Special Issue Paper

Statistics in Medicine

Received 28 July 2016, Accepted 18 November 2016 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7196

Bayesian nonparametric clustering in phylogenetics: modeling antigenic evolution in influenza

Gabriela B. Cybis,[a*†] Janet S. Sinsheimer,[b,c,d] Trevor Bedford,[e] Andrew Rambaut,[f,g] Philippe Lemey[h] and Marc A. Suchard[b,c,d]

Influenza is responsible for up to 500,000 deaths every year, and antigenic variability represents much of its epidemiological burden. To visualize antigenic differences across many viral strains, antigenic cartography methods use multidimensional scaling on binding assay data to map influenza antigenicity onto a low-dimensional space. Analysis of such assay data ideally leads to natural clustering of influenza strains of similar antigenicity that correlate with sequence evolution. To understand the dynamics of these antigenic groups, we present a framework that jointly models genetic and antigenic evolution by combining multidimensional scaling of binding assay data, Bayesian phylogenetic machinery and nonparametric clustering methods. We propose a phylogenetic Chinese restaurant process that extends the current process to incorporate the phylogenetic dependency structure between strains in the modeling of antigenic clusters. With this method, we are able to use the genetic information to better understand the evolution of antigenicity throughout epidemics, as shown in applications of this model to H1N1 influenza.

Keywords: Bayesian nonparametric mixture models; phylodynamics; antigenic cartography

1. Introduction

Every year, 10% to 20% of the world population is affected by influenza epidemics, with a death toll of approximately half a million people. Additionally, there is an enormous disease burden associated with influenza, with economic losses reaching an estimated 87 billion dollars for seasonal influenza in the USA alone [1–3]. The World Health Organization carefully monitors the influenza epidemics and defines strategies regarding disease control, including vaccine design. One of the main challenges for vaccination is the continuous evolution of viral immunogenic proteins to evade immune response, known as antigenic drift. To be effective, vaccines must be designed to specifically match the antigenic types circulating after they are administered, and these do not necessarily coincide with those circulating at the time of design. Consequently, an understanding of how viral antigenicity evolves over time is paramount for the efficacy of future influenza vaccination campaigns. In this paper, we present methodology to study this evolutionary process through the perspective of clusters of viruses with similar antigenicity and their relation to genetic evolution.

To characterize changes in antigenicity, researchers use information on how strongly the viral proteins from various strains interact with sera from hosts that are immunologically challenged by a specific,

[a]*Department of Statistics, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil*
[b]*Department of Biomathematics, David Geffen School of Medicine at UCLA, Los Angeles, California, U.S.A.*
[c]*Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, California, U.S.A.*
[d]*Department of Biostatistics, UCLA Fielding School of Public Health, Los Angeles, California, U.S.A.*
[e]*Fred Hutchinson Cancer Research Center, Seattle, Washington, U.S.A.*
[f]*Institute of Evolutionary Biology, University of Edinburgh, Ashworth Laboratories, Edinburgh, U.K.*
[g]*Fogarty International Center, National Institutes of Health, Bethesda, Maryland, U.S.A.*
[h]*Department of Microbiology and Immunology, Rega Institute, University of Leuven, Leuven, Belgium*
**Correspondence to: Gabriela B. Cybis, Department of Statistics, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil.*
[†]*E-mail: gabriela.cybis@ufrgs.br*

often different viral strain. This information is traditionally obtained from binding assays that measure the affinity of the two main immunogenic viral proteins, hemagglutinin (HA) and neuraminidase, to the sera [4]. Antibodies that bind to these proteins prevent attachment of the virus to red blood cells, a process called hemagglutination. The resulting assay data are sparse matrices of hemagglutination inhibition (HI) titers, the highest dilution of serum that inhibits hemagglutination; these are censored, intervaled, and noisy measurements. To visualize patterns across many viral strains, antigenic cartography methods use these binding assay data to map influenza antigenicity onto a low-dimensional space (with generally two dimensions). In these maps, points that are close together represent antigenically similar strains [5, 6].

Since their introduction in 2004, antigenic cartography methods have gained significant popularity and are currently among the tools used by World Health Organization for vaccine strain selection [7]. Antigenic cartography methods have been used to study differences in antigenic evolution between neuraminidase and HA proteins, the history of seasonal influenza in the last 35 years, and vaccination strategies for avian influenza [5, 8, 9]. Although the initial motivation was seasonal influenza, antigenic cartography methods have also found applications in horse and swine influenza, as well as in other diseases such as rabies, malaria, and dengue [7, 10].

Antigenic cartography studies in influenza have shown that the circulating strains tend to form aggregates based on similar antigenicity. Clusters are temporally oriented, so that in most years the circulating influenza strains belong to only one or two distinct clusters. Additionally, these studies show that antigenic evolution correlates with genetic evolution. Nevertheless, this correlation is not perfect. While genetic evolution is approximately continuous over time, antigenic evolution seems to be more punctuated [7], reflecting evolution at a restricted number of key antigenic sites [11].

For all its potential impact, methods that explicitly model the correlation between genetic evolution and antigenicity, as measured through antigenic cartography maps, have been conspicuously absent. Bedford *et al.* [12] lay groundwork for this type of study by presenting a probabilistic model for the multidimensional scaling of binding assay data. They use a Bayesian phylogenetic framework to connect molecular evolution to an antigenic map that lacks clustering. We build upon this work to create our phylogenetic antigenic clustering method. Our goal is to use the genetic information to better understand the evolution of antigenicity, the emergence of new antigenic groups, and the molecular changes that give rise to new clusters.

Our model focuses on the antigenic clusters. Because there are potentially an infinite number of them and the interactions of the evolutionary processes that govern antigenic evolution are not simply defined parametrically, we opt to use a nonparametric model. A canonical choice for modeling nonparametric clustering would be the Dirichlet mixture model, where the likelihood of the data is a mixture of normal distributions with the normal components representing the clusters [13]. However, in the Dirichlet mixture model, the data points are assumed to be exchangeable and there is no flexibility for a dependency structure. This is not appropriate for antigenic data, because the viral strains are related to each other through evolution. The resulting history of common ancestry, with some strains having more recently diverged than others, can be captured by a phylogenetic tree. Thus, for this problem, a method that considers the dependency structure between samples in determining the probabilities of clustering would be more adequate.

Recent developments in Bayesian nonparametrics have made it feasible to account for the dependency structure required to incorporate phylogenetic data in this clustering problem. Miller, Griffins, and Jordan [14] present a variation on the Indian buffet process that incorporates a non-exchangeable prior in the form of a tree. Dahl [15] develops a modification of the Chinese restaurant process (CRP) that considers distances between data points for computing the probabilities of cluster arrangements. Blei and Frazier [16] build an alternative representation of the CRP that also incorporates a distance matrix but presents a more efficient sampling scheme. Both CRP extensions reduce to the original CRP by an appropriate choice of parameters. We build upon the distance-dependent Chinese restaurant process (ddCRP) [16] as a clustering method for defining antigenic groupings. In our phylogenetic Chinese restaurant process (pCRP), distances between data points are informed by the phylogenetic tree.

