Predicting the Evolution of Influenza A

by

Reatha Sandie

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
In partial fulfillment of the requirements
For the M.Sc. degree in
Biology, specialization Bioinformatics

Department of Biology
Faculty of Science
University of Ottawa

Abstract

Vaccination against the Influenza A virus (IAV) is often an important and critical task for much of the population, as IAV causes yearly epidemics, and can cause even deadlier pandemics. Designing the vaccine requires an understanding of the current major circulating strains of Influenza, as well as an understanding of how those strains could change over time to become either less harmful or more deadly, or simply die out completely. An error in the prediction process can lead to a non-immunized population at risk of epidemics, or even a pandemic. Presented here is a posterior predictive approach to generate emerging influenza strains based on a realistic genomic model that incorporates natural features of viral evolution such as selection and recombination. Also introduced is a sequence sampling scheme to relieve the computational burden of the posterior predictive analysis by clustering sequences based on their pairwise similarity. Finally, the impact of "evolutionary accidents" that take the form of bursts of evolution and or of recombination on the predictive power of our procedure is tested. An analysis of the impact of these bursts is carried out in a retrospective study that focuses on the unexpected emergence of a new H3N2 strain in the 2007-08 influenza season. Measuring the R^2 values of both pairwise and patristic distances, the model reaches a predictive power of $\sim$40%, but is not able to simulate the emergence of the target Brisbane/10/2007 sequence with a high probability. The inclusion of "evolutionary accidents" improved the algorithm's ability to predict HA sequences, but the prediction power of the NA gene remained low.

Résumé

La vaccination contre le virus de la grippe ou virus influenza A (IAV) est souvent une tâche importante et essentielle pour une bonne partie de la population, comme l'IAV provoque des épidémies annuelles, et peut causer des pandémies encore plus mortelles. La conception du vaccin nécessite une compréhension des souches circulantes de la grippe, ainsi que du comment ces souches pourraient changer au cours du temps pour devenir soit moins nocives ou plus mortelle, ou simplement disparaître complètement. Une erreur dans le processus de prédiction peut entraîner une population non immunisée vers une épidémie, voire même vers une pandémie. Nous présentons ici une approche prédictive *a posteriori* afin de générer des souches grippales émergentes. Notre approche est basée sur un modèle réaliste qui intègre les caractéristiques naturelles de l'évolution virale, comme la sélection et la recombinaison. Aussi nous introduisons un système d'échantillonnage des séquences pour diminuer la charge de calcul de l'analyse prédictive *a posteriori*, grâce à un regroupement des séquences basé sur leur similarité par paires. Enfin, l'impact des "accidents de l'évolution," prennant la forme de salves évolutives et ou de recombinaison, sur le pouvoir prédictif de notre procédure est testé. Une analyse de l'impact de ces salves est effectuée par d'une étude rétrospective qui se concentre sur l'émergence inattendue d'une nouvelle souche H3N2 de la grippe au cours de la saison 2007-08. En utilisant les valeurs de R^2 entre distances par paires et patristiques comme mesure de la qualité de nos prédictions, nous montrons que le modèle atteint une puissance prédictive de $\sim$40%, mais n'est pas capable de simuler l'apparition de la séquence cible Brisbane/10/2007 avec une forte probabilité. L'inclusion "d'accidents évolutifs" améliore la capacité de l'algorithme à prédire les séquences HA, mais le pouvoir prédictif pour le gène NA reste faible.

Acknowledgements

Great thanks goes to Dr Stéphane Aris-Brosou for all the time and help he gave me over the course of this project. I would also like to express my appreciation and thanks to my committee and examiners for all their time and effort: Dr Guy Drouin, Dr Xuhua Xia and Dr Root Gorelick. Various lab members have helped greatly over the course of this project, namely Dr Rob Carter, Dr Nicolas Rodrigue, Brady Tracey and Gareth Palidwor. Thanks also goes out to NSERC and the University of Ottawa for providing the funding necessary to support this project.

And lastly, I would like to acknowledge all the wonderful help and advice I received from another committee member, Dr George Carmody, who will be sadly missed.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Virus basics

Viruses are considered to be obligate intracellular parasites, due to their dependency on host cells to replicate. The life cycle of most viruses involve the infection of a cell, after which the virus then co-opts the cell's machinery in order to generate large numbers of virions for release and further infection. Most viruses have evolved unique strategies to infect and take control of a cell, including a preference for cell types. Influenza viruses invade a cell by way of the salicylic acid chain on the surface of cells found in the respiratory tract [1], while the majority of rhinoviruses (common cold) infect respiratory epithelial cells by way of the intracellular adhesion molecule-1 (ICAM-1)[2].

The most basic of viral structures consists solely of a viral capsid and a small genome that often solely codes for structural proteins, and is thus dependant on the target cell for reproduction machinery. The overall size of a virus can vary depending on the amount of viral genetic material and the complexity of the virus itself. The overall shape of the virus can be spherical (eg. hepadnaviridae), ovoid (eg. poxviridae) or elongated (eg. filoviridae) [3], just as the type of genetic material can change as the virus family changes. Virus

genomes are not required to be composed of DNA to target and infect a healthy cell. Over 70% of viruses are composed solely of RNA, rather than DNA, and the genomes can be found in many different configurations from double stranded (ds) to single stranded (ss), positive sense or negative sense, and linear, circular or segmented [3]. Typically, the larger virus species are more complex, and tend to be composed of dsDNA, whereas the smaller virus species tend to be less complex and contain either positive or negative sense ssRNA[4, 5]. RNA viruses can range from 3kb to 30kb in size[5], while DNA viruses such as the Mimivirus can be as large as 1.2Mb[6].

Compared to a living cell, viruses are comparatively small, ranging from the circovirus (17nm) to the poxvirus (300nm) and the worm-like filovirus (2500nm) [3]. Their small size can likely be attributed to the need to be smaller than the target cell, to ensure infection. Due to a virus' use of host machinery for reproduction purposes, many viruses have little to no self-contained replication genes or machinery.

Some viruses are relatively simple, such as the porcine circovirus, a ssDNA virus which contains 2kb of genetic material that codes for three genes, a capsid and two replicase proteins [7]. Others are more complex, including the herpes simplex virus, which has a complex assortment of glycoproteins on the viral surface and encompasses the genetic code for nearly 80 gene products in 150kb of dsDNA [8, 3] or the aforementioned Mimivirus which has 911 coding genes [6].

The host range of viruses also differ between virus species, and every branch of life has the capacity to be infected by one virus or another. Viruses themselves have the potential to be infected by a different virus, though so far only a single virophage has been discovered[9]. The ability of a virus to infect more than one species is not a rare trait, as even the most basic virus phylogeny will cross the species barrier [1]. However, a virus able to infect two or more species with the same ease of infection is much more

rare as the biology and biomolecules of the hosts differ. An excellent example of this is the Influenza H5N1 subtype which can infect both avian species and mammals [1, 10], though its ability to replicate efficiently in different species appears to be determined by a single amino acid in the PB2 gene [1]. Additionally, in humans, the hemagglutinin of Influenza, responsible for host cell entry, attaches itself to the α-2,6-sialic acid of lung epithelials, but in avian species the α-2,3-sialic acid found deeper in the respiratory tract is the glycoprotein of choice [1]. At this time, wild-type H5N1 has not made the leap from being a highly pathogenic avian influenza to a highly pathogenic human influenza. However, several research groups have made an effort to illustrate the ease with which the human variant could emerge, by engineering a highly transmissible H5N1 virus in a laboratory environment capable of infecting mammal species, and illustrating how few mutations would be needed for the shift from avian to mammalian virus. [11].

1.2 Influenza A and its importance in public health

Part of the *Orthomyxoviridae* family, influenza viruses are known for their high mutation rates and the ability to infect a variety of mammalian species, including humans. Influenza viruses encompass three of the five genera of the *Orthomyxoviridae*, named Influenza A, B and C, and all have unique characteristics differentiating them from each other.

The Influenza A virus (IAV) is the most well-known of all influenza species, containing 12 proteins found across 8 separate negative, single-stranded RNA segments (Figure 1.1). The hemagglutinin (HA) gene [12], which is responsible for host cell recognition and cell entry [13, 14], and the neuraminidase (NA) protein [12], which allows the release of the newly constructed virus particles from the host cell [13], are cell surface glycoproteins, and are two of the more well known genes found in IAV. The largest of the ssRNA

segments code for the proteins involved in the polymerase complex, PB2, PB1 and PA, and all are involved in the replication of the genomic viral RNA (vRNA) [15, 14]. PB1 is the only known IAV gene with alternate reading frames, which code for the protein PB1-F2 and its truncated version labelled N40, which while nonessential, have been found to negatively affect virus replication when absent [16]. The Influenza virus also contains genetic material that codes for the structural proteins M1(matrix protein), M2 (membrane protein that forms ion channels), as well as a nucleoprotein (NP) that binds to the viral RNA [14]. In addition, Influenza has two non-structural proteins. NS1 is the smallest of all the genes, and is involved in the transport, splicing and translation of host cell RNA, while NS2 or NEP (nuclear export protein) has been shown to export small viral proteins from the cell, by interacting with a human chromosome region maintenance protein (CRM1) [13].

Influenza B and C are similar to Influenza A in either structure or genome arrangement, but not identical. Influenza B virus (IBV) is structurally almost identical to Influenza A, but differ in gene number and function. Specifically, NB and BM2 in IBV functionally replace M2 in Influenza A [17]. The Influenza C virus (ICV), in comparison, is easily distinguishable from either A or B by an extensive clumping of virion particles on the surface of infected cells [17], and while ICV is compositionally similar, it contains only a single surface glycoprotein, the hemmagglutinin-esterase-fusion protein (HEF), which reduces the number of genomic segments in ICV compared to IAV and IBV [17]. There is also a notable decrease in mutation rate between IAV and IBV [18], as well as species specificity, despite being similar in both genome size and composition. Influenza B generally targets only humans [18] while Influenza C is rarely found in humans, preferring mammalian species such as pigs and dogs [19]. These differences in both prospective hosts and mutability/adaptation prevent the large-scale antigen shifting seen in Influenza

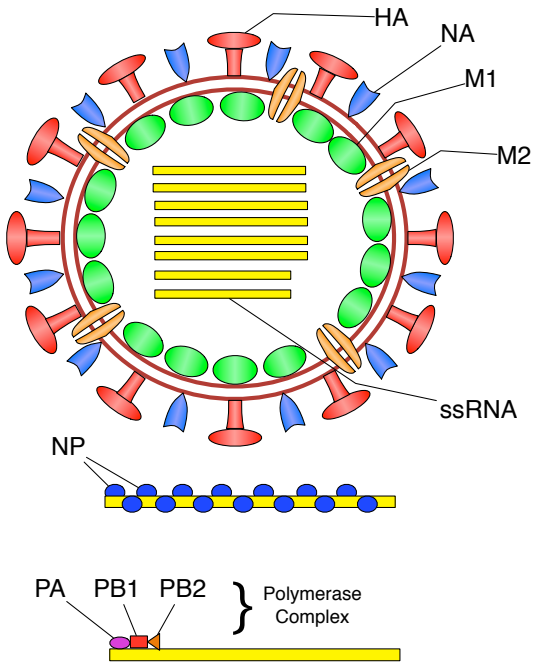

Figure 1.1: **Influenza A Virus.** Eight negative-sense single-stranded RNA segments encode all the proteins needed for the virus life cycle. The virus surface is covered with two antigens, hemagglutinin (HA) and neuraminidase (NA) glycoprotein, which are used for cell entry and cell lysis, respectively. PB1, PB2 and PA proteins form the polymerase complex used in viral RNA replication. The matrix protein (M1) retains the shape and structure of the virus capsid, while the M2 membrane protein opens up ion channels throughout the membrane. A nuclear protein (NP) functions to stabilize viral RNA. The non-structural protein NS1 is involved in the translation process of host RNA, while NS2 has a putative function as a nuclear export protein.

