Mathematisches Forschungsinstitut Oberwolfach

Report No. 7/2018

DOI: 10.4171/OWR/2018/7

Design and Analysis of Infectious Disease Studies

Organised by
Martin Eichner, Tübingen
M. Elizabeth Halloran, Seattle
Philip O'Neill, Nottingham

18 February – 24 February 2018

Abstract. This was the fifth workshop on mathematical and statistical methods on the transmission of infectious diseases. Building on epidemiologic models which were the subject of earlier workshops, this workshop concentrated on disentangling who infected whom by analysing high-resolution genomic data of pathogens which were routinely collected during disease outbreaks. Following the trail of the small mutations which continuously occur in different places of the pathogens' genomes, mathematical tools and computational algorithms were used to reconstruct transmission trees and contact networks.

Mathematics Subject Classification (2010): 05C05, 05C12, 37E25, 37N25, 62-07, 62F15, 62H30, 62M05, 62M09, 62N01, 62N05, 62P10, 92-04, 92C60, 92D20, 92D30.

Introduction by the Organisers

The workshop *Design and Analysis of Infectious Disease Studies*, organized by Martin Eichner (Tübingen), M. Elizabeth Halloran (Seattle, USA) and Philip O'Neill (Nottingham, UK), was well attended with 50 participants with broad geographic representation. The participants came from Australia, New Zealand, Singapore, USA, Brazil, and several countries in Europe, including the UK, Germany, Sweden, Denmark, Finland, Italy, Belgium, and the Netherlands. Fourteen of the 50 participants were women. Over 20 of the participants were at MFO for the first time. Professor Klaus Dietz of Tübingen also attended the meeting on Thursday. Professor Dietz was the original organizer of the precursor of this meeting. One of the planned participants, Professor John Edmunds of the London School of Hygiene and Tropical Medicine, was invited to Buckingham Palace to receive a prize in the name of the school from Queen Elizabeth II for work done during the Ebola outbreak in West Africa. Thus, he did not attend the MFO workshop this time with great regret.

The focus of the workshop was on integrating genomic data on pathogens with dynamic epidemiological analysis of infectious disease data either in the endemic or outbreak setting. This is a particularly exciting and challenging area for the analysis of infectious disease data. Now that sequencing the RNA or DNA of viruses, bacteria and other pathogens has become very inexpensive, such data are being obtained from most field studies of infectious diseases. This type of data and evolutionary analysis can contribute a lot to determining who infected whom. Such insight can contribute greatly to public health interventions. The analysis of such data poses statistical, mathematical, theoretical, and computational challenges all at the same time.

There were 21 talks of about one hour length, including discussion. There were five talks by the OWLG recipients in one session, each of 12 minutes length. Some of the talks spoke more about the statistical models that were being developed to do such analyses. Other talks dealt with details of computational algorithms. In one talk, the speaker said that once they had set up the analysis, the computation ran for two months. In a final talk during the last discussion, a young colleague who is trained as a pure mathematician presented a chalk talk on her newest results on a new metric on distance of transmission trees that she has proposed. All talks on these related subjects produced active discussions during and after the talks. Other topics included the relevance of social contact patterns for spread of infectious diseases, survival analysis of observational data, and disease burden in transient situations.

There was much discussion in the breaks and in the free periods. A few of the important exchanges are reported here. In one talk, the speaker said that he was using a greedy algorithm to solve the problem of the graph among contacts where a disease was spreading. Two of the participants recognized that this was an example of arborescence in graph theory. Thus, there is a theory allowing for a more general solution to the problem. One of the colleagues has been heading up a method to combine genomic data with transmission analysis. She has developed a software to do this analysis. One of the participants was able to convince her that his method using survival analysis to formulate the problem was a more principled approach. She is now looking into integrating his approach into her software. Other topics that elicited considerable debate were forward versus backward simulation, whether multi-scale approximations are needed, and to what extent graph theory is applicable to the current problems.

A sign of the excitement and growth in this field is that many speakers announced a number of available postdoctoral positions. The number ranged from one to five. Some participants who were not giving talks also announced available postdoctoral positions. Thus the field is growing enormously. There was some discussion on the side about how to increase the pipeline of students entering the field.

One initiative to come out of the meeting is a plan to write a paper on challenges of infectious disease modeling. A one hour discussion of interested parties included about half of the workshop participants. The opportunity to participate in this paper is open to all participants. A dropbox has been set up to coordinate the writing of the paper. The role of the workshop at MFO will be credited in the paper.

Of potential interest was the participation of Dr. Michelle Kendall, a young mathematician who has just moved from Imperial College to Oxford University. She is married to Dr. Ed Kendall, who is the son of Wilfrid Kendall, and grandson of David Kendall, both of whom are (were) well-known mathematicians, who also had been at MFO. Dr. Michelle Kendall attended the workshop with her husband and her 15-month old daughter Sophia. Thus, this was a fourth-generation presence at MFO for the Kendalls. Professor David Kendall was also the PhD advisor of Denis Mollison, one of the other participants of this workshop, who is also the workshop photographer.

Due to a storm in January that had knocked down a number of trees in the forest, we were discouraged by the MFO administration to take the usual Wednesday afternoon hike to St. Roman. Instead the official hike went to the Museum für Mineralien and Mathematik, an hour's walk from MFO. A guided tour of the museum was provided by Prof. Dr. Stephan Klaus, Scientific Administrator of MFO. Those who went also had Schwarzwälderkirschtorte nearby. About half the participants took the unofficial hike through the forest to St. Roman.

On Thursday evening, we had our usual musical talent show and cultural event in the lovely music room available at MFO. Dr. Lorenzo Pellis organized the program. The evening's program is presented below.

Acknowledgement: The MFO and the workshop organizers would like to thank the National Science Foundation for supporting the participation of junior researchers in the workshop by the grant DMS-1641185, "US Junior Oberwolfach Fellows". Moreover, the MFO and the workshop organizers would like to thank the Simons Foundation for supporting Mick Roberts in the "Simons Visiting Professors" program at the MFO.

MFO Talent show, February 22, 2018, 8:00 pm

Kari Auranen (piano)	**Einojuhani Rautavaara**	Two excerpts from: Fiddlers (op. 1)
Denis Mollison (voice)	**W. H. Auden**	September 1, 1939
Lorenzo Pellis (flute) Michiel van Boven (piano)	**Carlo Gounod** **Johann Sebastian Bach**	Celebre Ave Maria
Lorenzo Pellis (flute) Michiel van Boven (piano)	**Johann Sebastian Bach**	Preludio I from: Das Wohltemperierte Klavier
Mick Roberts (voice)	**Tim Upperton**	The truth about Palmerston North
Denis Mollison (voice) Lorenzo Pellis (flute) Kari Auranen (piano)	**Antonio Vivaldi**	Sol da te mio dolce amore from: Orlando Furioso
Lorenzo Pellis (flute) Jason Xu (piano)	**Antonio Vivaldi**	Largo from: The four seasons: Winter
Mick Roberts (voice)	**Margaret Mahy**	Down the back of the chair
Lorenzo Pellis (flute) Caroline Colijn (piano)	**Telemann**	Cantabile from: Essercizii Musici
Caroline Colijn (voice & piano)	**Kilgore Trout**	Bitter Sea
Kari Auranen Betz Halloran (4-handed piano)	**Igor Strawinsky**	Berceuse Finale from: Firebird
	Interval	
Lorenzo Pellis (flute) Kari Auranen (piano)	**Gaetano Donizetti**	Sonata for flute and piano
Markus Schwehm (video equipment)	**Markus Schwehm**	Migration: III
Chris Wymant (voice) Gavin Gibson (guitar)	**Renan Luce**	La lettre
Ira Longini (voice) Mirjam Kretzschmar (voice) Gavin Gibson (guitar)	**Lowell George** **Carlos Puebla**	Willin' Hasta Siempre Comandante
Blowin' In The Wind Ensemble (bottles) and the audience (voice)	**Leonard Lipton** **Peter Yarrow**	Puff the Magic Dragon
Swedish Midsummer Night Ensemble and everybody else	**Traditional**	Swedish contribution to international music

Workshop: Design and Analysis of Infectious Disease Studies

Table of Contents

Abstracts

Defining clusters, predicting missing cases, estimating and comparing transmission trees with genomic data

Caroline Colijn
(joint work with Yuanwei Xu and James Stimson)

There has been much promise that sequencing pathogens in short-term outbreak situations will lead to great improvements in our understanding of transmission. Knowing who is transmitting, what variants are transmitted, when and where they are transmitting and covariates associated with transmission can inform improved control policies. When pathogens accrue genetic variation as they spread from host to host, this information can be used to infer transmission events. However, the relationship between who infected whom and the pathogen sequences is not straightforward. Variation in the case timing, complexities in the pathogen mutation process, our ability to detect genetic variation, pathogen populations inside hosts and other factors mean that it is not possible to directly obtain transmission events from either pathogen sequences or phylogenetic trees.

One first step in this domain of genetic epidemiology is to group sequences into sets (called clusters) of sequences that are thought to have originated from the same individual (the source of the outbreak). Currently, many researchers use a cutoff: if two sequences s_i and s_j differ at fewer than K sites, an edge is made between i and j; the clusters are the connected components of the resulting graph. In tuberculosis, the cutoff value K has ranged from 2-3 to 50-100. This approach is simplistic, and does not account for the mutation process of the pathogen, the generation time between infections, selection (eg drug resistance), in-host diversity and variation in bioinformatics pipelines.

We propose an alternative which builds some of these into account. We proceed as follows: given sequences s_i and s_j (from hosts i and j), we estimate the likelihood of the height h of a two-tip tree (in units of time) under a sequence evolution process λ. We then estimate the probability that m transmissions occurred in the total evolutionary time separating i and j. For this we use a transmission process β; the simplest approach is for λ and β to be Poisson processes with constant rates. We then connect s_i and s_j if the probability that there were more than m transmissions between i and j is smaller than a stated threshold p. We extend the process λ to model loci under selection evolving at a higher rate than loci not under selection.

Given a cluster c, we can construct a phylogenetic tree T_c using its sequences. The next question is how these can be analysed to estimate who infected whom and when. There are a number of statistical tools for this problem, including our previous method TransPhylo [1]. TransPhylo uses an MCMC approach, starting with a timed phylogenetic tree T reconstructed from the sequences, and augmenting this tree with each host's infection time. It handles in-host diversity and unsampled cases. Here we extend TransPhylo to proceed simultaneously on a *set* of clusters, sharing parameters across them all. This allows us to compare clusters and highlight those with high posterior numbers of unsampled cases.

At this stage, we have many posterior transmission trees for each cluster. It is challenging to check whether this posterior is unimodal and to summarise it. We present a metric on transmission trees [2]. The metric $d(T_1, T_2)$ is the usual Euclidean distance between two vectors v_1 and v_2. These vectors contain the number of steps between the source of the outbreak and the most recent common infector of cases k and n, for all pairs k, n in a cluster. We illustrate how the metric allows visualisation of posterior sets of transmission trees, and allows us to select geometric median trees.

References

[1] X. Didelot, C. Fraser, J. Gardy, C. Colijn, *Genomic Infectious Disease Epidemiology in Partially Sampled and Ongoing Outbreaks*, Molecular Biology and Evolution, **34** (2017), 997-1007.

[2] M. Kendall, D. Ayabina, Y. Xu, J. Stimson, C. Colijn. *Estimating Transmission from Genetic and Epidemiological Data: A Metric to Compare Transmission Trees*, Statistical Science, **33** (2018), 70-85.

Fitting stochastic epidemic models to incidence time series and gene genealogies

Vladimir N. Minin

(joint work with Jon Fintzi, Jon Wakefield, Kari Auranen, Mingwei Tang, Trevor Bedford and Gytis Dudas)

Introduction

Stochastic epidemic models describe how infectious diseases spread through a population of interest. These models are constructed by first assigning individuals to compartments (e.g., susceptible, infectious, and recovered) and then defining a stochastic process that governs the evolution of sizes of these compartments through time. Here, we propose a new strategy for fitting these models to data, which turns out to be a challenging task. The main difficulty is that even the most vigilant infectious disease surveillance programs offer only noisy snapshots of the number of infected individuals in the population. We present a Bayesian data augmentation strategy that makes statistical inference with stochastic epidemic models computationally tractable. Besides standard incidence data, our approach can also handle more exotic data types, such as genealogies/phylogenies of infectious disease agent genetic sequences collected during outbreak monitoring. We present results of using our new approach to fit stochastic epidemic models to data from outbreaks of influenza and Ebola viruses.

Hidden stochastic epidemic models

Let $\mathbf{Y}$ represent data collected during or after an infectious disease outbreak. In this work, we will assume that $\mathbf{Y}$ contains either underreported incidence data or a genealogy of infectious disease molecular sequences collected from a sample of infectious hosts. We assume a Markov stochastic model that divides the population into finite number compartments (e.g., susceptible, infectious, removed) and governed by parameters $\boldsymbol{\theta}$ (e.g., infectivity and recovery rates). Let $\mathbf{X} = (X_{t_0}, \dots, X_{t_n})$ be the compartment sizes recorded at times $t_0, \dots, t_n$, where these times can be times at which incidence data are collected or a suitable regular grid of time points when we use a pathogen genealogy as data. We are interested in the posterior distribution

$$\Pr(\boldsymbol{\theta} \mid \mathbf{Y}) \propto \Pr(\mathbf{Y} \mid \boldsymbol{\theta})\Pr(\boldsymbol{\theta}),$$

where

$$\Pr(\mathbf{Y} \mid \boldsymbol{\theta}) = \sum_{\mathbf{X}} \left(\Pr(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\theta}) \left[\Pr(\mathbf{X}_{t_0} \mid \boldsymbol{\theta}) \prod_{l=1}^{n} p(\mathbf{X}_{t_l} \mid \mathbf{X}_{t_{l-1}}, \boldsymbol{\theta}) \right] \right)$$

is the observed data likelihood and $\Pr(\boldsymbol{\theta})$ is the prior density of model parameters. The likelihood above is computationally intractable, because the state space of $\mathbf{X}_t$ is too large even for moderately high population size N.

