

# Predicting the Emergence of H3N2 Influenza Viruses Reveals Contrasted Modes of Evolution of HA and NA Antigens

Reatha Sandie · Stéphane Aris-Brosou

Received: 24 May 2013 / Accepted: 6 December 2013 / Published online: 17 December 2013  
© Springer Science+Business Media New York 2013

**Abstract** Vaccine design for rapidly changing viruses is based on empirical surveillance of strains circulating in a given season to assess those that will most likely spread during the next season. The choice of which strains to include in the vaccine is critical, as an erroneous decision can lead to a nonimmunized human population that will then be at risk in the face of an epidemic or, worse, a pandemic. Here, we present the first steps toward a very general phylogenetic approach to predict the emergence of novel viruses. Our genomic model builds upon natural features of viral evolution such as selection and recombination / reassortment, and incorporates episodic bursts of evolution and or of recombination. As a proof-of-concept, we assess the performance of this model in a retrospective study, focusing: (i) on the emergence of an unexpected H3N2 influenza strain in 2007, and (ii) on a longitudinal design. Based on the analysis of hemagglutinin (HA) and neuraminidase (NA) genes, our results show a lack of predictive power in both experimental designs, but shed light on the mode of evolution of these two antigens:

(i) supporting the lack of significance of recombination in the evolution of this influenza virus, and (ii) showing that HA evolves episodically while NA changes gradually.

**Keywords** Bayesian inference · Recombination · Punctuated evolution · Viral evolution · Influenza H3N2

## Introduction

One of the reasons why viruses are so prone to causing epidemics stems from their high genetic diversity which, in the case of influenza A viruses (IAVs), is in part due to their high mutation rate and to their segmented genome, comprised of eight negative single stranded RNA molecules. The ten to twelve proteins encoded across these segments (Wise et al. 2009) play different roles in the life cycle of the virus and thus undergo different selective pressures. The two most studied proteins are the hemagglutinin (HA) and the neuraminidase (NA), responsible for host cell recognition and entry, and for facilitating virus release from infected cells, respectively (*e.g.*, Neumann et al. 2009). The antigenic properties of these two cell surface antigens are used to name and classify IAV subtypes, which can form almost all possible combinations between the 17 known subtypes of HA and the ten known for NA (Tong et al. 2012). While most of these subtypes have been observed in wild waterfowl, the most prevalent subtypes in the human population are H1N1 and H3N2, with H3N2 dominating human epidemics since its emergence in 1968 (Rambaut et al. 2008; Guan et al. 2010).

Each of the eight IAV segments can be exchanged between viruses coinfecting the same host, hereby leading to reassortant viruses. This process has the potential to change the antigenic properties of the virus in a dramatic

---

**Electronic supplementary material** The online version of this article (doi:10.1007/s00239-013-9608-6) contains supplementary material, which is available to authorized users.

---

R. Sandie  
Department of Biology, University of Ottawa, Ottawa,  
ON K1N 6N5, Canada

S. Aris-Brosou  
Department of Biology, Center for Advanced Research in  
Environmental Genomics, University of Ottawa, Ottawa, ON,  
Canada

S. Aris-Brosou (✉)  
Department of Mathematics and Statistics, University of Ottawa,  
Ottawa, ON K1N 6N5, Canada  
e-mail: sarisbro@uottawa.ca

way, leading to an “antigenic shift.” Such a process was at the origin of the 1968 H3N2 pandemic (Guan et al. 2010) or of the 2009 H1N1 pandemic (Smith et al. 2009), and can be experimentally shown to lead to the adaptation of swine and/or avian reassortant viruses to a mammalian host (Imai et al. 2012). In addition to reassortment, the genetic material of the virus also undergoes mutations which, in the absence of a proofreading mechanism, lead to substitution rates as high as  $\approx 5 \times 10^{-3}$  substitutions per site per year (Rambaut et al. 2008). Evidence, however, suggests that all subtypes do not evolve at the same rate. For instance, emergence of new H1N1 variants is often slow while H3N2 has often undergone rapid evolution and dissemination, as evidenced by the A/Sydney/5/97-like viruses that were detected in all parts of the world six months after their initial discovery (Hay et al. 2001).

In the face of these constantly evolving antigens, the human immune system detects the infecting virus and generates antibodies that contain current viruses and help prevent future infections. This immune machinery can easily recognize similar viral variants, but novel influenza strains that are antigenically different from their progenitors can trump the host immune system. Vaccines help boost the human immune system, but the rapid viral evolution described above demands that the composition of the influenza vaccine be updated every year. Currently, the trivalent vaccine targets two IAV subtypes (H3N2 and H1N1) and influenza B (Hay et al. 2001). Vaccine composition is reevaluated every year based on recommendations from the World Health Organization and other National Influenza Centers distributed around the world (see Hay et al. 2001)<sup>1</sup>. Candidate strains for vaccine composition are based on circulating viruses. Strain selection begins 8–10 months before the vaccine is available to the public, but the data used to determine candidate strains are inaccessible to the general public (Salzberg 2008) and are largely based on HA inhibition assays that often have poor resolution in distinguishing between strains (Plotkin et al. 2002).

While the strain-selection process can be effective and match as much as 91 % of the circulating viruses as in 2012–2013<sup>2</sup>, the selection process is imperfect. In 2007 for instance, a virulent H3N2 variant emerged in Australia and New Zealand a short time before the onset of the influenza season in the Southern hemisphere (April–September). To some extent, because the virulent strain was not the part of the 2007 vaccine, a widespread epidemic with a threefold increase in prevalence compared to regular seasons ensued (Owen et al. 2008). This then novel and highly infectious

strain, identified as Brisbane/10/2007, crossed the equator to North America just before the onset of the Northern hemisphere’s influenza season (November–March), eliciting a similar epidemic during the 2007–2008 season (Saks 2008).

The failure of the 2007 and 2007–2008 vaccine shows that there is a need for additional methods for determining which strain to include in the vaccine for each upcoming season. Computational methods have long been sought to predict the emergence of influenza viruses. An early method looked at nucleotide substitutions in codons of the HA gene undergoing positive selection and used a phylogenetic approach to determine which HA sequence was most likely to emerge (Bush et al. 1999). This type of *directional* evolution was later dismissed as it was determined that the evolution of HA genes tends to be more clustered than linear (Plotkin et al. 2002). The segmented structure of the influenza genome must also be taken into account when attempting to predict future influenza strains; as the rate of change is not constant across all segments (Holmes et al. 2005), epistatic interactions are likely to shape the virulence of a given virus (Neumann et al. 2009; Kryazhimskiy et al. 2011), and recombination / reassortment are key processes of the evolution of most viruses (*e.g.*, Holmes 2009, p. 48).

To address the current dearth of prediction tools for the emergence of novel viruses, we introduce a very general phylogenetic approach that takes both selection and recombination / reassortment into account. Because simulations only confirm that a model performs well in the absence of model misspecification, we put the model to test in the worst possible scenarios: (i) detecting the emergence of Brisbane/10/2007, and (ii) out of cluster prediction in a longitudinal study design. In each context, a retrospective analysis of HA and NA data sets shows that our model has a different but low predictive power for these two genes. We show that including punctual bursts of evolution in our model almost doubles predictive power for HA, but not for NA. In turn, this result suggests that the evolution of HA is more episodic than that of NA in H3N2 viruses.

## Methods

### Overview of the Model

The objective of the model is to generate a sample of sequences that have a high probability of emerging, given a set of observed sequences. Let us denote the observed sequences as  $(X_1, \dots, X_t) = X_{1:t}$ . If time  $t$  represents the current influenza season, then  $X_{1:t}$  represents a set of sequences sampled over the recent  $t$  seasons, and  $X_{t+1}$  represents the sequences sampled from the future season.

