Inferring Disease Contact Networks from Genetic Data

Frank Chen, Bryan Hooi

December 11, 2013

Abstract

The analysis of genetic sequence data collected during disease outbreaks has emerged as a promising new tool for understanding infectious disease dynamics and designing control measures against infectious diseases. Hence, there is a need for statistical methodologies that effectively integrate genetic data with other epidemiological data to perform inference on the underlying disease dynamics. In this project, we model contact networks using a random graph model, and tackle the problem of inferring parameters of the contact network over which diseases spread, using both genetic and other epidemiological data. We develop a Markov chain Monte Carlo algorithm to perform Bayesian inference for both the parameters of the stochastic epidemic model and the underlying contact network. We evaluate the algorithm by simulating random contact networks, simulating epidemics over these networks, simulating viral evolution sequences over the resulting epidemics, and comparing the inferences our algorithm makes on this data with the true parameters. Finally, we apply our approach to analyze 433 H1N1 viral genetic sequences drawn from the early stages of the H1N1 influenza pandemic.

1 Introduction

1.1 Compartment Models vs. Network Models

In the study of infectious diseases, the most common model has long been the compartment or "mean field" model, in which each infected individual is equally likely to spread the disease to any susceptible member of the population (Kermack & McKendrick, 1932). However, this approach ignores the fact that individuals are embedded in contact networks along which epidemics generally spread. The heterogeneity between individuals in contact networks has been shown to have important implications on epidemic dynamics as well as on the recommended strategies for controlling the spread of these epidemics (Anderson, 1991).

1.2 Inferring Epidemic Contact Networks

Due to the importance of the contact networks over which diseases spread, there are important public health implications in the problem of making use of past epidemic data to better understand these networks, informing how control measures are designed for future outbreaks. While many studies have studied contact networks by simulating from a variety of random graph models, relatively few have tackled the associated inference problem: of inferring the parameters of the contact network (such as the p in Erdos-Renyi $G(n, p)$ models) given epidemiological data. Among those that do are Britton and O'Neill (2002), which models the contact network as an Erdos-Renyi random graph, and uses a Markov chain Monte Carlo approach based on epidemic data to infer its parameters. This line of work has been advanced by Groendyke et al. (2010) to use the SEIR stochastic epidemic model, and further by Groendyke et al. (2011) to use exponential-family random graph models to model the transmission process over the contact network.

1.3 Genetic Sequence Data

As genetic sequence data collected during disease outbreaks becomes increasingly commonly available, such data has emerged as a promising tool for better understanding disease dynamics. So far, genetic data has primarily been used in the form of phylogenetic tree analysis (Grenfell et al., 2004). Others such as Jombart et al. (2011) use a primarily parsimony-based approach to derive the most likely ancestor of each patient's viral genetic sequence, us-

ing epidemiological data only to break ties. Ypma et al. (2010) construct a likelihood function combining genetic, spatial and temporal data, assuming these components are independent of one another to construct a Bayesian inference scheme to obtain posterior intervals for the viral mutation parameters and the transmission history of the disease. Morelli et al. (2012) uses a similar but more complex Bayesian inference scheme that allows dependence between the likelihood components.

2 Problem Description

2.1 Relation to existing work

Our project builds on the work of Jombart et al. (2011): we also use genetic data as a key input to understanding past epidemics. However, we explicitly model the contact network as a random graph and produce an inference scheme that infers the parameters of the network (such as the edge-generation probability in the case of the Erdos-Renyi model). In this way, we also build on Britton and O'Neill (2002), who also use perform inference on the parameters of random graph models, but only use the times of detection and infection of each patient as input, not genetic sequence data.

2.2 Problem statement

Given a collection of patients affected by a disease outbreak, along with their respective detection times, locations, and genetic sequences, what can we infer about the underlying contact network and disease outbreak parameters?

In typical real-life settings, many of the inputs (time, geographical and genetic data) are likely to be missing or uncertain, so the algorithm should be able to smoothly handle this.

2.3 Data

We evaluate our method using both simulations and actual sequence data. The data consists of 433 viral H1N1 genetic sequences sequenced at the hemagglutinin and neuraminidase genes, as well as the geographical (latitude/longitude) coordinates of the associated patients. The sequences are based on freely available data on GenBank (Benson et al., 2010), annotated and aligned by Jombart et al. (2011).

2.4 Methods

2.4.1 Model

In this section we describe the model we use for contact networks, epidemics, as well as underlying the observed data. The model, as well as the Bayesian inference scheme, is fairly similar to that used in Britton and O'Neill (2002), except that we make use of genetic data, while they used only infection time data. Another difference is that we also chose to only model patients who actually get infected, because this reduces the number of variables in the model, and because in actual applications, there is no clear method of estimating the number of patients who could have been infected by a disease but were not infected.

The population is assumed to consist of a fixed population N. We model the underlying, unobserved contact network structure by a random graph G over these nodes: the edges in G represent contact between the two patients. We model G using a $G(n, p)$ Erdos-Renyi model.

Next, we model epidemics over this contact network. Denote by H the directed graph corresponding to actual infective contacts: i.e. $e_{ij} \in H$ iff individual j was infected by individual i.

We use a simplified version of the standard SIR (Susceptible, Infectious, Recovered) model in epidemiology: in our case, at any time individuals can be in one of two states: Susceptible or Infectious. Infectious individuals can transmit infections to neighboring susceptible individuals by making infectious contacts: the time that passes between when individual i is infected and when i makes an infectious contact with a susceptible neighbor j follows an exponential distribution with fixed parameter β. Letting t_{ij} be the time taken for i to infect j (assuming j is not infected by another node), we can write this as:

$$t_{ij} = \begin{cases} Exponential(\beta) & \text{if } e_{ij} \in G \\ \infty & \text{if } e_{ij} \notin G \end{cases}$$

Since in general a susceptible node may have multiple infectious neighbors, we model it as being infected at the first time it receives an infectious contact from any of its neighbors. Thus, letting I_i be the infection time of node i, we have the following recurrence:

$$I_i = \begin{cases} 0 & \text{if } i = 1 \\ \min_{j:e_{ij} \in G}(I_j + t_{ji}) & \text{otherwise} \end{cases} \quad (1)$$

Finally, we define the process of viral genetic evolution over an epidemic. To prevent the creation of

a large number of latent variables, we will not model the sequences directly, but instead model the number of mutations, i.e. differences between sequences. In this way we neglect back mutation (the process whereby a mutated gene reverts to its previous state) as well as cases whfere multiple mutations occur on the same location, but since genetic sequences are typically fairly long, such occurrences have low probability.

We use the simple Jukes-Cantor (1969) model of evolution, which assumes equal base frequencies of the bases A, T, C and G, as well as equal mutation rates between any two of these bases. We use a Poisson process model which models genetic mutations as independent rare events, with the result that the number of mutations occurring during the transmission from i to j follows a Poisson distribution with parameter α. Since the amount of time passing in the intervening period between i and j's times of infection is $I_j - I_i$, thus letting $GeneDist$ be the genetic distance function, we have:

$$GeneDist(i,j) \sim Poisson(\alpha|I_j - I_i|) \quad \text{if } e_{ij} \in H$$

The distance between two nodes which are not directly connected in H is the sum of distances $GeneDist$ along the shortest path between i and j in the graph H.

Figure 1 shows the output of a simulation run involving 15 nodes.

2.4.2 Dijkstra-based Simulation Algorithm

In this section, we describe a simple approach based on Dijkstra's algorithm used to speed up the process of simulating epidemics using the model described in the previous subsection. This is especially useful in simulating large training sets involving many such simulated epidemics to use in training a prediction algorithm.

To show that this algorithm works, we simply note that the recurrence (1) defining infection times I_i is exactly the recurrence used by Dijkstra's algorithm to compute shortest paths from n to each node. Furthermore, the infector of node i is $argmin_{j:e_{ij}\in G}(I_j + t_{ji})$, which is the second-last node in the shortest path from n to i.

2.4.3 Direct prediction approach

We also use a different and more straightforward approach along the lines of Adar & Adamic (2005): we

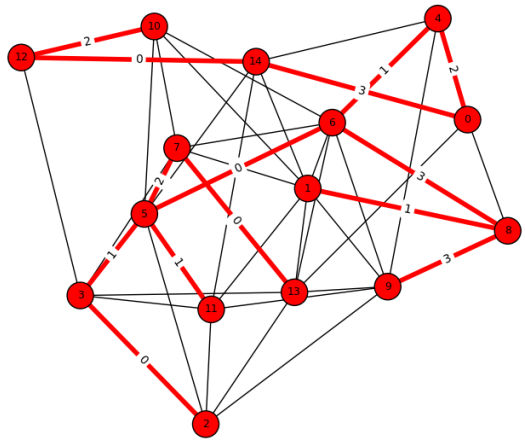

Figure 1: Output of a simulated contact network (edges in black), an epidemic run over this contact network (edges in red), and the numbers of mutations that occurred as the virus spread during the epidemic (edge labels). The first node affected by the epidemic is node 0 (upper-right). This simulation run has 15 nodes, $p = 0.4$, $\beta = 1, \alpha = 5.0$.

use the genetic and epidemiological data to predict the unknown contact network parameters using a machine learning model.

We first use the following features of the observed data to predict the unknown parameter p of Erdos-Renyi graphs:

- Latest infection time among all patients
- Sum of infection times
- Largest number of genetic differences between any two sequences
- Sum of genetic differences between each two sequences

We choose two prediction models, Random Forests and Gradient Boosting Machines - chosen due to their ability to fit nonlinear functions as well as interactions between predictors. We performed the model training using Python's scikit-learn library. (Pedregosa et al., 2011) We tuned the parameters slightly to ensure that the number of estimators in each case was sufficient - we used 500 estimators for the Random Rorest model and 300 for the Gradient Boosting model.

Algorithm 1 Algorithm for simulating epidemics

- **Input**: a contact network G
- **Output**: a simulated epidemic along the contact network

1. For each edge $ij \in G$, sample $t_{ij} \sim$ Exponential(β)
2. Choose a random starting node n to be the index case (i.e. the first infected patient)
3. Run Dijkstra's algorithm starting from n.
4. For each node i, the infection time of i is the shortest path length from n to i, and the node that infected i is the second-last node in the shortest path from n to i.

2.4.4 Bayesian Inference using Markov chain Monte Carlo

We also design a Markov chain Monte Carlo algorithm capable of performing inference on the parameters of a random contact graph based on observed genetic and epidemiological data. Letting M_θ be the model for the contact network (indexed by θ, for example, $\theta = p$ if we are using Erdos-Renyi $G(n,p)$ graphs), and letting:

- I_{AB} be the event in which patient A infects patient B;
- t_A be a random variable for the observed detection time for patient A;
- $Gene_A$ be the observed genetic sequence of patient A;
- $Data = (t, Gene)$ be the collection of observed data;
- L the likelihood function, and L_t and L_{gene} be components of the likelihood function based on temporal and genetic data, we have the following equation:

$$\begin{aligned} L(\theta, I_{AB}|Data) &= L_t(\theta, I_{AB}|t_A, t_B) \\ &\quad \times L_{gene}(\theta, I_{AB}|Gene_A, Gene_B) \end{aligned}$$

In this manner, we perform posterior inference for the model parameter θ.

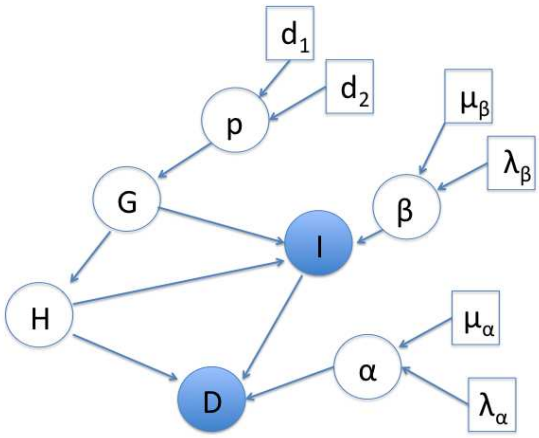

Figure 2: Bayesian DAG depicting variables in MCMC model. Circles are variables, squares are hyperparameters. Shaded circles represent observed data, while unshaded circles are unknown (latent variables). G is the contact network, H is the subgraph of epidemic edges, I the vector of infection times, and D the matrix of genetic distances. α is the genetic mutation rate and β is the epidemic transmission rate.

The following Bayesian directed acyclic graph defines the conditional independence relationships between the variables in our model:

We use a Gibbs sampling scheme. To make the inference process more efficient and simpler, we use conjugate priors for our parameters p, β, α: since each edge of G is drawn random based on a $Bernoulli(p)$ distribution, we set a $Beta(d_1, d_2)$ prior on p. This allows us a straightforwardly draw p from the conditional distribution given G, by simply sampling p from a $Beta(d_1 + |G|, d_2 + \frac{n(n-1)}{2} - |G|)$ distribution, where $|G|$ is the number of edges in G. Similarly, since β and α are rate parameters for an exponential distribution, we place a Gamma conjugate prior with hyperparameters as shown in the diagram.

This gives us the proposal steps used for p, α, β; for G, we sample each edge independently based on the prior p (as well as the likelihood based on the observed H and I. For H, we make a proposal by selecting each patient in turn, and resampling their ancestor from among the set of earlier-infected nodes by computing the conditional probability of each node as ancestor and sampling from this distribution.

More details of the exact equations used in the MCMC sampling and explanations are given in Ap-

pendix 1.

Unfortunately, we could not use BUGS (which would have been more simpler and more efficient) to do the MCMC inference, due to the unavoidable combinatorial structure in the problem: the genetic distance matrices and the infection times I are defined in terms of shortest paths, which cannot be expressed easily expressed in the form of a series of equations defining the generative model. As such, we implemented the MCMC inference manually. A difficulty we faced was that the algorithm was fairly slow - as such, we were not always able to run the algorithm for long enough to be reasonably convinced that it had converged to the stationary distribution. As a result, the chains tended to get stuck in local optima, resulting in performance that was somewhat worse than expected (described in Results section).

2.4.5 Predicting p based on conditional mean

Although the simplest way to estimate p based on our MCMC chains would be to simply average the sampled values of p in the chain, we instead a different approach that decreases the variance of predicting p. Rather than predicting the average value of p, we predict the average value of $E(p|...)$, i.e. the conditional mean of p given the rest of the variables. This has less variance because the prediction no longer incorporates the unavoidable variance arising from the process of sampling p.

2.5 Evaluation

We evaluate our method using simulation - we simulate random contact networks from a random graph model with known parameter, and a random epidemic using this contact network, as well as genetic sequences arising from viral evolution as the virus transmits along this epidemic. We evaluate each of our approaches by using them to compute estimates $\hat{\theta}$ for θ, then compare the estimates to the true value using mean squared error and mean absolute error. Repeating this process multiple times for multiple values of θ will allow us to generate plots to compare our methods against one another.

As for the H1N1 data, since we do not have ground truth in terms of the contact network or its parameters, we cannot directly evaluate the model in the same way. However, we can indirectly evaluate the model by using our learned model to infer epidemic parameters such as the basic reproduction number R_0. For H1N1 influenza, R_0 is fairly well measured in the literature, estimated to be between 1.7 and 1.8 for the early part of the US epidemic (White et al., 2009), which is the period our data is drawn from. We can thus evaluate our method by using it to compute posterior predictions for R_0 and evaluating how well this agrees with the literature. Methods for estimating R_0 are also of great interest in their own right because R_0 is highly relevant to disease modeling and planning control measures such as vaccination.

3 Results

We first trained the prediction models, random forests and gradient boosting machines, on 400 independent simulations (with different values of p), and tested on 80 more simulations.

Each simulation involved 40 patients. The mean squared error (MSE) and mean absolute error (MAE) of predicting p are as follows:

Model	MSE	MAE
Random Forest	0.011	0.074
Gradient Boosting	0.010	0.071
MCMC	0.053	0.18

Table 1: Mean squared error (MSE) and mean absolute error (MAE) of predicting p using random forests, gradient boosting machines, and Markov Chain Monte Carlo

Figure 2 plots the predictions of each method (y-axis) against the true values of p (x-axis). The higher error of MCMC is partly due to lack of convergence (due to its high computational cost, we had to run the chains for relatively small number of iterations). At the same time, we note that MCMC has the advantages of simultaneously performing inference on all the unknown parameters, as well as the full epidemic history, rather than just predicting p.

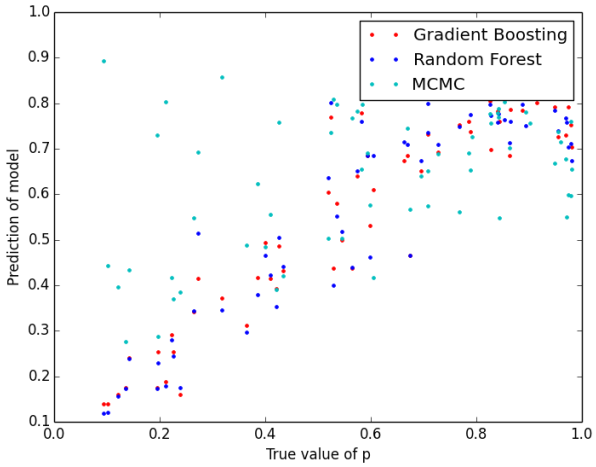

Figure 3: Predictions of random forest and gradient boosting machine models against true values when inferring the unknown parameter p in $G(n, p)$ random contact networks. Each point represents one simulation of 100 patients.

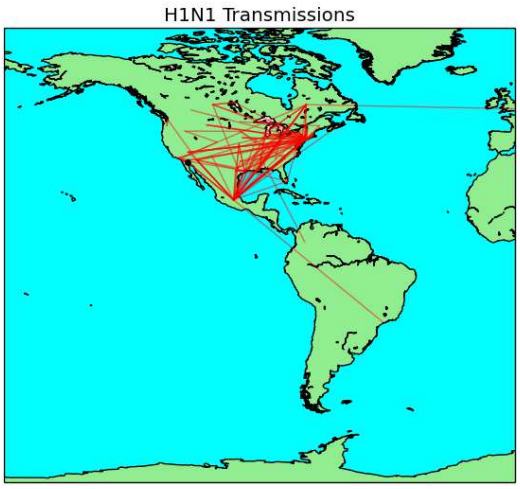

Figure 4: Transmissions inferred by the Markov chain Monte Carlo algorithm. The index case is shown as a black circle.

3.1 Variable Importance

Both the Random Forests and Gradient Boosting models allow us to evaluate the importance of each feature. In both cases, this works by measuring the importance of a feature based on the average decrease in squared error when we split the feature space according to that variable as part of the training process. A more detailed description of variable importance can be found in Hastie et al. (2005).

Model	F1	F2	F3	F4
Random Forest	0.12	0.27	0.28	0.32
Gradient Boosting	0.11	0.25	0.32	0.32

Table 2: Variable importances based on the two models for features. F1 and F2 are the infection time based features, while F3 and F4 are the genetic distance based features.

The max genetic distance between any two genetic sequences (F4) is the most important feature, while the sum of genetic distances (F3) is not far behind. The genetic distance-based features seem to be more predictive than the infection time-based features. This could be because the genetic distance-based features are able to capture the fact that two patients who were infected at around the same time may still be at a large distance away from each other (e.g. if the branches leading from patient 0 to these two patients separate very early), and this notion of distance is useful for predicting the overall number of edges in the graph. The infection time-based features, however, would only see that these two patients were infected at around the same time, without considering the actual graph-based distance between them.

3.2 H1N1 Influenza Analysis

We used the Markov chain Monte Carlo algorithm to jointly infer infer the most likely ancestors of each patient togethet with the unknown parameters (transmission rate β, genetic mutation rate α, contact network edge probability p). While the contact network is not an Erdos-Renyi graph, inferring p allows us to approximate the edge density in the graph.

The most likely ancestors for each patient define a series of inferred transmissions, plotted in Figure 3. They suggest that the disease started out in northern Mexico, then transmitted toward Mexico City, in which a sizable outbreak seems to have occurred. From there, it spread to a number of locations across the United States (as well as Brazil). We also observe a second large outbreak in New York, from which the disease again spread to many locations across the United States, particularly the northern parts, followed by spreading toward Europe and South America.

3.1.1 Estimation of Basic Reproductive Number R_0

Based on the output of the Markov Chain Monte Carlo algorithm, we can estimate the basic reproductive number R_0 of the epidemic.

Following Wallinga (2007), we estimate R_0 based on the exponential growth rate of the early part of the epidemic. Since there is no fixed method for choosing the 'window size' defining the early epidemic, we plot a curve displaying the estimated R_0 for each possible value of the window size.

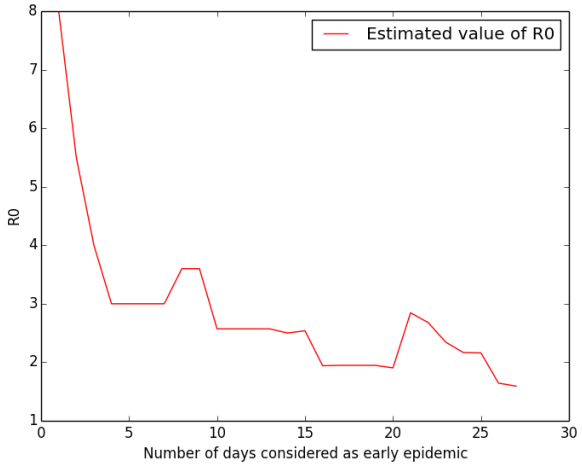

Figure 5: Estimated value of basic reproductive number R_0 against the choice of window size defining the early epidemic.

The figure shows that the estimated values generally range around 1.6 and 3.0, which, although it does not conflict with the values in the literature, are not particularly convincing due to their lasrge variance. This is probably due to our using only a single epidemic, which thereby contains only a small sample size in which patients are highly dependent. In contrast, the 1.7 to 1.8 value in the literature is based on a large number of outbreaks. Still, the fact that our method allows us to come up with a reasonable estimate for R_0 based only on genetic and infection time data suggests that extending our method to larger datasets is worthwhile.

Appendix 1

The following equations give the conditional distributions we use to sample each variable given the others. Let A_i be the ancestor of i, i.e. the node that infected node i. Let L be the length of each patient's genetic sequence. Let G_{ij} be 1 if edge i, j is in G, and 0 otherwise. Let D_{ij} be the genetic distance between node i and j, i.e. the number of differences between their genetic sequences.

$$
\begin{aligned}
P(p|...) &= Beta(d_1 + |G|, d_2 + \binom{n}{2} - |G|) \\
P(\beta|...) &= Gamma(\mu_\beta + n - 1, \lambda_\beta + \sum_{i,j \in G} |I_j - I_i|) \\
P(\alpha|...) &= Gamma(\mu_\alpha + \sum_{i,j \in H} D_{ij}, \lambda_\beta + \sum_{i,j \in H} |I_j - I_i|) \\
P(G_{ij} = 1|...) &\propto \begin{cases} 1 & \text{if } H_{ij} = 1 \\ \frac{p \exp(\beta |I_j - I_i|)}{(1-p) + p \exp(\beta |I_j - I_i|)} & \text{otherwise} \end{cases} \\
P(A_i = j|...) &\propto \begin{cases} (\frac{\alpha}{4L})^{D_{ij}} \exp(-\alpha |I_j - I_i|) & \text{if } I_j < I_i \\ 0 & \text{otherwise} \end{cases}
\end{aligned} \tag{2}
$$

We resample p, β and α based on the standard method of updating a Bernoulli variable with a Beta conjugate prior based on observed data, or Exponential variables with Gamma conjugate priors. The update for β comes from the fact that $\sum_{i,j \in G} |I_j - I_i|$ is the total time during which transmissions could have occurred, and $n - 1$ is the number of transmissions which actually occurred. Analogously, the update for α comes from the fact that $\sum_{i,j \in H} |I_j - I_i|$ is the total time during which genetic mutations could have occurred, and $\sum_{i,j \in H} D_{ij}$ is the number of genetic mutations which actually occurred.

We resample each edge G_{ij} independently as follows: if the edge is currently used for the epidemic, then its conditional probability is clearly 1; otherwise, the probability that the edge is present depends on its prior (p) multiplied by the likelihood based on the lack of a transmission that occurred along this edge, with transmission probability β and an elapsed time of $|I_j - I_i|$ in which a transmission could potentially have occured. As such, the probability of G_{ij} being present is proportional to $p \exp(\beta |I_j - I_i|)$, which we must normalize to get the exact probability.

