

# 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 10, 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  $i = 1, \dots, n$  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 occurring  $\Delta_t$  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 infector of  $i$ ;  $A_i = j$  indicates that  $j$  has infected  $i$

## Parameters

Parameters are indicated using greek letters:

- $\mu_1$ : rates of transitions.
- $\mu_2$ : rate of transversions, assumed proportional to  $\mu_1$  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:

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

which can be decomposed in:

$$p(s_i|t_i, T_i^{inf}, A_i, w, \mu_1, \mu_2)p(t_i, T_i^{inf}, A_i, w, \mu_1, \mu_2) \quad (2)$$

$$= p(s_i|t_i, T_i^{inf}, A_i, w, \mu_1, \mu_2)p(T_i^{inf}|t_i, A_i, w, \mu_1, \mu_2)p(t_i, A_i, w, \mu_1, \mu_2) \quad (3)$$

$$= \underbrace{p(s_i|t_i, T_i^{inf}, A_i, \mu_1, \mu_2)}_{\Omega_1} \underbrace{p(T_i^{inf}|A_i, w)}_{\Omega_2} \underbrace{p(t_i, A_i, w, \mu_1, \mu_2)}_{\Omega_3} \quad (4)$$

where  $\Omega_1$  is the genetic likelihood,  $\Omega_2$  if the epidemiological likelihood (from W&T), and  $\Omega_3$  is mixture of constants and priors.

$\Omega_1$  is computed as:

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

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

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

otherwise;  $\mathcal{P}(\cdot|\lambda)$  is the density of a Poisson distribution of parameter  $\lambda$ .

$\Omega_2$  is defined by the (known) distribution of the generation time:

$$\Omega_2 = \frac{w(t_i - t_{A_i})}{\sum_{k=1}^n w(t_i - t_{A_k})}$$

The term  $\Omega_3$  can be rewritten:

$$\Omega_3 = p(t_i, A_i, w, \mu_1, \mu_2) \quad (5)$$

$$= p(t_i, w)p(A_i, \mu_1, \mu_2) \quad (6)$$

$$= p(t_i, w)p(A_i)p(\mu_1)p(\kappa) \quad (7)$$

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  $1/(n-1)$ .  $p(\mu_1)$  and  $p(\kappa)$  are the priors for the mutations rates.

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

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

## Model with missing data

!!! Need to work on that one.

The basic model can be extended to incorporate the fact that some infectors have not been sampled, and therefore some infections cannot be reconstructed. We introduce the parameter  $\pi$ , which is the probability of having observed the infector of a given case. Then the expression (1) becomes:

$$p(s_i, t_i, T_i^{inf}, A_i, w, \mu_1, \mu_2, \pi) \quad (9)$$

which can be decomposed in:

$$p(s_i|t_i, T_i^{inf}, A_i, w, \mu_1, \mu_2, \pi) p(t_i, T_i^{inf}, A_i, w, \mu_1, \mu_2, \pi) \quad (10)$$

$$= p(s_i|t_i, T_i^{inf}, A_i, w, \mu_1, \mu_2, \pi) p(T_i^{inf}|t_i, A_i, w, \mu_1, \mu_2, \pi) p(t_i, A_i, w, \mu_1, \mu_2, \pi) \quad (11)$$

$$= \underbrace{p(s_i|t_i, T_i^{inf}, A_i, \mu_1, \mu_2, \pi)}_{\Omega_1^*} \underbrace{p(T_i^{inf}|A_i, w, \pi)}_{\Omega_2^*} \underbrace{p(t_i, A_i, w, \mu_1, \mu_2, \pi)}_{\Omega_3^*} \quad (12)$$

The first two terms are simply given by  $\Omega_1^* = \pi\Omega_1$  and  $\Omega_2^* = \pi\Omega_2$