<sup>1</sup> <http://www.who.int/csr/disease/influenza/vaccinerecommendations/en/>

<sup>2</sup> <http://www.cdc.gov/flu/about/season/flu-season-2012-2013.htm>

The quantity of interest here is the *posterior predictive probability* of the data at season  $t + 1$ , given the data observed between the recent  $t$  seasons, or  $p(X_{t+1}|X_{1:t})$ . This quantity can be decomposed as:

$$p(X_{t+1}|X_{1:t}) = \int_{\Theta} p(X_{t+1}|\theta)p(\theta|X_{1:t})d\theta \tag{1}$$

where  $\theta$  is a vector of nuisance parameters, typically the branch lengths of the phylogenetic tree and the parameters of the model of evolution, and where  $\Theta$  denotes the state space of  $\theta$ . Equation (1) represents the sum (integral) over the product of two probability density functions:  $p(X_{t+1}|\theta)$ , the likelihood of  $\theta$  given the future data, and  $p(\theta|X_{1:t})$ , the posterior distribution of the nuisance parameters  $\theta$  given the observed data. According to Bayes’ theorem, this posterior distribution is proportional to the product of the likelihood of  $\theta$  given the sampled data,  $p(X_{1:t}|\theta)$ , and a prior on nuisance parameters  $p(\theta)$ :

$$p(\theta|X_{1:t}) = \frac{p(X_{1:t}|\theta)p(\theta)}{p(X_{1:t})} \tag{2}$$

The posterior predictive probability (Eq. 1), therefore, summarizes the information about the probability of new (emerging) sequences given the likelihood, the prior, a model of evolution and the observed data. However, the integration in Eq. (1) cannot be done analytically. Instead, we resorted to a two-step procedure where we first sample  $\theta$  from the posterior distribution as in Eq. (2), and then use these sampled  $\theta$  values to simulate future sequences  $X_{t+1}$  (e.g., Pagel and Meade 2006; Liu and Pearl 2007; Liu et al. 2008).

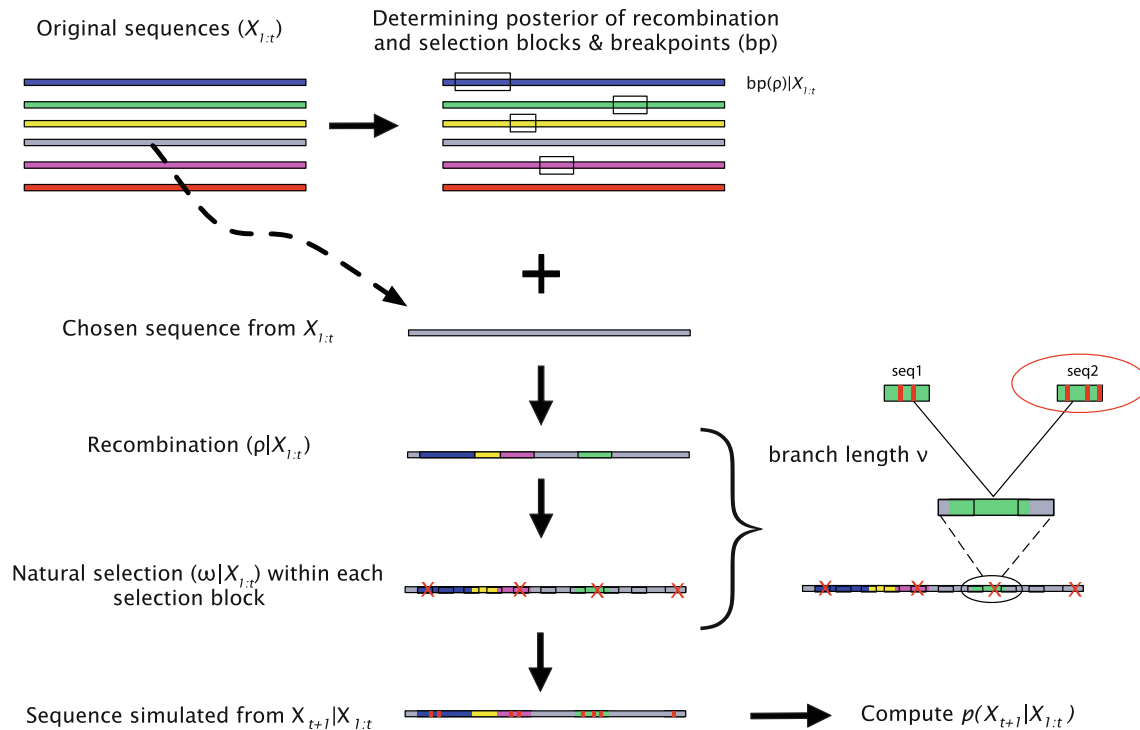
### Computational Details

In the first step (Fig. 1, top), the  $\theta$  values are drawn with the reversible-jump Markov chain Monte Carlo (rjMCMC) sampler implemented in OmegaMap ver. 0.5 (Wilson and McVean 2006). This model describes the evolution of codon data with selection and recombination under a standard coalescent prior (constant population size). The model has two parameters, collectively denoted as  $\theta$  in Eq. 1: a selection parameter  $\omega$ , which is the rate ratio of nonsynonymous to synonymous substitutions; and the population recombination rate  $\rho$ . Both can vary along the sequence by defining a block-like structure that segments an alignment of length  $L$  into at most  $L$  selection blocks and  $L-1$  recombination blocks. In both cases, the number of blocks is estimated from the data. The model was parameterized as follows. Prior distributions for  $\omega$  and  $\rho$  were set to have mean lengths of 20 and 74 codons, respectively, while in intensity these parameters were assumed to follow exponential priors centered on 1 for  $\omega(\exp(1))$  and 0.01 for

$\rho(\exp(1/10))$ . The model also includes nuisance parameters that are defined over the entire length of the alignment: the transition to transversion rate ratio  $\kappa \sim \exp(3)$ , the rate of synonymous transversion  $\mu \sim \exp(1/14)$ , and the insertion/deletion rate  $\phi \sim \exp(1/10)$ ; the specification of these priors followed Wilson and McVean (2006). Equilibrium codon frequencies were set to their empirical frequencies. The recombination model is asymmetric, as it assumes that one of the sampled sequences is a mosaic of the other sampled sequences; therefore, chains were run with ten random sequence orderings. Each sampler was run for  $10^7$  steps with a thinning of 100. Two independent runs were performed to check for convergence and to obtain the marginal distributions of  $\omega|X_{1:t}$ ,  $\rho|X_{1:t}$  as well as that of their respective block structures. Burn-in periods were empirically determined.

The second step (Fig. 1, below the “+” sign) performs the predictive simulation of future sequences  $X_{t+1}$  based on the  $\theta$  values sampled in the previous step. The procedure is initialized by estimating the average amount of evolution  $\bar{b}$  separating two sequences in  $X_{1:t}$ ; by so doing, we assume that the evolutionary process is on average time-homogeneous over the entire time window considered. Maximum likelihood pairwise branch length estimation is performed under the one-ratio codon model (Goldman and Yang 1994) with `codeml` (Yang 2007). Simulation of a sequence from  $p(X_{t+1}|X_{1:t})$  proceeds in two steps. First, a recombinant sequence is generated according to the recombination block structure sampled from  $p(\theta|X_{1:t})$ . For that purpose, a “master” sequence is first drawn at random; this draw is limited to the most recent sequences in  $X_{1:t}$ , i.e., those collected during the current season  $t$ . The positions of the recombination blocks are extracted from the output of OmegaMap. For each of these blocks, a corresponding block is drawn with probability  $[\rho|X_{1:t}]$  from one sequence taken at random with replacement from the most recent sequences in  $X_{1:t}$ . The blocks thus sampled are concatenated to form the recombinant sequence  $X_{t+1}^{\rho}$ . This recombinant sequence is then evolved following the block structure of the selection (codon) process, as sampled from  $p(\theta|X_{1:t})$ . Indels are first replaced by a random nucleotide (in practice, adenines) to give  $X_{t+1}^{\rho \text{ indels}}$ . To reproduce the among-site variation in  $\omega$  implemented in OmegaMap, each  $\omega|X_{1:t}$  block of  $X_{t+1}^{\rho \text{ indels}}$  is used as the root of a simulated two-sequence tree ( $seq_1 : \bar{b}, seq_2 : \bar{b}$ ) under the one-ratio codon model parameterized with  $(\omega|X_{1:t}, \kappa|X_{1:t})$ . One of these sequences is drawn at random to form the final sequence; indels are repositioned in this simulated sequence  $X_{t+1}^{\rho \text{ indels}}$  to give  $X_{t+1}$ . This process is repeated 100 times for each of the  $\theta$  values drawn from  $p(\theta|X_{1:t})$ .