A (see Section 1.5), which is one of the major reasons Influenza A is such a hazard to human populations.

This high genetic diversity found in IAV is due in most part to different selective pressures on each of the RNA segments, with the HA and NA genes showing the highest mutation rate of all the genes [12]. As these are the antigens recognized by the immune system, the reason for IAV's wide subtype diversity is readily apparent. As of 2012, there are 17 distinct HA subtypes [20], and 9 NA subtypes [13], and are the proteins used to

identify Influenza A subtypes from one another (ie. H3N1, H4N2, etc...). Waterfowl are a natural reservoir for influenza subtypes, containing multitudes of antigenically distinct influenza viruses that can occasionally jump to other species, though are often unable to further transmit, such as is seen with previous H5N1 variants [14].

1.3 Global influenza pandemics

Three major human pandemics have broken out in the last century: the 1918 Spanish flu (H1N1 subtype) was estimated to have killed anywhere from 20-50 million people in 1918-1920 [14], the 1957 Asian flu (H2N2) and the 1968 Hong Kong flu (H3N2). Despite the prevalence of both Influenza A and Influenza B, which account for nearly all of human infections [21, 22], the majority of epidemics and all pandemics are due to Influenza A [23].

Humans are most often infected by either H3N2 or H1N1, as the H1, H3, N1 and N2 are the only subtypes currently capable of routinely infecting humans, though H3N2 is the most common of the subtypes to infect humans until 2009 [24]. Despite the prolific nature of the 2009 H1N1 swine virus [25], new variants of H1N1 often emerge at a much slower pace than new H3N2 strains [23]. An H3N2 strain in 1997 provides evidence of this rapid evolution and dispersal, as within 6 months of the initial discovery, A/Sydney/5/97-like viruses were detected in all parts of the world [23].

1.4 Evolution of Influenza A viruses

Mutation, natural selection, reassortment and potentially recombination are all mechanism in the evolution of influenza A, and impact the different segments to different degrees. Influenza's own polymerase complex has no proof-reading ability [26], which

allows for numerous mutations to be easily incorporated into the genome. When these point mutations occur in key antigenic sites, they can lead to a complete shift in the antigenicity of the virus [27], thus allowing the virus to temporarily evade the immune response of the host organism. When viral populations grow large, natural selection completely determines the fate of these mutations. Continual selective pressure on the virus generates distinct strains of influenza to which the human immune system has no immunity [28]. This accumulation of mutations at key antigenic sites on the HA and NA genes is referred to as antigenic drift, a type of evolution that occurs one mutation at a time. Antigenic shift, however, occurs when there is a massive rearrangement of DNA in a virus, from such processes as reassortment.

Reassortment occurs only in segmented virus genomes when a host cell is infected by more than one strain of the virus, and is the exchange of one or more entire segments, potentially creating a novel or "reassorted" virus [29]. Homologous or "intrasegmental" recombination in IAV is a process that can occur when RNA from another source, such as a second influenza virus, is present in the infected cell, and results in the exchange of small sections of genetic material [29]. Such is what is believed to have occurred just prior to the emergence of the swine-derived H1N1 in 2009, the triple-reassortment of viruses capable of infecting human, swine and birds allowed for the emergence of a highly virulent influenza virus [30]. Reassortment and recombination are not limited to influenza, though it is more common in RNA viruses than with other virus species [3]. Reassortment is limited to multisegmented viruses, while recombination can occur in viruses that are either segmented or not [29].

The ability of an Influenza virus to recombine is under considerable debate. Some studies suggest that recombination has occurred in human and swine Influenza A viruses [31] and in the case of the 1918 Spanish Flu HA gene [32]. Yet, the validity of these

conclusions has been called into question [33], and the general consensus holds that recombination is either absent or very rare in Influenza A [34]. A study by Boni *et al.* [34] assessed the possibility of recombination occurring by testing all possible two-break point recombinations in a set of nearly 14,000 HA sequences. This was done with an algorithm which uses a three-sequence combination called triplets to determine whether any two sequences form the third through homologous recombination, called `3SEQ` [35]. Only two possible recombinant sequences were found out of more than 7 billion possible triplets using this algorithm, and both candidates were found to have 'parent' sequences from different decades. The explanation put forward by Boni *et al.* [34], that the 25 and 31 year gaps between the two possible parent sequences are due to lab sequencing errors, remains an unsupported hypothesis. In an attempt to explain the absence of evidence of recombination in Influenza, Holmes [36] hypothesizes that recombination may very well add numerous deleterious mutations, to the point where recombination rates might be selectively reduced and highly selected against.

The high variability in the HA and NA sequences from one generation to the next, which translates to different conformational shapes of the epitopes on the infecting virus, is what allows re-infection of the same host organism by Influenza viruses, and is also what makes vaccine construction a continuous and on-going process.

1.5 Vaccination

The effectiveness of the yearly flu vaccine lies in its ability to provide immunity against the current major circulating strains of Influenza. However, with the high mutability of the HA and NA antigens, the major circulating strain can change over the course of months. Antigenic shift and antigenic drift are the terms used when there is a change in the antigenic properties of the HA and or NA gene, either through mutation, reassort-

ment, or even recombination. Conformational changes in these two key antigens allow the influenza virus to infect host cells, bypassing the immune system unrecognized. Influenza vaccines are designed to mimic an influenza infection in a host organism, without triggering a systemic immune response. This allows the immune system to better recognize and fight future infections of similar viruses. The problem lies in constructing the correct type of vaccine, as a single influenza virus' antigenic recognition sites (HA and NA) can change several times over the course of a single year.

The adaptive nature of the Influenza virus is matched by that of the human immune system. The immune system rapidly detects the infecting virus and generates antibodies to prevent future infection with a similar variant. Thus, newly mutated Influenza strains that are antigenically different from their progenitors are able to successfully infect the same host as there will be no immunity to the new virus variant. This rapid evolution necessitates continual adjustment of the composition of the annual Influenza vaccine, which is comprised of three antigens against the HA surface receptor, belonging to H3N2, H1N1 and Influenza B [23], to provide maximum immunity against the proposed major circulating strain of each type. Due to the perpetually evolving nature of the Influenza virus, particularly the HA surface glycoprotein, the composition of each HA in the vaccine changes approximately every 2-5 years [37] depending on the strain, and thus the conglomeration of the three HA antigens are rarely the same from one year to the next.

Vaccine composition begins with strain recommendations from the World Health Organization (WHO) and other National Influenza Centers around the world [23]. Each candidate strain is chosen based on the prediction that it will be one of the dominant circulating strains in the coming flu season, as well as how well it represents those strains [23]. This process occurs 8-10 months before the vaccine is available to the public, and

well before the next flu season has begun. In addition, the data used to select probable candidate strains for the vaccine is unavailable to the public, as is the exact selection process [38], and selection is largely based on the hemagglutinin inhibition assay tests which often have poor resolution in distinguishing between strains [39]. Some experts suggest that testing for new influenza variants should take place later in the season, as these are the strains more likely to escape the immune system and become highly virulent [37]. However, vaccine production and quality assurance are lengthy and are the limiting steps for quick turnaround of an effective vaccine (Figure 1.2).

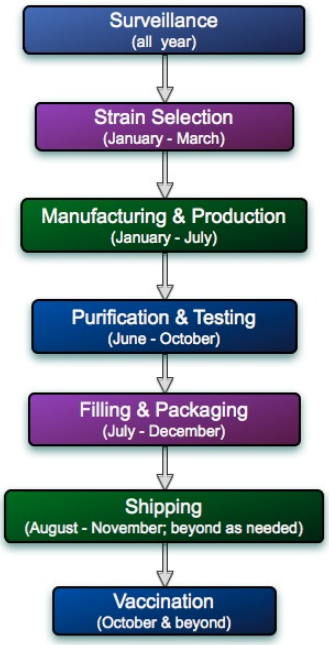

Figure 1.2: **Vaccine selection process.** The selection process and timeline for the North American yearly influenza vaccine, from strain selection to vaccine availability.

1.6 Importance of accurate strain prediction

Due to the long lead time between strain selection and vaccine availability, the possibility of a change in the major circulating strain increases. This shift, if it occurs, can cause the vaccine to become ineffective against one of the three targeted strains of the upcoming influenza season. Often in the northern hemisphere, at the time of strain selection in January/February, the bulk of the current season's epidemic has yet to hit. This prevents the selection committee from having the correct data from which to predict and design an effective vaccine for the next year's seasonal epidemics. This situation has occurred several times in recent years. An Influenza strain was discovered in Fuji in late 2002, which was antigenically distinct from the previous year's A/Sydney/5/97-like major H3N2 strain [40, 37]. The 2003-2004 influenza vaccine failed to protect against the newer A/Fujian/411/2002, resulting in a more severe season than any of the previous three years, as more than 95% of Influenza A infections in Canada were due to the new variant [21]. There is some suggestion that testing of possible major circulating strains should occur later in the Influenza season [37], but the lengthy vaccine production makes this relatively unfeasible.

A more recent example of this imperfect selection process was seen during the 2007-2008 flu season [22]. A virulent H3N2 variant emerged in 2007 in Australia and New Zealand a short time before the influenza season (April-September, Southern hemisphere) and disseminated quickly, causing a widespread epidemic with a three fold increase in the number of infected individuals [22] due to the fact that the virulent strain was not part of the current year's influenza vaccine. This new highly infectious strain, identified as Brisbane/10/2007, crossed the equator to North America for the Northern hemispheres influenza season (November-March), eliciting a similar epidemic during the 2007-2008 season [41].

1.7 Current approaches in predicting influenza emergences

The current strategy of fighting Influenza with yearly vaccines is not nearly as effective as it could be. The rapid evolution of the virus is simply one problem among many. With no way of producing vaccines from cell cultures, the process lies entirely on using chicken eggs as incubators, which themselves are vulnerable to H5N1.

No clear method of predicting the evolutionary path of a single Influenza strain has been found. This is likely due to the many factors that can affect a single subtype, such as geographical area, climate, access to susceptible hosts, herd immunity and prevalence of an effective vaccine. Each of these factors has an effect on the rate of change of each individual virus, all in different ways, and until the environmental and biological effects are understood, a successful model will continue to elude the research community. In addition, due to the segmented nature of the genome, rates of substitution affect each segment of the virus differently [42], greatly adding to the complexity of accurate predictions.

