Phylogeny Inference in the Presence of Hybridization and Incomplete Lineage Sorting

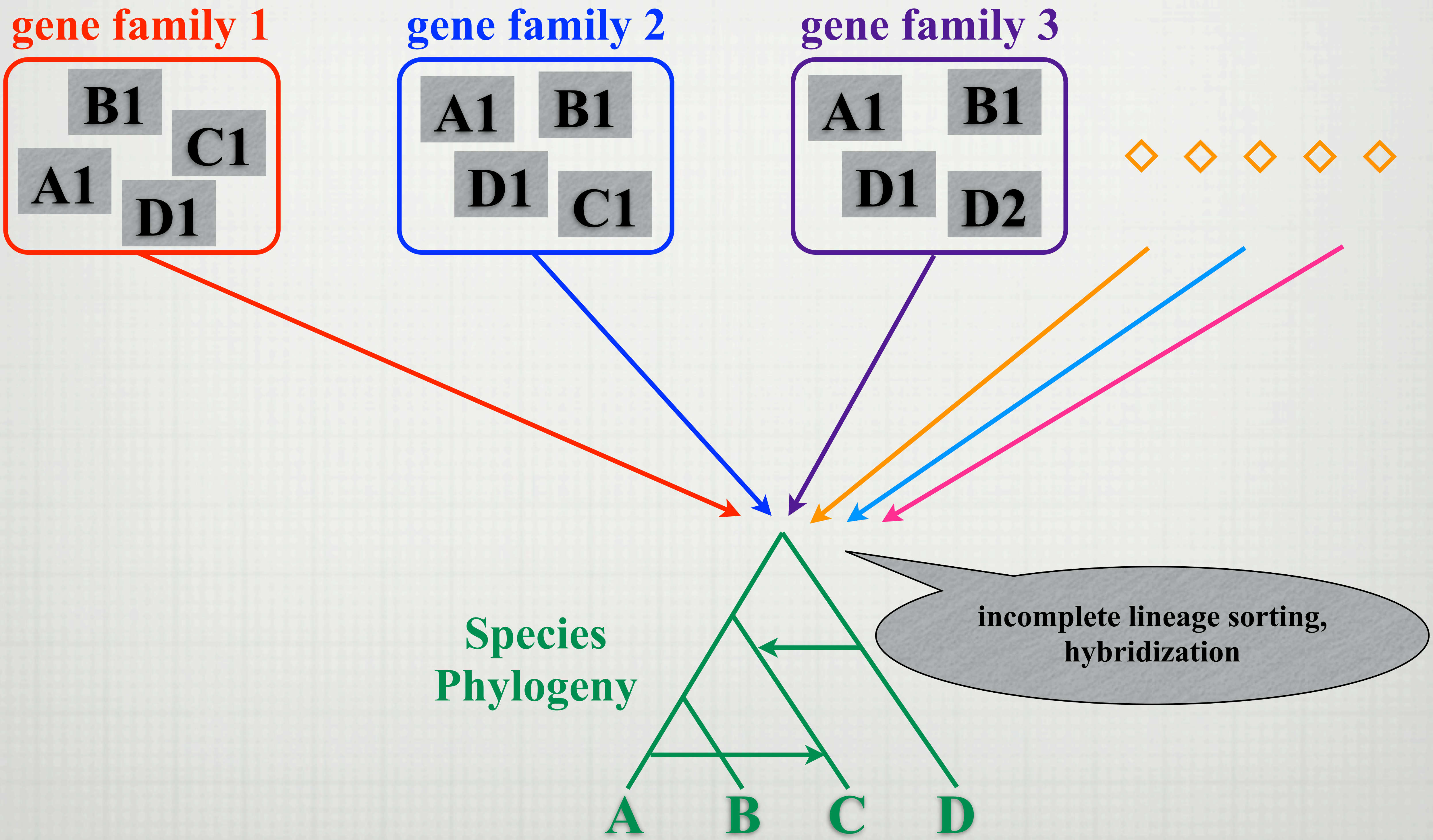
Luay Nakhleh
Department of Computer Science
Rice University

Symposium on New Methods for Phylogenomics and Metagenomics
The University of Texas at Austin
16 February 2013

OUTLINE

- □ Methods for **phylogenetic network** inference from **gene tree topologies** when both **incomplete lineage sorting (ILS)** and **hybridization** are at play
  - □ Part I: A probabilistic approach
  - □ Part II: A parsimony approach

PHYLOGENY RECONSTRUCTION IN THE POST-GENOMIC ERA



gene family 1

B1 C1 A1 D1

gene family 2

A1 B1 D1 C1

gene family 3

A1 B1 D1 D2

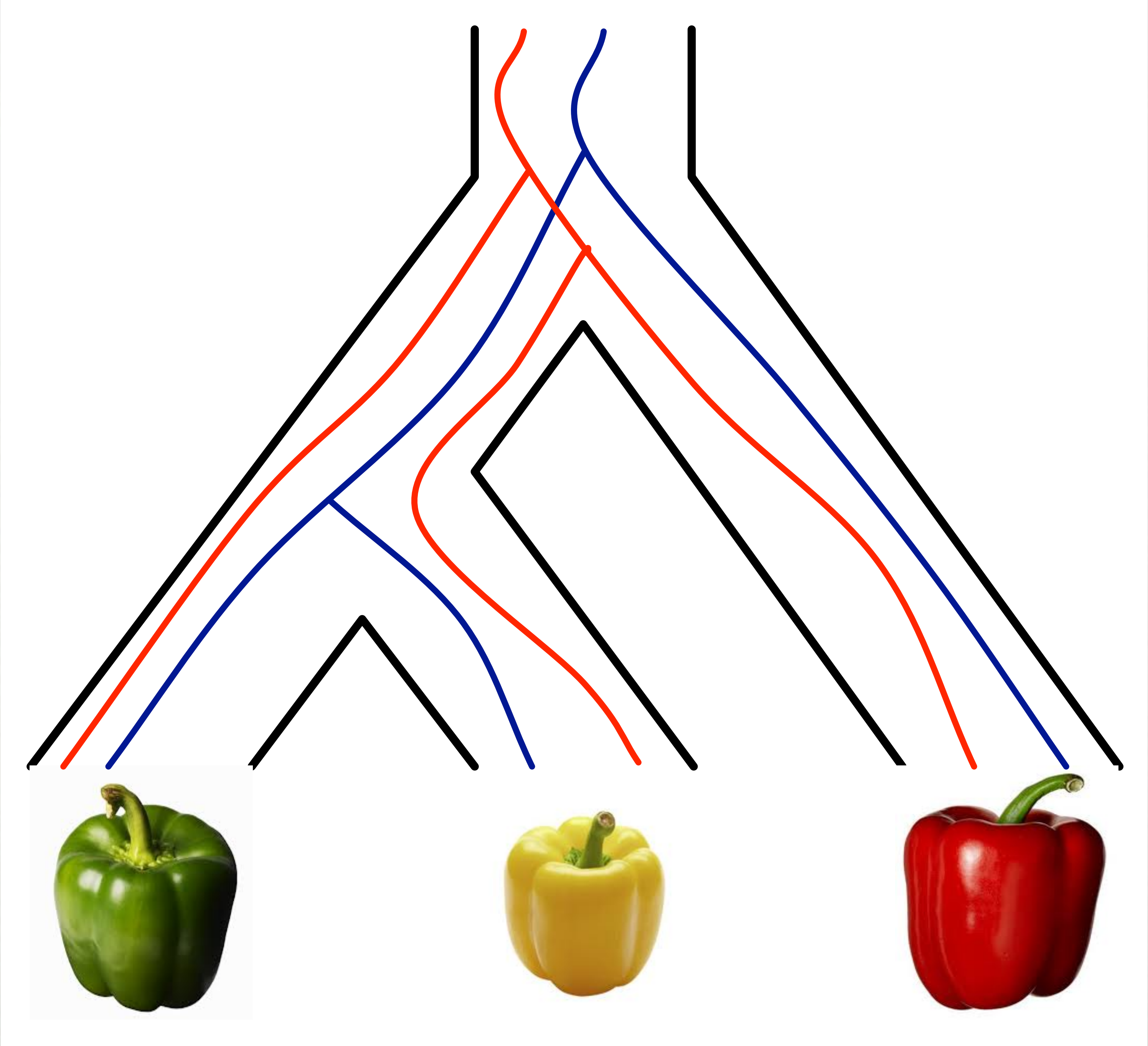
◇ ◇ ◇ ◇ ◇

Species Phylogeny

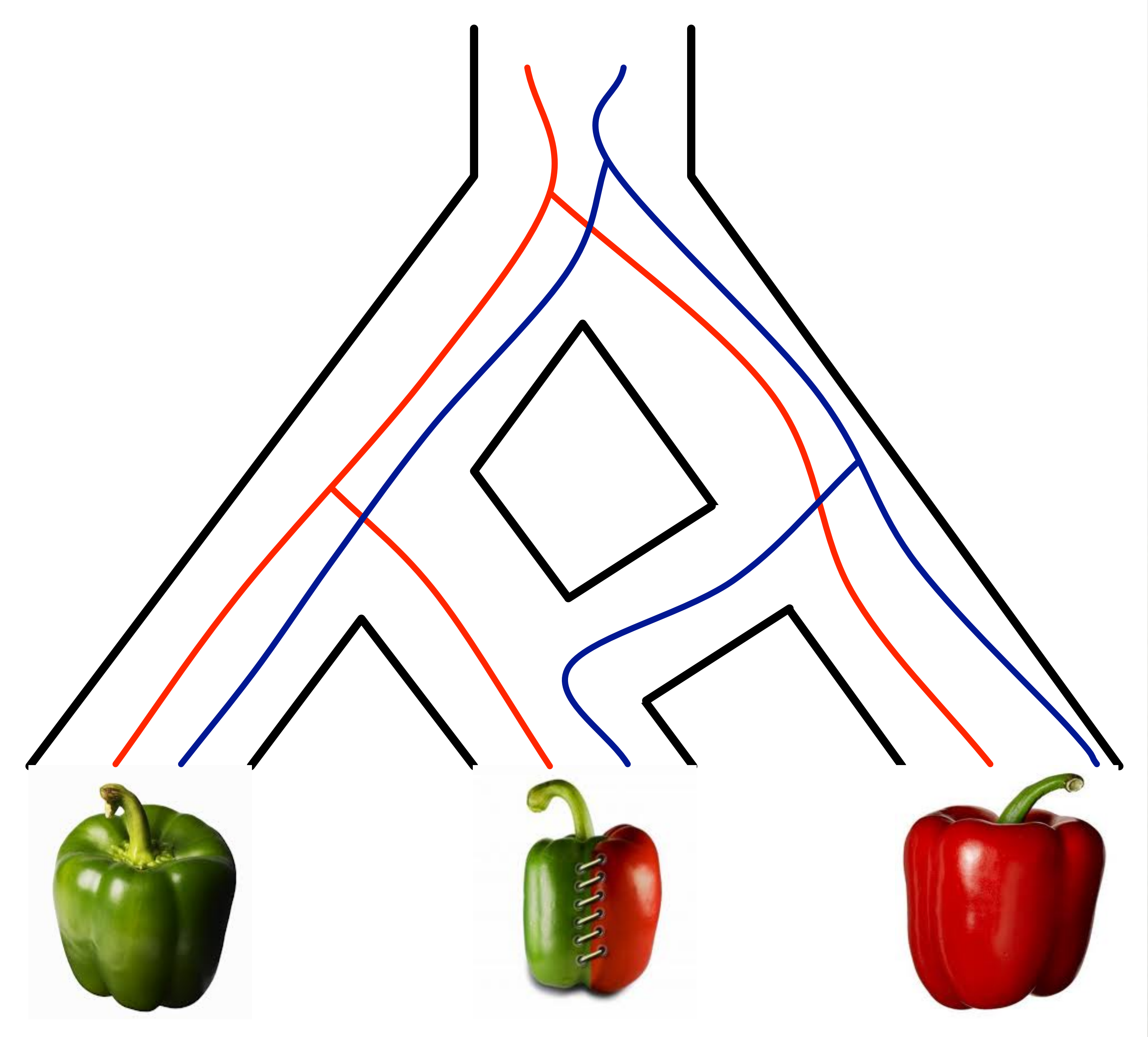
A B C D

incomplete lineage sorting, hybridization

INCOMPLETE LINEAGE SORTING (ILS)

