

# Comparing Evolutionary Trees

Katherine St. John

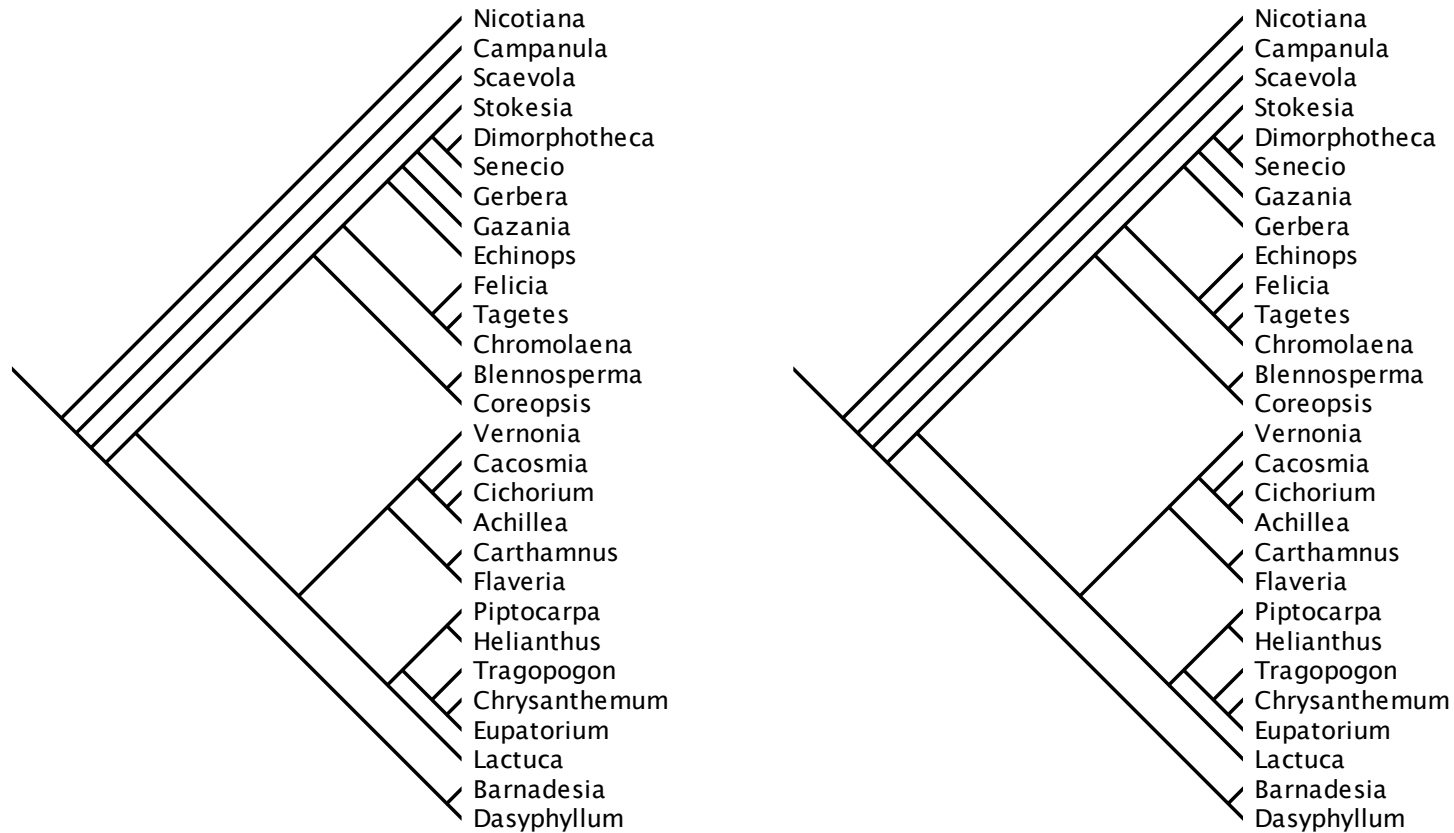
Lehman College and the Graduate Center

City University of New York

stjohn@lehman.cuny.edu

(Joint work with Maria Luisa Bonet)

# When are two trees similar?



# Distances Between Trees

- When are two trees similar?

