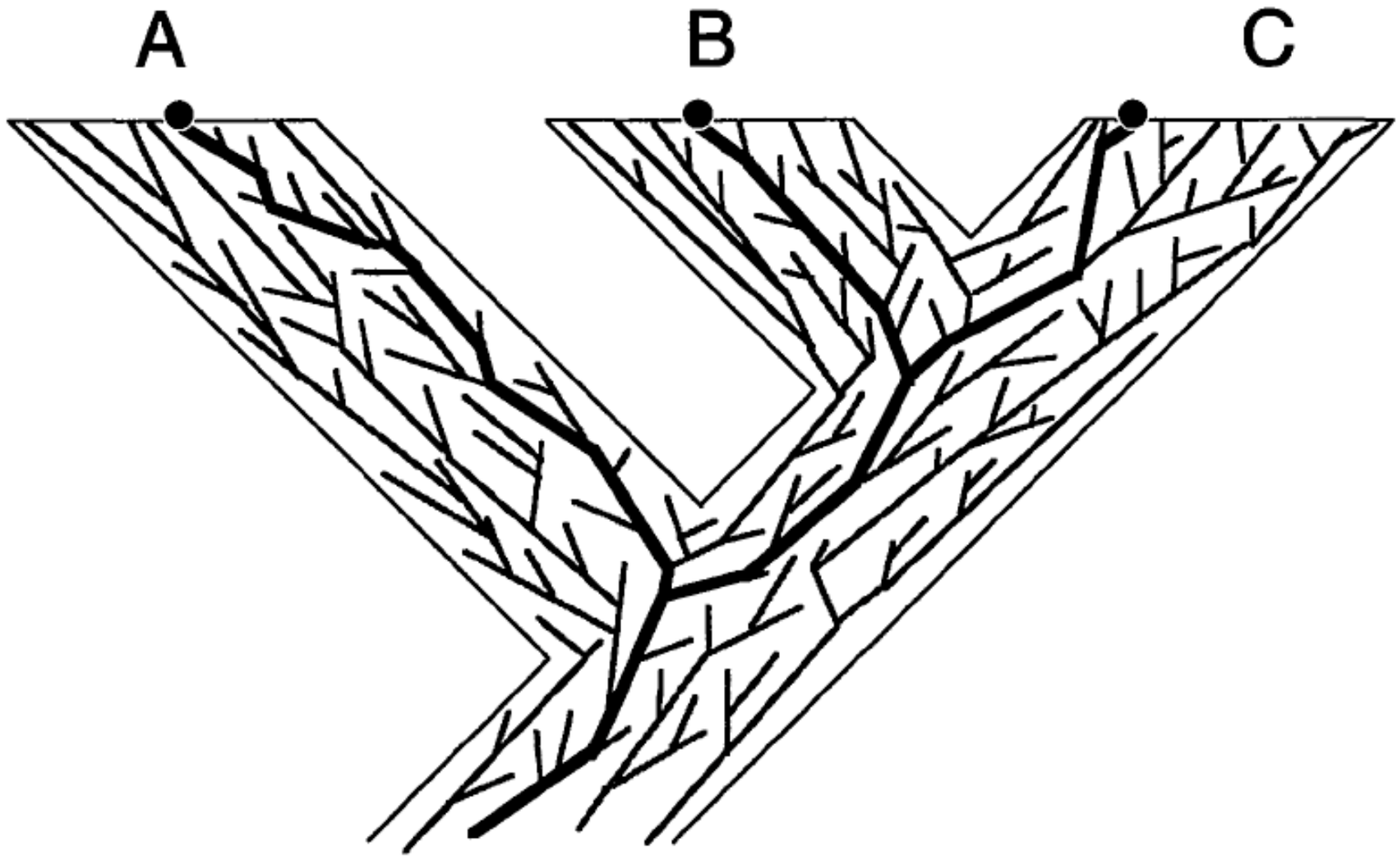
Integrating Sequence and Topology for Efficient and Accurate Detection of Horizontal Gene Transfer