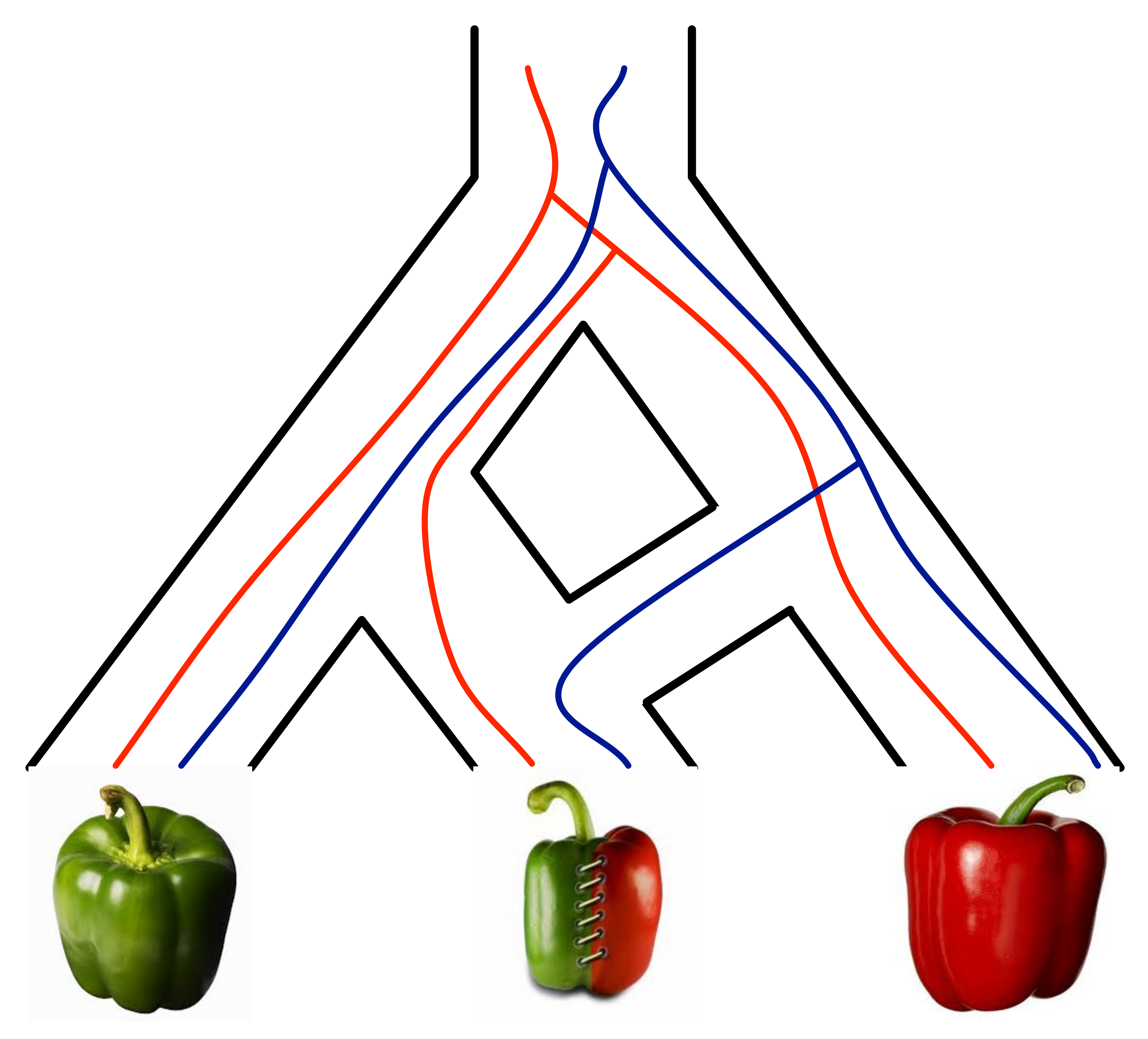


HYBRIDIZATION



source for “hybrid bell pepper”: http://blog.onesuite.com/index.php/blog/item/64-onesuite-the-hybrid-communications-service.html

HYBRIDIZATION AND ILS



source for “hybrid bell pepper”: http://blog.onesuite.com/index.php/blog/item/64-onesuite-the-hybrid-communications-service.html

A PROBABILISTIC APPROACH

$$L(\Psi|\mathcal{S}) = \prod_{S\in\mathcal{S}} \left[\sum_{T} [\mathbf{P}(S|T)\cdot\mathbf{P}(T|\Psi)]\right]$$

species phylogeny
and its parameters

sequences of
gene families

A PROBABILISTIC APPROACH

$$L(\Psi|\mathcal{S}) = \prod_{S\in\mathcal{S}} \left[\sum_{T} [\mathbf{P}(S|T) \cdot \mathbf{P}(T|\Psi)] \right]$$

species phylogeny
and its parameters

sequences of
gene families

If a gene tree has been inferred for each gene family, then:

$$L(\Psi|\mathcal{G}) = c \cdot \prod_{gt\in\mathcal{G}} \mathbf{P}(gt|\Psi)$$

A PROBABILISTIC APPROACH

$$L(\Psi|\mathcal{S}) = \prod_{S\in\mathcal{S}} \left[\sum_{T} [\mathbf{P}(S|T)\cdot\mathbf{P}(T|\Psi)] \right]$$

species phylogeny
and its parameters

sequences of
gene families

If a gene tree has been inferred for each gene family, then:

$$L(\Psi|\mathcal{G}) = c \cdot \prod_{gt\in\mathcal{G}} \mathbf{P}(gt|\Psi)$$

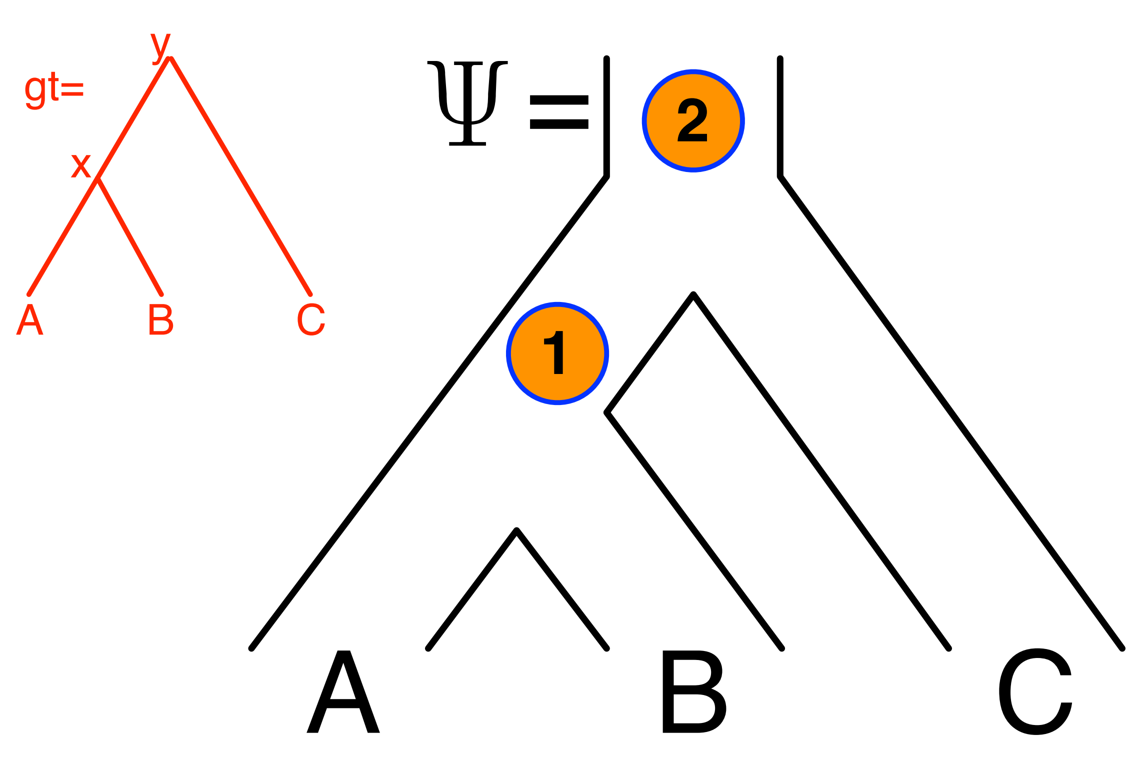
How do we compute $\mathbf{P}(gt|\Psi)$?

