An Application of the Modifiable Areal Unit Problem: Optimizing Cluster Method Parameters to Produce Predictive Data for HIV Outbreaks.

Connor Chato and Art FY Poon

Background

A popular approach to study HIV outbreaks is to cluster cases based on genetic similarity. However, there is no widely-used statistical criterion which optimizes the parameters for sequence-based clustering methods. The relationship between a cluster-defining similarity threshold and it's associated set of clusters can be analogized to the aggregation level in the Modifiable Areal Unit Problem (MAUP).

Hypothesis

Based on the selection of aggregation level for study partitions in MAUP, we present a statistical framework to optimize the similarity threshold for pairwise distance algorithm TN93 (http://github.com/veg/tn93). We hypothesize that defining this threshold includes case connections such that the most predictive clusters are defined for the purposes of public health.

Methods

We obtained 1,653 published HIV-1 pol sequences from Seattle, USA. The sequences were aligned using MAFFT and coupled with sampling dates from Genbank. Years ranged from 2000 to 2013, with 2013 cases reflecting cluster growth. TN93 obtained pairwise distances between sequences and an R script interpreted these distances as an annotated, undirected network, annotated. Edges between cases were included in this network based on cutoff *d*, which was modulated from 0 to 0.06 in steps of 0.001. Based on a Poisson-linked linear model with the cluster growth outcome predicted by cluster size, we calculated the Generalized Akaike Information Criterion (GAIC) for networks at each value of *d*.

Results

GAIC was minimized at *d* = 0.036; notably larger than values often used in literature. Common Values in literature fall within maximum deviance peaks.

Keywords:

HIV
Genetic Clustering
Bioinformatics
Outbreak Prediction
Public Health
Modifiable Areal Unit Problem