Computational Biology – Class Presentation

Gene Trees in Species Trees

Wayne P. Maddison

Published in *Systematic Biology*, Vol. 46, No.3 (Sept 1997), pp. 523-536

Overview

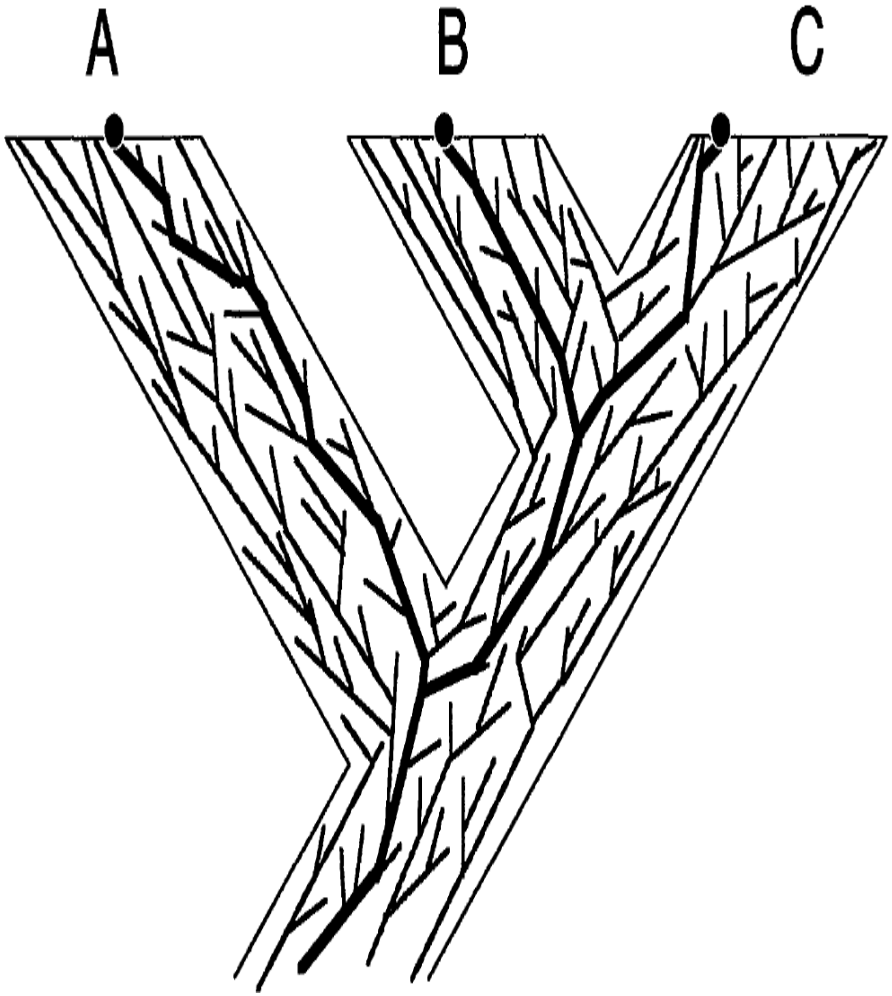
- Gene Trees and Species Trees
- Processes of Discord
  - Horizontal Transfer
  - Lineage Sorting or Deep Coalescence
  - Gene Duplication/Extinction
- Reconstructing Species Tree Using Gene Trees
- Using MLE for Deep Coalescence
- Species Phylogeny - New interpretation
- Conclusion

Gene Trees and Species Trees

- Species Trees / Phylogenetic tree
  - Pattern of branching of species lineages via the process of speciation
- Gene Trees
  - Formed due to gene replication.
  - Broken into several pieces. Due to recombinations within populations

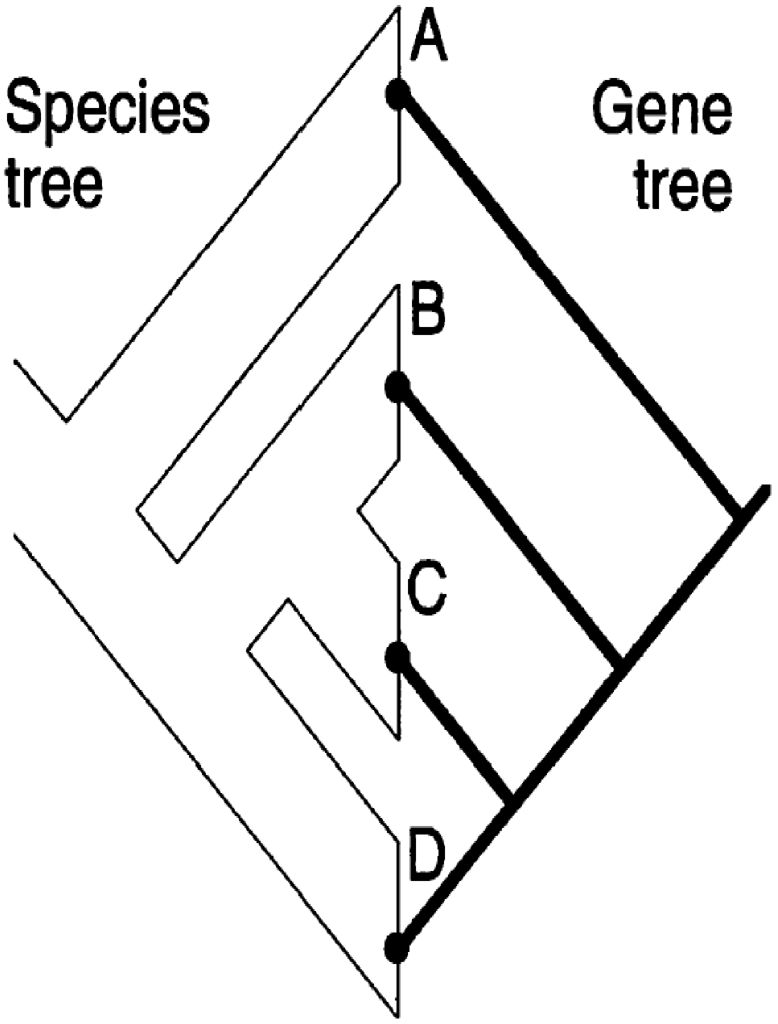
- What is the relationship between Species Trees and Genes Trees?

