QUENTIN: reconstruction of disease transmissions from viral quasispecies genomic data

Palash Sashittal

University of Illinois at Urbana-Champaign

Paper Review

October 30^{th}, 2018

Table of Contents

Interpreting Genomic Variation to Understand Disease Transmission

- Molecular Epidemiology has become an integral tool to understand transmission dynamics of pathogens
- There are several challenges in inferring transmission history

Challenges

- Intra-host population diversity
- Limited epidemiological data
- Viral dynamics complexity

Mutation rate

Time

	Virus	Bacteria
Chronic	HIV HCV	Bacteroides
Acute	Ebola Lassa	Cholerae

Contributions of this paper

- Computational approach for the inference of
  - transmission clusters
  - direction of transmission
  - source of outbreak and transmission history
- Uses complex intra-host viral evolution and utilizes properties of inter-host social networks applied to transmission networks
- Applied to experimental and simulated data of HCV outbreaks investigated by CDC in recent years

Definitions

- Genetic network $\mathcal{G}_g$: Undirected network where vertices correspond to viral genomes and edges connect genomes which differ by a single nucleotide
- Host network $\mathcal{G}_h$: Directed weighted graph with infected hosts as vertices and edges indicated possible transmission directions. The weights $\mathcal{W} = (W_e)_{e \in E(\mathcal{G}_h)}$ are equal to the genetic distance between corresponding viral populations
- Transmission tree $\mathcal{T}$: Rooted binary tree with leaves as the infected hosts and internal nodes represented transmission events.

Genetic Distance

- Consider viral populations P_1 and P_2, find $d_{1,2} = d(P_1, P_2)$
- Consider $\mathcal{G}_g$ that contains P_1 and P_2 and estimated unsampled variants *(using MJ)*
- Viral evolution is modeled as random process on a genetic network
- $d_{1,2}$ is expected time of evolution starting at vertices of P_1 to reach each vertex of P_2

Genetic Distance

- Let x^t be a vector with x_i^t is the expected number of virions with i^{th} genome at time t

$$x^t = \left(1 - \sum_{i=1}^{n} x_i^{t-1}/M\right)((1+r)I_n + qA)x^{t-1}$$

  where A is adjacency matrix, M is maximal population size, r and q are probabilities of replication and single mutation

- Distance computed by taken $x_i^0 = \delta_0$ if $i \in P_1$ and $x_i^0 = 0$ otherwise

$$d_{1,2} = \min\{t : x_i^t \geq \delta_0, \forall i \in P_2\}$$

- $d_{1,2}$ need not be equal to $d_{2,1}$, suggests directionality

$$W_{1,2} = d_{1,2}, \quad \text{if } d_{1,2} < d_{2,1}$$

  for $(1,2) \in E(\mathcal{G}_h)$

Transmission tree

- Weakly connected components of the host network are identified as transmission clusters
- Trees inside these clusters is inferred using MCMC

$$p(\mathcal{T}|\mathcal{G}_h, \mathcal{W}) \propto p(\mathcal{W}|\mathcal{G}_h, \mathcal{T})p(\mathcal{T}|\mathcal{G}_h)$$

- First term is the likelihood of $\mathcal{W}$ given $(\mathcal{G}_h, \mathcal{T})$ and second is the prior of $\mathcal{T}$
- $\mathcal{T}$ only includes tree topology
- Very interesting methods used to estimate both these terms

Likelihood

- Estimated by how closely $\mathcal{W}$ relates to the $\mathcal{T}$ topology
- Solve the following constrained least squares problem

$$\min \sum_{(i,j)\in E(\mathcal{G}_h)} (X_{i,j} - W_{i,j})^2 / W_{i,j}$$

$$\text{s.t.} \sum_{e\in C_i} \alpha_e x_e = \sum_{e\in C_{i+1}} \alpha_e x_e, \quad i = 1, \cdots, n-1$$

$$x_e > 0, \forall e \in E(\mathcal{T})$$

$$X_{i,j} = \sum_{e\in C_{i,j}} x_e$$

where $C_{i,j}$ is the path between leaves i and j and C_i is the path from leaf i to root. α_e is a time-scale parameter

- Finally

$$p(\mathcal{W}|\mathcal{G}_h, \mathcal{T}) = r(\mathcal{W}, \mathcal{X})$$

where r is the Pearson correlation which is interpreted as probability

Prior

- Assumption: Among trees that agree with $\mathcal{G}_h$, probability is proportional to 'scale-freeness'
- *s-metric* is used to measure 'scale-freeness'

$$s(\mathcal{G}_T) = \sum_{(i,j)\in E(\mathcal{G}_T)} d_i d_j$$

where d_i is undirected degree of vertex i

- Let s^* be the maximal s-metric among all graphs with the same degree distrubtion as $\mathcal{G}_T$.
- degree distribution not known a-priori, so consider graphs with same expected number of hubs k to get

$$s^*(k) = (\lfloor n/k \rfloor + k - 2)(n-1) + (k-1)(\lfloor n/k \rfloor - 1)n/k$$