Computational models are one tool that can be useful in the arena of virus evolution, as the repository of genetic data expands with each new sequence added. This would add another level of selection to the current selection procedure, providing a metric for strain suitability. It could also provide an additional filter to further narrow the list of candidate strains under consideration. One of the earlier prediction models using a phylogenetic approach identified nucleotide substitutions in actual HA genes, and used these rates and locations to determine what sequence was most likely to emerge next [43]. However, Plotkin *et al.* [39] later rebutted this type of linear evolution in regards to the HA gene, discovering that HA genes tend to be clustered rather than linear.

Koelle *et al.* [44] use a multi-strain model as the basis for their antigenic tempo model, which attempts to anticipate future Influenza variants based on the antigenic properties of each strain. Their model uses a set of strain-interactions based on work by Gog *et al.* [45], which are determined by tracking host immunity that is dependant on previous infections. However, while the Koelle model is able to predict clusters of future Influenza variants, the length and geographical location of circulation are both overestimated when compared to empirical evidence [44].

Some researchers are investigating other regions of the virus to target for immunization, relying on structural rather than antigenic properties. One example is Heiny *et al.* [46] who determined regions of high conservation across the entire Influenza genome, and throughout a variety of species, and discovered 50 regions in the structural proteins that were viable candidates to target for immunity. Another study looked at the use of drugs to counteract Influenza infection by targeting not only the antigenic components (HA and NA), but also the other major structural and non-structural proteins essential in viral infection and replication [13].

1.8 Objectives of this thesis

It has become evident in recent years that new tools are needed for current strain predictions. New virulent strains are emerging rapidly, such as the 2009 H1N1 swine subtype, infecting a population with little immunity. In addition, the failure of the 2007-2008 flu vaccine has shown that current methods cannot predict a sudden, virulent antigenic shift in the major circulating influenza strains. Additional methods need to be used to determine vaccine composition. The large repository of publicly available data is a valuable resource for researchers to test theories and models, however accounting for this rapid evolution of influenza viruses continue to be problematic.

Unlike previous models, the one described here attempts to take both recombination and natural selection into considerations, both together and singularly at differing rates. Rates of selection and recombination were computed through an analysis of a collection of sequences from previous years, and applied to a more current set of sequences in an effort to mimic past evolutionary forces. The computational time needed for a complete set of sequences was extremely intensive and unfeasible for its intended use, forcing the design of a method to pare down the sequences to a sampling that solely includes representative sequences. Sets of sequences from both HA and NA were examined, as well as a closer study of the original data sets HA and NA sequences from Influenza A. In-depth analysis showed a distinct and significant sequence change in the HA sequences between the 2003 and 2004 Influenza A strains.

Chapter 2

Methods

2.1 Outline

The objective of this study is to generate a sample of highly-probable future sequences of the protein-coding genes of the influenza virus, given some observed sequences. Let us denote these observed sequences as $(X_1, \ldots, X_t) = X_{1:t}$. If time t represents an infectious season for instance, then $X_{1:t}$ represents the sequences sampled between season 1 (some arbitrary point in the past) and season t, and X_{t+1} represents the data at season $t+1$. If t is the current season, X_{t+1} are the future sequences, expected to be circulating during the next season. The quantity of interest is now the 'posterior predictive probability' of the data at season $t+1$, given the data sampled between season 1 and season t, that is:

$$p(X_{t+1}|X_{1:t}) = \int_{\Theta} p(X_{t+1}|\theta)\, p(\theta|X_{1:t})\, \mathrm{d}\theta \tag{2.1}$$

where θ is a vector of nuisance parameters, typically the branch lengths of the phylogeny and the parameters of the model of evolution, and where Θ denotes the state space of θ. Equation (2.1) represents the sum (integral) over the product of two probability density

functions: the likelihood of θ given the future data: $p(X_{t+1}|\theta)$; and the posterior distribution of the nuisance parameters θ given the observed data: $p(\theta|X_{1:t})$. According to Bayes' theorem, this posterior distribution is proportional to the product of the likelihood of θ given the sampled data, $p(X_{1:t}|\theta)$, and a prior on nuisance parameters $p(\theta)$:

$$p(\theta|X_{1:t}) = \frac{p(X_{1:t}|\theta)\, p(\theta)}{p(X_{1:t})} \tag{2.2}$$

The posterior predictive probability (Eq. 2.1) 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. (2.1) cannot be solved analytically. Instead, we used the following two-step procedure, already used by others [47, 48, 49]: first, sample the θ values from the posterior distribution of Eq. (2.2); second, use these sampled θ values to simulate future sequences X_{t+1} (Figure 2.1).

2.2 Computational details

In the first step, the θ values were drawn with the reversible-jump Markov chain Monte Carlo (MCMC) sampler implemented in `OmegaMap` ver. 0.5 [50]. The model used in this approach combines a codon model with a coalescent process with recombination. Model parameters θ include a selection parameter ω, which is the ratio of nonsynonymous to synonymous substitutions, and the population recombination rate ρ. Both of these parameters are defined over an *a priori* block-like structure that segments the alignment of length L into up to L selection blocks and $L-1$ recombination blocks; in both cases, the number of blocks are estimated from the data. The model was parameterized as follows: Prior distributions were here set to have a mean length of 20 and 74 codons, with an exponential mean 1 and mean 0.01 distribution, respectively. The

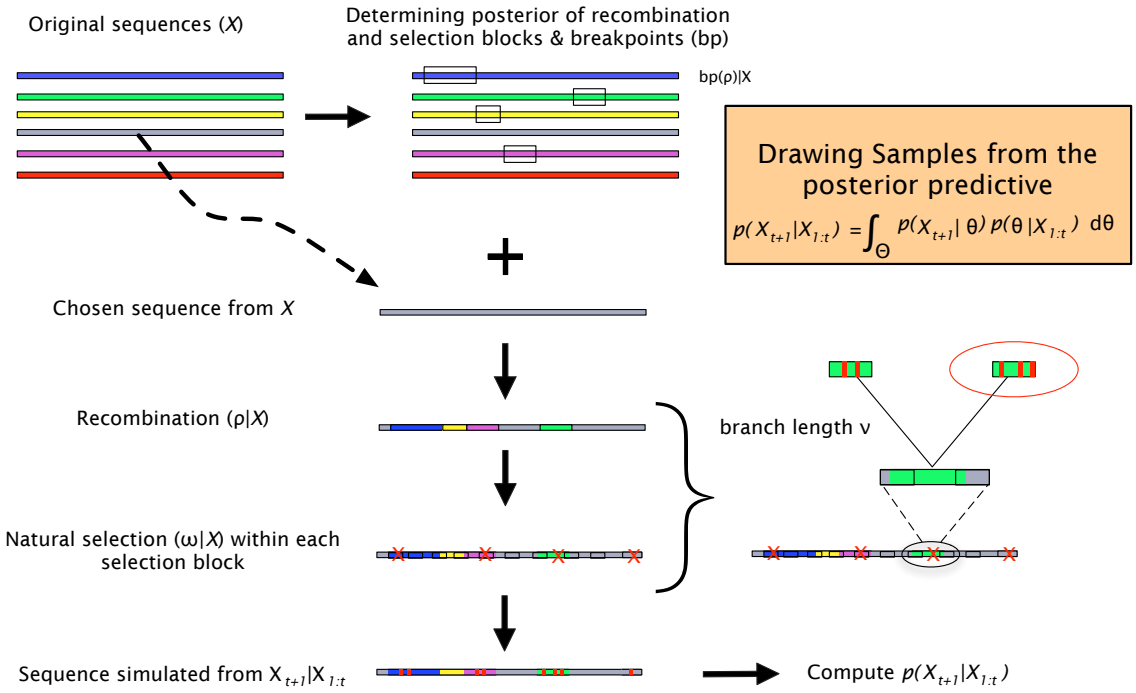

Figure 2.1: **Diagram of the simulation procedure.** Simulation method developed, beginning with the selection of a sequence from the original sample set. The sequence then undergoes a series of recombination events based on ρ, then each selection block is evolved based on ω. The log of the posterior probability of each simulated sequence is calculated before the simulation is repeated with a new sequence and parameters

model also includes some nuisance parameters that are defined over the entire length of the alignment: the transition-transversion ratio $\kappa \sim exp(1/3)$, the rate of synonymous transversion $\mu \sim exp(14.0)$, and the insertion/deletion rate $\phi \sim exp(10.0)$. Equilibrium codon frequencies were set to their empirical frequencies, as calculated by `codeml` [51]. The recombination model is asymmetric as it assumes that one of the sampled sequences is a mosaic of the other sampled sequence; therefore, chains were run with 10 random sequence orderings. Each sampler was run for 10^7 steps with a thinning of 100, which resulted in the MCMC sampler becoming the limiting step, computationally. Two independent runs were performed to check convergence and to obtain the marginal

distributions of $\omega|X_{1:t}$, $\rho|X_{1:t}$ and that of their respective block structures.

The second step consists in the predictive simulation of the new sequences X_{t+1} based upon the θ values sampled from the posterior distribution $p(\theta|X_{1:t})$. The simulation procedure is initialized by estimating the average amount of evolution $\overline{b}$ separating two sequences in $X_{1:t}$. Maximum likelihood pairwise branch length estimation is performed under the one-ratio codon model [52] using `codeml` from PAML ver. 4.0b [51]. Simulations *per se* proceed in two steps. First, a recombinant sequence is generated according to the recombination block structure sampled from $p(\theta|X_{1:t})$. More specifically, 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 year t. The coordinates of the recombination blocks sampled at a given generation of the MCMC sampler 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^{ρ}_{t+1}. Generation of the recombinant sequence was limited to a single recombination block replacement, keeping with the idea that a recombination event is rare enough that it will occur only once per generation. In the second step of the predictive simulation algorithm, this recombinant sequence is evolved following the block structure of the selection (codon) process, as sampled from $p(\theta|X_{1:t})$. Indel characters are first replaced by a random nucleotide (in practice, adenines) to give $X^{\rho\setminus indels}_{t+1}$. Each $\omega|X_{1:t}$ block of $X^{\rho\setminus indels}_{t+1}$ is used as the root of a simulated two-sequence tree $(seq_1 : \overline{b}, seq_2 : \overline{b})$ under the one-ratio codon model parameterized with $(\omega|X_{1:t}, \kappa|X_{1:t})$. Finally, indel characters are repositioned in the simulated sequence $X^{\rho\setminus indels}_{t+1}$ to give X_{t+1}. This process is repeated 100 times for each of the θ values drawn from $p(\theta|X_{1:t})$.

The last step consists in computing the likelihood of the alignment that includes

the simulated sequence: $X_{1:t}, X_{t+1}^{(t)}$. To speed computations up, only the selection block structure was taken into account, and the recombination block structure was ignored as there is little to no detectable recombination in IAVs. For each $\omega|X_{1:t}$ block drawn by the MCMC sampler, a matrix of maximum likelihood pairewise distances is first estimated under the one-ratio codon model [52] using `codeml`. This matrix is used to obtain an approximate tree for this block by weighted Neighbor-Joining [53, 54] as implemented in `weighbor`. Negative branch lengths are set to zero to avoid computational problems. 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 `weighbor` branch lengths. The log-likelihood of the predicted alignment is obtained by summing the log-likelihood values over the selection blocks.

Because computations involved in the last two steps are easily distributed, either on a multiprocessor computer or on a cluster, they are typically extremely quick to perform (of the order of a few days for the data analyzed below). The main computational bottleneck is in the first step, when samples are drawn from the posterior distribution (of the order of a couple of months for the same data).

