TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees

Paper Presentation

Xilin Yu

University of Illinois Urbana-Champaign

Nov 27th, 2018
Table of Contents

1 Background

2 The k-shrink Problem

3 Statistical Test Methods

4 Experiments and Evaluation

5 Discussion and Conclusion
# Table of Contents

1. **Background**

2. **The k-shrink Problem**

3. **Statistical Test Methods**

4. **Experiments and Evaluation**

5. **Discussion and Conclusion**
Problem: errors from early steps propagate to downstream

Possible sign of errors: unexpectedly long branches in inferred tree

Common species filtering methods
  - Rogue taxon removal (RTR, e.g., RogueNaRok)
  - Rooted filtering based on branch length (e.g., Rooted Pruning)

No clear definition of outlier (erroneous sequences)
  - causing unexpectedly long branches
  - causing discordance in gene trees
  - large edit distance to the rest – causing underalignment
  - low probability of generation by the profile HMM on the set
The k-shrink Problem

Definition

The diameter of a tree is the maximum distance between any two leaves.

Definition (The k-shrink problem)

Given a tree on $n$ leaves with branch lengths and $k \in [n]$, for every $i \in [k]$, find a set of $i$ leaves whose removal reduces the tree diameter maximally.

Definition

1. A diameter pair of vertices is any pair of vertices whose distance is equal to the diameter of the tree.
2. A reasonable removal is a removal of a leaf that belongs to some diameter pair.
3. A reasonable $k$-removing set is a set of $k$ leaves s.t. there is an ordering $x_1, \ldots, x_k$ s.t. the removal of $x_i$ is reasonable after removing all of $x_1, \ldots, x_{i-1}$.
Polytime Algorithm for the k-shrink Problem

- Suppose for simplicity, there is only one diameter pair for the tree and any tree obtained by restricting it to a subset of leaves.
- DAG of reasonable k-removal sets

\[ k = 0 \]
\[ k = 1 \]
\[ k = 2 \]
\[ k = 3 \]
\[ k = 4 \]
Theorem

Any $k$-removing set that maximally reduces the tree diameter is a reasonable $k$-removing set.

Theorem

There are $k + 1$ reasonable $k$-removing sets.

The above implies $O(k^2)$ nodes in the graph to search. Poly sized graph + poly time to find each node gives poly time algorithm.

The paper shows that the above holds for general trees as well.
Table of Contents

1 Background

2 The k-shrink Problem

3 Statistical Test Methods

4 Experiments and Evaluation

5 Discussion and Conclusion
Notations for Statistics

- $d_i =$ minimum tree diameter after removing $i$ leaves.
- $\text{OPT}_i =$ set of $i$ leafs to remove to achieve $d_i$.
- $\nu_i = \frac{d_i-1}{d_i}, \Delta_i = \log \nu_i$
- the signature of a leaf $x$ is max $\Delta_i$ such that $x \in \text{OPT}_i$

Outliers are those with abnormally large signature
Three Statistical Tests

| Tests         | # of trees | # of density functions | sequences to remove                                                                 |
|--------------|------------|------------------------|--------------------------------------------------------------------------------------|
| Per-gene     | 1          | 1                      | sequence whose signature has cumulative density $\geq 1 - \alpha$                     |
| All-gene     | multiple   | 1                      | sequence from gene trees in which signature has cumulative density $\geq 1 - \alpha$ |
| Per-species  | multiple   | n                      | sequence from gene trees in which signature has cumulative density $\geq 1 - \alpha$ for that species |

$\alpha = \text{false positive tolerance parameter}$
Experiments

Data Set

| Data   | Sequences | Genes | Outgroups |
|--------|-----------|-------|-----------|
| Plants | 104       | 852   | 4         |
| Mammal | 37        | 424   | 1         |
| Insects| 144       | 1478  | 9         |
| Cannon | 78        | 213   | 5         |
| Rouse  | 26        | 393   | 4         |
| Frogs  | 164       | 95    | 8         |
| HIV    | 648       | 1     | 7+2       |

Methods

1. TreeShrink (with 20 different $\alpha$ values)
2. RogueNaRok (with 20 different weights)
3. Rooted filtering (with 20 cut offs using different number of deviations)

Evaluation criterion: gene tree discordance and taxon occupancy
### Table 2 The impact of the three tests of TreeShrink on taxon occupancy

| Dataset   | Method      | Portion of data removed(%) | Portion of outgroups removed(%) |
|-----------|-------------|----------------------------|---------------------------------|
| Plants    | Per-gene    | 3.3                        | 29.9                            |
|           | All-gene    | 2.5                        | 12.8                            |
|           | Per-species | 4.9                        | 5.1                             |
| Mammals   | Per-gene    | 0.6                        | 11.8                            |
|           | All-gene    | 1.2                        | 17.0                            |
|           | Per-species | 3.6                        | 4.7                             |
| Cannon    | Per-gene    | 1.4                        | 6.2                             |
|           | All-gene    | 1.3                        | 4.7                             |
|           | Per-species | 3.5                        | 5.0                             |
| Rouse     | Per-gene    | 1.3                        | 1.9                             |
|           | All-gene    | 1.2                        | 1.1                             |
|           | Per-species | 4.0                        | 4.5                             |
| Insects   | Per-gene    | 1.2                        | 6.6                             |
|           | All-gene    | 0.8                        | 2.9                             |
|           | Per-species | 4.3                        | 5.0                             |
| Frogs     | Per-gene    | 1.3                        | 26.7                            |
|           | All-gene    | 0.8                        | 15.9                            |
|           | Per-species | 2.7                        | 4.5                             |
Evaluation

(a) Delta MS vs Proportion of species retained
(b) Delta MS vs Proportion of species retained
(c) Delta MS vs Proportion of species retained
(d) Delta MS vs Proportion of species retained
Evaluation

TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees

Nov 27th, 2018
Table of Contents

1 Background

2 The k-shrink Problem

3 Statistical Test Methods

4 Experiments and Evaluation

5 Discussion and Conclusion
The k-shrink problem is solvable in polynomial time.

The per-species test is most effective among the three tests, but it also demands more data.

Treeshrink works better than rooted filtering on majority of data sets in the test.

Treeshrink and RogueNaRok have different target and can complement each other.

Treeshrink is scalable: $10^6$ sequences in 28 minutes.