# Distances Between Trees

- When are two trees similar?
- Many ways to measure the distance between trees.  
For example:

# Distances Between Trees

- When are two trees similar?
- Many ways to measure the distance between trees.  
For example:
  - Count the number of branches that are different (false positives and false negatives).

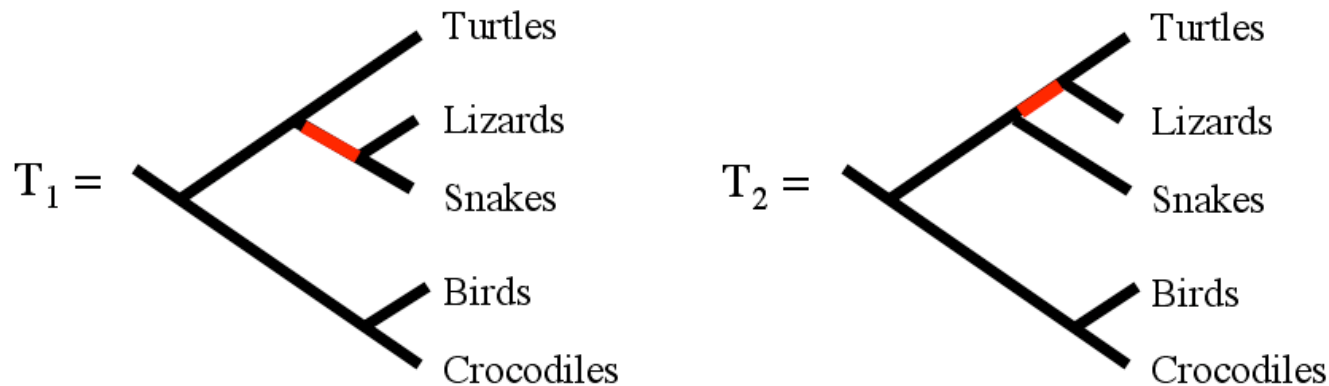
# Distances Between Trees

- When are two trees similar?
- Many ways to measure the distance between trees.  
For example:
  - Count the number of branches that are different (false positives and false negatives).
  - Give credit for getting the details (subtrees) correct.

# Distances Between Trees

- When are two trees similar?
- Many ways to measure the distance between trees.  
For example:
  - Count the number of branches that are different (false positives and false negatives).
  - Give credit for getting the details (subtrees) correct.
  - Give credit for getting the overall structure correct.

# Distances Between Trees

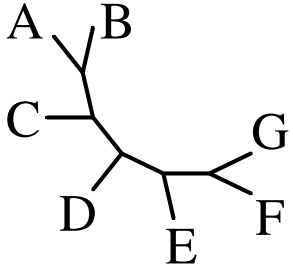


- Robinson-Foulds distance: # of branches that occur in only one tree.
- Calculate in  $O(n)$  time using Day's Algorithm (1985).
- Extends naturally to weighted trees.



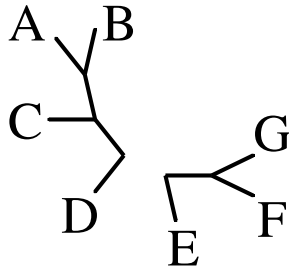
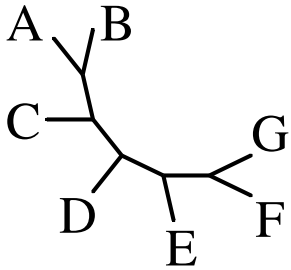
# TBR Distance

- Tree-Bisection-Reconnect (TBR) Move:



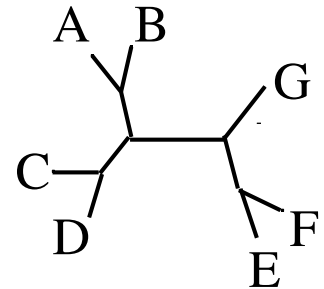
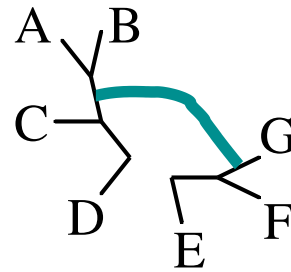
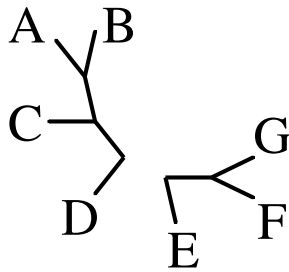
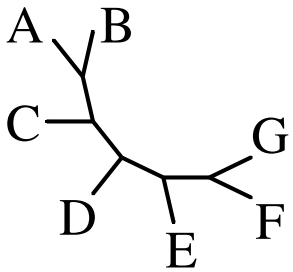
# TBR Distance

- Tree-Bisection-Reconnect (TBR) Move:



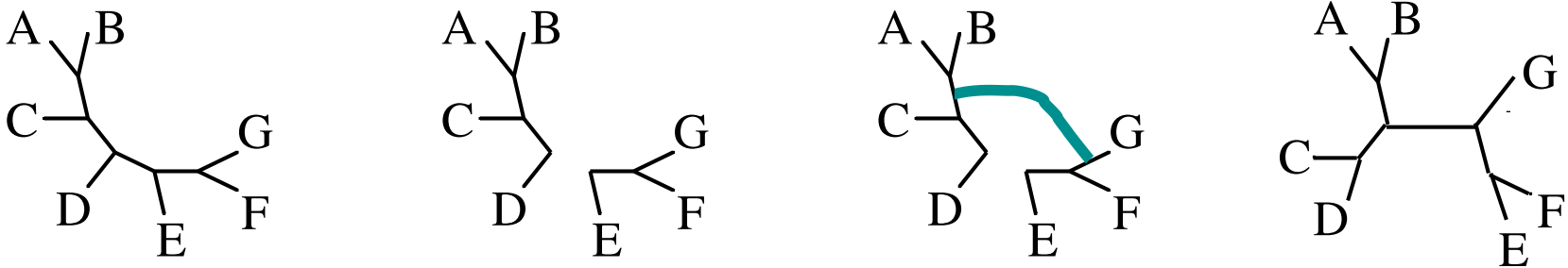
# TBR Distance

- Tree-Bisection-Reconnect (TBR) Move:



# TBR Distance

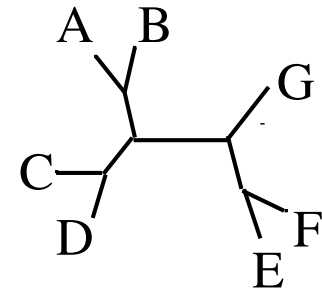
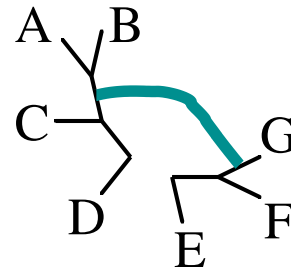
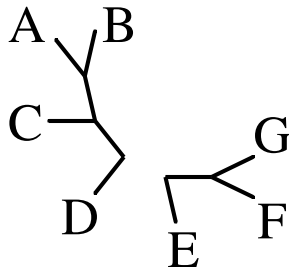
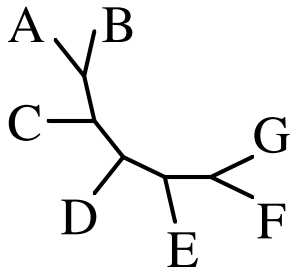
- Tree-Bisection-Reconnect (TBR) Move:



- The TBR distance between two trees is the minimal number of TBR moves needed to transform the first tree into the second tree.