The simulated sequence were then used as queries in BLASTn searches [55] against a local copy of the influenza database (`ftp://ftp.ncbi.nih.gov/genomes/INFLUENZA`), downloaded in July 2009. This made it possible to know the identity (year and country of sampling, subtype and accession number) of the most similar sequences present in the database, and check whether our algorithm is capable of sampling sequences from the future with a high probability. The BLAST programs are available at `http://www.ncbi.nlm.nih.gov/Ftp`.

2.3 Sequence data for the retrospective study

Individual protein coding sequences for both the HA and NA genes were downloaded from the influenza Virus Resource [56] in July 2008. Only unique, full-length sequences from 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 period of time.

2.4 Sampling method

The first calculation step of the model (Section 2.2) is very computationally intensive, which made it necessary to determine a method of reducing the computational load of the original HA and NA sequence data sets. Thus, a sampling method was devised to generate a series of sequences that are representative of the whole of the diversity of the original, larger data set (Figure 2.2). To accomplish this, the sequences were first aligned with `MUSCLE` [57], then the distances between sequences were calculated in a pairwise fashion with `PAUP` [58] using a maximum likelihood estimator for distance under the GTR + Γ model of evolution. The resultant distance matrix was used to cluster similar sequences using the nearest neighbor method implemented in `DOTUR` [59], after which sequences with 95% similarity or more were removed. The first sequence identified from each of the clusters was taken as the representative sequence for that cluster.

2.5 Altering recombination rates and branch lengths

Changes within the simulation program were used to study the effect of recombination rates and increasing branch lengths on the simulated influenza sequences. In addition,

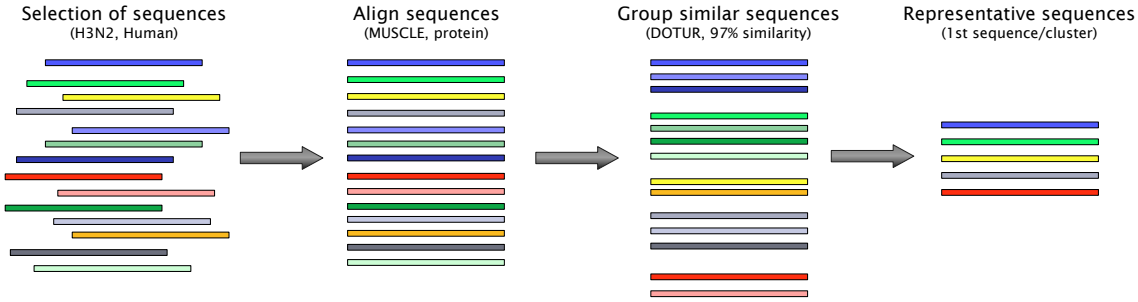

Figure 2.2: **Diagram of sampling process.** Sequences from the Influenza Virus Resource at NCBI were selected from human influenza strains, subtype H3N2, from 2002-2007, equaling 555 (HA) and 498 (NA) sequences. Their protein sequences were aligned, then pairwise distances estimated under GTR + Γ, and finally clustered using nearest neighbor method. The first sequence in each cluster was taken as the representative sequence, resulting in 19 (HA) and 30 (NA) sequences.

the process was altered to allow a single recombination event along the sequences, instead of multiple recombination blocks, which is more in line with natural influenza evolution.

The recombination rates determined by `omegaMap` were used to determine if recombination occurred at each iteration. To study the effects of a single recombination event, the recombination rate was multiplied by factors of 1, 2 and 5. This allowed for an increasing probability of recombination in each iteration. Branch length was also altered to evaluate the effects of increasing branch length on the three different recombination rates. The average branch length used in determining the phylogenetic trees during simulated evolution, were multiplied by factors of 1, 2, 5, and 10.

A phylogenetic tree was created for each of the 12 runs (3 recombination factors x 4 different branch lengths) using `weighbor` on a pairwise distance matrix calculated with `Paup` using a gamma distribution. The top 100 simulated sequences from each of these runs were added to the original HA or NA sequences to create these trees. There were multiples cases of identical sequences simulated in the different runs (ie. 2 or more instances of identical sequences), but were not removed from the data set.

2.6 Assessment of predictive power

In order to assess the predictive power of our model with bursts of evolution (rate multiplier ν increasing from 1 to 10) and / or of our model with bursts of recombination (ϱ increasing from1 to 5), each sequence sampled from the posterior predictive distribution was placed on a phylogenetic tree along with the original sample sequences from the "current sampling period" (HA: 19 sequences, NA: 30 sequences). These Neighbor-Joining trees were reconstructed using maximum likelihood pairwise distances estimated under the general-time reversible substitution model with among-site rate variation modeled as a Γ distribution. 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 our approach has good predictive power, then we expect a significant relationship between the probability of the generated sequences and their distance to the target sequence: highly probable sequences should be very similar to the target sequence and show a small, ideally zero, distance. Predictive power was then quantified by the R^2 of the regression, which is the common practice in linear regression. Slopes were tested as described in [60, p. 493].

Chapter 3

Preliminary Results

3.1 Preliminary work

Initially, the purpose of this process was to determine if new Influenza strains could be predicted from a selected sampling of strains. An earlier test run of 72 HA sequences proved that a full sampling of sequences was too time and processor intensive to be of any practical use. The MCMC inherent in the `omegaMap` program was using upwards of 32G of RAM after 3 months of running continuously. Thus, an alternative approach was developed to minimize the sampling field and shorten the run time (Section 2.4).

3.2 Sequence analysis and prediction under the base model

Full-length HA and NA sequences were extracted from NCBI, and used as the original data sets in the following analyses (Section 2.3). In this retrospective study, the "current sampling period" is set to cover the 5-year period spanning from 2002 to 2007. The

objective is to predict the unexpected emergence of the Brisbane/10/2007 strain.

Due to the large size of the sequence alignments covering this period of time (HA: 555 sequences, NA: 498 sequences), we clustered sequences with at least 95% similarity with `DOTUR` [59], so that the size of each data set be reduced while still maintaining most of the diversity found in each data set. A single sequence from each resultant cluster was arbitrarily chosen as the representative sequence for that group, save for the Brisbane/10/2007 strain, chosen to represent its own cluster. This sampling 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. Two smaller data sets were constructed by using one sequence from each cluster as a representative sequence. The Brisbane/10/2007 sequence was used as the representative sequence from that cluster. In addition, another two data sets were constructed by removing the Brisbane/10/2007 sequence from each of the HA and NA sets before clustering, to determine what effect the addition of the target strain would have on the resulting simulated sequences. This resulted in 4 data sets: HA+ (19 sequences), HA- (18 sequences), NA+ (30 sequences) & NA- (29 sequences).

As evolutionary pressures are exerted on each individual segment to varying degrees, the reduced HA and NA data sets were analyzed independently of each other under the posterior predictive model. Briefly, sequences were sampled from their posterior predictive distribution given a multiple sequence alignment of "current" sequences in two steps, a computational scheme taking inspiration from others [47, 48, 49]. First, posterior distributions of model parameters were obtained. In order to use a realistic and general model of viral evolution that includes both selection and recombination, we used the codon model implemented in `omegaMap` [50]. Here, the parameters of interest are the nonsynonymous to synonymous rate ratio (ω) and the recombination rate (ρ).

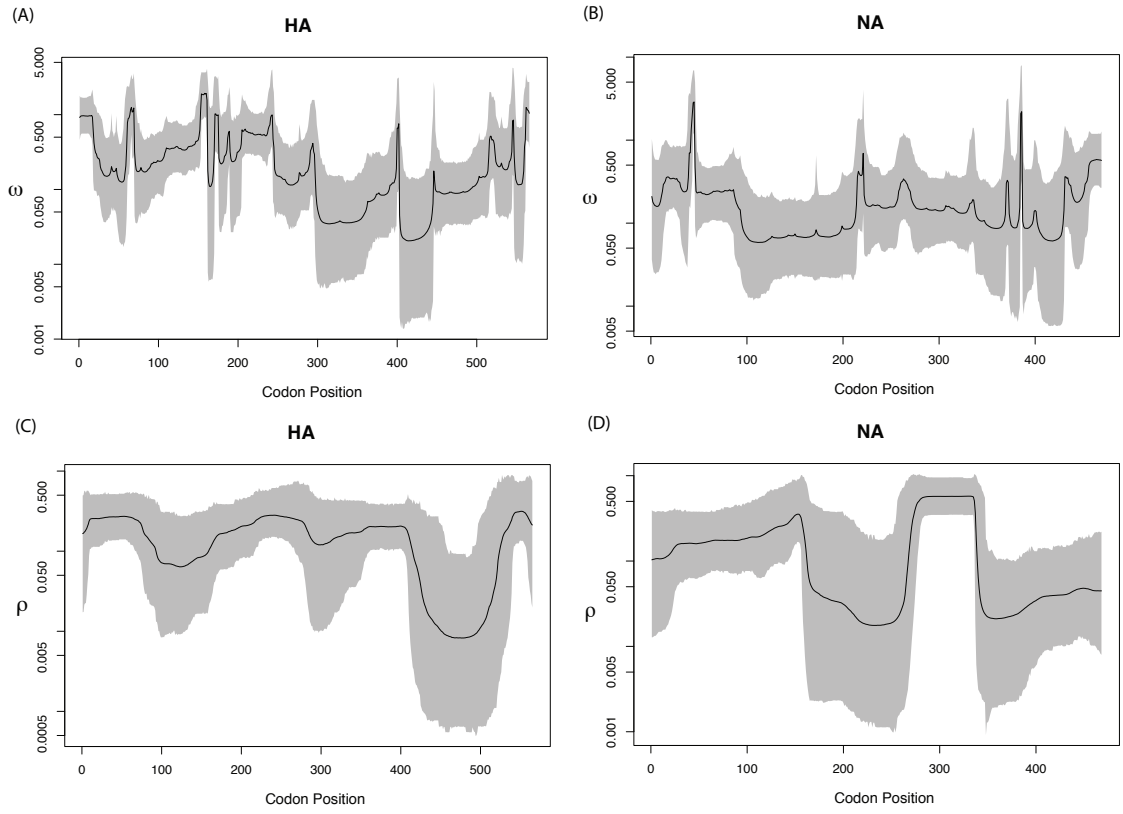

Figure 3.1: **Posterior distributions of ω and ρ.** Posterior distribution and 95% credibility intervals for the selection (ω) and recombination (ρ) parameters along the sequence length of both genes. (A) ω across the length of the HA sequence, (B) ω across the length of the NA sequence, (C) ρ across the length of the HA sequence, and (D) ρ across the length of the NA sequence.

