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 $i = 1, \ldots, n$ we note the data:

- s_i: the genetic sequence obtained for patient i
- t_i: the collection time for s_i
- $d_{i,j}$: 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
- $w(\Delta_t)$: likelihood function for a secondary infection occuring Δ_t unit times after the primary infection; we assume $w(\Delta_t) = 0$ for $\Delta_t \leq 0$.

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 parent node 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.

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

Parameters

Parameters are indicated using greek letters:

- μ_1: rates of transitions.
- μ_2: rate of transversions, assumed proportional to μ_1 so that $\mu_2 = \kappa\mu_1$.

Basic model

This model assumes that all ancestries have been sampled. The posterior distribution for patient i is proportional to the joint distribution:

$$p(s_i, t_i, T_i^{inf}, A_i, K_i, w, \mu_1, \mu_2) \tag{1}$$

which can be decomposed in:

$$\begin{aligned}
& p(s_i|t_i, T_i^{inf}, A_i, K_i, w, \mu_1, \mu_2) p(t_i, T_i^{inf}, A_i, K_i, w, \mu_1, \mu_2) && (2)\\
= \; & p(s_i|t_i, T_i^{inf}, A_i, K_i, w, \mu_1, \mu_2) p(T_i^{inf}|t_i, A_i, K_i, w, \mu_1, \mu_2) p(t_i, A_i, K_i, w, \mu_1, \mu_2) && (3)\\
= \; & \underbrace{p(s_i|t_i, T_i^{inf}, A_i, K_i, \mu_1, \mu_2)}_{\Omega_i^1} \underbrace{p(T_i^{inf}|A_i, K_i, w)}_{\Omega_i^2} \underbrace{p(t_i, A_i, K_i, w, \mu_1, \mu_2)}_{\Omega_i^3} && (4)
\end{aligned}$$

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:

$$\underbrace{\mathcal{P}\left(d_{i,A_i}|(t_i - t_{A_i})l_{i,A_i}\mu_1\right)}_{\text{transitions}} \times \underbrace{\mathcal{P}\left(g_{i,A_i}|(t_i - t_{A_i})l_{i,A_i}\mu_2\right)}_{\text{transversions}}$$

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

$$\underbrace{\mathcal{P}\left(d_{i,A_i}|(t_{A_i} - T_i^{inf} + t_i - T_i^{inf})l_{i,A_i}\mu_1\right)}_{\text{transitions}} \times \underbrace{\mathcal{P}\left(g_{i,A_i}|(t_{A_i} - T_i^{inf} + t_i - T_i^{inf})l_{i,A_i}\mu_2\right)}_{\text{transversions}}$$

otherwise; $\mathcal{P}(.|\lambda)$ 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 defined by the (known) distribution of the generation time:

$$\Omega_i^2 = w^{(K_i)}(t_i - t_{A_i})$$

with $w^{(k)} = \underbrace{w * w * \ldots * w}_{k \text{ times}}$, 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 term Ω_i^3 can be rewritten:

$$\begin{aligned} \Omega_i^3 &= p(t_i, A_i, K_i, w, \mu_1, \mu_2) && (5)\\ &= p(t_i, w) p(A_i, K_i, \mu_1, \mu_2) && (6)\\ &= p(t_i, w) p(A_i) p(K_i) p(\mu_1) p(\kappa) && (7) \end{aligned}$$

as the different components are independent. $p(t_i, w)$ 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. In a first simple approach, this could be set to $p(K_i) = \mathbf{1}_{\{K_i=1\}}$, where $\mathbf{1}_{\{.\}}$ denotes the indicator function, defined by $\mathbf{1}_{\{X\}} = 1$ if X is true, and 0 otherwise. In a second approach, a Poisson distribution could be used to allow unobserved intermediate cases. $p(\mu_1)$ and $p(\kappa)$ are the priors for the mutations rates.

The global posterior distribution is proportional to the product of (1) over all individuals:

$$\prod_{i=1}^{n} p(s_i, t_i, T_i^{inf}, A_i, K_i, w, \mu_1, \mu_2) \quad (8)$$