An alternative approach to combine phylogenetic and antigenic data for this problem would be to first perform the multidimensional scaling on a combination of phylogenetic distances and antigenic data and then directly apply the standard CRP to the resulting map locations. However, this conditioning argument would introduce bias into the model, which we avoid by formulating a single, coherent probabilistic model.

In summary, our model follows Bedford *et al.* [12] in generating an antigenic map from binding assay data. The virus locations in the antigenic map are parameters of their probabilistic multidimensional scal-

ing model. We define the pCRP as a clustering prior for these location parameters. This prior assigns each viral strain to one antigenic cluster, such that probabilities of clusters are a function of phylogenetic relatedness. By jointly modeling the cartographic map, antigenic clustering, and molecular evolution, we effectively incorporate uncertainty about mapping, the unobserved phylogeny, and evolutionary parameters. Therefore, we are able to jointly estimate distributions for the antigenic map and clustering while assessing how these relate to molecular evolution. In the following section, we present our model and the sampling scheme that allows us to draw inference from it. Then, in Section 3, we illustrate applications of this model to H1N1 influenza. A discussion of the results, modeling, and future directions is presented in Section 4.

2. Methods

Consider a dataset of aligned molecular sequences $\mathbf{S}$ from N influenza strains and an $N \times M$ cross-reactivity matrix $\mathbf{H} = \{h_{ij}\}$ originating from HI assays for the N viral strains and M challenging sera. To assess antigenic similarities between viruses, HI assays measure the reactivity of one viral strain to serum-containing antibodies raised against another. These assessments are made through serial dilutions, and the cross-reactivity measure h_{ij} represents $\log_2$ of the largest dilution titer at which serum j is effective against viral strain i.

We model the sequence data $\mathbf{S}$ using standard Bayesian phylogenetics models [17] that include, among other evolutionary parameters $\boldsymbol{\theta}$, an unobserved phylogenetic tree F that represents the evolutionary relationship between the N viruses. This phylogenetic tree is a bifurcating, directed graph with N terminal nodes of degree 1 that correspond to the tips of the tree, $N-2$ internal nodes of degree 3, and a root node of degree 2. The edge weights between nodes are termed branch lengths and track-elapsed evolutionary time. Conditional on F, we assume independence between the sequence data $\mathbf{S}$ and assay data $\mathbf{H}$.

In our model, the assay data $\mathbf{H}$ are used to generate an antigenic cartography map of the N viruses. We then model the locations of the virus on this map as a pCRP that clusters viruses based on antigenic and phylogenetic distances, linking $\mathbf{H}$ and $\mathbf{S}$. In order to create a two-dimensional antigenic map that preserves the relationships represented in $\mathbf{H}$, we employ a Bayesian multidimensional scaling (BMDS) method [18] adapted from Bedford *et al.* [12].

2.1. *Bayesian multidimensional scaling of* $\mathbf{H}$

Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)^t$ be the $N \times 2$ matrix of virus locations in the antigenic cartography map, such that $\mathbf{X}_i = (x_{i1}, x_{i2})$ for $i = 1, \ldots, N$, where t indicates the matrix transpose. Likewise, let $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_M)^t$ represent the $M \times 2$ analogous matrix of antigenic coordinates for the sera, with $\mathbf{Y}_j = (y_{j1}, y_{j2})$ for $j = 1, \ldots, M$. BMDS is a probabilistic approach that estimates locations $\mathbf{X}$ and $\mathbf{Y}$ by matching immunologic distances from $\mathbf{H}$ to distances in the antigenic map [12].

If h'_{ij} represents the theoretical titer at which reactivity ceases between virus i and serum j on a continuous scale, the immunologic distance can be defined as

$$\Delta_{ij} = s_j - h'_{ij} \tag{1}$$

in which $s_j = \max\{h_{1j}, \ldots, h_{Nj}\}$ represents a fixed serum effect. Additionally, let $\delta_{ij} = ||\mathbf{X}_i - \mathbf{Y}_j||_2$ represent the Euclidean distance between $\mathbf{X}_i$ and $\mathbf{Y}_j$. BMDS assumes that the HI titers are normally distributed with variance φ^2 and mean such that the expected value for the immunologic distance Δ_{ij} is the map distance δ_{ij}; thus,

$$h'_{ij} \sim \mathcal{N}(s_j - \delta_{ij}, \varphi^2), \tag{2}$$

where $\mathcal{N}(\mu, \sigma^2)$ represents the normal distribution with mean μ and variance σ^2.

However, the observed h_{ij} are integer values representing the last titer in the serial dilution at which reactivity was detected; consequently, the matrix $\mathbf{H}$ is censored and intervaled. To circumvent this issue, we adopt the interpretation of an observed titer h_{ij} as a threshold that reports that reactivity has ceased somewhere between the titers h_{ij} and $h_{ij}+1$. Thus, the likelihood of an observed titer can be computed as

$$p(h_{ij}|\mathbf{X}_i, \mathbf{Y}_j, \varphi^2) = \phi\left(\frac{h_{ij} + \delta_{ij} - s_j + 1}{\varphi}\right) - \phi\left(\frac{h_{ij} + \delta_{ij} - s_j}{\varphi}\right), \tag{3}$$

where $\phi(\cdot)$ is the standard normal cumulative distribution function. When a serum and virus pair is not reactive for the smallest titer in the assay, the likelihood is defined analogously as an open interval, through the lower tail probability. Assuming independence between the observations in $\mathbf{H}$, the joint likelihood can be computed as

$$p(\mathbf{H}|\mathbf{X}, \mathbf{Y}, \varphi^2) = \prod_{(i,j)\in\mathcal{I}} p(h_{ij}|\mathrm{X}_i, \mathbf{Y}_j, \varphi^2), \tag{4}$$

where $\mathcal{I}$ is the set of virus/serum pairs (i, j) for which observations are available. We note that, because of experimental constraints, most of the $N \times M$ comparisons cannot be made, leading to an incomplete matrix $\mathbf{H}$ and identifiability issues in the model.

We address identifiability by adopting the drift prior of Bedford *et al.* [12] on the serum locations. The drift prior assumes that locations are normally distributed, and their expected values increase linearly with date of sampling along antigenic dimension 1. Thus

$$y_{j1} \sim \mathcal{N}(\beta t_j, \sigma_y^2) \text{ and } y_{j2} \sim \mathcal{N}(0, \sigma_y^2), \tag{5}$$

for $j = 1, \ldots, M$, where t_i is the difference between time of sampling of serum j and that of the oldest virus or serum in the sample, β is the drift parameter, and σ_y^2 is the serum prior variance. This choice of prior is motivated by the observation that antigenic distances in influenza tend to increase over time [5].