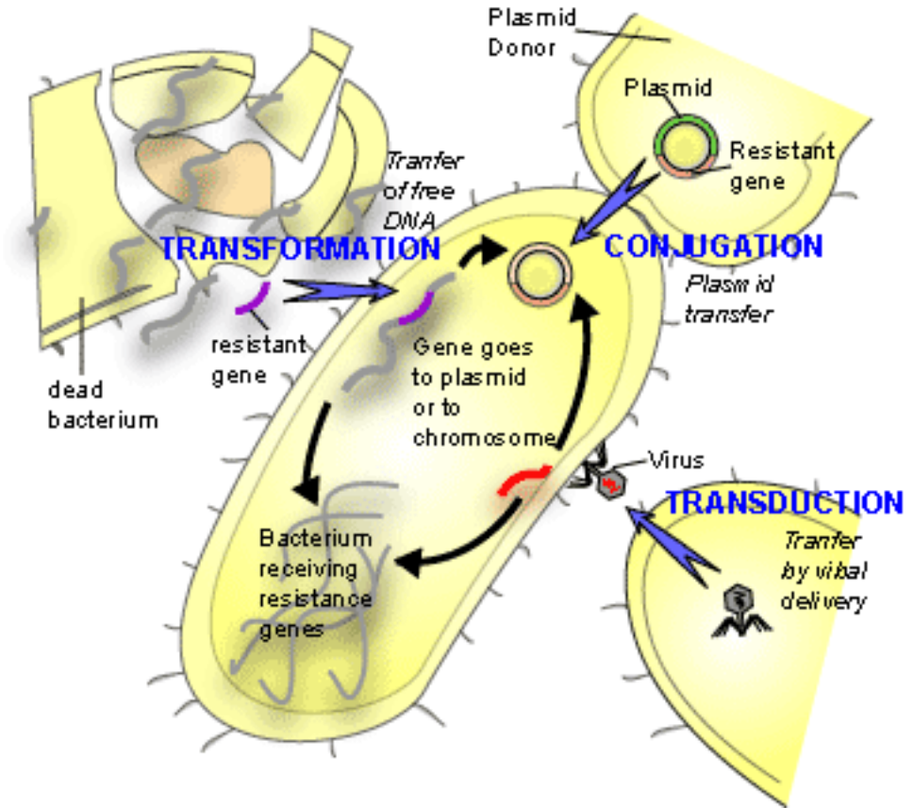
Cuong Than, Guohua Jin, and Luay Nakhleh

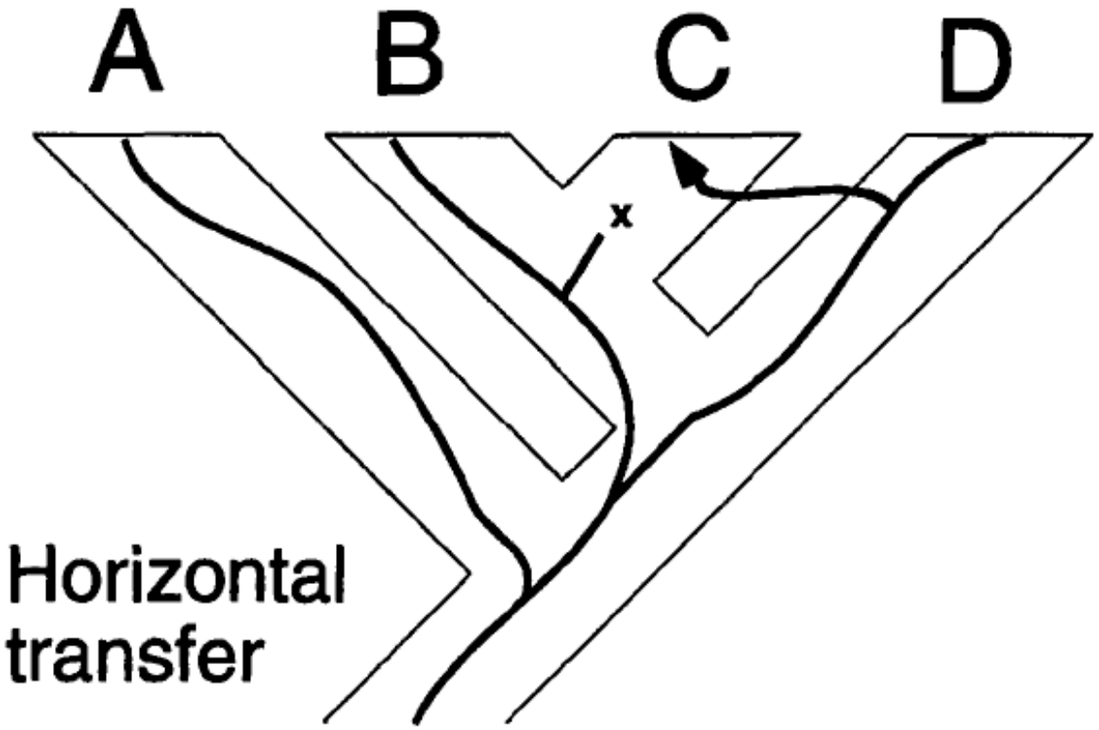
Department of Computer Science
Rice University

