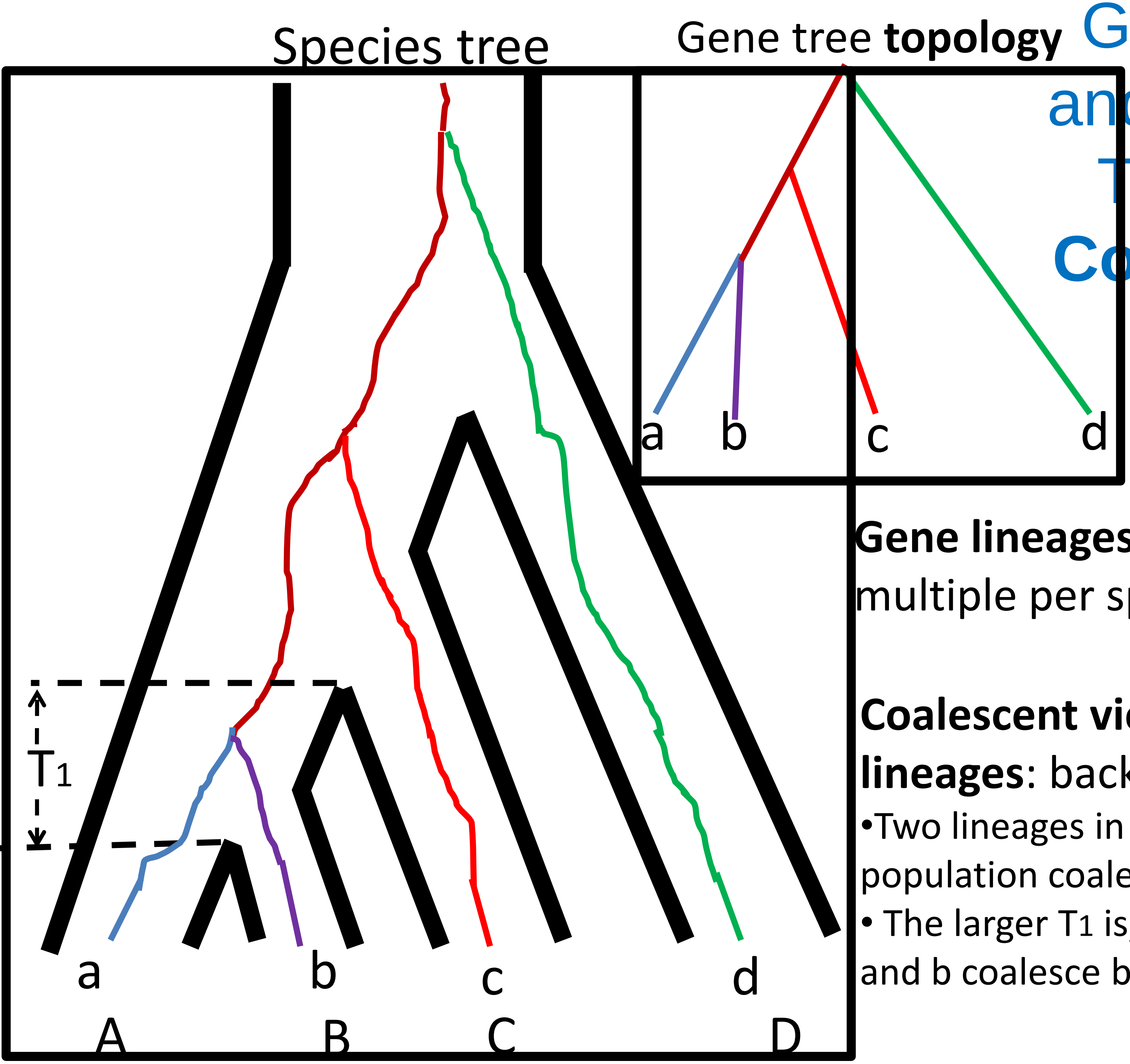
Coalescent-based Species Tree Inference from Gene Tree Topologies Under Incomplete Lineage Sorting by Maximum Likelihood

