Tools for comparing and choosing between alternative phylogenetic inferences

Steven Woolley, Samuel Harrington and Alan Templeton

Washington University in St. Louis, Missouri

Outline

- Motivation

- Current Progress

- Future

Motivation

- Many phylogenetic inference algorithms, file formats, software applications, etc.

- Many of these are not simply trees.

- Tools for comparing and/or visualizing trees/networks are in their infancy.

- How can we choose between alternatives without knowing how (or if) they differ on our data?

Example…

- Different software, different output…
- Does it matter which method is used?
- When does it matter?
- How much does it matter?
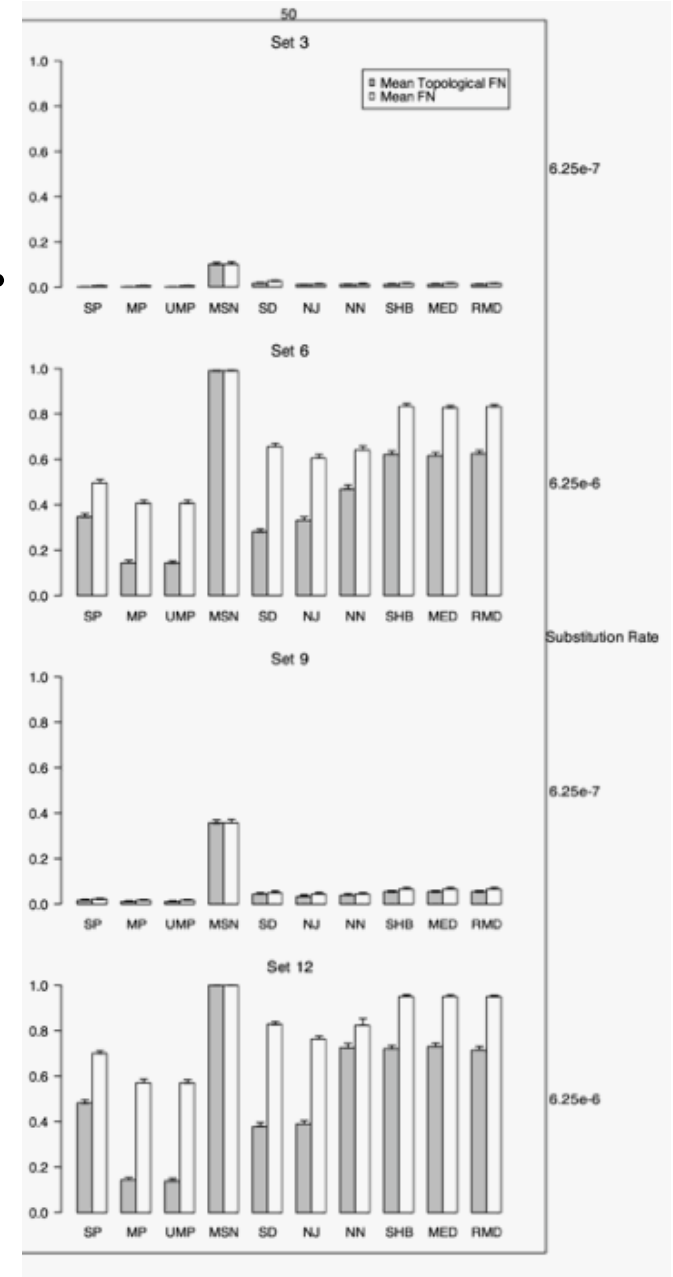
Woolley et al 2008, PLoS ONE

How to compare?

- With trees, we have measures such as RF score, Branch Score, etc.

- With networks, several measures have been proposed but…

  - Are different methods even comparable???
  - Which measure is best and in what circumstance?
  - Will a measure work when comparing inferences from disparate software?
  - What about visualization?

Current Progress

- Skipping the issue of what comparison measure is best…

- For our comparison study, we measured whether the simulated topologies and/or branch lengths were "contained" within the inferred tree/network.

Huh???

- Enumerate trees from inference (N)
- Set of trees simulated (T)
- Calculate fraction of trees/topologies in N but not in T and vice versa. (Type I and II errors)

Implementation

- Input:
  - two inferred networks or trees (leaf sets must match)
    - importers available for Splitstree, Neighbornet, shrub-gc, newick, ms simulation output, extended newick, TCS, Union of maximum parsimony trees, and more.
- Output:
  - Fraction of trees/topologies in only one or in both
  - Various summary statistics related to measures (mean branch difference, number of contained trees, etc.)

So?? What does it mean?

- Tells whether one phylogeny contains one or more exact trees or topologies of the other.
- But… doesn't really give a sense of where they might differ.

Visualizing differences

- Showing where 2 phylogenies (potentially networks) differ or are the same.
- 2 simple simple algorithms tried:
  - Match first by node label (where possible) and then iteratively, by "matching" nodes with similar nearest neighbors.
  - Match first by node label, then by similarity of (possibly weighted) distances from a node to all other already "matched" nodes.

Visualizing 2

- Matched branches/nodes are shown in black
- Branches/nodes present in one but not the other phylogeny are colored differently.

Future

- Better visualization
- More formats (or fewer hopefully?)
- More measures
- More simulataneous comparisons (not just pairwise)
- Software is (almost) available… you can find it by googling "steven woolley" or emailing me at: stevenwoolley@wustl.edu

Acknowledgements

- MIEP organizers
- Alan Templeton
- Sam Harrington
- My family---------->
- Funding
  - NSF Graduate Research Fellowship
  - WashU Young Scientist Training program