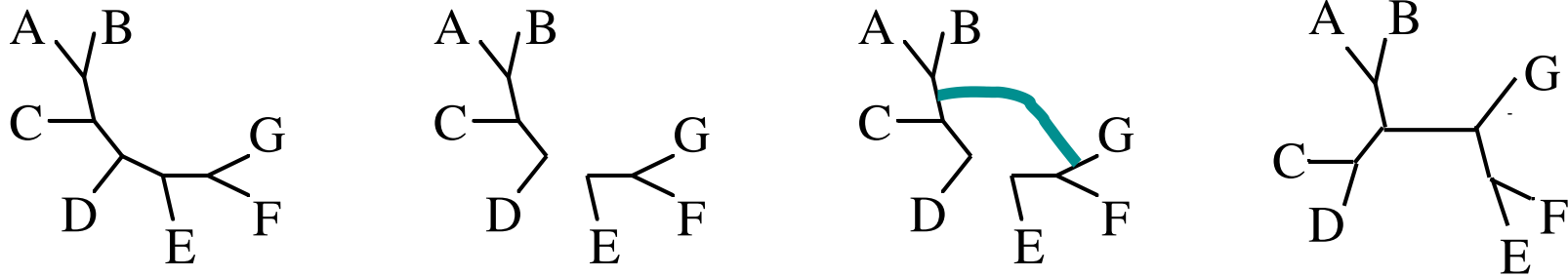
# TBR Distance

- Tree-Bisection-Reconnect (TBR):



# TBR Distance

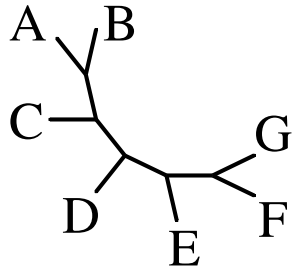
- Tree-Bisection-Reconnect (TBR):



- TBR is NP-hard. (Allen & Steel '01)
- Many attempts, but no approximation algorithm with guaranteed bounds.

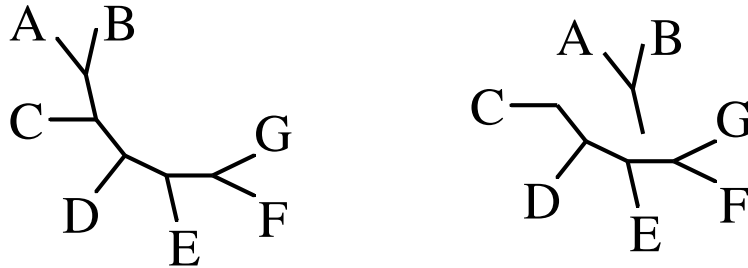
# SPR Distance

- Subtree-prune-regraft (SPR):



# SPR Distance

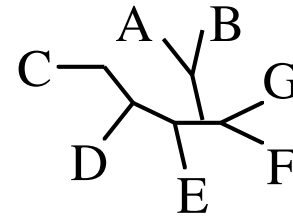
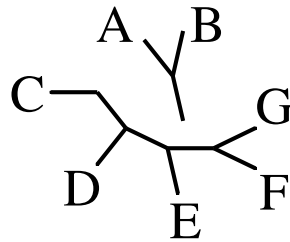
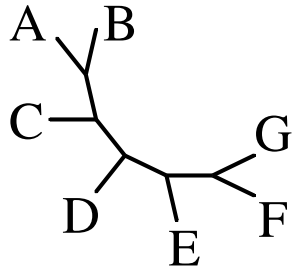
- Subtree-prune-regraft (SPR):





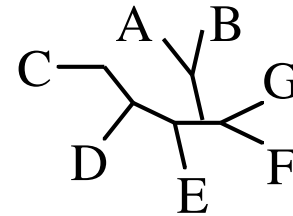
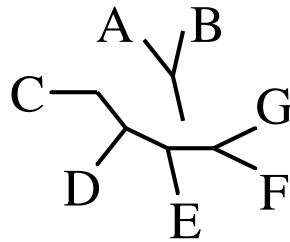
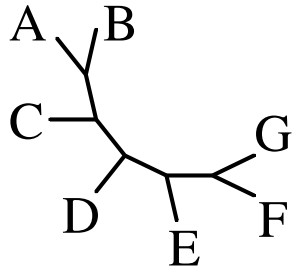
# SPR Distance

- Subtree-prune-regraft (SPR):