Yufeng Wu
Dept. of Computer Science & Engineering
University of Connecticut, USA

Gene Tree and Species Tree: the **Coalescent** View

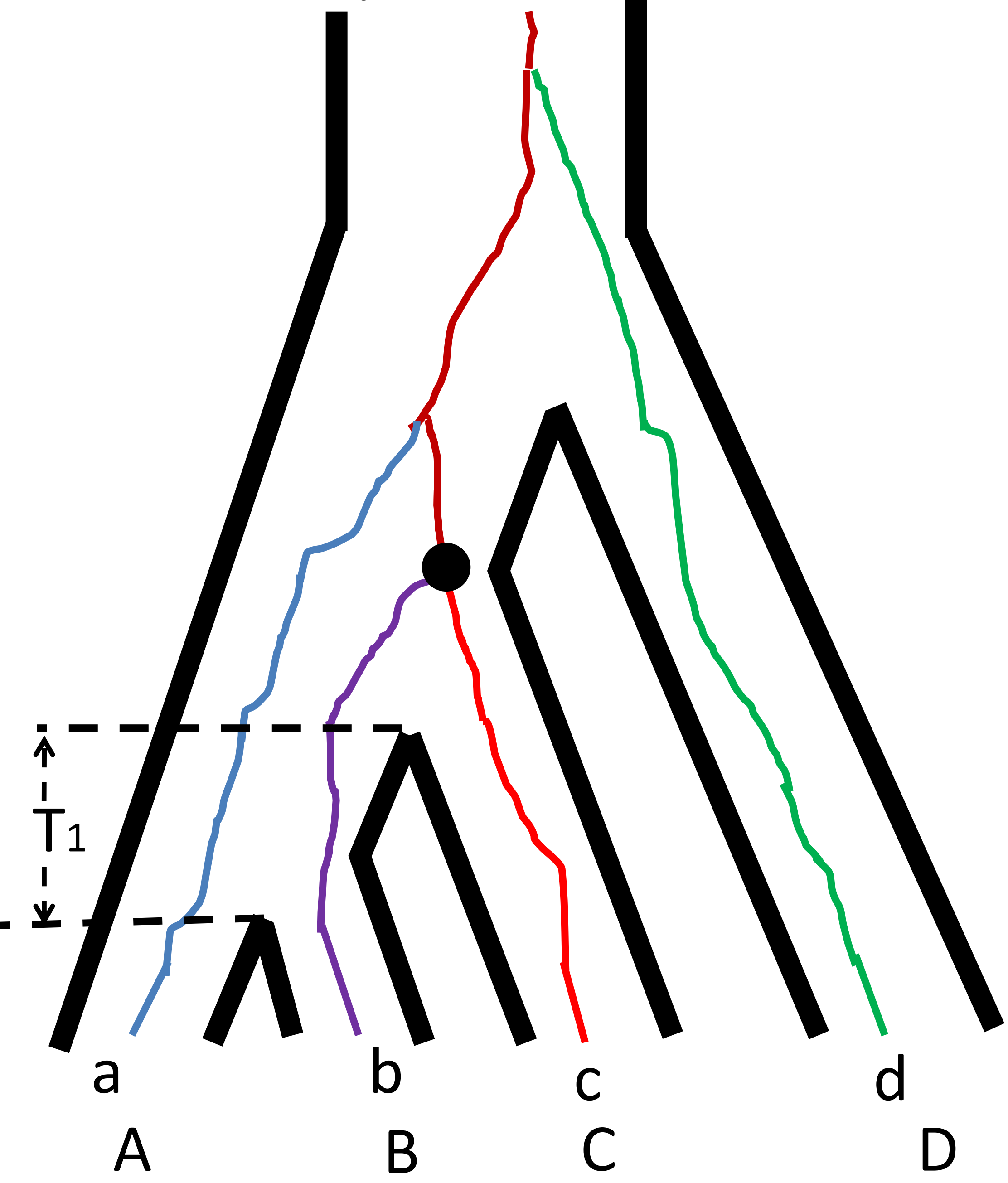


Gene lineages: can be multiple per species.

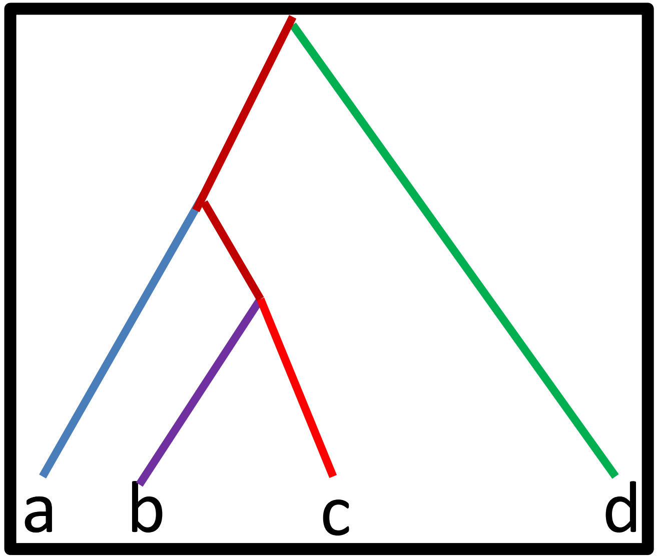
Coalescent view of gene lineages: backwards in time

- Two lineages in the *same* population coalesce stochastically
- The larger T_1 is, the *more* likely a and b coalesce before LCA(A,B,C).

Species tree



A *different* gene tree topology



Different causes of topologically different gene tree and species tree.

Incomplete lineage sorting: gene lineages fail to coalesce within the species boundary.

Lineages a and b fail to coalesce within T1. The smaller T1 is, the more likely this happens.

Gene Tree Probability

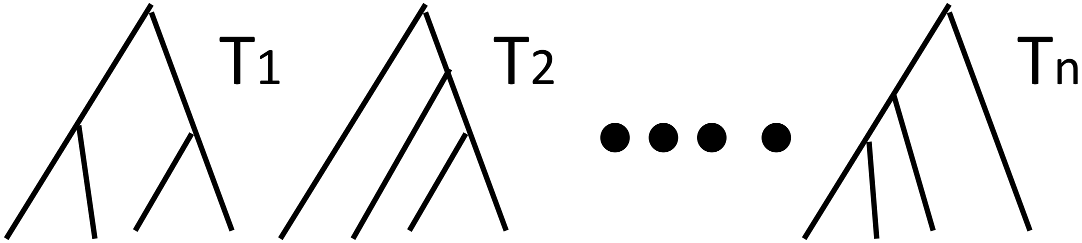
For a species tree, **any** gene tree *topology* can arise, but with probability.

For species tree T_s (with branch length) and a gene tree **topology** T_g:

Gene tree probability $P(T_g|T_s)$: probability of observing a gene tree topology T_g for species tree T_s under coalescent theory.

- The larger $P(T_g|T_s)$ is, the more likely T_g will be observed.

What is the use of gene tree probability?