Gene Trees and Species Trees



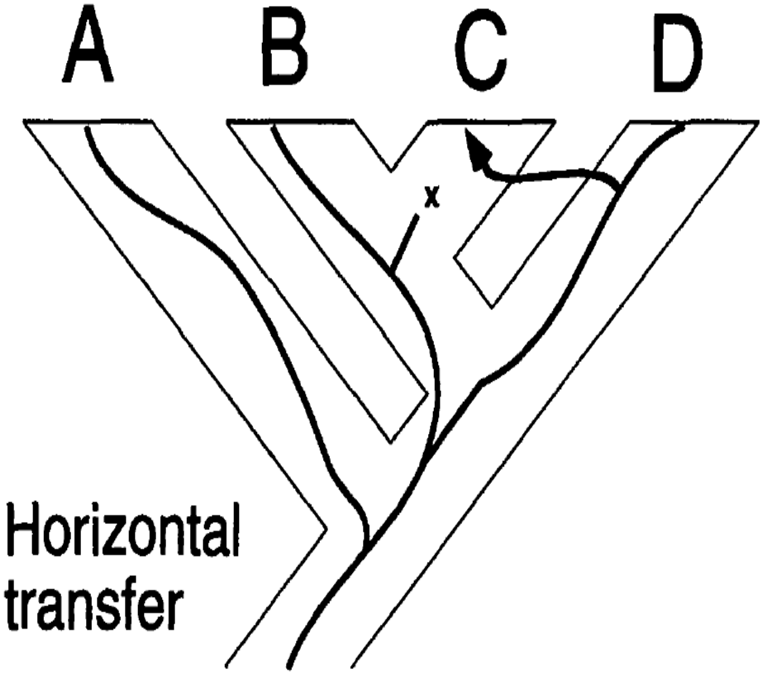
- Gene copies within reproductive communities are also split
- Gene trees – contained within branches of species phylogeny

Discord between Gene Trees and Species Trees



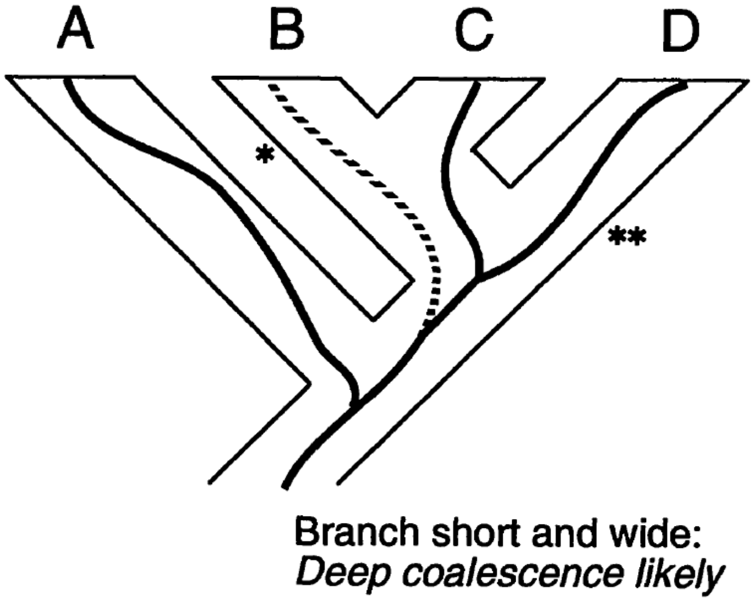
- If in agreement - gene copies should show same branching topology as species tree
- Gene trees has more terminals when more gene copies or extant gene copies are considered

Horizontal Transfer



- Renegade genes breaking confines of species lineages
- Occurs by isolated hybridization events across phylogeny and by vector transfer
- Successful transfer less likely the more phylogentically distant the original and receiving species

