Marianne Bjorner
BMI 826 - Fall 2020
Project Proposal

Topic: Comparing graphical and phylogenetic models of outbreak transmission trees

1. Introduction:

The CDC currently has several systems for identifying disease clusters during outbreak investigations. Laboratory networks such as PulseNet and CalciNet are both systems linked to NORS (National Outbreak Reporting System)[5, 6]. PulseNet uses a combination of molecular biology methods including pulsed-field gel electrophoresis (PFGE), whole genome sequencing (WGS), and multiple loci VNTR analysis (MLVA) to subtype bacterial pathogens present in laboratory samples[5]. With this data, PuleNet conducts passive surveillance of these bacterial pathogens nationally[5]. Similarly, CalciNet is used for norovirus detection[6].

While these current methods may cluster diseases based on similar genetic information, they often may not take the extra step of analyzing these laboratory samples to infer the minutiae of disease genealogy, instead opting for a wider scope of analysis in the form of phylogenies[10]. One drawback of using classical phylogenetic structures is that the branch points are not associated with a given sample, but merely suggest the presence of a common ancestor[1]. When ancestors and descendents are used to construct a phylogenetic tree, this results in an inappropriate resulting structure that obscures the true relationships between samples.

Uncovering disease genealogy using a graphical structure that directly depicts ancestors would elucidate information regarding disease origin, which becomes especially important when one transmission route includes person-to-person spread. In practice, this information can be used to halt the progression of disease outbreaks. I aim to use genetic data to construct phylogenetic trees or genealogic graphs as a proxy for transmission trees and outbreak reconstruction. I will be using previously described graphical[1, 14] and phylogenetic tree methods[3, 8] and compare their performance.

2. Approach:

Methods/Algorithms:

(1) and (2) build Graph Structures: where each node is a sample, and edge weights indicate genetic distance/similarity between samples.

(1) SeqTrack algorithm (Jombart et al. 2011)[1] as found in the *adegenet* R package. This may also include using tools found within the *APE* R package, developed by Paradis et al[2]. This constructs the most likely genealogy in the form of a graph structure. It constructs trees using optimum branching methods to identify the most likely ancestors of each sample. This likelihood estimation leverages temporal data and maximum parsimony between genetic samples to infer disease genealogy.
(2) The R package *outbreaker2*[13, 14, 15] contains software useful for computing transmission chains within a Bayesian framework, using MCMC methods. These maximize a combination of a genetic pseudo-likelihood function and an epidemiological pseudo-likelihood function in order to produce the most likely graph structure.

(3) and (4) build Phylogenetic Tree Structures: where each leaf node represents a sample, but branch points are hypothesized ancestors (not samples).

(3) BEAST: Bayesian evolutionary analysis by sampling trees developed by Drummond et. al. (2007)[3]. This method uses MCMC as its basis for determining likely phylogenetic tree structures. This is in java and available for download here.[4]

(4) The PhyML 3.0 algorithm[7,8] developed by Guindon et al., which infers phylogenies based on maximum likelihood methods. Source code is available on Github[7]. Online execution of PhyML is also possible[16, 17].

Criteria used to compare methods:
Simulated data attained through methods described by Jombart et. al.[1] and the *outbreak2* package from Campbell et. al.[14] will be used to develop a ground truth. Outputs of the methods above will be compared to the expected outcome of the simulated data. Ancestral relationships are the most important when it comes to tracing disease transmission, so this will be the primary criterion for accuracy measurement. Since phylogenetic trees do not directly connect parents to children as graph structures, branch lengths will be used for comparison purposes.

Datasets that these methods will be applied to:
I will start with simulated datasets annotated with temporal information. Then I will use datasets which focus on H1N1, the 2009 swine flu. Both simulated and real datasets will consist of nucleotide sequences and temporal data for each sample.

(1) Jombart et. al. 2011[1] has several datasets of note, including a described method implemented in R for attaining simulated DNA data. These simulated data are constructed using either structured dispersal or random diffusion. The simulated data will serve as the ground truth for comparison purposes.
(2) The *outbreak2* R package has a built-in *fake_outbreak* simulated dataset[19].
(3) Jombart et. al. 2011[1] will provide H1N1 datasets
    (a) Hemagglutinin (HA) sequences: "Supplementary Dataset 1"
    (b) Neuraminidase (NA) sequences: "Supplementary Dataset 2"
    (c) Temporal and Geographical Data for corresponding HA and NA sequenced: "Supplementary Dataset 3"

3. Significance:

Because of its current real-world applications to disease surveillance, correctly and efficiently determining disease genealogy is vital to tracing a disease to its source[18]. In the case of food-borne illnesses and outbreaks, this is especially important as identification of the source can lead to the implementation of new safety measures and regulations, and elucidate disease transmission pathways[5, ,6]. Disease genealogy also helps predict the true spread of a pathogen within communities based on expected vs. observed mutation rates[11].