3.3 Initial generation of simulated sequences using predictive algorithm

Output from the first step showed evidence of varying selective pressures (Figure 3.1a-b) and recombination levels (Figure 3.1c-d) across the entire length of both the HA and NA alignments. In the second step of our model, these posterior distributions of ω and ρ were used to generate, for each individual data set, gene sequences drawn from their target

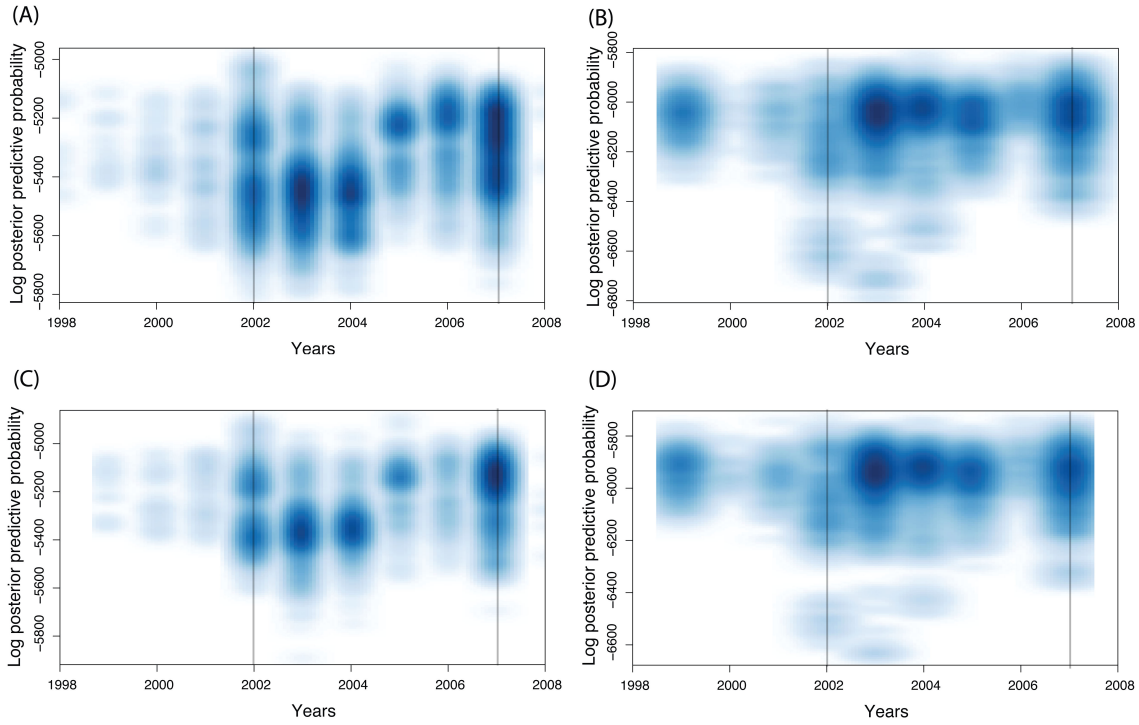

Figure 3.2: **Density plots of simulated sequences through time.** Probability of the predicted sequences as a function of emergence times. The density of the log posterior predictive probability was plotted against the predicted sequences as BLASTn-identified by year for each of the (A) HA, (B) NA data sets that include the target sequences, and (C) and (D) without the target sequences.

posterior predictive distribution. The length of the branch leading to the predicted sequences was set to the average branch length of the tree containing sequences from the "current sampling period". The posterior predictive probability of each predicted sequence was then computed. These predicted sequences were then labeled according to their top BLASTn hit against a local influenza database extracted from NCBI. We note that only taking the top hit is a very stringent labeling criterion.

The ω and ρ probabilities were used to generate sets of simulated sequences for each of the four data sets used in this study. A large proportion of the simulated sequences were identified in the original 2002-2007 time frame from which the original sequences

Table 3.1: Total sequences identified by year from the initial algorithm run, showing time periods before 2002 and during 2008.

Data Sets	**Before 2002**	**% Before 2002**	**2008**	**% 2008**
HA+	185	1.3%	3	0.02%
HA-	403	1.1%	8	0.02%
NA+	2148	9.1%	6	0.02%
NA-	4337	10.0%	0	0.00%

Table 3.2: Number of simulated Brisbane/10/2007 strains found in each of the four data sets.

Data Sets	**#Sequences Simulated**	**#Brisbane/10/2007**	**% Brisbane/10/2007**
HA+	14,035	0	0.0%
HA-	35,363	0	0.0%
NA+	23,532	916	3.9%
NA-	43,429	1000	2.3%

came from (Figure 3.2).

All four data sets contained sequences simulated by our prediction model that were identified as being from before 2002 (Table 3.1). The HA data sets showed fewer simulated sequences from before 2002 than the NA data sets (1.1-1.3% and 9.1-10%, respectively). The simulated sequences identified as 2008 Influenza strains showed no clear preference between HA and NA data sets, unlike the 2002 sequences. In fact, there were very few 2008 sequences simulated in any of the four sequence sets, ranging from 0 to 8 sequences in total (Table 3.1).

In addition, there was a distinct preference in the number of simulated Brisbane/10/2007 sequence between the HA and NA data sets (Table 3.2). Neither of the HA data sets predicted any Brisbane/10/2007 sequences, while the NA+ and NA- data sets simulated the Brisbane/10/2007 sequence at 3.4% and 2.3% respectively.

Table 3.3: Number of clusters generated in each of the four combined data sets, in four different top percentiles (1.0%, 0.5%, 0.15%, 0.01%).

Data Sets	**1.0%**	**0.5%**	**0.1%**	**0.01%**
HA+	138	174	29	20
HA-	290	143	44	21
NA+	100	77	45	32
NA-	161	124	53	32

3.4 Analysis and phylogenetic results of clustered sequences

Simulated sequences with the highest posterior predictive probability were selected from each of the four data sets for further study. Sequences from the top 1.0%, 0.5%, 0.1% and 0.01% were selected and clustered with the original HA and NA data sets (555 and 498 sequences, respectively) using the sampling method described in Section 2.4 (Table 3.3). The Brisbane/10/2007 cluster was analyzed in each of the four data sets at 1.0%. In the HA+ data set, there were five simulated sequences in the Brisbane/10/2007 cluster, all 5 identified as Brisbane/2006 strains, and 387 sequences from the original 555 data set. No other data set contained any simulated sequences in the Brisbane/10/2007 cluster, and the inclusion or exclusion of the Brisbane/10/2007 strain showed no change in simulations of sequences.

3.5 Effect of increasing branch lengths on simulated sequences

Branch lengths were used as an internal variable representing a unit of time in the prediction model when phylogenetic trees were created, to evolve the recombinant sequence.

They were multiplied by 10, 20, 50 and 100 times what was used previously (1×), resulting in fold increases in simulated evolution. The HA+ data set was used to test what effect increasing the branch lengths would have on the simulation of sequences.

The simulated sequences showed a wider dispersion across years as the branch lengths increased from 10 to 100 times what was originally used (Figure 3.3). There was an increase in the number of strains from before 2002, but also an increase in strains found in 2008. The posterior predictive probability also decreases as the branch lengths increase. The number of clusters also increased with branch length (10× = 44 clusters, 20× = 46 clusters, 50× = 257 clusters, 100× = 409 clusters) showing an increase in diversity as branch lengths increased. When the top 1.0% of sequences were clustered with the original HA (555) sequences, there were still no simulated sequences seen in the Brisbane/10/2007 cluster.

Phylogenetic trees were constructed in the same way as described in Section 3.4, using representative sequences from the clusters and the original HA sequences (Figure 3.4). The trees seen at both 50× and 100× show extreme branch lengths in the simulated sequences, the majority of which tend to cluster together on the tree. At 10× and 20×, the Brisbane/10/2007 clade was not nearly as distinct as previous trees have shown, as it appeared to cluster more in the majority than in a distinct clade.

3.6 Preliminary findings

The model was able to simulate sequences BLASTn identified as the Brisbane/10/2007 strain for the NA data set, 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 top 5% of the distribution) using either the HA or NA data sets (Figure 3.5). On the other side of the prediction spectrum, both the HA and NA posterior pre-

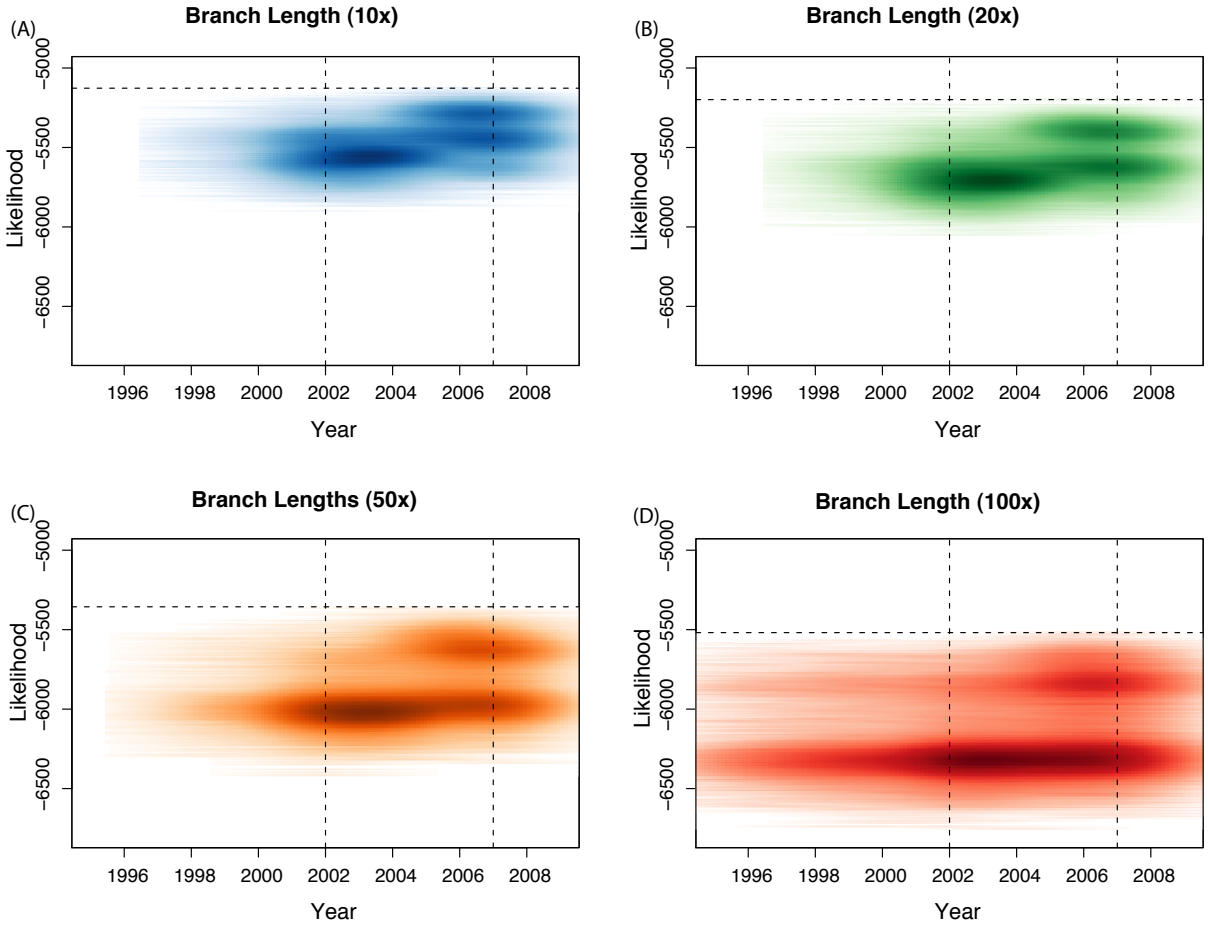

Figure 3.3: Density plots of different branch lengths used in simulating sequences using the HA+ set of sequences. Sequences are identified by year and plotted vs. the calculated log likelihood for each different branch length used: (A)10×, (B)20×, (C)50× and (D)100×.