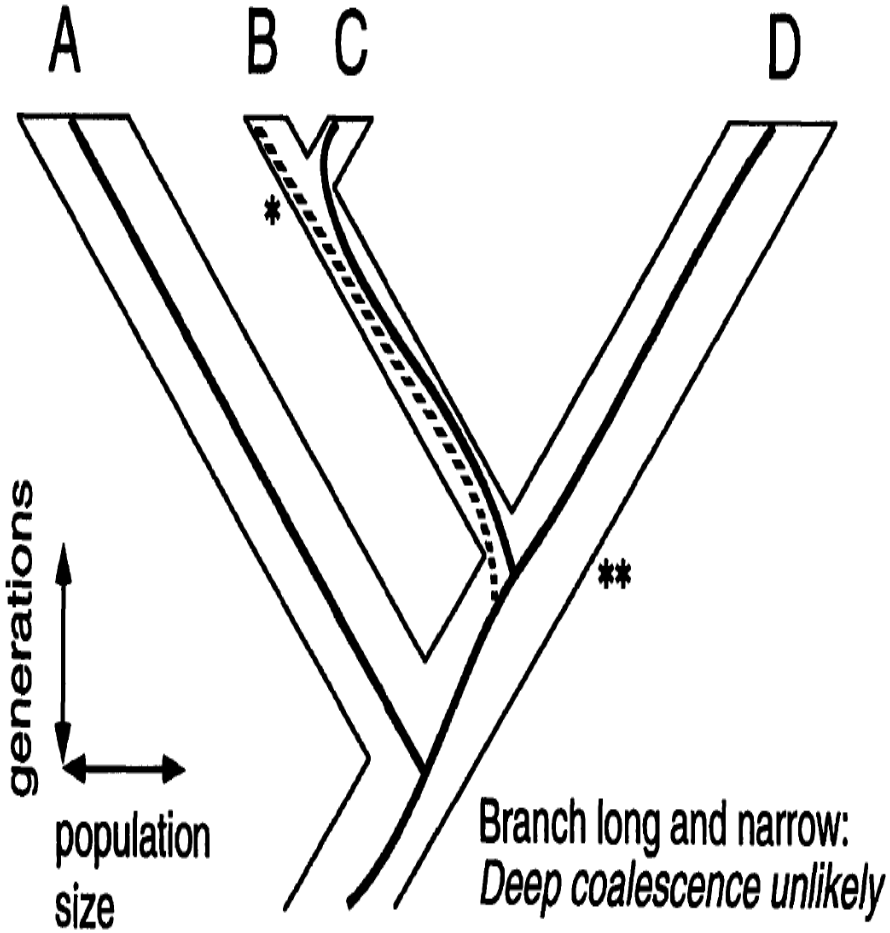
Lineage Sorting or Deep Coalescence



- Common ancestry of gene copies at a single locus extends deeper than speciation events

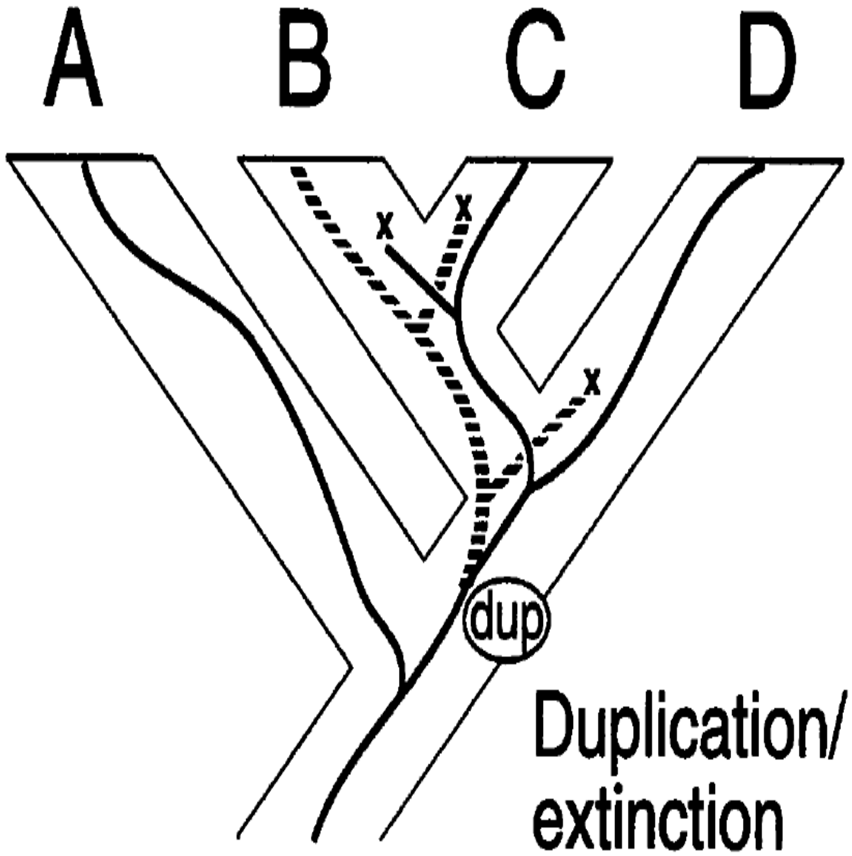
- Coalescence theory – views this process backwards in time
- With large population, gene copies take many generation to 'coalesce'

Lineage Sorting or Deep Coalescence



- Larger the population size and shorter the phylogenetic branch, lesser the chances to coalesce before speciation event

Gene Duplication and Extinction



- Gene duplication yields a second locus
- Both evolve, but only one survive

Gene Duplication and Extinction

- Different from deep coalescence
  - Does not depend on population size

- Reason:
  - Different gene copies are competing for the same locus in deep coalescence
  - Two copies at different loci in the genome – not competing for the same site in gene duplication