Trees T_i are inferred from gene sequences

Likelihood: $L(T_1,T_2,\ldots,T_n) = P(T_1|T_s) \bullet P(T_2|T_s) \bullet \ldots \bullet P(T_n|T_s)$

Assumption: trees T_i are independent of each other

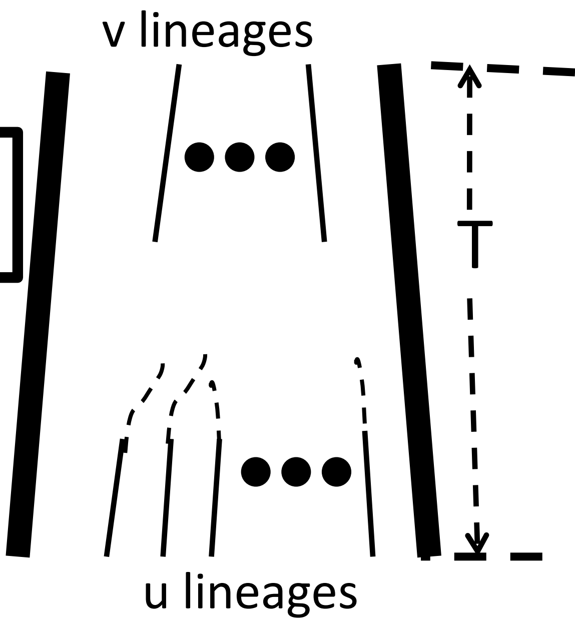
MLE: T_s that maximizes $L(T_1,T_2,\ldots,T_n)$. Standard local search for MLE.

Key: efficient computation of the gene tree probability.

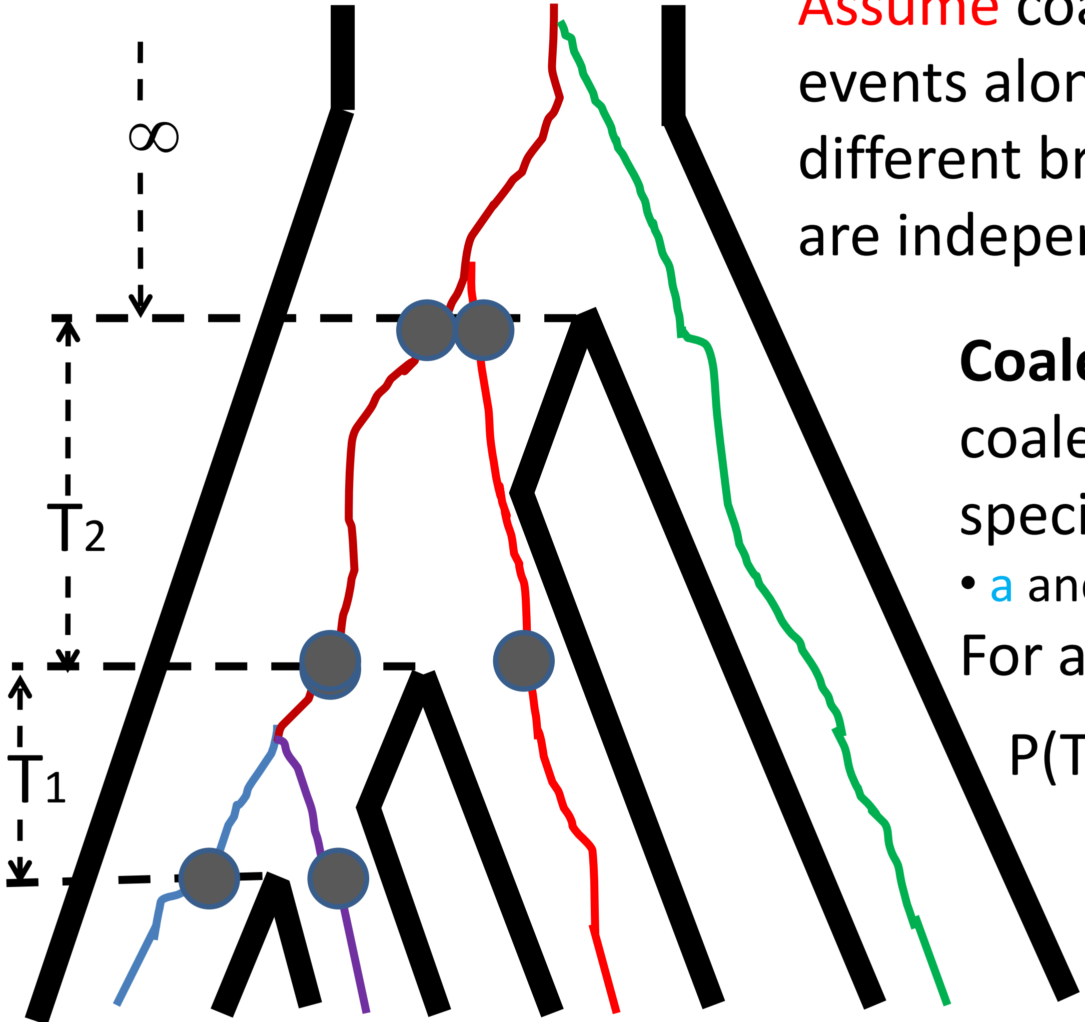
An algorithm for Gene Tree Probability (Degnan and Salter, 2005)

$p_{uv}(T)$: the probability of u (not labeled) lineages coalesce to v lineages within time T is:

$$\boxed{p_{uv}(T)=\sum_{k=v}^{u} e^{-k(k-1)T/2}\frac{(2k-1)(-1)^{k-v}}{v!(k-v)!(v+k-1)}\times\prod_{y=0}^{k-1}\frac{(v+y)(u-y)}{(u+y)}}$$



Assume coalescent events along different branch are independent.