dictive sets 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 simulated sequences. Note that this persistence of circulating sequences might be less pronounced for the HA gene (Figure 3.5a) than for the NA gene (Figure 3.5b). Similar results were obtained when the target sequences were not included in the analysis (Figure 3.2) for HA ($\chi^2_{48} = 54$, p-value $= 0.2559$; Fisher exact test: p-value $= 1$) and for NA ($\chi^2_{49} = 56$, p-value $= 0.2289$; Fisher exact test: p-value $= 1$). Therefore,

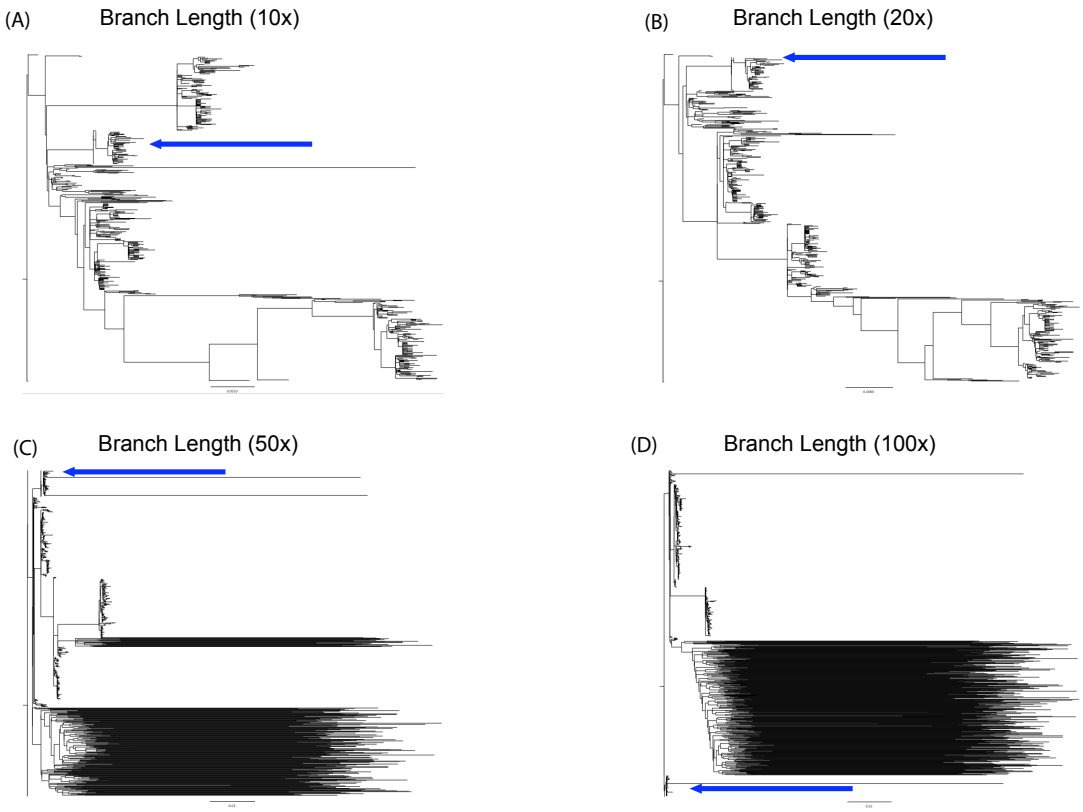

Figure 3.4: Phylogenetic trees constructed using representative simulated sequences added to the original set of HA sequences. The branch lengths used: (A) ×10 (B) ×20 (C) ×50 (D) ×100

the predictive results are unlikely to be affected by the presence of these sequences, which are kept for the rest of the analyses.

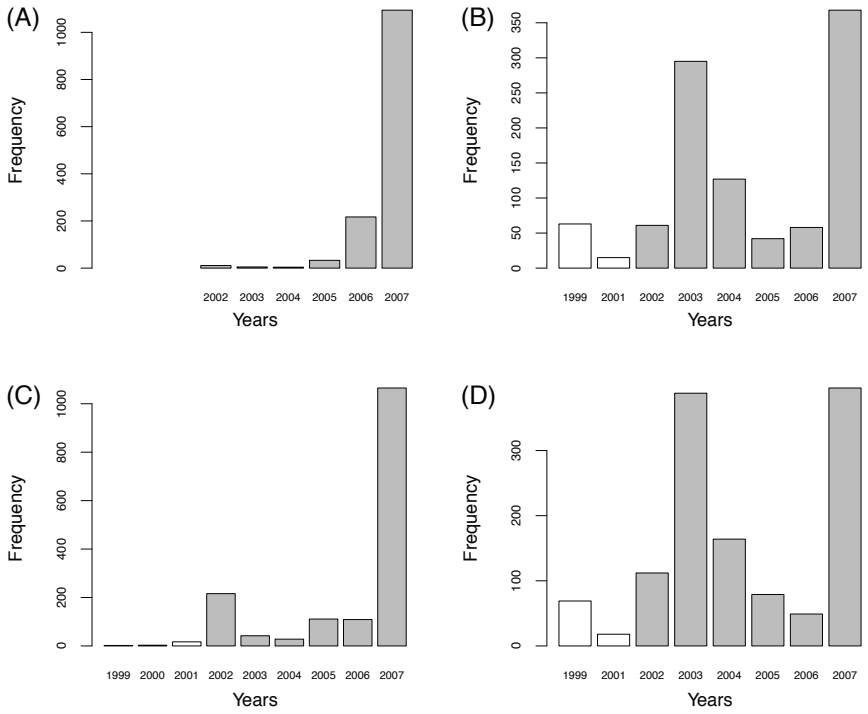

Figure 3.5: **Distribution of the BLASTn-identified sequences in the top 5% of the posterior predictive distribution.** Results are presented for data sets including the target sequence for (A) HA and (B) NA, and the data set excluding the target sequence for (C) HA and (D) NA. Shaded bars represent sequences BLASTn-identified as coming from the "current sampling period" (2002-2007), while empty bars represent sequences coming from outside of this period.

Chapter 4

Predicting the emergence of Influenza A viruses

4.1 Patristic distance as a measure of predictive power

In order to quantify the predictive power of the model, the log posterior probabilities of the simulated sequences were plotted against the patristic distances to the target strain. The R^2 value of the regression was calculated (proportion of the variance explained by the linear fit) for each set of simulated sequences. Table 4.1 shows that the base model has a predictive power of 26% for HA and 18% for NA. In the remainder of the text, we evaluate the impact of the sequences included in our "current sampling period", and assess means to improve the predictive power of the base model.

4.2 Effect of duration of "current sampling period"

The impact of the range of dates from which sequences were included in the "current sampling period" were investigated. The hypothesis here was that the predictive power

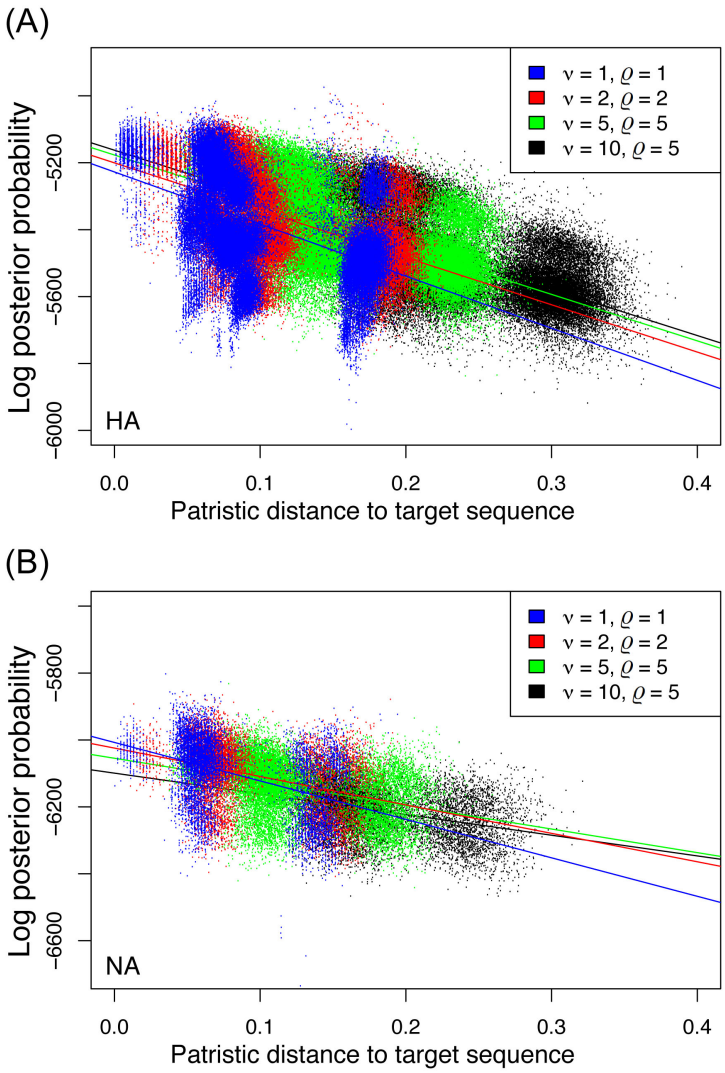

Figure 4.1: **Patristic Distances of both recombination rates and branch lengths.** Log posterior probabilities plotted against patristic distance between each simulated sequence and the target Brisbane/10/2007 sequences, both for HA (A) and NA (B).

of our model would be high if the evolutionary process is stationary during the sampled time period. Alternatively, reducing the year range might decrease power in the same way that decreasing sample size in a linear regression decreases 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

Table 4.1: R^2 values for the linear regressions of log posterior predictive probabilities against patristic distances to the target sequence.

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

2002 and 2007. A phylogenetic tree of the original 555 HA sequences (not showed) highlighted a distinct clade of sequences that were sufficiently distant from the rest of the sequences to potentially pose problems in the prediction model. Closer inspection revealed that the entirety of this clade was comprised of sequences from 2002 and 2003, suggesting that an evolutionary shift might have occurred after 2003. Therefore, all 2002-2003 sequences were removed from both the HA and NA data sets, and the posterior predictive algorithm was then run on the remaining 2004-2007 strains. This provided a 3-year time span of data, as opposed to the original 5-year span. In addition, a data set containing only the 2005 sequences was also used, as 2005 was the year that contained the largest number of sequences in the original data sets. As a result, it was possible to compare the effectiveness of the predictive method across three year ranges: 1 year (2005), 3 years (2004-2007), 5 years (2002-2007).

For each of the three ranges of years, we could not assess the predictive power of the model by plotting the log posterior probabilities of the simulated sequences against the patristic distances to the target strain as in the previous section. The 2005 data did not contain the target sequence (Brisbane/10/2007), resulting in no calculable patristic distances for this range. Thus, we calculated pairwise distances between each pair of

sequences for all three year ranges.