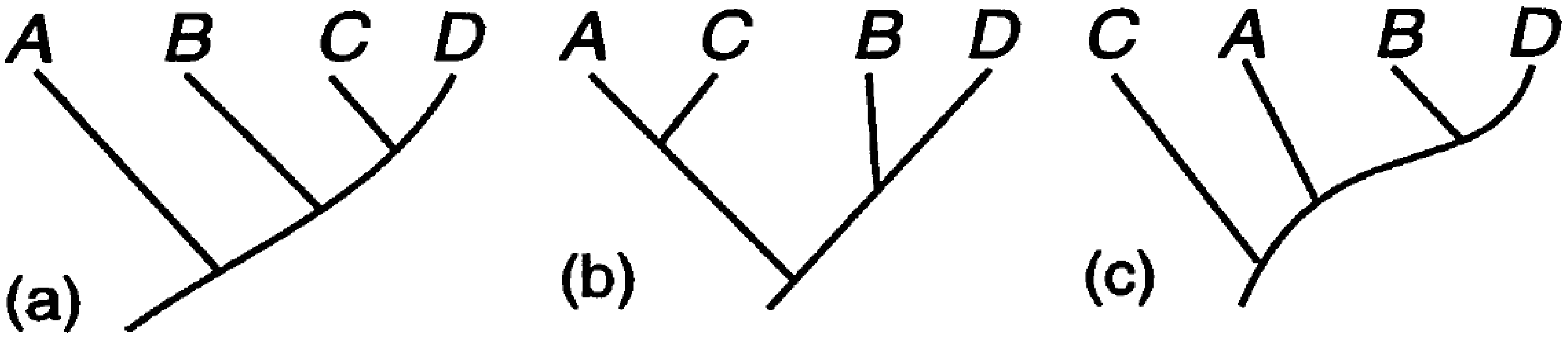
Reconstructing Species Trees from Gene Trees

- Gene trees can disagree with their containing species trees
- Even if gene tree is reconstructed correctly, reconstructed species trees can be wrong
- Gene tree – good indicator for small population size relative to phylogenetic branch length
- Simplest approach - choose the commonest gene tree. Does this works?

Reconstructing Species Trees from Gene Trees

- Simplest approach could fail to reflect the overall support for another tree if a series of less common trees were nearly in agreement and together were much more numerous than the modal tree
- Parsimony procedure to explain evolutionary events

Reconstructing Species Trees from Gene Trees



- Example
  - Ten gene loci sampled from species A, B, C and D
  - (a) and (b) occurs in three genes each
  - (c) occurs in four genes

Reconstructing Species Trees
[Considering Horizontal Transfer]

- Simplest approach – count number of transfers required. But how?
- Common Algorithms used
  - Brooks parsimony analysis
  - Nearest Neighbor Interchange metric (NNI)

Reconstructing Species Trees [Considering Horizontal Transfer]

- Brooks parsimony analysis
  - Uses algorithm for mapping character state trees onto species trees
  - Considers transfer events as character state changes
- Drawback
  - Their cost of gene transfer – depends on distance between entering and resident genes

Reconstructing Species Trees [Considering Horizontal Transfer]

- Nearest Neighbor Interchange
  - Counts branch moves to convert one tree into another
- Drawback
  - Horizontal transfer in nature is not restricted to series of neighbor events