We resample A_i (the ancestor of node i) as follows: if node j has a later infection time than i, the probability that j is the ancestor of i is 0. Otherwise, since all nodes have the same transmission rate β, there is no difference in the likelihood of each potential ancestor j transmitting to i. As such, the only remaining differences in likelihood between can-

didates for the ancestor of i come from the likelihood of the observed genetic sequences: intuitively, genetic sequences which are more similar to i's sequence are more likely to come from an ancestor of i.

Thus, we will evaluate the likelihoods of each potential ancestor j of i based on their genetic sequences. Since genetic mutation follows a Poisson process with mean α, the probability of having D_{ij} genetic differences between i and j is proportional to $\frac{\alpha^{D_{ij}} \exp(-\alpha|I_j - I_i|)}{(D_{ij})!}$, but this gives the combined likelihood of all possible genetic sequences with D_{ij} differences; since we have observed a particular one of these, we divide the Poisson probability by the number of sequences having D_{ij} genetic differences from i's genetic sequence. Since each mutation can occur in $4L$ ways and we want D_{ij} sequences but we ignore the order of these mutations, this number is $\frac{(4L)^{D_{ij}}}{(D_{ij})!}$ assuming the genetic sequence length L is large compared to the number of mutations. Dividing the Poisson probability by the number of sequences gives the result in (2). Intuitively L appears in the equation because the larger L is, the less likely it is that two mutations would occur at the exact same location, and thus the more importance the algorithm places on recovering ancestries that minimize the total number of genetic mutations. This is seen in equation (2) in the fact that the larger L is, the more the algorithm places weight on choosing an ancestry with low D_{ij}.

References

[1] Eytan Adar and Lada A Adamic. Tracking information epidemics in blogspace. In *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, pages 207–214. IEEE, 2005.

[2] RM Anderson and RM May. Infectious disease of humans. *Dynamics and control*, 1991.

[3] Dennis A Benson, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. Genbank. *Nucleic acids research*, 38(suppl 1):D46–D51, 2010.

[4] Tom Britton and Philip D. O'neill. Bayesian inference for stochastic epidemics in populations with random social structure. *Scandinavian Journal of Statistics*, 29(3):375–390, 2002.

[5] Bryan T Grenfell, Oliver G Pybus, Julia R Gog, James LN Wood, Janet M Daly, Jenny A Mumford, and Edward C Holmes. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, 303(5656):327–332, 2004.

[6] Chris Groendyke. *Inference for Social Networks Based on Epidemic Data*. PhD thesis, The Pennsylvania State University, 2011.

[7] Chris Groendyke, David Welch, and David R Hunter. Bayesian inference for contact networks given epidemic data. *Scandinavian Journal of Statistics*, 38(3):600–616, 2011.

[8] Chris Groendyke, David Welch, and David R. Hunter. A network-based analysis of the 1861 hagelloch measles data. *Biometrics*, 68(3):755–765, 2012.

[9] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

[10] T. Jombart, R. M. Eggo, P. J. Dodd, and F. Balloux. Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity*, 106(2):383–390, February 2011.

[11] William O Kermack and Anderson G McKendrick. Contributions to the mathematical theory of epidemics. ii. the problem of endemicity. *Proceedings of the Royal society of London. Series A*, 138(834):55–83, 1932.

[12] Marco J Morelli, Gaël Thébaud, Joël Chadœuf, Donald P King, Daniel T Haydon, and Samuel Soubeyrand. A bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS computational biology*, 8(11):e1002768, 2012.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[14] Jacco Wallinga and Marc Lipsitch. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences*, 274(1609):599–604, 2007.

[15] Laura Forsberg White, Jacco Wallinga, Lyn Finelli, Carrie Reed, Steven Riley, Marc Lipsitch, and Marcello Pagano. Estimation of the reproductive number

and the serial interval in early phase of the 2009 influenza a/h1n1 pandemic in the usa. *Influenza and Other Respiratory Viruses*, 3(6):267–276, 2009.

[16] Jaewon Yang and Jure Leskovec. Modeling information diffusion in implicit networks. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 599–608. IEEE, 2010.

[17] RJF Ypma, AMA Bataille, A Stegeman, G Koch, J Wallinga, and WM Van Ballegooijen. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proceedings of the Royal Society B: Biological Sciences*, 279(1728):444–450, 2012.