The results show that the 3-year and 1-year analysis have higher probabilities than the 5-year study performed above (Figure 4.2 for HA; NA not shown), which is expected since the 5-year alignment is larger. Crucially, the regression slope of the 3-year data set is smaller (in absolute value) than that of the full 5-year data set, which suggests a decrease in predictive power with a smaller data set. The R^2 value for the 5-year data set stands at 0.37, and drops to 0.11 for the 3-year data set. Following the same trend, the 1-year data set showed no predictive power (negative slope, not significantly different from 0, with $R^2 = 6.5 \times 10^{-4}$), and only generated sequences extremely far from the target sequence (pairwise distances > 0.4). Therefore, on the basis of this retrospective study, it seems that longer time spans for the "current sampling period" improve the predictive power of our model. Therefore in the rest of this study, only use the original 5-year data sets.

The posterior predictive model predicted the persistent circulation of Influenza A strains, which is consistent with previous work [42]. Even if the predictive power was not impressive (Table 4.1), our model failed to predict the emergence of the unexpected target strain (Brisbane/10/2007) with a high probability. To improve the prediction of the "unexpected" target strain, we incorporated punctual bursts of evolution and of recombination into our base model.

4.3 Effect of punctual bursts of evolution

For the reasons exposed above, we only consider the 5-year data sets, for which the "current sampling period" covers 2002 to 2007. 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 ν, mimicking a punctual burst of evolution. By default, this length

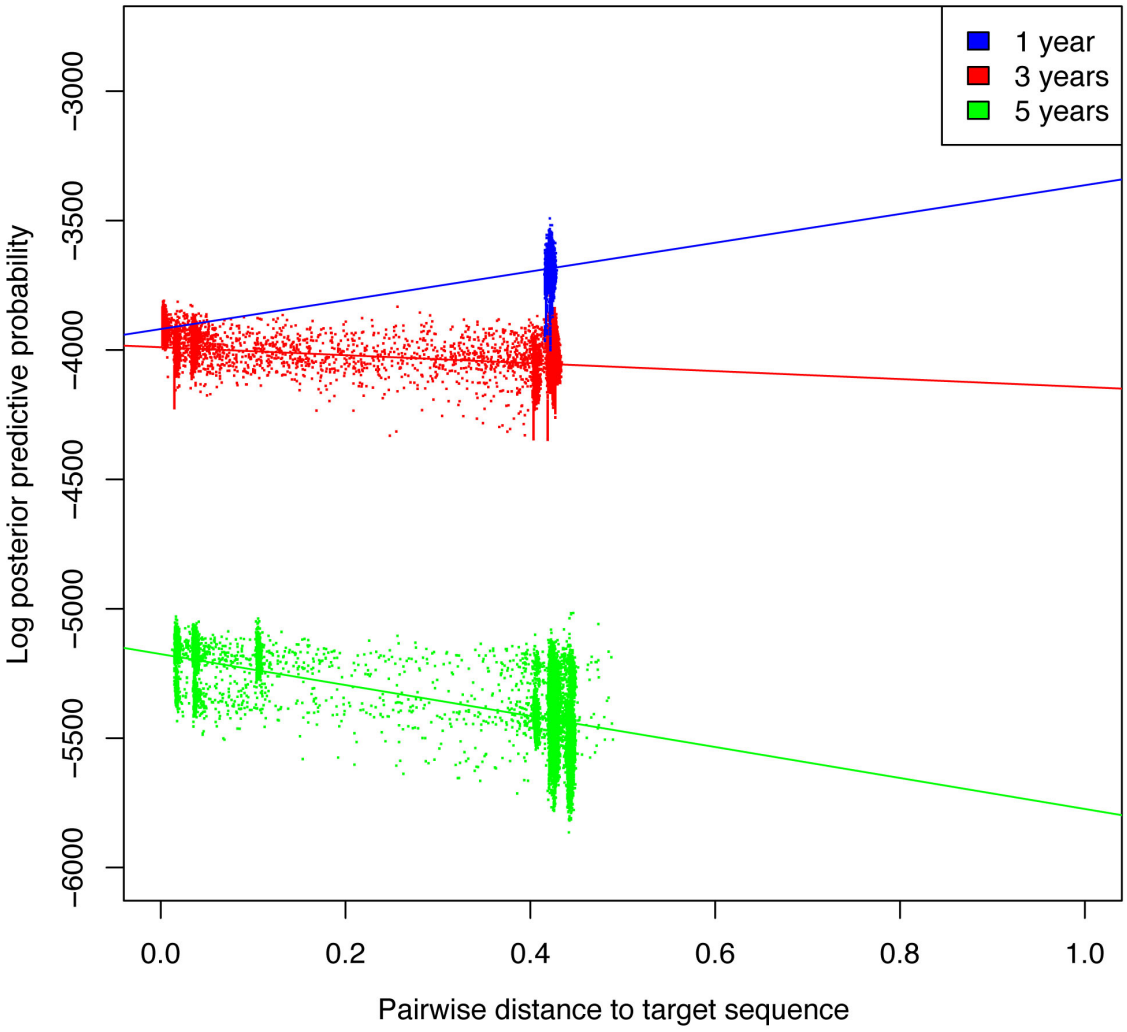

Figure 4.2: **Posterior predictive power of data subsampled for times.** Quantification of predictive power of each change in the original year range of sequence selection.

is set to the average branch length of the tree containing only the sequences from the "current sampling period" (see Methods). The scaling factor ν was set to 1, 2, 5 and 10 for both sets of sequences. A phylogenetic tree was reconstructed using the sequences from the "current sampling period" and the top 100 simulated sequences with the highest posterior predictive probability (Figure 4.3; results not shown for NA). As expected, the number of easily identifiable clades on these phylogenetic trees increased with the branch

length multiplier ν, showing a greater diversity among the simulated sequences.

Again, the predictive power of the model was assessed by plotting the log posterior probabilities of the simulated sequences against the patristic distances to the target strain. The results show a very significant negative relationship between posterior predictive probabilities and patristic distances for all ν multipliers, both for HA (Figure 4.1a) and NA (Figure 4.1b). These significant relationships demonstrate that the model is able to predict sequences that (i) have a relatively high probability and (ii) that are close to the sequences that actually emerged in nature, that is, the target Brisbane/10/2007 sequence. Table 4.2 further shows that for HA, the average probability is increasing with ν (see also Figure 4.1a), while the slopes show a small but significant decrease in absolute value (Table 4.3). The pattern is similar for NA, where the slopes are progressively decreasing (in absolute value) with ν, but to a much larger extent (Table 4.2 and 4.3). These results suggest that the inclusion of bursts of evolution in our posterior predictive model helps predict HA sequences. Indeed, the R^2 values of the regressions increase to almost 40% as ν increases (Table 4.1). On the other hand, the inclusion of bursts of evolution makes our prediction of NA sequences worst, as R^2 values decrease with increasing ν (Table 4.1). Reciprocally, these results suggest that the evolution of HA sequences during that period of time for the H3N2 subtype might be characterized by episodic bursts of evolution (at least in the case of the emergence of Brisbane/10/2007), while the evolution of NA might be more gradual.

4.4 Effect of bursts of recombination

Although homologous (intra-segmental) recombination is generally not considered to be a major driver of the evolution of Influenza A viruses [34], the posterior analysis shows that recombination rates are variable across the genes analyzed here (Figure 3.1), so

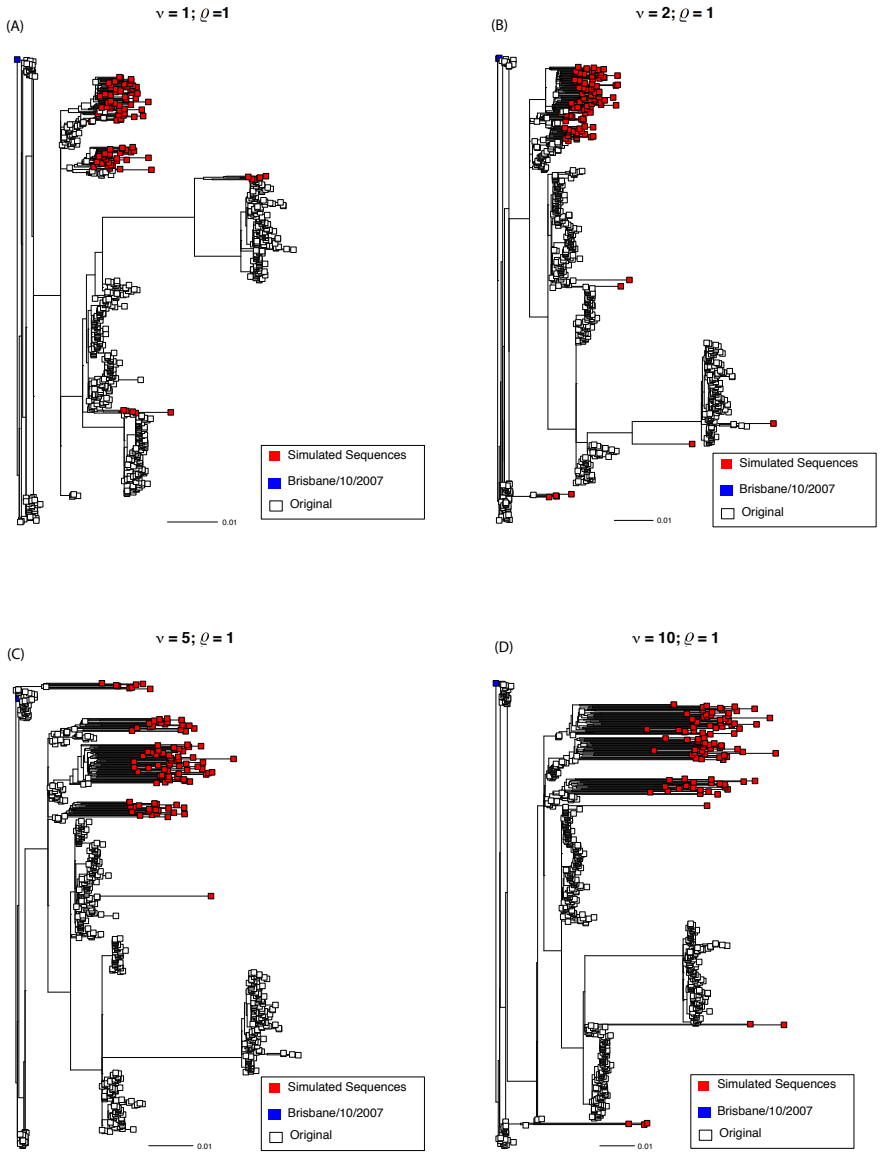

Figure 4.3: **Phylogenetic Trees.** Phylogenetic trees containing the top 100 simulated sequences plus the original 555 HA sequences, constructed by weighted Neighbor-Joining, and artificially rooted with the Brisbane/10/2007 sequence for four branch lengths multipliers: (A) $\nu = 1$, (B) $\nu = 2$, (C) $\nu = 5$, (D) $\nu = 10$. Box colors indicate the origin of sequences: red for simulated sequences, white for the original 555 sequences and blue for the Brisbane/10/2007 sequence.

Table 4.2: Slopes for the regressions of log posterior predictive probabilities against patristic distances to the target sequence.

