Phylogenetic Error-Correction for Viral Transmission Network Inference

Mukul S. Bansal

Department of Computer Science and Engineering,
University of Connecticut,
USA

CAME
August 20, 2017

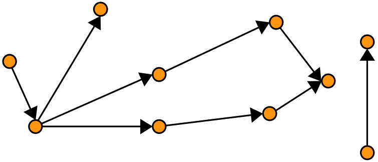
Disease Transmission Networks

Problem

How did a given infectious disease spread from individual to individual?

- Input: Viral/bacterial sequences taken from a population of infected hosts.
- Output: A directed network where each node is an individual host and each directed edge represents direct (or indirect, through unsampled individuals) transmission.

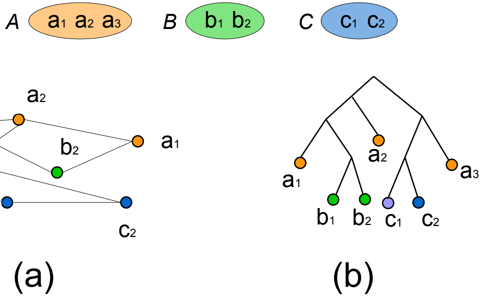
We assume that we have a sample of sequences from each host (not just a consensus sequence).



Inference of Transmission Networks

- Can build relatedness graph.
- Can do phylogenetic analysis of the sequences.

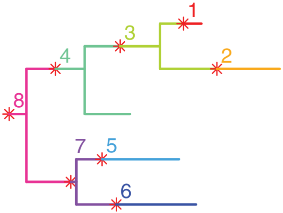
The relatedness graph or phylogenetic tree is then processed to estimate the transmission network



Phylogeny-based Inference of Transmission Networks

Two main approaches:

1. Construct phylogeny from sequences and use it for clustering sequences.
   - Identifies outbreaks, suggests possible transmissions.
2. Use epidemiological model of viral sequence evolution to label internal nodes of a phylogenetic tree with individuals. Perform MCMC search for the phylogeny and the labeling, either separately or simultaneously. E.g., Didelot et al., 2014; Hall et al., 2015; Didelot et al., 2017.
   - Gives full transmission history (but makes many simplifying assumptions).



Didelot et al., 2017

Challenges with using phylogenies

1. Phylogenies can be highly error-prone and uncertain.
   - Small sequences with insufficient information.
   - Slow/fast rates of evolution, or short/long branches.
2. Incorrect phylogeny → incorrect inferences.
3. Estimation of phylogenies and labelings using MCMC is highly computationally intensive → Not scalable.

Thus, current methods are either highly error-prone or computationally intensive or both.

Phylogenetic Error-Correction

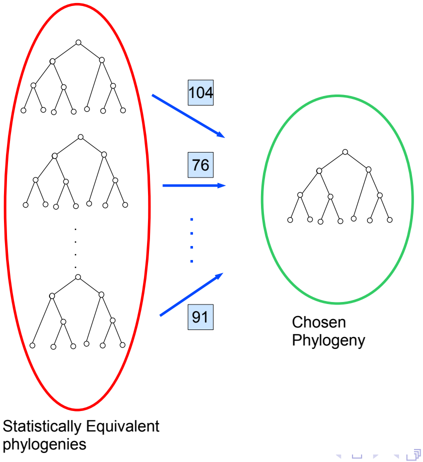
Goal

Construct viral sequence phylogenies more accurately.

- We introduce the first computational method for error-correcting viral sequence phylogenies: TreeFix-VP (viral phylogeny)
- More accurate phylogenies will lead to:
  - More accurate clustering and outbreak/transmission inference.
  - Remove need for MCMC or co-estimation to estimate phylogeny, leading to greatly improved scalability.

TreeFix-VP

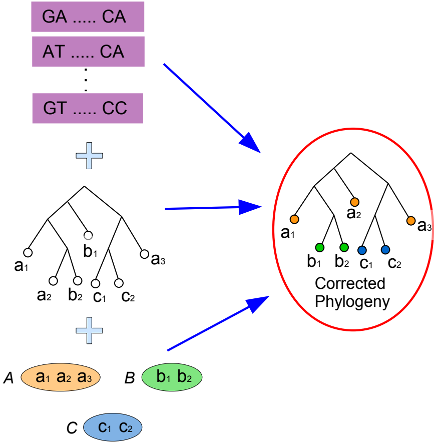
Idea: Search over *candidate* phylogenies and choose one with lowest *cost*.



Overview of Algorithm

Input: ML viral phylogeny, multiple sequence alignment, and host assignment for each sequence.

Output: Reconstructed (error-corrected) viral phylogeny.



