Construction of semi-Markov genetic-space-time SEIR models and inference

Samuel Soubeyrand

Soumission

Journal de la Société Française de Statistique

arXiv: math.PR/00000000

Construction of semi-Markov genetic-space-time SEIR models and inference

Titre: Construction de modèles stochastiques génético-spatio-temporels et inférence

Samuel Soubeyrand[1]

Abstract: Identifying transmission links of an infectious disease through a host population is critical to understanding its epidemiology and informing measures for its control. Infected hosts close together in their locations and timings are often thought to be linked, but timing and locations alone are usually consistent with many different scenarios of *who infected whom*. To infer more reliably who-transmitted-to-whom over the course of a disease outbreak caused by a fast-evolving pathogen, pathogen genomic data have been combined with spatial and temporal data. However, the manner to combine these data remains today a modeling and statistical challenge.

One of the approaches recently proposed is based on an extension of stochastic Susceptible-Exposed-Infectious-Removed (SEIR) models. In this article, we present this extension that combines (i) an individual-based, spatial, semi-Markov SEIR model for the spatio-temporal dynamics of the pathogen, and (ii) a Markovian evolutionary model for the temporal evolution of genetic sequences of the pathogen. The resulting model is a state-space model including latent vectors of high dimension. Then, we describe a new algorithm that allows an approximate Bayesian inference of model parameters and latent variables. Finally, the capacity of the estimation algorithm to reconstruct transmission trees (i.e. who infected whom) is assessed with a simulation study. We especially investigate how the inference method performs when only a fraction of pathogen genomic data is available.

Résumé : Identifier les évènements de transmission d'une maladie infectieuse dans une population hôte est essentiel pour comprendre son épidémiologie et améliorer les mesures de lutte contre la maladie. Les hôtes infectés proches spatialement et temporellement sont souvent supposés être liés, mais les données temporelles et spatiales seules sont généralement compatibles avec de nombreux scénarios de *qui a infecté qui*. Pour inférer de manière plus précise qui a infecté qui au cours d'une épidémie causée par un pathogène à évolution rapide, des données génomiques sur le pathogène ont été associées aux données spatiales et temporelles. Cependant, la manière d'associer ces données reste aujourd'hui un défit en terme de modélisation et de statistique.

Une des approches récemment développées est basée sur une extension des modèles stochastiques Susceptible-Exposé-Infectieux-Retiré (SEIR). Dans cet article, nous présentons cette extension qui associe (i) un modèle SEIR individu-centré, spatial et semi-markovien pour la dynamique spatio-temporelle du pathogène, et (ii) un modèle markovien d'évolution temporelle des séquences génétiques du pathogène. Le modèle résultant est un modèle à espace d'états incorporant des vecteurs latents de grande dimension. Ensuite, nous décrivons un nouvel algorithme permettant de mener une inférence bayésienne approchée des paramètres du modèle et des variables latentes. Enfin, la capacité de l'algorithme d'estimation à reconstruire les arbres de transmission (c-à-d qui a infecté qui) est évaluée avec une étude simulatoire. Nous nous intéressons tout particulièrement aux performances de la méthode d'inférence lorsque seulement une fraction des données génomiques sur le pathogène est observée.

Keywords: Bayesian estimation, Genomic data, Spatiotemporal data, State-space model, Susceptible-Exposed-Infectious-Removed model

Mots-clés : Estimation bayésienne, Données génomiques, Données spatiotemporelles, Modèle à espace d'état, Modèle Susceptible-Exposé-Infectieux-Retiré

AMS 2000 subject classifications: 62H11, 62P10, 97K80

[1] INRA, UR546 Biostatistics and Spatial Processes, 84914 Avignon.
E-mail: Samuel.Soubeyrand@avignon.inra.fr

Soumis au Journal de la Société Française de Statistique
File: seir.tex, compiled with jsfds, version : 2009/12/09
date: September 29, 2015

1. Introduction

Fast-evolving pathogens such as RNA viruses can cause human, animal and plant epidemics of high impact in developing and developed countries alike. For instance, hepatitis E causes 3 million acute cases each year worldwide (>60% in South and East Asia), of which 57,000 are fatal (World Health Organization, 2013). The 2014 Ebola outbreak, for which 4,507 cases were reported between 2013-12-30 and 2014-09-14, presented a real risk of continued expansion in fall 2014 (World Health Organization, 2014). During the foot-and-mouth outbreak in Great Britain in 2001, some 6 million animals were culled and the total cost was estimated to be in excess of £8 billion (Anderson et al., 1996; Haydon et al., 2004). At the global scale and over three decades, the overall cost of sharka (infecting trees of the genus *Prunus*) was estimated to exceed €10 billion (Cambra et al., 2006).

In order to minimize these social, environmental and economic costs, we need to most effectively control infectious diseases and thus to better understand how pathogens spread within host populations, yet this is something we know remarkably little about. Cases close together in their locations and timings are often thought to be linked, but timings and locations alone are usually consistent with many different scenarios of *who infected whom*. For fast-evolving pathogens such as the RNA viruses given as examples in the previous paragraph, pathogen genomic data can advantageously complete spatial and temporal data. The genome of such pathogens evolves so quickly relative to the rate that they are transmitted, that even over single short epidemics we can identify which hosts contain pathogens that are most closely related to each other. This information is valuable because when combined with the spatial and timing data it should help us infer more reliably who-transmitted-to-who over the course of a disease outbreak. However, doing this so that spatial, temporal and genetic data are appropriately combined remains a major statistical challenge. Furthermore, because sequencing genetic material has become so affordable, new statistical methods combining spatial, temporal and genetic data and estimating transmission trees will become very important for future epidemiology.

In recent years, the methodological challenge of reconstructing the dynamics of epidemics using (spatio-)temporal information and pathogen genetic information has been mainly addressed with two different but complementary approaches. The first approach is based on phylogeny / phylogeography, birth-death processes (often approximated by coalescent models) and BEAST (Hall and Rambaut, 2014; Lemey et al., 2009, 2010; Pybus et al., 2012; Rambaut et al., 2008; Rasmussen et al., 2011; Stadler and Bonhoeffer, 2013). This approach is used to estimate parameters related to the pathogen demographic process: the rate of spatial spread of the pathogen, the rate of switching between various host types, and the rate of evolution over time. The second approach is based on stochastic, spatiotemporal and evolutionary SEIR (Susceptible, Exposed, Infectious, Removed) models (Jombart et al., 2014; Morelli et al., 2012; Mollentze et al., 2014; Ypma et al., 2012, 2013). This modeling and inference approach combines heterogeneous and multi-scale processes and data: it links the epidemiological scale —or host population(s) scale— and the micro-evolutionary scale —or pathogen genome scale. The space-time-genetic SEIR models explicitly recognize the host population structure, and the epidemiological processes that govern the interaction of host and pathogen. They enable inferences to be made about epidemiological processes, specifically the transmission tree reflecting *who infected whom* (at the host, premise or population resolution), and other parameters (e.g. incubation durations, heterogeneity among

Soumis au Journal de la Société Française de Statistique
File: seir.tex, compiled with jsfds, version : 2009/12/09
date: September 29, 2015

hosts in susceptibility and force of infection, rates of evolution and spatial spread *per transmission event*, and sizes of infected populations).

In this article, we are interested in the second approach and the associated estimation algorithms allowing inference about transmission dynamics based on spatial, temporal and genetic data.

So far, the space-time-genetic SEIR approach has been developed in only a few articles (in generalist journals and journals of epidemiology and computational biology) that demonstrated that the approach is promising for reconstructing transmission trees and for estimating epidemiological parameters. However, this approach can be improved in several directions. One of these directions, on which we focus on in this article, concerns the inference method. In this article, we present the space-time-genetic SEIR model introduced by Mollentze et al. (2014) as a semi-Markov model (Section 2), we propose a new MCMC algorithm to estimate the model parameters and the transmission tree (Section 3; the originality of this algorithm lies in the reconstruction of transmitted pathogen sequences), and we assess the performance of the new algorithm with a simulation study (Section 4). In the simulation study, we especially assess the performance of the algorithm when pathogen genomic data are available for all sampled cases or for only a fraction of these cases. This introductory article gives probabilists and statisticians an invitation for enriching the class of space-time-genetic SEIR models and developing efficient inference algorithms in order to carry out high-impact analyses of infectious diseases due to viruses.

2. Model

The genetic-space-time SEIR model, presented below in Subsection 2.6, is a combination of a semi-Markov epidemic model and a Markovian evolutionary model. The following subsections shows how these submodels and the resulting genetic-space-time SEIR model are constructed.

2.1. Discrete-state, continuous-time Markovian SEIR model

Here, we consider a classical SEIR model describing the temporal dynamics of numbers of susceptible, exposed, infectious and removed individuals in a population affected by a pathogen (Britton and Giardina, 2015, propose in this special issue of the journal an overview of statistical inference applied to epidemic models of this kind). Time is viewed as a continuous variable. Let $\mathbf{S}(t)$, $\mathbf{E}(t)$, $\mathbf{I}(t)$ and $\mathbf{R}(t)$ in $\mathbb{N}$ respectively denote the numbers of susceptible, exposed, infectious and removed individuals at time $t \geq 0$. The sum of these quantities is equal to the instantaneous total size of the population $\mathbf{T}(t) \in \mathbb{N}$, i.e. $\mathbf{S}(t)+\mathbf{E}(t)+\mathbf{I}(t)+\mathbf{R}(t) = \mathbf{T}(t)$ for any time $t \geq 0$.

In general, many different events can cause a change in the population pattern $(\mathbf{S}(t), \mathbf{E}(t), \mathbf{I}(t), \mathbf{R}(t))$, for instance the birth and death of susceptibles, the infection of susceptibles, the death of exposed individuals, the end of exposed stage (coinciding with the beginning of the infectious stage), the death in infectious individuals, and the end of infectious stage (coinciding with the beginning of removed stage).

For the sake of simplicity, in this article, we consider only three possible events, namely infection, end of exposed stage and end of infectious stage. Corresponding transition rates are provided in Table 1. In this model, the risk of infection is a combination of a basic risk whose rate is $\alpha_0 \mathbf{S}(t)$, and an endogenous risk whose rate $\alpha_1 \mathbf{S}(t)\mathbf{I}(t)/\mathbf{T}(t)$ is proportional to the number of infectious individuals in the population of interest. The basic risk may correspond to exogenous

pathogen sources. For instance, in the case of zoonoses (i.e. diseases that can be transmitted from animals to humans) if the population of interest is the set of humans, the basic risk may correspond to animals infecting humans.

