Reconstructing transmission trees from genetic data: a Bayesian approach

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

September 13, 2012

The model, in a nutshell

We seek a simple probabilistic model allowing to reconstruct the transmission tree (who infected whom) of disease outbreaks based on RNA/DNA sequences sampled at given time points. This model is designed for densely sampled outbreaks of diseases with fairly short (epidemiological) generation times and moderate genetic diversity (typically, genomes should accumulate zero, one or maybe two mutations per generation of infection). For instance, the method should be relevant for influenza, but HIV is clearly out of the scope of the approach.

The model is inspired by *SeqTrack* in some of the key assumptions it makes:

- within-host evolution is considered negligible and mutations only occur during transmission events
- a single pathogen is considered for each patient (no multi-infection, no within-host diversity)
- reverse mutations are negligible

However, our model aims at improving *SeqTrack* in several respects:

- a Bayesian framework allowing parameter estimation and incorporating prior information
- the use of the generation time to compute the likelihood (cf Wallinga & Teunis)
- the ability to accommodate unobserved cases
- the incorporation of infection dates in the transmission model (as augmented data)
- the ability to incorporate multiple index cases

In a first approach, we assume that the generation time follows a known distribution. The elements we aim to infer are the transmission tree, the dates of infections, and the mutation rates.

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.
- α_i: the most recent ancestor (MRA) of i in the sample; $\alpha_i = j$ indicates that j has infected i, either directly, or with one or several intermediate generations, which were unobserved. Imported cases are coded as $\alpha_i = 0$, and are considered as fixed and known. We note the tree topology $\alpha = \{\alpha_1, \dots, \alpha_n\}$.
- 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 $\kappa = \{\kappa_1, \dots, \kappa_n\}$.

Functions

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

- $d(s_i, s_j)$: the number of transitions between s_i and s_j.
- $g(s_i, s_j)$: the number of transversions between s_i and s_j.
- $l(s_i, s_j)$: 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 $\Delta_t \geq \Delta_{max}$.
- f_w: 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 the time of collection.

Parameters

This model assumes that cases are ordered by increasing infection dates ($T_i^{inf} \leq T_{i+1}^{inf}$). Parameters are indicated using greek letters:

- μ_1: rates of transitions, given per site and per transmission event.
- 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.
- π: the proportion of observed cases.

Model

Likelihood

The posterior distribution is proportional to the joint distribution:

$$
\begin{aligned}
& p(\{s_i, t_i, T_i^{inf}\}_{(i=1,\dots,n)}, \alpha, \kappa, w, \mu_1, \gamma, \pi) && (1)\\
= \quad & p(\{s_i, t_i, T_i^{inf}, \alpha_i, \kappa_i\}_{(i=1,\dots,n)} | w, \mu_1, \gamma, \pi) \times p(w, \mu_1, \gamma, \pi) && (2)
\end{aligned}
$$

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

$$
\begin{aligned}
& p(\{s_i, t_i, T_i^{inf}, \alpha_i, \kappa_i\}_{(i=1,\dots,n)} | w, \mu_1, \gamma, \pi) && (3)\\
= \quad & \prod_{i=2}^{n} p(s_i, t_i, T_i^{inf}, \alpha_i, \kappa_i | \{s_k, t_k, T_k^{inf}\}_{(k=1,\dots,i-1)}, w, \mu_1, \gamma, \pi) \times p(s_1, t_1, T_1^{inf}, \alpha_1, \kappa_1 | w, \pi) && (4)\\
= \quad & \prod_{i=2}^{n} p(s_i, t_i, T_i^{inf}, \alpha_i, \kappa_i | s_{\alpha_i}, t_{\alpha_i}, T_{\alpha_i}^{inf}, w, \mu_1, \gamma, \pi) \times p(s_1, t_1, T_1^{inf}, \alpha_1, \kappa_1 | w, \pi) && (5)\\
= \quad & \prod_{i=2}^{n} p(s_i, t_i, T_i^{inf}, \alpha_i, \kappa_i | s_{\alpha_i}, t_{\alpha_i}, T_{\alpha_i}^{inf}, w, \mu_1, \gamma, \pi) && (6)\\
& \times p(t_1 | T_1^{inf}, w) p(\alpha_1) p(s_1) p(T_1^{inf}) p(\kappa_1) && (7)
\end{aligned}
$$