$\mathbf{P}(gt|\Psi)$

- The probability of observing the gene tree topolpogy gt given species phylogeny Ψ
- Three cases:
  - Under the coalescent
  - Under HGT
  - Under both

$\mathbf{P}(gt|\Psi)$ UNDER THE COALESCENT

- □ Denote by H_Ψ(gt) the set of all coalescent histories of species tree Ψ and gene tree topology gt



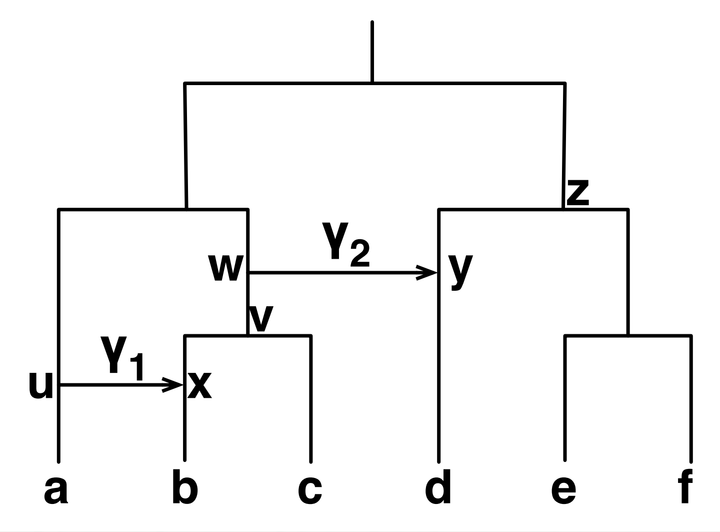
$$H_\Psi(gt) = \{(1,2),(2,2)\}$$

$\mathbf{P}(gt|\Psi)$ UNDER THE COALESCENT

□ Degnan and Salter (Evolution, 2005) gave the mass probability function of a gene tree topology gt for a given species tree with topology Ψ and vector of branch lengths **λ:**

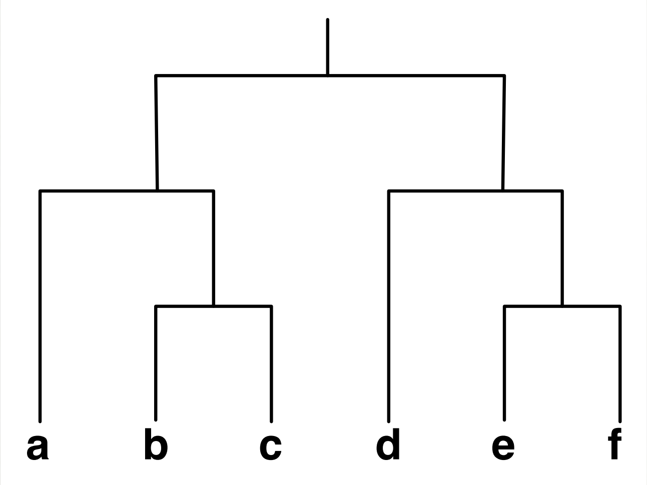
$$P_{\Psi,\lambda}(gt) = \sum_{h \in H_\Psi(gt)} \frac{w(h)}{d(h)} \prod_{b=1}^{n-2} \frac{w_b(h)}{d_b(h)} p_{u_b(h) v_b(h)}(\lambda_b)$$

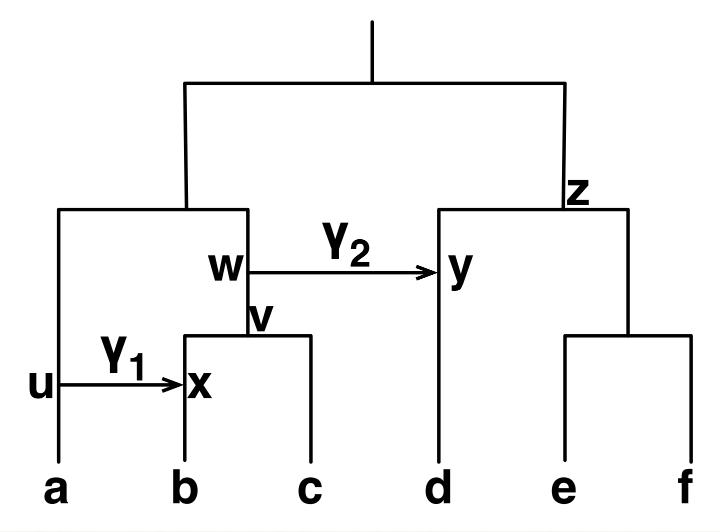
$\mathbf{P}(gt|\Psi)$ UNDER HYBRIDIZATION



$\mathbf{P}(gt|\Psi)$ UNDER HYBRIDIZATION

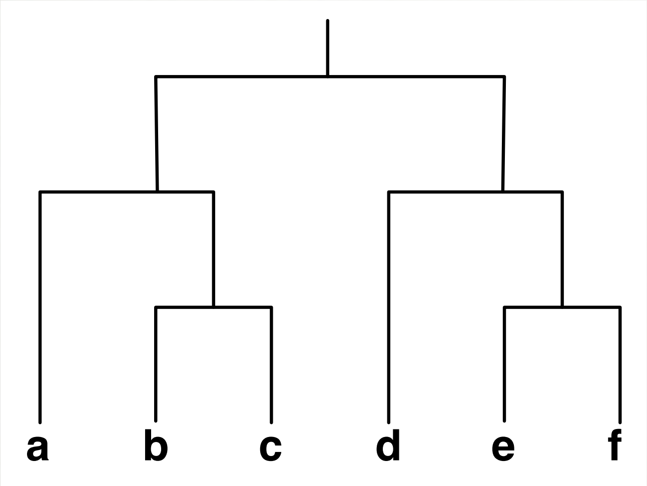
$$P_{N,\gamma_1,\gamma_2}(gt) = (1-\gamma_1)(1-\gamma_2)$$

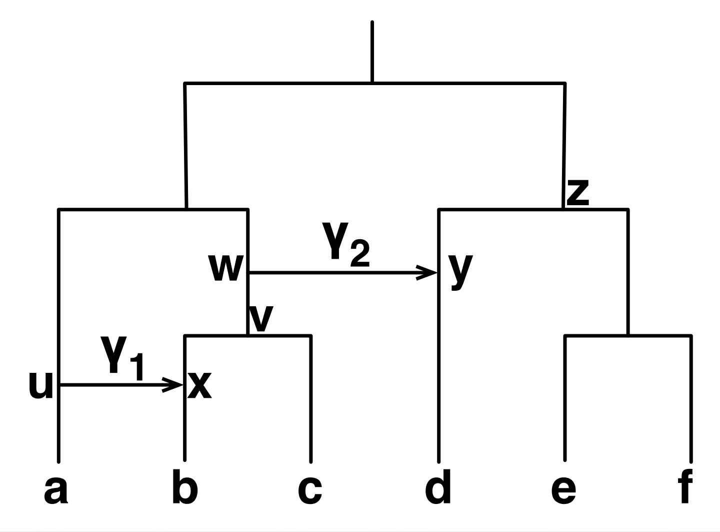


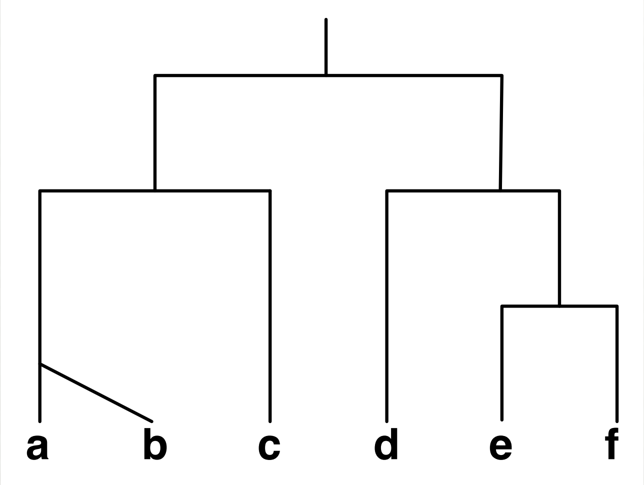


$\mathbf{P}(gt|\Psi)$ UNDER HYBRIDIZATION

$$P_{N,\gamma_1,\gamma_2}(gt) = (1-\gamma_1)(1-\gamma_2)$$



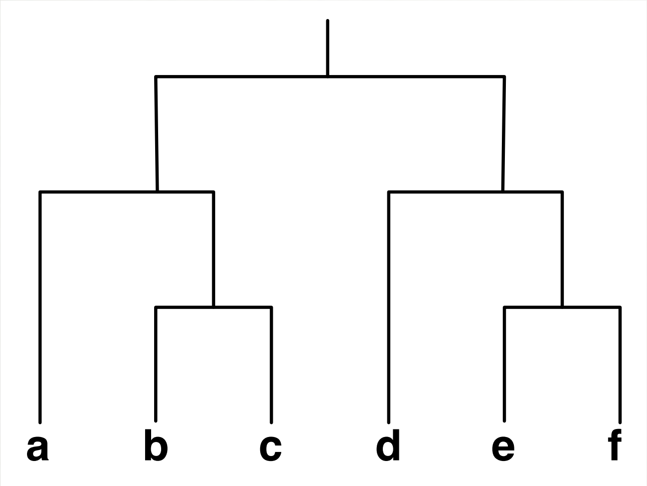




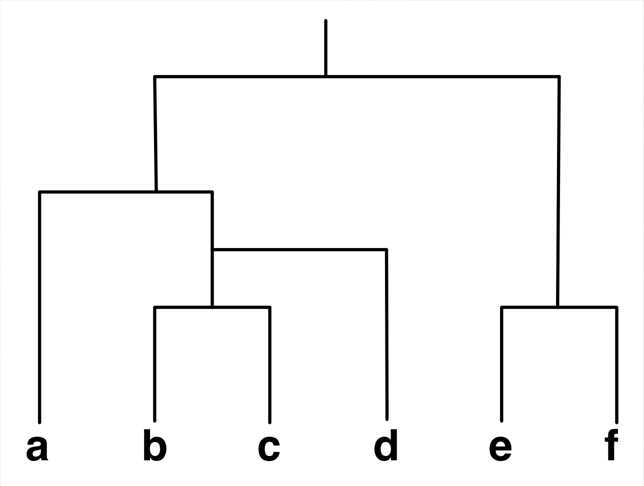
$$P_{N,\gamma_1,\gamma_2}(gt) = \gamma_1(1-\gamma_2)$$

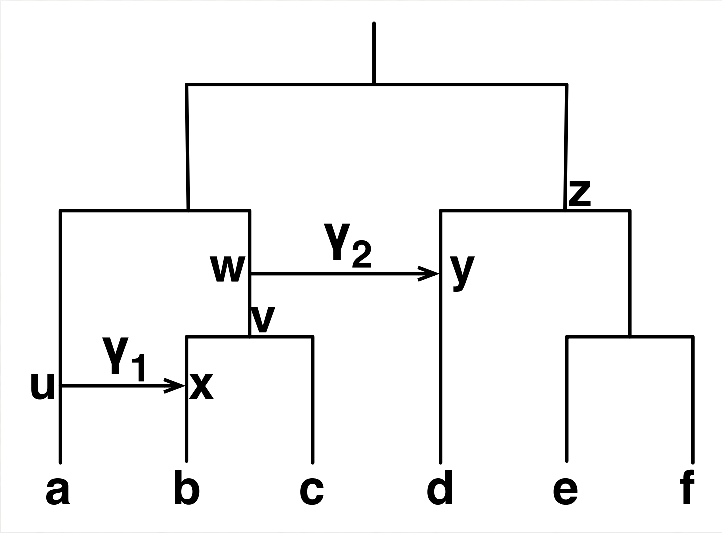
$\mathbf{P}(gt|\Psi)$ UNDER HYBRIDIZATION