TABLE 1. *Possible events and corresponding transition rates for the discrete-state, continuous-time Markovian SEIR model.*

Description	Event	Rate
Infection	$\mathbf{S} \to \mathbf{S}-1$ & $\mathbf{E} \to \mathbf{E}+1$	$\alpha_0\mathbf{S} + \alpha_1\mathbf{SI}/\mathbf{T}$
End of exposed stage	$\mathbf{E} \to \mathbf{E}-1$ & $\mathbf{I} \to \mathbf{I}+1$	$\beta\mathbf{E}$
End of infectious stage	$\mathbf{I} \to \mathbf{I}-1$ & $\mathbf{R} \to \mathbf{R}+1$	$\delta\mathbf{I}$

2.2. *Spatial extension*

Now, consider a space decomposed into $n \in \mathbb{N}^*$ districts. In each district $k \in \{1,\dots,n\}$, the size $\mathbf{T}_k(t)$ of the resident population at time $t \geq 0$ is the sum of local numbers of susceptible, exposed, infectious and removed individuals denoted by $\mathbf{S}_k(t)$, $\mathbf{E}_k(t)$, $\mathbf{I}_k(t)$ and $\mathbf{R}_k(t)$, respectively. By assuming that contacts between individuals of different districts are possible, the local risk of infection is a combination of a basic risk whose rate is $\alpha_0\mathbf{S}_k(t)$, a local endogenous risk whose rate is $\alpha_1\mathbf{S}_k(t)\mathbf{I}_k(t)/\mathbf{T}_k(t)$, and a distant endogenous risk whose rate is $\alpha_1\mathbf{S}_k(t)\sum_{j\neq k} w_{kj}\mathbf{I}_j(t)/\mathbf{T}_j(t)$. In the latter rate, the weight w_{kj} is a measure of the intensity of contacts between individuals of districts k and j. By convention, the intensity of contacts between individuals residing in the same district is equal to one. The spatial, discrete-state, continuous-time Markovian SEIR model considered in this section is defined by events and rates provided in Table 2.

TABLE 2. *Possible events and corresponding transition rates for the spatial, discrete-state, continuous-time Markovian SEIR model.*

Description	Event	Rate
Infection	$\mathbf{S}_k \to \mathbf{S}_k-1$ & $\mathbf{E}_k \to \mathbf{E}_k+1$	$\alpha_0\mathbf{S}_k + \alpha_1\mathbf{S}_k\mathbf{I}_k/\mathbf{T}_k + \alpha_1\mathbf{S}_k\sum_{j\neq k} w_{kj}\mathbf{I}_j/\mathbf{T}_j$
End of exposed stage	$\mathbf{E}_k \to \mathbf{E}_k-1$ & $\mathbf{I}_k \to \mathbf{I}_k+1$	$\beta\mathbf{E}_k$
End of infectious stage	$\mathbf{I}_k \to \mathbf{I}_k-1$ & $\mathbf{R}_k \to \mathbf{R}_k+1$	$\delta\mathbf{I}_k$

2.3. *Particular case: individual-based version of the model*

Now, let us consider a particular case of the previous model: assume that for all $k \in \{1,\dots,n\}$, the size $\mathbf{T}(t) \equiv 1$. Thus, districts are replaced by single individuals, and the model becomes an individual-based model where the dynamics of the epidemics is modeled at the individual resolution. In addition, values of $\mathbf{S}_k(t)$, $\mathbf{E}_k(t)$, $\mathbf{I}_k(t)$ and $\mathbf{R}_k(t)$ are in $\{0,1\}$ and their sum is equal to one whatever t. By assuming that each individual $k \in \{1,\dots,n\}$ is located at x_k in the planar space $\mathbb{R}^2$, events and rates shown in Table 2 can be re-written as in Table 3. The location x_k can be viewed as the central or main location of k. We can see in Table 3 that the local endogenous risk disappeared from the expression of the rate of infection since a susceptible individual cannot infect himself (another justification is that $\mathbf{S}_k(t)\mathbf{I}_k(t) = 0, \forall t \geq 0$). Moreover, the rate corresponding to the distant endogenous risk is now written $\alpha_1\sum_{j\neq k} w(x_j - x_k)\mathbf{I}_j$ where w is a kernel whose value

depends on the relative locations of individuals k and j. In ecology and epidemiology, w is called a dispersal kernel, and $w(x_j - x_k)$ measures the intensity of contact between individuals k and j. A classical and flexible class of dispersal kernels is formed by the power-exponential kernels parametrized by $\alpha_2 = (\alpha_{2,1}, \alpha_{2,2})$ and satisfying, for all $x \in \mathbb{R}^2$:

$$w(x) = \frac{\alpha_{2,2}}{2\pi(\alpha_{2,1})^2\Gamma\left(\frac{2}{\alpha_{2,2}}\right)} \exp\left\{-\left(\frac{||x||}{\alpha_{2,1}}\right)^{\alpha_{2,2}}\right\}, \tag{1}$$

where $||x||$ is the Euclidean distance between the origine of the planar space and x. Thus, the measure of the intensity of contact between individuals k and j decreases with the distance separating the central locations of k and j. Note that in (1), the constant before the exponential ensures that the integral of w over $\mathbb{R}^2$ is equal to one.

TABLE 3. *Possible events and corresponding transition rates for the individual-based, spatial, discrete-state, continuous-time Markovian SEIR model.*

Description	Event	Rate
Infection	$\mathbf{S}_k : 1 \to 0$ & $\mathbf{E}_k : 0 \to 1$	$\alpha_0 + \alpha_1 \sum_{j \neq k} w(x_j - x_k)\mathbf{I}_j$
End of exposed stage	$\mathbf{E}_k : 1 \to 0$ & $\mathbf{I}_k : 0 \to 1$	β
End of infectious stage	$\mathbf{I}_k : 1 \to 0$ & $\mathbf{R}_k : 0 \to 1$	δ

2.4. Semi-Markov extension of the individual-based model

In the Markovian model presented in the previous subsection, the times spent by individuals in the exposed and the infectious states are exponentially distributed. Depending on the context, this may be viewed as an unrealistic assumption. For instance, the exposed duration, that corresponds to a latency or incubation duration, is usually not exponentially distributed but has a distribution with a mode away from zero (e.g. see Hampson et al., 2009). Semi-Markov models (Barbu and Limnios, 2008) offer a framework to handle non-exponential durations in some of the possible states. Thus, in this subsection, we introduce a semi-Markov model where durations in the exposed and infectious stages are independently drawn under gamma distributions (see Table 4). The draws are also independent from the duration in the susceptible stage.

Remark concerning the durations in the susceptible and removed states: the durations between successive changes of the susceptible state in the models of Subsection 2.1-2.4 are not exponentially distributed in general because the corresponding transition rates randomly vary in time (due to variations in the infectious states; see Appendix A). In the individual based models of Subsection 2.3-2.4, the removed states are absorbant and the corresponding durations are therefore infinite.

TABLE 4. *Possible events and corresponding transition rates or distributions for the individual-based, spatial, discrete-state, continuous-time, semi-Markov SEIR model.*

Description	Event	Rate	Distribution
Infection	$\mathbf{S}_k : 1 \to 0$ & $\mathbf{E}_k : 0 \to 1$	$\alpha_0 + \alpha_1 \sum_{j \neq k} w(x_j - x_k)\mathbf{I}_j$	
End of exposed stage	$\mathbf{E}_k : 1 \to 0$ & $\mathbf{I}_k : 0 \to 1$		$\Gamma(\beta_1, \beta_2)$
End of infectious stage	$\mathbf{I}_k : 1 \to 0$ & $\mathbf{R}_k : 0 \to 1$		$\Gamma(\delta_1, \delta_2)$

2.5. *Markovian evolutionary model for a pathogen sequence*

Now, suppose that the pathogen under consideration is an RNA virus that can evolves with time. More specifically, we suppose that mutations can occur at s sites of the viral sequence between the four possible nucleobases that are adenine (A), cytosine (C), guanine (G) and uracil (U). We assume that mutations in different sites are independent but mutation rates vary as functions of the current nucleobases and the substituting nucleobases as in the 3-parameters Kimura substitution model (Kimura, 1981). Here, *mutation* and *substitution* are synonyms. Thus, at the s sites of the sequence under mutation, the mutation processes follow independent, discrete-state, continuous-time Markovian models that are defined by events and rates provided in Table 5.

Based on this model, Kimura (1981) considered two sequences evolving from a hidden common ancestral sequence, they built a system of ordinary differential equations governing the probabilities of obtaining in the two sequences at any time all the possible pairs of nucleotides (AA, AC, AG, AU, CG, ...), and they solved the system analytically to obtain the expressions of the probabilities as a function of the time of evolution (further technical details are provided in Takahata and Kimura, 1981).

In our framework, the times of evolution since a common ancestor of two observed sequences are generally different because the sequences are observed at varying times. However, because the mutation rates do not depend on the direction of the mutation (e.g. the same rate applies for A→G and G→A), the direction of the evolution does not matter and, formally, only the evolutionary time lag separating the two sequences matters. Thus, the expressions provided by Kimura (1981) can be used in our framework to compute the expected proportions of the numbers of transitions, type-1 transversions, type-2 transversions and unchanged nucleobases over an evolutionary time lag Δ separating two sequences. These expected proportions are:

$$\rho = (\rho_1, \rho_2, \rho_3, \rho_4) = \frac{1}{4}\big(1 - e_1 - e_2 + e_3,\ 1 - e_1 + e_2 - e_3,\ 1 + e_1 - e_2 - e_3,\ 1 + e_1 + e_2 + e_3\big), \tag{2}$$

where $e_1 = \exp\{-2(\mu_1 + \mu_2)\Delta\}$, $e_2 = \exp\{-2(\mu_1 + \mu_3)\Delta\}$, $e_3 = \exp\{-2(\mu_2 + \mu_3)\Delta\}$, and μ_1, μ_2 and μ_3 are the genetic substitution rates per nucleotide per day, for transitions, type-1 transversions and type-2 transversions, respectively.

In addition, we make the following distributional assumption: the numbers of observed transitions, type-1 transversions, type-2 transversions and unchanged nucleobases over an evolutionary time lag Δ separating two sequences are distributed from a multinomial distribution, say $P_{\mu,s}(\cdot \mid \Delta)$, with size equal to the length s of the observed sequence fragment and with the vector of probabilities ρ given by Equation (2). Thus, for any nonnegative integers m_1, m_2, m_3, m_4 whose sum is s,

$$P_{\mu,s}\{(m_1, m_2, m_3, m_4) \mid \Delta\} = \frac{(s!) \times \rho_1^{m_1}\rho_2^{m_2}\rho_3^{m_3}\rho_4^{m_4}}{(m_1!) \times (m_2!) \times (m_3!) \times (m_4!)}.$$

2.6. *Genetic-space-time SEIR model*

The genetic-space-time SEIR model, whose structure is illustrated by Figure 1, is obtained by combining the semi-Markov SEIR model of Subsection 2.4 and the Markovian evolutionary model of Subsection 2.5. The two models are combined under the following list of assumptions.