Linear noise approximation

To overcome the likelihood intractability, we first use standard Bayesian data augmentation and develop a Markov chain Monte Carlo (MCMC) algorithm targeting the augmented posterior

$$\Pr(\boldsymbol{\theta}, \mathbf{X} \mid \mathbf{Y}) \propto \Pr(\mathbf{Y}, \mathbf{X} \mid \boldsymbol{\theta})\Pr(\boldsymbol{\theta}),$$

where

$$\Pr(\mathbf{Y}, \mathbf{X} \mid \boldsymbol{\theta}) = \Pr(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\theta}) \left[\Pr(\mathbf{X}_{t_0} \mid \boldsymbol{\theta}) \prod_{l=1}^{n} p(\mathbf{X}_{t_l} \mid \mathbf{X}_{t_{i-1}}, \boldsymbol{\theta}) \right].$$

Data augmentation itself does not solve the likelihood intractability problem, because transition densities $p(\mathbf{X}_{t_l} \mid \mathbf{X}_{t_{l-1}}, \boldsymbol{\theta})$ are computationally intractable for most stochastic epidemic models. We use linear noise approximation (LNA) to replace $p(\mathbf{X}_{t_l} \mid \mathbf{X}_{t_{l-1}}, \boldsymbol{\theta})$ with a suitable Gaussian density [2]. As a result, we are left with a latent Gaussian model with a non-Gaussian conditional density of the observed data. We extend the LNA approach to fitting stochastic epidemic models to incidence and genealogical data and equip it with a modern MCMC sampler — elliptical slice sampling algorithm.

Swine flu in Finland: inference from underreported incidence data

We use our LNA approach to fit an SEIR-type model to weekly incidence of mild flu cases from a national surveillance system from April 2009–2010 and weekly vaccination counts from a national campaign initiated in mid-October 2009. These data

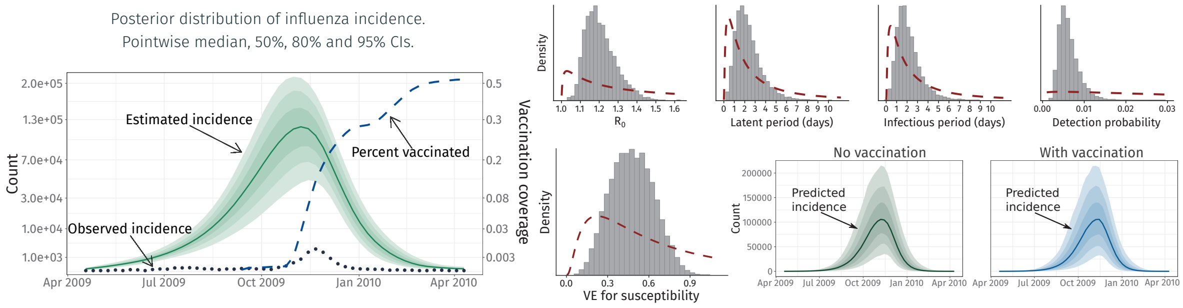

FIGURE 1. Results of fitting flu stochastic epidemic model to data.

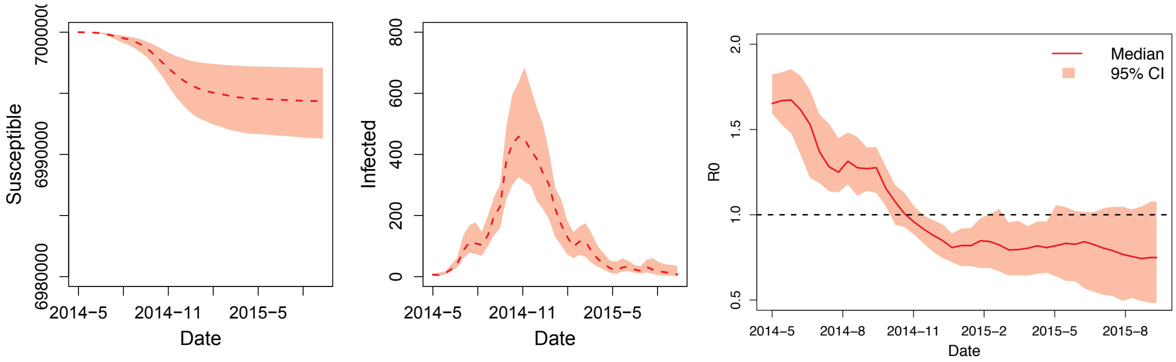

FIGURE 2. Results of fitting an SIR stochastic epidemic model to Ebola genealogy.

were originally analyzed in [4] using a discrete-time stochastic epidemic model. One of the main parameters of interest is the vaccine efficacy for susceptibility (VE), which we estimate by the posterior median of 0.48 with 95% credible interval (0.20, 0.78). According to our estimates, the vaccination campaign started when the population-level incidence was already high, so the effect of the campaign on the total number of infected individuals was minimal.

Ebola outbreak in West Africa: inference from inferred genealogy

Using the study of Ebola phylodynamics in [1] as a starting point, we reconstruct a genealogy of Ebola virus sequences sampled from Sierra Lione, where sequence data strongly indicate a single introduction of the virus into the country during the outbreak. We fit an SIR model with a time-varying effective reproduction number $R_0(t)$, which we estimate nonparametrically, and a constant recovery period. Our main finding is that the R_0 dropped below 1.0 by November 2014, probably due to interventions and change of behavior of individuals in the population.

References

[1] G. Dudas, L.M. Carvalho, T. Bedford, A.J. Tatem, G. Baele, N.R. Faria, D.J. Park, J.T. Ladner, A. Arias, D. Asogun, et al., *Virus genomes reveal factors that spread and sustained the ebola epidemic*, Nature, **544** (2017), 309–315.

[2] P. Fearnhead, V. Giagos, C Sherlock, *Inference for reaction networks using the linear noise approximation*, Biometrics, **70** (2014), 457–466.

[3] D.A. Rasmussen, O. Ratmann, K. Koelle, *Inference for nonlinear epidemiological models using genealogies and time series*, PLoS Computational Biology, **7** (2011), e1002136.

[4] M. Shubin, M.A. Lebedev, O. Lyytikäinen, K. Auranen, *Revealing the true incidence of pandemic A (H1N1) pdm09 influenza in Finland during the first two seasons—an analysis based on a dynamic transmission model*, PLoS Computational Biology, **12** (2016), e1004803.

[5] R.A. Smith, E.L. Ionides, A.A. King, *Infectious disease dynamics inferred from genetic data via sequential Monte Carlo*, Molecular Biology and Evolution, **34** (2017), 2065–2084.

[6] E.M. Volz, S.L. Kosakovsky Pond, M.J. Ward, A.J. Leigh Brown, S.D.W. Frost, *Phylodynamics of infectious disease epidemics*, Genetics, **183** (2009), 1421–1430.

Statistical inference of epidemiological parameters: what is the value of virus phylogenies?

Tom Britton
(joint work with Federica Giardina)

In the current project we are concerned with making inference for infectious disease outbreaks. Having HIV in mind, our focus lies in estimating the reproduction number (under current preventive measures) R_{curr}, the population size of the risk group N, and the fraction p_{diag} of cases that are diagnosed. We study several related epidemic models: SIR-closed with treatment, SI-open with treatment, SIR-open with treatment and SIS-open with treatment, where these models depend on if considering a short outbreak in a closed community, or over a longer period allowing for demography. For all models, a fraction p_{diag} of the infected cases are diagnosed and treated. We make the strong assumption of having a community of homogeneously mixing individuals that are also similar in terms of infectivity and susceptibility. The underlying reason for making this simplifying assumption is that more realistic assumptions will make the analysis even harder, and that when new HIV outbreaks are observed, the underlying structure of the community within which the virus has spread, is usually not well studied.

We study two different sources of data: either that we "only" observe the times of diagnosis, or that this information is available *and* that diagnosed individuals are sequenced (to be used for reconstructing the underlying virus phylogeny). For the latter case, we make the simplifying idealistic assumption that the virus phylogeny is reconstructed without any uncertainty, and that it coincides exactly with the transmission tree except that the direction of infection is lost. Clearly this is unrealistic, but will give us an upper bound on the precision of parameter estimates as compared to reality. We denote the two data sets considered by d-data and d&$\mathcal{G}$-data respectively, where $\mathcal{G}$ stands for geneaology or genetic data.

The inference is performed using MCMC and making use of the recursive method of Tanja Stadler for computing the likelihood of a partially sampled phylogeny.

The main conclusions of our analysis, obtained through simulation studies are:

– The precision of parameter estimates are more or less the same for the d-data and d&$\mathcal{G}$-data as well as for the different epidemic models. As a consequence, taking sequences and inferring the phylogeny is of little use (recall that we are assuming there are no heterogeneities).

– If the population at risk has known size N, then estimation of R_{curr} and p_{diag} is consistent, meaning that uncertainty decreases to 0 as N gets large. The same result holds true if instead p_{diag} is known and N is estimated.

– Estimation of both N and p_{diag} (as well as R_{curr}) is not feasible, meaning that a range of combinations of N and p_{diag} are consistent with both data sets. The two parameters are hence not separately identifiable.

A consequence of the results is that, if the population size N as well as the sampling fraction p_{diag} are unknown (as is common in e.g. HIV local outbreak situations), then traditional epi-data (d-data) or even epi-data together with the the virus phylogeny (d&$\mathcal{G}$-data) is not sufficient for disentangling the two.

If the population is heterogeneous in one or several ways (as is highly likely), we believe that some such heterogeneities may very well be inferred with higher precision for the d&$\mathcal{G}$-data as compared to the d-data. However, we do not believe this data will give consistent estimation of p_{diag} and N separately for this situation either.

This is work in progress.

R_0 for SIR epidemics in structured populations

Pieter Trapman
(joint work with Frank Ball, Lorenzo Pellis and Carolina Fransson)

For the analysis of SIR (Susceptible → Infectious → Recovered) epidemics in homogeneously mixing populations, the basic reproduction number R_0 is a key quantity for interpreting the potential for the disease to spread. R_0 is often defined as the expected number of infections caused by a typically infected individual, in a mostly susceptible population. R_0 may be interpreted as the offspring mean of an approximating branching process and because of that R_0 is a threshold parameter, in the sense that introduction of the infection leads to a large outbreak with positive probability if and only if $R_0 > 1$.

For epidemics in structured populations, such as a population partitioned in "households", in which individuals contact each other a higher intensity, it is often possible to find an approximating branching process at another level than that of individuals. E.g. for a household epidemic, one may see the first infected cases in households as the particles of a branching process. The offspring mean of this branching process is still a threshold parameter for the epidemic, but it can no longer be interpreted as the expected number of individuals a typical infected individual infects.

In this talk we consider ways of defining an approximating branching process for epidemics in two types of structured populations, from which we can find the expected number of individuals infected by a typically infected individuals.

For "household epidemics" we still consider only initial cases within a household as the particles in the approximating branching process. But now we keep track of the number of person to person transmissions that are needed to infect individuals represented by particles, by saying that the age of a mother particle at the birth of a child particle is this number of person to person transmissions needed to go from the initial infection in one household to the initial infection in a second household. For this branching process it is straightforward to compute the Malthusian parameter α and then to show that $R_0 = e^{\alpha}$.

A second model that we consider is a configuration model random graph with clustering. In this model individuals are represented by vertices and contacts are only possible along the edges of the graph. We say that each vertex has a random number (distributed as S) of "single stubs" assigned to it, and of a random number (distributed as T) of pairs of "triangle stubs". The random numbers assigned to different vertices are independent. We then create the edges by pairing the single stubs uniformly at random and create the triangles by uniformly grouping the pairs of triangle stubs into groups of three pairs of stubs.

Naive branching process approximations in the same spirit as for household-epidemics do not work for this model, because possibly random infectious periods create dependencies in the offspring of siblings and parents in the approximating process. However, by considering three-types of vertices: (i) vertices infected through single edges and through a triangle edge either (ii) with or (iii) not with the third vertex in the triangle still susceptible, we obtain a proper multi-type branching process. For this branching process we can compute the largest eigenvalue of the mean-offspring matrix, which corresponds to the expected number of infections caused by a typically infected individual.

References

[1] L. Pellis, F. Ball, P. Trapman, *Reproduction numbers for epidemic models with households and other social structures. I. Definition and calculation of R0*, Mathematical biosciences, **235.1** (2012), 85–97.

[2] F. Ball, L. Pellis, P. Trapman, *Reproduction numbers for epidemic models with households and other social structures II: comparisons and implications for vaccination*, Mathematical biosciences, **274** (2016), 108-139.

[3] C. Fransson, P. Trapman, *SIR epidemics and vaccination on random graphs with clustering*, arXiv preprint arXiv, (2018), 1802.05011.

phyloscanner: Automated phylogenetics with NGS reads from multiple hosts reveals transmission, multiple infection, recombination and contamination

Chris Wymant

(joint work with Matthew Hall, Oliver Ratmann, David Bonsall, Tanya Golubchik, Mariateresa de Cesare, Astrid Gall, Marion Cornelissen, Christophe Fraser, the STOP-HCV Consortium, the Maela Pneumococcal Collaboration and the BEEHIVE Collaboration)

Phylogenetics allows us to learn more about the spread of infectious disease by identifying individuals whose pathogens are closely related, from which we infer that the individuals are close in a chain of transmission. This identification of closely related pathogens requires pathogen sequences to be be determined accurately, so that small differences can be meaningfully interpreted. For human immunodeficiency virus (HIV), reconstructing whole genome sequences from the 'reads' (short fragments of sequence) produced by next-generation sequencing has been technically challenging. In particular, mapping (aligning) reads to a reference sequence leads to biased loss of information; this bias can distort epidemiological and evolutionary conclusions. *De novo* assembly avoids this bias by effectively aligning the reads to themselves, producing a set of sequences called contigs. However contigs provide only a partial summary of the reads, misassembly may result in their having an incorrect structure, and no information is available at parts of the genome where contigs could not be assembled.