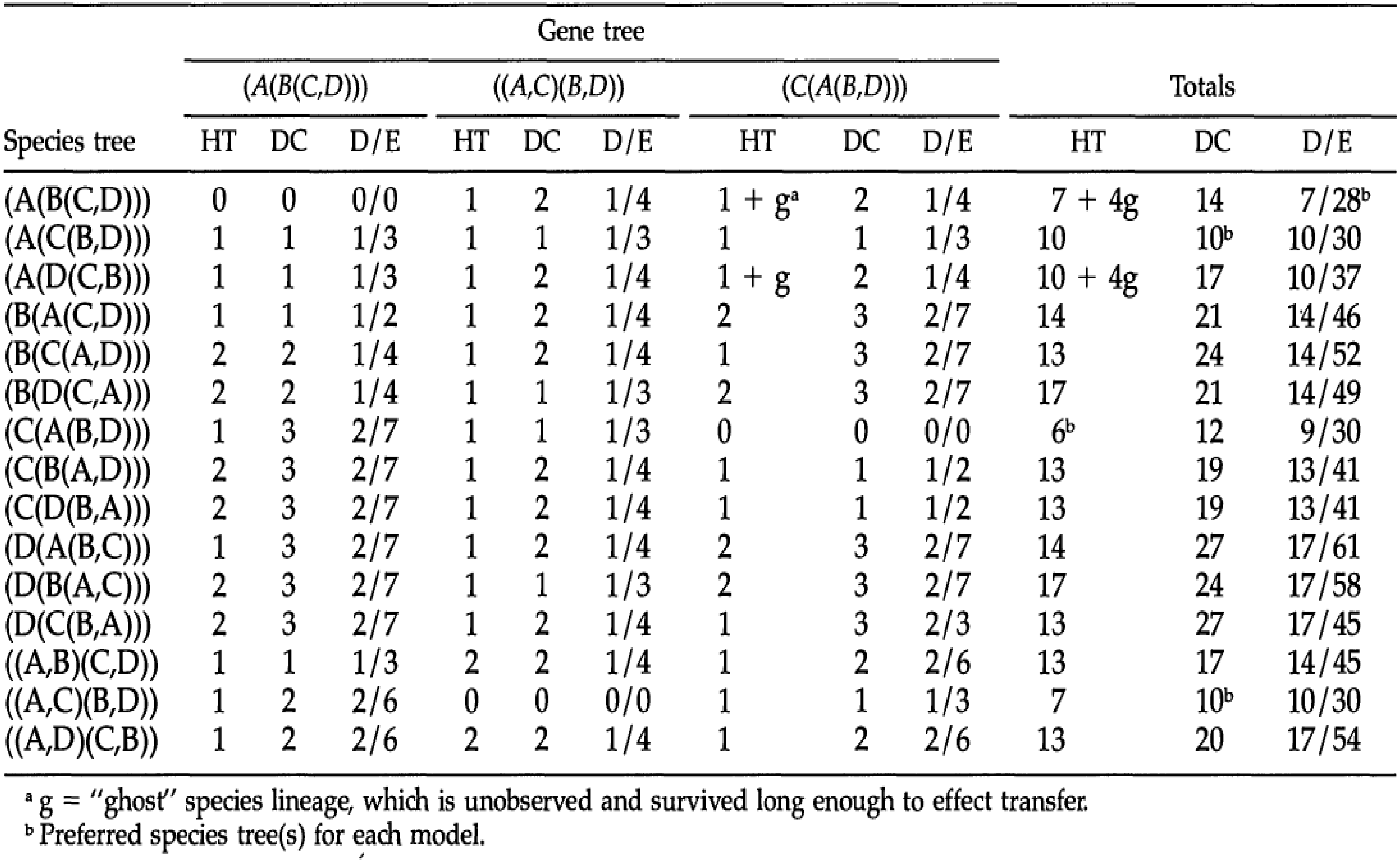
Reconstructing Species Trees [Considering Horizontal Transfer]

- Required – Method that
  - counts minimal number of branch moves required to convert one tree into another
  - Does not violate linear time order
    eg. Move from branch A to branch B and from descendant of B to ancestor of A

Reconstructing Species Trees [Considering Horizontal Transfer] Maddison's Approach

- Direct manual examination of species trees and gene trees
- If gene tree and species tree match, number of transfer events = 0
- Complete explanation – requires complicated events to happen (like extinct species to split early and survive long enough to transfer gene - ghost lineage)

Reconstructing Species Trees
[Considering Horizontal Transfer]

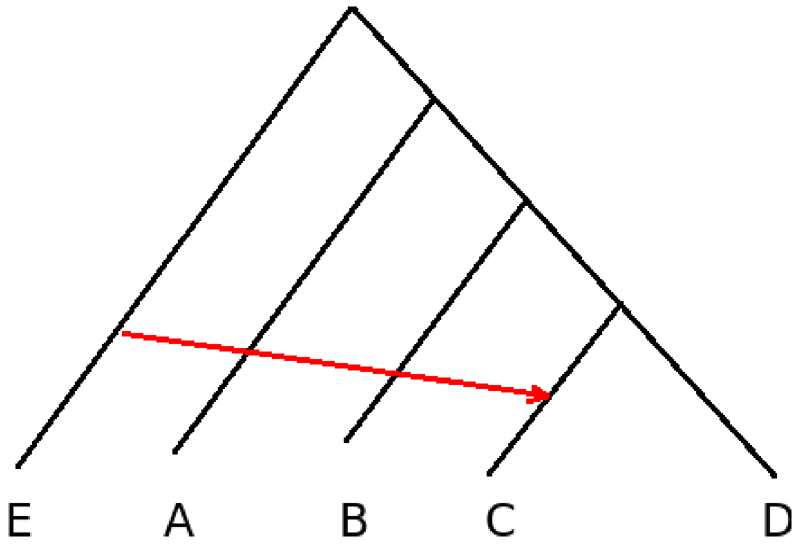


	Gene tree											
	(*A*(*B*(*C*,*D*)))			((*A*,*C*)(*B*,*D*))			(*C*(*A*(*B*,*D*)))			Totals		
Species tree	HT	DC	D/E	HT	DC	D/E	HT	DC	D/E	HT	DC	D/E
(A(B(C,D)))	0	0	0/0	1	2	1/4	1 + g[a]	2	1/4	7 + 4g	14	7/28[b]
(A(C(B,D)))	1	1	1/3	1	1	1/3	1	1	1/3	10	10[b]	10/30
(A(D(C,B)))	1	1	1/3	1	2	1/4	1 + g	2	1/4	10 + 4g	17	10/37
(B(A(C,D)))	1	1	1/2	1	2	1/4	2	3	2/7	14	21	14/46
(B(C(A,D)))	2	2	1/4	1	2	1/4	1	3	2/7	13	24	14/52
(B(D(C,A)))	2	2	1/4	1	1	1/3	2	3	2/7	17	21	14/49
(C(A(B,D)))	1	3	2/7	1	1	1/3	0	0	0/0	6[b]	12	9/30
(C(B(A,D)))	2	3	2/7	1	2	1/4	1	1	1/2	13	19	13/41
(C(D(B,A)))	2	3	2/7	1	2	1/4	1	1	1/2	13	19	13/41
(D(A(B,C)))	1	3	2/7	1	2	1/4	2	3	2/7	14	27	17/61
(D(B(A,C)))	2	3	2/7	1	1	1/3	2	3	2/7	17	24	17/58
(D(C(B,A)))	2	3	2/7	1	2	1/4	1	3	2/3	13	27	17/45
((A,B)(C,D))	1	1	1/3	2	2	1/4	1	2	2/6	13	17	14/45
((A,C)(B,D))	1	2	2/6	0	0	0/0	1	1	1/3	7	10[b]	10/30
((A,D)(C,B))	1	2	2/6	2	2	1/4	1	2	2/6	13	20	17/54