Overview of Algorithm

Step 1: Start from input ML tree and search in its local neighborhood for trees that are “statistically equivalent” to ML tree by SH-test and have a lower cost.

Step 2: Repeat the local search step above using the best tree found so far.

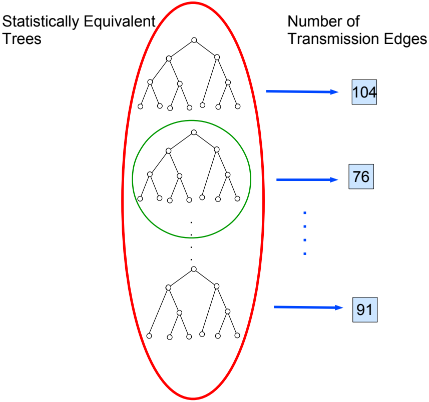
Step 3: Terminate after certain number of search steps.

How can we define the *cost* of a candidate phylogeny?

Defining the cost of a viral phylogeny

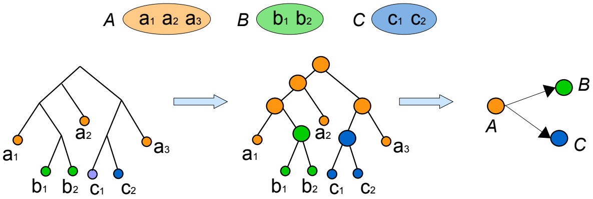
Requirements: Should be *biologically meaningful* and *efficiently computable*.

Idea: Compute *minimum number of required transmission events*. The lower the better.



Computing the Minimum Number of Transmission Events

1. Label leaves with individuals.
2. Assign individuals to internal nodes and use parsimony with individuals as character states.
3. Use Fitch's or Sankoff's algorithms for the small parsimony problem. Complexity O(Size of tree × number of individuals).
4. Edges labeled with different individuals at its end points represent transmission edges.



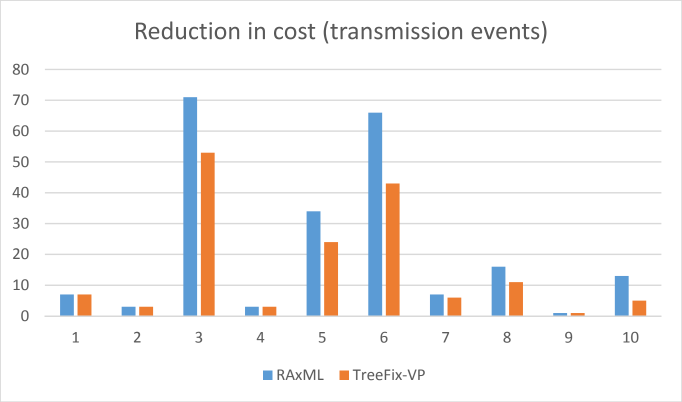
Performance Evaluation

Dataset:

- 142 intra-host HCV populations from 33 outbreaks (provided by CDC),
- Outbreaks contain from 2 to 19 samples, and
- A few dozen to a few hundred sequences.
- True transmission history known for 10 of the outbreaks.

Error-Correction Greatly Reduces Noise

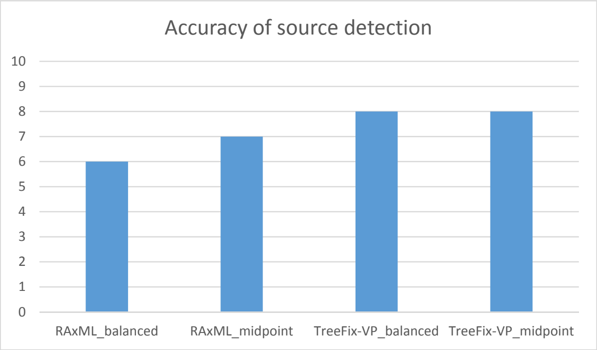
Reduction in the minimum number of transmission events for the 10 outbreaks with known transmission histories.



Error-Correction Leads to Improved Inferences

Accuracy of source inference for the 10 outbreaks with known transmission histories.

- Parsimony assignment at root is assumed to be the source.
- Phylogeny is rooted using either midpoint rooting or rooting on edge that best balances total branch lengths.



Summary

- TreeFix-VP: Statistically informed, fast and scalable, easy to use.
- Can lead to more accurate inference of transmission events and more scalable analyses.
- Next step: Test using thorough simulation framework.

Acknowledgements

Student: Chengchen Zhang
Collaborators: Ion Mandoiu, Alex Zelikovsky, Pavel Skums, Yury Khudyakov.
Funding: NSF award CCF 1618347

Questions!