TABLE 5. *Possible events and corresponding substitution rates for the Markovian evolutionary model. Letters A, C, G and U denotes nucleobases adenine, cytosine, guanine and thymine, respectively.*

Description	Event	Rate
Transition	A→G or G→A or C→U or U→C	μ_1
Transversion of type 1	A→U or U→A or C→G or G→C	μ_2
Transversion of type 2	A→C or C→A or G→U or U→G	μ_3

– The disease reservoir (i.e. exogenous source) is assumed to simply consist of one virus sequence S_{exo} dated at time $t_{\text{exo}} \in \mathbb{R}$.
– We assume that, at any time, there is only one sequence of the virus (i.e. one viral variant) per infected individual. The sequence in individual k at time t, for t such that $\mathbf{E}_k(t) = 1$ or $\mathbf{I}_k(t) = 1$, is denoted by $S_k(t)$ and is a vector of letters A, C, G and U.
– When a susceptible individual k is infected by an infectious individual j at time t, the sequence $S_j(t)$ is transmitted to k, i.e. $S_k(t) = S_j(t)$.
– For any individual, mutations of the virus sequence during the exposed and infectious stages are assumed to be independent from the epidemiological dynamics. When an individual is removed, the sequence is fixed. This is illustrated in Figure 2 where the length of the sequence under mutation is $s = 2$.
– Virus mutations in different infected individuals are assumed to be conditionally independent given the virus sequences at the infection times.

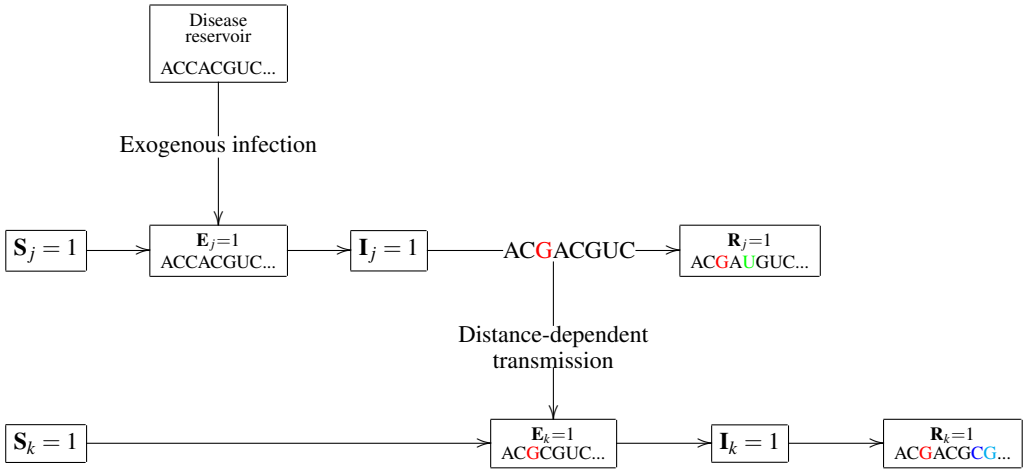


FIGURE 1. *Diagram illustrating the combination of the semi-Markov SEIR model and the Markovian evolutionary model. Individual j is infected by the disease reservoir with virus sequence "ACCACGUC...". Then, j becomes infectious and when j infects k, the sequence in j has evolved (C at the 3rd base mutated to G). Finally, the sequences in j and k independently evolve.*

Soumis au Journal de la Société Française de Statistique
File: seir.tex, compiled with jsfds, version : 2009/12/09
date: September 29, 2015

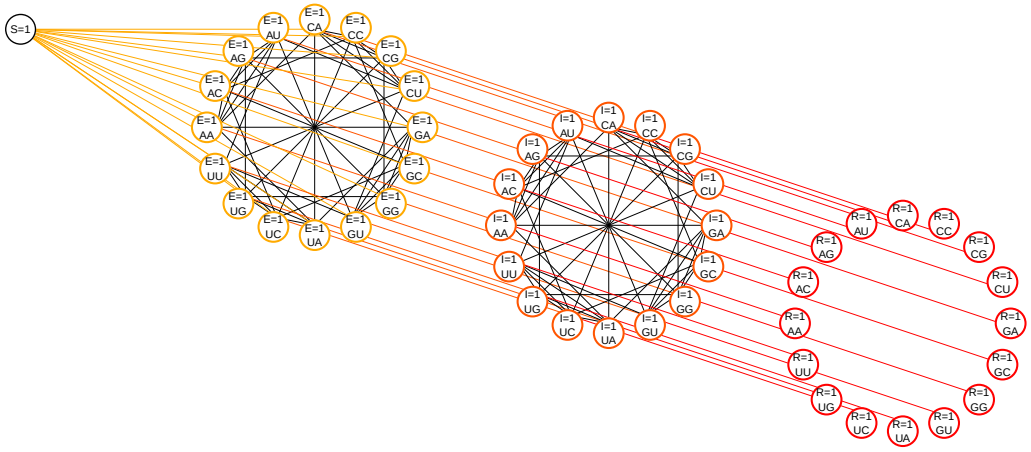


FIGURE 2. *Possible transitions for an individual in the genetic-space-time SEIR model, where the virus sequence is of length* $s = 2$. *Transitions from* $\mathbf{S} = 1$ *(susceptible) to* $\mathbf{E} = 1$ *(exposed), from* $\mathbf{E} = 1$ *to* $\mathbf{I} = 1$ *(infectious) and from* $\mathbf{I} = 1$ *to* $\mathbf{R} = 1$ *(removed) are irreversible, whereas transitions corresponding to substitutions in nucleobases are reversible.*

3. Estimation methods

3.1. Data structure

We consider the same type of data than in Mollentze et al. (2014) who analyzed transmissions of rabies within a South African population of dogs, jackals and livestock. Thus,

- Data are collected in a spatio-temporal observation window included in the whole spatial domain and in the whole time frame covered by the disease dynamics (which is either an epidemic or an endemic dynamics);
- Among all the infected cases in the spatio-temporal observation window, only a subset is observed (thus, sources of infection can be unobserved, and the unobserved sources can be inside or outside the observation window);
- The removed state corresponds to the death of the individual;
- Sampled individuals are observed when they die, that is to say at the transition from $\mathbf{I} = 1$ to $\mathbf{R} = 1$;
- Virus sequences that are available correspond to the states of the sequences at the death times (usually only a fragment of the sequence is available, the same fragment for all inividuals);
- The central locations $x_1, \ldots, x_n$ of sampled individuals are assumed to be the locations at the death;
- The sequence S_{exo} is assumed to be known (in real cases, S_{exo} can be reconstructed by a phylogenetic analysis of available genetic sequences: S_{exo} is typically the reconstructed sequence of the most recent common ancestor; see Mollentze et al., 2014).

Compared to the amount of variables in the model, data are particularly sparse. Indeed, looking at Figure 1, data consist of the sequence of the disease reservoir and locations, times and sequences collected when observed individuals are in state $\mathbf{R} = 1$. It has to be noted that, usually, only a

fraction of infected individuals are observed. Now, consider Figure 2, infected individuals and pathogen sequences are observed in one of the state $\mathbf{R} = 1$ whereas they previously evolved in a high-dimension space of states (37 states in Figure 2; $1 + 2 \times 4^s$ states in general, where s is the length of the observed sequence fragment).

3.2. Posterior distribution

Following Morelli et al. (2012) and Mollentze et al. (2014), we consider the joint posterior distribution $p(J, T^{inf}, L, D, \theta \mid data)$ of the transmission tree J, infection times $T^{inf} = (T_1^{inf}, \ldots, T_n^{inf})$, exposed (or latency) durations $L = (L_1, \ldots, L_n)$, infectious durations $D = (D_1, \ldots, D_n)$, and parameters θ that contains infection and dispersal parameters $\alpha = (\alpha_0, \alpha_1, \alpha_{2,1}, \alpha_{2,2})$, latency parameters $\beta = (\beta_1, \beta_2)$, infectiousness parameters $\delta = (\delta_1, \delta_2)$, mutation parameters $\mu = (\mu_1, \mu_2, \mu_3)$ and the date t_{exo} of the exogenous sequence S_{exo}.

The transmission tree J is a function from $\{1, \ldots, n\}$ to $\{0, 1, \ldots, n\}$ that states who infected whom: an observed individual i is infected by a pathogen source $j = J(i)$ that is either another observed individual $j \in \{1, \ldots, n\}$, $j \neq i$, or the disease reservoir (exogenous source) denoted by 0.

Data are removal times $T^{end} = (T_1^{end}, \ldots, T_n^{end})$, central locations $X = (x_1, \ldots, x_n)$ and observed sequences $S^{end} = \{S_1(T_1^{end}), \ldots, S_n(T_n^{end})\}$. The posterior distribution is:

$$\begin{aligned} p(J, T^{inf}, L, D, \theta \mid data) &= p(J, T^{inf}, L, D, \theta \mid S^{end}, T^{end}, X, S_{\text{exo}}) \\ &\propto p(S^{end} \mid J, T^{inf}, L, D, \theta, T^{end}, X, S_{\text{exo}}) p(J, T^{inf}, L, D, \theta \mid T^{end}, X, S_{\text{exo}}) \\ &= p(S^{end} \mid J, T^{inf}, L, D, \theta, T^{end}, X, S_{\text{exo}}) p(J, T^{inf} \mid L, D, \theta, T^{end}, X, S_{\text{exo}}) \\ &\quad \times p(L, D \mid \theta, T^{end}, X, S_{\text{exo}}) p(\theta), \end{aligned} \tag{3}$$

where $\propto$ means "proportional to" (the multiplicative constant does not depend on the unknowns $(J, T^{inf}, L, D, \theta)$), $p(S^{end} \mid J, T^{inf}, L, D, \theta, T^{end}, X, S_{\text{exo}})$ is called the genetic likelihood, $p(J, T^{inf} \mid L, D, \theta, T^{end}, X, S_{\text{exo}})$ is called the transmission likelihood, $p(L, D \mid \theta, T^{end}, X, S_{\text{exo}})$ is the distribution of latency and infectious durations and $p(\theta)$ is the prior distribution of parameters.

The prior distribution of parameters will be specified in Subsection 4.2. Based on Subsection 2.4, the distribution of latency and infectious durations is simply the product of gamma probabilities:

$$\begin{aligned} p(L, D \mid \theta, T^{end}, X, S_{\text{exo}}) &= p(L, D \mid \theta) \\ &= \prod_{i=1}^{I} \gamma(L_i; \beta_1, \beta_2) \gamma(D_i; \delta_1, \delta_2), \end{aligned}$$

where $\gamma(\cdot; a, b)$ is the probability distribution function of the gamma distribution parameterized by (a, b).