Gene Trees Within the Branches of a Species Tree

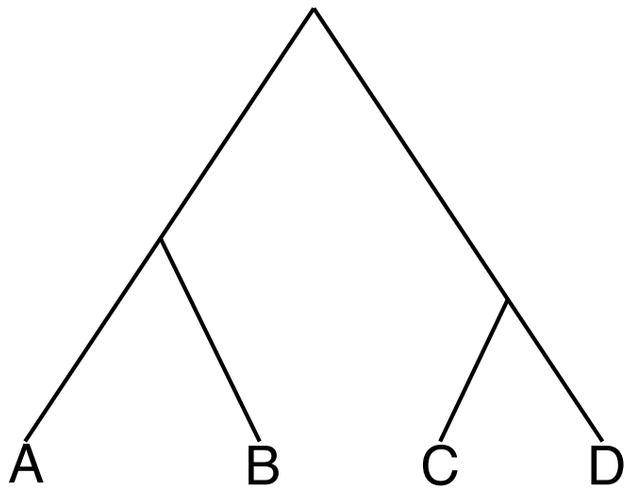


Horizontal Gene Transfer and Tree Incongruence

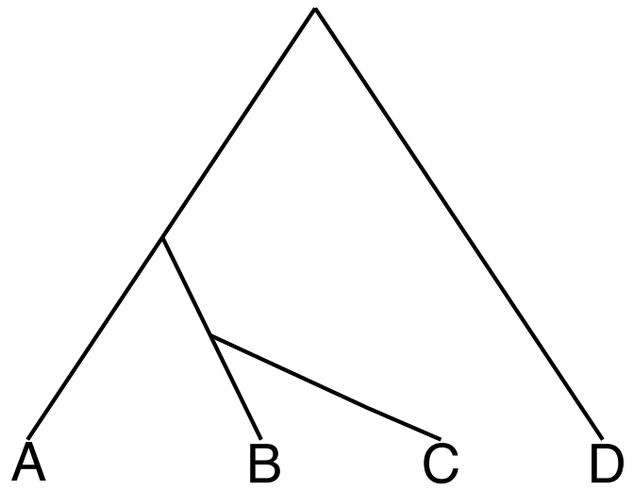




Horizontal Gene Transfer Detection

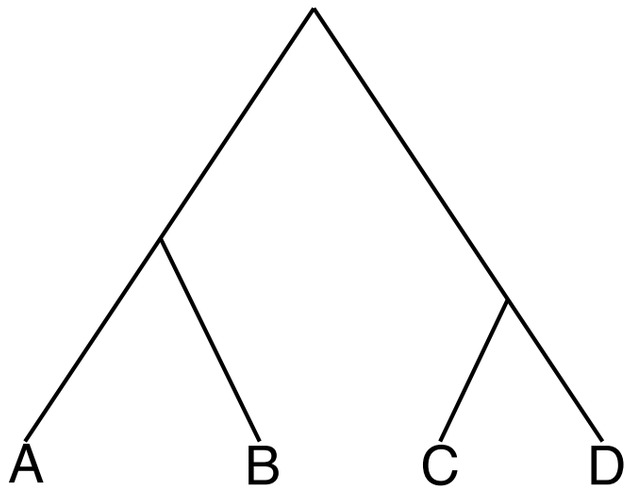


species tree

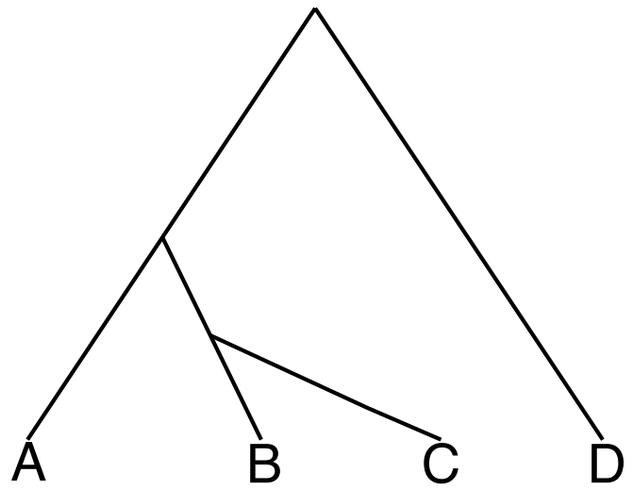