[a] g = "ghost" species lineage, which is unobserved and survived long enough to effect transfer.

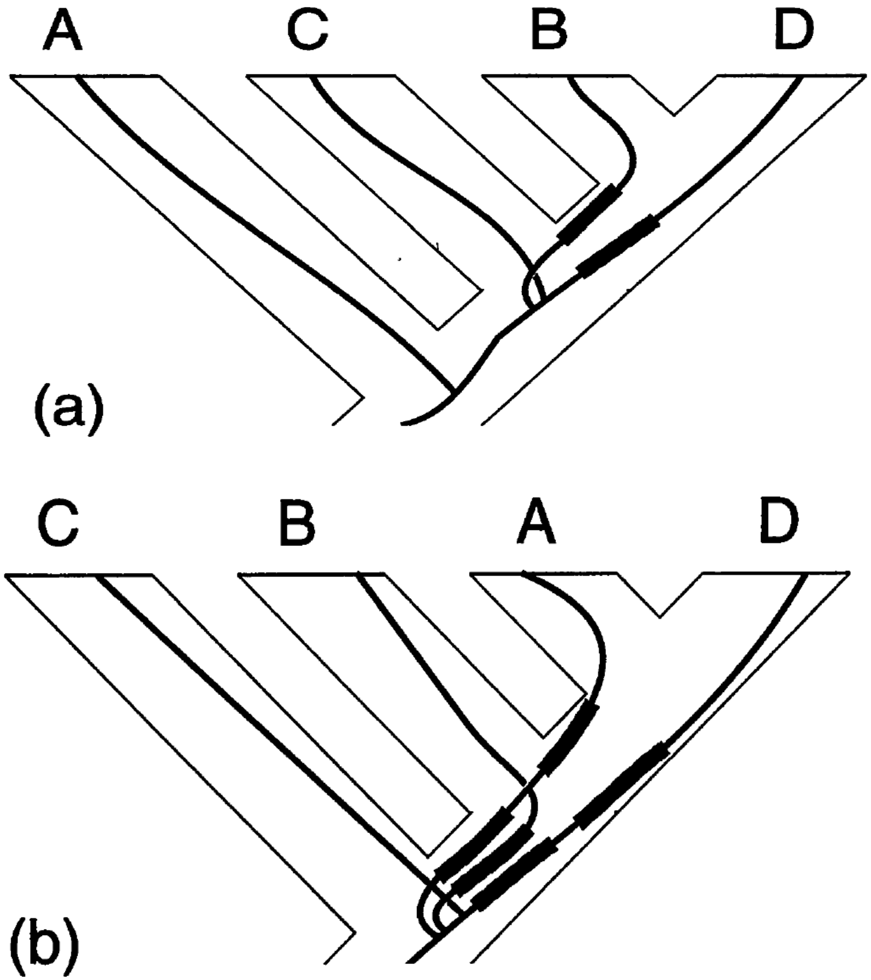
[b] Preferred species tree(s) for each model.

Reconstructing Species Trees [Considering Horizontal Transfer]



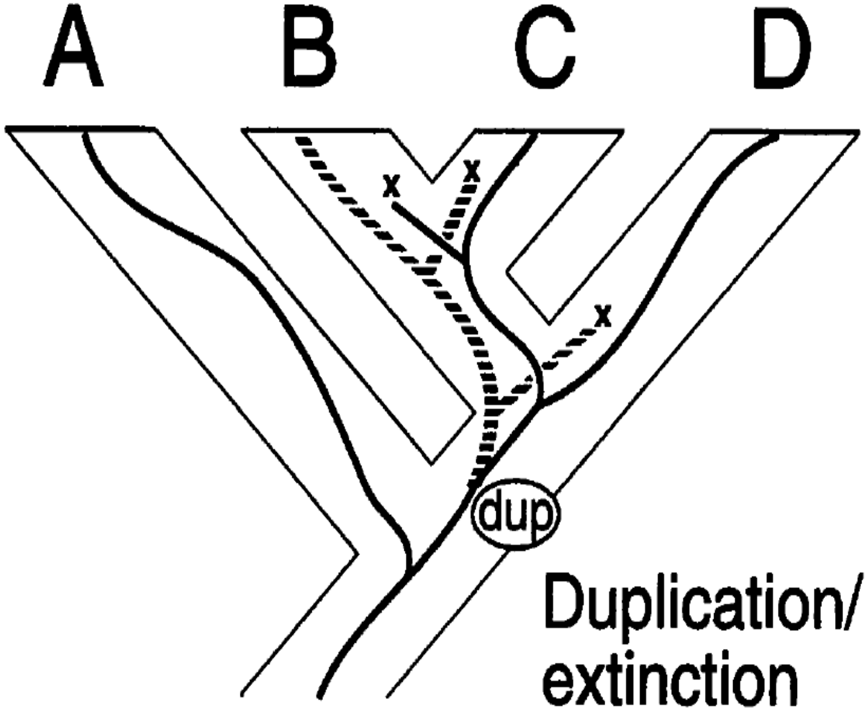
- Here in this case E is a “ghost” species lineage
- E could have gone extinct after the horizontal transfer to C