In Subsection 3.3, we detail the transmission likelihood and show how the incompleteness of epidemiological data, i.e. the missing infecting hosts, is handled. In Subsections 3.4, 3.5 and 3.6, we detail the genetic likelihood and show how the incompleteness of genetic data, i.e. the missing pathogen sequences, is handled. The genetic likelihood can be formally written as a function of the unobserved virus sequences at the infection times $S_k(T_k^{inf})$, $k = 1, \ldots, n$. To avoid

to handle latent vectors $S_k(T_k^{inf})$ (the dimension of these unknowns is $n \times s$), Morelli et al. (2012) and Mollentze et al. (2014) replaced the genetic likelihood by a pseudo-likelihood in their MCMC algorithm; see Subsection 3.5. In Subsection 3.6, we propose a new approximation of the genetic likelihood whose performance will be assessed in Section 4.

3.3. Transmission likelihood

Following Mollentze et al. (2014), the transmission likelihood $p(J, T^{inf} \mid L, D, \theta, T^{end}, X, S_{\text{exo}})$ can be written:

$$\begin{aligned} p(J, T^{inf} \mid L, D, \theta, T^{end}, X, S_{\text{exo}}) &= p\left(J(1), T_1^{inf} \mid L, D, \theta, T^{end}, X\right) \\ &\times \prod_{i=2}^{I} p\left(J(i), T_i^{inf} \mid J\{1:(i-1)\}, T_{1:(i-1)}^{inf}, L, D, \theta, T^{end}, X\right), \end{aligned} \tag{4}$$

where the index i is sorted with respect to increasing infection times T_i^{inf}, $J\{1:(i-1)\} = (J(1), \ldots, J(i-1))$ for $i > 1$, $T_{1:(i-1)}^{inf} = (T_1^{inf}, \ldots, T_{i-1}^{inf})$ for $i > 1$, and by assuming that the transmission dynamics does not depend on the exogenous sequence S_{exo}.

Each host has the same chance $(1/I)$ to be infected first (by an external source $J(1) = 0$), and its infection time is assumed to be less than or equal to the first observation time ($\min\{T^{end}\}$):

$$p\left(J(1), T_1^{inf} \mid L, D, \theta, T^{end}, X\right) = \frac{1}{I} \times \mathbf{1}(T_1^{inf} \leq \min\{T^{end}\}),$$

where $\mathbf{1}$ is the indicator function. Subsequent infections (i.e. for $i > 1$) occur with the following probabilities derived from assumptions made in Subsections 2.4 and 2.6:

$$\begin{aligned} &p\left(J(i), T_i^{inf} \mid J\{1:(i-1)\}, T_{1:(i-1)}^{inf}, L, D, \theta, T^{end}, X\right) \\ &= \exp\left(-\alpha_0(T_i^{inf} - T_1^{inf}) - \int_{T_1^{inf}}^{T_i^{inf}} \sum_{j=1}^{i-1} \alpha_1 \mathbf{1}(T_j^{inf} + L_j \leq t \leq T_j^{end}) w(x_j - x_i) dt\right) \\ &\times \left(\alpha_0 \mathbf{1}\{J(i) = 0\} + \alpha_1 \mathbf{1}(T_{J(i)}^{inf} + L_{J(i)} \leq T_i^{inf} \leq T_{J(i)}^{obs}) w(x_{J(i)} - x_i) \mathbf{1}\{J(i) > 0\}\right) \end{aligned}$$

where the exponential term is the probability that host i has not been infected between times T_1^{inf} and T_i^{inf}, and the second term is the probability density that host i has been infected by $J(i)$ at time T_i^{inf}. Here, if $J(i) > 0$ the source is observed, while the source is external to the dataset (an introduction) if $J(i) = 0$. α_0 is the infection strength of the exogenous sources, assumed to be constant in time and space, α_1 is the infection strength of an observed source, and w is the parametric dispersal kernel specified in Equation (1).

3.4. Genetic likelihood

By assuming that the evolution of pathogen sequences does not depend on the transition from exposed to infectious states and does not depend on individual locations, the genetic likelihood

$p(S^{end} \mid J, T^{inf}, L, D, \theta, T^{end}, X, S_{\text{exo}})$ reduces to $p(S^{end} \mid J, T^{inf}, \theta, T^{end}, S_{\text{exo}})$. This distribution was written by Mollentze et al. (2014) as a sum over all possible transmitted sequences $S_i^{inf} = S_i(T_i^{inf})$ at times T_i^{inf} for all i such that $J(i) > 0$:

$$\begin{aligned} & p(S^{end} \mid J, T^{inf}, \theta, T^{end}, S_{\text{exo}}) \\ & = \sum_{\substack{\{S_i^{inf} \in \mathbb{S}:\ i=1,\ldots,n \\ \text{and } J(i)>0\}}} \left\{ \left(\prod_{i=1}^{n} P_{\mu,s}\{M(S_i^{end}, S^{\dagger}_{\text{prec}(i,obs)}) \mid \Delta = T_i^{end} - T^{\dagger}_{\text{prec}(i,obs)}\} \right) \right. \\ & \left. \times \left(\prod_{\substack{i=1 \\ J(i)>0}}^{n} P_{\mu,s}\{M(S_i^{inf}, S^{*}_{\text{prec}(i,inf)}) \mid \Delta = T_i^{inf} - T^{*}_{\text{prec}(i,inf)}\} \right) \right\}. \end{aligned} \tag{5}$$

In Equation (5), the first series of factors accounts for the probabilities of the number of substitutions between an observed sequence and the immediately preceding unobserved, transmitted sequence or the sequence of the exogenous source (last double arrows on every horizontal lines of Figure 3A). The second series of factors accounts for the probabilities of the number of substitutions between each transmitted sequence and the transmitted sequence immediately preceding in time or the sequence of the exogenous source (every double arrows except the last one on every horizontal lines of Figure 3A).

Let us now detail terms in Equation (5): $\mathbb{S}$ is the set of all possible sequences (the size of $\mathbb{S}$ is 4^s, where s is the length of the sequence fragment); $M(S', S)$ is the vector of the numbers of transitions, type-1 transversions, type-2 transversions and unchanged nucleobases between S and S'; $P_{\mu,s}\{M(S', S) \mid \Delta = T' - T\}$ is the probability given by the multinomial distribution in Subsection 2.5.

The subscript $\text{prec}(i, obs)$ can take two types of values. First case: if $J(i) = 0$ and i did not infected any other observed host, then $\text{prec}(i, obs) = \text{exo}$, $S^{\dagger}_{\text{prec}(i,obs)} = S_{\text{exo}}$ and $T^{\dagger}_{\text{prec}(i,obs)} = t_{\text{exo}}$. Second case: if $J(i) > 0$ or if i infected another observed host, then $\text{prec}(i, obs)$ denotes the host whose node of infection belongs to the tree path from the root of the tree to the observation of i (at time T_i^{end}) and whose infection is just preceding the observation of i ($S^{\dagger}_{\text{prec}(i,obs)}$ is the transmitted sequence $S^{inf}_{\text{prec}(i,obs)}$ at the infection time $T^{\dagger}_{\text{prec}(i,obs)} = T^{inf}_{\text{prec}(i,obs)}$). The node of infection of a given host k is defined as the point on the tree at which "the branch leading to the observation of k" and "the branch leading to the observation of the infecting host $J(k)$" diverged. The tree path from one point of the tree to another is defined as the most direct broken line on the graph connecting the two points. In the second case, if $J(i) > 0$ and i did not infect any other host, then $\text{prec}(i, obs)$ is i itself.

The subscript $\text{prec}(i, inf)$ can also take two types of values. First case: if $J(J(i)) = 0$ and $J(i)$ did not infect any other observed host before the infection of i at T_i^{inf}, then $\text{prec}(i, inf) = \text{exo}$, $S^{*}_{\text{prec}(i,inf)} = S_{\text{exo}}$ and $T^{*}_{\text{prec}(i,inf)} = t_{\text{exo}}$. Second case: if $J(J(i)) > 0$ or if $J(i)$ infected another observed host before the infection of i at T_i^{inf}, then $\text{prec}(i, inf)$ denotes the host whose node of infection belongs to the tree path from the root of the tree to the infection of i (at time T_i^{inf}) and whose infection is just preceding the infection of i ($S^{*}_{\text{prec}(i,inf)}$ is the transmitted sequence $S^{inf}_{\text{prec}(i,inf)}$ at the infection time $T^{*}_{\text{prec}(i,inf)} = T^{inf}_{\text{prec}(i,inf)}$).

3.5. *Genetic pseudo-likelihood*

To reduce the complexity of the inference algorithm, Mollentze et al. (2014) used a conditional pseudo-distribution of S^{end}, noted $\tilde{p}(S^{end} \mid J, T^{inf}, \theta, T^{end}, S_{\text{exo}})$, instead of the exact conditional distribution given by Equation (5). The conditional pseudo-distribution does not depend on the extra latent vectors $\{S_i^{inf} : i = 1, \ldots, n,\ J(i) > 0\}$ appearing in Equation (5).

With index i being sorted with respect to increasing infection times T_i^{inf}, the distribution $p(S^{end} \mid J, T^{inf}, \theta, T^{end}, S_{\text{exo}})$ can be written:

$$\begin{aligned} p(S^{end} \mid J, T^{inf}, \theta, T^{end}, S_{\text{exo}}) = & p(S_1^{end} \mid J, T^{inf}, \theta, T^{end}, S_{\text{exo}}) \\ & \times \prod_{i=2}^{n} p(S_i^{end} \mid S_{1:(i-1)}^{end}, J, T^{inf}, \theta, T^{end}, S_{\text{exo}}), \end{aligned} \tag{6}$$

where $S_{1:(i-1)}^{end}$ is the set of observed sequences from hosts $1, \ldots, i-1$.