gene tree

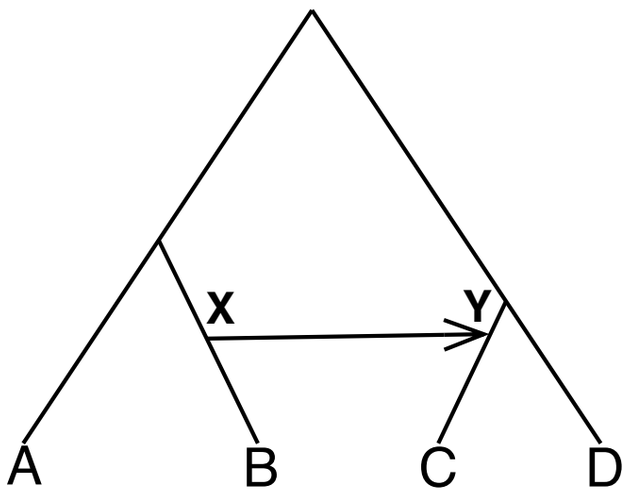
Horizontal Gene Transfer Detection



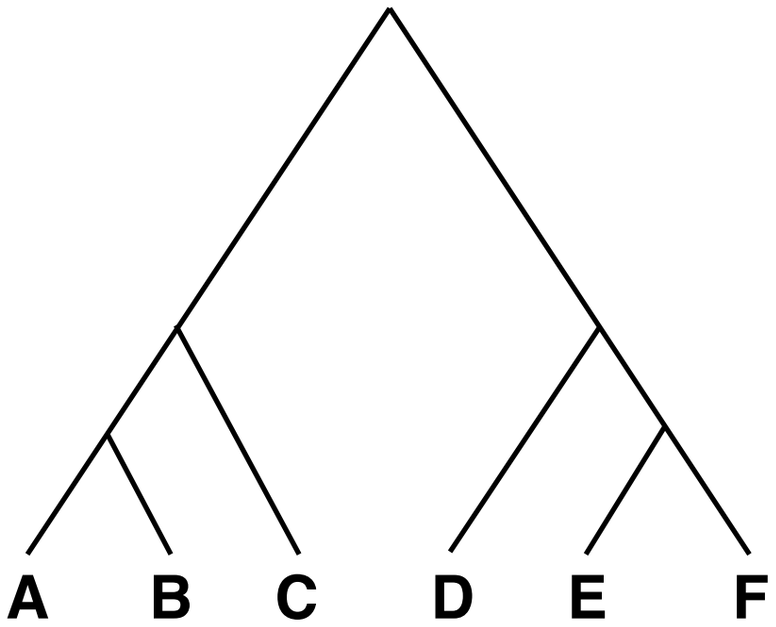
species tree



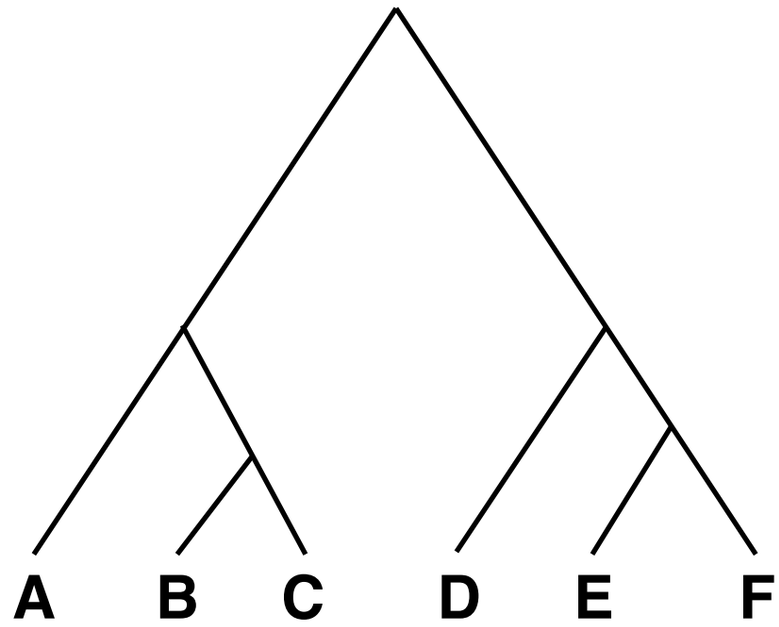
gene tree



Multiple Solutions for HGT Detection

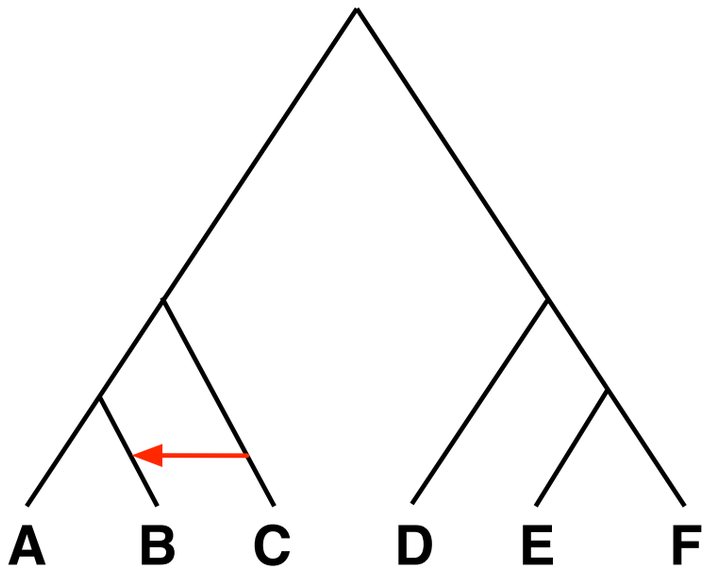


species tree

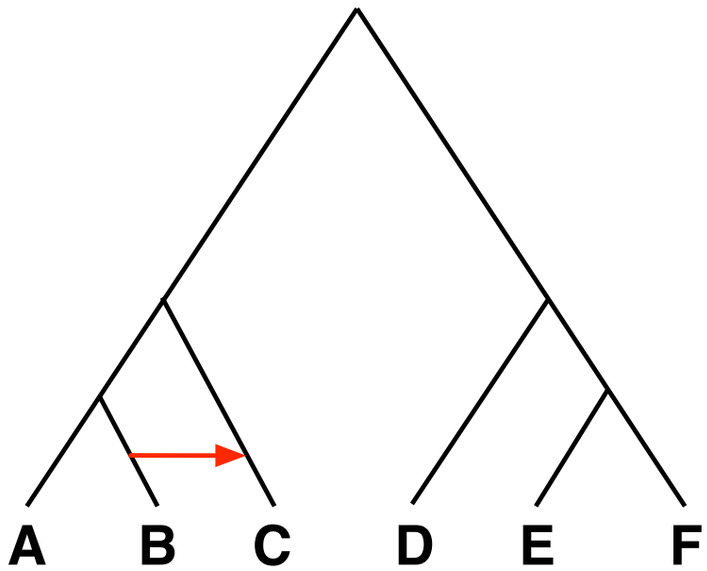


