = \mathcal{A}_1|\mathcal{A}_2|\cdots|\mathcal{A}_k$, a partition of $\mathcal{A}$
4. a, b two leaves in $\mathcal{A}$
5. $\mathcal{A}' = \mathcal{A} \setminus \{a, b\}$, (possibly the empty set)
6. $\mathcal{C} = \mathcal{X} \setminus \mathcal{A}$, ($\mathcal{C}$ must be nonempty)
7. let $v = MRCA(\mathcal{A})$ in the species tree.

A special invariant

Lemma

Let $\mathcal{A} \subsetneq \mathcal{X}$ be a subset of taxa with at least two elements, and $\mathcal{C} \subseteq \mathcal{X} \setminus \mathcal{A}$ a non-empty set of taxa not in $\mathcal{A}$. For some $a, b \in \mathcal{A}$, let $\mathcal{A}' = \mathcal{A} \setminus \{a, b\}$. Then if $\mathcal{A}$ is a clade on σ,

$$\left(\sum_{\mathcal{S} \subseteq \mathcal{A}'} \mathbb{P}_\sigma(\mathcal{S}_g \cup \{A\} \cup \mathcal{C}_g) \right) - \left(\sum_{\mathcal{S} \subseteq \mathcal{A}'} \mathbb{P}_\sigma(\mathcal{S}_g \cup \{B\} \cup \mathcal{C}_g) \right) = 0.$$

HCG invariant

Example, consider the species tree $(((h, c), g), o)$.

Since $\mathcal{A} = \{h, c, g\}$ is a clade, we should be able to find a special invariant associated with each choice of $a, b \in \mathcal{A}$.

Suppose we choose $a = h$, $b = g$. Then $\mathcal{A}' = \mathcal{A} \setminus \{h, g\} = \{c\}$. Also let $\mathcal{C} = \mathcal{X} \setminus \{h, c, g\} = \{o\}$.

The subsets $\mathcal{S}$ used in the sums are $\mathcal{S} = \{c\}$ and $\mathcal{S} = \emptyset$. Thus, the invariant is

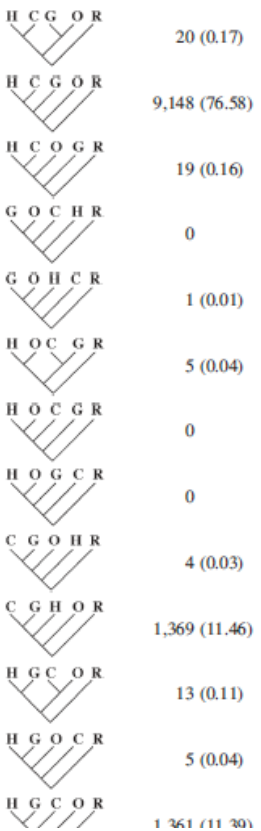
$$\begin{aligned} &\big[\mathbb{P}(\emptyset \cup \{H\} \cup \{O\}) + \mathbb{P}(\{C\} \cup \{H\} \cup \{O\})\big] \\ &\quad - \big[\mathbb{P}(\emptyset \cup \{G\} \cup \{O\}) + \mathbb{P}(\{C\} \cup \{G\} \cup \{O\})\big] \\ = &\big[\mathbb{P}(\{HO\}) + \mathbb{P}(\{CHO\})\big] - \big[\mathbb{P}(\{GO\}) + \mathbb{P}(\{CGO\})\big] = 0 \end{aligned}$$

Just for fun

Let's see how this invariant does on the Ebersberger data. The clades $\{HO\}, \{CHO\}, \{GO\}, \{CGO\}$ were observed 5, 19, 4, and 20 times respectively. This gives

$$\frac{(5+19)-(4+20)}{11945} = \frac{0}{11945}$$

From Ebersberger et al., 2007, *Mol. Biol. Evol.*



Why does the special invariant work?

The strategy of the proof is to find two sets of gene tree clades, $\mathscr{C}_1$ and $\mathscr{C}_2$, such that

$$\sum_{\mathcal{C}_g \in \mathscr{C}_1} \mathbb{P}_\sigma[\mathcal{C}_g | \pi(\mathcal{A})] - \sum_{\mathcal{C}_g \in \mathscr{C}_2} \mathbb{P}_\sigma[\mathcal{C}_g | \pi(\mathcal{A})] = 0$$

Excluding clades

Lemma

Let ψ be a rooted binary species tree topology on $\mathcal{X}$, where $\mathcal{X} = \mathcal{A} \sqcup \mathcal{D}$ is a disjoint union of non-empty subsets. If $\mathcal{A}$ is not a clade on ψ then for all choices of edge lengths λ except those in some set of measure zero there exists some $\mathcal{C} \subseteq \mathcal{D}$, $a, b \in \mathcal{A}$, such that the corresponding special clade invariants do not vanish on the clade probabilities arising under the multispecies coalescent on $\sigma = (\psi, \lambda)$.

The special invariants allow us to identify clades while the Lemma allows us to show that a set of taxa is not a clade (except possibly on a set of measure 0), so we are done.

Main result

Theorem

Let ψ be a rooted binary species tree topology on $\mathcal{X}$. For generic choices of edge lengths λ, ψ can be identified from the probabilities of clades under the multispecies coalsecent on $\sigma = (\psi, \lambda)$.

Q: What assumptions from the multispecies coalescent have we used?

Miscellaneous other results

1. Any clade with probability greater than $1/3$ must be on the species tree
2. The most probable 2-clade is a 2-clade on the species tree
3. The most probable 3-clade is not necessarily a clade on the species tree, even if the species tree has a 3-clade
4. If $\mathcal{A}$ is a clade, and $x \in \mathcal{A}$ and $y \in \mathcal{X} \setminus \mathcal{A}$, then $\mathbb{P}[\mathcal{A}] > \mathbb{P}[(\mathcal{A} \setminus \{x\}) \cup \{y\}]$.

Recovering branch lengths

We've only shown the identifiability of species tree topologies. Can branch lengths be recovered also?

Once the species tree topology is known, we can write down a system of equations for the clade probabilities as functions of the species tree branch lengths.

Table 1: Probabilities of clades under three 4-taxon species trees. $X = \exp(-x), Y = \exp(-y)$.

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

Conclusion

I'd bet a bottle of whiskey that branch lengths are also identifiable for arbitrary binary species trees.