To complete specification of the BMDS model, we select gamma prior distributions for the multidimensional scaling precision $1/\varphi^2$, and the precision $1/\sigma_y^2$ of the serum drift prior. We also adopt a diffuse gamma prior on the serum drift parameter β. Finally, the prior distribution of virus locations $\mathbf{X}$ on the antigenic map is given by the pCRP, thus connecting BMDS to antigenic clustering.

2.2. *Phylogenetic Chinese restaurant process*

The CRP [19] is a stochastic process that can be used to generate samples from the Dirichlet process. It is usually understood through the following analogy: customers arrive in turn at a restaurant with an infinite number of tables and choose a table to sit at according to a predefined distribution. After the arrival of N customers, their distribution in this Chinese restaurant represents a random partition of customers into table groups. The ddCRP modifies the CRP to consider affinities between customers for table assignment [16]. We use this feature to incorporate phylogenetic distances into our model.

In our pCRP, each customer represents one of the N viral strains. Each table represents one antigenic cluster, so that all customers assigned to the same table represent viruses in the same antigenic cluster. Even though theoretically the pCRP has an infinite number of potential clusters, in one realization only a finite number K is observed.

The dependency structure between customers is represented by the phylogenetic distance matrix $\mathbf{D} = \{d_{il}\}$, in which d_{il} is computed as the sum of branch lengths separating viruses i and l on the tree F. The effect that the phylogenetic distances $\mathbf{D}$ have on antigenic clustering is modulated through a decay function $f(\cdot)$. We adopt the form $f(d) = 1/d^{\lambda}$, which for positive λ takes large distances and transforms them into low probabilities of belonging to the same cluster. The parameter λ can regulate the effects of differences in the d_{il}'s that may span many orders of magnitude, especially for asynchronous data, dampening or heightening their differences. When $\lambda = 0$, we have no phylogenetic effect on the clustering, and the pCRP becomes the standard CRP. We choose the simplest formulation for $f(\cdot)$ that carries these properties.

Under the Chinese restaurant analogy, the customer groupings in this pCRP act on two levels. On the first level, each customer chooses another customer with whom he would like to sit and forms one single directional link. Customer i forms a link to customer l with probability proportional to $f(d_{il})$. Alternatively, the customer might choose not to form a link with any other customer and form an auto-link instead, with probability proportional to α. If c_i represents the customer to which customer i is linked, and $\mathbf{c} = \{c_i\}$, then

$$p(\mathbf{c}|\mathbf{D}, \lambda, \alpha) = \prod_{i=1}^{N} \frac{\mathbf{1}_{\{c_i=i\}}\alpha + \mathbf{1}_{\{c_i\neq i\}} f(d_{i,c_i})}{\alpha + \sum_{\ell\neq i} f(d_{i,\ell})}. \tag{6}$$

On the second level, the set of customer links is converted into table assignments through the transform $z(\mathbf{c})$ that takes all connected customers and assigns them to the same table.

In our antigenic cartography setting, this translates into a set of links $\mathbf{c}$ between viruses that is resolved through $z(\mathbf{c})$ into antigenic cluster associations. The virus locations $\mathbf{X}$ in the antigenic map are modeled

as a mixture of normal distributions, where the mixture components are given by the antigenic clusters. Thus, to each antigenic cluster corresponds one mean vector $\boldsymbol{\mu}_k$ and one precision matrix $\boldsymbol{\Lambda}_k$, such that, if $\mathbf{z^k}$ represents the viruses in antigenic cluster k, then

$$\mathbf{X}_{\mathbf{z^k}}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k \sim \mathcal{MVN}(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}) \tag{7}$$

for $k = 1, \ldots, K$. Here, $\mathcal{MVN}(\cdot)$ represents the multivariate normal distribution.

For convenience, we use the conjugate normal–Wishart prior for the mean and precision parameters of the mixture components; thus,

$$\begin{aligned} \boldsymbol{\Lambda}_k &\sim \mathcal{W}(\mathbf{T}_0, u_0) \\ \boldsymbol{\mu}_k &\sim \mathcal{MVN}\left(\mathbf{m}_0, (\kappa_0 \boldsymbol{\Lambda}_k)^{-1}\right), \end{aligned} \tag{8}$$

for $k = 1, \ldots, K$. Here, $\mathcal{W}(\mathbf{T}, u)$ is the Wishart distribution with scale matrix $\mathbf{T}$ and ν degrees of freedom. For ease of notation, we collect all the hyperparameters for the cluster normal distributions in $\mathbf{G}_0 = \{\mathbf{m}_0, \kappa_0, \mathbf{T}_0, u_0\}$. Exploiting the conjugate prior, we can analytically integrate out the cluster mean and precision parameters and compute the density of the viral locations in antigenic cluster k as

$$p(x_{\mathbf{z^k}}|\mathbf{G}_0, \mathbf{c}) = \frac{1}{\pi^{N_k}} \frac{\Gamma_2(u_k/2)}{\Gamma_2(u_0/2)} \frac{|\mathbf{T}_k|^{u_k/2}}{|\mathbf{T}_0|^{u_0/2}} \frac{\kappa_0}{\kappa_k}, \tag{9}$$

for $k = 1, \ldots, K$. Here, N_k is the number of viruses in cluster k and

$$\begin{aligned} u_k &= u_0 + N_k, \\ \kappa_k &= \kappa_0 + N_k. \end{aligned} \tag{10}$$

Also, the posterior scale matrix can be obtained through

$$\begin{aligned} \mathbf{T}_k^{-1} = \mathbf{T}_0^{-1} &+ \sum_{\mathbf{z^k}} (\mathbf{X}_i - \bar{\mathbf{X}}_{\mathbf{z^k}})(\mathbf{X}_i - \bar{\mathbf{X}}_{\mathbf{z^k}})^t \\ &+ \frac{\kappa_0 N_k}{\kappa_0 + N_k} (\bar{\mathbf{X}}_{\mathbf{z^k}} - \mathbf{m}_0)(\bar{\mathbf{X}}_{\mathbf{z^k}} - \mathbf{m}_0)^t, \end{aligned} \tag{11}$$

where $\bar{\mathbf{X}}_{\mathbf{z^k}}$ represents the mean location for viruses in cluster k and Γ_2 is the two-dimensional Gamma function. For further information on expression (9), we refer the reader to the Supporting Information.

Combining expressions (6) and (9), the joint density of the virus location matrix $\mathbf{X}$ and link vector $\mathbf{c}$ can be computed as

$$p(\mathbf{X}, \mathbf{c}|\mathbf{D}, \lambda, \alpha, \mathbf{G}_0) = p(\mathbf{c}|\mathbf{D}, \lambda, \alpha) \prod_{k=1}^{K} p(\mathbf{X}_{\mathbf{z^k}}|\mathbf{G}_0, \mathbf{c}). \tag{12}$$

We assume a priori that the tuning parameter λ of the decay function is normally distributed with zero mean and the concentration parameter α has an exponential distribution. Figure 1 presents a schematic representation of the pCRP.