Finally, the likelihood of the alignment that includes the simulated sequence is computed. To speed computations up, only the selection block structure was taken into



**Fig. 1** Workflow of the algorithm used to draw sequences from the posterior predictive distribution. The first step of the algorithm computes the posterior distribution  $p(\theta|X_{1:t})$  with the rjMCMC sampler implemented in OmegaMap. The second step of the algorithm samples from  $p(X_{t+1}|X_{1:t})$  by first drawing a sequence from  $X_{1:t}$ ; this sequence is used to generate a recombinant sequence and is

then evolved by an amount  $v$  under a codon model; this simulation step gives rise to two sequences, seq1 and seq2; one of these two sequences is drawn at random (red oval) to generate  $X_{t+1}$ . The posterior predictive probability is then computed as described in the text (Color figure online)

account (the recombination block structure was ignored). For each  $\omega|X_{1:t}$  block drawn by the rjMCMC sampler, a matrix of maximum likelihood pairwise distances is first estimated under the one-ratio codon model (Goldman and Yang 1994), still using `codeml`. This matrix is used to obtain an approximate tree for this block by weighted Neighbor-Joining as implemented in `weighor` (Bruno et al. 2000). Negative branch lengths are set to zero to avoid computational problems. A maximum likelihood tree could also be obtained for greater accuracy, for instance using `codeml`, but this approach is expected to increase the computational burden. The log-likelihood of each block is computed with `codeml` by reusing the parameters drawn from the posterior distribution ( $\omega|X_{1:t}$  and  $\kappa|X_{1:t}$ ) and the `weighor` branch lengths. The log-likelihood of the predicted alignment is obtained by summing the log-likelihood values over the selection blocks.

Computations involved in the last two steps are easily distributed, either on a shared memory / multicore computer or on a computer cluster. Therefore, they are typically quick to perform (of the order of a few days for the data analyzed below after parallelization of the algorithm on a cluster). The main computational bottleneck is in the first step, when samples are drawn from the posterior

distribution (of the order of a few weeks for the same data even on a large shared memory computer).

Episodic bursts of evolution or of recombination were incorporated into the model as follows. In the base model described above (Fig. 1), the simulated sequences are evolved on a two-sequence tree in which the branch lengths are both set to the average branch length  $\bar{b}|X_{1:t}$  within each selection block, while recombination follows the sampled  $\rho|X_{1:t}$  within each recombination block. Episodic bursts of evolution are then emulated by multiplying  $\bar{b}|X_{1:t}$  by a scaling factor denoted  $v$ , while episodic bursts of recombination are generated by multiplying  $\rho|X_{1:t}$  by a scaling factor  $\rho$ .

#### Identification of the Simulated Sequences

The simulated sequences were then used as queries in BLASTn searches (Altschul et al. 1990) against a local copy of the influenza database<sup>3</sup>. As a result, it is possible to infer the identity (year and country of sampling, subtype and accession number) of the most similar sequences present in the database, and check whether the algorithm is

<sup>3</sup> <ftp://ftp.ncbi.nih.gov/genomes/INFLUENZA>

capable of sampling sequences from the future with a high probability.

### Sequence Data for the Retrospective Studies

Individual protein-coding sequences for both the HA and NA genes were downloaded from the influenza Virus Resource (Bao et al. 2008). Only unique, full-length sequences collected between 2002–2007 were used, resulting in 555 HA sequences and 498 NA sequences. As these sequences were not limited to a particular geographic area, they represent the worldwide diversity of sampled influenza viruses available during this entire period of time.

Two designs were used to assess predictive power: one analyzing the whole period (2002–2007), and one analyzing the data year by year (longitudinal analyses). In the first design, due to the relatively large size of these sequence alignments, we clustered sequences with at least 95 % similarity (Abdussamad and Aris-Brosou 2011), so that the size of each data set is reduced while maintaining most of the diversity found in each data set. A single sequence from each resultant cluster was randomly picked as the representative sequence for that group, save for the Brisbane/10/2007 strain, which was set to represent its own cluster. This subsampling of each data set resulted in alignments comprising 19 HA and 30 NA sequences, which represent most of the diversity found in the original pool of sequences. In the second design, 24 sequences were randomly sampled for each year for both HA and NA data sets. The algorithm was then run on each year  $X_t$  to predict sequences circulating during  $X_{t+1}$ ; note that each year in the Northern hemisphere overlaps with two seasons, with the majority of sequences coming from the second half of the first season.

In order to assess the predictive power of our model, we constructed Neighbor-Joining trees of the original and simulated sequences together. These trees were obtained using maximum likelihood pairwise distances estimated under the general-time reversible substitution model with among-site rate variation modeled with a discrete  $\Gamma$  distribution (e.g., Aris-Brosou and Rodrigue 2012). For each of the resulting trees, we computed the patristic distance between the simulated sequence and the target Brisbane/10/2007 sequence (both for HA and NA). If the model has good predictive power, then we expect that highly probable sequences will be very similar to the target sequence and hence show a significant relationship between the probability of the generated sequences and their distance to the target sequence. Predictive power was then quantified by computing the  $R^2$  value of the regression (proportion of the variance explained by the linear model). Equality of slopes was tested with an  $F$  test (Sokal and Rohlf 2011, p. 513). Trees in Fig. S3 were reconstructed by maximum likelihood using *fasttree* (Price et al. 2010) under the

GTR +  $\Gamma$  model of evolution; support values are based on the SH-like  $P$ -values from the approximate likelihood ratio test (Anisimova and Gascuel 2006).