Coalescent history: specify **each** coalescent event occur at which species tree branch, e.g.:

- a and b coalesce within T1

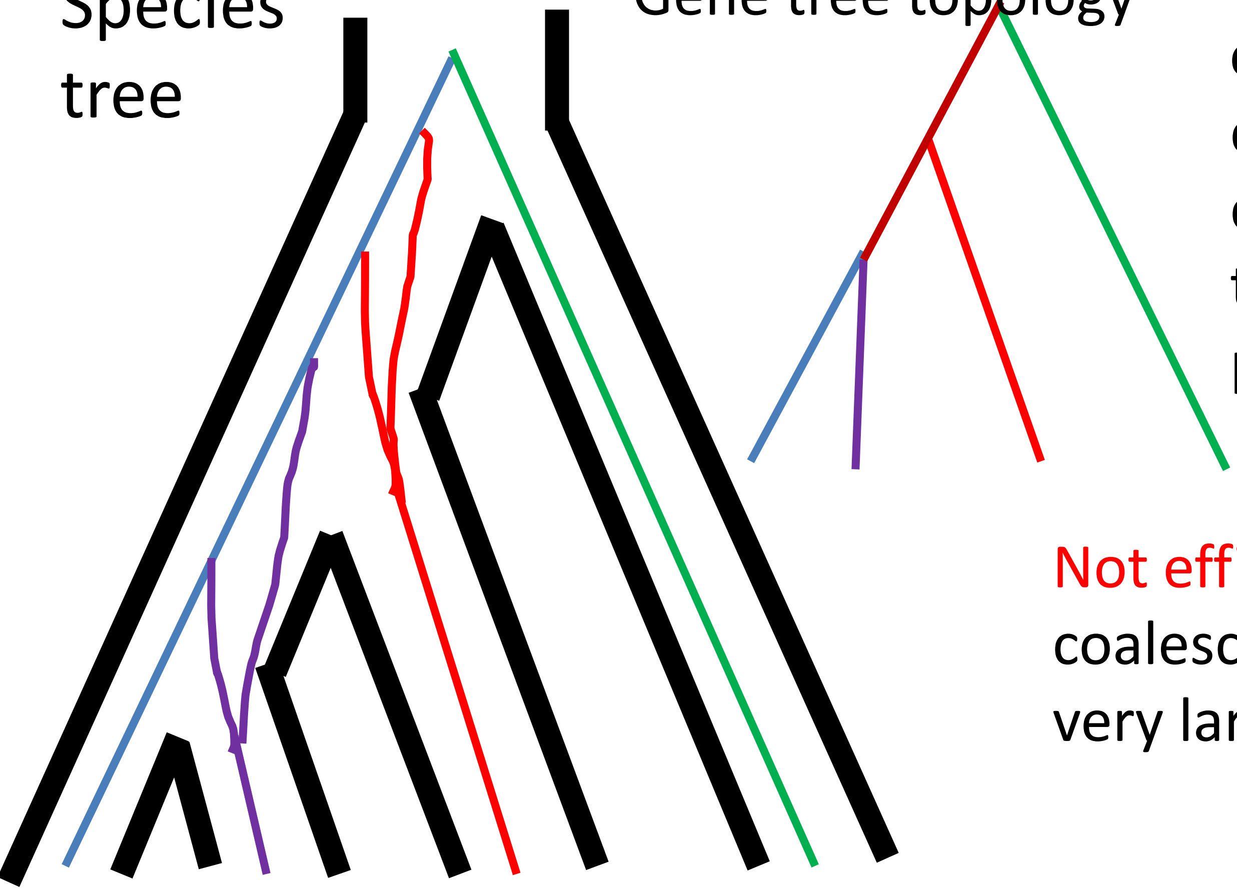
For a fixed coalescent history **H**:

$$P(Tg, H|Ts) = \boxed{p_{21}}(T_1) * \boxed{p_{22}}(T_2) * p_{31}(\infty)$$

* C (combinatorial factor)

Degnan and Salter’s algorithm is not scalable

Main challenge in computing gene tree probability: coalescent history H is not known for a given gene tree topology and so need to consider **all** possible coalescent histories.