For the first infected host,

$$p(S_1^{end} \mid J, T^{inf}, \theta, T^{end}, S_{\text{exo}}) = P_{\mu,s}\{M(S_1^{end}, S_{\text{exo}}) \mid \Delta = T_1^{end} - t_{\text{exo}}\}.$$

For the other hosts infected by the exogenous source (i.e. for $i > 1$ such that $J(i) = 0$),

$$p(S_i^{end} \mid S_{1:(i-1)}^{end}, J, T^{inf}, \theta, T^{end}, S_{\text{exo}}) = P_{\mu,s}\{M(S_i^{end}, S_{\text{exo}}) \mid \Delta = T_i^{end} - t_{\text{exo}}\}.$$

For hosts infected by observed hosts (i.e. for $i > 1$ such that $J(i) > 0$), we replaced the conditional probability $p(S_i^{end} \mid S_{1:(i-1)}^{end}, J, T^{inf}, \theta, T^{end}, S_{\text{exo}})$ of S_i^{end} given sequences S_j^{end} $(j = 1, \ldots, i-1)$ by the product, up to a power, of the conditional probabilities of S_i^{end} given *each* sequence S_j^{end} such that $j \in 1, \ldots, i-1$ and j is in the transmission chain leading to i (the latter condition is mathematically written: $\exists m \in \mathbb{N}^*, J^m(i) = j$, where J^m consists of composing J with itself m times):

$$\left(\prod_{\substack{j=1 \\ \exists m \in \mathbb{N}^*, J^m(i) = j}}^{i-1} P_{m,s}\{M(S_i^{end}, S_j^{end}) \mid \Delta = \mid T_i^{end} - T_{\text{div}(i,j)}^{inf} \mid + \mid T_j^{end} - T_{\text{div}(i,j)}^{inf} \mid\} \right)^{1/\eta_i}, \tag{7}$$

where η_i is the number of terms in the product, $T_{\text{div}(i,j)}^{end}$ denotes the infection time at which the chain of infection leading to i and the chain of infection leading to j diverged ($T_{\text{div}(i,j)}^{inf}$ is one of the latent variables in T^{inf}) and $\Delta = \mid T_i^{end} - T_{\text{div}(i,j)}^{inf} \mid + \mid T_j^{end} - T_{\text{div}(i,j)}^{inf} \mid$ is the evolutionary duration separating the observation of S_i^{end} and S_j^{end}. The use of the power $1/\eta_i$ is a way to get a quantity homogeneous to a single probability and not to a product of probabilities whatever the length of the transmission chain leading to i. Therefore, the hosts have similar weights in the pseudo-distribution given below. The computation of Equation (7) is illustrated by Figure 3B.

Thus, the conditional pseudo-distribution of S^{end}, or in other words the genetic pseudo-likelihood, satisfies:

$$\begin{aligned}
&\tilde{p}(S^{end} \mid J, T^{inf}, \theta, T^{end}, S_{\text{exo}}) \\
&= \prod_{\substack{i=1 \\ J(i)=0}}^{n} P_{\mu,s}\{M(S_i^{end}, S_{\text{exo}}) \mid \Delta = T_i^{end} - t_{\text{exo}}\} \\
&\quad \times \prod_{\substack{i=1 \\ J(i)>0}}^{n} \left(\prod_{\substack{j=1 \\ \exists m \in \mathbb{N}^*, J^m(i)=j}}^{i-1} P_{m,s}\{M(S_i^{end}, S_j^{end}) \mid \Delta = \mid T_i^{end} - T_{\text{div}(i,j)}^{inf} \mid + \mid T_j^{end} - T_{\text{div}(i,j)}^{inf} \mid\} \right)^{1/\eta_i} .
\end{aligned} \tag{8}$$

3.6. *Genetic approximate likelihood*

The genetic likelihood given by Equation (5) is a sum over all possible transmitted sequences S_i^{inf} for i such that $J(i) > 0$. In the genetic approximate likelihood, we use a fixed value $\breve{S}_i^{inf}$ for S_i^{inf} that is deterministically reconstructed conditionally on J, S_{exo} and S^{end}. The reconstruction is based on the parsimony principle commonly used in phylogenetics: the most parsimonious reconstruction of $\{S_i^{inf} : i = 1, \ldots, n, J(i) > 0\}$ is the one that requires the fewest evolutionary changes (i.e. the fewest substitutions of nucleobases; see Tuffley and Steel, 1997, for a formal definition). The genetic approximate likelihood is given by:

$$\begin{aligned}
&\breve{p}(S^{end} \mid J, T^{inf}, \theta, T^{end}, S_{\text{exo}}) \\
&\quad = \left(\prod_{i=1}^{n} P_{\mu,s}\{M(S_i^{end}, \breve{S}^{\dagger}_{\text{prec}(i,obs)}) \mid \Delta = T_i^{end} - T^{\dagger}_{\text{prec}(i,obs)}\} \right) \\
&\qquad \times \left(\prod_{\substack{i=1 \\ J(i)>0}}^{n} P_{\mu,s}\{M(\breve{S}_i^{inf}, \breve{S}^{*}_{\text{prec}(i,inf)}) \mid \Delta = T_i^{inf} - T^{*}_{\text{prec}(i,inf)}\} \right) ,
\end{aligned} \tag{9}$$

where $\breve{S}^{\dagger}_{\text{prec}(i,obs)}$ and $\breve{S}^{*}_{\text{prec}(i,inf)}$ are computed like $S^{\dagger}_{\text{prec}(i,obs)}$ and $S^{*}_{\text{prec}(i,inf)}$ in Subsection 3.4 except that S_i^{inf} is replaced by $\breve{S}_i^{inf}$ for all i such that $J(i) > 0$.

The set of sequences $\{\breve{S}_i^{inf} : i = 1, \ldots, n, J(i) > 0\}$ is computed by applying Algorithm 2 that calls Algorithm 1 several times.

Algorithm 1 computes one of the most parsimonious sequence for the node of bifurcation of an ancestral sequence G^{anc} into two sequences $G^{(1)}$ and $G^{(2)}$ (there can be several most parsimonious sequences for the node; see Discussion section). In our setting (and in Algorithm 2), nodes of bifurcation coincide with infection events. The function that provides G^{node} by carrying out calculations in Algorithm 1 is denoted by $\mathbf{G} : (G^{anc}, G^{(1)}, G^{(2)}) \mapsto \mathbf{G}(G^{anc}, G^{(1)}, G^{(2)})$.

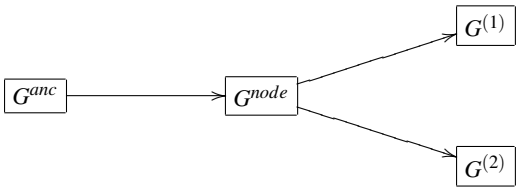
Algorithm 2 shows how the fixed value $\{\breve{S}_i^{inf} : i = 1, \ldots, n, J(i) > 0\}$ are obtained given J, S_{exo} and S^{end}. Note that the operation $G^{anc} \Leftarrow \breve{S}_j^{inf}$ that uses one of the output of the algorithm is

feasible because index i is sorted with respect to increasing infection times T_i^{inf} (i.e. $\check{S}_j^{inf}$ has been already computed when the operation $G^{anc} \Leftarrow \check{S}_j^{inf}$ is performed).

Remark: Algorithm 2 is not demonstrated to yield the most parsimonious reconstruction of the transmitted sequences. However, it is expected to yield a relatively parsimonious reconstruction because it is based on successive calls of Algorithm 1, which leads to the most parsimonious sequence for the node of bifurcation of an ancestral sequence G^{anc} into two sequences $G^{(1)}$ and $G^{(2)}$.

Algorithm 1 Most parsimonious reconstruction of the sequence $G^{node} = (g_1^{node}, \ldots, g_s^{node})$ of the node of bifurcation from an ancestral sequence $G^{anc} = (g_1^{anc}, \ldots, g_s^{anc})$ to two sequences $G^{(1)} = (g_1^{(1)}, \ldots, g_s^{(1)})$ and $G^{(2)} = (g_1^{(2)}, \ldots, g_s^{(2)})$. These sequences, that are shown on the diagram below on the right of the algorithm, are vectors of size s whose components take values in a finite countable set.

for $b = 1, \ldots, s$ **do**
  if $g_b^{(1)} = g_b^{(2)}$ **then**
    $g_b^{node} \Leftarrow g_b^{(1)}$
  else
    $g_b^{node} \Leftarrow g_b^{anc}$
  end if
end for



Algorithm 2 Reconstruction of sequences $\{S_i^{inf} : i = 1, \ldots, n, J(i) > 0\}$ conditionally on J, S_{exo} and S^{end}. Index i is sorted with respect to increasing infection times T_i^{inf}.

for $j = 1, \ldots, (n-1)$ **do**
  if $J(j) = 0$ **then**
    $G^{anc} \Leftarrow S_{\text{exo}}$
  else
    $G^{anc} \Leftarrow \check{S}_j^{inf}$
  end if
  for $i = 1, \ldots, (n-1)$ **do**
    if $J(i) = j$ **then**
      $G^{(1)} \Leftarrow S_j^{end}$
      $G^{(2)} \Leftarrow S_i^{end}$
      $\check{S}_i^{inf} \Leftarrow \mathbf{G}(G^{anc}, G^{(1)}, G^{(2)})$
      $G^{anc} \Leftarrow \check{S}_i^{inf}$
    end if
  end for
end for

3.7. MCMC

To estimate parameters and latent variables, in particular the transmission tree J, Mollentze et al. (2014) built an MCMC algorithm that samples in:

$$\begin{aligned}\tilde{p}(J,T^{inf},L,D,\theta \mid data)&\\ \propto \tilde{p}(S^{end} \mid J,T^{inf},L,D,\theta,T^{end},X,S_{\text{exo}})&p(J,T^{inf} \mid L,D,\theta,T^{end},X,S_{\text{exo}})\\ \times p(L,D \mid \theta,T^{end},X,S_{\text{exo}})&p(\theta),\end{aligned} \tag{10}$$

Equation (10) is analogue to Equation (3) that gives the posterior distribution of parameters and latent variables, except that the genetic likelihood (Eq. (5)) is replaced by the genetic pseudo-likelihood (Eq. (8)).

Here, we propose to estimate parameters and latent variables with an MCMC algorithm that samples in:

$$\begin{aligned}\breve{p}(J,T^{inf},L,D,\theta \mid data)&\\ \propto \breve{p}(S^{end} \mid J,T^{inf},L,D,\theta,T^{end},X,S_{\text{exo}})&p(J,T^{inf} \mid L,D,\theta,T^{end},X,S_{\text{exo}})\\ \times p(L,D \mid \theta,T^{end},X,S_{\text{exo}})&p(\theta),\end{aligned} \tag{11}$$

Equation (11) is analogue to Equation (3), except that the genetic likelihood (Eq. (5)) is replaced by the genetic approximate likelihood (Eq. (9)).