Reconstructing Species Trees
[Considering Lineage Sorting]



- Number of 'extra' gene lineages on species branches – used for MP
- Bold lines – coexisting gene lineages that fail to coalescence
  - Extra lineage Count: number of gene lineages minus one

Reconstructing Species Trees [Considering Duplication / Extinction]



- Goodman et al
    - Developed algorithm that counts duplication and extinction events when fitting gene tree on species tree

Considering all events together

- Gene tree may be mapped onto species tree invoking bit of all three events
- Problems
  - Algorithm difficulties of assessing multitude of possible scenarios that could be used to fit any given gene tree
  - Difficulty of weighting these different events

Considering all events together

- Each event depend on different circumstances for its occurrence
- Example
  - With longer generations, deep coalescence not favored
  - If vectors and other means are unavailable, horizontal transfer unlikely
  - At small scales, near species level, gene duplication unlikely

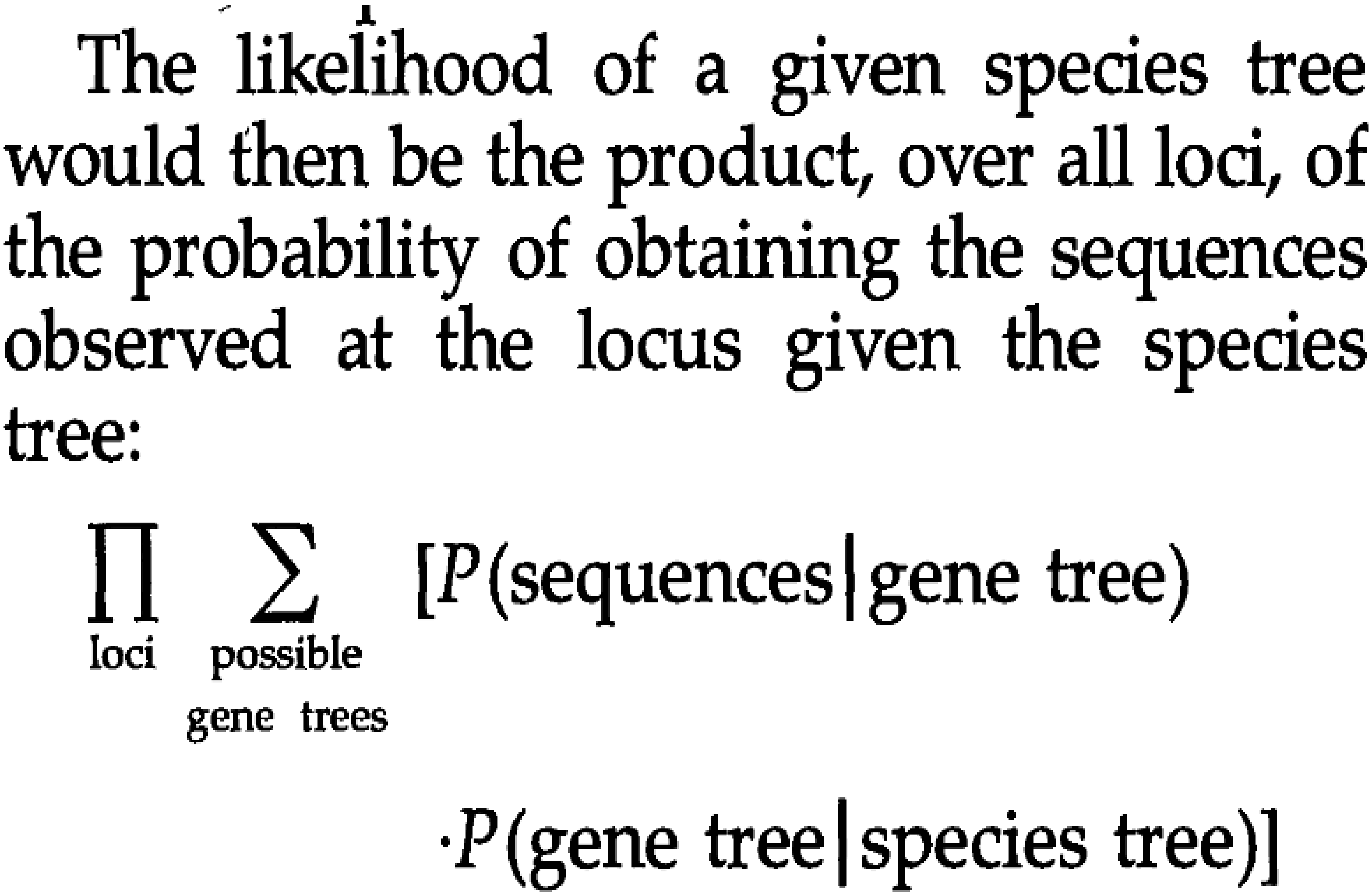
Considering all events together

- Depending on process assumed, most parsimonious species tree could be chosen by counting minimal numbers of transfers, extra gene lineages or duplication and extinction events
- multiple species occupying an area coexist without competition – then behavior more like duplicated genes
- If competition exist, behavior more like alleles at a locus