Star network of k hubs with degree $\lfloor n/k \rfloor - 1$

- Finally

$$p(\mathcal{T}|\mathcal{G}_h) = \begin{cases} \kappa \exp\left(-\rho|1 - \frac{s(\mathcal{G}_T)}{s^*(k)}|\right), & \text{if } \mathcal{T} \text{ agrees with } \mathcal{G}_h \\ 0, & \text{otherwise} \end{cases}$$

Results

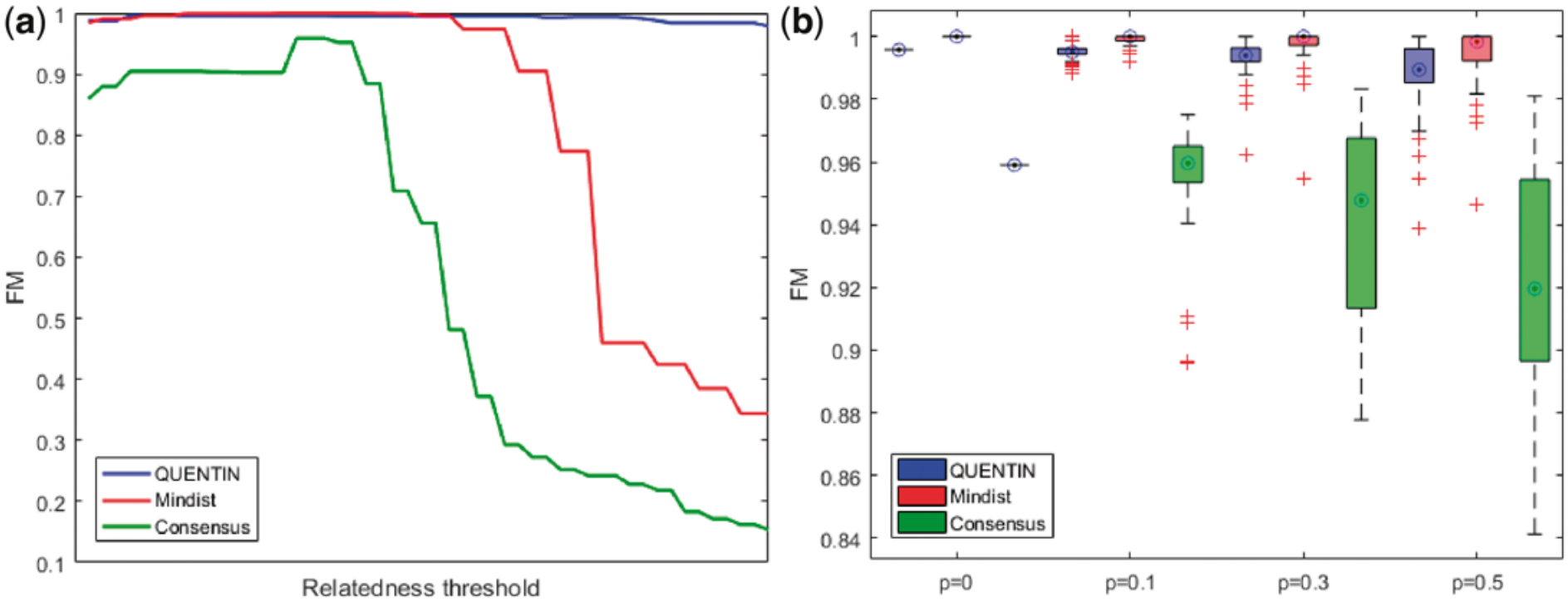

Figure: Robustness results. (a) Threshold variation robustness (b) Sampling robustness

Summary

- This paper introduces very interesting ideas used for computing/estimating
  - Distance between genome populations
  - Likelihood of distances given tree topology
  - Prior on the transmission tree topology
- Molecular Epidemiology is a very interesting application of the tools we were taught in the class

Questions (BACKUP)

- Q: How are viral transmission networks and generalized social networks related? I understand the basic idea of a host corresponding to a person, but there are probably a lot of confounding factors like immune systems or resistance in particular hosts that could make these networks considerably different.
- A: General properties like power-law distribution and presence of hubs are observed. While computing prior, transmission networks which are close to scale-free are considered more probable. They concede in the paper that standard measures of 'scale-freeness' are poorly applicable to real transmission networks.
- Q: I don't see any information on how long QUENTIN takes to complete, or how long the other methods take to complete. It seems very computationally intensive, so could a heuristic similar to this method be created that perhaps considers these temporal epidemiological data?
- A: