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 8, 2012

The model, in a nutshell

We seek a simple probabilistic model allowing to reconstruct the transmission tree (who infected whom) of a densely sampled disease outbreak based on RNA/DNA sequences sampled at given time points. This model is designed for densely sampled outbreaks with fairly short 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
- 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)

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)

In a first approach, we assume that the generation time follows a known distribution. This could be relaxed in a more complex model where parameters of this distribution would be estimated. The elements we aim to infer are the transmission tree and the mutation rates.

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.

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:

- 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 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.

Model

Likelihood

The posterior distribution is proportional to the joint distribution:

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

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

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

The term $p(s_1, t_1, T_1^{inf})$ is the probability of the data of 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 ($i = 2, \dots, n$) is:

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

which can be decomposed into:

$$\begin{align} & p(s_i | t_i, T_i^{inf}, s_{\alpha_i}, t_{\alpha_i}, T_{\alpha_i}^{inf}, w, \alpha_i, \kappa_i, \mu_1, \gamma) \notag\\ & \times p(t_i | T_i^{inf}, s_{\alpha_i}, t_{\alpha_i}, T_{\alpha_i}^{inf}, w, \alpha_i, \kappa_i, \mu_1, \gamma) \notag\\ & \times p(T_i^{inf} | s_{\alpha_i}, t_{\alpha_i}, T_{\alpha_i}^{inf}, w, \alpha_i, \kappa_i, \mu_1, \gamma) \notag\\ = \quad & \underbrace{p(s_i | \alpha_i, s_{\alpha_i}, \kappa_i, \mu_1, \gamma)}_{\Omega_i^1} \times \underbrace{p(t_i | T_i^{inf}, w) p(T_i^{inf} | \alpha_i, T_{\alpha_i}^{inf}, \kappa_i, w)}_{\Omega_i^2} \tag{7} \end{align}$$

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):

$$\underbrace{\mathcal{B}\left(d(s_i, s_{\alpha_i}) | l(s_i, s_{\alpha_i})\kappa_i, \mu_1\right)}_{\text{transitions}} \times \underbrace{\mathcal{B}\left(g(s_i, s_{\alpha_i}) | l(s_i, s_{\alpha_i})\kappa_i, \gamma\mu_1\right)}_{\text{transversions}} \tag{8}$$

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

$$\underbrace{\mathcal{P}\left(d(s_i, s_{\alpha_i}) | l(s_i, s_{\alpha_i})\kappa_i\mu_1\right)}_{\text{transitions}} \times \underbrace{\mathcal{P}\left(g(s_i, s_{\alpha_i}) | l(s_i, s_{\alpha_i})\kappa_i\gamma\mu_1\right)}_{\text{transversions}} \tag{9}$$

where $\mathcal{P}(.|\lambda)$ is the density of a Poisson distribution of parameter λ.

Note that when the genetic likelihood cannot be computed (i.e. s_i or s_j is missing, or the two sequences have no typed nucleotide position in common), it can be replaced by the average likelihood of the other Ω_i^1

Ω_i^2 is determined by the distribution of the generation time, and the dates of collection and infection:

$$\begin{aligned} \Omega_i^2 &= p(t_i | T_i^{inf}, w) \times p(T_i^{inf} | \alpha_i, T_{\alpha_i}^{inf}, \kappa_i, w) \\ &= f_w(t_i - T_i^{inf}) \times w^{(\kappa_i)}(T_i^{inf} - T_{\alpha_i}^{inf}) \qquad (10) \end{aligned}$$

where the first term is the likelihood of the collection date, and the second, the likelihood of the infection time. $w^{(k)} = \underbrace{w * w * \ldots * w}_{k \text{ times}}$, where $*$ denotes the convolution operator, defined, for two positive discrete distributions a and b, by $(a * b)(t) = \sum_{u=0}^{+\infty} a(t-u)\, b(u)$.

Priors

For all model parameters, independent prior distributions have been 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.
- $p(\alpha_i) = \mathbf{1}_{\{k \neq i\}} \frac{1}{n-1}$.
- $p(\kappa_i - 1) = \mathcal{NB}(1, \pi)$, 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, $p(\kappa_i) = \mathbf{1}_{\{\kappa_i = 1\}}$.
- $p(\mu_1) = Unif(0, 1)$.
- $p(\gamma)$ set so that μ_2 has a uniform prior on $[0; 1]$ (to be worked out).