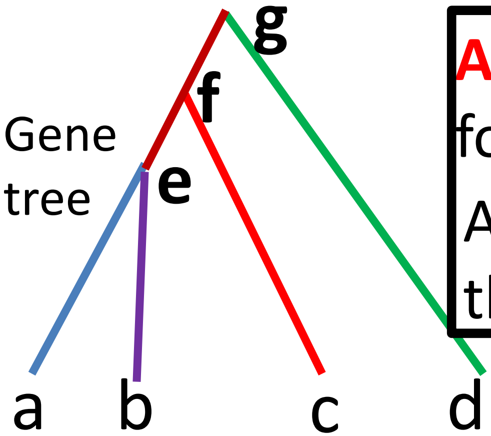
Species tree

Gene tree topology

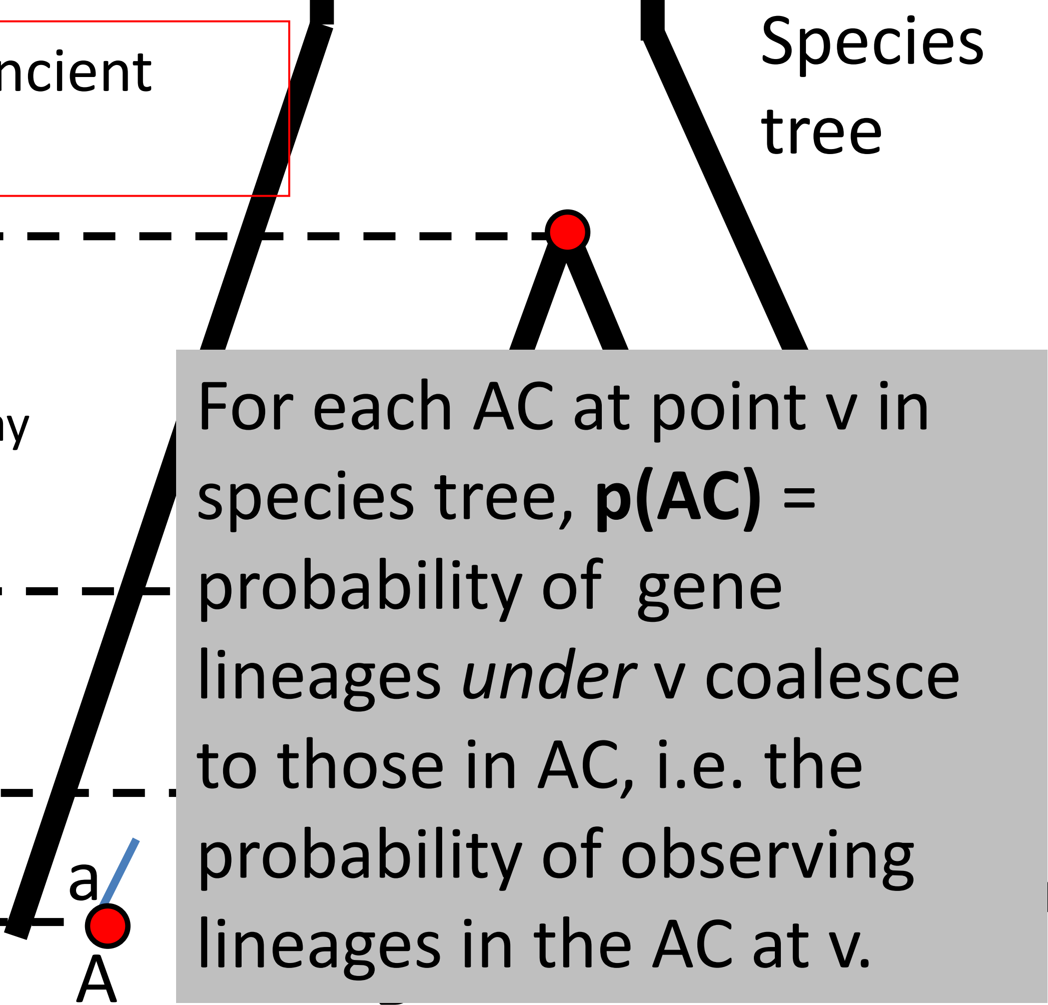
Degnan and Salter: **enumerate** all *feasible* coalescent histories consistent with the gene tree topology; sum their probabilities.

Not efficient: number of coalescent histories can be very large (later).

Key Concept: Ancestral Configurations



Ancestral configuration (AC), foundation of our method:
At a *position* of *species* tree, the set of gene lineages alive.



Possible ACs **right** at (but more ancient than) speciation:

{a,b,c,d}, {e,c,d}, {f,d}

Gene lineages a, b, c may remain *un-coalesced* or may coalesce to e, c or may coalesce to f

{a,b,c}, {e,c}

Gene lineages a, b may remain *un-coalesced* or may coalesce to e

{a,b}

No coalescence between lineages right at speciation time

{a}

For each AC at point v in species tree, **p(AC)** = probability of gene lineages *under* v coalesce to those in AC, i.e. the probability of observing lineages in the AC at v.

Recurrence of Ancestral Configuration Probability

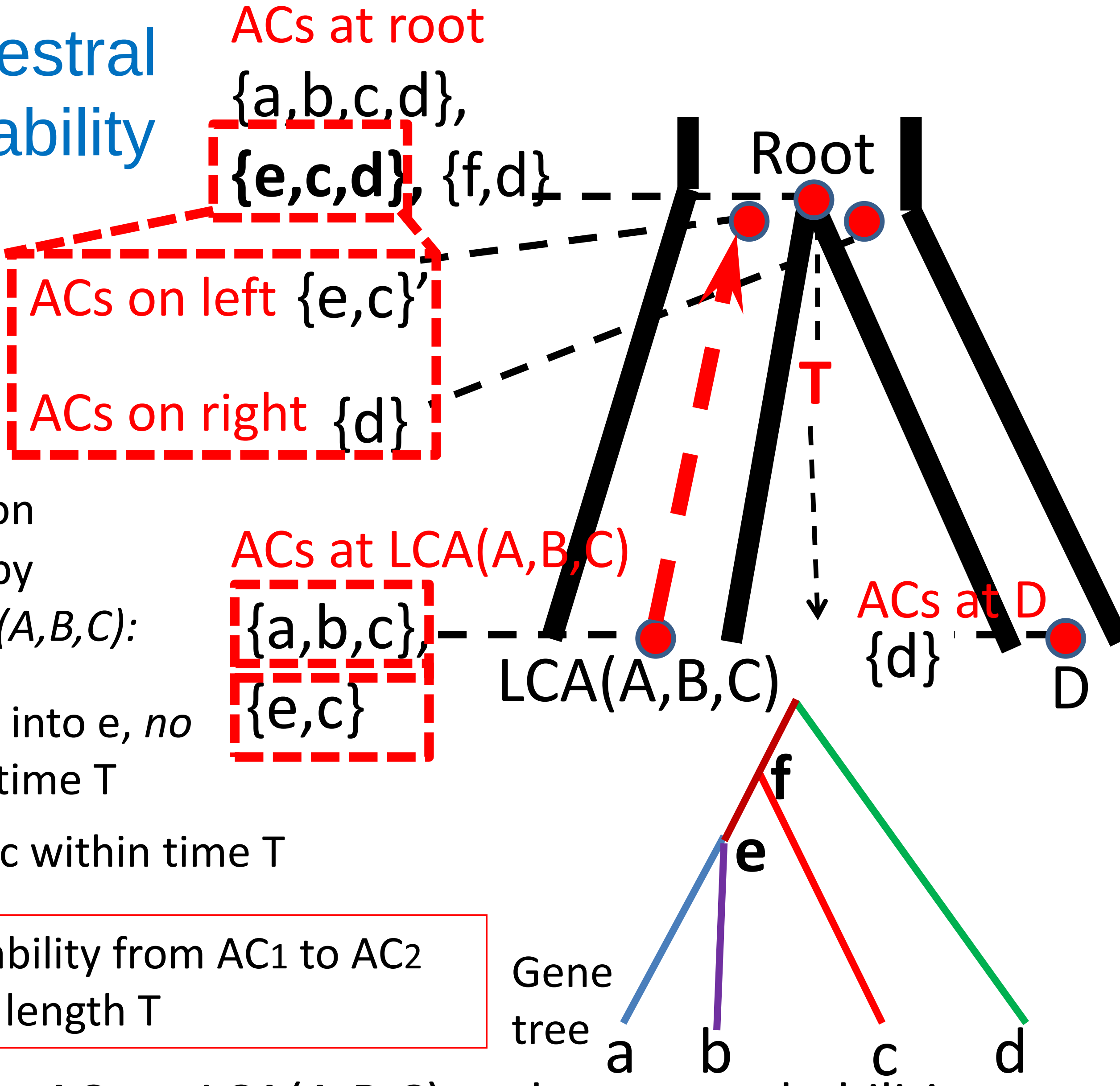
At root, {e,c,d}: derived by merging {e,c}' on the left and {d} on the right.