We developed the tool `shiver` [2] to address these problems: pre-processing the reads for quality and contamination, then mapping them to a reference tailored to the sample using corrected contigs supplemented with the user's choice of existing reference sequences. On a test data set of 65 existing publicly available samples and 50 new samples from the BEEHIVE project (*Bridging the Evolution and Epidemiology of HIV in Europe*), the consensus sequence called using `shiver` was systematically superior to that derived by mapping the same reads to the closest of 3249 real references from the comprehensive Los Alamos National Laboratory database online (http://www.hiv.lanl.gov/). A median of 43 bases were called differently and more accurately (supported by higher coverage), at the cost of 1 base called differently and less accurately.

Having accurately reconstructed our HIV genomes, we turned our attention towards inference of transmission in our data and more generally. Most molecular epidemiological analyses have used only one pathogen sequence from each sampled infected host. Phylogenies of these sequences only show whose pathogens are closely related, not who has infected whom. It is possible to infer transmission direction from such a phylogeny, if a transmission model can be fitted to the phylogeny together with additional epidemiological data, such as estimated times of infection. This relies on the availability and accuracy of both the epidemiological data and the transmission model.

In reference [1] it was demonstrated that using phylogenies of multiple genotypes from each host infected with HIV-1, and ancestral reconstruction of the virus's host state, it is possible to infer the direction of transmission without additional epidemiological data. Our public software tool `phyloscanner` automates inference of phylogenies showing within- and between-host evolution at once. For next-generation sequencing data, phylogenies can constructed using reads in sliding windows along the whole genome. `phyloscanner` identifies and removes likely contaminant sequences, quantifies within-host diversity, identifies individuals infected by multiple strains, and finds signals of strains recombining. It performs ancestral reconstruction of the pathogen's host state, providing unprecedented resolution into the transmission process, allowing inference of the direction of transmission from sequence data alone. We illustrate `phyloscanner` on small illustrative datasets of HIV sequenced on Illumina and Roche 454 platforms, HCV sequenced with the Oxford Nanopore MinION platform, and *Streptococcus pneumoniae* with sequences from multiple colonies per individual.

The listener (reader) is referred to Oliver Ratmann's talk (abstract) for an application of `phyloscanner` to large-scale population data, and the resulting lessons learned about transmission patterns.

References

[1] E.O. Romero-Severson, I. Bulla, T. Leitner, *Phylogenetically resolving epidemiologic linkage*, Proc Natl Acad Sci USA, **113** (2016), 2690–2695.

[2] C. Wymant, F. Blanquart, T. Golubchik, A. Gall, M. Bakker, D. Bezemer, N.J. Croucher, M. Hall, M. Hillebregt, S.H. Ong, O. Ratmann, J. Albert, N. Bannert, J. Fellay, K. Fransen, A. Gourlay, M.K. Grabowski, B. Gunsenheimer-Bartmeyer, H.F. Gunthard, P. Kivela, R. Kouyos, I. Laeyendecker, K. Liitsola, L. Meyer, K. Porter, M. Ristola, A. van Sighem, B. Berkhout, M. Cornelissen, P. Kellam, P. Reiss, C. Fraser, The BEEHIVE Collaboration, *Easy and Accurate Reconstruction of Whole HIV Genomes from Short-Read Sequence Data with SHIVER*, Virus Evolution, (2018), in press.

[3] C. Wymant, M. Hall, O. Ratmann, D. Bonsall, T. Golubchik, M. de Cesare, A. Gall, M. Cornelissen, C. Fraser, STOP-HCV Consortium, The Maela Pneumococcal Collaboration, The BEEHIVE Collaboration, *PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity*, Molecular Biology and Evolution, (2017).

Combining transmission and evolutionary models to reconstruct trees with genomic data

Don Klinkenberg

(joint work with Xavier Didelot, Caroline Colijn, Jantien Backer and Jacco Wallinga)

Reconstructed outbreaks (who infected whom) help to understand transmission and guide future outbreak control. Pathogen sequence data can improve reconstruction, but within-host strain diversity can make it a challenge to take all uncertainty into account. We have developed a model for (neutral) pathogen evolution during infectious disease outbreaks, and developed an mcmc sampling scheme to do outbreak analysis using sampling times and sequences, i.e. to infer who infected whom, and when. The method is available as an R-package, phybreak [1]. In the package we implemented a variety of mcmc moves to traverse the tree space. The presentation in this workshop focused on the development and design of these moves.

The model in phybreak is a description of a transmission tree and associated phylogenetic tree. The current state of the tree is described by the observed sampling times $\mathbf{S}$ and the unobserved infection times $\mathbf{I}$, infectors $\mathbf{M}$, and phylogenetic tree P, which can be subdivided into minitrees P_i within each host i. The sampling times are associated with the observed sequences $\mathbf{G}$. The likelihood is a product of four terms, for the generation interval distribution, the sampling interval distribution, the within-host minitrees, and the mutation process on the phylogenetic tree.

The original design to propose alternative trees is based on the idea of picking one host, the focal host, removing its infector and infection time and proposing a new infection time and infector based on the sampling time of the focal host and current state of infection times of the other hosts. This is the active part of the move. Because this active part disrupts the minitrees in the focal host and the old and new infectors, it is followed by a reactive part of resimulating these minitrees.

This proposal was designed to focus on the transmission tree rewiring, so that it will be possible to accommodate additional epidemiological information in the proposal. That information could be censoring of infection times due to negative test results or hospital admission dates, or censoring of possible infectors because of information on location of patients, e.g. wards in a hospital.

It turned out that in some circumstances, this procedure resulted in poor mixing of the mcmc chain. Poor mixing is caused by frequent rejection of proposals, and rejection was in this case caused by proposing phylogenetic minitrees that resulting in a higher parsimony score of the complete phylogenetic tree, i.e. a larger number of mutations needed to explain the tree topology. Frequent rejection occurred if the data contained many SNPs (single nucleotide polymorphisms, i.e. differences between samples), or if the minitrees underwent too many changes. The latter especially occurred if hosts were sampled multiple times, or if the transmission bottleneck was allowed to be wide, leading to multiple parallel lineages moving between the hosts.

The first problem (many SNPs) was solved by designing proposals with changes to only a single minitree whilst keeping the transmission tree intact, or changes to the transmission tree keeping the phylogenetic tree intact. The latter starts with proposing a new infection time for the focal host, moving it on the phylogenetic tree, and checking if that results in a consistent new transmission tree. If not, it is tried to resolve the inconsistency by proposing a new infection time for another host. Because of the possibility of inconsistencies which have to be resolved, this procedure is less elegant than the initial design, but it did significantly improve mcmc mixing.

The second problem (too many topological changes in a single proposal) was tackled by redesigning the reactive part of the original proposal design: the resimulation of minitrees. First, in the old infector it appeared possible to only remove the lineages coming from the focal host, but keep the rest of the minitree intact. Second, in the new infector it appeared possible to keep the old minitree, and attach the new lineages coming from the focal host to that minitree through simulation. Thus, the only changes in the minitrees in the old and new infectors concerned the lineages coming from the focal host. Third, we redesigned the proposal for the minitree of the focal host itself, by keeping the topology of the minitree intact and only resimulating the coalescent times.

With the current suite of proposals for the transmission and phylogenetic trees, most mixing problems have been resolved. There is one exception, when all problems come together: if hosts have multiple samples and there is indication of a wide bottleneck, the proposal keeping the phylogenetic tree intact does not work. If in that case the data contain many SNPs, mixing remains less efficient. Apart from further work on improving mcmc mixing, future directions will cover the possibility to use additional epidemiological data, and to carry out analyses with host covariates, aiming at better understanding of infectivity and transmission and guide outbreak control.

References

[1] D. Klinkenberg, J.A. Backer, X. Didelot, C. Colijn, J. Wallinga, *Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks*, PLOS Computational Biology, **13** (2017), e1005495.

Assessment and refinement of epidemic and phylogenetic models

Gavin Gibson

(joint work with Max Lau, George Streftaris, Glenn Marion and Colin Worby)

A major challenge in modelling the dynamics of epidemics as they spread in space and time is that of fitting models to partial observations of the process and assessing the validity of the modelling assumptions in the light of the data. For example, in the case of the SEIR spatio-temporal models typically used to model the spread of arboreal pathogens such as citrus canker, it is of particular importance to understand the nature of the spatial kernel function $K(d;\kappa)$ which characterises the dependence of the infectious challenge presented by an infected individual to a susceptible on the distance d between the two. Beliefs about $K(d;\kappa)$ heavily influence the design of control strategies, for example based on removal of hosts within some specified radius of a newly discovered infection. Similar questions arise when these models are applied to human or veterinary pathogens such as FMD.

This talk explores the question of how knowledge of $K(d;\kappa)$ might be enhanced by the availability of genetic information on pathogen strains, in the form of sequence data, in addition to information on the location of infections. By including sequences of strains passed during transmission events, the unobserved transition times, and unobserved transmission graph as additional parameters within a data-augmented Bayesian analysis, it is possible to integrate the genetic and epidemic data. Using simulated spatio-temporal epidemics we show that the information in the posterior on the spatial kernel function and the transmission graph may be considerably enhanced by genetic information when the majority of infections are sampled. It should of course be noted that such enhancements are dependent on there being sufficient evolution of the pathogen for the genetic data to be informative. Results on this topic presented in the talk appear in [1].

A second topic described is that of assessing modelling assumptions. In particular, we have used a very simple model for pathogen molecular evolution that assumes that, at any time, there is a single dominant strain in the pathogen population. Within-host diversity of the pathogen population is therefore ignored. In ongoing work (with Max Lau and Colin Worby (Princeton U)) we adapt the latent-residual approach, previously used for example in [2], [3], [4]. Specifically we define a latent process of $U(0,1)$ i.i.d. random variables that can be imputed in the data-augmented MCMC and to which a classical test (here the Anderson-Darling test) of fit to the $U(0,1)$ can be applied. By appropriate construction of the residual process, a test that is sensitive to mis-specification of the genetic model can be tailored. Some preliminary results using simulated data (with correctly specified and mis-specified models for molecular evolution) are presented to illustrate the potential value of the approach. In a further enhancement, we suggest how 'marks' can be allocated to the imputed residuals such that selecting residuals with high values of the marks may further increase the sensitivity of the resulting tests.

References

[1] M.S.Y. Lau, G. Marion, G. Streftaris, G.J. Gibson, *A Systematic Bayesian Integration of Epidemiological and Genetic Data*, PLoS Comp. Bio., **11** (2015), e1004633.

[2] G.J. Gibson, W. Otten, J.A.N. Filipe, A. Cook, G. Marion, C.A. Gilligan, *Bayesian estimation for percolation models of disease spread in plant populations*, Statistics & Computing, **16** (2006), 391–402.

[3] M.S.Y. Lau, G. Marion, G. Streftaris, G.J. Gibson, *New model diagnostics for spatio-temporal systems in epidemiology and ecology*, J. Roy. Soc. Interface, **11** (2014), 20131093.

[4] G.J. Gibson, G. Streftaris, D. Thong, *Comparison and Assessment of Epidemic Models*, Statistical Science, **33** (2018), 19–33.

Reconstructing transmission trees for communicable diseases using densely sampled genetic data

Theodore Kypraios

(joint work with Colin Worby, Rosanna Cassidy, Ben Cooper and Philip O'Neill)

A fundamental aim in the analysis of infectious disease epidemics is to identify who infected whom, however, achieving this is challenging, since transmission dynamics are generally unobserved. A probabilistic estimation of the transmission tree based on all available data offers many potential benefits. In particular, this can lead to improved understanding of transmission dynamics, provide a mechanism to quantify factors associated with heightened transmissibility and susceptibility to carriage and infection, and help identify effective interventions to reduce transmission. Pathogen typing can be used to cluster genetically similar isolate samples, which can rule out potential transmission routes. Whole genome sequence (WGS) data offers maximal discriminatory power through the identification of individual point mutations, or single nucleotide polymorphisms (SNPs), potentially leading to more accurate transmission tree reconstructions than hitherto possible. However, the joint analysis of genetic and surveillance data poses several challenges, as the relationship between epidemic and evolutionary dynamics is complex.

We focus on individual-level transmission, using high-frequency genomic samples from a subpopulation (eg. hospital, school, jail, farm, community), with the aim of reconstructing transmission routes. We describe a generalized approach to transmission tree reconstruction that overcomes these limitations and makes use of both molecular typing information and known exposure data. A key novelty of our approach is that we model the genetic distances between sequences rather than the microevolution of the sequences themselves. This offers a flexible framework in which multiple independent introductions of the pathogen and within-host diversity may be considered, as well as the transmission process itself. This approach avoids the need to make any assumptions about the within-host pathogen population dynamics, which in general, are poorly understood. Furthermore, our proposed framework allows data to be simulated forward in time, a feature lacking in the majority of existing methods (with reverse time simulation typically required in phylogenetic methods, and only an incomplete set of genetic distances simulated from other approaches), which is of fundamental importance in predictive modelling and model evaluation.

A data augmented Markov chain Monte Carlo algorithm was used to sample over the transmission trees, providing a posterior probability for any given transmission route. We illustrate the predictive performance of our methodology using simulated data, demonstrating high sensitivity and specificity, particularly for rapidly mutating pathogens with low transmissibility. We present an analysis on data collected during an outbreak of methicillin-resistant *Staphylococcus aureus* in a hospital, identifying probable transmission routes and estimating epidemiological parameters. Our approach overcomes limitations of previous methods, providing a framework with the flexibility to allow for unobserved infection times, multiple independent introductions of the pathogen, and within-host genetic diversity, as well as allowing forward simulation.

References

[1] C.J. Worby, P.D. O'Neill, T. Kypraios, J.V. Robotham, D. De Angelis, E.J. Cartwright, S.J. Peacock, B.S. Cooper, *Reconstructing transmission trees for communicable diseases using densely sampled genetic data*, Ann Appl Stat, **10** (2016), 395–417.

