A model of MRSA transmission incorporating epidemiological and genetic data

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

May 4, 2012

Observed data (Y)

Observed data are all denoted with $*$ as a superscript in order to distinguish them from the augmented data.

For each patient $i = 1, \ldots, N$ admitted to one of the wards in the study period (from time 1 to time T), we denote

- w_i^* the ward where the patient is admitted (1 for adult ICU, 2 for paediatric ICU)
- k_i^* the number of times the patient is admitted (1 if no readmission)
- A_i^* and D_i^* vectors containing the times of admission and discharge from the ward
- P_i^* and N_i^* vectors containing the times of positive and negative swabs (positive defined as any of the samples taken is positive ; negative defined as all samples taken are negative).
- p_i^* and n_i^* the size of those vectors, ie the number of positive and negative swabs.
- $S_i^* = \{s_i^{1^*}, \ldots, s_i^{m_i^*}\}$ a set of m_i^* genetic sequences of MRSA isolated in patient i at times $T_i^* = \{t_i^{1^*}, \ldots, t_i^{m_i^*}\}$; collection dates T_i^* are ordered so that $t_i^{k^*} \leq t_i^{k+1^*}$.
- $d^*_{s_i^k, s_j^q}$ the number of transitions between sequence k isolated in patient i and sequence q isolated in patient j.
- $g^*_{s_i^k, s_j^q}$ the number of transversions between sequence k isolated in patient i and sequence q isolated in patient j.
- $l^*_{s_i^k, s_j^q}$ the number of typed nucleotides common to sequences $s_i^{k^*}$ and $s_j^{q^*}$.

Augmented (unobserved) data (Z)

For each patient i admitted to one of the wards in the study period, we denote

- C_i the colonisation time (we assume no supercolonisations)
- E_i the time of end of colonisation.

- j_i the index if the case who infected patient i (with value -1 if patient i was infected outside the wards and -2 if he was not infected during the period of observation).
- w_i the ward where the patient is admitted (1 for adult ICU, 2 for paediatric ICU)
- k_i the number of times the patient is admitted (1 if no readmission)
- A_i and D_i vectors containing the times of admission and discharge from the ward
- $S_i = \{s_i^1, \ldots, s_i^{m_i}\}$ a set of m_i genetic sequences of MRSA isolated in patient i at times $T_i = \{t_i^1, \ldots, t_i^{m_i}\}$; collection dates T_i are ordered so that $t_i^k \leq t_i^{k+1}$.
- $d_{s_i^k, s_j^q}$ the number of transitions between sequence k isolated in patient i and sequence q isolated in patient j.
- $g_{s_i^k, s_j^q}$ the number of transversions between sequence k isolated in patient i and sequence q isolated in patient j.
- $l_{s_i^k, s_j^q}$ the number of typed nucleotides common to sequences s_i^k and s_j^q.

We denote $I_w(t) = \sum_{i=1}^{N} \mathbf{1}_{\{w_i = w\}} \mathbf{1}_{\{C_i \leq t < E_i\}} \sum_{l=1}^{k_i} \mathbf{1}_{\{A_i[l] \leq t < D_i[l]\}}$ the number of patients in ward w who are colonized at time t.

Parameters (θ)

Parameters of the model are:

- β: a 2 by 2 matrix containing $\beta_{i \leftarrow j}$, the person to person transmission rate from ward j to ward i
- $\lambda_{\text{ward}\leftarrow\text{out}}$: the force of infection from outside the 2 wards applied to patients in the wards
- $\lambda_{\text{out}\leftarrow\text{out}}$: the force of infection from outside the 2 wards applied to patients when they are not in the wards (eg inbetween two admissions)
- ψ: the specificity of the testing, ie the probability of getting a negative test given uncolonized (assumed 100%)
- ϕ: the sensitivity of the testing, ie the probability of getting a positive test given colonized
- π: the probability of being already colonized at first admission
- μ, σ: the mean and standard deviation of the duration of colonization.
- ν_1, ν_2: the rate of transitions ($A \leftrightarrow G$ and $C \leftrightarrow T$) and transversions (other changes) of the DNA sequences. In practice, we will use $\nu_2 = \kappa \nu_1$ with $\kappa \in \mathbb{R}_+$.
- α: the "within-host pathogenic diversity", defined as the number of pathogenic lineages infecting a given patient; α is assumed to follow a Poisson distribution with mean μ_α (hyperparameter); all lineages are considered as likely to be have been sequenced.