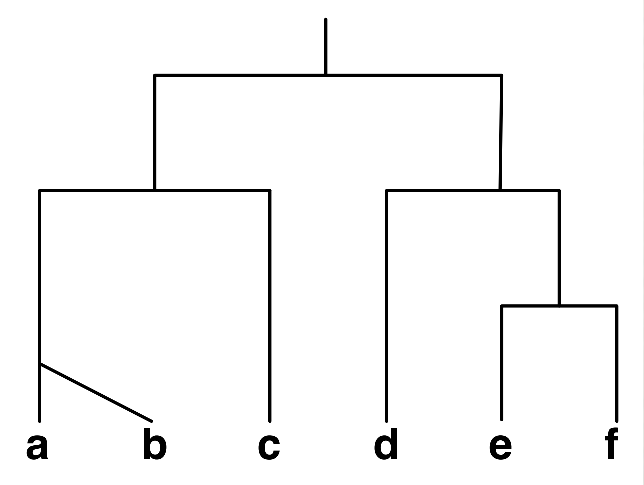
$P_{N,\gamma_1,\gamma_2}(gt) = (1-\gamma_1)(1-\gamma_2)$



$P_{N,\gamma_1,\gamma_2}(gt) = (1-\gamma_1)\gamma_2$



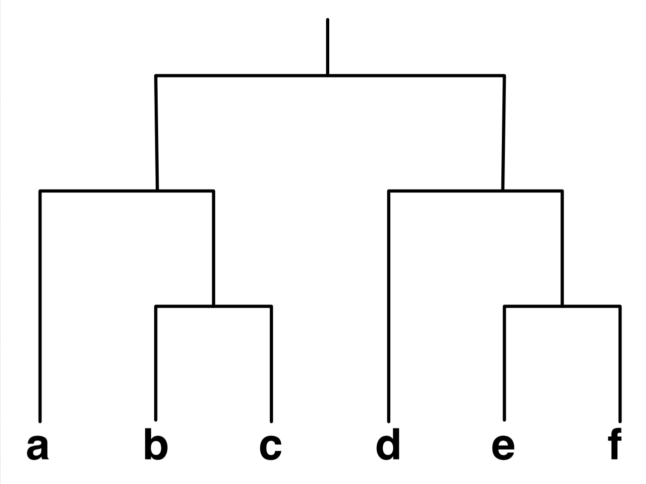




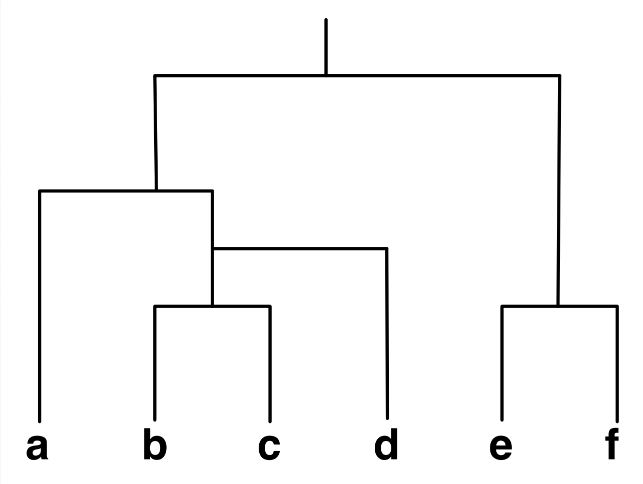
$P_{N,\gamma_1,\gamma_2}(gt) = \gamma_1(1-\gamma_2)$

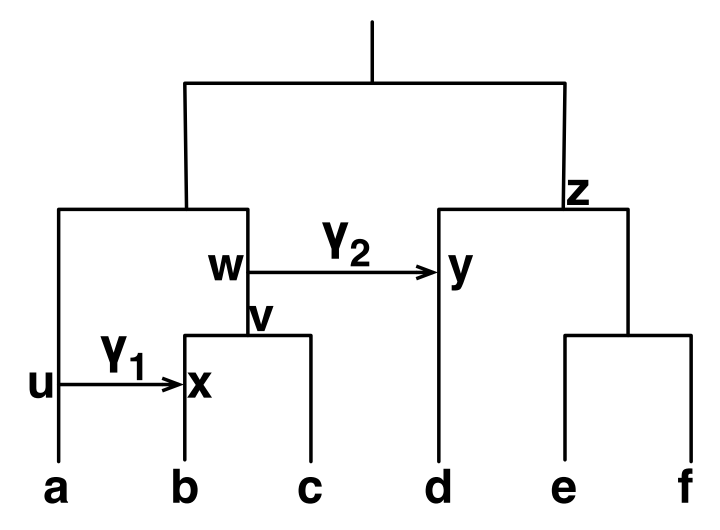
$\mathbf{P}(gt|\Psi)$ UNDER HYBRIDIZATION

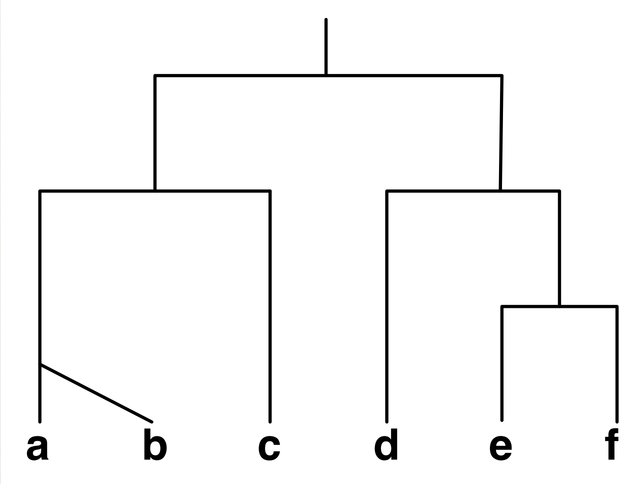
$$P_{N,\gamma_1,\gamma_2}(gt) = (1-\gamma_1)(1-\gamma_2)$$



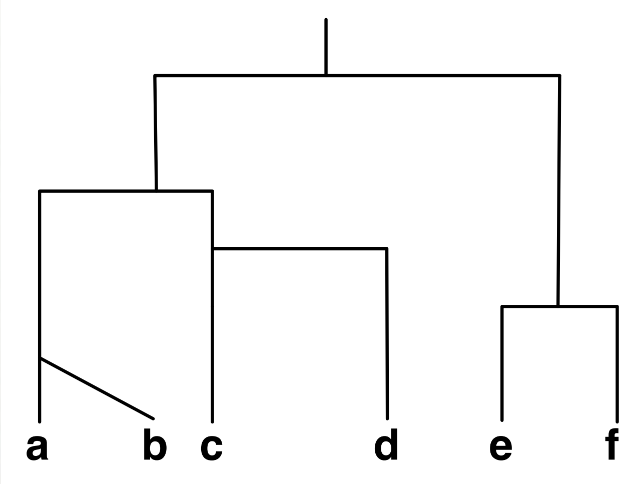
$$P_{N,\gamma_1,\gamma_2}(gt) = (1-\gamma_1)\gamma_2$$