Likelihood-based inference for dynamic systems, with phylodynamic applications

Edward L. Ionides

Sequential Monte Carlo (SMC) algorithms enable computation of the likelihood function for general partially observed Markov process (POMP) models. A POMP model consists of a latent Markov process observed via a collection of noisy measurements. A collection of independent POMP models with some shared parameters is called a PanelPOMP model. A POMP model for which the Markov process has a tree-valued structure appropriate for disease transmission modeling, and the measurements include genetic sequence data, we call a GenPOMP. We discuss advances in the theory and practice of inference for POMP, PanelPOMP and GenPOMP models. From a data analysis perspective, we demonstrate the pomp and panelPomp R packages and the genPomp C++ program. From a theoretical perspective, we discuss four theorems and one conjecture relevant to likelihood-based statistical inference via SMC for nonlinear POMP models in general, and PanelPOMP or GenPOMP models in particular.

Theorem 1. *Convergence of the IF1 algorithm* [2]
[4]. An iterated filtering algorithm perturbs the parameter vector of a POMP model, carries out a filtering operation via SMC, uses this SMC filtering to update the parameter vector, and then repeats with diminishing perturbations. The IF1 algorithm updates using a weighted average of filtered perturbed parameters.

Theorem 2. *Convergence of the IF2 algorithm* [5]. The IF2 algorithm has a similar structure to IF1, but rather than updating via a weighted average it uses the filtered perturbed parameter vector at the end of one filtering iteration as the start of the next. The convergence of IF2 is proved using entirely different theoretical techniques to IF1. In practice, the IF2 algorithm is found to be superior.

Theorem 3. *Convergence of the PIF algorithm* [1]. An extension of the IF2 algorithm to PanelPOMP models is called the PIF algorithm. PIF inherits its convergence theory from that of IF2.

Theorem 4. *Convergence of the GenSMC algorithm* [8]. Various computational techniues were developed to obtain a computationally tractable variant of SMC for GenPOMP models, which we call GenSMC. We show that asymptotic guaranteed for GenSMC are provided by the general theory of SMC.

Conjecture 1. *Monte Carlo adjusted profile confidence intervals* [3]. For large and complex models, such as PanelPOMPs and GenPOMPs, iterated filtering maximization and SMC evaluation of the likelihood function has considerable Monte Carlo error. We present methodology to obtain proper confidence intervals despite this Monte Carlo uncertainty. The methodology is derived heuristically, and shown to work by a simulation study. We therefore conjecture that it enjoys asymptotic theoretical guarantees.

References

[1] C. Bretó, E.L. Ionides, A.A. King, *Panel data analysis via mechanistic models*, Arxiv (2018), 1801.05695.

[2] E.L. Ionides, C. Bretó, A.A. King, *Inference for nonlinear dynamical systems*, Proceedings of the National Academy of Sciences of the USA, **103** (2006), 18438–18443.

[3] E.L. Ionides, C. Breto, J. Park, R.A. Smith, A.A. King, *Monte Carlo profile confidence intervals for dynamic systems*, Journal of the Royal Society Interface, **14** (2017), 1–10.

[4] E.L. Ionides, A. Bhadra, Y. Atchadé, A.A. King, *Iterated filtering*, Annals of Statistics, **39** (2015), 1776–1802.

[5] E.L. Ionides, D. Nguyen, Y. Atchadé, S. Stoev, A.A. King, *Inference for dynamic and latent variable models via iterated, perturbed Bayes maps*, Proceedings of the National Academy of Sciences of the USA, **112** (2015), 719–724.

[6] A.A. King, E.L. Ionides, M. Pascual, M.J. Bouma, *Inapparent infections and cholera dynamics*, Nature, **454** (2008), 877–880.

[7] A.A. King, D. Nguyen, E.L. Ionides, *Statistical inference for partially observed Markov processes via the R package pomp*, Journal of Statistical Software, **69** (2016), 1–43.

[8] R.A. Smith, E.L. Ionides, A.A. King, *Infectious disease dynamics inferred from genetic data via sequential Monte Carlo*, Molecular Biology and Evolution, **34** (2017), 2065–2084.

Reconstructing HIV-1 Transmission Networks from deep sequence data in an African setting: Are fishing communities the predominant source of new HIV infections in Rakai, Uganda?

Oliver Ratmann

(joint work with Kate Grabowski, Matthew Hall, Tanya Golubchik, Chris Wymant, Joseph Kagaayi, Godfrey Kigozi, Thomas Quinn, Maria Wawer, Oliver Laeyendecker, David Serwadda, Ronald Gray, Christophe Fraser on behalf of the PANGEA consortium and the Rakai Health Sciences Program)

Background and Methods

Targeting combination HIV prevention (CHP) to areas of high HIV prevalence is considered a cost-efficient and essential strategy for reducing HIV incidence in sub-Saharan Africa. Since 2014, the Ugandan National Antiretroviral Therapy Guidelines recommend targeted CHP to fishing communities on Lake Victoria, with an estimated HIV prevalence 25%-40%, partly based on the assumption that fishing sites are a major source of HIV transmissions to the inland populations; however the validity of this assumption has not been empirically evaluated.

Between August 2011 and Oct 2014, individuals aged 15-49 years in 40 communities in Rakai District, Uganda, were tested for HIV and interviewed (sociodemographic, behavioral and health information). Households were geocoded, and communities were classified as Lake Victoria fishing (n=4), agrarian (n=27), or main road trading (n=9) communities. Viral RNA from newly diagnosed participants was deep sequenced via Illumina instruments.

Here, we present and validate the phyloscanner software package for viral phylogenetic analysis from deep sequencing output in African settings [1]. The method is available at https://github.com/BDI-pathogens/phyloscanner. The input for the software package are viral read fragments that were mapped across the HIV genome for each individual, and are typically relatively short, 250bp. Owing to the deeper resolution of this approach into HIV-1 quasi-species within individuals, we hypothesised that deep sequencing data could improve phylogenetic inference of transmission events and the direction of transmission at the population level. We tested this hypothesis on individuals from Rakai District Uganda that could be deep sequenced. After validation, the software was then used to reconstruct HIV transmission networks, and to infer the direction of HIV transmission from deep sequence data. Finally, transmission flows between fishing sites and agrarian/trading communities were estimated with Bayesian multi-level models.

RESULTS

Of 23,719 individuals surveyed, 6205 were HIV-positive, 4309 (69%) were antiretroviral naive, of whom 2,803 (65%) were sequenced. 359 phylogenetically likely transmission events involving 676 individuals were reconstructed, with an estimated, expected 16% [11%-23%] of false reconstructions. Direction of transmission could be inferred in 241 likely transmission events, with an estimated, expected 14% [7%-26%] of false directions. Only 3/241 transmission events occurred from fishing sites to agrarian/trading communities. Adjusting for differences in participation and sequence sampling by age and community, an estimated 34.3% [28.6%-40.5%] of transmissions occurred within fishing sites, 58.0% [51.2%-64.6%] within agrarian/trading communities, 3.4% [1.7%-6%] from fishing sites to agrarian/trading communities, and 4.0% [2%-7.2%] from agrarian/trading communities to fishing sites.

CONCLUSIONS

Deep sequencing data provides unprecedented opportunities for characterising HIV-1 transmission at the population-level, though cannot prove transmission at the individual-level. HIV is infrequently transmitted from 4 high-prevalence fishing sites to the population in 36 agrarian/trading communities further inland, based on population-level NGS viral phylogenetic analysis. Our results suggest that targeted CHP to Lake Victoria fishing sites would not mitigate the broader HIV epidemic. Further studies in sub-Saharan Africa are needed to assess the strategy of targeting CHP to various high prevalence hotspots.

References

[1] C. Wymant, M. Hall, O. Ratmann, D. Bonsall, T. Golubchik, M. de Cesare, A. Gall, M. Cornelissen, C. Fraser, STOP-HCV Consortium, The Maela Pneumococcal Collaboration, The BEEHIVE Collaboration, *PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity*, Molecular Biology and Evolution, msx304 (2017), https://doi.org/10.1093/molbev/msx304.

Distinguishing introductions from local transmission

Simon Frost

Infectious diseases know no borders. While in principle, we could model the entire at-risk population, this presents problems due to potentially high heterogeneity, limitations on computational resources to simulate or fit the model, and limited data on which to base large-scale models. Consequently, we tend to consider smaller subpopulations, which necessitates considering the relative roles of introductions of infections from outside the study population versus local transmission. Introductions may occur either by movement of infected individuals (so-called 'distributed infectives') or by infection of individuals in the study population by individuals outside the study population ('distributed contacts'). Distinguishing introductions from local transmission is vital for obtaining accurate estimates of the basic reproductive number R_0, and critically affects the choice of intervention strategies.

Multiple data sources can be harnessed in order to tease apart introductions from local transmission. In some scenarios, case onset data can be used, especially when the infection is maintained by introductions. One such example is influenza A H7N9, which has a reservoir in poultry but is poorly adapted to transmisesion in humans. Kucharski et al. [5] employed a discrete-time model that estimated the rate of introductions. To test robustness of the results to model choice, I fitted a continuous time immigration-birth-death process using a partially observed Markov process (POMP, [4]) to three datasets of H7N9 from Shanghai, Jiangsu, and Zhejiang. In two out of three datasets, the methods were largely concordant; however, the inferred number of jumps was significantly higher using POMP than using the discrete-time model. A distinct advantage of the POMP approach is that more complex mechanistic models can be used. When applied to time series of measles cases from an epidemic that occurred in Wales in 2012/2013, there was strong evidence of continued introductions into schools, necessitating the development of models that couple the dynamics in different schools and in the general community.

Genetic data are particularly informative in endemic situations. In some cases, naïve approaches may suffice to count the number of introductions. For example, in a large scale analysis of influenza A H1N1 sequences from birds and swine, we found that there were three cross-species transmissions from birds to swine, with only one transmission leading to a substantial number of subsequent infections in swine [1].

In general, cross-species transmissions are harder to resolve using simple methods. Two widely used approaches, parsimony and the Mk model, that consider subpopulations as discrete traits that 'evolve' independently along branches of the phylogeny, are extremely sensitive to sampling bias. Such bias is extreme in many zoonotic infections, including MERS Coronavirus (in camels and humans), SARS Coronavirus (in civet cats and humans), and Lassa Fever Virus (in multimammate rats and humans). In these examples, humans are sampled earlier in time and more intensively than the reservoir species. Applying trait-based models to these data mistakenly infers a substantial rate of introductions *from* humans *to* the reservoir. If the direction of introductions is known *a priori*, it is possible to improve the performance of these methods, although a substantial number of samples from the reservoir may be needed in order to avoid dramatically underestimating the number of introductions. One such case where this may be satisfied is for HIV, where there are large number of publicly available sequences taken from many countries around the world, which can be treated as a source of introduction into a small subpopulation of interest. When applied to a dataset of nearly 3000 sequences from middle Tennessee from the United States, we infer almost 1000 separate introductions. Rarely, however, we find large clusters of cases that may represent ongoing chains of local transmission that have gone unnoticed against the backdrop of a fairly constant rate of new HIV infections.

In most cases, however, we will not know the direction of transmission beforehand. Structured coalescent models, in which the topology and branch lengths of the tree may be dependent on the subpopulation, are known to be more robust to sampling biases [2], but their widespread use has been limited to due dramatically increased computational demands compared to trait-based methods. We re-analyse a dataset of MERS Coronavirus, which transmits from camels and humans, that had previously been fitted using a standard structured coalescent model [3]. Using an approximate (pseudo)likelhood for an ‘island‘ model, we correctly infer the direction of transmission for zoonotic infections without making *a priori* assumptions, with the analysis taking seconds as opposed to months. However, the inferred ancestral states in the phylogeny are sensitive to the assumption of distributed infectives versus distributed contacts.

References

[1] V. Bourret, J. Lyall, S.D.W. Frost, A. Teillaud, C.A. Smith, S. Leclaire, J.Fu, S. Gandon, J.L. Guérin, L.S. Tiley, *Adaptation of avian influenza virus to a swine host*, Virus Evolution, **3** (2017), vex007.

[2] N. De Maio, C.H. Wu, K.M. O'Reilly, D. Wilson, *New routes to phylogeography: a Bayesian structured coalescent approximation*, PLoS Genetics, **11** (2015), e1005421.

[3] G. Dudas, L.M. Carvalho, A. Rambaut, T. Bedford, *MERS-CoV spillover at the camel-human interface*, eLife, **7** (2018), 7.

[4] A.A. King, D. Nguyen, E.L. Ionides, *Statistical inference for partially observed Markov processes via the R package pomp*, Journal of Statistical Software, **69** (2016), 1–43.

[5] A. Kucharski, H. Mills, A. Pinsent, C. Fraser, M. Van Kerkhove, C.A. Donnelly, S. Riley, *Distinguishing between reservoir exposure and human-to-human transmission for emerging pathogens using case onset data*, PLoS Currents, **6** (2014), 6.

Theoretical tactics to help sequence and serology data transform infection modeling

James S. Koopman
(joint work with Xinyu Zhang and Carl Simon)

We have shown that risk fluctuation at a frequency of months to years strongly alters the potential of universal test and treat strategies (T&T) or and/or pre-exposure prophylaxis (PrEP) and especially combined T&T and PrEP to eliminate HIV transmission. But there is little population level data that measures risk fluctuation. Therefore, using simulations, we show that HIV genome sequences from population samples of infected individuals contain information about how fluctuating risk behavior is generating high infection rates in men who have sex with men (MSM). Our analysis illustrates how methods that help develop theory about sequence pattern generation will be useful for addressing this and other dynamic process issues. New methods developed by Smith *et al.* allow direct fitting of models to sequence data [1]. These methods overcome pitfalls in making epidemiologic inferences from phylogenies. But to be fully informative, fitting models to sequence data should help understand the processes in the model that affect both the model behavior and the generation of sequence patterns. That is needed because validating any inference made from fitting a model to data requires that realistically relaxing simplifying assumptions does not change the inference. Since the ways to realistically relax simplifying assumptions are nearly limitless, a key step in validating model-based inferences is to develop theory that elucidates aspects of a model to which inferences may be most sensitive. It takes more than just following the fitting steps in Smith et al. to do this [1].