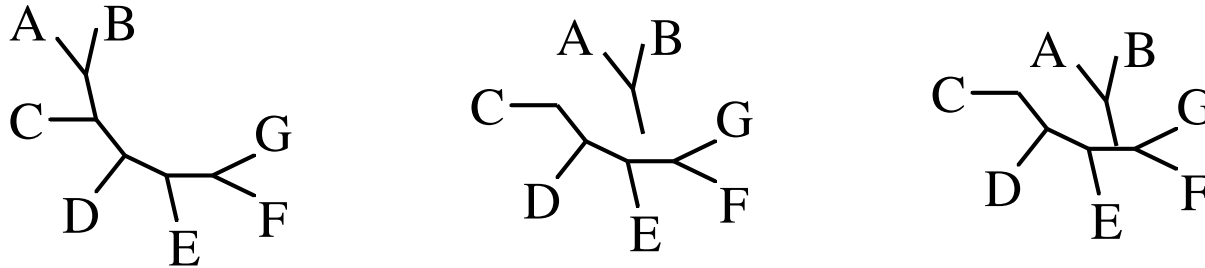
# SPR Distance

- Subtree-prune-regraft (SPR):



# SPR Distance

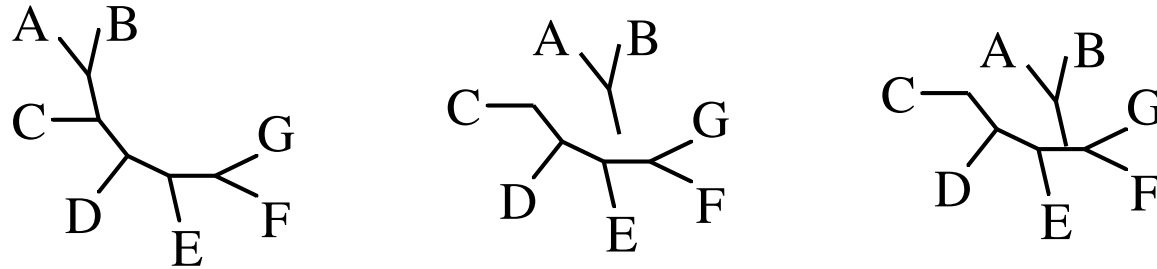
- Subtree-prune-regraft (SPR):



- The SPR distance between two trees is the minimal number of SPR moves needed to transform the first tree into the second tree.

# SPR Distance

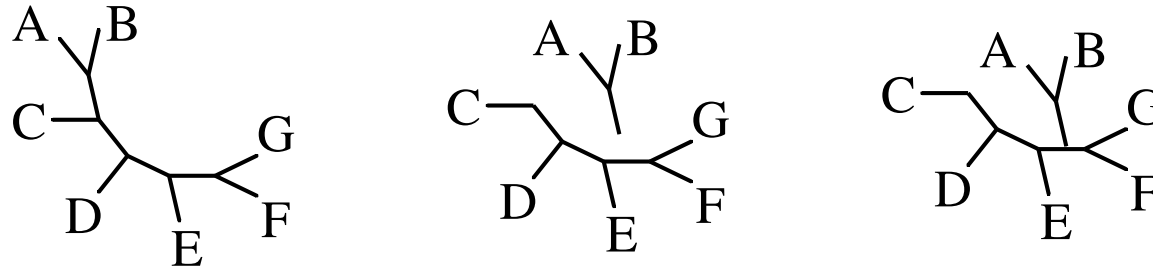
- Subtree-prune-regraft (SPR):



- SPR for rooted trees is NP-hard. (Bordewich & Semple '05)

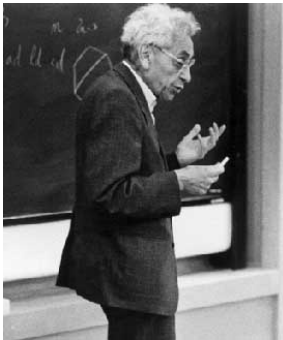
# SPR Distance

- Subtree-prune-regraft (SPR):



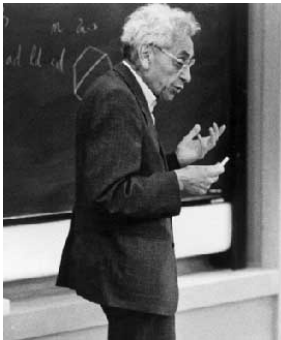
- SPR for rooted trees is NP-hard. (Bordewich & Semple '05)
- Approximation algorithm for SPR on rooted trees.  
(Bonnet, St. John, Amenta, & Mahindru '05)

# \$100 Problem



Paul Erdős gave numerous prizes for favorite problems in combinatorics, graph theory, and number theory.