## Results

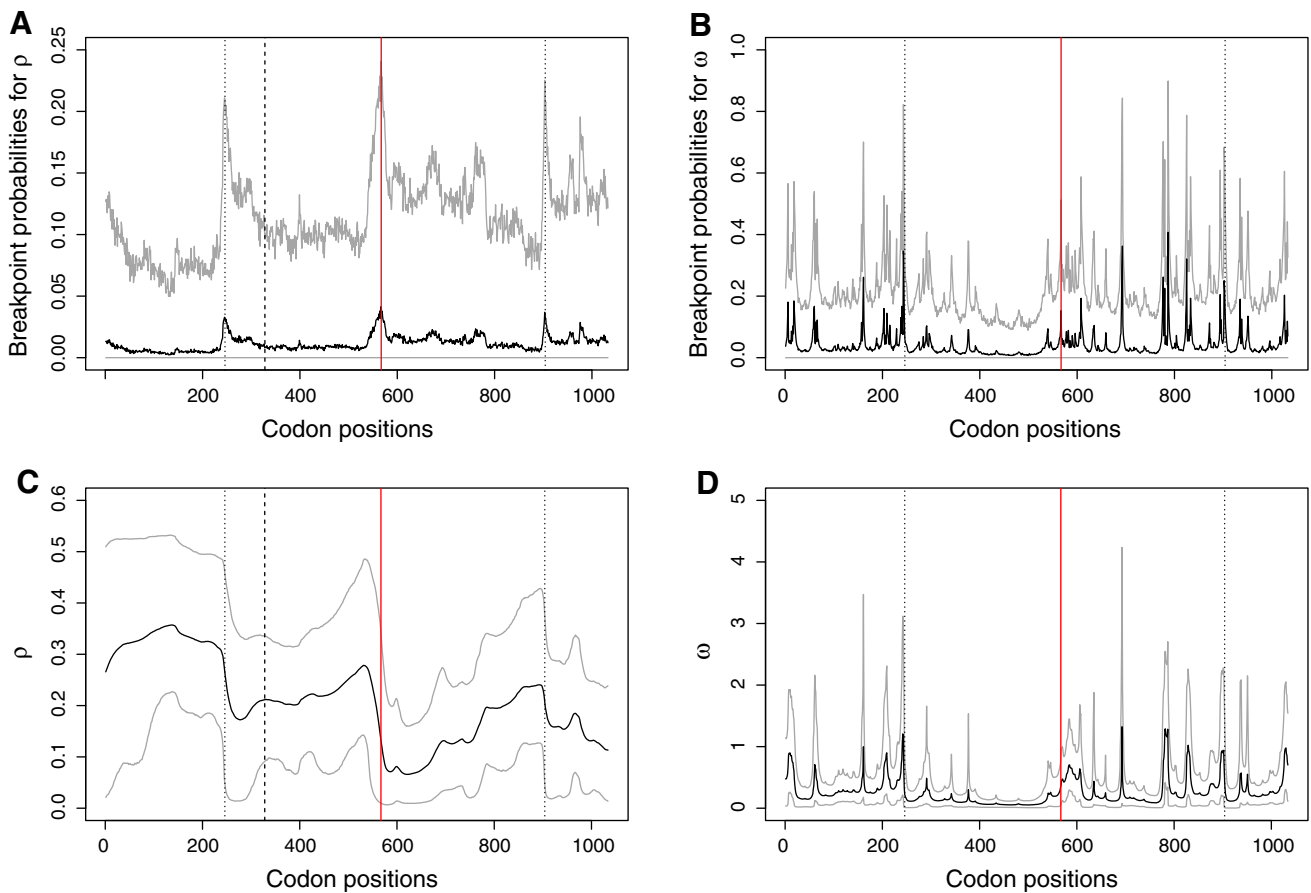
### Prediction Under the Base Model

Full-length HA and NA sequences were extracted from NCBI. In the first retrospective study, the current sampling period was set to cover the 6-year period spanning 2002–2007. The objective was to test if the emergence of the Brisbane/10/2007 strain, our *target sequences*, can be predicted from first principles of molecular evolution, involving selection and recombination / reassortment.

The reduced HA and NA data sets (after clustering) were first analyzed together to demonstrate the possibility of identifying “breakpoints” in concatenated data sets. The results show that the algorithm is able to recover the concatenation point between the two genes as the most probable breakpoint (at codon position 567 in Fig. 2a). Two other codon positions have high breakpoint probabilities (dotted lines in Fig. 2a); they do not correspond to subunit limits: in HA for instance, the limit between HA1 (encoding the globular part of the antigen) and HA2 (encoding the transmembrane domain) is at position 328 (dashed line in Fig. 2a). However, these two peaks of breakpoint probabilities correspond to positions of elevated  $\omega$  rate ratios (Fig. 2b, d), which is due to the confounding signals of recombination and selection (Anisimova et al. 2003).

To alleviate the computational cost of the algorithm, from here on, we analyzed the HA and NA genes independently of each other, so that  $\rho$  represents intragenic recombination and not reassortment or a mixture of these two processes. Output from the first step of the computation (Fig. 1) showed evidence of varying selective pressures (Fig. S1a, b) and recombination levels (Fig. S1c, d) across the entire length of both the HA and NA alignments (see also Fig. 2c, d). In the second step of our model (Fig. 1), these posterior distributions of  $\omega$  and  $\rho$  were used to generate gene sequences drawn from their target posterior predictive distribution. A large proportion of the simulated sequences, for both data sets, was identified with BLASTn to be from the 2002–2007 time frame, our “current sampling period” (Fig. S2). The model was able to simulate sequences BLASTn-identified as the Brisbane/10/2007 strain for the NA data set (521/16,317 = 3.2 % in one particular run of the MCMC sampler), but not for HA. In addition, no simulated sequences were BLASTn-identified as coming from 2008 or later with a high posterior predictive probability (in the





**Fig. 2** Concatenated analysis of HA and NA gene segments. All panels represent posterior estimates along the concatenated sequence of breakpoint probabilities for recombination blocks (a), selection blocks (b), posterior estimates of recombination (c), and selection (d). Gray lines indicate the 5 and 95 % percentiles of the posterior means (in black). HA run from codon position 1–567 (vertical red line),

while NA runs from position 568–1035. Dotted vertical lines mark the peaks of recombination breakpoint probabilities within each gene at position 246 for HA and 336 (904–568) for NA; for HA, this position differs from the HA1/HA2 boundary (vertical dashed line) (Color figure online)

top 5 % of the distribution) using either the HA or NA data set (Fig. 3). On the other side of the prediction spectrum, both the HA and NA analyses contained simulated sequences that were already circulating well before 2002, illustrating the wide range of diversity simulated by the model, as well as the potentially long persistence time of viral sequences. Note that, this persistence of circulating sequences might be less pronounced for the HA gene (Fig. 3a) than for the NA gene (Fig. 3b). Results were robust to the inclusion of the target sequences in the “current sampling period” (Fig. S2) for HA (distribution of top-scoring simulated sequences over the years:  $\chi^2_{48} = 54$ ,  $P = 0.2559$ ; Fisher exact test:  $P = 1.000$ ) and for NA ( $\chi^2_{49} = 56$ ,  $P = 0.2289$ ; Fisher exact test:  $P = 1.0000$ ).