As explained above, sampling in $\tilde{p}(J,T^{inf},L,D,\theta \mid data)$ instead of sampling in the exact posterior avoids us to handle the latent transmitted viral sequences. In contrast, sampling in $\breve{p}(J,T^{inf},L,D,\theta \mid data)$ requires us to handle these latent sequences but, using Algorithm 2, they are deterministically calculated given the values of J, S_{exo} and S^{end}. In both approaches, avoiding to stochastically update the transmitted viral sequences along the MCMC algorithm (with a Metropolis-Hastings sampler for example) significantly reduces the complexity of the algorithm. Note that computation costs are comparable when the pseudo-likelihood and the approximate likelihood are used: sampling in $\breve{p}(J,T^{inf},L,D,\theta \mid data)$ only took us about 1.05 times the duration required for sampling in $\tilde{p}(J,T^{inf},L,D,\theta \mid data)$.

For the application presented in the next section, we used both MCMC algorithms (the one based the pseudo-likelihood and the one based on the approximate likelihood). Proposal distributions and initialization were approximately similar than those used by Mollentze et al. (2014). It has to be noted that the transmission tree J was initialized such that $J(i) = 0$ for all $i \in \{1,\ldots,n\}$ (i.e. all individuals are initially infected by the exogenous source). Each MCMC algorithm was run for 100,000 iterations. Following a burn-in of 10,000 iterations, we sampled every 250th iterations.

4. Application

We carried out a simulation study to assess the efficiency of the two MCMC algorithms presented in Subsection 3.7. In this article, we assessed the algorithm efficiency by focusing on the estimation of the transmission tree J. We also evaluated how this efficiency is decreased when a fraction of genetic data are missing (i.e. when pathogen is not collected on some of the observed individuals and, therefore, not sequenced). Regarding the estimation of model parameters that is not studied here, Morelli et al. (2012) and Mollentze et al. (2014) provided extensive results on the coverage of parameters by posterior intervals provided by our approach.

Soumis au Journal de la Société Française de Statistique
File: seir.tex, compiled with jsfds, version : 2009/12/09
date: September 29, 2015

4.1. Simulation model

For each simulation, the epidemic is initiated at time zero with one infected host localized at the origin $(0,0)$ and 119 susceptible hosts uniformly and randomly localized in the $[0.0,0.3] \times [0.0,0.1]$ rectangle. At time zero, the sequence of the virus in the infected host purely consists of A nucleobases. The incubation parameters are $\beta_1 = 50$ (mean) and $\beta_2 = 5$ (standard deviation). The infectious duration parameters are $\delta_1 = 10$ (mean) and $\delta_2 = 1$ (standard deviation). The unit of β_1, β_2, δ_1 and δ_2 is an arbitrary time unit that is typically the day. The infection strength of each infectious host is $\alpha_1 = 10^4$. The basic risk parameter is $\alpha_0 = 0$. The dispersal parameters are $\alpha_{2,1} = 0.002$ (scale) and $\alpha_{2,2} = 1$ (shape). The substitution rates are $\mu_1 = \mu_2 = \mu_3 = 2 \times 10^{-5}$ (in substitutions per nucleotide per time unit). This value is rather high but is included in current values observed for RNA viruses (Hanada et al., 2004; Jenkins et al., 2002). Only a fraction of the $n = 120$ hosts is sampled: hosts with x-coordinate larger than 0.2 have a probability equal to 3/4 to be sampled, hosts with x-coordinate lower than or equal to 0.2 are not sampled (the sampling region is reduced to the square $[0.2,0.3] \times [0.0,0.1]$). For genetic data, a sequence fragment of length $s = 800$ sites is observed. The values $\beta_1 = 50$, $\delta_1 = 10$, $\mu_1 = \mu_2 = \mu_3 = 2 \times 10^{-5}$ and $s = 800$ lead, in expectation, to 0.96 substitutions per infected host over the observed sequence fragment. Figure 4 presents an example of simulated epidemic.

The simulation model is not exactly the same than the model used for the inference. Indeed, there is a difference concerning infection from external sources. In the simulation model, the external sources are handled in a realistic manner: the external sources are infectious hosts within or outside the sampling region generating infectious risks that are local in time and space. In the inference model, exogenous infections are due to a single source that has a constant infection strength (measured by α_0) in time and space and that is characterized by the sequence S_{exo} at time t_{exo}. This trick allows us to not have to explicitly deal with multiple unobserved sources (and therefore supplementary latent variables) in the estimation algorithm. It has to be noted that the parameter α_0 has not the same meaning in the simulation model and the inference model.

4.2. Prior distributions

Independent prior distributions were used for all parameters. Exponential priors with mean values 10^6 were chosen for α_0 and α_1. An exponential prior with mean value 1 was chosen for $\alpha_{2,1}$, while an informative gamma prior distribution with mean 1 and standard deviation 1 was specified for $\alpha_{2,2}$. Doing so allows classical kernels to have a non-negligible weight, in particular the thin-tailed normal kernel ($\alpha_{2,2} = 2$), the exponential kernel ($\alpha_{2,2} = 1$) and the fat-tailed kernel corresponding to $\alpha_{2,2} = 0.5$.

For the parameters governing incubation and infectious periods, we used very narrow prior distributions centered around true parameter values. Thus, we consider a situation where accurate information about these quantities are available (e.g. see Hampson et al., 2009, for such information in the case of canine rabies). Narrow prior distributions were used to prevent the inference algorithm from creating direct connections by extending the incubation and infectious periods when one or more intermediate cases have not been sampled (Note that in a situation where all infected cases are observed, the parameters governing incubation and infectious periods can be estimated with our approach Morelli et al., 2012). A narrow gamma prior with mean 50

and standard deviation 0.01 was specified for β_1. A narrow gamma prior with mean 5 days and standard deviation 0.01 was specified for β_2. A narrow gamma prior with mean 10 and standard deviation 0.01 was specified for δ_1. A narrow gamma prior with mean 1 and standard deviation 0.01 was specified for δ_2.

Exponential prior distributions with mean parameter $m = 2 \times 10^{-5}$ substitutions per nucleotide per day were used for the substitution rates μ_1, μ_2 and μ_3. A normal prior distribution with mean 0 and standard deviation 10 was specified for t_{exo}.

4.3. Results

Fifty data sets were independently simulated using the procedure described above. For each data set, we applied four different estimation algorithms:

- in the first one, we used the genetic pseudo-likelihood and all genetic data;
- in the second one, we used the genetic approximate likelihood and all genetic data;
- in the third one, we used the genetic approximate likelihood and 50% of genetic data (50% of the available sequences were randomly selected and used for the inference, the other sequences were ignored);
- in the fourth one, we used the genetic approximate likelihood and 25% of genetic data (25% of the available sequences were randomly selected and used for the inference, the other sequences were ignored).

It has to be noted that the genetic sampling effort (which varies) is different from the epidemiological sampling effort (which is constant in expectation): the sampling scheme described in Subsection 4.1 leads to observe about 3/4 of the infected hosts located within the study region; then, the virus is sampled and sequenced for either 100%, 50% or 25% of the observed hosts.

Comparing the two first algorithms will allow us to assess which of the approaches based on the pseudo-likelihood and the approximate likelihood is more accurate for reconstructing the transmission tree in the simulation setting described in Subsection 4.1.

Comparing the three last algorithms will allow us to assess how the reconstruction of the transmission tree relies on the genetic data, by keeping constant the other components of the reconstruction (i.e. the epidemiological data and the inference algorithm).

The algorithms were compared by assessing their ability to correctly estimate endogenous and exogenous transmissions. Endogenous transmissions designate direct transmissions between observed cases whereas exogenous transmissions designate infections of observed cases by unobserved cases. Three criteria were used to compare the performance of the algorithms. They are defined as follows, for each simulation and each algorithm:

- Average posterior probability of true endogenous transmissions:

$$\mathscr{C}_1 = \frac{1}{\sum_{i=1}^{n} \mathbf{1}\{J^{true}(i) > 0\}} \sum_{i=1}^{n} \mathbf{1}\{J^{true}(i) > 0\} \hat{p}\{J(i) = J^{true}(i) \mid data\},$$

  where J^{true} is the true transmission tree that was simulated, $\hat{p}\{J(i) = J^{true}(i) \mid data\}$ is the MCMC-based assessment of the posterior probability that the infecting source of i is the

true one,

$$\hat{p}\{J(i) = J^{true}(i) \mid data\} = \frac{1}{Z}\sum_{z=1}^{Z} \mathbf{1}\{J^{(z)}(i) = J^{true}(i)\}\},$$

$J^{(z)}$ is the z-th state of J in the posterior sample obtained from the MCMC, and Z is the size of the posterior sample of $(J, T^{inf}, L, D, \theta)$;

– Average posterior probability of true exogenous transmissions:

$$\mathscr{C}_2 = \frac{1}{\sum_{i=1}^{n} \mathbf{1}\{J^{true}(i) = 0\}} \sum_{i=1}^{n} \mathbf{1}\{J^{true}(i) = 0\}\hat{p}\{J(i) = J^{true}(i) \mid data\};$$

– Average posterior probability of false exogenous transmissions:

$$\mathscr{C}_3 = \frac{1}{\sum_{i=1}^{n} \mathbf{1}\{J^{true}(i) > 0\}} \sum_{i=1}^{n} \mathbf{1}\{J^{true}(i) > 0\}\hat{p}\{J(i) = 0 \mid data\},$$

where $\hat{p}\{J(i) = 0 \mid data\}$ is the MCMC-based assessment of the posterior probability that the infecting source of i is an external source,

$$\hat{p}\{J(i) = 0 \mid data\} = \frac{1}{Z}\sum_{z=1}^{Z} \mathbf{1}\{J^{(z)}(i) = 0\}\}.$$

Table 6 provides the means and standard deviations of $\mathscr{C}_1$, $\mathscr{C}_2$ and $\mathscr{C}_3$ computed from the 50 simulations and using the four estimation algorithms. Using the genetic approximate likelihood instead of the pseudo-likelihood allowed us to significantly improve the identification of true endogenous sources when all genetic data were used (the mean of $\mathscr{C}_1$ was increased from 0.75 to 0.82; p-value of pairwise t-test: 0.018).

When the genetic sampling effort was decreased from 100% to 50% and 25%, the correct identification of both endogenous and exogenous sources was more uncertain but remained relatively high. Indeed, in average, there were 31 observed hosts in the simulated data sets and consequently, for each infected host there were 30 possible endogenous sources and 1 possible exogenous source. Therefore, obtaining mean values for $\mathscr{C}_1$ and $\mathscr{C}_2$ equal to 0.36 and 0.72 indicates that, in average, the true source is identified with a relatively high probability. We can notice that the decrease in the identification of true endogenous transmissions is mostly counterbalanced by an increase in the incorrect detection of exogenous transmissions. Thus, in general, when an endogenous source is not identified correctly, the algorithm tends to assign the infection to an external source but not to an incorrect endogenous source.