Given the sparse datasets typical of outbreak laboratory samples, graphical methods should ideally infer a transmission tree that accurately depicts ancestral relationships. When combined with other forms of epidemiological data, this is vital to the determination of a disease's source. WGS data, which PulseNet already collects, has shown useful in the endeavor of disease tracing[12]. Future expansion of PulseNet to construct genealogical transmission trees will become increasingly useful as WGS methods continue to decrease in cost and increase in quality and accuracy. With this, faster outbreak source identification will aid in halting outbreaks at early stages, resulting in less human suffering[18].

Once complete, I hope to compare the efficiency and accuracy of disease genealogy methods. With this I will assess limitations of either the models themselves or the Bayesian methods they are built upon. As this is my first foray into conducting phylogenetic or genealogic graphical analyses, I expect to run into some hurdles related to configuring real datasets and optimally tuning initial parameters of the provided algorithms.

4. References:

1. T Jombart, R M Eggo, P J Dodd and F Balloux *Reconstructing disease outbreaks from genetic data: a graph approach*. Heredity. 2011. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3183872/
2. E Paradis, J Claude, K Strimmer *APE: Analyses of phylogenetics and evolution in R language. Bioinformatics*. 2004. https://pubmed.ncbi.nlm.nih.gov/14734327/
3. A J Drummond and A Rambaut. *BEAST: Bayesian evolutionary analysis by sampling trees.* BMC Evol. Biol. 2007. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2247476/
4. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ & Rambaut A *Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10* Virus Evolution 4, vey016. 2018. DOI:10.1093/ve/vey016
5. *"PulseNet Methods."* Pathogens and Protocols | Pulsenet. CDC. https://www.cdc.gov/pulsenet/pathogens/index.html
6. Responding to Norovirus Outbreaks. CDC. https://www.cdc.gov/norovirus/trends-outbreaks/responding.html
7. PhyML - Phylogenetic estimation using (Maximum) Likelihood. Github. https://github.com/stephaneguindon/phyml
8. S Guindon, J F Dufayard, V Lefort, M Anisimova, W Hordijk, and O Gascuel. *New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0*. 2008. http://www.atgc-montpellier.fr/download/papers/phyml_2010.pdf
9. B Swaminathan, T J Barrett, S B Hunter, R V Tauxe, and the CDC PulseNet Task Force *PulseNet: The Molecular Subtyping Network for Foodborne Bacterial Disease Surveillance, United States.* Emerging Infectious Diseases. 2001. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2631779/pdf/11384513.pdf
10. E Vega, L Barclay, N Gregoricus, K Williams, D Lee, and J V. *Novel Surveillance Network for Norovirus Gastroenteritis Outbreaks, United States.* Emerging Infectious Diseases. 2011. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3381557/
11. R K Pathan, M Biswas, and M U Khandaker *Time series prediction of COVID-19 by mutation rate analysis using recurrent neural network-based LSTM model.* Chaos Solitons Fractals. 2020. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7293453/
12. M P Powers *Whole Genome Sequencing Used to Trace COVID-19 Outbreaks*. The Association of Public Health Laboratories. 2020. https://www.aphlblog.org/whole-genome-sequencing-used-to-trace-covid-19-outbreaks/
13. *outbreaker2: Bayesian Reconstruction of Disease outbreaks by Combining Epidemiologic and Genomic Data* https://cran.r-project.org/web/packages/outbreaker2/index.html
14. T Jombart, A Cori, X Didelot, S Cauchemez, C Fraser, N Ferguson. *Bayesian Reconstruction of Disease outbreaks by Combining Epidemiologic and Genomic Data.* PLOS Computational Biology. 2014. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003457
15. F Campbell, X Didelot, R Fitzjohn, N Ferguson, A Cori, and T Jombart. *Outbreaker2: a modular platform for outbreak reconstruction.* BMC Bioinformatics. 2018. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6196407/
16. Genomics Lab. *PhyML for Phylogenies.* YouTube. 2020. https://www.youtube.com/watch?v=sfOC0FhTiYw&ab_channel=GenomicsLab

17. *PhyML 3.0: new algorithms, methods, and utilities.* ATCG Bioinformatics Platform. http://www.atgc-montpellier.fr/phyml/
18. National Research Council (US) Committee on Achieving Sustainable Global Capacity for Surveillance and Response to Emerging Diseases of Zoonotic Origin; Keusch GT, Pappaioanou M, Gonzalez MC, et al., editors. Sustaining Global Surveillance and Response to Emerging Zoonotic Diseases. Washington (DC): National Academies Press (US); 2009. 4, Achieving an Effective Zoonotic Disease Surveillance System. https://www.ncbi.nlm.nih.gov/books/NBK215315/
19. T Jombart. I*ntroduction to outbreaker2.* R Epidemic Consortium. 2020. http://www.repidemicsconsortium.org/outbreaker2/articles/introduction.html