#phylodynamics-jc week 3

Fitting compartment models using coalescent approaches.

Marlin Figgins

August 13, 2020

Week 3 goals

We'll be doing a whirlwind tour of the following three papers in order to paint a picture of the mathematics underlying phylodynamics methods.

- Phylodynamics of Infectious Disease Epidemics
- Complex Population Dynamics and the Coalescent Under Neutrality
- Inferring the Source of Transmission with Phylogenetic Data

Section 1

Volz 2009: Phylodynamics of Infectious Disease Epidemics

Takeaways

“We present a formalism for unifying the inference of infected population sizes from genetic sequences and mathematical models of infectious disease in populations.”

In practice, we are able to fit epidemiological models to a phylogeny of viral sequences and make inferences regarding the disease dynamics.

Methods

There are several practical questions that we seek to answer with these methods.

- If n individuals from a total infected population of $|I|$ are sampled at time T, how many lineages existed at time $t < T$.
- How many of the lineages at time t have surviving progeny at time T?

Coalescent model for SIR

Suppose that we're given a human population of size N, the SIR dynamics of this population are described by the series of differential equations

$$\frac{dS}{dt} = -\beta SI \tag{1}$$

$$\frac{dI}{dt} = \beta SI - \gamma I \tag{2}$$

$$\frac{dR}{dt} = \gamma I, \tag{3}$$

where S, I, R denote the fraction of the population which are susceptible, infected, and recovered respectively.

Probability of observing coalescence

Given a population of size N with k lineages, the probability that these lineages merge is well approximated by $\binom{k}{2} = \frac{k(k-1)}{2}$ if N is large relative to k. Therefore, given a coalescent event occurs between the infected individuals, the probability of observing this event in the n sampled lineages is given by

$$p_C = \binom{n}{2} \Big/ \binom{|I|}{2} = \frac{n(n-1)}{|I|\left(|I|-1\right)}.$$

Probability of sampled ancestors

If we define a function $A(t, T)$ which describes the fraction of the individuals at time t with sampled progeny at T. We'll use this definition alongside our early computation to find the probability of a transmission causing us to observe a coalescent event.

$$p_c(t, T) = \left(\frac{A(t, T)}{I(t)} \right)^2 ,$$

since the total number of lineages in a population with total size N is given by $A(t, T)N$ and the number of infected individuals is $I(t)N$. Therefore, we can compute the function $A(t, T)$ using the following ODE:

$$-\frac{dA}{dt} = -\beta SI \cdot \left(\frac{A(t, T)}{I(t)} \right)^2 .$$

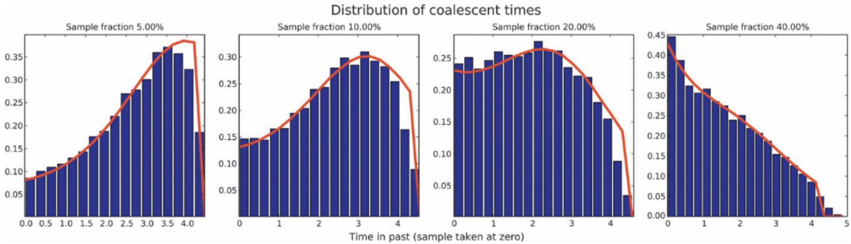
Distribution of coalescent events

The ancestor function allows us to define the fraction of coalescent events which have occurred by time τ between times t and T as

$$\mathbb{P}(t < \tau < T) = F(\tau) = \frac{A(T,T) - A(\tau,T)}{A(T,T) - A(t,T)}.$$

This forms a cumulative distribution function of coalescent times. Differentiating then gives the probability density function

$$f(\tau) = -\frac{dA}{dt}(\tau) \cdot \frac{1}{A(T,T) - A(t,T)}.$$



Fitting epidemic models to sequence data

Suppose we're given branching times $t_1, t_2, \ldots, t_{n-1}$ for a phylogeny of n sequences. Then, we can write a log-likelihood for our branching times as

$$\begin{aligned}\Lambda(t_1, \ldots, t_{n-1} \mid \theta) &= \sum_{i=1}^{n-1} \log(f(\tau)) && (4)\\ &= \sum_{i=1}^{n-1} \log\left(-\frac{dA}{dt}(t_i)\right) - (n-1)\log(A(T,T) - A(t,T)) && (5)\end{aligned}$$