We illustrate an approach to opening up the black box of fitting procedures. First, we show that risk fluctuation affects the shapes of phylogenies in a unique manner not reproduced by varying other model parameters. To understand how this occurs, we postulate two mechanisms through which risk fluctuation generates effects on the basic reproduction number (R_0), on endemic prevalence, and on phylogeny shapes. Mechanism 1 brings susceptible individuals into contact with acutely infected individuals. Mechanism 2 reduces the high contact rates that led MSM to get infected so that they become less likely to transmit to others. Mechanism 1 raises prevalence but does not affect R_0. It has unique effects on tree shapes that reduce Sackin's index, raise cherry counts, and cause tree branches to form fewer but larger clusters from the tree root to the tree leaves. Mechanism 2, in contrast, raises Sackin's index, reduces cherry counts and reduces the chances that large clusters grow near the tree root while increasing the chances that small clusters grow further from the tree root. These two mechanisms vary differently across the possible ranges of risk fluctuation. This gives them a unique effect that cannot be reproduced by other sets of parameters. Thus, this work illustrates that sequences have the potential to reflect risk fluctuation effects and indicate the expected effects of T&T and PrEP. As models get more realistically complex, additional mechanisms will be needed. And finding distinct mechanisms generating dynamics and tree shapes will be more challenging. But doing so will add great value to fitting models to sequences.

References

[1] R.A. Smith, E.L. Ionides, A.A. King, *Infectious Disease Dynamics Inferred from Genetic Data via Sequential Monte Carlo*, Mol. Biol. Evol., **34** (2017), 2065–2084.

The role of host and pathogen population structure in the dynamics of multi-drug resistance

Sonja Lehtinen

(joint work with Francois Blanquart, Marc Lipsitch, Christophe Fraser and the Maela Pneumococcal Collaboration)

Understanding the short- and long-term dynamics of drug and multi-drug resistance is important for public health. Yet, there are pervasive trends in resistance dynamics that have not been fully explained. Firstly, antibiotic sensitive and resistant strains coexist robustly, despite prolonged selection pressure from antibiotics. Secondly, resistance to different antibiotics tends to co-occur on the same strains, leading to high frequencies of multi-drug resistance (MDR).

First, we present a model in which coexistence is maintained by variation in duration of carriage within the pathogen population (e.g. pneumococcal serotypes differing in duration of carriage) because the fitness effect of resistance depends on duration of carriage. Second, we show that this model is structurally similar to other plausible models of coexistence where the coexistence-maintaining mechanism is based on variation in the fitness benefit of resistance. Models with this structure also give rise to high MDR frequencies, because resistance against all antibiotics is concentrated in the sub-populations where the fitness advantage gained from resistance is high.

We find that predictions from this model are qualitatively consistent with trends observed in multiple *Streptococcus pneumoniae* datasets. This model provides a parsimonious explanation for the pervasiveness of high MDR frequencies and allows us to reconcile this trend with observed long-term stability in the prevalence of resistance.

Is your family pet a source of antibiotic resistance?

Mick Roberts

Antibiotics are used extensively to control infections in domestic pets, either as a course of oral tablets or a single injection. There are several methods by which bacteria can develop resistant strains, including mutation during reproduction and horizontal gene transfer [1]. We present a model for the development of antibiotic resistance within a single host animal.

Let $X = X_w + X_m + X_r$ represent the total bacteria in a single host animal, where X_w are the wild-type bacteria, X_m are the bacteria that are resistant to infection due to a mutation, and X_r are the bacteria that are resistant to infection due to the presence of a plasmid. The dynamics are described by the equations

$$\begin{aligned}\frac{dX_w}{dt} &= \nu_w F(X)X_w - \gamma_w X_w + \sigma X_r - \eta X_w X_r \\ \frac{dX_m}{dt} &= \epsilon F(X)X_w + \nu_m F(X)X_m - \gamma_m X_m \\ \frac{dX_r}{dt} &= \nu_r F(X)X_r - \gamma_r X_r - \sigma X_r + \eta X_w X_r\end{aligned} \tag{1}$$

where $F(X)$ accounts for saturation: $F(0) = 1$ and $F'(X) \leq 0$. For our analysis we define three reproduction numbers: $R_w = \nu_w/\gamma_w$; $R_m = \nu_m/\gamma_m$; and $R_r = \nu_r/\left(\gamma_r + \sigma\right)$. When only one mechanism for resistance operates, equations 1 reduce to a two-dimensional system, solutions are then confined to a bounded region of the positive quadrant and no periodic solutions are possible.

If $X_r \equiv 0$ selection is by mutation only. The trivial steady state $X_w = X_m = 0$ is stable if $R_w < 1$ and $R_m < 1$. The semi-trivial steady state $(X_w, X_m) = \left(0, X_m^{\#}\right)$, where $R_m F(X_m^{\#}) = 1$ exists if $R_m > 1$; and is stable if $R_m > R_w$. The non-trivial steady state where $R_w F(X^*) = 1$ exists and is stable if $R_w > 1$ and $R_w > R_m$. Here

$$X_w^* = \frac{R_w - R_m}{R_w - R_m + \epsilon/\gamma_m} X^* \qquad X_m^* = \frac{\epsilon/\gamma_m}{R_w - R_m + \epsilon/\gamma_m} X^*$$

The dynamics are summarised in Figure 1A. It can be seen in the figure that if $R_m < 1$ and R_w is reduced (lower horizontal broken line), then while $R_w > 1$ a non-trivial steady state with both types present exists and is stable, but for $R_w < 1$ the trivial state with no bacteria present is stable. If $R_m > 1$ and R_w is reduced (upper horizontal broken line), then the realised steady state changes from the non-trivial to the semi-trivial, with only the mutant X_m present.

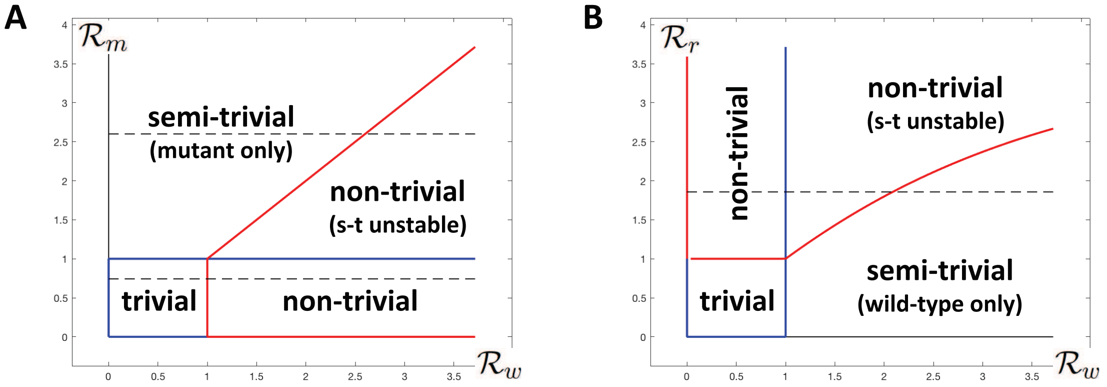

FIGURE 1. A: Regions in the (R_w, R_m) plane where the possible steady states of the model with resistance due to mutation only exist and are stable. B: Regions in the (R_w, R_r) plane where the possible steady states of the model with resistance due to a plasmid only exist and are stable.

If $X_m \equiv 0$ selection is by horizontal gene transfer only. The trivial steady state $X_w = X_r = 0$ is stable if $R_w < 1$ and $R_r < 1$. Define

$$X_w^\dagger = \frac{\nu_r}{\eta}\left(\frac{1}{R_r} - \frac{1}{R_w}\right)$$

The semi-trivial steady state $(X_w, X_r) = \big(X_w^\#, 0\big)$, where $R_w F(X_w^\#) = 1$ exists when $R_w > 1$; and is stable if $X_w^\# < X_w^\dagger$. A non-trivial steady state $(X_w, X_r) = (X_w^*, X_r^*)$ exists and is stable when $R_w < 1$ and $R_r > 1$; and when $R_w > 1$ and $X_w^\# > X_w^\dagger$. The dynamics are summarised in Figure 1B. It can be seen in the figure that if R_w is reduced along the horizontal broken line, then the realised steady state changes from the semi-trivial, with only the wild-type X_w present to the non-trivial with both types present.

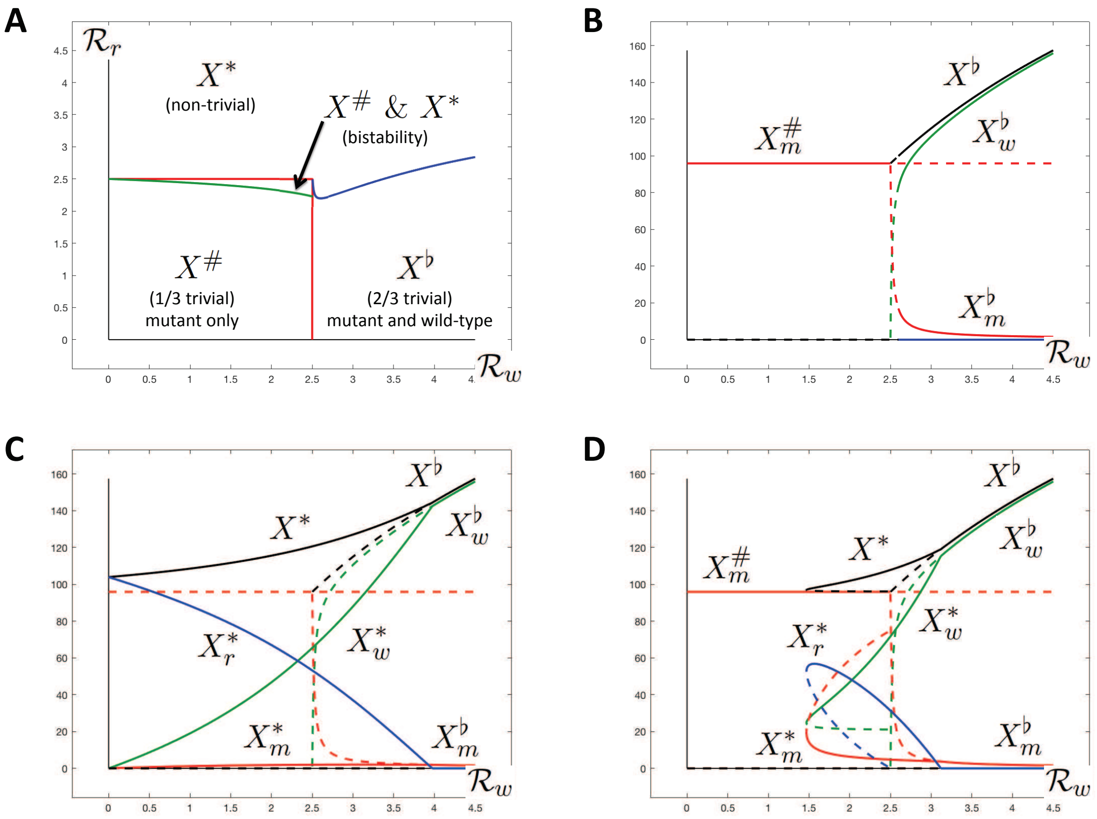

FIGURE 2. Summary of the dynamics of Equations 1 when $R_m = 2.5$. A: Regions in the (R_w, R_r) plane where steady states of the model exist and are stable. B-D: Bifurcation diagram showing steady states as a function of R_w when B: $R_r = 2.1$; C: $R_r = 2.7$; and D: $R_r = 2.4$. Parameter values are based on those in [2]-[5].

When both mechanisms for resistance are present then the dynamics are more complicated. The trivial steady state is stable if $R_w < 1$, $R_m < 1$ and $R_r < 1$. The $\frac{2}{3}$-trivial steady state $(X_w, X_m, X_r) = \big(0, X_m^\#, 0\big)$, where $R_m F(X_m^\#) = 1$ exists when $R_m > 1$; and is stable if $R_w < R_m$ and $R_r < R_m$. The $\frac{1}{3}$-trivial steady state $(X_w, X_m, X_r) = \left(X_w^\flat, X_m^\flat, 0\right)$, where $R_w F(X^\flat) = 1$ and $X_w^\flat + X_m^\flat = X^\flat$, exists when $R_w > 1$ and $R_w > R_m$; and is stable if $X_w^\flat < X_w^\dagger$. Non-trivial steady states have been found numerically.

As an illustration Figure 2 summarises the results when $R_m = 2.5$. In particular, in Figure 2A not only are there regions in the (R_w, R_r) plane where each steady state (apart from the trivial) exists and is stable, but there is a region where bistability has been shown to exist. Figure 2B shows a bifurcation diagram with $R_r \ll R_m$. As R_w is reduced, then the realised steady state changes from the $\frac{2}{3}$-trivial with the wild-type X_w and the mutant X_m present, to the $\frac{1}{3}$-trivial with only X_m present. Figure 2C shows a bifurcation diagram with $R_r > R_m$. As R_w is reduced, then the realised steady state changes from the $\frac{2}{3}$-trivial, to the non-trivial with all three types present. For the bifurcation diagram in Figure 2D, R_r is less than but close to R_m. The figure shows a region of bistability for $R_w < R_m$. For large values of R_w the $\frac{2}{3}$-trivial state is present and stable. As R_w is reduced the non-trivial steady state with all three types emerges. However, reducing R_w even further the non-trivial state disappears in a saddle node bifurcation, and the only steady state is the $\frac{1}{3}$-trivial with only X_m present.

We have presented and analysed a within-host model for the dynamics of antibiotic resistance. We have shown that if resistance arises due to mutation, then control measures that reduce R_w, the reproduction number for the wild-type bacteria, either drive the bacterial population to extinction or select for the mutant. If resistance arises due to the presence of a plasmid, then control measures either drive the bacterial population to extinction or result in both the wild-type and resistant bacteria being present. Finally, if both mechanisms are present, the dynamics are more complicated and control measures could result in hysteresis effects and selection for resistance.

Acknowledgements. The author wishes to thank Jackie Benschop, Sara Burgess, Nigel French, Jonathan Marshall and Leah Toombs-Ruane for advice.

References

[1] D.I. Andersson, D. Hughes, *Antibiotic resistance and its cost: is it possible to reverse resistance?*, Nature Reviews Microbiology, **8** (2010), 260–271.