Species Tree Using MLE

- Considering deep coalescence
  - Coalescence theory provides probability to coalesce
  - Using parameter related to length versus width of the branch
  - Likelihood of species tree – calculated using probability of obtaining set of gene trees from proposed species phylogeny
  - Likelihood depends on
    - Coalescence theory
    - Model of nucleotide evolution

Species Tree Using MLE



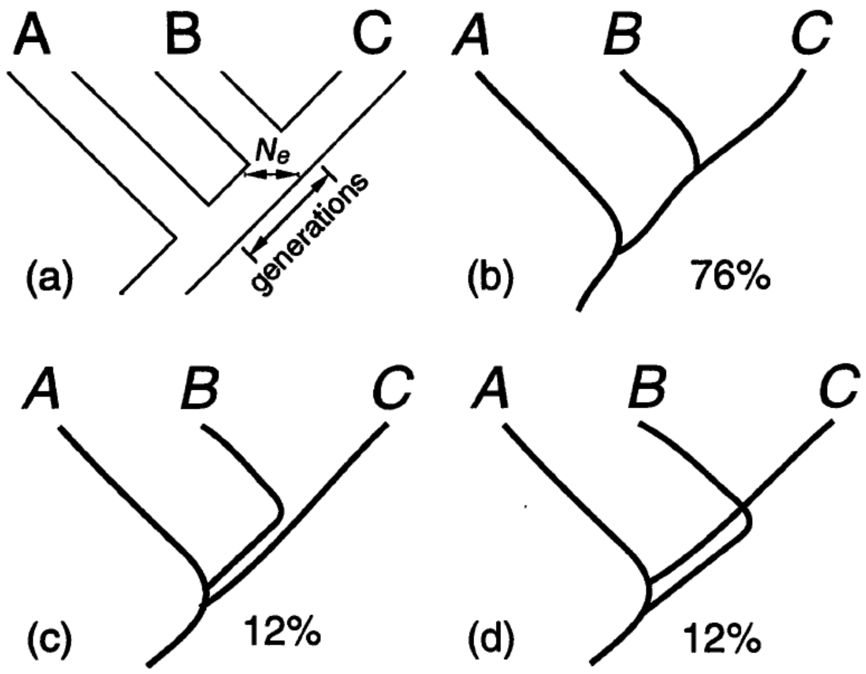
The likelihood of a given species tree would then be the product, over all loci, of the probability of obtaining the sequences observed at the locus given the species tree:

$$\prod_{\text{loci}} \sum_{\substack{\text{possible}\\ \text{gene trees}}} [P(\text{sequences}|\text{gene tree}) \cdot P(\text{gene tree}|\text{species tree})]$$

Species Tree Using MLE

- Probability of occurrence of gene trees given species trees – from coalescence theory
- Probability of occurrence of sequences given gene trees – from model of nucleotide evolution
- Searching ML tree – extremely tedious due to consideration of every species tree for all gene trees
- Felsenstein pointed out possibility of sampling among gene trees in proportion to coalescence probabilities

Phylogeny as a Cloud of Gene Histories



- Using Coalescence theory, ancestors of two sampled gene copies B and C will coalesce with each other about 76 % of the time
- Species phylogeny composed 76% of fig b, 12% of fig c ...

Phylogeny as a Cloud of Gene Histories

- Phylogeny – statistical distribution of samples from population
- Uncertainty – intrinsic part of phylogeny
- Analogous to uncertainty in determining exact position of an electron

Phylogeny as a Cloud of Gene Histories

- To prevent fuzz, can we delimit our species broadly enough?
- Unlikely
  - Since species delimitations would broaden considerably – if gene coalescence exclusivity were applied in its strictest conceivable form
  - May guarantee phylogenetic cleanliness, but tree could still misbehave in its deeper areas
    - because trees of different genes can still disagree on how two species relate

Phylogeny as a Model of Probabilities of Interbreeding

- Phylogeny more than composite of gene histories
- Under model of realized gene histories, species tree consists of gene trees
- Under a model of species tree as fragmenting container that stands apart from and constrains descent of gene trees

Overview

- Gene Trees and Species Trees
- Processes of Discord
  - Horizontal Transfer
  - Lineage Sorting or Deep Coalescence
  - Gene Duplication/Extinction
- Reconstructing Species Tree Using Gene Trees
- Using MLE for Deep Coalescence
- Species Phylogeny - New interpretation
- Conclusion

Conclusion

- Other concepts of phylogeny
  - Could be viewed as extended pedigree of individual organisms, a summary of realized matings
- Important conclusion
  - “History of genetic descent does not take the form of a simple tree with stick-like branches”
  - Need to confront the composite, cloud like nature of genetic history