gene tree

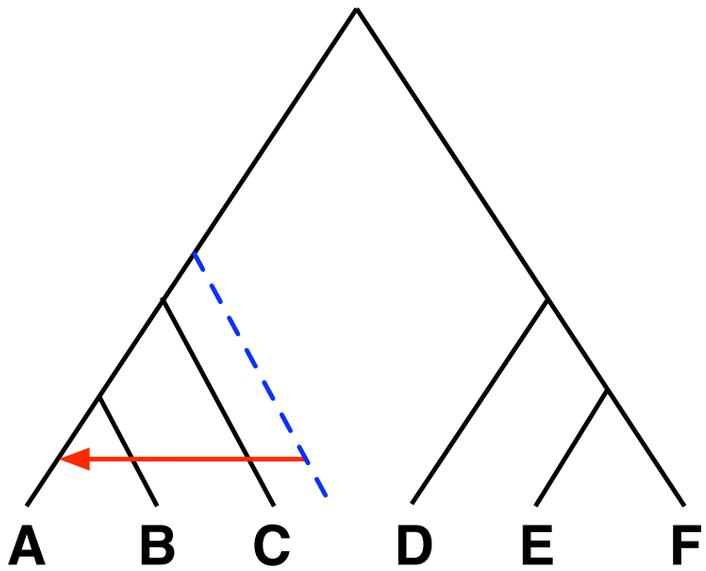
Multiple Solutions for HGT Detection



scenario 1



scenario 2



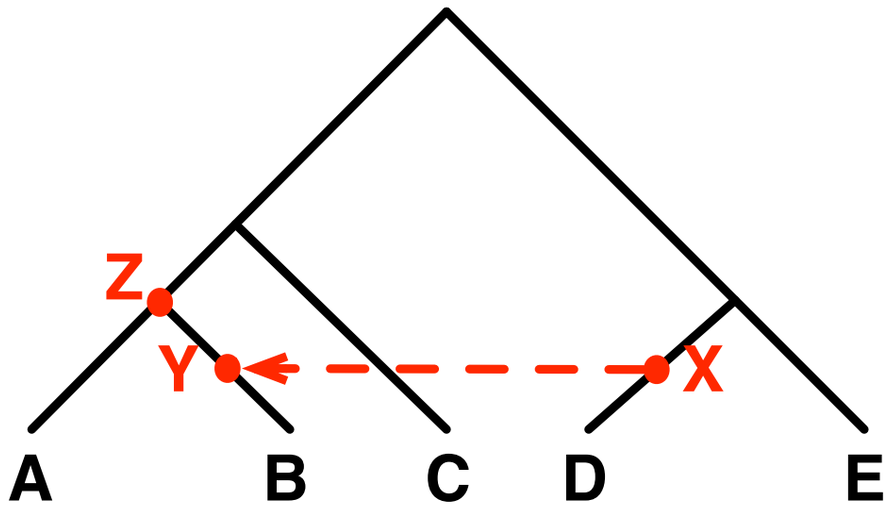
scenario 3

How can we improve the accuracy of HGT detection?

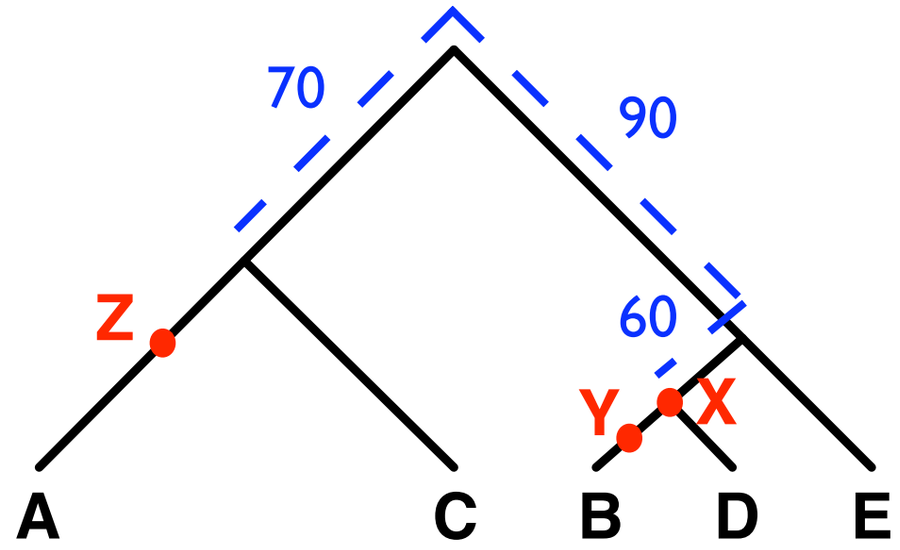
Our Method: Assessing the Support of HGT Edges