2.3. *Inference*

The posterior distribution for our model can be expressed as

$$\begin{aligned} p(\mathbf{c}, \mathbf{X}, \mathbf{Y}, F, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\eta}|\mathbf{H}, \mathbf{S}) \propto\ & p(\mathbf{H}|\mathbf{X}, \mathbf{Y}, \boldsymbol{\psi}) \times p(\mathbf{X}, \mathbf{c}|F, \boldsymbol{\eta}) \\ & \times p(\mathbf{Y}, \boldsymbol{\psi}) \times p(\boldsymbol{\eta}) \times p(\mathbf{S}, \boldsymbol{\theta}, F), \end{aligned} \tag{13}$$

where $\boldsymbol{\eta} = \{\alpha, \lambda\}$ collects the parameters of the pCRP and $\boldsymbol{\psi} = \{\varphi^2, \beta, \sigma_y^2\}$ collects parameters of BMDS. To learn about this distribution, we employ Markov chain Monte Carlo. We exploit a random-scan Metropolis-within-Gibbs scheme. For the tree F and other phylogenetic parameters $\boldsymbol{\theta}$ modeling sequence evolution, we employ standard Bayesian phylogenetic algorithms [17] based on Metropolis–Hastings parameter proposals. For the parameters $\mathbf{X}$, $\mathbf{Y}$, and $\boldsymbol{\psi}$ of the BMDS model, we follow Bedford *et al.* [12] in using Metropolis–Hastings transition kernels.

A central issue for this model is sampling for the links $\mathbf{c}$ between viruses and consequently the cluster assignments. For this parameter, we employ a Gibbs sampling scheme akin to the one of Blei and Frazier

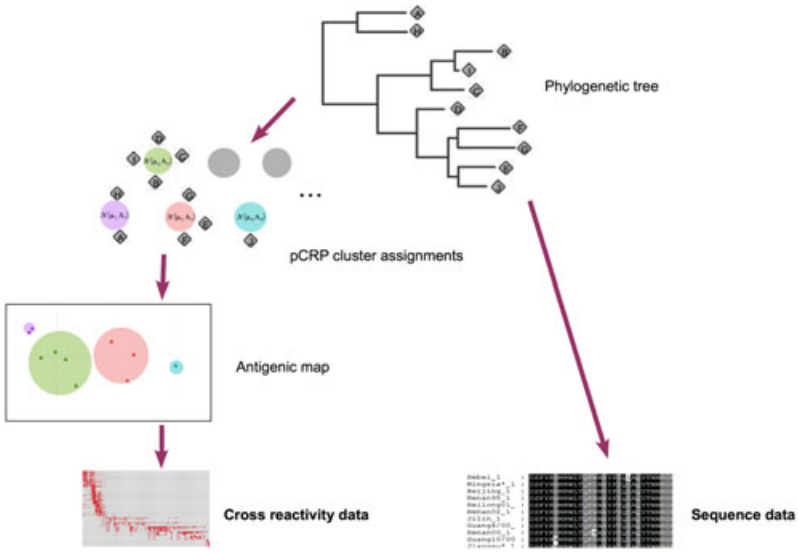

Figure 1. Schematic representation of the phylogenetic Chinese restaurant process (pCRP): The sequence data are modeled through a phylogenetic tree F. The tree also informs the probabilities of cluster assignments. Virus map locations $\mathbf{X}$ for each cluster are modeled as a Gaussian mixture component with mean $\boldsymbol{\mu}_k$ and precision $\boldsymbol{\Lambda}_k$. Finally, binding assay data $\mathbf{H}$ are modeled through Bayesian multidimensional scaling to generate the antigenic map.

[16] developed for the ddCRP. This Gibbs sampler explores the space of possible links between viruses by replacing at random one link at each step.

This individual link update to $\mathbf{c}$ can be understood in two steps. First, a virus i is chosen at random from a uniform distribution and the corresponding link is removed, resulting in a partition of $\mathbf{X}$ induced by the remaining links $\mathbf{c}_{-i}$. Subsequently, a new link c_i^* is created, rendering a new partition of $\mathbf{X}$ induced by $\mathbf{c}^* = (\mathbf{c}_{-i} \cup c_i^*)$.

Up to a constant, the conditional distribution of c_i^* can be computed as

$$\begin{aligned} p(c_i^*|\mathbf{c}_{-i}, \mathbf{X}, \mathbf{G}_0, \boldsymbol{\eta}, \mathbf{D}) &= \frac{p(\mathbf{c}^*|\mathbf{X}, \mathbf{G}_0, \boldsymbol{\eta}, \mathbf{D})}{p(\mathbf{c}_{-i}|\mathbf{X}, \mathbf{G}_0, \boldsymbol{\eta}, \mathbf{D})} \\ &\propto \left(\mathbf{1}_{\{c_i^*=i\}}\alpha + \mathbf{1}_{\{c_i^*\neq i\}} f(d_{i,c_i})\right) \frac{p(\mathbf{X}|\mathbf{G}_0, \mathbf{c}^*)}{p(\mathbf{X}|\mathbf{G}_0, \mathbf{c}_{-i})}. \end{aligned} \tag{14}$$

The creation of a new link i can have two possible effects on the virus partitions. First, virus i might form a new link to another virus that already belongs to the same cluster or form an auto-link: This operation does not change the data partition and thus does not affect the $p(\mathbf{X}|\mathbf{G}_0, \mathbf{c})$. Alternatively, virus i might form a link to a virus from another cluster: This would join the two clusters and therefore affect $p(\mathbf{X}|\mathbf{G}_0, \mathbf{c})$.

With these partition changes in mind, and noting that

$$p(\mathbf{X}|\mathbf{G}_0, \mathbf{c}) = \prod_{\ell=1}^{K} p(\mathbf{X}_{\mathbf{z}^\ell}|\mathbf{G}_0, \mathbf{c}), \tag{15}$$

we can obtain the full conditional distribution for the Gibbs sampler as proportional to

$$p(c_i^*|\mathbf{c}_{-i}, \mathbf{X}, \mathbf{G}_0, \boldsymbol{\eta}) \propto \begin{cases} \alpha & \text{if } c_i^* = i \\ f(d_{i\ell}) & \text{if } c_i^* = \ell \text{ does not join two clusters} \\ f(d_{i\ell}) \frac{p(\mathbf{X}_{z^{k+\ell}}|\mathbf{G}_0, \mathbf{c}_{-i})}{p(\mathbf{X}_{z^\ell}|\mathbf{G}_0, \mathbf{c}_{-i}) \times p(\mathbf{X}_{\mathbf{z}^k}|\mathbf{G}_0, \mathbf{c}_{-i})} & \text{if } c_i^* = \ell \text{ joins clusters } k \text{ and } \ell. \end{cases} \tag{16}$$

Additionally, we use Metropolis–Hastings schemes to sample λ and α.