# \$100 Problem



Paul Erdos gave numerous prizes for favorite problems in combinatorics, graph theory, and number theory.

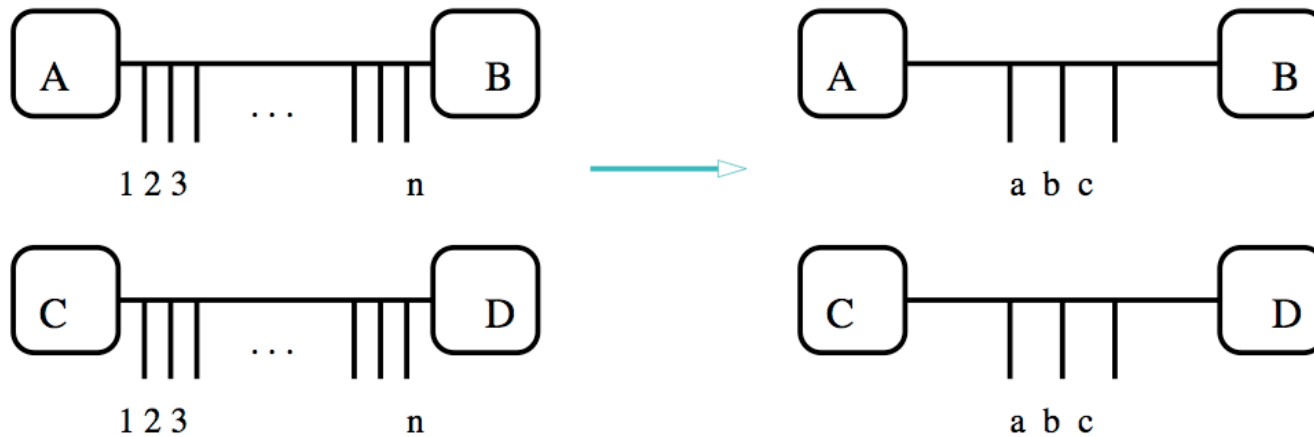


Following that tradition, Mike Steel has offered prizes for his favorite problems in mathematical phylogenetics.

# \$100 Problem



Does shrinking common subchains  
preserve SPR distance?  
(Implies fixed parameter tractability)





# \$100 Problem



Does shrinking common subchains  
preserve SPR distance?  
(Implies fixed parameter tractability)

- Allen & Steel '01: Showed true for TBR.

# \$100 Problem



Does shrinking common subchains  
preserve SPR distance?  
(Implies fixed parameter tractability)

- Allen & Steel '01: Showed true for TBR.
- Bordewich & Semple '05: Showed true for rooted trees.

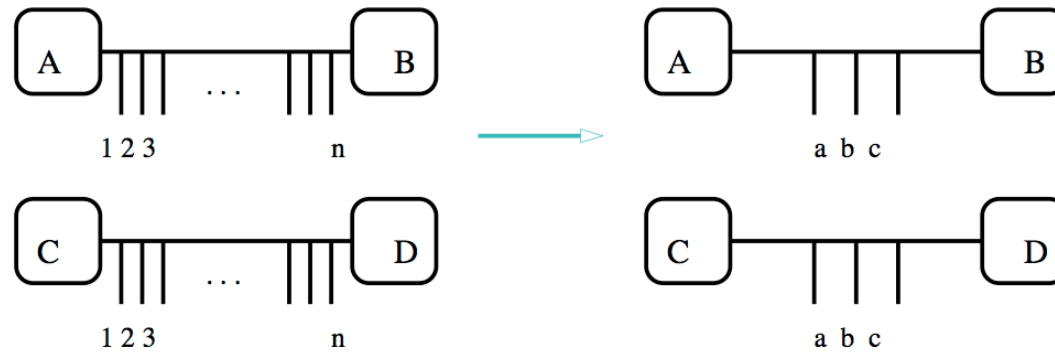
# \$100 Problem



Does shrinking common subchains  
preserve SPR distance?  
(Implies fixed parameter tractability)

- Allen & Steel '01: Showed true for TBR.
- Bordewich & Semple '05: Showed true for rooted trees.
- What about unrooted trees?

