Phylogenetic Clustering of HIV and its Connection to Sexual Contact Networks

Luc Villandré, Aurélie Labbe, David A. Stephens

Department of Epidemiology, Biostatistics and Occupational Health, Montreal, Canada

The Human Immunodeficiency Virus (HIV)

- HIV Type 1 (HIV-1) became pandemic in the 1980s.
- Prevalence low in the developed world, but high among men who have sex with men (MSM) (PHAC M-Track Survey, 2012).
- No vaccine or cure for HIV, but antiretroviral therapy (ART) pushes back onset of AIDS and death considerably.

Tracking HIV Transmission in MSMs

- Tracking HIV transmission in MSMs is difficult.
- Systematic contact tracing is generally unfeasible because of anonymous sex partners, inability to recall sexual encounters, cultural taboos.
- *Phylogenetics* allows inference of transmission based on genotyping data.
- HIV genotyping data available after drug-resistance testing.

Relevance of Transmission Clusters

- Transmission cluster: group of infections sharing a "close" common ancestor.
- Heuristic definition: set of DNA sequences similar at the genetic level and resulting from transmission events close in time.
- HIV epidemic in Montreal shows high levels of clustering (Brenner et al. 2011).
- Clusters have their own growth dynamics (and are growing).
- Randomly targeted intervention will not affect HIV incidence in MSMs (Leigh Brown et al., 2011).

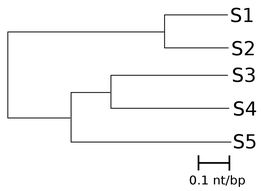
Rationale and Research Question

- HIV prevalence in MSMs in Montreal est. at 13% (PHAC M-Track Survey, 2012).
- HIV clusters because of constraints on its transmission, e.g. viral load must be high ($> 100,000$ copies/mL).
- Understanding clusters necessary to control HIV transmission among MSMs.
- How do transmission clusters relate to sexual contact networks?

Basics of Phylogenetics

- **Phylogeny**: Tree-like representation of the ancestry of DNA sequences.
- **Phylogenetics**: Field of stat. genetics concerned with inference of phylogenies.
- Model input: DNA sequences.
- Goal of inference: Phylogeny (not clusters).

Example: A Phylogeny



- Rooted phylogeny for a sample of 5 DNA sequences, "S1" to "S5".
- Branches = lineages: merge at common ancestor.
- Dist. unit: **Expected** number of nucleotide substitutions per base pair ("nt/bp").
- The raw number of differences between sequences underestimates distance.
- Markov mutation process: reported distance is an expectation.
- Dist. est. vary based on the structure of the assumed transition rate matrix.

Phylo. estimation

- *Tree-searching algorithms* start with a simple, sub-optimal phylo., and recursively explore the space of possible trees to find "better" phylo..
- Heuristic algorithms are used: they propose moves in the tree space, score them, and accept or reject each one based on this score (usually greedily).
- We retrieve the phylo. that maximizes optimality criterion, e.g. max. likelihood.

Epidemics and networks

- Infectious disease models often assume *random mixing* (Keeling et al. 2005).
- In HIV, # of contacts per individual smaller than the pop. size: no random mixing.
- Individuals have a fixed # of contacts $\rightarrow$ network structure.

Simulation algorithm

1. Simulate a network of one of three kinds: Erdos-Renyi (ER), Watts-Strogatz (WS), and Barabasi-Albert (BA),
2. Introduce HIV by selecting a node at random,
3. Generate transmissions, transmission time over each edge exponentially distributed,
4. Shut off nodes once diagnostic occurs (fixed time after infection),
5. Stop simulating transmissions once 50% of nodes are diagnosed,
6. Draw the phylogeny and simulate DNA samples over it,
7. Cluster the resulting sample after excluding undiagnosed cases.

Inference of Transmission Clusters

1. Derive est. of pairwise gen. dist. matrix (K80 model) and obtain *WPGMA* tree,
2. Refine dist. matrix. with obtained phylogeny,
3. Use new dist. matrix to partition sample: hierarchical clustering, average linkage,
4. For subset of epidemics, compare these clusters to those resulting from max. likelihood tree searching, in terms of clustering accuracy.
5. Cluster other simulated samples with the method that offers the best trade-off between speed/clustering accuracy.