[2] P. Komp Lindgren, L.L. Marcusson, D. Sandvang, N. Frimodt-Møller, D. Hughes, *Biological cost of single and multiple norfloxacin resistance mutations in* Escherichia coli *implicated in urinary tract infections*, Antimicrobial Agents and Chemotherapy, **49** (2005), 2343–2351.

[3] B.R. Levin, F.M. Stewart, V.A. Rice, *The kinetics of conjugative plasmid transmission: fit of a simple mass action model*, Plasmid, **2** (1979), 247–260.

[4] A.J. Lopatkin, H.R. Meredith, J.K. Srimani, C. Pfeiffer, R. Durrett, L. You, *Persistence and reversal of plasmid-mediated antibiotic resistance*, Nature Communications, **8** (2017), 1689.

[5] C. Myhrvold, J.W. Kotula, W.M. Hicks, N.J. Conway, P.A. Silver, *A distributed cell division counter reveals growth dynamics in the gut microbiota*, Nature Communications, **6** (2015), 10039.

The TransMID project: social contact patterns relevant for the spread of infectious diseases: past, present and future

Niel Hens
(joint work with the SIMID consortium (www.simid.be) and its collaborators)

Social contact data are increasingly being used to improve our understanding on how close contact infectious diseases spread from person to person, and to help guide effective policies on disease prevention and control. It is - to a large extent - within this context that the TransMID project, an ERC consolidator grant, was born. Here, I first describe the TransMID project after which I summarise the results of a systematic review of social contact surveys. I also describe a social contact data sharing initiative and I end with describing the fist household-based contact survey focussing on testing the often-made random mixing assumption in households.

TransMID focuses on the development of novel methods to estimate key epidemiological parameters from both serological and social contact data, with the aim to significantly expand the range of public health questions that can be adequately addressed using such data. Using new statistical and mathematical theory and newly collected as well as readily available serological and social contact data, fundamental mathematical and epidemiological challenges are addressed: (a) frequency and density dependent mass action relating potential effective contacts to transmission dynamics in (sub)populations of different sizes with an empirical assessment using readily available contact data, (b) behavioural and temporal variations in contact patterns and their impact on the dynamics of infectious diseases, (c) close contact household networks and the assumption of homogeneous mixing within households, (d) estimating parameters from multivariate and serial cross-sectional serological data taking temporal effects and heterogeneity in acquisition into account in combination with the use of social contact data, and (e) finally the design of sero- and social contact surveys with specific focus on serial cross-sectional surveys.

Based on a systematic review of the study design, statistical analyses and outcomes of the many social contact surveys that have been published, we found that surveys for collecting empirical contact data in a population have been conducted widely in many countries, but mostly in high-income countries. The surveys present a wide range of study designs. Throughout, we found that the overall contact patterns were remarkable robust to the study details. By taking the most common approach in each aspect of design (e.g. sampling schemes, data collection, definition of contact), we could identify a ‘common practice’ approach that can be used to facilitate comparison between studies and for benchmarking future studies.

Within TransMID, we are making social contact data available via `http://www.socialcontactdata.org` which links to data repositories in which the data are structured according to a relational database structure. These data can then be analysed using different methods (Van de Kassteele et al. 2017;

Camarda and Hens, 2003) made available in the R-package ‘socialmixr? first developed by Sebastian Funk (London School of Hygiene and Tropical Medicine, London, UK).

Finally, I present results on the first social contact survey specifically designed to study contact networks within households (Goeyvaerts et al., 2017). We found a high degree of clustering and, specifically on weekdays, decreasing connectedness with increasing household size. Epidemic simulation results suggest that within-household contact density is the main driver of differences in epidemic spread between complete and empirical-based household contact networks. The homogeneous mixing assumption may therefore be an adequate characterisation of the within-household contact structure for the purpose of epidemic simulation. However, ignoring the contact density when inferring from an epidemic model will result in biased estimates of within-household transmission rates. Further research on the implementation of within-household contact networks in epidemic models is necessary.

References

[1] G. Camarda, N. Hens, *Modelling social contact data: a smoothing constrained approach*, Proceedings of the 13th IWSM, (2013).

[2] N. Goeyvaerts, E. Santermans, G. Potter, A. Torneri, K.V. Kerckhove, L. Willem, M. Aerts, P. Beutels, N. Hens, *Household Members Do Not Contact Each Other at Random: Implications for Infectious Disease Modelling*, bioRxiv, (2017), 220202.

[3] J. van de Kassteele, J. van Eijkeren, J. Wallinga, *Efficient estimation of age-specific social contact rates between men and women*, Ann. Appl. Stat., **11** (2017), 320–339.

Survival biases lead to flawed conclusions in observational treatment studies of influenza patients

Martin Wolkewitz

(joint work with Martin Schumacher)

Background and Objective

Several observational studies reported that Oseltamivir (Tamiflu) reduced mortality in infected and hospitalized patients. Because of the restriction of observation to hospital stay and time-dependent treatment assignment, such findings were prone to common types of survival bias (length, time-dependent and competing risk bias).

Methods

British hospital data from the Influenza Clinical Information Network (FLU-CIN) study group were used which included 1,391 patients with confirmed pandemic influenza A/H1N1 2009 infection. We used a multistate model approach with following states: hospital admission, Oseltamivir treatment, discharge, and death. Time origin is influenza onset. We displayed individual data, risk sets, hazards, and probabilities from multistate models to study the impact of these three common survival biases.

Results

The correct hazard ratio of Oseltamivir for death was 1.03 (95% confidence interval [CI]: 0.64-1.66) and for discharge 1.89 (95% CI: 1.65-2.16). Length bias increased both hazard ratios (HRs): HR (death)=1.82 (95% CI: 1.12-2.98) and HR (discharge)=4.44 (95% CI: 3.90-5.05), whereas the time-dependent bias reduced them: HR (death)=0.62 (95% CI: 0.39-1.00) and HR (discharge)=0.85 (95% CI: 0.75-0.97). Length and time-dependent bias were less pronounced in terms of probabilities. Ignoring discharge as a competing event for hospital death led to a remarkable overestimation of hospital mortality and failed to detect the reducing effect of Oseltamivir on hospital stay.

Conclusions

The impact of each of the three survival biases was remarkable, and it can make neuraminidase inhibitors appear more effective or even harmful. Incorrect and misclassified risk sets were the primary sources of biased hazard rates.

References

[1] M. Wolkewitz, M. Schumacher, *Neuraminidase Inhibitors and Hospital Mortality in British Patients with H1N1 Influenza A: A Re-Analysis of Observational Data*, PLoS ONE, **11** (2016), 9:e0160430.

[2] M. Wolkewitz, M. Schumacher, *Survival biases lead to flawed conclusions in observational treatment studies of influenza patients.* J Clin Epidemiol, **84** (2017), 121–129.

Coalescent methods for integrating genomics and epidemiology

Daniel J. Wilson

Exploiting pathogen genomes to reconstruct transmission between populations and within outbreaks represents a powerful tool in the fight against infectious disease. However, the statistical models and inference methods that allow the exploitation of pathogen genomes are complex, so simplifying assumptions and approximations highly appealing. Sometimes these short-cuts are indispensable for practical purposes, but other times they can ignore important complexities of real outbreaks, such as within-host evolution and non-sampled patients, or they can be biased, overly confident, inefficient and downright misleading. In this talk I describe how my group has used coalescent-based models from population genetics to motivate the development of new methods for phylogeographic and outbreak inference. Using examples from real outbreaks of Ebola virus, avian influenza virus, foot and mouth disease virus and *Klebsiella pneumoniae*, together with detailed simulations, I compare existing methods to the new tools we have developed, and show how different methods can lead to very different epidemiological interpretations concerning transmission. Notably, we have found that phylogeographic inference of transmission based on the popular discrete trait analysis (DTA, also known as mugration) is extremely unreliable and sensitive to biased sampling. We have developed BASTA (BAyesian STructured coalescent Approximation), a phylogeography approach implemented in BEAST2 that combines the accuracy of methods based on the structured coalescent with the computational efficiency required to handle more than just few populations, and SCOTTI (Structured COalescent Transmission Tree Inference), a method for outbreak inference that takes forward the BASTA approach by modelling individual cases as separate subpopulations. BASTA overcomes the limitations of the DTA approach to phylogeography while SCOTTI outperforms the popular Outbreaker software while incorporating complexities not previously handled by existing tools. As genomics takes on an increasingly prominent role informing the control and prevention of infectious diseases, it will be vital to balance model robustness with statistical power to deliver statistical methods that provide reliable insights into transmission history.

References

[1] N. De Maio, C.J. Worby, D.J. Wilson, N. Stoesser, *Bayesian Reconstruction of Transmission within Outbreaks using Genomic Variants*, bioRxiv, (2017), 213819.

[2] N. De Maio, C.-H. Wu, D.J. Wilson, *SCOTTI: efficient reconstruction of transmission within outbreaks with the structured coalescent*, PLoS computational biology, **12** (2016), e1005130.

[3] N. De Maio, C.-H. Wu, K.M. O'Reilly, D. Wilson, *New routes to phylogeography: a Bayesian structured coalescent approximation*, PLoS genetics, **11** (2015), e1005421.

[4] B. Dearlove, D.J. Wilson, **Coalescent inference for infectious disease: meta-analysis of hepatitis C**, Phil. Trans. R. Soc. B, **368** (2013), 20120314.

Linking geostatistical and transmission models for neglected tropical diseases in Africa

Simon Spencer

(joint work with Panayiota Touloupou and Déirdre Hollingsworth)

In 2012, pharmaceutical companies, donors, endemic countries and non-governmental organisations committed to control, eliminate or eradicate 10 neglected tropical diseases by 2020. As 2020 approaches we aim to use geostatistical modelling coupled with mathematical models for the spread of infection to make projections about whether these ambitious goals can be met. We have focussed on 3 diseases: lymphatic filariasis (which leads to elephantiasis), onchocerciasis (which leads to river blindness) and Gambian human African Trypanosomiasis (HAT, which leads to sleeping sickness).

The key to our methodology lies in producing a large batch of simulations that cover all of the prevalence levels estimated in the geostatistical mapping, which was fitted to pre-control prevalence survey data. At each stage of the analysis we attempt to account for the uncertainties present. We begin by drawing simulations from the prior distribution of the model parameters, including the population size. For each set of parameters, the transmission model was then simulated to endemic equilibrium to obtain the pre-control prevalence according to the model. For each pixel in the map, we reweighted the simulations (via the empirical Radon-Nikadym derivative) to obtain the posterior prevalence distribution estimated from the geostatistical mapping. Finally we ran the simulations forward in time under different intervention strategies producing a weighted distribution of outcomes for each pixel in the map.

We illustrated our approach with results for lymphatic filariasis in Ethiopia, using 3 different transmission models from the Neglected Tropical Diseases Modeling Consortium. For each model we consider 4 intervention strategies: no intervention, annual Mass Drug Administration (MDA) with 65% coverage, annual MDA with 80% coverage and biannual MDA with 65% coverage. We find that most, but not all, implementation units can have greater than 90% probability of meeting the 2020 goal of having prevalence less than 1%, however in many regions the intensity of the intervention strategy must be increased to achieve this.

In future, we plan to extend our approach to a whole Africa analysis for both lymphatic filariasis and onchocerciasis. For HAT, our simulation-based approach is not appropriate due to the dependence between the emerging data and the trajectory of the epidemic. This occurs because infected individuals identified by large-scale surveys are treated, reducing the infection pressure and so any model simulations must be specific to each location. We are working on adaptive and computationally scalable Markov chain Monte Carlo techniques to fit models to individual health areas across Africa.

Population Genomics, Price Decomposition, Proximate and Ultimate Causes of Vector Competence

Claudio Struchiner
(joint work with Jose MC Ribeiro and Bruno Arca)

The completion of a set of reference genome assemblies for 16 species of Anopheles mosquitoes allows for the exploration of the capacity exhibited by some of these mosquito species to serve as vectors for malaria parasites. Vectorial capacity is a highly variable trait determined by differences in mosquito physiology, molecular biology, and behavior. By taking advantage of this genome-wide comparative framework, we explore the distinct evolutionary patterns that emerge when contrasting classes of genes implicated in the ability of mosquitoes to transmit malaria parasites to humans, as well as their chromosomal locations that might indicate closely related species intermixing.

Genes evolve by the accumulation of neutral mutations by random drift, and the fixation of adaptive mutations by selection. Tools proposed to detect these mechanisms develop around the idea of comparing the number of amino-acid replacement (non-synonymous) substitutions to synonymous substitutions in orthologous coding sequences, a notable example being the McDonald and Kreitman (MK) approach. The MK framework has been recently expanded to address genome-wide patterns of molecular variation using thousands of genes simultaneously. By fitting the latter models to genomic data from a Cameroonian An. gambiae population having An. stephensi as an outgroup, we are able to estimate key parameters, for about ten thousand genes, that drive the evolutionary history of the phenotypic properties of these mosquitoes, in particular, their competence as vectors.

The massive amount of data generated in the steps above may lead to the exploration of the proximate causes (developmental, physiological, or chemical mechanism) that triggers the competence to transmit diseases and the ultimate causes (why) that drive the fixation of this phenotypic trait in certain mosquito species. Still a work in progress, our exploration of the evolutionary significance of proximate and ultimate causal factors of vector competence among mosquitoes relies on the proximate causal decomposition provided by the Price equation and the unified language of causation provided by graph theory. In the previous two-step process, Price's framework is used to decompose the change in the mean phenotypic value from one generation to another into a set of underlying causal factors that includes selection and reproduction. The parameters estimated from the genome-wide analysis enter as input empirical data in this decomposition. Causal graph theory provides a bridge between causal statements and probability, and formalizes the interpretation of causal models involving the concepts of proximate and ultimate causes. This work is an attempt to add an explicit evolutionary genetic layer to the components that enter the concept of vectorial capacity, a summary measure of the relative importance of the parameters involved in vector-borne transmission dynamics.