Even though the mean and precision of the normal components that represent the clusters have been analytically integrated out of the pCRP posterior distribution, it may be of interest to generate posterior

samples for these parameters. By exploiting the conjugate structure of the model, these parameters can be directly obtained from their posterior distribution given the cluster assignments $\mathbf{c}$ and hyperparameters $\mathbf{G}_0$ and antigenic locations $\mathbf{X}$. Thus, we can sample directly from

$$\begin{aligned} \boldsymbol{\Lambda}_k &\sim \mathcal{W}(\mathbf{T}_k, u_k) \\ \boldsymbol{\mu}_k &\sim \mathcal{MVN}\left(\mathbf{m}_k, (\kappa_k \boldsymbol{\Lambda}_k)^{-1}\right), \end{aligned} \tag{17}$$

where u_k, κ_k and $\mathbf{T}_k$ have been defined respectively in expressions (10) and (11). Additionally, we have $\mathbf{m}_k = \frac{\kappa_0 \mathbf{m}_0 + N_k \bar{\mathbf{X}}_{\mathbf{z}\mathbf{k}}}{\kappa_0 + N_k}$.

3. H1N1 influenza

We examine a collection of $N = 115$ H1N1 influenza viral strains along with $M = 77$ sera, curated in Bedford *et al.* [12]. The dataset encompasses 1882 cross-reactivity measurements and nucleotide sequence data from the HA gene for each of the 115 viruses. The isolates have a wide geographic distribution and were sampled between the years 1977 and 2009.

Figure 2(a) presents the maximum a posteriori estimate for the antigenic map, with viruses color coded according to antigenic cluster assignments. As expected, our analysis indicates a time directionality in the clusters, with older strains in the clusters on the left side of the antigenic map and younger strains at the right side of the plot. Antigenic groups of viral strains correspond relatively well to phylogenetic clustering. This can be seen in Figure 2(b) that shows the viral phylogenetic tree with tip annotations color coded according to the antigenic clusters of Figure 2(a). The observation that recent clusters have larger numbers of viruses is mainly a reflection of unequal temporal sampling.

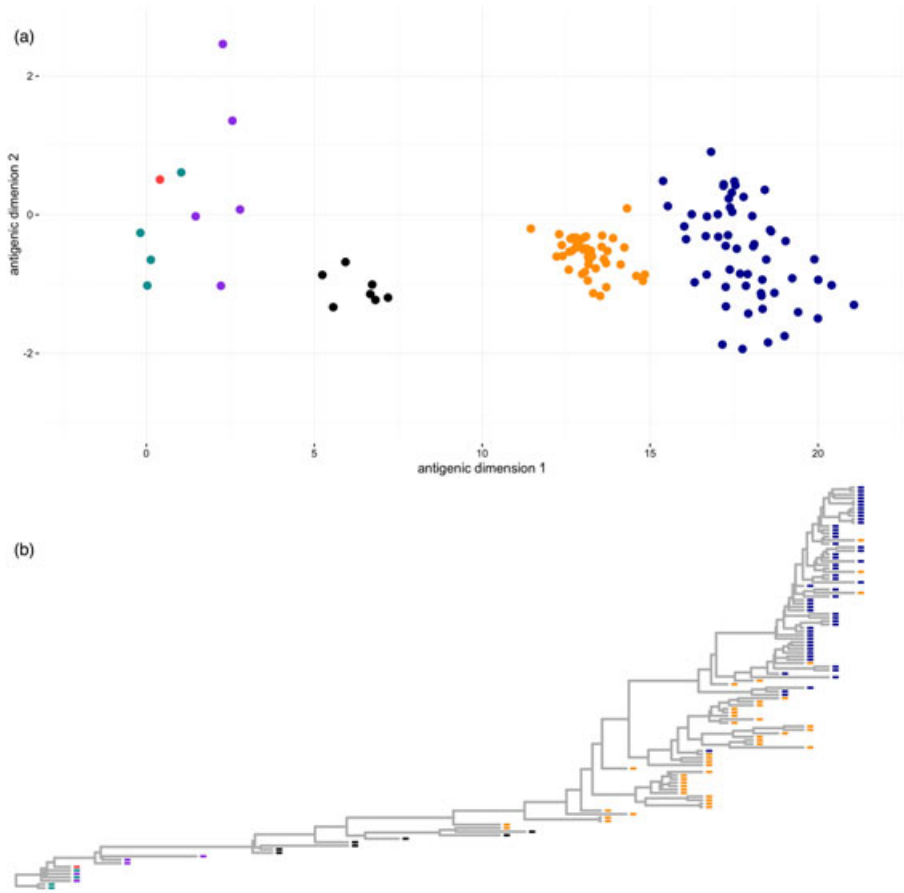

Figure 2. (a) Maximum a posteriori (MAP) estimates of virus locations $\mathbf{X}$ in the antigenic map. (b) Maximum clade credibility tree for H1N1 influenza viruses. In both figures, strains are color coded according to MAP cluster assignments.

For the time frame covered in the sample, we infer high posterior probability for the presence of four or five antigenic clusters (Figure 3(a)). For strains sampled after 1985, we clearly identify three antigenic clusters: The first, shown in black in Figure 2, contains strains sampled between 1986 and 1996; the second cluster, in orange, is composed of viruses ranging from 1995 to 2009; and the third antigenic cluster, in blue, has strains from 2006 to 2009. All strains in this period can be clearly associated with one of these antigenic clusters based on their posterior distribution of cluster assignments. The only notable exception is strain A/HongKong/1252/2000 that has similar posterior probabilities of being assigned to the orange and blue clusters (posterior probabilities between 0.48 and 0.54 of co-assignment to strains represented in blue or orange in Figure 2). The latter three antigenic clusters can be clearly seen in Figure 3(b) that presents a heatmap of posterior probabilities of cluster co-assignments. Here, strains are not ordered temporally, but arranged to highlight cluster associations.

The uncertainty on the number of antigenic clusters mainly reflects the mapping of viruses from 1977 to 1983. In this 7-year period, there is considerable antigenic variation that cannot be easily resolved into antigenic clusters (Figure 2). Although these strains are clearly distinct from later clusters, there is considerable uncertainty on whether they should all be grouped together (Figure 3(b)). A better sampling of this time period may help resolve the issue.