Clustering accuracy estimation

- From the true (known) phylogeny, we obtain a reference set of clusters.
- We compare clusters from simulated samples with this reference.
- We measure accuracy with the *corrected Rand Index*.
- Index measures the proportion of correctly co-clustered and separated pairs of elements, adjusted for chance.

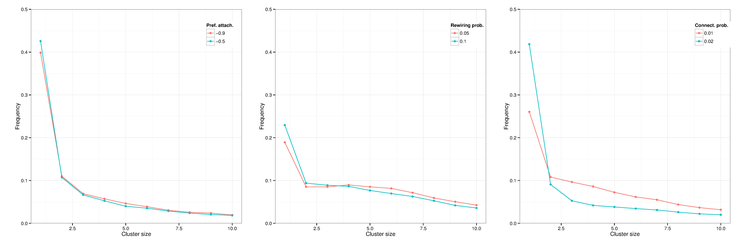
Cluster recovery (50 samples)

Network type	Network parameter	Corr. Rand Index			
		ML		WPGMA	
		Mean	St. Err.	Mean	St. Err.
WS	0.05	0.656	0.002	0.663	0.002
	0.10	0.690	0.002	0.639	0.002
BA	0.10	0.523	0.003	0.491	0.002
	0.50	0.482	0.004	0.461	0.003
ER	0.01	0.649	0.002	0.629	0.002
	0.02	0.544	0.004	0.511	0.003

Table: Mean corrected Rand indices and their SEs over 50 simulated epidemics for WPGMA and maximum likelihood tree-searching clustering algorithms.

WPGMA offers reasonable accuracy compared to ML tree searching, and is much faster $\rightarrow$ we use WPGMA clustering.

Cluster size distributions

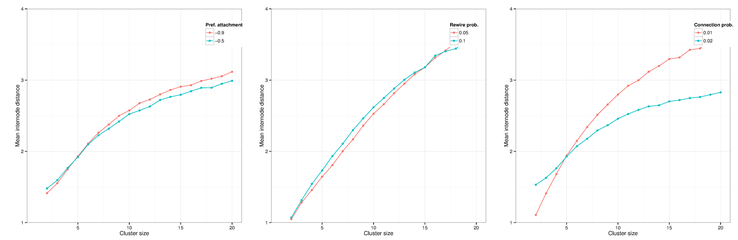


(a) BA network (b) WS network (c) ER network

Figure: Cluster size dist. for epidemics on BA, WS, and ER networks.

- 1000 samples, 150 diagnosed cases, cutpoint $= 0.035$..
- Distributions vary largely between net. types.
- Differences in cluster size dist. within net. types are small, except in ER.

Cluster shape - Mean pairwise dist. between nodes within clusters



(a) BA network (b) WS network (c) ER network

Figure: Mean pairwise distance between nodes against cluster size for HIV outbreaks on BA, WS, and ER networks. 50% of the population, 150 nodes, are diagnosed.

- Do clusters take different shapes across net. param.?.
- Yes, but differences are small, except in ER networks.

Notes and conclusions

- WPGMA is fast and offers a reasonable trade-off between speed and clustering accuracy.
- Network tuning parameters affect the cluster size distributions and the shape of clusters, especially in ER networks.
- Differences in reported means are small overall.
- Means stable, but distributions have high variance $\rightarrow$ unclear how much a single set of inferred clusters could reveal about contact network.
- Considering a larger range of parameter values could be worthwhile, but clustering accuracy becomes an issue.

Bibliography

We performed all computations in R v3.0.2, with the ape and phangorn packages. We plotted graphs with the ggplot2 package.

Brenner, B. G. *et al.* (2011), J Infect Dis, 204, 1115-1119.

Keeling, M. J. *et al.*. (2005), J. R. Soc. Interface, 22, 295-307.

Leigh Brown *et al.* (2011), J Infect Dis, 204, 1463-1469.

PHAC M-track survey (2012), `http://library.catie.ca/pdf/ATI-20000s/26403.pdf`.

Schliep, K. (2011), Bioinformatics, 27, 592-593.

Mail: `luc.villandre@mail.mcgill.ca`