$p(t_1|T_1^{inf}, w)$ is the probability of the first collection time given the first infection time. The term $p(\alpha_1)p(s_1)p(T_1^{inf})p(\kappa_1)$ is treated as a constant. The term for case i ($i = 2, \dots, n$) is:

$$
p(s_i, t_i, T_i^{inf}, \alpha_i, \kappa_i | s_{\alpha_i}, t_{\alpha_i}, T_{\alpha_i}^{inf}, w, \mu_1, \gamma, \pi) \quad (8)
$$

which can be decomposed into:

$$
\begin{aligned}
& p(s_i | t_i, T_i^{inf}, \alpha_i, \kappa_i, s_{\alpha_i}, t_{\alpha_i}, T_{\alpha_i}^{inf}, w, \mu_1, \gamma, \pi)\\
& \times p(t_i | T_i^{inf}, \alpha_i, \kappa_i, s_{\alpha_i}, t_{\alpha_i}, T_{\alpha_i}^{inf}, w, \mu_1, \gamma, \pi)\\
& \times p(T_i^{inf} | \alpha_i, \kappa_i, s_{\alpha_i}, t_{\alpha_i}, T_{\alpha_i}^{inf}, w, \mu_1, \gamma, \pi)\\
& \times p(\kappa_i | s_{\alpha_i}, t_{\alpha_i}, T_{\alpha_i}^{inf}, w, \mu_1, \gamma, \pi)\\
= \quad & \underbrace{p(s_i | \alpha_i, s_{\alpha_i}, \kappa_i, \mu_1, \gamma)}_{\Omega_i^1} \times \underbrace{p(t_i | T_i^{inf}, w) p(T_i^{inf} | \alpha_i, T_{\alpha_i}^{inf}, \kappa_i, w) p(\kappa_i | \pi)}_{\Omega_i^2} && (9)
\end{aligned}
$$

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

As mutations only occur during transmission events, the expected divergence between two isolates is determined by the number of generations separating these two isolates, and Ω_i^1 is computed as (cf Kimura 1980):

$$
\underbrace{\mathcal{B}\left(d(s_i, s_{\alpha_i}) | l(s_i, s_{\alpha_i})\kappa_i, \mu_1\right)}_{\text{transitions}} \times \underbrace{\mathcal{B}\left(g(s_i, s_{\alpha_i}) | l(s_i, s_{\alpha_i})\kappa_i, \gamma\mu_1\right)}_{\text{transversions}} \quad (10)
$$

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

$$
\underbrace{\mathcal{P}\left(d(s_i, s_{\alpha_i}) | l(s_i, s_{\alpha_i})\kappa_i\mu_1\right)}_{\text{transitions}} \times \underbrace{\mathcal{P}\left(g(s_i, s_{\alpha_i}) | l(s_i, s_{\alpha_i})\kappa_i\gamma\mu_1\right)}_{\text{transversions}} \quad (11)
$$

where $\mathcal{P}(.|\lambda)$ is the density of a Poisson distribution of parameter λ. In the absence of genetic information (including imported cases where $\alpha_i = 0$), $\Omega_i^1 = 1$.

Ω_i^2 is determined by the distribution of the generation time, the dates of collection and infection, and the proportion of unobserved or unsampled cases. It is computed as:

$$\begin{aligned}\Omega_i^2 &= p(t_i|T_i^{inf}, w) \times p(T_i^{inf}|\alpha_i, T_{\alpha_i}^{inf}, \kappa_i, w) \times p(\kappa_i|\pi) \\ &= f_w(t_i - T_i^{inf}) \times w^{(\kappa_i)}(T_i^{inf} - T_{\alpha_i}^{inf}) \times f_{\mathcal{NB}}(1|\kappa_i - 1, \pi)\end{aligned} \tag{12}$$

where the first term is the likelihood of the collection date, the second, the likelihood of the infection time, the third, the probability of external cases, and the last, the probability of unobserved intermediate cases. $w^{(k)} = \underbrace{w * w * \ldots * w}_{k \text{ times}}$, where $*$ denotes the convolution operator, defined, for two positive discrete distributions a and b, by $(a * b)(t) = \sum_{u=0}^{t} a(t-u)\, b(u)$. $f_{\mathcal{NB}}(1|r, p)$ is the probability mass function of a negative binomial distribution indicating the probability of getting 1 success after r failures, with a probability of success p (here, "*success*" refers to successful sampling of a case, and "failure" to an unsampled case). In the case of imported cases ($\alpha_i = 0$), Ω_i^2 is simply defined by the collection date, with a uniform distribution probability for T_i^{inf} over the the time span of the outbreak:

$$\Omega_i^2 = p(t_i|T_i^{inf}, w) \times p(T_i^{inf}|\alpha_i, T_{\alpha_i}^{inf}, \kappa_i, w) = f_w(t_i - T_i^{inf})\frac{1}{D} \tag{13}$$

where D is the fixed time span of the outbreak (approximated by the timespan of the collection dates).

Priors

For all model parameters, independent prior distributions have been chosen:

- $p(w) = \mathbf{1}_{\{w=w_0\}}$: the distribution of the generation time is fixed to a given distribution (w_0) by default; this can be parameterized later in a more complex model.
- $p(\mu_1) = Unif(0, 1)$.
- $p(\gamma) = log\mathcal{N}(1, 1.25)$ where $log\mathcal{N}(\mu, \sigma)$ is the log-normal distribution with mean μ and standard deviation σ (values are the default in BEAST).
- $p(\pi) = \beta(a, b)$: the proportion of unobserved cases is assumed to follow a Beta distribution of fixed parameters (e.g. $p(\pi) = \beta(10, 1)$ for an a priori densely sampled outbreak.