5. Discussion

In this article, we presented an extension of stochastic SEIR models that combines (i) an individual-based, spatial, semi-Markov SEIR model for the spatio-temporal dynamics of the pathogen, and (ii) a Markovian evolutionary model for the temporal evolution of genetic sequences of the pathogen. The resulting model is a state-space model including many latent variables. The high

TABLE 6. *Means and standard deviations of average posterior probabilities of true endogenous transmissions ($\mathscr{C}_1$), true exogenous transmissions ($\mathscr{C}_2$) and false exogenous transmissions ($\mathscr{C}_3$). Means and standard deviations were computed from 50 simulations & estimations. Endogenous transmissions designate direct transmissions between observed cases whereas exogenous transmissions designate infections of observed cases by unobserved cases. Four estimation algorithms were applied, one based on the genetic pseudo-likelihood and three based on the genetic approximate likelihood with three different levels of genetic sampling effort. The genetic sampling effort is the percentage of observed cases for which the virus was sequenced and used in the inference.*

Method	Pseudo-likelihood	Approximate likelihood		
Genetic sampling effort (%)	100	100	50	25
$\mathscr{C}_1$ (true endogenous transmissions)	0.75 (0.20)	0.82 (0.20)	0.48 (0.27)	0.36 (0.31)
$\mathscr{C}_2$ (true exogenous transmissions)	0.80 (0.18)	0.80 (0.18)	0.72 (0.21)	0.72 (0.21)
$\mathscr{C}_3$ (false exogenous transmissions)	0.14 (0.18)	0.11 (0.17)	0.26 (0.26)	0.34 (0.33)

dimension of the set of latent variables lead to inference difficulties. The main difficulty concerns the manner to handle latent genetic sequences (i.e. the transmitted pathogen sequences at infection times). We proposed a new algorithm including a new way to deal with latent genetic sequences. We assessed this new algorithm with respect to its capacity to reconstruct transmission trees (i.e. who infected whom). This new algorithm outperforms the one proposed by Mollentze et al. (2014). We also assessed the performance of the algorithm when pathogen sequences are observed for only a fraction of sampled hosts. The results that we obtained show that including more genetic data significantly improves the reconstruction of transmission trees. It has to be noted that similar studies could be carried out to investigate, for example, the influence of spatial data, contact tracing information, substitution rate heterogeneity and dissemination capacities in the reconstruction of transmission trees, but also in the estimation of key epidemic parameters and latent variables such as unobserved infection times. These influence analyses were out of the scope of the present introductory article but they are obviously of major interest.

Several directions can be explored to improve the space-time-genetic SEIR approach. One of these directions concerns the computation / approximation of the genetic likelihood. We based our approximation on the reconstruction of transmitted genetic sequences based on the parsimony principle. While the parsimony principle has drawbacks in phylogenetic analyses (Brocchieri, 2001), it can be viewed as a nonparametric approach that is robust to model misspecifications (Kolaczkowski and Thornton, 2004). One of its main interests is its low computation cost with respect to the costs of probabilistic alternatives reviewed in Brocchieri (2001). However, we can improve how the parsimony principle is applied in our approach. Let us give below an example of possible improvement (notations are those of Section 3.6). If $g_b^{(1)} \neq g_b^{(2)}$ and $g_b^{(1)} \neq g_b^{anc}$ and $g_b^{(2)} \neq g_b^{anc}$, then our algorithm states $g_b^{node} = g_b^{anc}$. This value for g_b^{node} requires two substitutions: one from g_b^{node} to $g_b^{(1)}$ and one from g_b^{node} to $g_b^{(2)}$. Fixing $g_b^{node} = g_b^{(1)}$ (resp. $g_b^{(2)}$) also requires two substitutions: one from g_b^{anc} to g_b^{node} and one from g_b^{node} to $g_b^{(2)}$ (resp. $g_b^{(1)}$). Thus, the parsimony principle does not lead to a unique solution. The criterion using the differences in nucleobases could be augmented by a criterion on temporal lengths of branches joining G^{anc}— G^{node}, G^{node}— $G^{(1)}$, and G^{node}— $G^{(2)}$. For example, we could proceed as follows: If $g_b^{(1)} \neq g_b^{(2)}$ and $g_b^{(1)} \neq g_b^{anc}$ and $g_b^{(2)} \neq g_b^{anc}$, then g_b^{node} is equal to the bth nucleobase of the genome among $\{G^{anc}, G^{(1)}, G^{(2)}\}$ that is the nearest in time from G^{node}. This could lead to a clear improvement when G^{anc} is S_{exo}, because the date t_{exo} of S_{exo} is often far away from dates of

bifurcation nodes coinciding with infection times of observed individuals.

The approach presented in this article for reconstructing the virus sequences at the internal nodes (and its possible improvement described above) exploits the knowledge of the transmission tree J (more exactly the state of J along the MCMC algorithm) and the knowledge of a common ancestral sequence S_{exo}. The knowledge of J implies that the topology of the evolutionary tree is known (at each stage of the MCMC algorithm) and, consequently, that we have to solve the so-called small parsimony problem (i.e. finding the most parsimonious labeling of the internal vertices in an evolutionary tree; Jones and Pevzner, 2004, chap. 10). This problem is classically solved by the algorithms of Fitch (1971), Hartigan (1973) and Sankoff and Rousseau (1975) that consist of two stages: (1) a *bottom-up phase* where candidate states for the internal nodes are determined and eventually associated with a score; and (2) a *top-down refinement* where one makes a choice for the sequence state at the tree root and one decides between the candidate states for the other internal nodes by minimizing a total score. The additional knowledge of a common ancestral sequence S_{exo} allows us to simply perform a *top-down phase* for reconstructing the internal nodes. Thus, Algorithm 2 that is conditional on J and S_{exo} allows us to divide the number of operations at least by a factor 2 compared with the algorithms mentioned above. In practice, S_{exo} will often have to be preliminary reconstructed, but this will be done once and not at each update in the MCMC algorithm that lead to the modification of the genetic likelihood. It has to be noted that Algorithm 2 is not demonstrated to yield the most parsimonious reconstruction of transmitted sequences, but is expected to yield a relatively parsimonious reconstruction because it is based on successive partial reconstructions (i.e. successive calls of Algorithm 1) and each of these partial reconstructions is the most parsimonious.

The reconstruction of evolutionary trees and ancestral sequences has also been tackled with likelihood-based methods such as the maximum likelihood, empirical Bayes and hierarchical Bayes approaches (Hanson-Smith et al., 2010; Huelsenbeck and Bollback, 2001; Pirie et al., 2012). Despite several similarities between (i) the construction of the likelihood defined in these approaches and (ii) the construction of our likelihood, these likelihoods are different because we make specific assumptions about the dependencies between temporal, spatial and genetic variables. However, from an algorithmic point of view, the numerical approaches developed for the reconstruction of evolutionary trees and ancestral sequences could be adapted to our setting. Thus, we could replace our mixed likelihood/parsimony-based inference algorithm by a fully likelihood-based algorithm. Future studies could be carried out in this direction to compare the performance of these algorithms.

As we mentioned in the introduction, this article gives probabilists and statisticians an invitation for enriching the class of space-time-genetic SEIR models and developing efficient inference algorithms in order to carry out high-impact analyses of infectious diseases due to viruses. This invitation is particularly relevant because, today, (i) our connected world favors in some ways the spread of infectious diseases, and (ii) sequencing genetic material has become so affordable that they are available for many infectious diseases. New statistical methods combining spatial, temporal and genetic data and estimating transmission trees will become very important for future epidemiology.

Soumis au Journal de la Société Française de Statistique
File: seir.tex, compiled with jsfds, version : 2009/12/09
date: September 29, 2015

Acknowledgements

A large part of the content in this manuscript is a probabilistic and statistical formalization of previous articles, namely Morelli et al. (2012) and Mollentze et al. (2014), and the author wish to thanks all co-authors of these articles. This work was funded by the ModEE grant (SPE division of INRA).

Appendix A: On the distribution of susceptible state durations

In each SEIR model presented in Section 2, the transition rate governing the changes of the susceptible state $\mathbf{S} \to \mathbf{S} - 1$ (resp. $\mathbf{S}_k \to \mathbf{S}_k - 1$) is a function of the infectious state(s) $\mathbf{I}$ (resp. $\{\mathbf{I}_j : j \neq k\}$). Suppose that these infectious states are constant in time (which is not true in general), then the duration required for a change in $\mathbf{S}$ or $\mathbf{S}_k$ is exponentially distributed. However, $\mathbf{I}$ and $\{\mathbf{I}_j : j \neq k\}$ are randomly varying. Therefore, (i) the transition rate governing the changes of the susceptible state $\mathbf{S}$ or $\mathbf{S}_k$ is a random time-varying function and (ii) the duration required for a change in $\mathbf{S}$ or $\mathbf{S}_k$ is not exponential in general.

For instance, consider the stochastic time process $\{\mathbf{S}(t) : t \geq 0\}$ defined in Subsection 2.1. The times at which $\mathbf{S}$ changes (i.e. makes a negative jump of one unit) form a Cox process (or doubly stochastic point process; Daley and Vere-Jones, 2003; Grandell, 1976) where the stochastic intensity function is $t \mapsto \alpha_0 \mathbf{S}(t) + \alpha_1 \mathbf{S}(t)\mathbf{I}(t)/\mathbf{T}(t)$. Note that this intensity function is zero once $\mathbf{S}(t) = 0$. Closed (and non-exponential) forms of the distribution of inter-point times have only been obtained for specific Cox processes (e.g. Dassios and Jang, 2008). In the general case, computing the distribution of inter-point times requires an integration over the possible realizations of the intensity function.

References

Anderson, R. M., Donnelly, C. A., Ferguson, N. M., Woolhouse, M. E., Watt, C., Udy, H., MaWhinney, S., Dunstan, S., Southwood, T., Wilesmith, J., et al. (1996). Transmission dynamics and epidemiology of BSE in British cattle. *Nature*, 382:779–788.

Barbu, V. S. and Limnios, N. (2008). *Semi-Markov chains and hidden semi-Markov models toward applications — Their Use in Reliability and DNA Analysis*, volume 191 of *Lecture Notes in Statistics*. Springer.

Britton, T. and Giardina, F. (2015). Introduction to statistical inference for infectious diseases. *Journal de la SFDS*.

Brocchieri, L. (2001). Phylogenetic inferences from molecular sequences: Review and critique. *Theoretical Population Biology*, 59:27–40.

Cambra, M., Capote, N., Myrta, A., and Llácer, G. (2006). Plum pox virus and the estimated costs associated with sharka disease. *EPPO Bulletin*, 36:202–204.

Daley, D. and Vere-Jones, D. (2003). An introduction to the theory of point processes, volume i: Elementary theory and methods of probability and its applications.

Dassios, A. and Jang, J. (2008). The distribution of the interval between events of a cox process with shot noise intensity. *International Journal of Stochastic Analysis*, 2008:367170.