{e,c}': AC right after speciation towards LCA(A,B,C), resulted by coalescences from ACs at *LCA(A,B,C):*

{a,b,c}: coalescence of a and b into e, *no* coalescence of e and c within time T

{e,c}: *no* coalescence of e and c within time T

Pt(AC1,AC2,T): transition probability from AC1 to AC2 along a species tree branch of length T



ACs at root

{a,b,c,d},

{e,c,d}, {f,d}

ACs on left {e,c}'

ACs on right {d}

ACs at LCA(A,B,C)

{a,b,c},

{e,c}

ACs at D

{d}

Prob. of {e,c}' depends on ACs at LCA(A,B,C) and trans. probabilities:

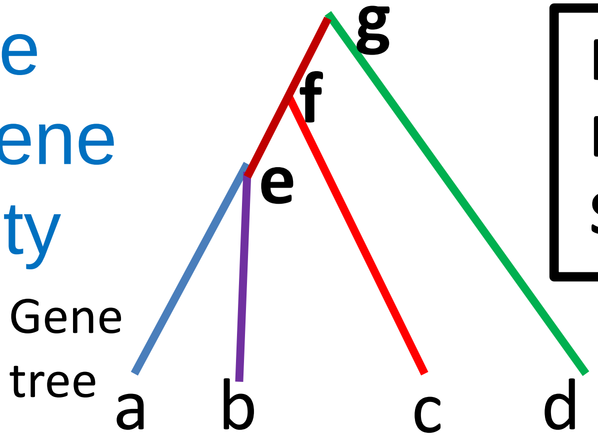
$$p(\{e,c\}') = \underbrace{\boxed{p(\{a,b,c\})*Pt(\{a,b,c\},\{e,c\},T)}}_{\text{a and b coalesce into e within T}} + \underbrace{\boxed{p(\{e,c\})*Pt(\{e,c\},\{e,c\},T)}}_{\text{a and b do not coalesce into e within T}}$$

Peeling-style algorithm for gene tree probability

Recurrence of ACs

Bottom up approach:

Start at leaves and move up



At root of species tree:

P({e,c,d})= p({e,c}') * p({d}) =

(p({a,b,c})Pt({a,b,c},{e,c},T2) +

p({e,c})*Pt({e,c},{e,c},T2)) * p({d})

....

At divergence of A, B and C:

P({a,b,c})=p({a,b})*Pt({a,b},{a,b},T1) * p({c})