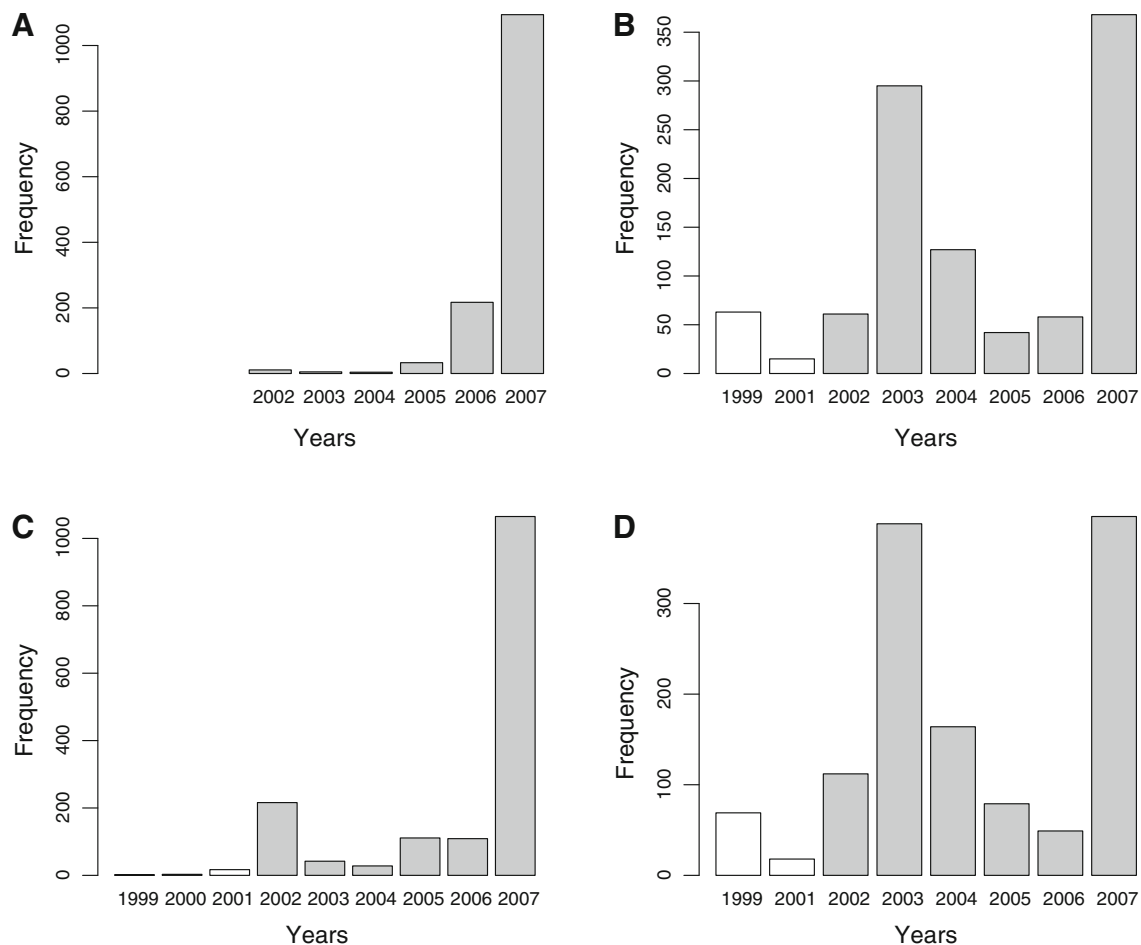
In order to quantify the predictive power of the model, we plotted the log-posterior probabilities of the simulated sequences against the patristic distances to the target strain (Fig. 4) and calculated the  $R^2$  value of the regression.

Table 1 shows that the base model has a predictive power of 26 % for HA and 18 % for NA.

#### Effect of Duration of Current Sampling Period

We then investigated the impact of the duration of the current sampling period. The hypothesis here was that the longer this time window, the higher the predictive power—assuming that the evolutionary process is stationary during the sampled time period, and that no multiple clades of H3N2 viruses were co-circulating. Alternatively, reducing the duration of the sampling period should decrease predictive power.

This hypothesis was evaluated by subsampling the original data sets of 555 (for HA) and 498 (for NA) sequences according to time. The original data were sampled from 2002 and 2007. A phylogenetic tree of the original 555 HA sequences highlighted multiple clusters of sequences that were sufficiently distant from the rest of the sequences to



**Fig. 3** Distribution of the BLASTn-identified sequences in the top 5 % of the posterior predictive distribution. Results are presented for data sets including the target sequence for (a) HA and (b) NA, and the data set excluding the target sequence for (c) HA and (d) NA. Shaded

bars represent sequences BLASTn-identified as coming from the “current sampling period” (2002–2007), while empty bars represent sequences coming from outside of this period

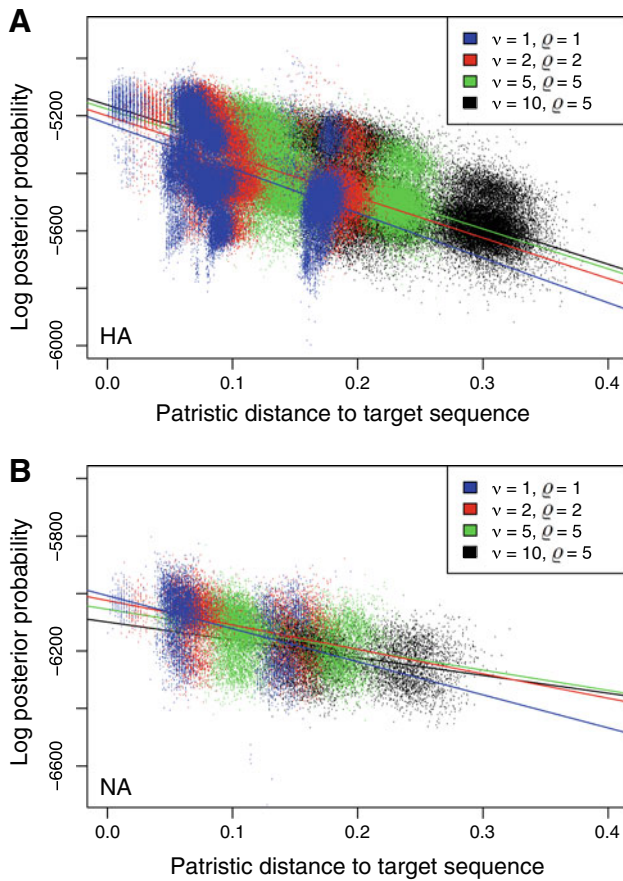
pose potential problems to the prediction model (Fig. S3). Closer inspection revealed that these clusters consisted of sequences circulating each year, suggesting that multiple evolutionary shifts might occur every year. Therefore, the most distant sequences, circulating in 2002 and 2003, were removed from both the HA and NA data sets, and the posterior predictive algorithm was then run on the remaining 2004–2007 sequences. This provided a 4-year time span of data, as opposed to the original 6-year span. Furthermore, a data set containing only the 2005 sequences was also extracted, as 2005 was the year that contained the largest number of sequences in the original data sets. As a result, we could compare the effectiveness of the predictive method across three sampling durations: 1 year (2005), 4 years (2004–2007), and 6 years (2002–2007). Note that, all three sampling durations aim at making out of cluster prediction.

The results show that the 4-year and 1-year analyses have higher probabilities than the 6-year study performed

above (Fig. S4 for HA), which is expected since the 6-year alignment is larger. More critically, the slope of the regression on the 4-year data set is smaller (in absolute value) than that of the full 6-year data set, which suggests a decrease in predictive power with smaller data sets (shorter sampling durations). Indeed, the  $R^2$  value for the 6-year data set stands at 0.37, and drops to 0.11 for the 4-year data set ( $P < 0.0001$ ) and to  $6.5 \times 10^{-4}$  for the 1-year data set ( $P < 0.0001$ ). Therefore, longer sampling durations improve the predictive (and out of cluster) power of our model. As a result, we only used the original 6-year data sets in the rest of this study.

#### Effect of Punctual Bursts of Evolution

So far, the model had a rather small predictive power, probably due to the structure of H3N2 circulation with each year forming its own cluster. However, cluster-shift or the emergence of “unexpected” strains might be due to



**Fig. 4** Patristic distances of both recombination rates and branch lengths. Log-posterior probabilities plotted against patristic distance between each simulated sequence and the target Brisbane/10/2007 sequences, both for HA (a) and NA (b)

**Table 1**  $R^2$  values for the linear regressions of log-posterior predictive probabilities against patristic distances to the target sequence

	$v = 1$	$v = 2$	$v = 5$	$v = 10$
HA				
$\rho = 1$	0.2646	0.3014	0.3812	0.3782
$\rho = 2$	0.2455	0.2901	0.3647	0.3712
$\rho = 5$	0.2327	0.2743	0.3490	0.3677
NA				
$\rho = 1$	0.1840	0.1281	0.1311	0.1324
$\rho = 2$	0.1620	0.1300	0.1231	0.1313
$\rho = 5$	0.1562	0.1173	0.1264	0.1308

episodes of accelerated evolution. To test this hypothesis, we incorporated punctual bursts of evolution and of recombination into the base model. Both the HA and NA data sets were used to test the effect of increasing the length of the branch leading to the predicted sequences by a factor  $v$ , hereby mimicking a punctual burst of evolution. By default, this length is set to the average branch length of

**Table 2** Slopes for the regressions of log-posterior predictive probabilities against patristic distances to the target sequence

	$v = 1$	$v = 2$	$v = 5$	$v = 10$
HA				
$\rho = 1$	-1557.39	-1442.62	-1445.97	-1419.97
$\rho = 2$	-1488.93	-1414.07	-1420.37	-1399.90
$\rho = 5$	-1437.57	-1371.95	-1388.85	-1385.06
NA				
$\rho = 1$	-1149.83	-854.64	-734.02	-622.88
$\rho = 2$	-1040.27	-849.57	-704.74	-616.66
$\rho = 5$	-1012.37	-793.47	-705.28	-618.56

the tree containing only the sequences from the current sampling period. This rate multiplier  $v$  was set to 1, 2, 5, and 10 for both data sets, HA and NA.

