## Perfect phylogeny: Gusfield's algorithm

Find and work with a partner

1. Data from "The Perfect Phylogeny Problem" by David Fernández-Baca

species	1	2	3	4	5	6
lamprey	0	0	0	0	0	1
shark	1	1	0	1	0	0
salmon	1	1	1	1	0	0
lizard	1	1	1	0	1	0

Step 2: write mutation numbers for each row

 Step 1: sort columns high to low

 species

 lamprey

 shark

 salmon

 lizard

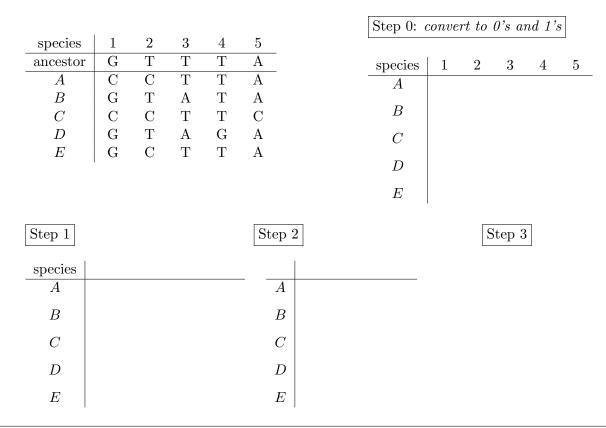
 Step 3: create tree from root to leaves

lamprey

shark salmon

lizard

2. Data from "Algorithms on Strings, Trees, and Sequences" by Dan Gusfield



3. Let  $O_i$  be the set of samples that have mutation *i*. For example, in question (2) on the previous page,  $O_2 = \{A, C, E\}$ . What must be true about  $O_i$  and  $O_j$  (for all pairs of mutations *i* and *j*) for a perfect phylogeny to be guaranteed? Hint: try to relate  $O_i$  with its corresponding mutation *i* on the final tree.

4. All possibilities for two sites ("Four gamete test"). Does a perfect phylogeny exist?

sample	1	2
A	0	0
B	0	1
C	1	0
D	1	1

\_

5. Human data from Michael F. Hammer, Nature (1995). Does a perfect phylogeny exist?

sample	1	2	3	4
A	0	0	0	0
B	0	1	0	0
C	1	0	0	0
D	1	0	1	1
E	1	0	0	1

## EXTRA PRACTICE

- 6. What sorting algorithm could we use to sort the columns, considering that we could have an arbitrary number of species?
- 7. What is the runtime of Gusfield's algorithm in terms of the number of species n and the number of mutations m?