Adding a spatial element to Outbreaker

January 20, 2015

1 Introduction

1.1 Aims

The aim of this project is to see if group structure or contact structure data can be added to the Outbreaker model by NAME HERE to help us infer information about the transmission dynamics between different groups or to further improve the process of inferring the most likely transmission tree for an outbreak. In certain outbreak scenarios it may be possible to separate the cases of infection into two or more distinct groups, the groups could be defined by what ward a patient is on in a hospital, what class a child is in at school or perhaps whether a case is a child or an adult. An interesting question in this scenario might be whether individuals in the same group are more likely to infect each other and less likely to infect individuals in the other group. We will aim to add a framework to the outbreaker model that allows further outbreak parameters to be estimated which may help us answer this question. Additionally, we will see how we can incorporate contract tracing data into outbreaker and how this can be used to help the statistical inference. In this first section I will introduce the Outbreaker package and go through the previous body of work which led up to the development of Outbreaker.

The Outbreaker package uses a Markov chain Monte Carlo process to try and sample from the posterior probability distribution of various parameters and pieces of augmented data. Specifically, it uses DNA sequence data and case data collected from an outbreak along with a generation time distribution and a time-to-collection distribution to infer the likely infector of each case. These likely ancestors can be combined into a transmission tree and if we are fairly confident in the truth of our assembled transmission tree then we can infer further properties about the outbreak from it such as the rate of mutation of nucleotides in the DNA sequence of the pathogen and the effective reproduction numbers of individuals through time.

1.2 A Brief History

The Outbreaker model builds upon previous methods that use a similar process of assigning a likelihood value to transmission trees (out of potentially thousands

of configurations) and then searching for the tree with the maximum likelihood value or using the likelihood of a tree to sample from the posterior distribution of the trees given the data we have collected. The earliest implementation of this approach was Haydon et al, 2002, who proposed a likelihood function for transmission trees that describe the transmission of foot-and-mouth disease between farms in the UK. Their likelihood function for each transmission event is a product of two independent terms. The first term gives a likelihood value based upon whether the timings of the infection between two farms match well. We can assign a value based upon how well the period farm A was infectious during overlaps with the predicted time period that farm B was infected during. The better these time periods overlap, the more likely we think that this infection event is. The second term gives a likelihood value based on how far apart the two farms are, seeing as animals on different farms do not freely mix (especially during an outbreak) then there is only the possibility of an aerial infection which is more plausible the closer two farms are. For each tree we can consider the likelihood of all of the separate events in the tree and come to an overall likelihood for the tree, Haydon et al proposed an algorithm which would work towards producing the transmission tree with the highest overall likelihood.

As genetic sequencing became faster and cheaper, Cottam et al could expand this model by including a genetic likelihood term. Now that most infected cases could also provide a DNA sequence of the pathogen then the DNA sequences could be compared to see if they can help figure out the ancestor of each case. Infectious diseases mutate quickly and therefore mutations in the DNA sequences can occur during one generation. We can compare the DNA sequences and produce a likelihood that one case is the ancestor of another case depending on how similar the two sampled DNA sequences are, the more similar two sequences are the more likely that one is an ancestor of the other. Cottam et al proceeded by selecting the transmission tree configurations which had the highest genetic likelihoods and then using the previous epidemiological likelihood (based on infection and collection times) to choose 4 final trees which accounted for 95% of the sum of the likelihood for every possible tree. Further alterations to the approach were made by Ypma et al ,2012, who combined the genetic and epidemiological likelihoods into a single term, therefore removing the assumption that the two likelihoods are independent. This is an important assumption to consider because more mutations will occur in a DNA sequence over longer periods of time so we expect that there will be some correlation between the generation times and the number of mutations that we find.

In the same year Morelli et al (2012) used a Markov Chain Monte Carlo (MCMC) approach to sample from the posterior distribution of transmission trees given the collected data. This technique begins with a tree and then moves the suspected ancestors of each case around according to some rules to form a new tree. The likelihood of both of these trees are calculated and the new tree is accepted a sample with a probability based on the ratio of the two likelihoods. The chain and the movement rules are constructed so that the probability of accepting a tree as a sample is equal to the posterior probability of the tree given the data. We can then look at the trees with the highest

posterior probabilities or consider the posterior probability that one case is an ancestor of another.

The Outbreaker model is a combination of these approaches, it uses an MCMC approach with independent genetic and epidemiological likelihoods. It also allows for unsampled cases to occur between two cases and a more complex account of the DNA sequence mutations. Unlike the previous approaches it can also be generalised into a package for the programming language R which means it can be run on personal computers by people with less technical computing skills within a reasonable amount of time. In the next section I will briefly introduce Markov Chain Monte Carlo processes before describing the Outbreaker process in more detail.

2 The Outbreaker Model

2.1 Markov Chain Monte Carlo Processes

Markov Chain Monte Carlo processes are a combination of two statistical tools, the easiest of the two is Monte Carlo methods. Monte Carlo methods aim to approximate values such as the expected value of a probability distribution by using a large number of samples from that distribution. To work out the expected value of a probability distribution analytically we would integrate over every possible value in the distribution multiplied by the probability of it occurring. Using Monte Carlo methods we would approximate this integration by sampling from the probability distribution thousands of times and taking the mean average of the results. Put simply, Monte Carlo is a way of approximating values of interest given lots of the right samples.

Markov Chains play the role of providing these samples, a Markov Chain is a chain of subsequent states where the next state in the chain is decided by probabilities which are only determined by the current state. A simple discrete Markov Chain may have 3 states called A, B and C, if we are currently in state A then whatever state we move to next only depends on the probability of moving from state A to states B or C. If I move to state B then the fact that I was just previously in state A does not play a role in my decision making about what state I will go to next. If we run certain Markov Chains for long periods of time we will find that they settle into a stationary distribution which determines what the probability of the chain being in each state at each time is. If we have a stationary distribution of interest, such as $p(X = A) = p(X = B) = p(X = C) = \frac{1}{3}$ then we can devise a Markov Chain that will have this distribution as its stationary distribution, run it for a while until it converges upon this steady state, and then use the values in the Markov Chain as samples from our distribution of interest.

We can use the Metropolis-Hastings algorithm to easily find Markov Chains that have a specified stationary distribution, therefore our stationary distribution could be very complex and we could still use a fairly simple Markov Chain to sample from it. We can then use these stationary distribution samples

in our Monte Carlo methods to make approximations about the distribution. Metropolis-Hastings can also be used in a Bayesian setting by specifying the stationary distribution as a posterior distribution of interest. This means that instead of having to find a posterior distribution analytically we can instead use an MCMC process to sample from it and then make inferences about it. This is what the Outbreaker model does in the specific context of finding the posterior distribution of transmission trees when given outbreak data.

2.2 Outbreaker Model