The results show a very significant negative relationship between posterior predictive probabilities and patristic distances for all  $v$  multipliers, both for HA (Fig. 4a) and NA (Fig. 4b). Our model is, therefore, able to predict sequences that have a relatively high probability. Table A.1 further shows that for HA, the average probability is increasing with  $v$  (see also Fig. 4a), while the slopes show a small but significant decrease in absolute value (Table A.1). The pattern is similar for NA, where the slopes are progressively decreasing (in absolute value) with  $v$ , but to a much larger extent (Table 2 and Table A.1). As a result, the inclusion of bursts of evolution in the model helps out of cluster prediction for HA but not NA sequences. Indeed, the  $R^2$  values of the regressions for HA increase to almost 40 % as  $v$  increases (Table 1). On the other hand, the inclusion of bursts of evolution makes our prediction of NA sequences worse, as  $R^2$  values decrease with increasing  $v$  (Table 1). This shows that the evolution of HA sequences during that period of time for the H3N2 subtype was characterized by episodic bursts of evolution (at least between 2002 and 2007), while the evolution of NA was more gradual.

#### Effect of Punctual Bursts of Recombination

Homologous (intrasgmental) recombination is generally considered to be insignificant in IAVs (Nelson and Holmes 2007; Boni et al. 2008). In order to assess the impact of recombination on the emergence of the target Brisbane/10/2007 strain under a different perspective, we incorporated a burst of recombination in our model. Branch length multipliers  $v$  were first kept constant and set to 1, while recombination rates along the branch leading to the simulated sequences were multiplied by a factor  $\rho$  that was varied from 1 to 5, a value of 5 meaning that recombination



rates leading to the predicted sequences were 5 times larger than those sampled from the rest of the tree.

Predictive power was assessed again by plotting log-posterior probabilities of the simulated sequences against patristic distances to the target strain. These regressions are highly significant for both the HA and the NA genes ( $P < 0.0001$ ; Fig. 4). However, increasing the recombination rate multiplier  $\rho$  essentially led to unchanged or even decreasing predictive power, both for HA and NA (Table 1). Therefore, our results confirm the general consensus that homologous (intrasegmental) recombination is not a significant process in the evolution of IAVs, at least in the case of the Brisbane/10/2007 strain.

### Joint Effect of Bursts of Evolution and of Recombination

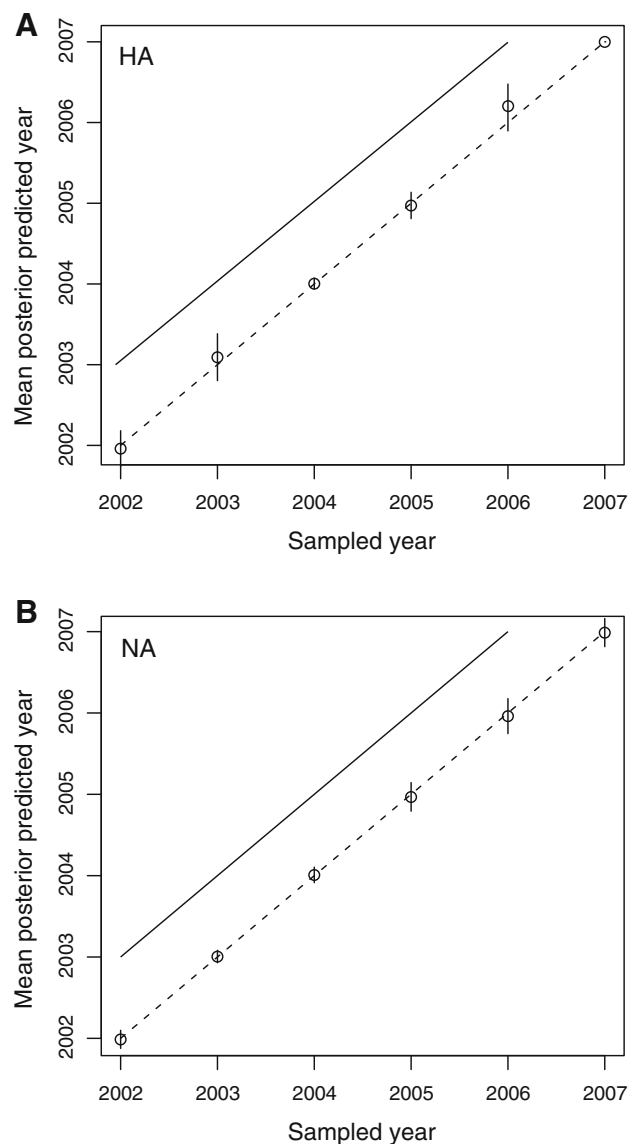
Despite the lack of evidence for any effect of recombination in the emergence of the Brisbane/10/2007 strain, we tested the hypothesis of a potential interaction between bursts of evolution and bursts of recombination. The model was then run on all combinations of multipliers for branch lengths ( $\nu$  set to 1, 2, 5, and 10) and recombination rates ( $\rho$  set to 1, 2, and 5). Phylogenetic trees were constructed as above for each of the  $4 \times 3 = 12$  possible combinations.

The computation of patristic distances for the different  $\nu$  and  $\rho$  combinations supported the pattern of increased sequence diversity both for HA (Fig. 4a) and NA (Fig. 4b). Consistently with the results found when varying  $\nu$  or  $\rho$  independently, the HA gene proved to be more responsive than NA to a joint increase in  $\nu$  and  $\rho$ , while the impact of bursts of recombination was negligible in both cases (Table 1). These results again support the hypothesis that the evolution of HA is mostly driven bursts of rates of evolution, and that (intrasegmental) recombination did not play any role in the evolution of these two genes.

### Longitudinal Analyses

In the preceding sections, we assessed predictive power with respect to one specific strain, Brisbane/10/2007. A more general way to assess predictive power is to monitor posterior predictive distributions longitudinally in time. We used a sliding window of width 1 year, from year  $y = 2002$  to  $y = 2007$ , to try and predict year  $y + 1$  in each case. Note that, in this longitudinal design, there is no “target sequence”: the goal is to be able to predict sequences that will be circulating in the upcoming year ( $y + 1$ ).

For computational reasons, we capped the number of sequences to 24 in each year. We then ran the complete algorithm on each of the 6 years, for both HA and NA. Given the results in Table 1, we used for our simulations two sets of rate multipliers: a first set with  $\nu = 1$  and



**Fig. 5** Distribution of BLASTn identified sequences in the longitudinal analyses. Analyses were performed on each individual year from 2002 to 2007 for (a) HA (2006 missing) and (b) NA. *Open circles* average predicted year in the top 1 % of the posterior predictive distribution. *Vertical bars* are 1 SD. *Solid line* expectation under the hypothesis that the analysis can predict emerging sequences in the following year. *Broken line* first diagonal

$\rho = 1$ , and a second set with  $\nu = 5$  and  $\rho = 1$ , which is the rate multipliers’ setting that corresponds to the largest posterior predictive power for HA.