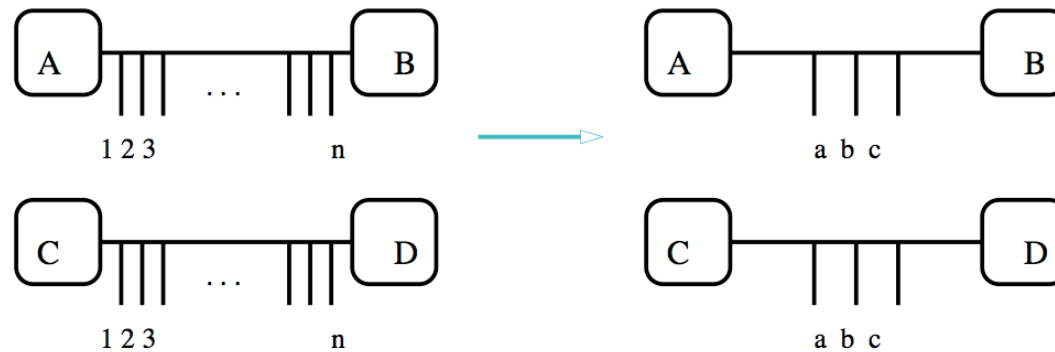
# Lower Bounds for Subchain Reduction



Hickey *et al.* showed that:

- uSPR is NP-hard.

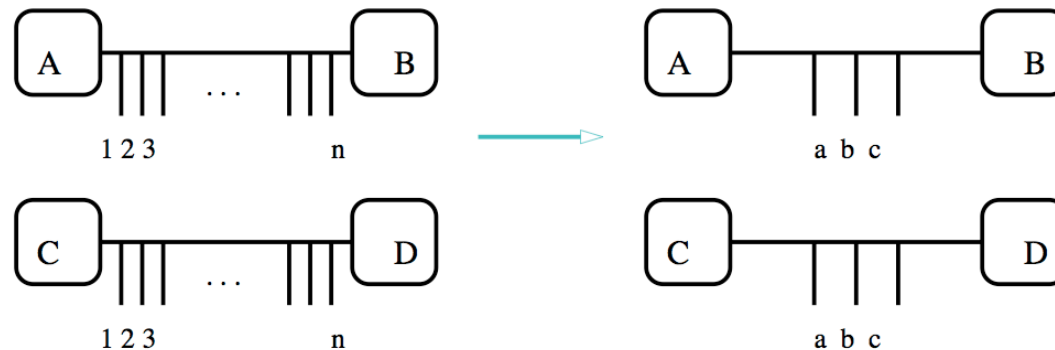
# Lower Bounds for Subchain Reduction



Hickey *et al.* showed that:

- uSPR is NP-hard.
- $d_{uSPR}(T_1^n, T_2^n) - 2 \leq d_{uSPR}(T_1^3, T_2^3)$ .

# Lower Bounds for Subchain Reduction



Hickey *et al.* showed that:

- uSPR is NP-hard.
- $d_{uSPR}(T_1^n, T_2^n) - 2 \leq d_{uSPR}(T_1^3, T_2^3)$ .
- We improve this to:  $d_{uSPR}(T_1^n, T_2^n) - 1 \leq d_{uSPR}(T_1^3, T_2^3)$

# uSPR is FPT

We show that uSPR is fixed parameter tractable with respect to parameter  $k$ :

That is, for distance  $k$ , it can be decided in  $p(n)f(k)$  time that two trees are distance  $k$ ,

- where  $p$  is a polynomial in  $n$ , the number of leaves in the tree, and
- $f(k)$  does not depend on  $n$ .