References

[1] D.E. Neafsey et al., *Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes*, Science, **347** (2015), doi:10.1126/science.1258522

Pairwise survival analysis: Contact intervals, regression, and phylogenetics

Eben Kenah

When integrating epidemiologic data with pathogen phylogenetics, the likelihood for the transmission model is often a branching-process likelihood based on a generation interval distribution. We show that a misspecified likelihood can lead to severely biased estimates with or without a pathogen phylogeny. Writing the likelihood as a survival likelihood with failure times in pairs – a process we call pairwise survival analysis – accounts for time spent at risk of infection and leads to more accurate estimation and source attribution than approaches based on branching processes and generation intervals.

In the ordered pair ij, the *contact interval* τ_{ij} from i to j is the time from the onset of infectiousness in i until infectious contact from i to j, where infectious contact is defined as a contact sufficient to infect j if j is susceptible. The contact interval distribution provides a useful summary of infectious disease transmission. The probability of infectious contact from i to j is $S_{ij}(\iota_i$ where $S_{ij}(\tau) = \Pr(\tau_{ij} > \tau)$

is the survival function of the contact interval and ι_i is the infectious period of i. The probability that j receives infectious contact from i in the infectiousness age interval $(\tau, \tau + \mathrm{d}\tau]$ is $h_{ij}(\tau)\mathrm{d}\tau$, where $h_{ij}(\tau)$ is the hazard function of the contact interval distribution. Thus, $h_{ij}(\tau)$ gives us the instantaneous infectiousness of i as a function of infectiousness age. These properties of the contact interval distribution make it a useful tool for understanding stochastic epidemic models. In a homogeneous epidemic models on a configuration-model network, the basic reproduction number is

$$R_0 = \mathbb{E}\left[\frac{D(D-1)}{\mathbb{E}[D]} S(\iota)\right] \tag{1}$$

where the expectation is taken over the joint distribution of the degree D and the infectious period ι. In a mass-action model where the hazard of infectious contact in a given pair is inversely proportional to the population size, we get

$$R_0 = \mathbb{E}\big[H(\iota)\big] \tag{2}$$

where $H(\tau) = -\ln S(\tau)$ is the cumulative hazard function of the contact interval distribution. In the stochastic Kermack-McKendrick model where $S'(t) = -\beta S(t)I(t)$, the contact interval distribution is exponential(β) and $R_0 = \beta\mathbb{E}[\iota]$. When $I'(t) = -\gamma I(t)$, the infectious period is exponential(γ) so $R_0 = \beta\gamma^{-1}$.

Consider a simple example where individuals A, B, and C are infected at times $t_A < t_B < t_C$ such that A and B are both possible infectors of C. For simplicity, assume there is no latent period. In a branching process model, the likelihood for the data is the sum of the likelihood contributions of the two possible transmission trees:

$$g_{AB}g_{AC} + g_{AB}g_{BC} = g_{AB}(g_{AC} + g_{BC}) \tag{3}$$

where $g_{XY} = g(t_Y - t_X)$ and g is the probability density function (PDF) of the contact interval distribution. The pairwise survival likelihood is

$$h_{AB}(h_{AC} + h_{BC})S_{AB}S_{AC}S_{BC} \tag{4}$$

where $h_{XY} = h(t_Y - t_X)$ and $S_{XY} = S(t_Y - t_X)$. This is also a sum over the two possible transmission trees. The survival terms account for person-time spent at risk of infection prior to infection. To calculate the probability that person X infected person C, we take a likelihood ratio with all likelihood contributions for transmission trees where X infected C on top and the total likelihood on the bottom. For generation interval pdfs, this gives us

$$\frac{g_{XC}}{g_{AB} + g_{BC}}. \tag{5}$$

The pairwise survival likelihood gives us

$$\frac{h_{XC}}{h_{AB} + h_{BC}}. \tag{6}$$

In a model with a constant hazard of infectious contact, we get the $p_{AC} < p_{BC}$ using generation intervals and $p_{AC} = p_{BC}$ using pairwise survival analysis. The latter result is known to be correct for this model.

One common motivation for using the branching process approach is to obtain a likelihood that does not depend on observations of uninfected individuals. Under current epidemiological practice, such data is often not available during an outbreak. In a mass-action model in a population of size n, we get an asymptotic log likelihood of the form

$$\ln h_{AB} + \ln(h_{AB} + h_{BC}) - \Big(H(\iota_A) + H(\iota_B) + H(\iota_C)\Big) \tag{7}$$

where ι_X is the infectious period in individual X. This likelihood does not depend on any data about uninfected individuals. In order for inference using generation intervals to work, the cumulative hazard terms must be near zero. However, this requires $R_0 \approx 0$. The pairwise survival likelihood works under the much weaker assumption of mass-action and negligible depletion of susceptibles.

As an example of the flexibility of pairwise survival analysis, we describe a pairwise accelerated failure time model that can be used to estimate covariate effects on infectiousness and susceptibility. In this model, the rate parameter for infectious contact from i to j has the form

$$\lambda_{ij} = \exp\big(\beta_{\text{int}}^\top X_{ij}\big)\lambda_0 \tag{8}$$

where β_{int} is an unknown coefficient vector, X_{ij} is a vector of individual or pairwise covariates, and λ_0 is a baseline rate parameter. This is called the internal transmission model. To account for the risk of infectious contact from outside the observed population, there is also an external transmission model where

$$\lambda_{0j} = \exp\big(\beta_{\text{ext}}^\top X_{0j}\big)\mu_0 \tag{9}$$

where β_{ext} is an unknown coefficient vector, X_{0j} is a vector of individual-level covariates, and μ_0 is a baseline rate parameter. Covariates in X_{ij} and X_{0j} can be unique to each model or shared. This model – modified to account for the buildup of immunity – will be used to estimate the efficacy of the Ebola vaccine based on the WHO ring vaccination trial in Guinea. This trial collected data on individuals exposed to infection who escaped as well as Ebola virus genetic sequences.

Finally, we describe a pruning (peeling) algorithm for calculating an approximate likelihood using both epidemiologic data and a pathogen phylogeny. This algorithm calculates an unweighted sum of all transmission trees consistent with a pathogen phylogeny. The true likelihood is a weighted sum, but it cannot be calculated quickly using any known pruning algorithm. One possible application of this algorithm is to run an Markov chain Monte Carlo (MCMC) algorithm in the approximate likelihood and use importance weighting to perform Bayesian inference using the true likelihood. Pathogen genetics can improve statistical efficiency and reduce bias, but this depends on good epidemiologic study design and a good likelihood for transmission.

Disease Burden

Johannes Müller
(joint work with Mirjam Kretzschmar)

1. Introduction

Public health authorities have the difficulty to decide how to distribute resources to the control of several diseases in parallel. The problem here is that – strictly spoken – diseases are not comparable.

One way out is a purely economic view. Intervention measures (vaccination, screening, contact tracing,. . .) have some costs. By means of intervention measures, cases can be prevented, and therewith money can be saved (treatment, sick leaves,. . .). If we subtract the costs from the savings we obtain an economic measure that can be used as a basis for decisions.

The economic approach clearly does not cover all relevant aspects. Of course, we have costs for a control measure. However, the primary aim is not to reduce the economic impact of a disease but to reduce detrimental effects. Disease burden is another idea that aims at quantifying these detrimental effects. In the last consequence, however, this is for sure not possible. Nevertheless, decisions have to be made, and it is for sure advantageous to have some rational basis for decisions. The idea of disease burden is to measure the effects of a disease in missed life years [1, 2].
• A person who dies due to a disease would otherwise life a certain number of further years. This aspect is captured by the years of life lost due to premature death (YLL).
• A person who is diseased has some (more or less severe) symptoms. These symptoms induce the inability to do what he/she would do otherwise. There is a reduction of the effective life time. This effect is measured in years of life lost due to disability (YLD).
Therewith we define the disease burden.

$$\text{Disease Burden} := \text{YLL}+\text{YLD}$$

The disease burden can be defined via incidence or via prevalence. In equilibrium situations (stationary state), the two definitions coincide. In this talk we focus on the extension of this concept from stationary to more dynamic situations, that is, to exponentially increasing/decreasing populations. Therewith it is possible to address the changing age structure in developed countries (population becomes older).

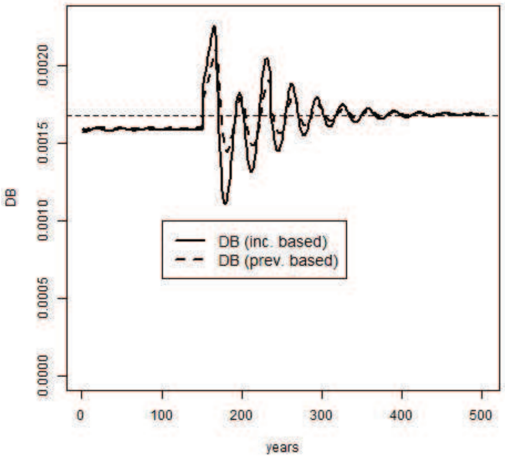

Figure 1. Disease burden.

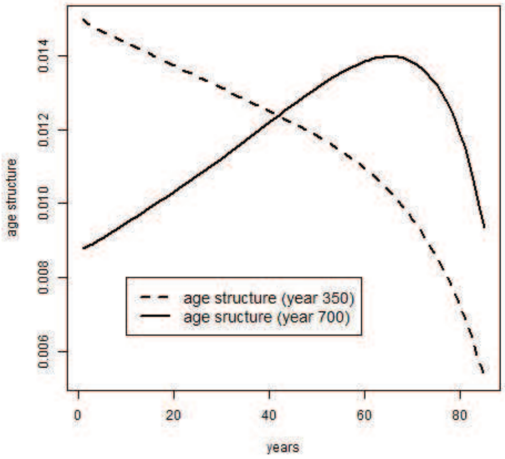

Figure 2. Age structure (years 350 and 700)

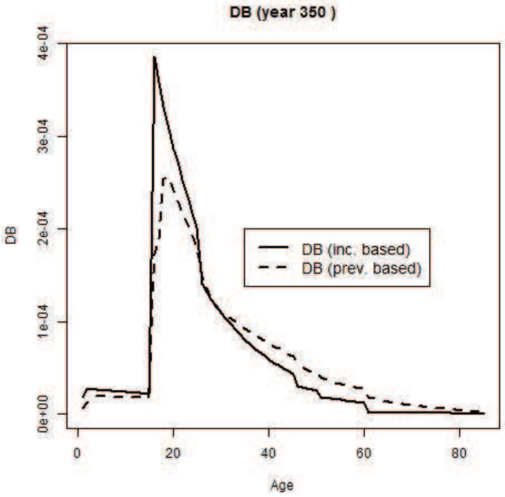

Figure 3. Age structured disease burden at year 350

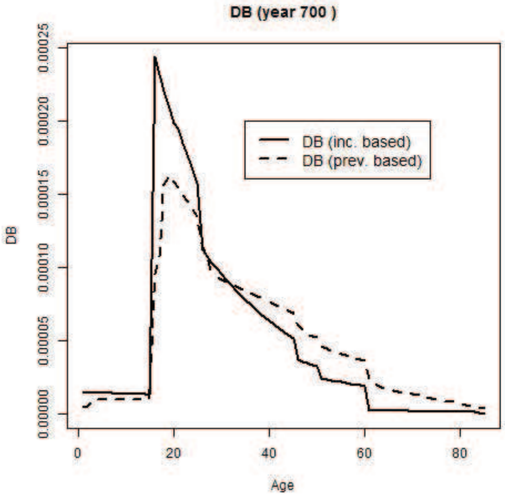

Figure 4. Age structured disease burden at year 700

2. Definition of disease burden

We consider the age-structured model ($\mathbf{e}_i$ denotes the i'th unit vector, $\mathbf{e} = (1, \ldots, 1)^T$)

$$\begin{aligned}
(\partial_t + \partial_a)S(a,t) &= -\Lambda S(a,t) - \mu(a)S \\
(\partial_t + \partial_a)X(a,t) &= \Lambda S(a,t)\ \mathbf{e}_1 + AX - \mu(a)X - D(a)X \\
\Lambda(a,t) &= \int_0^\infty k(a,b)I(b,t)/P(t)\,db \\
P(t) &= \int_0^\infty S(a,t) + \mathbf{e}^T X(a,t)\,da \\
S(0,t) &= \int_0^\infty b(a)(S(a,t) + \mathbf{e}^T X(a,t))\,da \\
X(0,t) &= 0
\end{aligned}$$

where X is a vector that incorporates the infected class and all subsequent classes (e.g. chronic infecteds, recovereds, etc.). A is an M-Matrix, $\mathbf{e}^T A = 0$, and $D(a)$ is a diagonal matrix indicating the additional mortality.

Let $Y(a; a_0)$ denote the fate of an individual infected at age a_0,

$$\frac{d}{da}Y(a;a_0) = AY(a;a_0) - \mu(a)Y(a;a_0) - D(a)Y(a;a_0), \quad Y(a_0;a_0) = \mathbf{e}_1.$$

We now introduce the functions $yll_{ind}(a_0,t)$ and $yld_{ind}(a_0,t)$ (the expected YLL resp. YLD for a single person infected at time t at age a_0). Therewith we define the prevalence DB_{prev} and incidence DB_{inc} based disease burden per capita. The vector $\mathbf{W}$ consist of the disability weights that measure the effective reduction of the live years due to disability.

$$\begin{aligned}
yll_{ind}(a_0,t) &= \int_{a_0}^\infty \left(e^{-\int_{a_0}^a \mu(\tau)\,d\tau} - \mathbf{e}^T Y(a;a_0)\right)\frac{P(t)}{P(t+a-a_0)}da \\
yld_{ind}(a_0,t) &= \int_{a_0}^\infty \mathbf{W}^T Y(a;a_0)\ \frac{P(t)}{P(t+a-a_0)}da. \\
DB_{inc}(t) &= \int_0^\infty \Big((yll_{ind}(a,t)) + yld_{ind}(a,t)\Big)\Lambda(a)S(a,t)/P(t)\,da, \\
DB_{prev}(t) &= \int_0^\infty \Bigg(\mathbf{e}^T\, D(a)(X(t,a)/P(t))\int_a^\infty e^{-\int_a^{a'}\mu(\tau)\,d\tau}\frac{P(t)}{P(t+a'-a_0)}da' \\
&\qquad +\mathbf{W}^T X(t,a)/P(t)\Bigg)\,da.
\end{aligned}$$

If the population size is constant, the definition coincides with the classical definitions. Typically, in the long run the age structured model tends to an exponentially growing (decreasing) solution, ($S(a,t) = e^{\lambda t}S(a)$ etc.). In this case we can prove the following theorem.