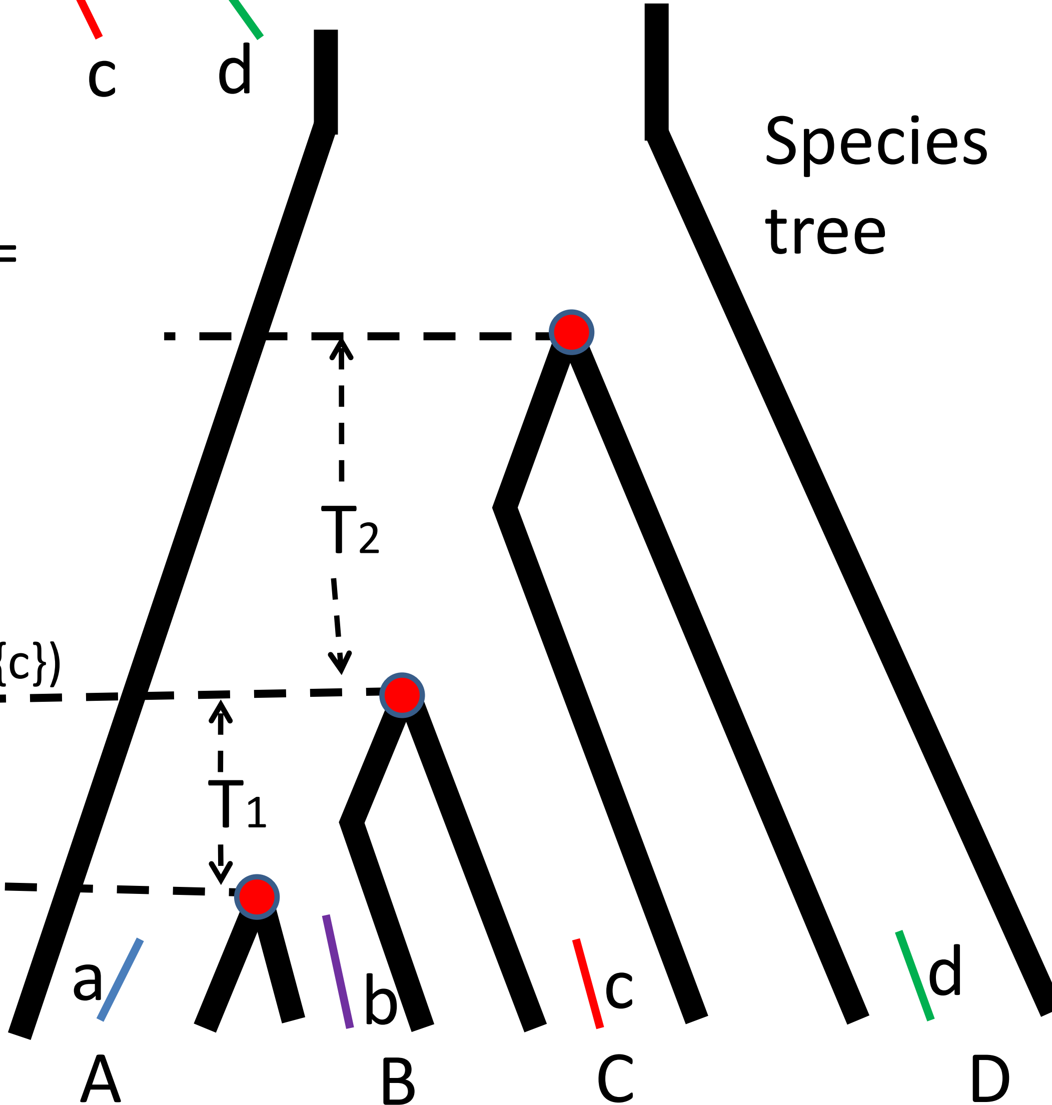
P({e,c})= p({a,b})*Pt({a,b},{e},T1) * p({c})

At divergence of A and B:

P({a,b})=p({a})*p({b})=1.0

At leaves of species tree:

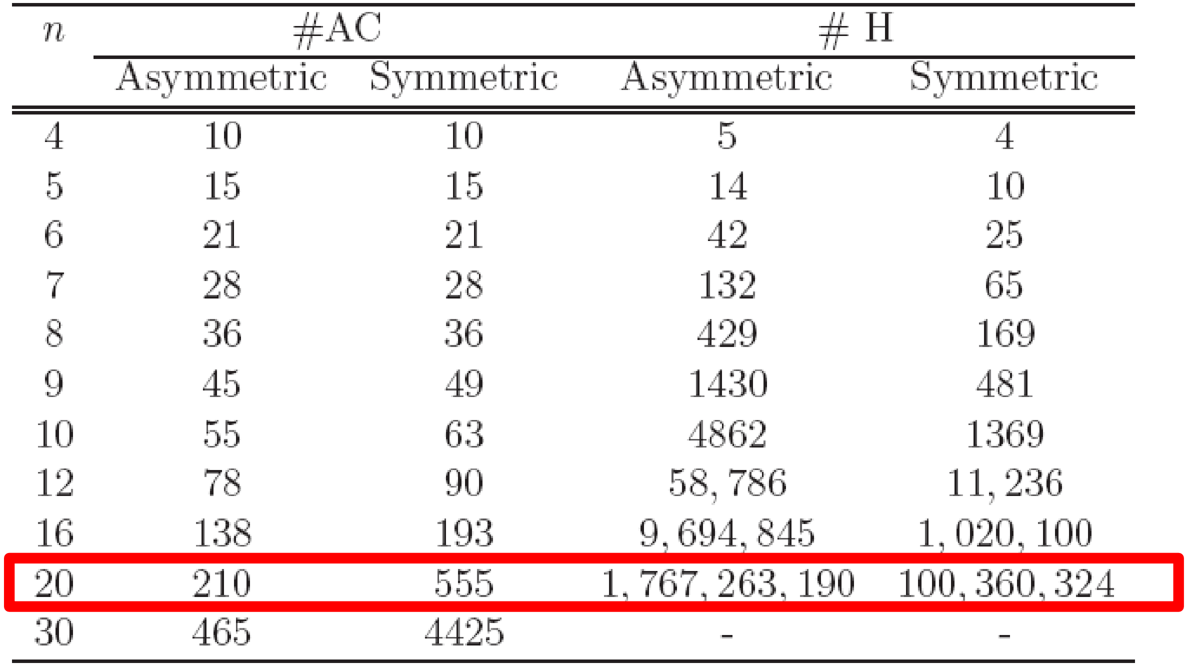
p({a})=p({b})=p({c})=p({d})=1.0



For *identical* gene/species trees with n leaves

	Number of ACs (our method)	**Number of histories (Degnan and Salter)**
Maximal *asymmetric* trees	n(n+1)/2	The Catalan number (exponential in n)
Maximal symmetric trees	≤ $\underline{(2n\text{-}1)n^2/2}$	Appear to be also exponential in n

Counting the number of ACs and histories



n	#AC		# H	
	Asymmetric	Symmetric	Asymmetric	Symmetric
4	10	10	5	4
5	15	15	14	10
6	21	21	42	25
7	28	28	132	65
8	36	36	429	169
9	45	49	1430	481
10	55	63	4862	1369
12	78	90	58, 786	11, 236
16	138	193	9, 694, 845	1, 020, 100
20	210	555	1, 767, 263, 190	100, 360, 324
30	465	4425	-	-

Unfortunately, the number of ACs can still be **exponential** in n for certain types of trees.