Fitch, W. M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Biology*, 20:406–416.

Grandell, J. (1976). *Doubly stochastic Poisson processes*, volume 529. Springer, Berlin.

Hall, M. and Rambaut, A. (2014). Epidemic reconstruction in a phylogenetics framework: transmission trees as partitions. *arXiv:1406.0428*.

Hampson, K., Dushoff, J., Cleaveland, S., Haydon, D. T., Kaare, M., Packer, C., and Dobson, A. (2009). Transmission dynamics and prospects for the elimination of canine rabies. *PLoS Biology*, 7:e1000053.

Soumis au Journal de la Société Française de Statistique
File: seir.tex, compiled with jsfds, version : 2009/12/09
date: September 29, 2015

Hanada, K., Suzuki, Y., and Gojobori, T. (2004). A large variation in the rates of synonymous substitution for rna viruses and its relationship to a diversity of viral infection and transmission modes. *Molecular biology and evolution*, 21:1074–1080.

Hanson-Smith, V., Kolaczkowski, B., and Thornton, J. W. (2010). Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Molecular biology and evolution*, 27:1988–1999.

Hartigan, J. A. (1973). Minimum mutation fits to a given tree. *Biometrics*, pages 53–65.

Haydon, D. T., Kao, R. R., and Kitching, R. P. (2004). The UK foot-and-mouth disease outbreak — the aftermath. *Nature Reviews Microbiology*, 2:675–681.

Huelsenbeck, J. P. and Bollback, J. P. (2001). Empirical and hierarchical bayesian estimation of ancestral states. *Systematic biology*, 50:351–366.

Jenkins, G. M., Rambaut, A., Pybus, O. G., and Holmes, E. C. (2002). Rates of molecular evolution in rna viruses: a quantitative phylogenetic analysis. *Journal of molecular evolution*, 54:156–165.

Jombart, T., Cori, A., Didelot, X., Cauchemez, S., Fraser, C., and Ferguson, N. (2014). Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Computational Biology*, 10:e1003457.

Jones, N. C. and Pevzner, P. (2004). *An introduction to bioinformatics algorithms*. MIT press.

Kimura, M. (1981). Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl Acad. Sci.*, 78:454–458.

Kolaczkowski, B. and Thornton, J. W. (2004). Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature*, 431:980–984.

Lemey, P., Rambaut, A., Drummond, A. J., and Suchard, M. A. (2009). Bayesian phylogeography finds its roots. *PLoS Comput. Biol.*, 5:e1000520.

Lemey, P., Rambaut, A., Welch, J. J., and Suchard, M. A. (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.*, 27:1877–1885.

Mollentze, N., Nel, L. H., Townsend, S., le Roux, K., Hampson, K., Haydon, D. T., and Soubeyrand, S. (2014). A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proceedings of the Royal Society B*, 281:20133251.

Morelli, M. J., Thébaud, G., Chadœ uf, J., King, D. P., Haydon, D. T., and Soubeyrand, S. (2012). A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *Plos Computation Biology*, 8:e1002768.

Pirie, M. D., Humphreys, A. M., Antonelli, A., Galley, C., and Linder, H. P. (2012). Model uncertainty in ancestral area reconstruction: a parsimonious solution? *Taxon*, 61:652–664.

Pybus, O. G., Suchard, M. A., Lemey, P., Bernardin, F. J., Rambaut, A., Crawford, F. W., Gray, R. R., Arinaminpathy, N., Stramer, S. L., Busch, M. P., and Delwart, E. L. (2012). Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proceedings of the National Academy of Sciences*, 109:15066–15071.

Rambaut, A., Pybus, O. G., Nelson, M. I., Viboud, C., Taubenberger, J. K., and Holmes, E. C. (2008). The genomic and epidemiological dynamics of human influenza A virus. *Nature*, 453:615–619.

Rasmussen, D. A., Ratmann, O., and Koelle, K. (2011). Inference for nonlinear epidemiological models using genealogies and time series. *PLoS computational biology*, 7:e1002136.

Sankoff, D. and Rousseau, P. (1975). Locating the vertices of a steiner tree in an arbitrary metric space. *Mathematical Programming*, 9:240–246.

Stadler, T. and Bonhoeffer, S. (2013). Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Phil. Trans. Roy. Soc. B*, 368:20120198.

Takahata, N. and Kimura, M. (1981). A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. *Genetics*, 98:641–657.

Tuffley, C. and Steel, M. (1997). Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bulletin of Mathematical Biology*, 59:581–607.

Ypma, R. J., Jonges, M., Bataille, A., Stegeman, A., Koch, G., van Boven, M., Koopmans, M., van Ballegooijen, W. M., and Wallinga, J. (2013). Genetic data provide evidence for wind-mediated transmission of highly pathogenic avian influenza. *Journal of Infectious Diseases*, 207:730–735.

Ypma, R. J. F., Bataille, A. M. A., Stegeman, A., Koch, G., Wallinga, J., and van Ballegooijen, W. M. (2012). Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proceedings of the Royal Society B*, 279:444–450.

Soumis au Journal de la Société Française de Statistique
File: seir.tex, compiled with jsfds, version : 2009/12/09
date: September 29, 2015

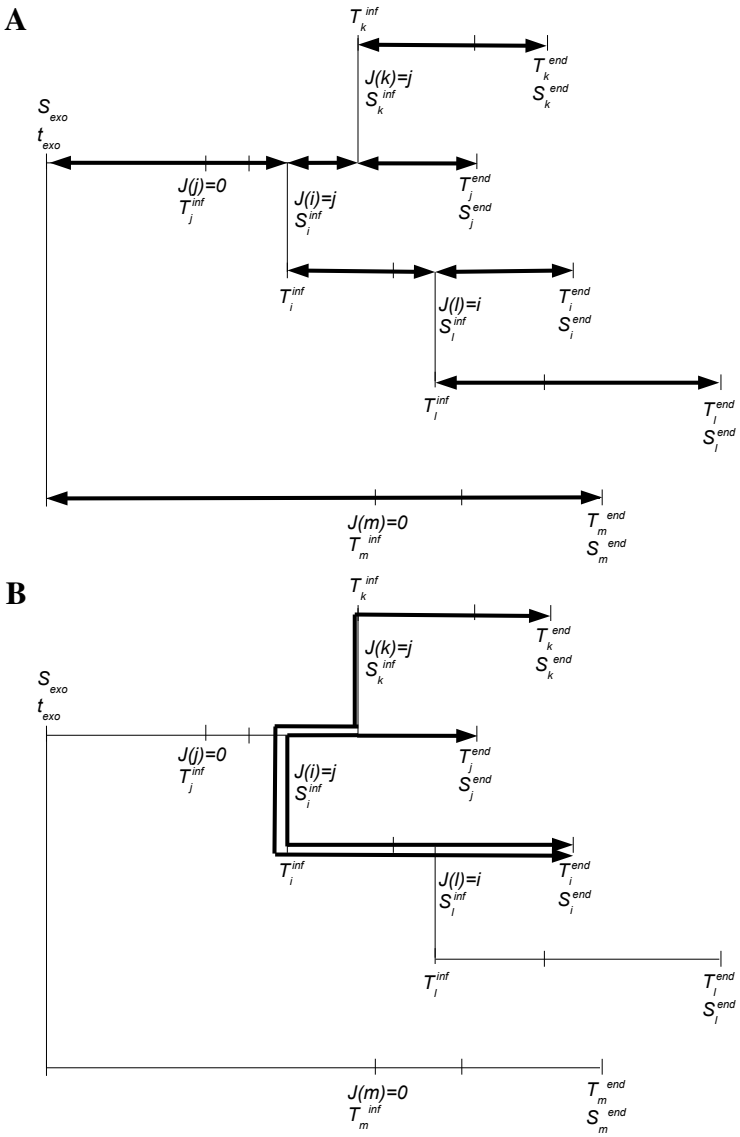


FIGURE 3. *Example of transmission tree with five individuals i, j, k, l and m, where j and m are infected by the exogenous source and i, k and l are infected by observed individuals. On horizontal lines, that are time lines, the first tick indicates the infection time, the second the end of exposed stage, and the third the observation time. A: Evolution of sequences are independent on tree segments marked with double arrows; the sequences S^{inf} at infection times appear in the genetic likelihood given in Subsection 3.4 and in its approximation given in Subsection 3.6. B: In the genetic pseudo-likelihood of Subsection 3.5, the probability of observing S_i^{inf}, for example, is assessed by computing the probability of evolution between S_i^{inf} and S_k^{inf} and the probability of evolution between S_i^{inf} and S_j^{inf}, without taking into account the dependence due to the overlapping of these two evolutions.*

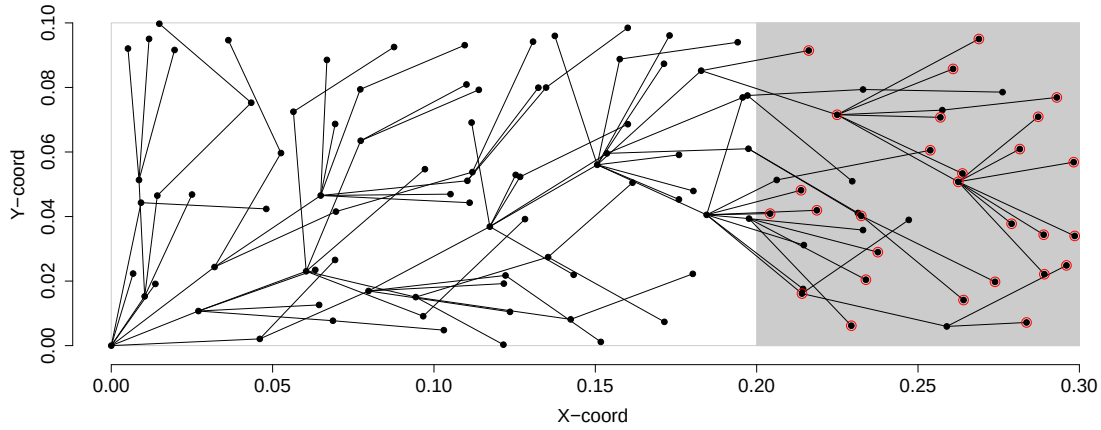


FIGURE 4. *Example of simulated epidemic. Black dots represent infected hosts, while black segments represent transmissions. Samples are only taken from a sub-region, colored in gray, and in this area, not all cases are detected or sampled. Infected hosts have a probability of 3/4 of being sampled. Points corresponding to sampled hosts are surrounded by red circles.*

Soumis au Journal de la Société Française de Statistique
File: seir.tex, compiled with jsfds, version : 2009/12/09
date: September 29, 2015