Bayesian Inference of Infectious Disease Transmission from Whole-Genome Sequence Data

Dreycey Albin (da39)

March 21, 2019

Scientific question

The scientific question and goal for this paper focuses on *using existing software in combination with a Bayesian method to infer a transmission network using whole-genome sequences.* This method is able incorporate within-host diversity, thereby accounting for multiple genetic isolates within a carrier. In addition, the method is able to incorporate pertinent epidemiological data.

Hypothesis tested

The hypothesis tested in this research is that *there method will be able to construct a timed-phylogenetic tree, and from this tree, infer a transmission network.* This involves a 2 step process, where the timed-tree is first developed using BEAST and ClonalFrame, and thereafter, their unique Bayesian-MCMC approach is used to build the transmission network. While previous methods have been developed for inferring transmission from sequence data, these methods fail to incorporate within-host genetic diversity for pathogens with long generation times. These previous methods also fail to produce time-calibrated trees.

Methods/Results

The authors apply this method to both simulated data, to gather performance metrics, and to data gathered from an M. tuberculosis outbreak.

(1) Simulated Data

Starting with the results for the simulated data, the method was able to effectively reconstruct simulated transmission trees. First, a transmission tree *without within host diversity* was constructed. This tree was simulated with a per year recovery rate of 2, and a per contact infectivity rate of 0.02. This gave a tree where the transmission network was synonymous with the coalescence. Next, a a transmission tree *with within host diversity* was constructed. When running their method on this tree, the authors were able to reconstruct a timed-tree with the basic topology of the simulated tree, but the within host diversity parameters were harder to capture. After selecting for posterior probabilities above 10 percent, they used Edmonds algorithm to find the most optimal branching tree. After running 100 simulations of the transmission tree with within-host diversity, a trend for the posterior probabilities became apparent, carrying 27 percent of the probability weight.

(2) Reconstructing the M. turberculosis outbreak data

The authors first constructed a BEAST phylogeny showing the that the first case could have infected, at most, eight other people. Thereafter, they tested their Baysian-MCMC to infer the transmission network. This gave a tree structure with many forward and reverse events. They decreased the number of bidirectional events by using available epidemiological data. This modified the posterior probabilities, and gave a tree showing "waves" of transmission. For most cases, these trees were able to capture the transmissions elucidated purely on epidemiological data. In addition, the method was able to produce transmission parameters similiar to what is expected for M. turberculosis.

Key implications of the results

The ajor implications for the research consist of making a Baysian method that is able to incorporate within-host diversity, genomic sequencing data, and epidemiological data to reconstruct a transmission tree. Using simulated data and real M. turburculosis data, this method is able to recapitulate the transmission network.