- Assign a support value to each HGT edge
- Branches of the gene tree often have bootstrap value
- Use those bootstrap values to evaluate HGT edge support

HGT Edge Support: Example



species tree + HGT edge



gene tree

Support for HGT edge (x, y): max{60, 90, 70} = 90

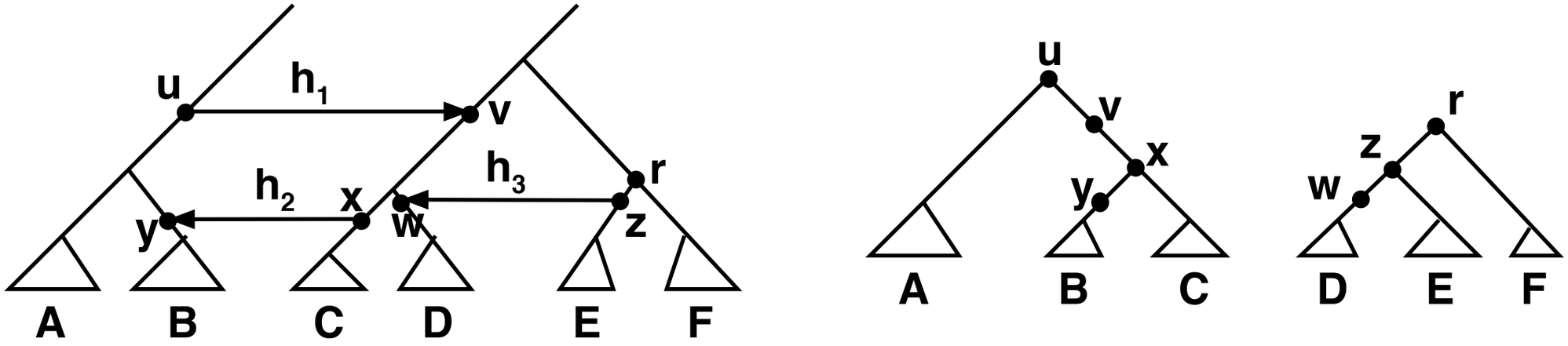
HGT Edge Support: Algorithm

- Create a network N from the species tree by adding HGT edges
- From N, create two trees:
  - ST': Keep HGT edges, while removing the other edges
  - ST": Remove HGT edges, while keeping the other edges