Statistical Model

In the following, $\mathbf{1}_{\{.\}}$ denotes the indicator function, defined by $\mathbf{1}_{\{X\}} = 1$ if X is true, and 0 otherwise.

The joint density of the observed data, the augmented data, and the model parameters is:

$$P\left(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\theta}\right) = P\left(\boldsymbol{Y} | \boldsymbol{Z}, \boldsymbol{\theta}\right) P\left(\boldsymbol{Z} | \boldsymbol{\theta}\right) P\left(\boldsymbol{\theta}\right)$$

where $P\left(\boldsymbol{Y} | \boldsymbol{Z}, \boldsymbol{\theta}\right)$, $P\left(\boldsymbol{Z} | \boldsymbol{\theta}\right)$ and $P\left(\boldsymbol{\theta}\right)$ refer to the observation level, the transmission level and the prior level respectively.

Observation level

The observation level ensures that the observed data are consistent with the augmented data:

$$\begin{aligned}
P\left(\boldsymbol{Y} | \boldsymbol{Z}, \boldsymbol{\theta}\right) = \quad \textstyle\prod_{i=1}^{N} \quad & \mathbf{1}_{\{C_i < E_i\}} \\
& \prod_{j=1}^{p_i^*} \left(\left(\mathbf{1}_{\left\{P_i^*[j] < C_i\right\}} + \mathbf{1}_{\left\{P_i^*[j] \geq E_i\right\}} \right) \times (1-\psi) + \mathbf{1}_{\left\{C_i \leq P_i^*[j] < E_i\right\}} \times \phi \right) \\
& \prod_{k=1}^{n_i^*} \left(\left(\mathbf{1}_{\left\{N_i^*[k] < C_i\right\}} + \mathbf{1}_{\left\{N_i^*[k] \geq E_i\right\}} \right) \times \psi + \mathbf{1}_{\left\{C_i \leq N_i^*[k] < E_i\right\}} \times (1-\phi) \right) \\
& \mathbf{1}_{\left\{w_i^* = w_i\right\}} \mathbf{1}_{\left\{k_i^* = k_i\right\}} \mathbf{1}_{\left\{A_i^* = A_i\right\}} \mathbf{1}_{\left\{D_i^* = D_i\right\}} \mathbf{1}_{\left\{m_i^* = m_i\right\}} \mathbf{1}_{\left\{S_i^* = S_i\right\}} \mathbf{1}_{\left\{T_i^* = T_i\right\}} \\
& \prod_{k=1}^{m_i^*} \prod_{j=1}^{N} \prod_{q=1}^{m_j^*} \mathbf{1}_{\left\{ {d^*}_{s_i^k, s_j^q} = d_{s_i^k, s_j^q} \right\}} \mathbf{1}_{\left\{ {g^*}_{s_i^k, s_j^q} = g_{s_i^k, s_j^q} \right\}} \mathbf{1}_{\left\{ {l^*}_{s_i^k, s_j^q} = l_{s_i^k, s_j^q} \right\}}
\end{aligned}$$

The second line describes the positive tests, which can be either false positives (first term) or true positives (second term). The third line describes the negative tests, which can be either true negatives (first term) or false negatives (second term).

Transmission level (discrete time version ; time step = half day or day ?)