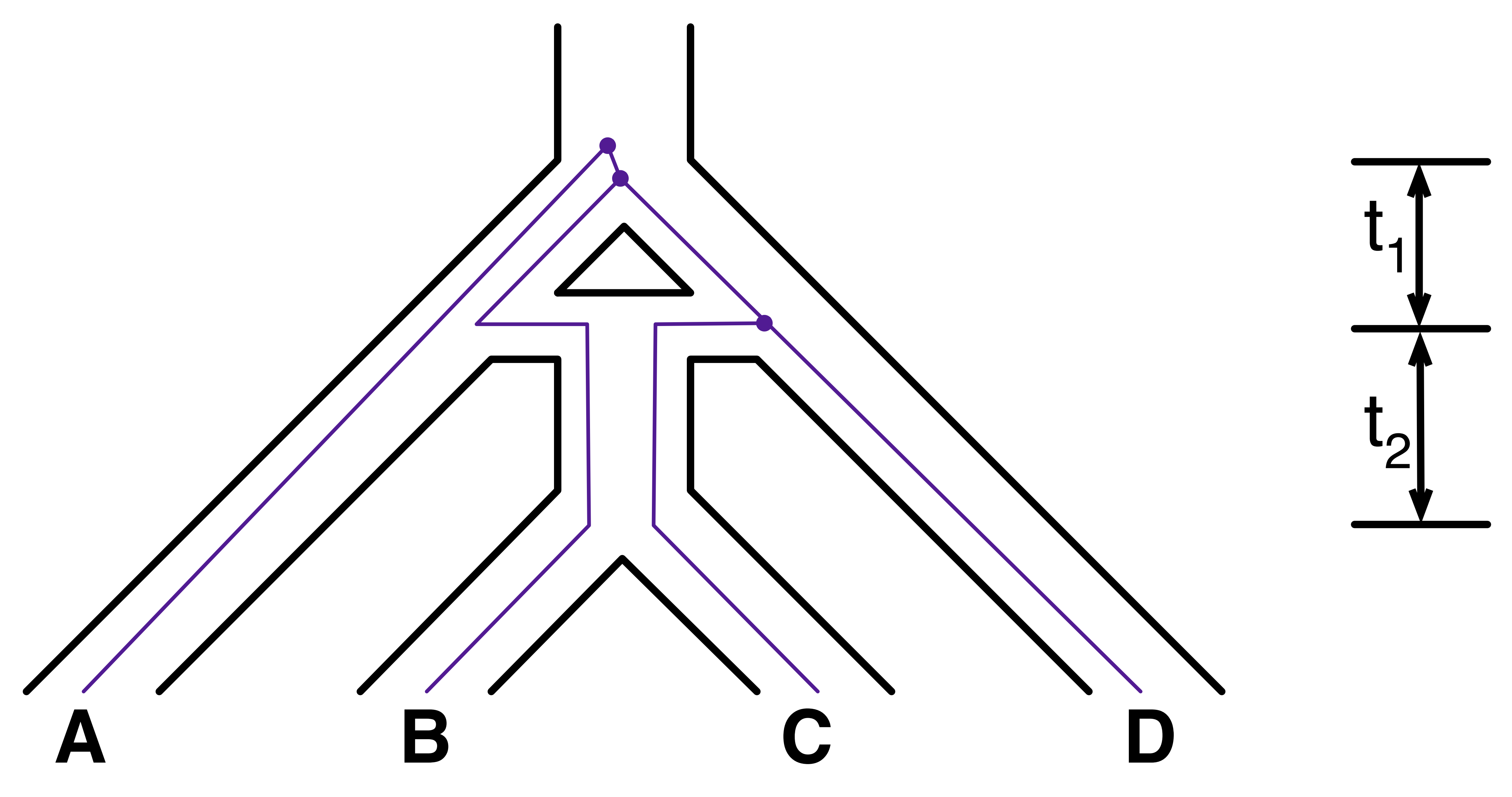


$$P_{N,\gamma_1,\gamma_2}(gt) = \gamma_1(1-\gamma_2)$$



$$P_{N,\gamma_1,\gamma_2}(gt) = \gamma_1\gamma_2$$

$\mathbf{P}(gt|\Psi)$ UNDER BOTH SCENARIOS

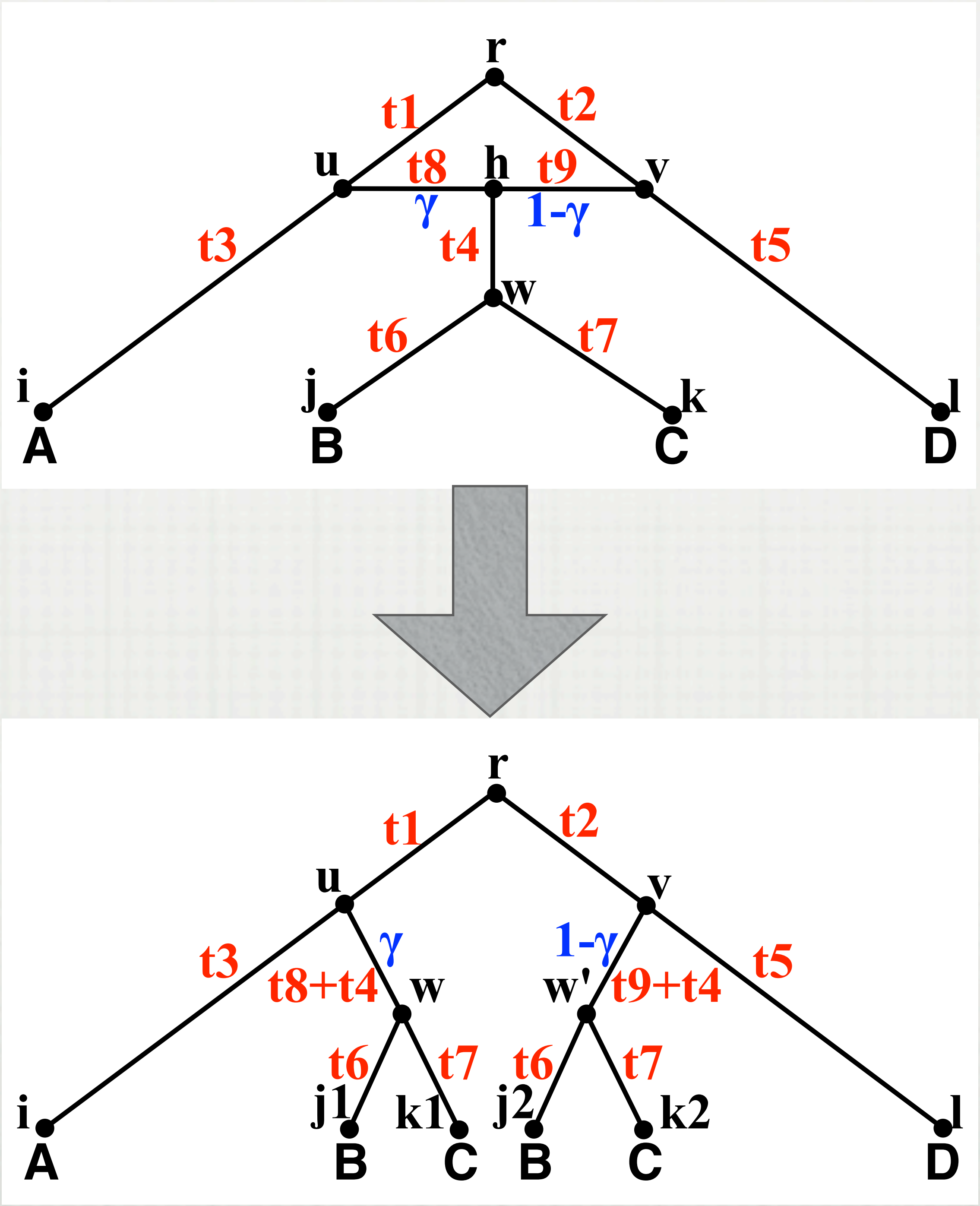


A SOLUTION

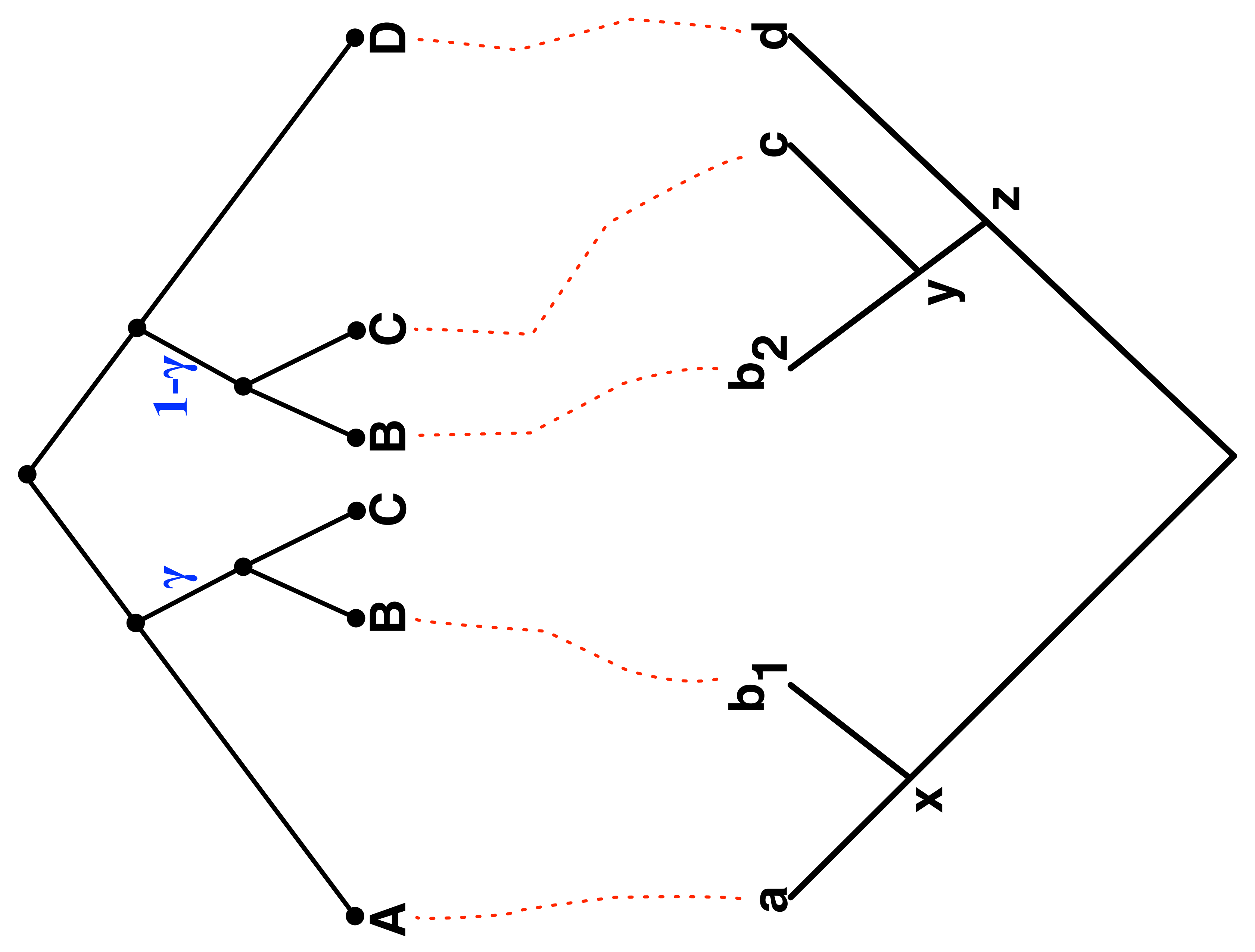
1. Convert the phylogenetic network N into a MUL-tree T
2. Consider all allele mappings from the leaves of gt to the leaves of T
3. For each allele mapping, compute the probability of observing gt, given T, and sum the probabilities.

[Yu, Degnan, Nakhleh, PLoS Genetics, 2012.]