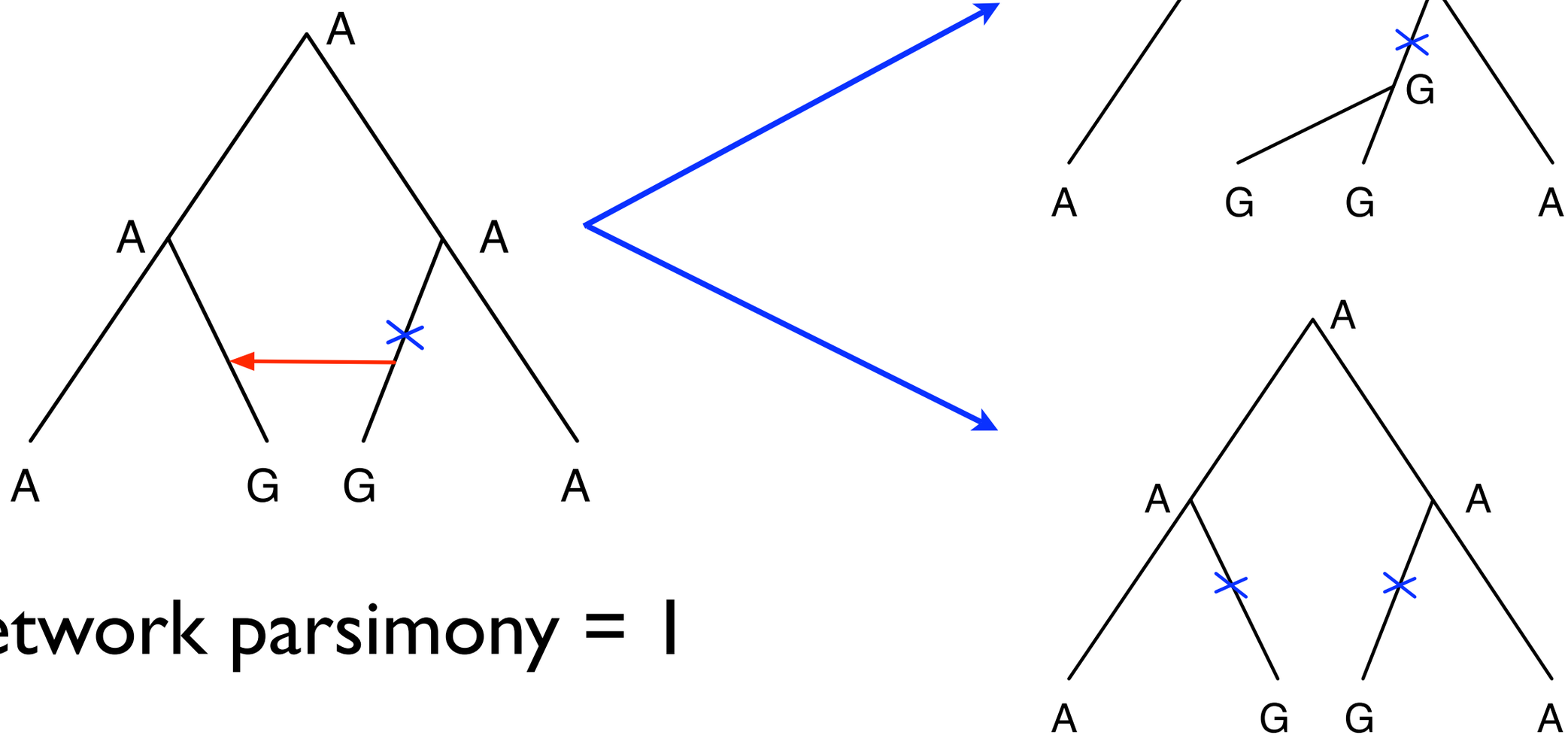
HGT Edge Support: Algorithm

- For HGT edge X → Y, find the moving clade P and its sister clade Q
- Finding P: let P = $L_{ST'}(Y)$
- Finding Q:
  - Let Y' = Y
  - Let p = parent of Y' in ST"
    - If $L_{ST'}(p) \neq \emptyset$, then Q = $L_{ST'}(p)$
    - If not, let Y' = p, and repeat
- Support = max of bootstrap values of edges from P to Q

HGT Edge Support: Illustration



Network Parsimony



Network parsimony = 1

Parsimony-based HGT Detection

- Input: Sequences for a group of species
- Output: A network (tree + HGT edges)
- Optimization criterion: Smallest network parsimony score

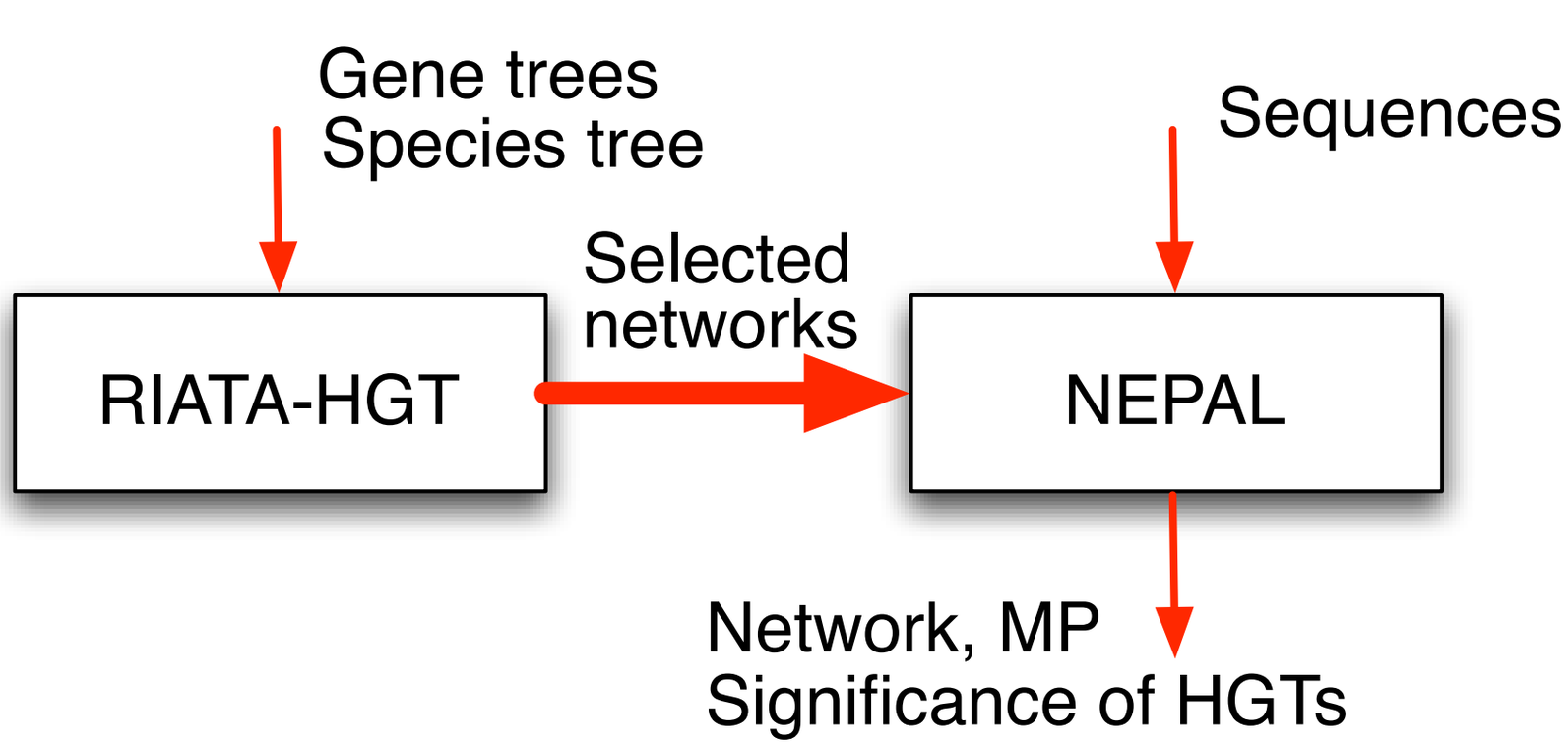
Parsimony-based HGT Detection

- Accurate, as shown in both simulated and biological data
- Slow, because it has to examine all possible networks