In the discrete version $A_i[k]$ is the first time step where individual i is in hospital (for his/her k_{th} stay), and $D_i[k]$ is the first time step where he/she is out of hospital (after his/her k_{th} stay). Individual i can transmit staph aureus from time step C_i to time step $E_i - 1$.

$$
\begin{aligned}
P\left(\boldsymbol{Z}|\boldsymbol{\theta}\right) &= P\left(\left\{C_i,E_i,j_i,w_i,k_i,A_i,D_i,m_i,S_i,T_i\right\}_{i=1\ldots,N},\left\{d_{s_i^k,s_j^q},g_{s_i^k,s_j^q},l_{s_i^k,s_j^q}\right\}_{\substack{i,j=1\ldots N\\k=1\ldots m_i\\q=1\ldots m_j}},\left\{I_w\left(t\right)\right\}_{\substack{w=1,2\\t=1\ldots,T}}|\boldsymbol{\theta}\right)\\
&\propto \prod_{t=1}^{T}\prod_{w=1}^{2}P\left(I_w\left(t\right)|\left\{E_i,C_i,w_i,k_i,A_i,D_i\right\}_{i=1\ldots,N/C_i\leq t},\boldsymbol{\theta}\right)\prod_{i=1}^{N}\mathbf{1}_{\left\{C_i=t\right\}}\\
&\quad\times P\left(E_i|C_i,\boldsymbol{\theta}\right)\\
&\quad\times P\left(S_i|m_i,T_i,\left\{d_{s_i^k,s_j^q},g_{s_i^k,s_j^q},l_{s_i^k,s_j^q}\right\}_{\substack{j=1\ldots N\\k=1\ldots m_i\\q=1\ldots m_j}},C_i,j_i,w_i,k_i,A_i,D_i,S_{j_i},m_{j_i},\boldsymbol{\theta}\right)\\
&\quad\times P\left(j_i|C_i,I_1\left(1\right),\ldots,I_1\left(t-1\right),I_2\left(1\right),\ldots,I_2\left(t-1\right),w_i,k_i,A_i,D_i,\boldsymbol{\theta}\right)\\
&\quad P\left(C_i|I_1\left(1\right),\ldots,I_1\left(t-1\right),I_2\left(1\right),\ldots,I_2\left(t-1\right),w_i,k_i,A_i,D_i,\boldsymbol{\theta}\right)
\end{aligned}
$$

with:

$$P\left(I_w\left(t\right)|\left\{E_i,C_i,w_i,k_i,A_i,D_i\right\}_{i=1\ldots,N/C_i\leq t},\boldsymbol{\theta}\right)=\mathbf{1}_{\left\{I_w(t)=\sum_{i=1}^{N}\mathbf{1}_{\left\{w_i=w\right\}}\mathbf{1}_{\left\{C_i\leq t<E_i\right\}}\sum_{l=1}^{k_i}\mathbf{1}_{\left\{A_i[l]\leq t<D_i[l]\right\}}\right\}}$$

$$P\left(E_i|C_i,\boldsymbol{\theta}\right)=\Phi_{\mu,\sigma}\left(E_i-C_i+0.5\right)-\Phi_{\mu,\sigma}\left(E_i-C_i-0.5\right)$$

where $\Phi_{\mu,\sigma}$ is the cumulative density function of a Gamma distribution with mean μ and standard deviation σ (we assume that the duration of colonisation is Gamma distributed)

$$P\left(S_i|m_i,T_i,\left\{d_{s_i^k,s_j^q},g_{s_i^k,s_j^q},l_{s_i^k,s_j^q}\right\}_{\substack{j=1\ldots N\\k=1\ldots m_i\\q=1\ldots m_j}},C_i,j_i,w_i,k_i,A_i,D_i,S_{j_i},m_{j_i},\boldsymbol{\theta}\right)=f_{i\leftarrow j_i}$$

$f_{i\leftarrow j}$ is the probability of observing sequences S_i at times T_i given that individual j infected individual i and that sequences S_j were observed at times T_j (see next section on the genetic likelihood). Similarly, $f_{\mathrm{i,ward}\leftarrow\mathrm{out}}$ and $f_{\mathrm{i,out}\leftarrow\mathrm{out}}$ are the probability of observing sequence s_i at time t_i given that individual i was infected from outside the wards ($j_i=-1$) while he was in or outsite hospital respectively.