Theorem *In the exponentially growing situation, we have*

$$DB_{inc}(t) = DB_{prev}(t).$$

3. Hepatitis B

We apply this definition to hepatitis B. We orient ourselves at the model as in [2], where parameters are chosen not completely realistically to exemplify the effect of an aging population. The birth rate is decreased at simulated year 350, such that the growth rate jumps from 0.004/year to the value of −0.0087/year. Before year 350 and eventually the incidence and prevalence based disease burden coincide, in the transient phase there is some difference (Fig. 1). As in a shrinking population there are more adults, and infection by hepatitis B is mainly transmitted by sexual contacts and needle sharing, the disease burden increases in time. Also the age structured disease burden reflects this fact (Figs. 3 and 4).

References

[1] S.A. McDonald, A. van Lier, D. Plass, M.E.E. Kretzschmar, *The impact of demographic change on the estimated future burden of infectious diseases: examples from hepatitis B and seasonal influenza in the Netherlands*, BMC Public Health, **12** (2012), 1046.

[2] D. Plass et al., *The disease burden of hepatitis B, influenza, measles and salmonellosis in Germany: first results of the Burden of Communicable Diseases in Europe Study*, Epidemiol. Infect., **142** (2014), 2024–2035.

A metric-based method for comparing transmission trees

Michelle Kendall

(joint work with Diepreye Ayabina, Yuanwei Xu, James Stimson and Caroline Colijn)

In the analysis of infectious disease outbreaks it is often important to be able to infer "who infected whom". The inferred links between infectors and infectees in an outbreak are commonly represented by a directed graph. If each infectee has at most one infector (in-degree ≤ 1), if there is a unique source case and if the graph forms a single connected component, then this graph is a tree and it is known as the *transmission tree.*

Accurate inference of the transmission tree of an outbreak is important for our understanding of pathogen dynamics and for public health strategies: determining whether there are certain individuals or locations causing high numbers of infections, identifying individuals at higher risk, determining which characteristics are associated with infectiousness, and analysing the efficacy of interventions.

However, inference of transmission trees is complicated. There are several choices to be made: the input data (genetic and/or epidemiological); the inference framework (maximum likelihood, Bayesian, etc.); data-collection assumptions such as the likely number of unsampled cases; and pathogen-specific assumptions such as the time between infection and becoming infectious, and the variety in the pathogen 'strains' transferred at the time of infection (size of bottleneck).

A wide range of methods and software are available for transmission tree inference. Typically, each methodological combination produces a different transmission tree; often the differences significantly alter the epidemiological story of the outbreak, with important implications for downstream analysis and public health decision making. Often the trees are too numerous and/or too large to compare simply by plotting them and examining their differences by eye. Although each tree captures meaningful signals in the data, it is often impractical to retain all inferred trees for onward analyses. Typically a *consensus* tree is calculated using a method like Edmond's algorithm. However, we demonstrate that this tree can differ significantly from each of the inferred trees.

To address these difficulties we propose a *metric* on transmission trees. A metric is a specific type of distance function; here we note that it describes a distance between two objects, giving a distance of zero if the objects are identical and larger values for objects which are more dissimilar. Our metric enables quantitative comparison of transmission trees, allowing us to compare and summarise various hypotheses of "who infected whom". By capturing differences in source case attribution, transmission direction between individuals, tree shape (linked to $R0$) and numbers of unsampled cases, it enables us to sort the trees in epidemiologically meaningful ways. It reveals whether multiple analyses (and/or a Bayesian posterior set of trees) are in broad agreement or whether they are multimodal, supporting different transmission histories each with comparable likelihoods.

We use multidimensional scaling (MDS) to project pairwise tree distances into a small number of dimensions for visualisation. We show that natural similarities and symmetries of the tree space are preserved in this projection, enabling straightforward identification of trends and clustering by eye, as well as through rigorous statistical techniques.

The metric can also be used to select a single representative *median* tree (or a representative median tree for each distinct "cluster" of transmission trees). Unlike a consensus tree, this tree will by construction be one of the candidate trees inferred by the analysis, with a corresponding likelihood. Finally, we briefly discuss some further applications of the method: it can be used to compare inferred transmission trees to a simulated "true tree", to test inference accuracy, to test for convergence in Bayesian posterior tree sets, and potentially to propose MCMC "moves" in tree space.

Full details are given in [1]. The functions are available in the R package *treespace* [2, 3], where there is also a vignette for reproducing each of our illustrative examples.

References

[1] M. Kendall, D. Ayabina, Y. Xu, J. Stimson, C. Colijn, *Estimating transmission from genetic and epidemiological data: a metric to compare transmission trees*, Statistical Science, **33** (2018), 70–85.

[2] T. Jombart, M. Kendall, J. Almagro-Garcia, C. Colijn, *treespace: Statistical Exploration of Landscapes of Phylogenetic Trees*, R package, version 1.1.2. (2018), https://CRAN.R-project.org/package=treespace.

[3] T. Jombart, M. Kendall, J. Almagro-Garcia, C. Colijn, *treespace: statistical exploration of landscapes of phylogenetic trees*, Molecular Ecology Resources, **17** (2017), 1385–1392.

Reporter: Martin Eichner

Participants

Prof. Dr. Kari Auranen
Department of Mathematics and
Statistics
Turku University
20014 University of Turku
FINLAND

Prof. Dr. Frank G. Ball
School of Mathematical Sciences
The University of Nottingham
University Park
Nottingham NG7 2RD
UNITED KINGDOM

Allison Black
Vaccine and Infectious Disease Division
Fred Hutchinson Cancer Research Center
Department of Epidemiology
Arnold Building, Room M1-B861
1100 Fairview Avenue N.
Seattle WA 98109
UNITED STATES

Dr. Martin Bootsma
Department of Mathematics
Utrecht University
P.O.Box 80.010
3508 TA Utrecht
NETHERLANDS

Johannes Bracher
Department of Epidemiology,
Biostatistics
and Prevention Institute (EBPI)
University of Zürich
Hirschengraben 84
8001 Zürich
SWITZERLAND

Kristin Bratton Nelson
Rollins School of Public Health
Emory University
1518 Clifton Road NE
Atlanta GA 30322
UNITED STATES

Prof. Dr. Tom Britton
Department of Mathematics
Stockholm University
10691 Stockholm
SWEDEN

Dr. Caroline Colijn
Department of Mathematics
Imperial College London
Huxley Building
180 Queen’s Gate
London SW7 2AZ
UNITED KINGDOM

Dr. Ben Cooper
Tropical Medicine Research Unit
Mahidol University
60th Anniv., Chalermprakiat Bldg.
420/6 Ratchawithi Road
Bangkok 10400
THAILAND

Prof. Dr. Martin Eichner
Institut für Klinische Epidemiologie und
Angewandte Biometrie
Universität Tübingen
Silcherstrasse 5
72070 Tübingen
GERMANY

Prof. Dr. Simon D.W. Frost
Department of Veterinary Medicine
University of Cambridge
Veterinary School
Madingley Road
Cambridge CB3 0ES
UNITED KINGDOM

Prof. Dr. Gavin J. Gibson
Department of Actuarial Mathematics
and Statistics
Heriot-Watt University
Riccarton
Edinburgh EH14 4AS
UNITED KINGDOM

Prof. Dr. M. Elizabeth Halloran
Department of Biostatistics
University of Washington and
Fred Hutchinson Cancer Research Center
1100 Fairview Ave. N, LE-400
Seattle, WA 98109-1024
UNITED STATES

Prof. Dr. Niel Hens
Center for Statistics
Hasselt University
Agoralaan Building D
3590 Diepenbeek
BELGIUM

Prof. Dr. Edward L. Ionides
Department of Statistics
University of Michigan
439 West Hall
1085 South University
Ann Arbor MI 48109-1107
UNITED STATES

Prof. Dr. Valerie S. Isham
Department of Statistical Science
University College London
Gower Street
London WC1E 6BT
UNITED KINGDOM

Prof. Dr. Niels Keiding
Section of Biostatistics
Kobenhavns Universitet
Oster Farimagsgade 5
P.O. Box 2099
1014 København K
DENMARK

Dr. Eben E. Kenah
Division of Biostatistics
College of Public Health
The Ohio State University
1841 Neil Avenue
Columbus, OH 43210
UNITED STATES

Dr. Michelle L. Kendall
Big Data Institute
University of Oxford
Li Ka Shing Centre for Health
Information
and Discovery
Old Road Campus
Oxford OX3 7LF
UNITED KINGDOM

Prof. Dr. Aaron King
Department of Ecology and
Evolutionary Biology
2019 Kraus Natural Science Building
830 North University Avenue
Ann Arbor MI 48109-1048
UNITED STATES

Dr. Don Klinkenberg
National Institute for Public Health
and the Environment
RIVM
P.O. Box 1
3720 BA Bilthoven
NETHERLANDS

Dr. Jim Koopman
Department of Epidemiology, SPH-1
University of Michigan
109 Observatory Street
Ann Arbor MI 48109
UNITED STATES

Dr. Mirjam Kretzschmar
Centre for Infectious Disease Control
RIVM
Antonie van Leeuwenhoeklaan 9
P.O.Box 1
3720 BA Bilthoven
NETHERLANDS

Dr. Theodore Kypraios
School of Mathematical Sciences
The University of Nottingham
University Park
Nottingham NG7 2RD
UNITED KINGDOM

Dr. Sonja Lehtinen
Big Data Institute
University of Oxford
Li Ka Shing Centre for Health
Information
and Discovery
Old Road Campus
Oxford OX3 7LF
UNITED KINGDOM

Dr. KaYin Leung
Department of Mathematics
Stockholm University
10691 Stockholm
SWEDEN

Prof. Dr. Ira M. Longini
Department of Biostatistics
University of Florida
452 Dauer Hall
22 Buckman Drive
P.O. Box 117450
Gainesville, FL 32610
UNITED STATES

Prof. Dr. Emma McBryde
Australian Institute of Tropical Health
and Medicine
James Cook University
Townsville QLD 4811
AUSTRALIA

Prof. Dr. Vladimir N. Minin
Department of Statistics
University of California, Irvine
Donald Bren Hall 2068
P.O. Box 354322
Irvine CA 92697-1250
UNITED STATES

Prof. Dr. Denis Mollison
The Laigh House
Inveresk
Musselburgh EH21 7TD
UNITED KINGDOM

Prof. Dr. Johannes Müller
Zentrum Mathematik
Technische Universität München
Boltzmannstrasse 3
85748 Garching bei München
GERMANY

Prof. Dr. Philip D. O'Neill
School of Mathematical Sciences
The University of Nottingham
University Park
Nottingham NG7 2RD
UNITED KINGDOM

Dr. Lorenzo Pellis
School of Mathematics
The University of Manchester
Alan Turing Building
Oxford Road
Manchester M13 9PL
UNITED KINGDOM

Kiesha Prem
Saw Swee Hock School of Public Health
Tahir Foundation Building
National University of Singapore
12 Science Drive 2, #09-01
117 549 Singapore
SINGAPORE

Dr. Oliver Ratmann
Faculty of Natural Sciences
Department of Mathematics and Statistics
Imperial College London
South Kensington Campus
525 Huxley Building
London SW7 2AZ
UNITED KINGDOM

Prof. Dr. Mick G. Roberts
Institute of Natural & Mathematical Sciences
Massey University
North Shore City Mail Centre
Private Bag 102904
Auckland
NEW ZEALAND

Dr. Gianpaolo Scalia-Tomba
Dipartimento di Matematica
Universita di Roma Tor Vergata
Via della Ricerca Scientif., 1
00133 Roma
ITALY

Dr. Markus Schwehm
ExploSYS GmbH
Otto-Hahn-Weg 6
70771 Leinfelden
GERMANY

Dr. Simon E. Spencer
Department of Statistics
University of Warwick
Coventry CV4 7AL
UNITED KINGDOM

Theresa Stocks
Department of Mathematics
Stockholm University
10691 Stockholm
SWEDEN

Dr. Claudio J. Struchiner
Escola Nacional de Saude Publica
Fundacao Oswaldo Cruz
Rua Bejamin Batista 22/202
Rio de Janeiro 22461-120
BRAZIL

Dr. Panayiota Touloupou
Zeeman Institute
Senate House
University of Warwick
Coventry CV4 7AL
UNITED KINGDOM

Dr. Pieter Trapman
Matematiska Institutionen
Stockholms Universitet
10691 Stockholm
SWEDEN

Dr. Michiel van Boven
Centre for Infectious Disease Control
National Institute for Public Health
and the Environment (RIVM)
PO Box 1
3720 BA Bilthoven
NETHERLANDS

Prof. Dr. Jacco Wallinga
Centre for Infectious Disease Epidemiology
National Institute for Public Health
and the Environment (RIVM)
P.O. Box 1
3720 BA Bilthoven
NETHERLANDS

Jessica Welding
Department of Mathematics and
Statistics
Fylde College
Lancaster University
Lancaster LA1 4YF
UNITED KINGDOM

Dr. Daniel Wilson
Nuffield Department of Medicine
University of Oxford
John Radcliffe Hospital
Oxford OX3 9DU
UNITED KINGDOM

Dr. Martin Wolkewitz
Institute for Medical Biometry and
Statistics
Medical Center University of Freiburg
Stefan-Maier-Strasse 26
79104 Freiburg i. Br.
GERMANY

Dr. Chris Wymant
Big Data Institute
University of Oxford
Li Ka Shing Centre for Health
Information
and Discovery
Old Road Campus
Oxford OX3 7LF
UNITED KINGDOM

Dr. Jason Q. Xu
Biomathematics Department
University of California, Los Angeles
Room 5303, Life Sciences
P.O. Box 951766
Los Angeles CA 90095-1766
UNITED STATES