Reconstructing transmission trees from genetic data: a Bayesian approach

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

June 7, 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, hence no within-host diversity. We also assume that all cases but the first one trace their ancestry back within the system studied (i.e., no infection from the outside except for the initial case). The generation time is assumed to follow a known distribution. The transmission tree and the mutation rates are the elements we want to infer.

Data and parameters

Data

For each patient $i = 1, \ldots, n$ 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.
- T_i^{ini}: time at which the most recent observed ancestor of i caused the initial infection of the lineage of i.

As a first simple approach, κ_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 $\Delta_t \leq 0$; while not a requirement in theory, in practice this function will be truncated at a value Δ_{max} so that $w(\Delta_t) = 0$ if $\Delta_t \geq \Delta_{max}$.

Parameters

Parameters are indicated using greek letters:

- α_i: the closest observed ancestor of i in the infection tree; $\alpha_i = j$ indicates that j has infected i, either directly, or with one or several intermediate generations, which were unobserved. We note the tree topology $\alpha = \{\alpha_2, \ldots, \alpha_n\}$.
- κ_i: an integer ≥ 1 indicating how many generations separate α_i and i: $\kappa_i = 1$ indicates that α_i infected i; $\kappa_i = 2$ indicates that α_i has infected an unobserved individual, who has in turn infected i. We note $\kappa = \{\kappa_2, \ldots, \kappa_n\}$.
- μ_1: rates of transitions, given per site and unit time (likely day).
- μ_2: rate of transversions, parametrised as $\mu_2 = \gamma\mu_1$ (with $\gamma \in \mathbb{R}_+$) to account for the correlation between the two rates.

Model

Likelihood

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 the joint distribution:

$$p(\{s_i, t_i, T_i^{inf}, T_i^{ini}\}_{(i=1,\ldots,n)}, w, \alpha, \kappa, \mu_1, \gamma) \tag{1}$$

$$= p(\{s_i, t_i, T_i^{inf}, T_i^{ini}\}_{(i=1,\ldots,n)} | w, \alpha, \kappa, \mu_1, \gamma) \times p(w, \alpha, \kappa, \mu_1, \gamma) \tag{2}$$

where the first term is the likelihood of observed and augmented data, and the second, the prior. The likelihood can be decomposed as:

$$p(\{s_i, t_i, T_i^{inf}, T_i^{ini}\}_{(i=1,\ldots,n)} | w, \alpha, \kappa, \mu_1, \gamma) \tag{3}$$

$$= \prod_{i=2}^{n} p(s_i, t_i, T_i^{inf}, T_i^{ini} | \{s_k, t_k, T_k^{inf}, T_k^{ini}\}_{(k=1,\ldots,i-1)}, w, \alpha, \kappa, \mu_1, \gamma)$$
$$\times p(s_1, t_1, T_1^{inf}, T_1^{ini}) \tag{4}$$

$$= \prod_{i=2}^{n} p(s_i, t_i, T_i^{inf}, T_i^{ini} | s_{\alpha_i}, t_{\alpha_i}, T_{\alpha_i}^{inf}, T_{\alpha_i}^{ini}, w, \alpha_i, \kappa_i, \mu_1, \gamma)$$
$$\times p(s_1, t_1, T_1^{inf}, T_1^{ini}) \tag{5}$$

The term $p(s_1, t_1, T_1^{inf}, T_1^{ini})$ is the probability of the data in the first case, treated as a constant. This will need to be modified if we explicitely model infections from outside the system. The term for case i ($i = 2, \ldots, n$) is:

$$p(s_i, t_i, T_i^{inf}, T_i^{ini} | s_{\alpha_i}, t_{\alpha_i}, T_{\alpha_i}^{inf}, T_{\alpha_i}^{ini}, w, \alpha_i, \kappa_i, \mu_1, \gamma) \tag{6}$$

which can be decomposed into:

$$\begin{aligned} & p(s_i|t_i, T_i^{inf}, T_i^{ini}, s_{\alpha_i}, t_{\alpha_i}, T_{\alpha_i}^{inf}, T_{\alpha_i}^{ini}, w, \alpha_i, \kappa_i, \mu_1, \gamma) \\ & \times p(t_i|T_i^{inf}, T_i^{ini}, s_{\alpha_i}, t_{\alpha_i}, T_{\alpha_i}^{inf}, T_{\alpha_i}^{ini}, w, \alpha_i, \kappa_i, \mu_1, \gamma) \\ & \times p(T_i^{inf}|T_i^{ini}, s_{\alpha_i}, t_{\alpha_i}, T_{\alpha_i}^{inf}, T_{\alpha_i}^{ini}, w, \alpha_i, \kappa_i, \mu_1, \gamma) \\ & \times p(T_i^{ini}|s_{\alpha_i}, t_{\alpha_i}, T_{\alpha_i}^{inf}, T_{\alpha_i}^{ini}, w, \alpha_i, \kappa_i, \mu_1, \gamma) \qquad (7) \\ = \; & \underbrace{p(s_i|t_i, T_i^{ini}, \alpha_i, s_{\alpha_i}, t_{\alpha_i}, \mu_1, \gamma)}_{\Omega_i^1} \\ & \times \underbrace{p(t_i|T_i^{inf}, w) p(T_i^{inf}|\kappa_i, T_i^{ini}, w) p(T_i^{ini}|T_{\alpha_i}^{inf}, w)}_{\Omega_i^2} \qquad (8) \end{aligned}$$

where Ω_i^1 is the genetic likelihood and Ω_i^2 if the epidemiological likelihood (derived from W&T).

Assuming that there is no within-host diversity, Ω_i^1 is computed as:

$$\underbrace{\mathcal{B}\left(d(i,\alpha_i)|(t_i - t_{\alpha_i})l(i,\alpha_i), \mu_1\right)}_{\text{transitions}} \times \underbrace{\mathcal{B}\left(g(i,\alpha_i)|(t_i - t_{\alpha_i})l(i,\alpha_i), \gamma\mu_1\right)}_{\text{transversions}} \tag{9}$$

if $t_{\alpha_i} \leq T_i^{ini}$, and as:

$$\underbrace{\mathcal{B}\left(d(i,\alpha_i)|(t_{\alpha_i} - T_i^{ini} + t_i - T_i^{ini})l(i,\alpha_i), \mu_1\right)}_{\text{transitions}} \times \underbrace{\mathcal{B}\left(g(i,\alpha_i)|(t_{\alpha_i} - T_i^{ini} + t_i - T_i^{ini})l(i,\alpha_i), \gamma\mu_1\right)}_{\text{transversions}} \tag{10}$$

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 determined by the (known) distribution of the generation time, and the collection and infection dates:

$$\begin{aligned} \Omega_i^2 &= p(t_i|T_i^{inf}, w) \times p(T_i^{inf}|\kappa_i, T_i^{ini}, w) \times p(T_i^{ini}|T_{\alpha_i}^{inf}, w) \\ &= f_w(t_i - T_i^{inf}) \times w^{(\kappa_i - 1)}(T_i^{inf} - T_i^{ini}) \times w(T_i^{ini} - T_{\alpha_i}^{inf}) \end{aligned} \tag{11}$$

where f_w is a function of the generation time distribution (w) indicating how likely it is to sequence an isolate at a given time after infection. By default, we set $f_w = w$, so that the probability of sequencing an isolate is proportional to the infectiousness of the host at this time. $w^{(k)}$ is the probability density function of the time between a primary infection and a subsequent infection k generations later. $w^{(k)} = \underbrace{w * w * \ldots * w}_{k \text{ times}}$, where $*$ denotes the convolution operator, defined, for two discrete distributions a and b, by $(a * b)(t) = \sum_{u=-\infty}^{+\infty} a(t-u)\, b(u)$. By convention, $w^{(0)}(t) = 1$. The third term is the likelihood of the secondary infection.

Priors

For all model parameters, independent prior distributions were chosen:

- $p(w) = \mathbf{1}_{\{w=w_0\}}$: the distribution of the generation time will be fixed to a given distribution (w_0) by default; this can be parameterized later in a more complex model.
- $p(\alpha_i)$: set to $1/(n-1)$.
- $p(\kappa_i - 1) = \mathcal{NB}(1, \pi)$, the probability mass function of a negative binomial distribution counting the number of unobserved cases before one observed case; π is the proportion of unobserved (unsampled) cases in the outbreak. If we assume that the entire outbreak has been sampled, $p(\kappa_i) = \mathbf{1}_{\{\kappa_i=1\}}$.
- $p(\mu_1) = Unif(0, 1)$.
- $p(\gamma) = Unif(0, 100)$.