$$P\left(j_i|C_i,I_1\left(1\right),\ldots,I_1\left(t-1\right),I_2\left(1\right),\ldots,I_2\left(t-1\right),w_i,k_i,A_i,D_i,\boldsymbol{\theta}\right)=$$
$$\begin{cases}\frac{\beta_{w_i\leftarrow w_{j_i}}}{\sum_{j\text{ colonized and in a ward at time }C_i}\beta_{w_i\leftarrow w_j}+\lambda_{\text{ward}\leftarrow\text{out}}} & \text{if individual } i \text{ is in a ward at time } C_i \text{ and } j_i\geq 0\\ \frac{\lambda_{\text{ward}\leftarrow\text{out}}}{\sum_{j\text{ colonized and in a ward at time }C_i}\beta_{w_i\leftarrow w_j}+\lambda_{\text{ward}\leftarrow\text{out}}} & \text{if individual } i \text{ is in a ward at time } C_i \text{ and } j_i=-1\\ 1 & \text{if individual } i \text{ is not in a ward at time } C_i \text{ and } j_i=-1\\ 0 & \text{otherwise}\end{cases}$$

HERE PROBLEM WHAT HAPPENS TO THOSE NEVER INFECTED ($j_i=-2$)?

$$P\left(C_i|I_1\left(1\right),\ldots,I_1\left(t-1\right),I_2\left(1\right),\ldots,I_2\left(t-1\right),w_i,k_i,A_i,D_i,\boldsymbol{\theta}\right)=\Omega_i^{(1)}+\Omega_i^{(2)}$$

where

$$\begin{aligned}
\Omega_i^{(1)} &= \pi \times \mathbf{1}_{\{C_i < A_i[1]\}} \\
\Omega_i^{(2)} &= (1-\pi) \times \mathbf{1}_{\{C_i \geq A_i[1]\}} \times e^{-\sum_{t=A_i[1]}^{C_i-1} \lambda_i(t)} \left(1 - e^{-\lambda_i(C_i)}\right)
\end{aligned}$$

$\Omega_i^{(1)}$ is the probability that individual i is colonized before his/her first admission in the wards ; $\Omega_i^{(2)}$ is the probability that individual i is colonized after his/her first admission in the wards. $\lambda_i\left(t\right)$ is the force of transmission applied to individual i at time t. It is equal to:

$$\begin{aligned}
\lambda_i^{(t)} &= \sum_{w=1}^{2} \beta_{w_i \leftarrow w} I_w\left(t\right) + \lambda_{\text{ward}\leftarrow\text{out}} \text{ if individual } i \text{ is in a ward at time } t \\
&= \lambda_{\text{out}\leftarrow\text{out}} \text{ otherwise}
\end{aligned}$$

Genetic likelihood

Ancestors and lineages

We say that B is an *ancestor* of A if and only if there is a path leading from B to A in the directed acyclic graph (DAG) representing the genealogy of A. Put simply, B is an ancestor of A if A derives from B. We will say that B is the *most recent ancestor* (MRA) of A if there is no other ancestor of A observed later in the considered set. A *lineage* is defined as a set of (temporally ordered) individuals $\{x_1, ..., x_n\}$ so that x_i is the MRA of x_{i+1} for $i = 1, \ldots, n-1$. For instance, in the lineage ($D \rightarrow C \rightarrow B \rightarrow A$), B, C, D are all ancestors of A and B is the MRA of A. The genetic likelihood of A given B, C, D will be defined as the probability of the observed mutations between B and A, and is not conditional on previous ancestries. The following statements (and thus the associated probabilities) are equivalent:

- A and B are from the same lineage and B is older than A
- B is an ancestor of A
- A is a descendent of B

Genetic likelihood of an infection

The genetic likelihood of the infection of i by j (noted $i \leftarrow j$) relies on how likely it is to observe the genetic differences between sequences in S_i and their most recent ancestors (MRA) in S_j. We first focus on the probability of observing a given sequence s_i^k in i given that j infected i. We note $\rho(s_i^k, s_j^q)$ the probability of observing s_i^k given an ancestor s_j^q (which is also the probability that s_j^q is an ancestor of s_i^k), defined as:

$$\rho(s_i^k, s_j^q) = \mathbf{1}_{\{t_i^k \geq t_j^q\}} \times \underbrace{\mathcal{P}\left(d_{s_i^k, s_j^q} | \nu_1 (t_i^k - t_j^q) l_{s_i^k, s_j^q}\right)}_{\text{transitions}} \times \underbrace{\mathcal{P}\left(g_{s_i^k, s_j^q} | \nu_2 (t_i^k - t_j^q) l_{s_i^k, s_j^q}\right)}_{\text{transversions}}$$

with:

- $\mathbf{1}_{\{\text{statement}\}}$: indicator function, 1 if 'statement' is true, 0 otherwise
- $\mathcal{P}(.|\lambda)$: the probability mass function of a Poisson distribution with parameter λ

. The three terms respectively correspond to the indicator function ensuring that s_j^q is older than s_i^k, the probability of the observed transitions ($d_{s_i^k,s_j^q}$), and the probability of the observed transversions ($g_{s_i^k,s_j^q}$).

We are now interested in $\xi(s_i^k, s_j^q)$,the probability that the sequence s_j^q is the MRA of s_i^k. This requires two elements: i) that s_j^q is an ancestor of s_i^k, and ii) that no ancestor of s_i^k has been collected after s_j^q. This is given by:

$$\xi(s_i^k, s_j^q) = \underbrace{\rho(s_i^k, s_j^q)}_{s_j^q \text{ ances. of } s_i^k} \times \prod_{r=q+1}^{m_j} \underbrace{(1-\rho(s_i^k, s_j^r))}_{s_j^r \text{ not ances. of } s_i^k}$$

The genetic likelihood also needs to account for the possibility that, due to the sampling process, no ancestor of s_i^k may have been isolated and sequenced in S_j. Assuming that all lineages are as likely to have been sequenced, the probability $\gamma(s_i^k, S_j)$ that S_j contains at least one ancestor of s_i^k is:

$$\gamma(s_i^k, S_j) = 1 - \mathcal{B}\left(0|\sum_{j=1}^{m_j} \mathbf{1}_{\{t_i^k \geq t_j^q\}}, 1/\alpha\right)$$

with:

- $\mathcal{B}(.|n, p)$: probability mass function of the Binomial distribution with n draws and probability p
- $\sum_{j=1}^{m_j} \mathbf{1}_{\{t_i^k \geq t_j^q\}}$: number of isolates sequenced in patient j and collected before the sequence s_i^k
- α: number of lineages in patient j

The probability $p(s_i^k|S_j, i \leftarrow j)$ of observing the sequence s_i^k given that patient j infected patient i can now be computed as:

$$p(s_i^k|S_j, i \leftarrow j) = (\underbrace{\gamma(s_i^k, S_j)}_{\text{ances. in } S_j} \times \underbrace{\sum_{q=1}^{m_j} \xi(s_i^k, s_j^q)}_{\text{prob. MRA for each } S_j}) + \underbrace{1-\gamma(s_i^k, S_j)}_{\text{ances. not sampled}}$$

The probability of observing the set of sequences S_i given that j infected i is simply computed as the product over all sequences in S_i:

$$p(S_i|S_j, i \leftarrow j) \quad = \quad \prod_{k=1}^{m_i} p(s_i^k|S_j, i \leftarrow j)$$

For the sake of simplicity, we shall refer to this quantity as $f_{i,j}$.

Assumptions of the genetic model

The genetic model makes a few key assumptions:

- different types of mutations happen independently
- all lineages within a host are as likely to have been sampled and sequenced; when lineages have different within-host population sizes, this may still be ensured by extensive swabbing and retaining the sequences of new haplotypes only.

Prior level

For all model parameters, independent prior distributions were chosen:

- uniform on $[0, 1]$ for ψ, ϕ, π, ν_1 and ν_2
- uniform on $[0, 100]$ for κ
- uniform on $[0 - 1000]$ for μ_α
- flat exponential (mean 1000) for all other parameters.

Parameter Estimation

A Markov chain Monte Carlo (MCMC) method was used to sample the joint posterior distribution $P(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\theta})$. ψ, ϕ and π were updated using the Gibbs sampler, and all other parameters using a Metropolis algorithm.