Figure 4 presents the posterior distribution of cluster means. This distribution has three concentrated peaks representing the means of the most recent antigenic clusters. It also presents a more diffuse peek

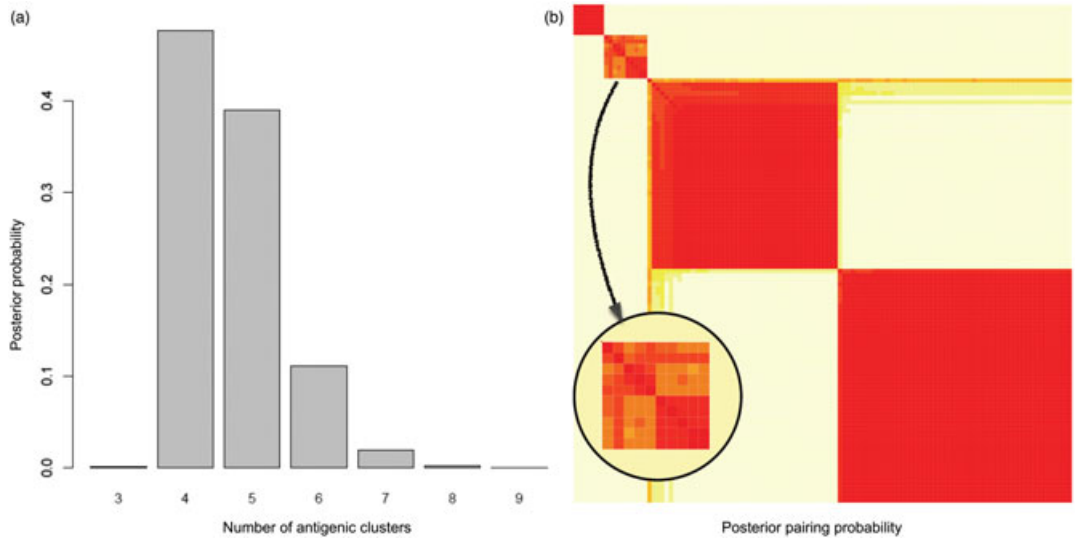

Figure 3. (a) Posterior distribution of the number of antigenic clusters. (b) Heatmap of the posterior probabilities of two viruses belonging to the same antigenic cluster. Red represents higher probabilities. Arrangement of the order of viruses is not chronologic but optimizes cluster visualization. Enhanced inset image shows internal structure of viruses from 1977 to 1983.

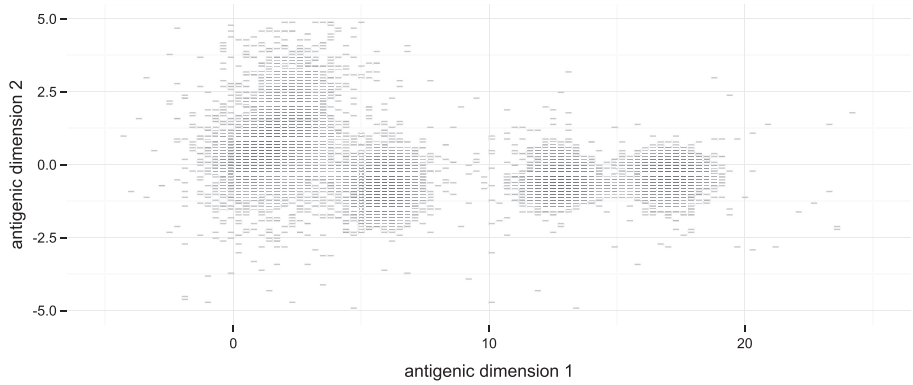

Figure 4. Posterior distribution of cluster means on antigenic map.

whose location corresponds to the viruses from 1977 to 1983. This last peak has a wider distribution, particularly along antigenic dimension 2, reflecting the clustering uncertainty in this period.

The drift parameter β of the prior distribution on serum locations can be seen as a measure of the overall linear change in antigenicity over time, because under this choice of prior, most of the antigenic change happens across antigenic dimension 1. We estimate a posterior mean of 0.511, with a 95% Bayesian credible interval of [0.469, 0.554] for β. These numbers are consistent with the findings of Bedford *et al.* [12] for their corresponding model. However, the rate of antigenic change is not constant through time, as is evident by the uneven spacing of clusters along antigenic dimension 1 (Figures 2 and 4).

We compare the posterior distribution of virus mappings for our model with that of the phylogenetic diffusion model [20] employed by Bedford *et al.* [12]. In their model, the prior distribution for the virus location $\mathbf{X}$ on the antigenic map is composed of a linear drift term and a continuous diffusion process along the phylogenetic tree. A supplementary video (https://phylogeography.github.io/videos/pCRP-comparison.pdf) contains dynamic representations of the posterior distribution of the antigenic map locations $\mathbf{X}$ for both models. While both models exhibit similar distributions along antigenic dimension 1 reflecting their corresponding drift priors, the overall aggregation of strains is quite different. The diffusion model generates a rather diffuse distribution of viruses on the map, while the pCRP produces antigenic maps in which viral strains have a tendency toward aggregation. This distinction is, without doubt, a consequence of modeling choices. We argue that better defined groupings on the antigenic map as well as explicit cluster association are important features of pCRP for the study of arising antigenic clusters.

Another important feature of the pCRP is the connection of antigenic clustering to molecular evolution. The tuning parameter λ of the decay function modulates the effect of the tree F on co-assignment probabilities. Our posterior mean estimate for λ was 0.3781, with a 95% Bayesian credible interval of [0.0184, 0.7559]. This implies that the probabilities of links between two viruses are less than inversely proportional to their phylogenetic distance. Consequently, the decay function has an attenuating effect on the large variability between phylogenetic distances induced by the trunk-like structure of the influenza phylogenetic tree. Finally, with greater than 95% posterior probability, we infer that $\lambda > 0$, highlighting the impact of phylogenetic distances on antigenic clustering.

4. Discussion

In this paper, we present a method for studying the interplay of antigenic and genetic evolution in influenza through the prism of antigenic clusters. We explicitly model the interaction between phylogenetics, antigenic clustering, and the multidimensional scaling. This allows us to jointly estimate the antigenic map and cluster assignments. We also demonstrate, in an application to H1N1 influenza, that phylogenetic relatedness is an important factor in antigenic clustering. Additionally, we show that our pCRP can lead not only to antigenic maps with better defined clusters but also to high confidence in cluster assignments for most viral strains.

The purpose of the multidimensional scaling method is to represent the structure of the HI titer data in a low-dimensional space to facilitate interpretation. Both Smith *et al.* [5] and Bedford *et al.* [12] have compared different space dimensionalities and found the two-dimensional plane to be optimal for their influenza data in terms of visualization and fit. For this reason, we develop our analysis with the two-dimensional antigenic map. Nevertheless, neither pCRP nor BMDS carry such a constraint and can easily accommodate other map dimensions. Further, the drift prior for serum locations $\mathbf{Y}$ on the antigenic map, although motivated by observations of increased antigenic distance over time [5], is important to address identifiability of map locations. This choice of prior yields a biologically interpretable rate parameter representing the overall antigenic change over time and has the effect of generating more linear antigenic maps. In Bedford *et al.* [12], inclusion of drift prior and a phylogenetic model not only improved identifiability of virus $\mathbf{X}$ and serum $\mathbf{Y}$ locations, it also improved model performance as measured through prediction errors.

Bayesian nonparametric modeling has found use in phylodynamics for estimating effective population sizes of ancestral populations [21, 22]. The nonparametric setting allows for a degree of flexibility for the shape of population size curve that would not be feasible with parametric forms. But nonparametric clustering methods have not yet been employed in phylodynamic studies of phenotypic traits such as antigenicity. Traditional antigenic cartography applications use K-means clustering algorithms to define the antigenic groups [5]. The pCRP, on the other hand, does not predefine the number of clusters, allowing

a potentially infinite number of clusters. This is a strength of the nonparametric clustering, and it is particularly relevant if we are interested in identifying the rise of new antigenic clusters.