# uSPR is FPT

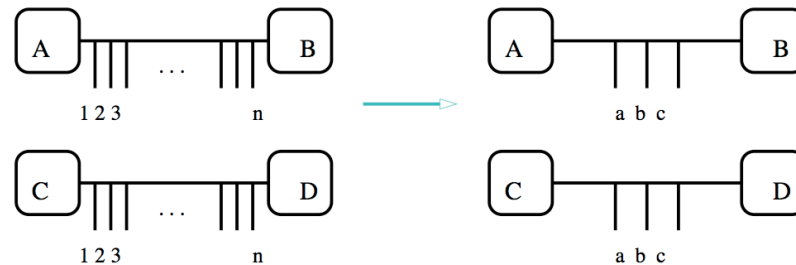
We show that uSPR is fixed parameter tractable.

Relies on two results:

- Lower bound for subchain reduction (either “ $-1$ ” or “ $-2$ ”)
- All the trees with more chain elements than the number of moves have the same distance (so only need to check  $f(k)$  number of extensions).



# Idea for Improving Lower Bounds



$$d_{uSPR}(T_1^n, T_2^n) - 1 \leq d_{uSPR}(T_1^3, T_2^3)$$

- Idea inspired from Hickey *et al*: treat common chains as subtrees to get “-2” bound.
- Go carefully by cases on a minimal set of moves to improve the bound to “-1”.

# Outline of FPT Proof

To check if trees are uSPR distance  $k$ :

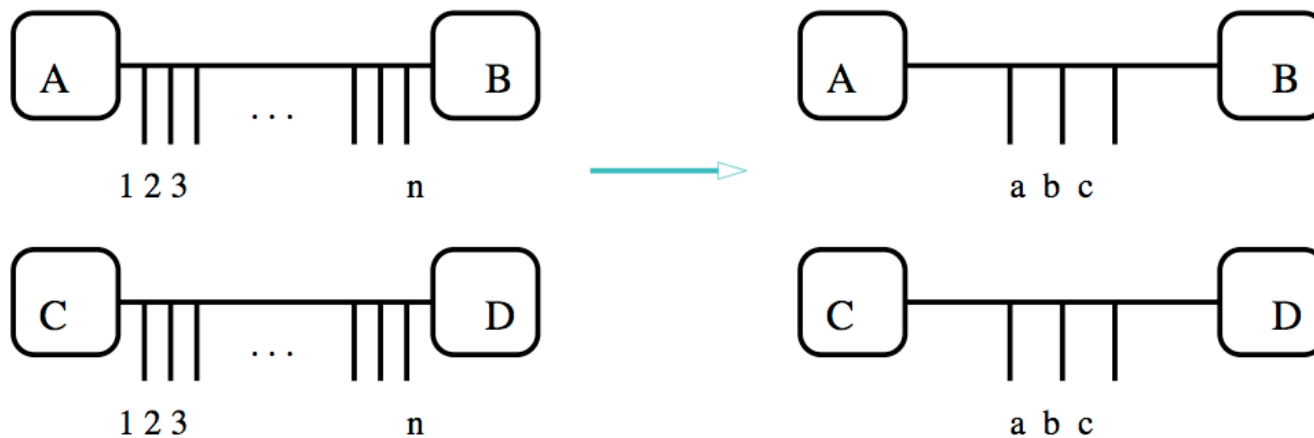
1. Reduce trees.  
linear in number of leaves
2. Check by brute force if there's a set of moves that transforms the trees. depends only on  $k$
3. For every set of possible moves, check if transforms reduced trees.  
depends only on  $k$
4. If it does, try extending to chains of length  $2k$ .  
depends only on  $k$

uSPR is FPT in parameter,  $k$ , the distance

# Open Problem



Does shrinking common subchains  
preserve SPR distance?  
(Implies fixed parameter tractability)



# Acknowledgements



*H. saintjohnianus*

(Image courtesy of NTBG)

- The Newton Institute and organizers of the special semester
- The National Science Foundation for their generous support