The results are essentially the same for the two settings of rate multipliers, so we only present the case where  $\nu = 1$  and  $\rho = 1$ . The results are shown for HA (Fig. 5a) and for NA (Fig. 5b). If the predictive power was significant, then the average predicted year in the top 1 % of the log-posterior predictive distribution would be on or above the solid line. This was not the case, as for every year analyzed, the

average predicted year in the top 1 % of the posterior predictive distribution was not significantly different from the current year. This result is, however, in line with those obtained above (Fig. S4), where predictive power was inexistent when the “current sampling period” is reduced to a single year. Further testing of the algorithm should focus on determining the optimal size of this sliding window.

## Discussion

The original motivation behind the development of the model presented here was to be able to predict the emergence of influenza viruses: (i) in a timely manner, and (ii) accurately. The computational time required by our model in a pilot phase was, at three months for about 500 sequences, far greater than what can be desired in practice (compared to about 20 days with the reduced data, with samplers run 10 times longer). The burden was essentially caused by the first step, where posterior distributions are estimated with *omegaMap*. The second step, being amendable to parallelization (each posterior predictive simulation being carried out as an independent thread), does not stand as a serious computational bottleneck. Therefore, a sampling method was used to produce smaller HA and NA data sets, while still preserving most of the existing sequence diversity (Abdussamad and Aris-Brosou 2011). However, this sampling method intrinsically discards information relative to haplotype frequencies, which may be critical to help predict emerging viruses. Yet, because genetic diversity of influenza viruses, as measured by effective population sizes scaled to generation time, is thought to be low (Rambaut et al. 2008), it is more likely that nonadaptive processes play a key role in the emergence of influenza viruses. If this nonadaptive hypothesis is correct, then our filtering of the data to represent most of the available sequence diversity circulating in a region or worldwide might be an efficient method to predict emerging viruses.

While the computational burden was reduced, the accuracy of the model as a prediction tool for emerging viruses was not impressive. The sequences generated by the predictive model reveal in particular that the majority of the high-probability sequences were generated from 2002 and 2007 (Fig. 3). This suggests that H3N2 strains continue to circulate for several years after their emergence (Holmes et al. 2005; Plotkin et al. 2002). The best prediction under the base model, which did not incorporate any burst of evolution or of recombination, was obtained by analyzing a window of 6 years. By including punctual bursts of evolution, the predictive power of the model increased from 25 to 40 % for HA, but not for NA the

power of which remained low at 12–28 % (Table 1). Therefore, while the forte of the current approach may not lie in its predictive power, the analysis reveals two key features about the mode of evolution of HA and NA in H3N2 viruses: (i) none of these genes undergo recombination (Nelson and Holmes 2007; Boni et al. 2008); and (ii) the evolution of HA is episodic in H3N2 viruses, undergoing sporadic bursts of evolution, while NA evolves gradually. This confirms a recent report that took a more direct approach to estimate bursts of evolution (Westgeest et al. 2012), based on a codon model explicitly allowing selection to change episodically (Kosakovsky Pond et al. 2011).

Irrespective of these confirmatory results, the general low accuracy of the approach presented here highlights the difficulties in predicting the emergence of influenza viruses in two nonexclusive situations: (i) long-term predictions of, and (ii) out of cluster viruses. In the case of long-term predictions, the stationarity assumption of our model is likely to be violated by changes in the mode of evolution of viruses. For instance, pandemic H1N1 viruses showed in 2009 increased  $\omega$  rate ratios, which were interpreted as an increase in surveillance and/or adaptation to the new human host (Smith et al. 2009). In the second situation, H3N2 viruses form clusters of co-circulating strains every year (Fig. S3) (see also Plotkin et al. 2002; Nelson et al. 2006), so that our approach here attempts to perform out of cluster prediction based on multicluster information. In this light, it becomes clear that making such out of cluster prediction is difficult (Fig. 5) as the underlying evolutionary process is nonstationary.

Compared to previous approaches, either rooted in phylogenetics (Bush et al. 1999; Plotkin et al. 2002; Ferguson and Anderson 2002) or in machine learning (Xia et al. 2009; Trtica-Majnaric et al. 2010; Lees et al. 2010; Ito et al. 2011), the predictive model we described has several unique features such as the incorporation of a more realistic model of natural selection along with a model of recombination / reassortment in a Bayesian phylogenetic framework. Our model would therefore be directly applicable to predicting the emergence of viruses that undergo intragenic recombination such as retroviruses (*e.g.*, Holmes 2009, p. 50), or the evolution of viruses with segmented genomes like influenza by concatenating segments in a whole-genome analysis. However, with the present algorithm, this kind of analysis of long (genomic) sequences is computationally prohibitive. In addition to incorporating heterogeneity of the evolutionary process (Le et al. 2008) and the interplay between mutation and selection (Rodrigue et al. 2010), a more fruitful set of extensions would be either to consider antigenic determinants, either by using predictive tools as in Abdussamad and Aris-Brosou (2011) or by using grammar models (Loose et al. 2006), or to take

the ecology of the virus and the spatial patterns of its spread into account. Demographic models usually adopt a different formal structure, being based on systems of partial differential equations (Ferguson et al. 2005), and are, therefore, difficult to incorporate into genetic models. One notable exception attempted to reconcile the outputs of the two approaches (Ferguson et al. 2003), but the authors did not attempt to predict emerging strains. More recent forays into spatial studies addressed the surveillance issue from a phylogenetic point of view (Wallace et al. 2007; Parks et al. 2009; Janies et al. 2010; Cybis et al. 2013). Although these tools have the potential to predict *where* a particular virus is likely to emerge (Janies et al. 2010), they do not attempt to predict *which* viral strain is likely to emerge. Finally, the development of predictors of epidemics and pandemics would clearly benefit from the release of a public database linking influenza genomes to a proxy of their phenotype, such as the results of hemagglutination inhibition assays (Smith et al. 2004). In order to increase the predictive power of the model presented here, special efforts will probably be required to combine spatial and immunological models with genetic models, without forgetting demographic modeling as well as the population genetics of the virus of interest.

**Acknowledgements** This study was funded by the Natural Sciences Research Council of Canada and by the Canada Foundation for Innovation (SAB). We thank two anonymous reviewers for comments that helped to improve this article.