One desirable property of Dirichlet processes is marginal invariance: The marginal distribution when one observation is removed is the same as the distribution of the process without that observation. The ddCRP, and consequently our model, are not marginally invariant [16]. It is still not clear what repercussion, if any, follows from this property. However, it could be said that the existence of a strain with a particular antigenic profile could alter the antigenic landscape and the selective pressure on other strains. Thus, clustering should be different if such a strain exists or not, independent of it being sampled. Yet the best we could expect is to have a representative sample of the antigenic landscape, because we could never observe all existing strains. The unbalanced temporal sampling of the H1N1 dataset gives some insight to the behavior of the pCRP regarding marginal invariance: The high sample density after 1996 does not lead to a larger number of clusters; instead, it reinforces the evidence for the existence of only two antigenic clusters in this period.

Antigenic cartography methods have been used to analyze other pathogens besides influenza, such as malaria and rabies. For these organisms, if the antigenic variability presents the cluster-like structure observed in influenza, our method could be instrumental to understanding the evolution of antigenicity and its relation to molecular evolution. More importantly, the fact that antigenic modeling is incorporated in the Bayesian phylogenetic context allows for joint estimation of the tree and the antigenic process. Demographic inference and geographic analysis are features already developed in this framework that can be jointly analyzed with antigenicity, leading to a more complete representation of the evolution of influenza both genotypically and phenotypically [23].

Through the posterior distribution of cluster assignments for an individual strain, we can assess the probability of recent viruses forming new antigenic clusters. Accurate detection of strains that are likely to seed new antigenic clusters is particularly useful for influenza surveillance and vaccine design. For this purpose, the effectiveness of our method in the definition and detection of new antigenic clusters should be further evaluated through prediction cross-validation strategies, as well as applications to different strains of influenza and larger datasets. Because of the nonparametric nature of this model, using larger datasets leads to significant increases in computational time and mixing problems, which could be partially addressed by the design of better proposal distributions for updating virus locations and pCRP links in the Markov chain Monte Carlo. In the H1N1 example, the high confidence in cluster assignments obtained through the pCRP is particularly encouraging, even though in this example recent strains were well nested in current antigenic clusters, and we have no indication of new cluster formation.

The methods presented in this paper have been implemented in the Bayesian phylogenetics software BEAST [17]. Additionally, the XML file containing the data and complete model specification for the H1N1 application is available at https://phylogeography.github.io/xml/H1N1.xml. For this analysis, we ran parallel chains up to 20,000,000 iterations, taking an average of 0.95 h per million states on a standard MacBook Pro with 2.3 GHz Intel Core i7 processor.

Appendix A

A. Normal–Wishart conjugate prior and marginal likelihood of X

For ease of notation, in this section, we drop the explicit dependency of all densities on hyperparameters $\mathbf{m}_0$, κ_0, $\mathbf{T}_0$, and u_0. Additionally, for this section, let $\mathbf{X}$ represent all the virus locations in the antigenic map for strains belonging to cluster k. Then, these locations are all generated by the same normal component, and their density is given by

$$p(\mathbf{X}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = (2\pi)^{-rN_k/2} \, |\boldsymbol{\Lambda}_k|^{N_k/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{N_k} (\mathbf{X}_i - \boldsymbol{\mu}_k)\boldsymbol{\Lambda}_k(\mathbf{X}_i - \boldsymbol{\mu}_k)' \right), \tag{A.1}$$

where r is the dimension of the antigenic map. All normal components share the same priors for the mean parameter $\boldsymbol{\mu}_k$ and precision matrix $\boldsymbol{\Lambda}_k$. We adopt the conjugate normal–Wishart prior, where

$$\begin{aligned} \boldsymbol{\Lambda}_k &\sim \mathcal{W}(u_0, \mathbf{T}_0) \text{ and} \\ \boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k &\sim \mathcal{MVN}(\mathbf{m}_0, (\kappa_0\boldsymbol{\Lambda}_k)^{-1}). \end{aligned} \tag{A.2}$$

Thus, the joint prior density can be expressed as

$$p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = P(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)p(\boldsymbol{\Lambda}_k) = \left(\frac{\kappa_0}{2\pi}\right)^{d/2} |\boldsymbol{\Lambda}_k|^{1/2} \exp\left(-\frac{\kappa_0}{2}(\boldsymbol{\mu}_k - \mathbf{m}_0)\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k - \mathbf{m}_0)^t\right) \times \frac{1}{Z_0}|\boldsymbol{\Lambda}_k|^{\frac{u_0-r-1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{T}_0^{-1}\boldsymbol{\Lambda}_k)\right), \tag{A.3}$$

where

$$Z_0 = 2^{\frac{u_0 r}{2}} |\mathbf{T}_0|^{\frac{u_0}{2}} \Gamma_r(u_0/2), \tag{A.4}$$

and $\Gamma_r(\cdot)$ is the multivariate gamma function

$$\Gamma_r(v) = \pi^{r(r-1)/4} \prod_{n=1}^{r} \Gamma(v + (1-n)/2). \tag{A.5}$$

In this conjugate model, the posterior distribution also assumes the form of a normal–Wishart distribution, and

$$\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k|\mathbf{X} \sim \mathcal{NW}(\mathbf{m}_k, \kappa_k, \mathbf{T}_k, u_k), \tag{A.6}$$

where $\mathcal{NW}(\cdot)$ represents the normal–Wishart distribution with density

$$p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k|\mathbf{X}) = \left(\frac{\kappa_k}{2\pi}\right)^{d/2} |\boldsymbol{\Lambda}_k|^{1/2} \exp\left(-\frac{\kappa_k}{2}(\boldsymbol{\mu}_k - \mathbf{m}_k)\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k - \mathbf{m}_k)^t\right) \times \frac{1}{Z_0}|\boldsymbol{\Lambda}_k|^{\frac{u_k-r-1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{T}_k^{-1}\boldsymbol{\Lambda}_k)\right). \tag{A.7}$$

Further,

$$Z_n = 2^{\frac{u_k r}{2}} |\mathbf{T}_k|^{\frac{u_k}{2}} \Gamma_r(u_k/2), \tag{A.8}$$

where $u_k = u_0 + N_k$, $\kappa_k = \kappa_0 + N_k$ and $\mathbf{T}_k^{-1} = \mathbf{T}_0^{-1} + \sum -N_k(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^t + \frac{\kappa_0 n}{\kappa_0+n}(\bar{\mathbf{X}} - \mathbf{m}_0)(\bar{\mathbf{X}} - \mathbf{m}_0)^t$ and $\mathbf{m}_k = \frac{\kappa_k \mathbf{m}_0 + N_k \bar{\mathbf{X}}}{\kappa_0 + N_k}$ [24]. The posterior precision for each cluster can be sampled from a $\mathcal{W}(\mathbf{T}_k, u_k)$ and $\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k, \mathbf{X} \sim \mathcal{MVN}(\mathbf{m}_k, (\kappa_k \boldsymbol{\Lambda}_k)^{-1})$.

