Reconstructing transmission trees from genetic data: a Bayesian approach

In alphabetic order:
Simon Cauchemez, Anne Cori, Xavier Didelot,
Neil Ferguson, Christophe Fraser, Thibaut Jombart,
...

May 11, 2012

Purpose of the model

We seek a probabilistic model allowing to reconstruct the transmission tree of a disease outbreak based on RNA/DNA sequences sampled at given time points. We consider a single pathogen and genetic sequence per infection. The generation time is assumed to follow a known distribution. The transmission tree and the mutation rates are the quantities we want to infer.

Data and parameters

Data

For each patient $i = 1, \dots, n$ we note the data:

- s_i: the genetic sequence obtained for patient i.
- t_i: the collection time for s_i (time is considered as a discrete variable).

Augmented data

Augmented data are noted using capital latin letters:

- T_i^{inf}: time at which patient i has been infected.
- A_i: the closest observed ancestor of i in the infection tree; $A_i = j$ indicates that j has infected i, either directly, or with one or several intermediate generations, which were unobserved.
- K_i: an integer ≥ 1 indicating how many generations separate A_i and i: $K_i = 2$ indicates that j has infected an unobserved individual, who has infected i.

As a first simple approach, K_i could be set to 1 for all i, hence assuming that the whole outbreak was observed.

Functions

We use the following functions of the data/augmented data:

- $d(i,j)$: the number of transitions between s_i and s_j.
- $g(i,j)$: the number of transversions between s_i and s_j.
- $l(i,j)$: the number of nucleotide positions typed in both s_i and s_j.
- $w(\Delta_t)$: generation time distribution (likelihood function for a secondary infection occuring Δ_t unit times after the primary infection); we assume $w(\Delta_t) = 0$ for $\Delta_t \leq 0$.

Parameters

Parameters are indicated using greek letters:

- μ_1: rates of transitions, given per site and unit time (likely day).
- μ_2: rate of transversions, parametrised as $\mu_2 = \kappa\mu_1$ to account for the correlation between the two rates.

Model

This model assumes that cases are ordered by increasing infection dates ($T_i^{inf} \leq T_{i+1}^{inf}$). The posterior distribution is proportional to:

$$\begin{aligned} & p(\{s_i, t_i, T_i^{inf}, A_i, K_i\}_{(i=1,\dots,n)}, w, \mu_1, \kappa) && (1) \\ = & \prod_{i=2}^{n} p(s_i, t_i, T_i^{inf}, A_i, K_i, w, \mu_1, \kappa | \{s_k, t_k, T_k^{inf}, A_k, K_k\}_{(k=1,\dots,i-1)}) \\ & \times p(s_1, t_1, T_1^{inf}, A_1, K_1, w, \mu_1, \kappa) && (2) \\ = & \prod_{i=2}^{n} p(s_i, t_i, T_i^{inf}, A_i, K_i, w, \mu_1, \kappa | s_{A_i}, t_{A_i}, T_{A_i}^{inf}) \\ & \times p(s_1, t_1, T_1^{inf}, A_1, K_1, w, \mu_1, \kappa) && (3) \end{aligned}$$

The term for case i ($i = 2, \dots, n$) is:

$$p(s_i, t_i, T_i^{inf}, A_i, K_i, w, \mu_1, \kappa | s_{A_i}, t_{A_i}, T_{A_i}^{inf}) \quad (4)$$

which can be decomposed into:

$$\begin{aligned} & p(s_i | t_i, T_i^{inf}, A_i, s_{A_i}, t_{A_i}, T_{A_i}^{inf}, K_i, w, \mu_1, \kappa) p(t_i | T_i^{inf}, A_i, s_{A_i}, t_{A_i}, T_{A_i}^{inf}, K_i, w, \mu_1, \kappa) \\ & \times p(T_i^{inf} | A_i, s_{A_i}, t_{A_i}, T_{A_i}^{inf}, K_i, w, \mu_1, \kappa) p(A_i, K_i, w, \mu_1, \kappa | s_{A_i}, t_{A_i}, T_{A_i}^{inf}) \\ = & \underbrace{p(s_i | t_i, T_i^{inf}, A_i, s_{A_i}, t_{A_i}, \mu_1, \kappa)}_{\Omega_i^1} \underbrace{p(t_i | T_i^{inf}, w) p(T_i^{inf} | A_i, T_{A_i}^{inf}, K_i, w)}_{\Omega_i^2} \underbrace{p(A_i, K_i, w, \mu_1, \kappa | s_{A_i}, t_{A_i}, T_{A_i}^{inf})}_{\Omega_i^3} \quad (5) \end{aligned}$$

where Ω_i^1 is the genetic likelihood, Ω_i^2 if the epidemiological likelihood (from W&T), and Ω_i^3 is mixture of constants and priors.

Ω_i^1 is computed as:

$$\underbrace{\mathcal{B}\left(d(i,A_i)|(t_i - t_{A_i})l(i,A_i),\mu_1\right)}_{\text{transitions}} \times \underbrace{\mathcal{B}\left(g(i,A_i)|(t_i - t_{A_i})l(i,A_i),\kappa\mu_1\right)}_{\text{transversions}} \tag{6}$$

if $t_{A_i} \leq T_i^{inf}$, and as:

$$\underbrace{\mathcal{B}\left(d(i,A_i)|(t_{A_i} - T_i^{inf} + t_i - T_i^{inf})l(i,A_i),\mu_1\right)}_{\text{transitions}} \times \underbrace{\mathcal{B}\left(g(i,A_i)|(t_{A_i} - T_i^{inf} + t_i - T_i^{inf})l(i,A_i),\kappa\mu_1\right)}_{\text{transversions}} \tag{7}$$

otherwise; $\mathcal{B}(.|n,p)$ is the probability mass function of a Binomial distribution with n draws and a probability p.

Ω_i^2 is defined by the (known) distribution of the generation time and the collection time:

$$\begin{aligned} \Omega_i^2 &= p(t_i|T_i^{inf}, w) \times p(T_i^{inf}|A_i, T_{A_i}^{inf}, K_i, w) \\ &= \mathbf{1}_{\{w(t_i - T_i^{inf})>0\}} \times w^{(K_i)}(T_i^{inf} - T_{A_i}^{inf}) \end{aligned} \tag{8}$$

with $\mathbf{1}$ the indicator function and $w^{(k)} = \underbrace{w * w * \ldots * w}_{k \text{ times}}$, where $*$ denotes the convolution operator, defined, for two discrete distributions a and b, by $(a*b)(t) = \sum_{s=-\infty}^{+\infty} a(t-s)\,b(s)$. The first term ensures that the augmented infection time (T_i^{inf}) is compatible with the collection time (t_i), while the second term is an extension of Wallinga & Teunis's model for unobserved intermediate infections.

The term Ω_i^3 can be rewritten:

$$\begin{aligned} \Omega_i^3 &= p(A_i, K_i, w, \mu_1, \kappa | s_{A_i}, t_{A_i}, T_{A_i}^{inf}) && (9) \\ &= p(A_i, K_i, w, \mu_1, \kappa) && (10) \\ &= p(w)p(A_i)p(K_i)p(\mu_1)p(\kappa) && (11) \end{aligned}$$

as the different components are independent. $p(w)$ is a constant and does not need to be known to sample from (1). $p(A_i)$ is the prior on ancestries, set to $1/(n-1)$. $p(K_i)$ is the prior on the number of generations from closest ancestries. If we assume that then entire outbreak has been sampled, this would be set to $p(K_i) = \mathbf{1}_{\{K_i=1\}}$. Alternatively, more flexibility would be gained by using a Poisson distribution to allow for unobserved intermediate cases. $p(\mu_1)$ and $p(\kappa)$ are the priors for these two parameters.