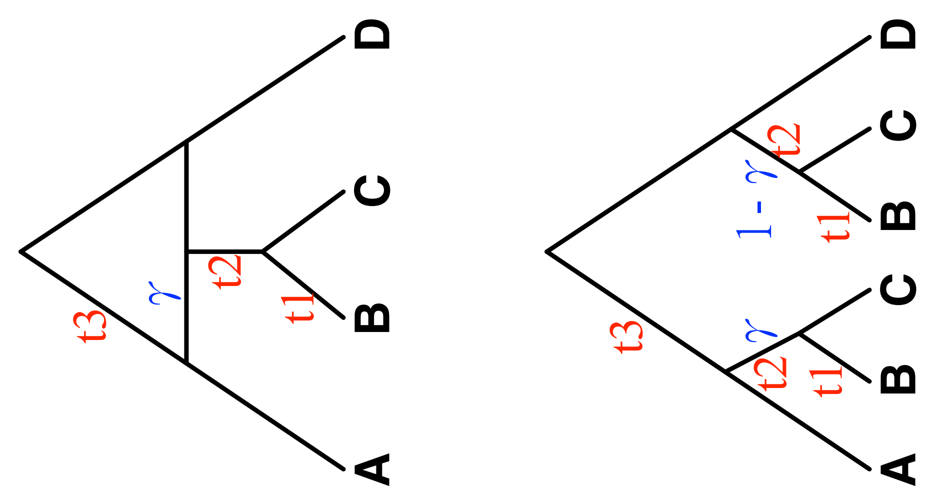
1. FROM A NETWORK TO A MUL-TREE



2. CONSIDER ALL ALLELE MAPPINGS FROM gt TO T



2. CONSIDER ALL ALLELE MAPPINGS FROM gt TO T



Phylogenetic network

MUL tree

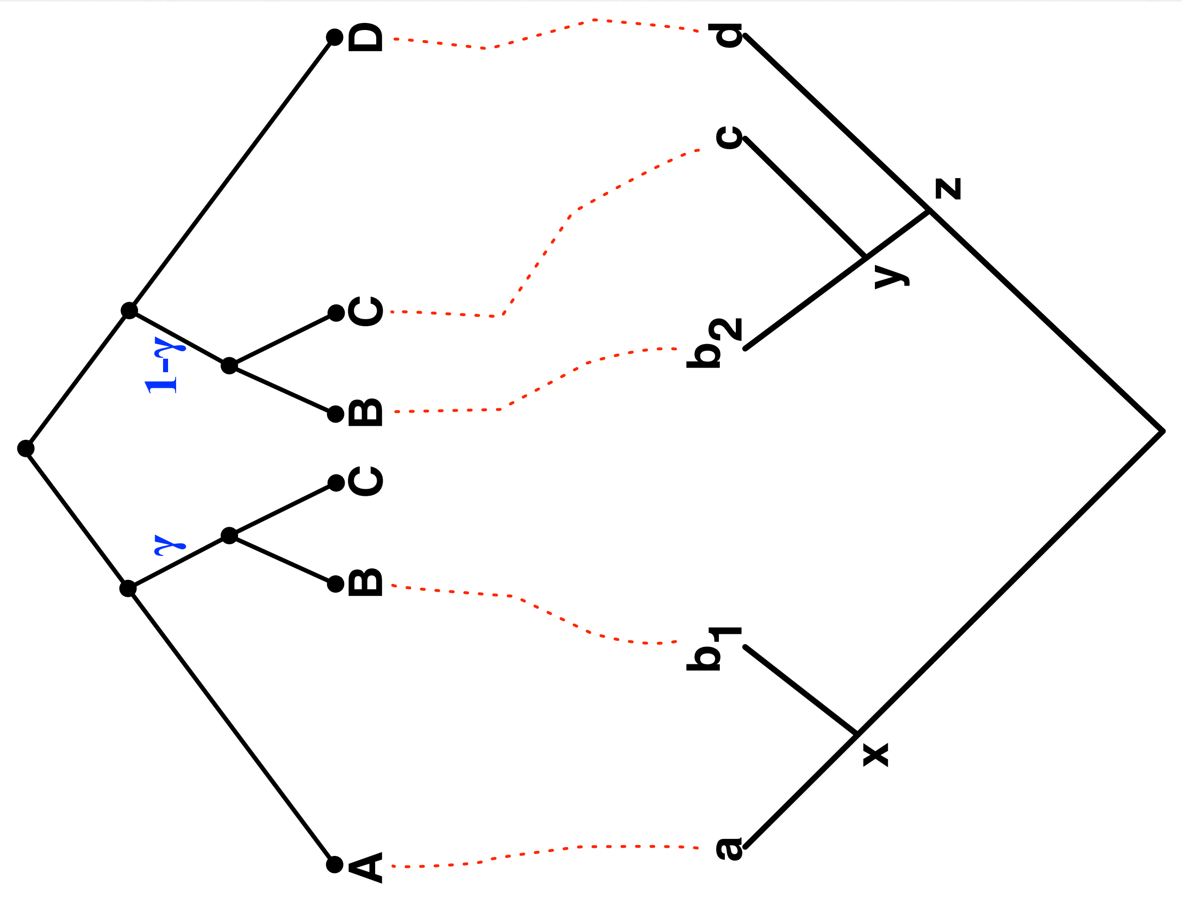
Valid allele mappings

f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
d	d	d	d	d	d	d	d
—	c	—	c	—	c	—	c
—	—	b_2	b_2	b_1	b_1	b_1 b_2	b_1 b_2
c	—	c	—	c	—	c	—
b_1 b_2	b_1 b_2	b_1	b_1	b_2	b_2	—	—
a	a	a	a	a	a	a	a

$$P_{N,\boldsymbol{\lambda},\gamma}(gt) = \sum_{f\in\mathcal{F}} P_{T,\boldsymbol{\lambda}',\gamma',f}(gt)$$

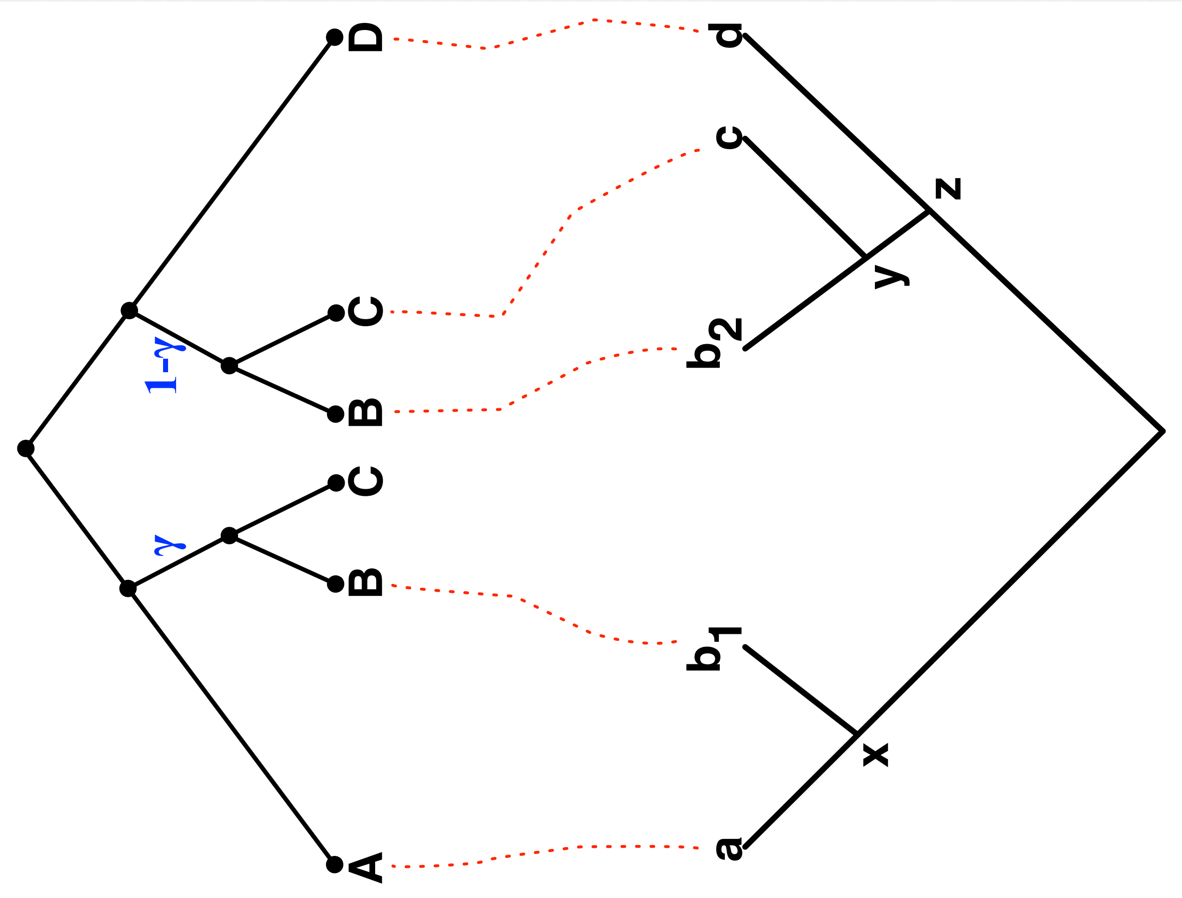
2. CONSIDER ALL ALLELE MAPPINGS FROM gt TO T

- We need to account for dependence among the branches of the MUL-tree



2. CONSIDER ALL ALLELE MAPPINGS FROM gt TO T

□ We need to account for dependence among the branches of the MUL-tree



□ The edge-mapping ϕ solves this problem.

3. THE PROBABILITY OF gt GIVEN MUL-TREE T

$$P_{T,\boldsymbol{\lambda}',\boldsymbol{\gamma}',f}(gt) = \sum_{h\in H_{T,f}(gt)} \frac{w(h)}{d(h)} \prod_{b=1}^{n-2} {\gamma_b'}^{v_b(h)} P_b'(h)$$

$$\prod_{b\in\phi^{-1}(b')} P_b'(h) = \left[\frac{1}{d_{b'}(h)} p_{u_{b'}(h)v_{b'}(h)}(\lambda_{b'}) \left[(u_{b'}(h)-v_{b'}(h))! \prod_{b\in\phi^{-1}(b')} \frac{w_b(h)}{(u_b(h)-v_b(h))!}\right]\right]$$

$$u_{b'}(h) = \sum_{b\in\phi^{-1}(b')} u_b(h) \qquad v_{b'}(h) = \sum_{b\in\phi^{-1}(b')} v_b(h)$$

ACCOUNTING FOR UNCERTAINTY IN GENE TREES

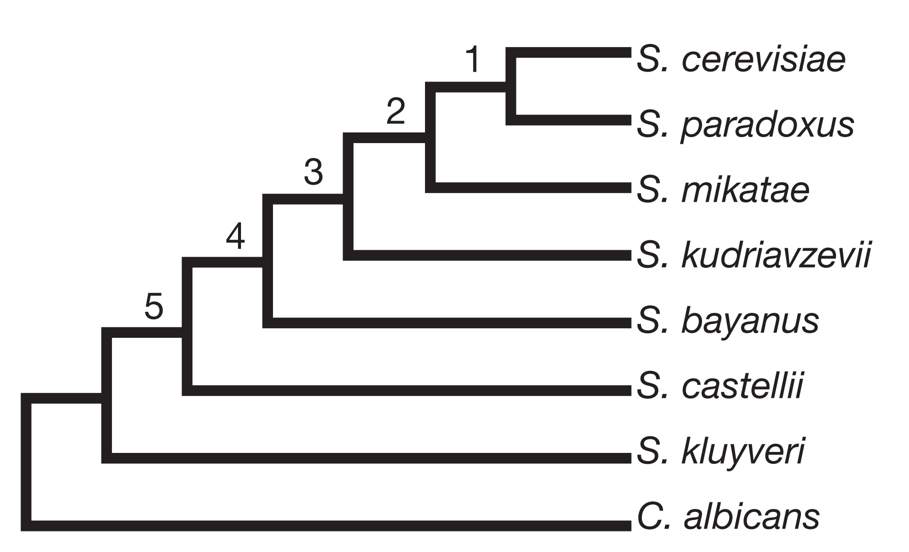
- We have implemented two methods for accounting for uncertainty in the estimated gene trees:
  - Using gene tree distributions: $L(N, \boldsymbol{\lambda}, \boldsymbol{\gamma} | \mathscr{G}) = \prod_{g \in \mathscr{G}} [\mathbf{P}_{N,\boldsymbol{\lambda},\boldsymbol{\gamma}}(G = g)]^{p_g}$
  - Using consensus trees: $$L(N, \boldsymbol{\lambda}, \boldsymbol{\gamma} | \mathscr{G}) = \prod_{g \in \mathscr{G}} \max_{g' \in b(g)} \{\mathbf{P}_{N,\boldsymbol{\lambda},\boldsymbol{\gamma}}(G = g')\}$$