We now compute the marginal likelihood of the data, integrating out the mean and precision parameters. Notice that

$$p(\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)}{p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k|\mathbf{X})}. \tag{A.9}$$

Combining expressions (A.1) and (A.7) and noting that $p(\mathbf{X})$ does not depend on $\boldsymbol{\mu}_k$ or $\boldsymbol{\Lambda}_k$, we can obtain the marginal likelihood as

$$p(\mathbf{X}) = \frac{(2\pi)^{-rN_k/2}\left(\frac{\kappa_0}{2\pi}\right)^{r/2}\frac{1}{Z_0}}{\left(\frac{\kappa_k}{2\pi}\right)^{r/2}\frac{1}{Z_n}} = \frac{\kappa_0^{r/2} Z_n}{(2\pi)^{rN_k/2}\kappa_k^{r/2} Z_0} = \frac{1}{\pi^{N_k r/2}} \frac{\Gamma_r(u_k/2)}{\Gamma_r(u_0/2)} \frac{|\mathbf{T}_k|^{u_k/2}}{|\mathbf{T}_0|^{u_0/2}} \left(\frac{\kappa_0}{\kappa_k}\right)^{r/2}. \tag{A.10}$$

When $r = 2$, this becomes the marginal likelihood of expression (9).

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 278433-PREDEMICS and ERC grant agreement no. 260864. This work was also supported by National Institutes of Health grants R01 AI107034 and R01 HG006139 and National Science Foundation grants DMS 1264153 and IIS 1251151. We thank Kenneth Lange, Christina Ramirez, and Jamie Lloyd-Smith for providing constructive feedback on an earlier version of this manuscript.

References

1. Stohr K. Influenza – WHO cares. *Lancet Infectious Diseases* 2002; **2**:517.
2. Molinari NAM, Ortega-Sanchez IR, Messonnier ML, Thompson WW, Wortley PM, Weintraub E, Bridges CB. The annual impact of seasonal influenza in the US: measuring disease burden and costs. *Vaccine* 2007; **25**(27):5086–5096.
3. Thompson WW, Comanor L, Shay DK. Epidemiology of seasonal influenza: use of surveillance data and statistical models to estimate the burden of disease. *Journal of Infectious Diseases* 2006; **194**(Supplement 2):S82–S91.
4. Hirst GK. Studies of antigenic differences among strains of influenza A by means of red cell agglutination. *Journal of Experimental Medicine* 1943; **78**(5):407–423.
5. Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus AD, Fouchier RA. Mapping the antigenic and genetic evolution of influenza virus. *Science* 2004; **305**(5682):371–376.
6. Cai Z, Zhang T, Wan XF. A computational framework for influenza antigenic cartography. *PLoS Computational Biology* 2010; **6**(10):e1000949.
7. Smith DJ, de Jong JC, Lapedes AS, Jones TC, Russell CA, Bestebroer TM, Rimmelzwaan GF, Osterhaus AD, Fouchier RA. Antigenic cartography of human and swine influenza A (H3N2) viruses. *Novel and Re-emerging Respiratory Viral Diseases* 2008; **699**:32–44.
8. Sandbulte MR, Westgeest KB, Gao J, Xu X, Klimov AI, Russell CA, Burke DF, Smith DJ, Fouchier RA, Eichelberger MC. Discordant antigenic drift of neuraminidase and hemagglutinin in H1N1 and H3N2 influenza viruses. *Proceedings of the National Academy of Sciences* 2011; **108**(51):20748–20753.
9. Fouchier RA, Smith DJ. Use of antigenic cartography in vaccine seed strain selection. *Avian Diseases* 2010; **54**(s1): 220–223.
10. Katzelnick LC, Fonville JM, Gromowski GD, Bustos Arriaga J, Green A, James SL, Lau L, Montoya M, Wang C, VanBlargan LA, Russell CA, Thu HM, Pierson TC, Buchy P, Aaskov JG, Muñoz-Jordán JL, Vasilakis N, Gibbons RV, Tesh RB, Osterhaus AD, Fouchier RA, Durbin A, Simmons CP, Holmes EC, Harris E, Whitehead SS, Smith DJ. Dengue viruses cluster antigenically but not as discrete serotypes. *Science* 2015; **349**(6254):1338–1343.
11. Koel BF, Burke DF, Bestebroer TM, van der Vliet S, Zondag GC, Vervaet G, Skepner E, Lewis NS, Spronken MI, Russell CA, Eropkin MY, Hurt AC, Barr IG, de Jong JC, Rimmelzwaan GF, Osterhaus AD, Fouchier RA, Smith DJ. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science* 2013; **342**(6161):976–979.
12. Bedford T, Suchard MA, Lemey P, Dudas G, Gregory V, Hay AJ, McCauley JW, Russell CA, Smith DJ, Rambaut A. Integrating influenza antigenic dynamics with molecular evolution. *eLife* 2014; **3**:e01914.
13. Antoniak CE. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics* 1974; **2**(6):1152–1174.
14. Miller KT, Griffiths T, Jordan MI. The phylogenetic Indian buffet process: a non-exchangeable nonparametric prior for latent features. In *the Proceedings of the Twenty-Fourth Annual Conference on Uncertainty in Artificial Inteligence (UAI-80)*. AUI Press: Corvallis, Oregon, 2008; 403–410.
15. Dahl DB. Distance-based probability distribution for set partitions with applications to Bayesian nonparametrics. *JSM Proceedings. Section on Bayesian Statistical Science, American Statistical Association*, Denver, Colorado, 2008.
16. Blei DM, Frazier PI. Distance dependent Chinese restaurant processes. *The Journal of Machine Learning Research* 2011; **12**:2461–2488.
17. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* 2012; **29**(8):1969–1973.
18. Oh MS, Raftery AE. Bayesian multidimensional scaling and choice of dimension. *Journal of the American Statistical Association* 2001; **96**(455):1031–1044.
19. Blackwell D, MacQueen JB. Ferguson distributions via Pólya urn schemes. *Annals of Statistics* 1973:353–355.
20. Lemey P, Rambaut A, Welch JJ, Suchard MA. Phylogeography takes a relaxed random walk in continuous space and time. *Molecular biology and evolution* 2010; **27**(8):1877–1885.
21. Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Molecular Biology and Evolution* 2013; **30**(3):713–724.
22. Palacios JA, Minin VN. Gaussian process-based Bayesian nonparametric inference of population size trajectories from gene genealogies. *Biometrics* 2013; **69**(1):8–18.
23. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. *PLoS Computational Biology* 2009; **5**(9):e1000520.
24. DeGroot MH. *Optimal Statistical Decisions*, Vol. 82. Wiley-Interscience, 2005.