Topology-based HGT Detection

- Fast in computing HGT edges
- Return many false positives

Our Method: Combining Topology and Parsimony-based Methods



TopSeq Algorithm

TopSeqHGTIdent(ST, GT, S)
INPUT: species tree ST, gene tree GT, and sequence dataset S
OUTPUT: network N with marked significance of each HGT

1 Let $\{N_1, \ldots, N_m\}$ be the set of all phylogenetic networks computed by RIATA-HGT, and let $H(N_i)$ be the set of HGT edges in N_i.
2 Let $\mathcal{H} = \cap_{i=1}^{m} H(N_i)$, and $R(N_i) = H(N_i) - \mathcal{H}$. In other words, $\mathcal{H}$ denotes the set of HGT edges that are shared by all networks, and $R(N_i)$, for $1 \leq i \leq m$, the set of HGT edges that are in N_i but not shared by all other networks.
3 Apply NEPAL to $N' = ST + \mathcal{H}$.
4 For each network N_i, $1 \leq i \leq m$, apply NEPAL by incrementally adding (in no particular order) the HGT edges in $R(N_i)$ to N', and compute the minimum parsimony length of the phylogenetic network.
5 Let $N = ST$, N_{opt} be the best network according to maximum parsimony criterion, that is $NCost(N_{opt}, S) = min_{i=1}^{m}(NCost(N_i, S))$.
Apply NEPAL by adding to ST each time one of the HGT events $h \in H(N_{opt})$ that results in the most significant drop in the parsimony score and let $N = N \cup h$. Stop this process when the drop is smaller than a specified threshold.

Experimental Data

- The biological data by Bergthorsson
  - HGT transfer to Amborella
  - 20 genes
  - Donors: Bryophytes, Moss, Eudicots, and Angiosperms

Experimental Results

	Bergthorsson *et al.*			MP			RIATA-HGT			RIATA-HGT+MP		
Gene	#HGTs	donor	SH	#HGTs	F?	#Nets	#HGTs	#Nets	#events	#HGTs	F?	#Nets
cox2	3	M	<0.001	1	Y	8482	9	4	12	1	Y	23
		E	NS		N	8482	9	4	12		N	23
		E	NS		N	8482	9	4	12		N	23
nad2	2	M	<0.001	1	Y	3500	7	6	11	1	Y	21
		E	NS		N	3500	7	6	11		N	21
nad4(exons)	1	M	<0.001	1	Y	1620	4	2	5	1	Y	9
nad4(ex4)	1	E	NS	2	Y	1832	6	3	8	2	Y	21
nad5	2	M	<0.001	1	Y	3292	6	6	9	1	Y	17
		A	0.025		N	3292	6	6	9		N	17
nad6	1	B	<0.001	1	Y	2484	6	3	8	1	N	15
nad7	2	M	<0.001	1	Y	2948	7	1	7	1	Y	13
		E	NS		N	2948	7	1	7		N	13
atp1	1	E	0.001	1	Y	2817	6	18	14	1	Y	27
atp8	1	E	0.008	2	Y	9059	5	6	11	1	Y	21
ccmB	1	E	NS	2	Y	66015	6	3	14	2	Y	101
ccmC	1	E	0.03	1	Y	2786	7	21	15	1	Y	29
ccmFN1	1	E	0.004	2	Y	4412	7	18	13	2	Y	46
cox3	1	A	NS	1	N	3466	8	15	18	1	N	52
nad1	1	E	<0.001	1	Y	2812	9	12	14	1	Y	27
rpl16	1	E	NS	3	Y	21632	10	27	23	1	Y	67
rps19	1	E	0.003	1	Y	1476	5	4	7	1	Y	13
sdh4	1	E	NS	3	Y	18670	9	18	18	3	Y	540
nad2intron	1	M	–	2	Y	5904	8	2	10	2	Y	44
nad5intron	1	M	–	2	Y	10280	9	5	18	2	Y	51
nad7intron	1	M	–	1	Y	3284	12	48	26	1	Y	51

Experimental Results

- NEPAL: 12 out of 13 HGTs were found
- RIATA-HGT: 12 out of 13 HGTs were found (the missing HGT is different)
- The support for 12 HGT edges were high, over 95%

Conclusions

- A method for assessing the support of HGT edges
- Integration of sequence-based and topology-based methods gives promising performance

Thank You!

http://bioinfo.cs.rice.edu