THE YEAST DATA SET OF ROKAS ET AL. (NATURE 2003)

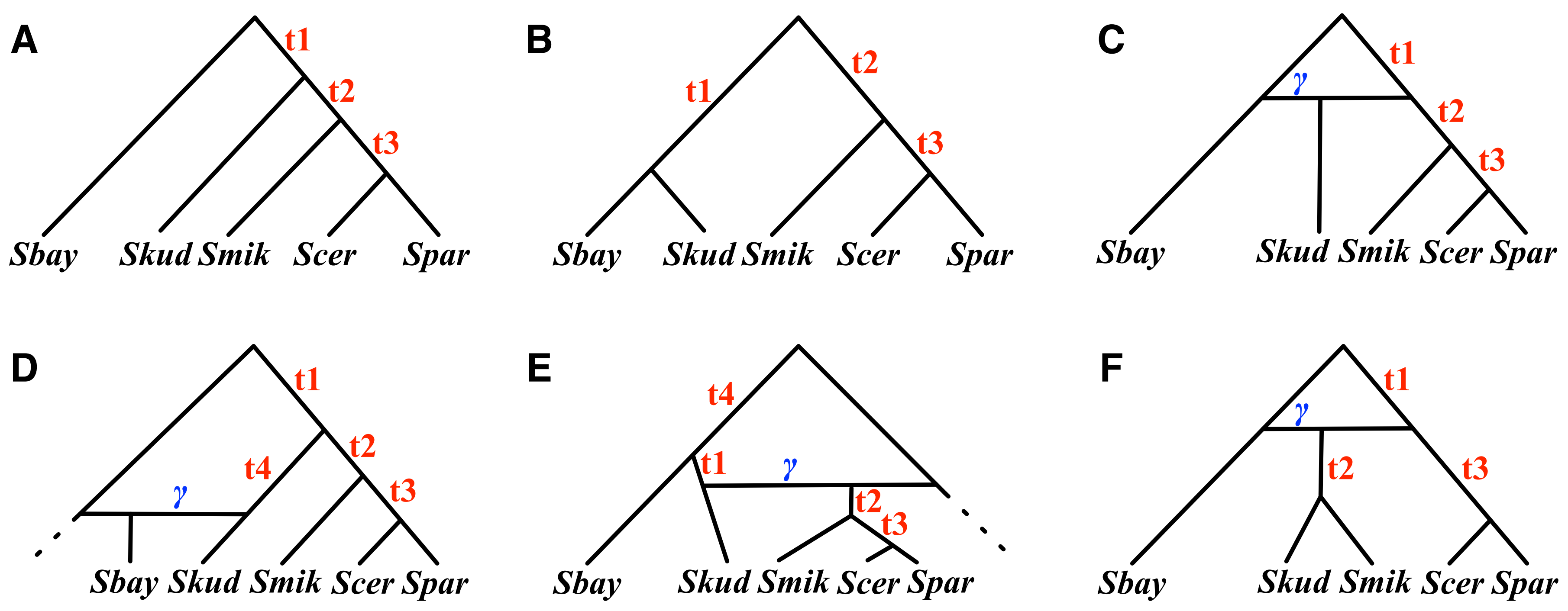
- □ The authors concatenated the sequences of 106 genes, and inferred a single species tree, which had 100% bootstrap support of all branches

- □ The method BEST inferred the same tree [Edwards et al., PNAS 2007]

- □ The MDC method inferred the same tree [Than&Nakhleh, PLoS Comp Bio 2009]



REANALYSIS OF THE YEAST DATA



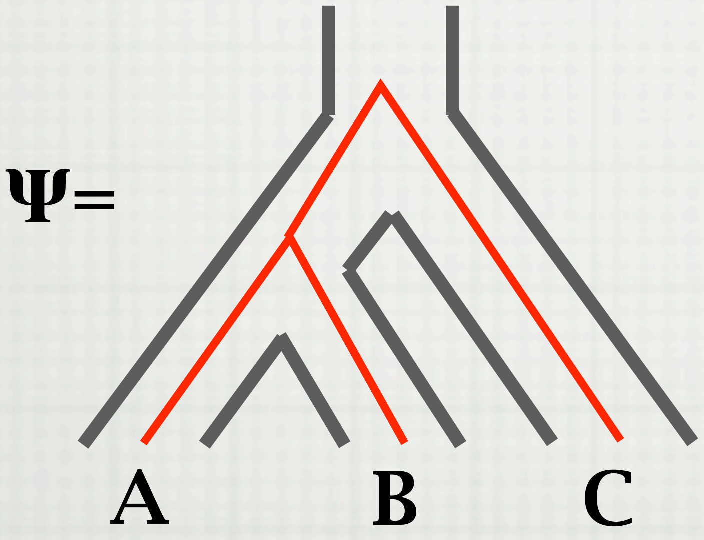
Species phylogeny	t_1	t_2	t_3	t_4	γ	$-lnL$	AIC	AICc	BIC
Fig. 3(**A**)	0.3	1.25	3.6	N/A	N/A	205	416	417	424
Fig. 3(**B**)	0.2	1.35	3.6	N/A	N/A	208	423	423	431
Fig. 3(**C**)	1.1	1.05	3.6	N/A	0.34	188	384	385	395
Fig. 3(**D**)	3.45	1.15	3.6	3.05	0.34	157	325	326	338
Fig. 3(**E**)	0.3	1.25	3.6	N/A	1.0	205	420	421	434
Fig. 3(**F**)	1.55	0.05	3.7	N/A	0.18	252	512	512	523

A PROBABILISTIC APPROACH

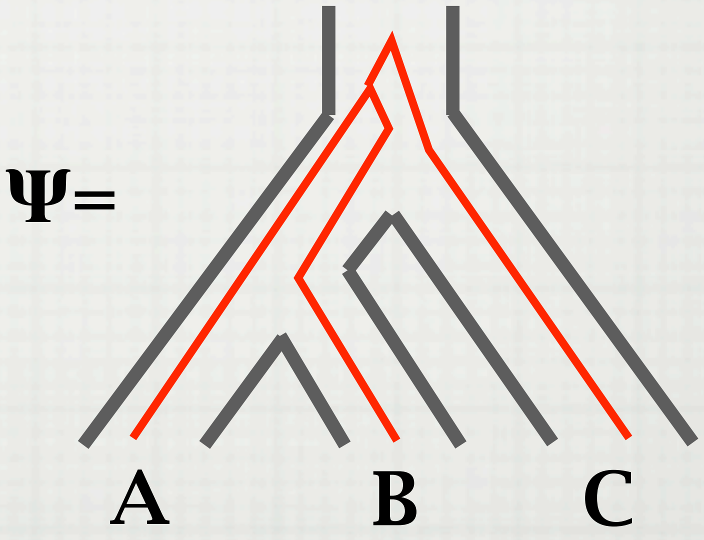
- The method produced very accurate results on synthetic data
- In addition, we currently have:
  - a faster method for computing gene tree probabilities, and
  - a method for inferring phylogenetic networks under the probabilistic method.

A PARSIMONY APPROACH

□ W. Maddison (Systematic Biology, 1997) proposed reconciling a gene tree with a species tree so as to minimize the "number of extra lineages" or "deep coalescences" (MDC).



0 extra lineages



1 extra lineage

A PARSIMONY APPROACH

□ Denote by XL(Ψ,gt,h) the number of extra lineages assuming coalescent history h gave rise to gene tree gt within the branches of species tree Ψ.

□ Then, W. Maddison's MDC cost for a given pair of species/gene tree is:

$$XL(\Psi, gt) = \min_{h \in H_\Psi(gt)} XL(\Psi, gt, h)$$

A PARSIMONY APPROACH

- The reconciliation problem under MDC is easy:
    - Map every clade in the gene tree to its MRCA in the species tree (the lca mapping)

A PARSIMONY APPROACH

□ The inference problem is hard

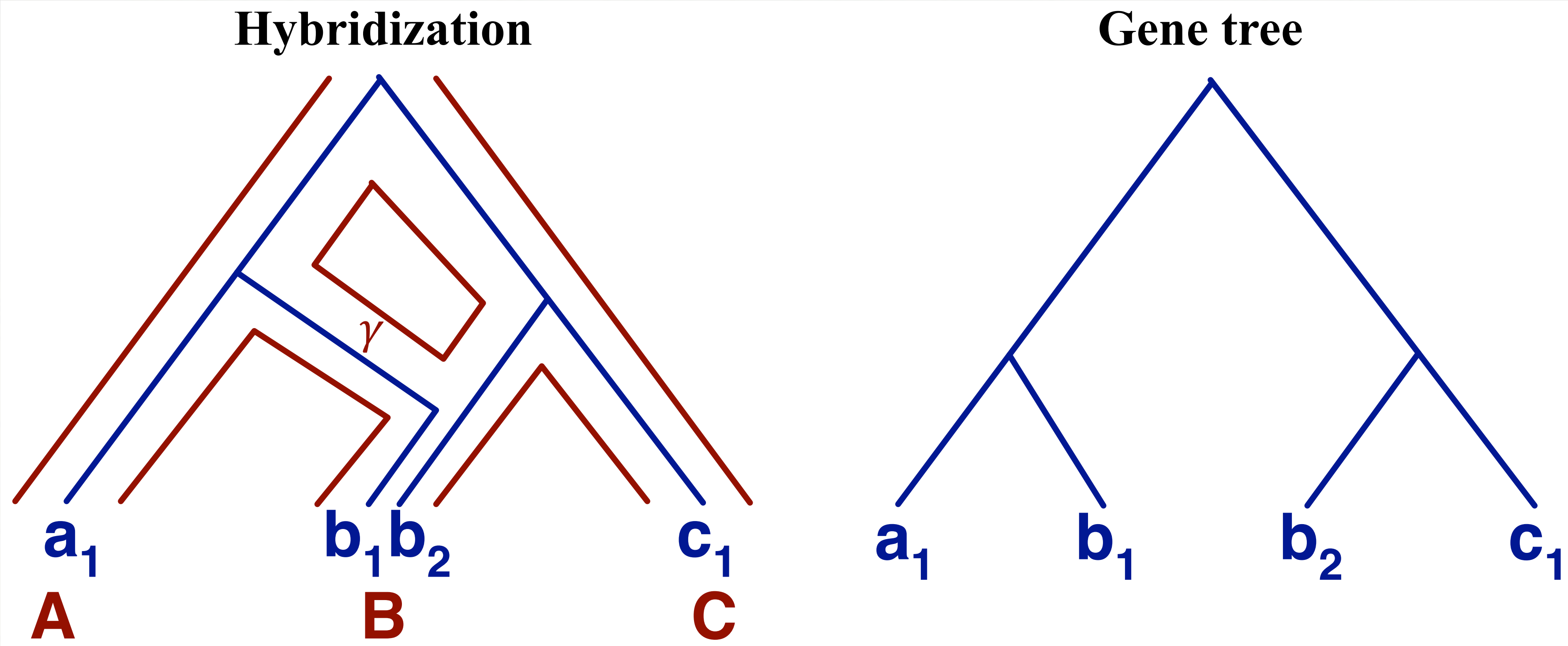
$$\Psi^* \leftarrow \mathrm{argmin}_{\Psi} \sum_{gt \in \mathcal{G}} XL(\Psi, gt)$$