Fitting epidemic models to sequence data

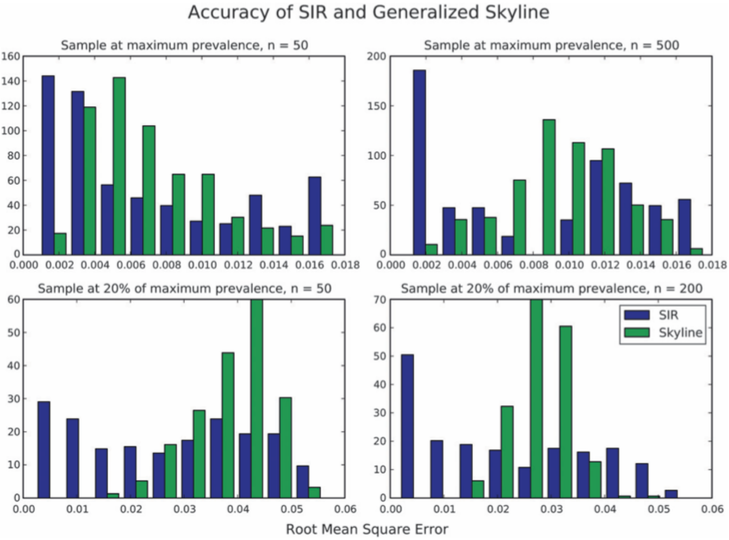


FIGURE 3.—Root mean square error of SIR and generalized skyline estimates of epidemic prevalence. Data are based on 300 simulated epidemics ($R_0 = 2$). RMSE is averaged over 100 time points.

Application: Fitting to HIV-1 sequences

Using the likelihood function defined above, the authors fit an SIR to a phylogeny of 55 HIV-1 sequences sampled in 1993. This involved using modified infection dynamics:

$$\frac{dS}{dt} = \mu - S^{\alpha}(\beta_1 + I_1 + \beta_2 I_2) - \mu S \tag{6}$$

$$\frac{dI_1}{dt} = S^{\alpha}(\beta_1 + I_1 + \beta_2 I_2) - \gamma_1 I_1 - \mu I_1 \tag{7}$$

$$\frac{dI_2}{dt} = \gamma_1 I_1 - \gamma_2 I_2 - \mu I_2. \tag{8}$$

Here, $\beta_{\cdot}, \gamma_{\cdot}$ are transmission rates and recovery rates respectively. A subscript of 1 refers to those individuals with an acute infection and subscript 2 refers to chronic infection.

Application: Fitting to HIV-1 sequences.

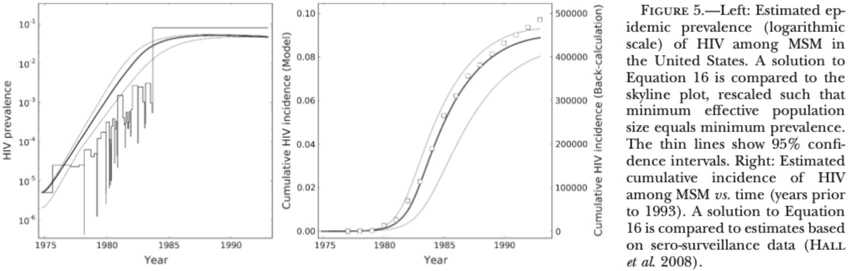


FIGURE 5.—Left: Estimated epidemic prevalence (logarithmic scale) of HIV among MSM in the United States. A solution to Equation 16 is compared to the skyline plot, rescaled such that minimum effective population size equals minimum prevalence. The thin lines show 95% confidence intervals. Right: Estimated cumulative incidence of HIV among MSM *vs.* time (years prior to 1993). A solution to Equation 16 is compared to estimates based on sero-surveillance data (HALL *et al.* 2008).

Figure 3: HIV prevelance and model fit.

Section 2

Volz 2012: Complex Population Dynamics and the Coalescent Under Neutrality

Takeaways

“A coalescent model is developed for a large class of populations such that the demographic history is described by a deterministic nonlinear dynamical system of arbitrary dimension. This class of demographic model differs from those typically used in population genetics. Birth and death rates are not fixed, and no assumptions are made regarding the fraction of the population sampled.”"

Approaching this from the perspective of a birth-death process. Useful for generating lineages.

As shown in Volz 2009, the rate of of coalescence for two extant lineages is:

$$\lambda_2(t) = \frac{2f(t)}{I(t)^2}$$

In the case of the simple SIR, this reduces to the familiar

$$\lambda_2(t) = \frac{2\beta S(t)}{I(t)}$$

Varying birthrate: Skyline estimates of effective population size will be biased for true population size

This is section “The effective number of infections”

Calculating the likelihood of a genealogy condition on f(s) and I(s)

Test

We can. . .

Test