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 6, 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 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.
- 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: $K_i = 1$ indicates that A_i infected i; $K_i = 2$ indicates that j has infected an unobserved individual, who has in turn infected i.
- T_i^{ini}: time at which A_i caused the initial infection of the lineage of i. For $K_i = 1$, $T_i^{inf} = T_i^{ini}$.

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 $\Delta_t \leq 0$; 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:

- 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 the joint distribution:

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

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

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

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

$$= \prod_{i=2}^{n} p(s_i, t_i, T_i^{inf}, T_i^{ini}, A_i, K_i|\{s_k, t_k, T_k^{inf}, T_k^{ini}, A_k, K_k|w, \mu_1, \kappa\}_{(k=1,\ldots,i-1)})$$

$$\times p(s_1, t_1, T_1^{inf}, T_1^{ini}, A_1, K_1|w, \mu_1, \kappa) \tag{4}$$

$$= \prod_{i=2}^{n} p(s_i, t_i, T_i^{inf}, T_i^{ini}, A_i, K_i|s_{A_i}, t_{A_i}, T_{A_i}^{inf}, T_{A_i}^{ini}, w, \mu_1, \kappa)$$

$$\times p(s_1, t_1, T_1^{inf}, T_1^{ini}, A_1, K_1|w, \mu_1, \kappa) \tag{5}$$

The term $p(s_1, t_1, T_1^{inf}, T_1^{ini}, A_1, K_1|w, \mu_1, \kappa)$ is the probability of observing the first data point given an unknown ancestor, an is assumed to be constant. This may be modified if we explicitely model infections from outside the system. The term for case i ($i = 2, \ldots, n$) is:

$$p(s_i, t_i, T_i^{inf}, T_i^{ini}, A_i, K_i|s_{A_i}, t_{A_i}, T_{A_i}^{inf}, T_{A_i}^{ini}, w, \mu_1, \kappa) \tag{6}$$

which can be decomposed into:

$$
\begin{aligned}
& p(s_i|t_i, T_i^{inf}, T_i^{ini}, A_i, K_i, s_{A_i}, t_{A_i}, T_{A_i}^{inf}, T_{A_i}^{ini}, w, \mu_1, \kappa) \\
& \times p(t_i|T_i^{inf}, T_i^{ini}, A_i, K_i, s_{A_i}, t_{A_i}, T_{A_i}^{inf}, T_{A_i}^{ini}, w, \mu_1, \kappa) \\
& \times p(T_i^{inf}|T_i^{ini}, A_i, K_i, s_{A_i}, t_{A_i}, T_{A_i}^{inf}, T_{A_i}^{ini}, w, \mu_1, \kappa) \\
& \times p(T_i^{ini}|A_i, K_i, s_{A_i}, t_{A_i}, T_{A_i}^{inf}, T_{A_i}^{ini}, w, \mu_1, \kappa) \\
& \times p(A_i, K_i|s_{A_i}, t_{A_i}, T_{A_i}^{inf}, T_{A_i}^{ini}, w, \mu_1, \kappa) \\
= \; & \underbrace{p(s_i|t_i, T_i^{ini}, A_i, s_{A_i}, t_{A_i}, \mu_1, \kappa)}_{\Omega_i^1} \\
& \times \underbrace{p(t_i|T_i^{inf}, w) p(T_i^{inf}|A_i, K_i, T_{A_i}^{inf}, w) p(T_i^{ini}|A_i, K_i, T_{A_i}^{inf}, w)}_{\Omega_i^2} \\
& \times \underbrace{p(A_i, K_i|s_{A_i}, t_{A_i}, T_{A_i}^{inf}, T_{A_i}^{ini}, w, \mu_1, \kappa)}_{\Omega_i^3}
\end{aligned}
\tag{7}
$$

where Ω_i^1 is the genetic likelihood, Ω_i^2 if the epidemiological likelihood (derived from W&T), and Ω_i^3 are priors for augmented data.

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

$$
\underbrace{\mathcal{B}\left(d(i, A_i)|(t_i - t_{A_i}) l(i, A_i), \mu_1\right)}_{\text{transitions}} \times \underbrace{\mathcal{B}\left(g(i, A_i)|(t_i - t_{A_i}) l(i, A_i), \kappa\mu_1\right)}_{\text{transversions}} \tag{8}
$$

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

$$
\underbrace{\mathcal{B}\left(d(i, A_i)|(t_{A_i} - T_i^{ini} + t_i - T_i^{ini}) l(i, A_i), \mu_1\right)}_{\text{transitions}} \times \underbrace{\mathcal{B}\left(g(i, A_i)|(t_{A_i} - T_i^{ini} + t_i - T_i^{ini}) l(i, A_i), \kappa\mu_1\right)}_{\text{transversions}} \tag{9}
$$

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 and infection dates:

$$
\begin{aligned}
\Omega_i^2 &= p(t_i|T_i^{inf}, w) \times p(T_i^{inf}|A_i, K_i, T_{A_i}^{inf}, w) \times p(T_i^{ini}|A_i, K_i, T_{A_i}^{inf}, w) \\
&= w(t_i - T_i^{inf}) \times w^{(K_i)}(T_i^{inf} - T_{A_i}^{inf}) \times w(T_i^{ini} - T_{A_i}^{inf})^{\mathbf{1}_{\{K_i > 1\}}}
\end{aligned}
\tag{10}
$$

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 $(a * b)(t) = \sum_{u=-\infty}^{+\infty} a(t-u)\, b(u)$. The first term assumes that the probability of sequencing an isolate at time t_i is proportional to the infectiousness of the host at this time. The second term is an extension of Wallinga & Teunis's model for $K_i + 1$ unobserved intermediate infections. The third term is a strict application of Wallinga & Teunis's model for the initial

infection (when $K_i > 1$).

The term Ω_i^3 can be rewritten as:

$$\begin{aligned} \Omega_i^3 &= p(A_i, K_i | s_{A_i}, t_{A_i}, T_{A_i}^{inf}, T_{A_i}^{ini}, w, \mu_1, \kappa) && (11) \\ &= p(A_i)p(K_i) && (12) \end{aligned}$$

as the different components are independent. $p(A_i)$ is the prior on ancestries, set to $1/(n-1)$. $p(K_i)$ is the prior on the number of unobserved transmission steps. This is given by a binomial distribution of parameter π, which is the proportion of observed (sampled) cases in the outbreak. If we assume that the entire outbreak has been sampled, this would be set to $p(K_i) = \mathbf{1}_{\{K_i=1\}}$.