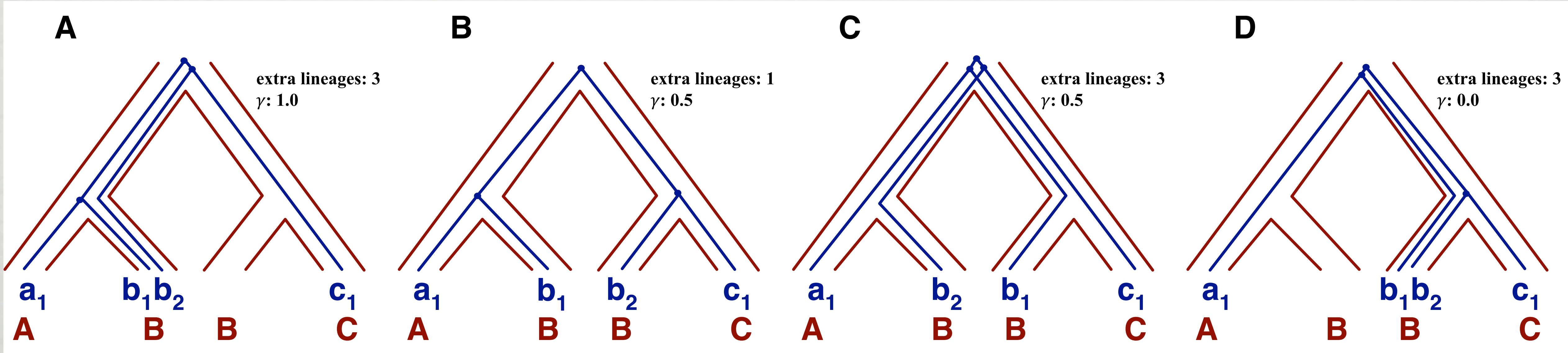
A PARSIMONY APPROACH

- □ Exact DP- and ILP-based solutions for inferring species trees:
  - □ When all gene trees are rooted, binary, with single allele per locus per species (Than&Nakhleh, PLoS Comp Bio 2009)
  - □ When the gene trees may be unrooted, non-binary, and zero or more alleles sampled per locus per species (Yu, Warnow, and Nakhleh, RECOMB 11 and JCB 11)

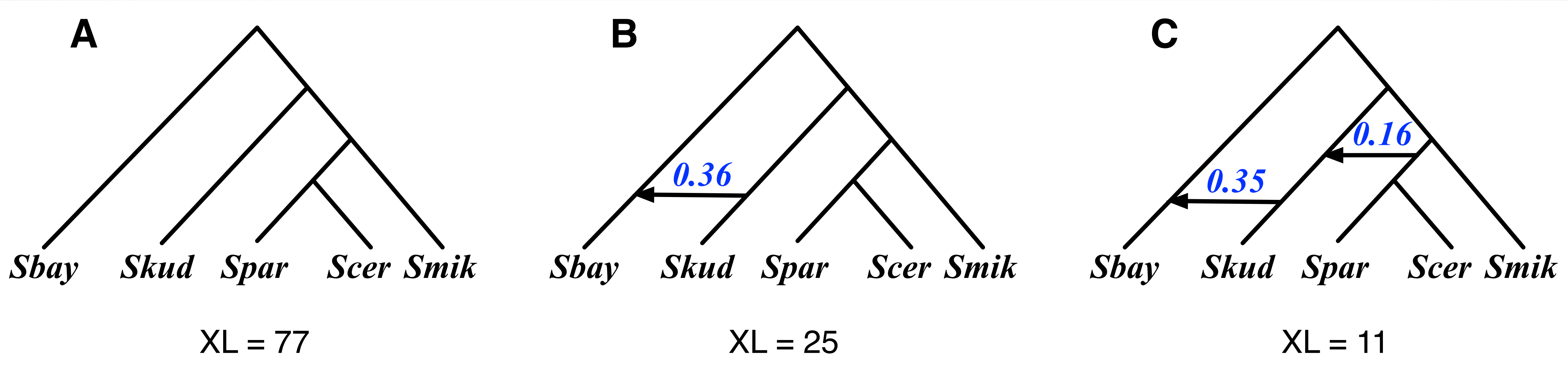
ILS + HYBRIDIZATION: A PARSIMONY APPROACH

- But, what about inference of species networks?
  - Solution for special cases (Yu, Than, Degnan, Nakhleh, Syst Biol 2011)
  - Solution for the general case (Yu, Barnett, Nakhleh, under review, 2012)

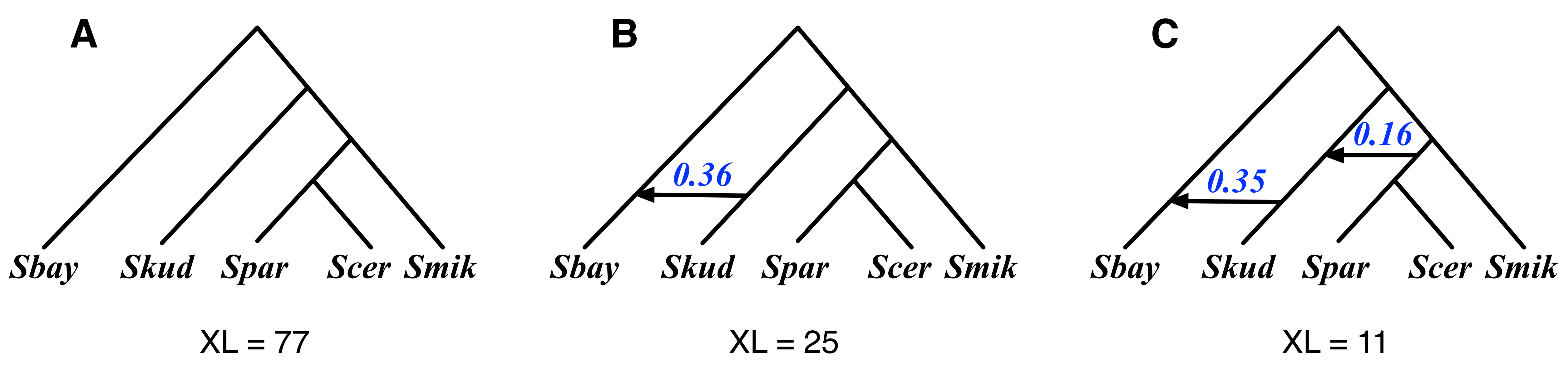

Hybridization
Gene tree
γ
a_1
$b_1 b_2$
c_1
A
B
C
a_1
b_1
b_2
c_1


A
extra lineages: 3
γ: 1.0
a_1 $b_1 b_2$ c_1
A B B C
B
extra lineages: 1
γ: 0.5
a_1 b_1 b_2 c_1
A B B C
C
extra lineages: 3
γ: 0.5
a_1 b_2 b_1 c_1
A B B C
D
extra lineages: 3
γ: 0.0
a_1 $b_1 b_2$ c_1
A B B C

REANALYSIS OF THE YEAST DATA

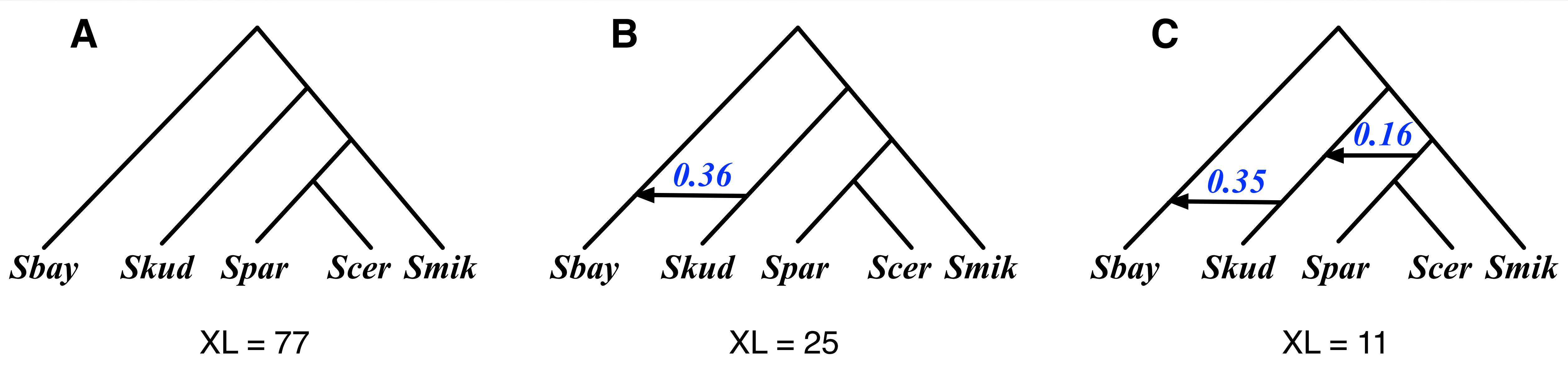


REANALYSIS OF THE YEAST DATA



Observe the decrease in XL as more reticulations are added!

REANALYSIS OF THE YEAST DATA



Observe the decrease in XL as more reticulations are added!

Have to account for network complexity!

ILS + HYBRIDIZATION: A PARSIMONY APPROACH

□ The parsimony approach does surprisingly well at (1) inferring the phylogenetic network topology, and (2) estimating inheritance probabilities, on synthetic data

□ Much faster than the probabilistic method

□ Suffers from the "model selection" problem (the more hybridization, the merrier!)

SUMMARY

□ Dealing with ILS and hybridization simultaneously, we have methods for

  □ computing gene tree probabilities

  □ inferring phylogenetic networks

  □ parsimonious reconciliation of gene trees

  □ parsimonious inference of phylogenetic networks

□ The most challenging task:

  □ how to achieve scalability of these methods to large data sets!

PHYLONET

- All the Methods are implemented in PhyloNet:
  - http://bioinfo.cs.rice.edu/phylonet
- Tutorial tomorrow, by Yun Yu

ACKNOWLEDGMENTS

- People:
  - **R.M. Barnett** (Rice), **J.H. Degnan** (Canterbury), **Jianrong Dong** (Rice), **Kevin Liu** (Rice), **Y. Yu** (Rice)
- Funding:
  - NSF, NIH, The Sloan Foundation, The Guggenheim Foundation

THANK YOU

HTTP://WWW.CS.RICE.EDU/~NAKHLEH