## References

- Abdussamad J, Aris-Brosou S (2011) The nonadaptive nature of the H1N1 2009 Swine Flu pandemic contrasts with the adaptive facilitation of transmission to a new host. *BMC Evol Biol* 11:6
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–10
- Anisimova M, Gascuel O (2006) Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst Biol* 55(4):539–52
- Anisimova M, Nielsen R, Yang Z (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* 164(3):1229–36
- Aris-Brosou S, Rodrigue N (2012) The essentials of computational molecular evolution. *Methods Mol Biol* 855:111–52
- Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, Ostell J, Lipman D (2008) The influenza virus resource at the National Center for Biotechnology Information. *J Virol* 82(2):596–601
- Boni MF, Zhou Y, Taubenberger JK, Holmes EC (2008) Homologous recombination is very rare or absent in human influenza A virus. *J Virol* 82(10):4807–11
- Bruno WJ, Succi ND, Halpern AL (2000) Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Mol Biol Evol* 17(1):189–97
- Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM (1999) Predicting the evolution of human influenza A. *Science* 286(5446):1921–5
- Cybis GB, Sinsheimer JS, Lemey P, Suchard MA (2013) Graph hierarchies for phylogeography. *Philos Trans R Soc Lond B Biol Sci* 368(1614):20120,206
- Ferguson NM, Anderson RM (2002) Predicting evolutionary change in the influenza A virus. *Nat Med* 8(6):562–3
- Ferguson NM, Galvani AP, Bush RM (2003) Ecological and immunological determinants of influenza evolution. *Nature* 422(6930):428–33
- Ferguson NM, Cummings DAT, Cauchemez S, Fraser C, Riley S, Meeyai A, Iamsirithaworn S, Burke DS (2005) Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature* 437(7056):209–14
- Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* 11(5):725–36
- Guan Y, Vijaykrishna D, Bahl J, Zhu H, Wang J, Smith GJD (2010) The emergence of pandemic influenza viruses. *Protein Cell* 1(1):9–13
- Hay AJ, Gregory V, Douglas AR, Lin YP (2001) The evolution of human influenza viruses. *Philos Trans R Soc Lond B Biol Sci* 356(1416):1861–70
- Holmes EC (2009) The evolution and emergence of RNA viruses. Oxford series in ecology and evolution Oxford University Press, Oxford
- Holmes EC, Ghedin E, Miller N, Taylor J, Bao Y, St George K, Grenfell BT, Salzberg SL, Fraser CM, Lipman DJ, Taubenberger JK (2005) Whole-genome analysis of an influenza H5 HA reveals multiple persistent lineages and reassortment among recent H3N2 viruses. *PLoS Biol* 3(9):e300
- Imai M, Watanabe T, Hatta M, Das SC, Ozawa M, Shinya K, Zhong G, Hanson A, Katsura H, Watanabe S, Li C, Kawakami E, Yamada S, Kiso M, Suzuki Y, Maher EA, Neumann G, Kawaoka Y (2012) Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets. *Nature* 486(7403):420–8
- Ito K, Igarashi M, Miyazaki Y, Murakami T, Iida S, Kida H, Takada A (2011) Gnarled-trunk evolutionary model of influenza A virus hemagglutinin. *PLoS One* 6(10):e25,953
- Janies DA, Treseder T, Alexandrov B, Habib F, Chen J, Ferreira R, Catalyürek U, Varón A, Wheeler WC (2010) The Supramap project: linking pathogen genomes with geography to fight emergent infectious diseases. *Cladistics* 26:1–6
- Kosakovsky Pond SL, Murrell B, Fourment M, Frost SDW, Delpont W, Scheffler K (2011) A random effects branch-site model for detecting episodic diversifying selection. *Mol Biol Evol* 28(11):3033–43
- Kryazhimskiy S, Dushoff J, Bazykin GA, Plotkin JB (2011) Prevalence of epistasis in the evolution of influenza A surface proteins. *PLoS Genet* 7(2):e1001,301
- Le SQ, Lartillot N, Gascuel O (2008) Phylogenetic mixture models for proteins. *Philos Trans R Soc Lond B Biol Sci* 363(1512):3965–76
- Lees WD, Moss DS, Shepherd AJ (2010) A computational analysis of the antigenic properties of haemagglutinin in influenza A H3N2. *Bioinformatics* 26(11):1403–8
- Liu L, Pearl DK (2007) Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst Biol* 56(3):504–14
- Liu L, Pearl DK, Brumfield RT, Edwards SV (2008) Estimating species trees using multiple-allele DNA sequence data. *Evolution* 62(8):2080–91
- Loose C, Jensen K, Rigoutsos I, Stephanopoulos G (2006) A linguistic model for the rational design of antimicrobial peptides. *Nature* 443(7113):867–9
- Nelson M, Simonsen L, Viboud C, Miller M, Taylor J, George K, Griesemer S, Ghedin E, Sengamalay N, Spiro D, Volkov I,

- Grenfell B, Lipman D, Taubenberger J, Holmes E (2006) Stochastic processes are key determinants of short-term evolution in influenza A virus. *PLoS Pathog* 2(12):e125
- Nelson MI, Holmes EC (2007) The evolution of epidemic influenza. *Nat Rev Genet* 8(3):196–205
- Neumann G, Noda T, Kawaoka Y (2009) Emergence and pandemic potential of swine-origin H1N1 influenza virus. *Nature* 459(7249):931–9
- Owen R, Barr IG, Pengilly A, Liu C, Paterson B, Kaczmarek M, National Influenza Surveillance Scheme (2008) Annual report of the national influenza surveillance scheme, 2007. *Commun Dis Intell* 32(2):208–26
- Pagel M, Meade A (2006) Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *Am Nat* 167(6):808–25
- Parks DH, Porter M, Churcher S, Wang S, Blouin C, Whalley J, Brooks S, Beiko RG (2009) GenGIS: A geospatial information system for genomic data. *Genome Res* 19(10):1896–904
- Plotkin JB, Dushoff J, Levin SA (2002) Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. *Proc Natl Acad Sci USA* 99(9):6263–8
- Price MN, Dehal PS, Arkin AP (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 5(3):e9490
- Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC (2008) The genomic and epidemiological dynamics of human influenza A virus. *Nature* 453(7195):615–9
- Rodrigue N, Philippe H, Lartillot N (2010) Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc Natl Acad Sci USA* 107(10):4629–34
- Saks M (2008) Was this a bad flu season or what?. *Emerg Med News* 30(7):14
- Salzberg S (2008) The contents of the syringe. *Nature* 454(7201):160–1
- Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus ADME, Fouchier RAM (2004) Mapping the antigenic and genetic evolution of influenza virus. *Science* 305(5682):371–6
- Smith GJD, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG, Ma SK, Cheung CL, Raghvani J, Bhatt S, Peiris JSM, Guan Y, Rambaut A (2009) Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* 459(7250):1122–5
- Sokal RR, Rohlf FJ (2011) *Biometry: the principles and practice of statistics in biological research*, 4th edn. W. H. Freeman and Co., New York
- Tong S, Li Y, Rivailler P, Conrardy C, Castillo DAA, Chen LM, Recuenco S, Ellison JA, Davis CT, York IA, Turmelle AS, Moran D, Rogers S, Shi M, Tao Y, Weil MR, Tang K, Rowe LA, Sammons S, Xu X, Frace M, Lindblade KA, Cox NJ, Anderson LJ, Rupprecht CE, Donis RO (2012) A distinct lineage of influenza A virus from bats. *Proc Natl Acad Sci USA* 109(11):4269–74
- Trtica-Majnaric L, Zekic-Susac M, Sarlija N, Vitale B (2010) Prediction of influenza vaccination outcome by neural networks and logistic regression. *J Biomed Inform* 43(5):774–81
- Wallace RG, Hodac H, Lathrop RH, Fitch WM (2007) A statistical phylogeography of influenza A H5N1. *Proc Natl Acad Sci USA* 104(11):4473–8
- Westgeest KB, de Graaf M, Fourment M, Bestebroer TM, van Beek R, Spronken MIJ, de Jong JC, Rimmelzwaan GF, Russell CA, Osterhaus ADME, Smith GJD, Smith DJ, Fouchier RAM (2012) Genetic evolution of the neuraminidase of influenza A (H3N2) viruses from 1968 to 2009 and its correspondence to haemagglutinin evolution. *J Gen Virol* 93(Pt 9):1996–2007
- Wilson DJ, McVean G (2006) Estimating diversifying selection and functional constraint in the presence of recombination. *Genetics* 172(3):1411–25
- Wise HM, Foeglein A, Sun J, Dalton RM, Patel S, Howard W, Anderson EC, Barclay WS, Digard P (2009) A complicated message: Identification of a novel PB1-related protein translated from influenza A virus segment 2 mRNA. *J Virol* 83(16):8021–31
- Xia Z, Jin G, Zhu J, Zhou R (2009) Using a mutual information-based site transition network to map the genetic evolution of influenza A/H3N2 virus. *Bioinformatics* 25(18):2309–17
- Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24(8):1586–91