Simulation

Implementation: program **STELLS**, our new MLE species tree inference based on gene tree probability computation.

- Given a set of gene tree topologies, find the MLE of the species tree under the coalescent model
- Can also compute gene tree probability for a given species tree

Simulation: simulate *k* gene trees for a given species tree. Simulate gene sequences for the gene trees. Infer gene trees from gene sequences.

Inference **error**: normalized *Robinson-Foulds* distance between inferred species tree and the true species tree.

Compare STELLS (our method) with:
STEM: an existing maximum likelihood approach

Accuracy of Species Inference with MLE

n: number of taxa.

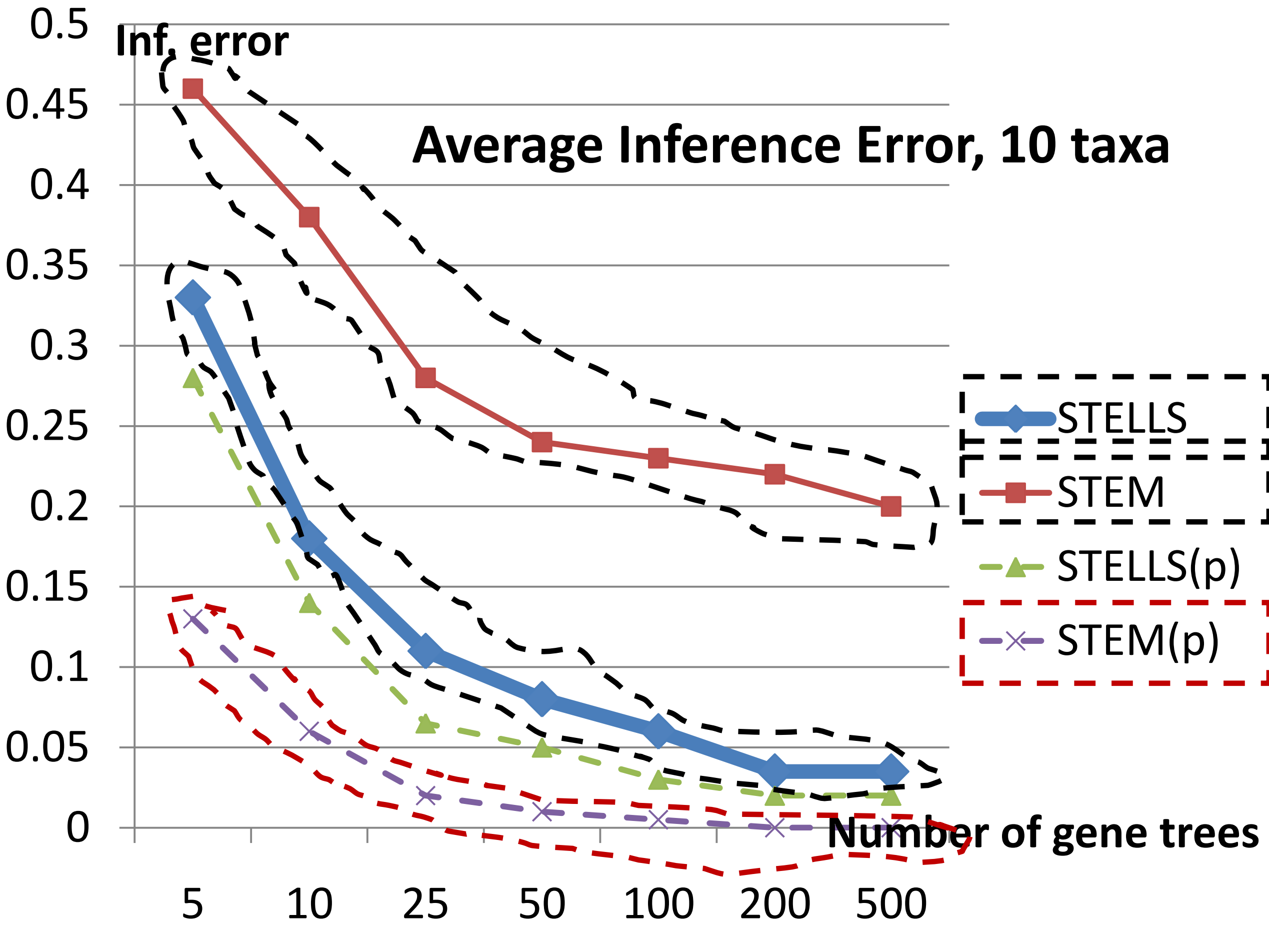
Horizontal axis: number of gene trees

Dashed lines: results with **perfect** gene trees

Solid lines: results with **inferred** gene trees

STELLS (our method) is generally more accurate (but slower) than STEM, especially with noisy gene trees.

STELLS also allows multiple gene lineages for a single species.



Acknowledgement

- More information available at: **http://www.engr.uconn.edu/~ywu**
- Research supported by National Science Foundation