		$\nu = 1$	$\nu = 5$	$\nu = 5$	$\nu = 10$
HA	$\varrho = 1$	-1557.390	-1442.617	-1445.967	-1419.968
	$\varrho = 2$	-1488.930	-1414.069	-1420.374	-1399.896
	$\varrho = 5$	-1437.566	-1371.948	-1388.851	-1385.064
NA	$\varrho = 1$	-1149.831	-854.641	-734.015	-622.877
	$\varrho = 2$	-1040.274	-849.571	-704.738	-616.660
	$\varrho = 5$	-1012.370	-793.469	-705.283	-618.557

Table 4.3: *P*-values for the pairwise comparison of slopes for the regressions of log posterior predictive probabilities against patristic distances to the target sequence when ν was altered. Results for HA are presented in the lower triangular matrix, while those for NA are above the diagonal.

	–	$\varrho = 1\ \nu = 1$	$\varrho = 1\ \nu = 2$	$\varrho = 1\ \nu = 5$	$\varrho = 1\ \nu = 10$	
	$\varrho = 1\ \nu = 1$	–	1.62×10^{-13}	3.67×10^{-35}	3.86×10^{-64}	NA
	$\varrho = 1\ \nu = 2$	4.68×10^{-16}	–	1.23×10^{-5}	3.77×10^{-21}	
	$\varrho = 1\ \nu = 5$	3.05×10^{-17}	0.7770	–	1.56×10^{-7}	
HA	$\varrho = 1\ \nu = 10$	8.80×10^{-26}	0.0535	0.0140	–	

that recombination could, in principle, play a role in the emergence of novel influenza viruses. In order to assess the impact of recombination on the emergence of the target Brisbane/10/2007 strain, a burst of recombination was included in our model and its impact on our predictive power was evaluated as described above. Branch length multipliers ν were first kept constant and set to 1, while recombination rates along the branch leading to the simulated sequences were multiplied by a factor ϱ 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 the log posterior probabilities of the

simulated sequences against the patristic distances to the target strain. These regressions were highly significant for both the HA and the NA genes (Figure 4.1). However, increasing the recombination rate multiplier ϱ led essentially to unchanged or even decreasing predictive power, both for HA and NA (Table 4.1). Therefore, consistently with the current consensus that homologous (intra-segmental) recombination is not significant in the evolution of Influenza A viruses, the results suggest that this recombination process does not improve the predictive power of the Brisbane/10/2007 strain.

4.5 Joint effect of bursts of evolution and of recombination

Of the rate increase in branch lengths and in recombination rates studied separately above, only the former led to an improved predictive power, in particular for the HA gene. The joint effect of increasing ν and ϱ on the predictive power of our model was investigated. The model was run on all combinations of branch lengths (ν set to 1, 2, 5, 10) and recombination rates (ϱ set to 1, 2, 5). Phylogenetic trees were constructed as above for each of the $4 \times 3 = 12$ possible combinations.

The resulting trees showed an expected increase in sequence diversity as both the branch length multiplier ν and the recombination rate multiplier ϱ were increased (not shown). The computation of patristic distances for the different ν and ϱ combinations supported the pattern of increased sequence diversity both for HA (Figure 4.1a) and NA (Figure 4.1b). Supporting the results found when varying ν or ϱ independently, the HA gene proved to be more responsive than the NA gene to a joint increase in ν and ϱ, while the impact of bursts of recombination was negligible (Table 4.1). These results again support the hypothesis that the evolution of HA of strain Brisbane/10/2007 was mostly

driven by a burst of rates of evolution.

Chapter 5

Conclusions

5.1 Recombination not present in Influenza A

In this study, the prediction model was built with the idea of using the model on other RNA viruses, many of which undergo homologous recombination. As the question of homologous recombination in Influenza A is still under debate, testing of the model at different rates of recombination was done. While there are several studies supporting the presence of homologous recombination in Influenza A [31, 32], the current consensus [33, 34] is that recombination is not present among Influenza A viruses, which the results of this study have supported. This is seen in the R^2 values in Table 4.1, where increasing recombination rates had no positive effect on the sequences produced by the prediction model, and in fact had a slight negative effect in terms of R^2 in both the HA and NA values. Overall, taking into account the presence of recombination in the model showed no increase in the ability to predict future Influenza A strains.

5.2 Principal findings

Time constraints are a large part of vaccine production, so the development of a model that can help predict emerging influenza strains quickly and accurately is essential. The computational time required by the model in the initial test was, at 3 months, far greater than what was desired. 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 computational bottleneck. Therefore, a sampling method was devised to produce smaller HA and NA data sets, while still preserving most of the existing sequence diversity. However, this sampling method to reduce data sets intrinsically discards information relative to haplotype frequencies. It is possible that this information is 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 [24], is it more likely that nonadaptive processes, such as drift, play a key role in the emergence of influenza viruses. If this hypothesis on the mode of evolution of influenza viruses is correct, the 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 influenza viruses.

The sequences generated by the predictive model from the initial two data sets (HA, NA) reveal in particular that the majority of the high-probability sequences were generated from 2002 and 2007 (Figure 3.5), the original sampling period for the HA and NA sequences. This suggests that influenza strains continue to circulate for several years after emergence. However, the predictive model does assume that the key genetic parameters, ω and ρ, are stationary over time. This stationarity assumption could increase the length of time our model assumes an influenza strain persists in the population. In spite

of this potential bias, previous studies have documented the persistence of strains in natural populations of Influenza A viruses [39, 42]. One first possibility that was explored here was to reduce the time interval for inclusion of sequences in the "current sampling period", with the expectation that such a reduction would help alleviate the stationarity dependence; however, this shortening of the "current sampling period" led to a reduction in predictive power (Table 4.1). The explored alternative, to include historical 'accidents' such as bursts of evolution and recombination, increased the predictive power from 25% to about 40% for HA, but not for NA, whose power remained low at 12-28% (Table 4.1). Note that the inclusion of the target sequence in the "current sampling period" did not affect of results, which means that the approach is robust to the sampled sequences. Since only the HA protein is used in the vaccine, the better results for HA than for NA are very encouraging.

It is difficult to compare this predictive power, defined as an R^2 value, with other measures employed in machine learning for epitope or binding prediction (e.g., [61]), which generally report for that purpose the area under the receiver operating characteristic (ROC) curve. In spite of having been developed to evaluate binary classifiers, recent developments have paved the way to making ROC analysis amendable to sequence data [62]. However, these approaches only indicate that the predictive model is suitable for building good classifiers, not that a given classifier is going to be efficient [63]. As a result, there was no attempt to use this approach to quantify the predictive power of this model. However, it is clear that in the retrospective study, there was a failure to simulate the target sequences with a high probability. Altogether, the results suggest that in spite of genomic realism, the model has a number of shortcomings.

Validating this method based on its prediction of one particular strain does not seem sufficient, given that the Brisbane/10/2007 strain was not found to be predicted by any

other models, suggesting that its emergence is due to a rare set of circumstances. A better method of validation would be to test this prediction model using more common or more abundant strains, to determine if minor shifts in Influenza populations can be detected on a yearly basis, rather than more drastic shifts as seen in 2007. There is also a need to validate the method using other rapidly emerging strains, such as the Fujian/411/2002 strain that emerged in late 2002 in a similar rapid emergence as the Brisbane/10/2007 strain. A useful predictive model would need to predict not only these new, rapidly emerging strains, but also the more common strains that do not require a complete change in the Influenza vaccine.

While the main focus here was on the sequence as a whole, it would be interesting to focus more closely on known sites of genetic interest such as drug resistance, areas of higher likelihood of mutation [42] or known viral epitopes. In a recent study of the 2009 H1N1 swine flu outbreak [64], it was found that one of the key mutations that gave the virus its high virulence, was found on a loop-forming of an epitope. This is a more epidemiological approach to strain prediction focussing on the specific areas of IAVs that directly affect virulence.

In terms of advantages over previous models [43, 65, 39], the predictive model described here has several unique features such as the incorporation of a more realistic model of natural evolution, and a model of recombination and/or reassortment. Although not exploited here, this model can be used to predict the emergence of whole strains, not just individual protein-coding genes. To do so, concatenated alignments of all 10 segments should be constructed, analyzed with `omegaMap` to find posterior probabilities of the parameters of interest (ω and ρ) as well as the "recombination" breakpoints. These breakpoints are expected to correspond to the limits of the different segments of the Influenza A genome, so that recombination and reassortment become similar processes

from the standpoint of the model. The rest of the analysis follows what is described above. This kind of analysis was not attempted for computational reasons, as the first step of the algorithm (with `omegaMap`) is currently the limiting step, both in terms of memory footprint (we were limited to 32GB of RAM) and running time. It is expected that further developments will make this step more efficient.

5.3 Future directions

In order to minimize computational time, as well as aid in increasing predictive power, it would be necessary to investigate strain prediction on a year-by-year basis, rather than as year ranges. As such, sequence prediction would begin with only Influenza sequences identified in a single influenza season, but with the ω and ρ derived from the overall set of sequences, and proceed year by year to determine the accuracy of the prediction model. Though the use of a single year range (2005 sequences, Section 4.2, Figure 4.2) proved unusable in this case, it may be possible to effectively use sequences from a single season to predict the following year's sequences, rather than a single sequence two or more years in the future. One factor that would need to be taken into account is that influenza seasons differ based on location (North America, October-March; Oceania, April-September) [41], and whether the definition of an influenza season is a calendar year (January-Decenber) or the time period of increased influenza infections (October-February for the northern hemisphere).

However, as with previous genetic models (see [65]), including grammar models [66], this approach fails 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 (e.g., [67]), and are therefore difficult to incorporate into genetic models. One notable exception attempted to reconcile the

outputs of the two approaches, but did not attempt to predict emerging strains [68]. More recent forays in spatial studies address the surveillance issue from a phylogenetic point of view [69, 70, 71]. Although these tools have the potential to predict where a particular known virus is likely to emerge [71], they do not attempt to predict which viruses are likely to emerge. Finally, the development of predictors of epidemics and pandemics would clearly benefit from the release of a publicly available database linking influenza genomes to a proxy of their phenotype, such as the results of hemagglutination inhibition assays [40]. 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 modelling and of course the consideration of the population genetics of the viruses of interest.

5.4 Concluding statement

Overall, the model described here only has a moderate predictive power (up to ∼40% in terms of R^2 values) when attempting to generate the unexpected sequence of interest (Brisbane/10/2007) with a high posterior predictive probability. This may be biologically relevant, since highly pathogenic sequences do not emerge frequently, and the probability of generating the exact sequence is relatively low. In the majority of the phylogenetic trees generated by this model, the Brisbane/10/2007 sequence tended to be located in a separate clade, distinct from the rest of the tree. This showed that the Brisbane/10/2007 strain differed from the majority of other H3N2 strains, and likely emerged through a different process than other H3N2 variants, and a process that departs from (i) stationarity (continuation of past biological processes), (ii) busts of evolution (in particular for NA), (iii) bursts of recombination / reassortment or (iv) a mixture of bursts of evolution and recombination. It should be noted that, as in previous prediction attempts (e.g., [43]),

it was assumed that viral effective populations sizes were "large enough" for selection to be possible. These results may therefore hint at the driving role of nonadaptive events, such as drift, in the emergence of highly pathogenic strains of Influenza A viruses, which is consistent with previous estimates of ancestral population sizes of Influenza A viruses [24].

Bibliography

[1] Kuiken T: **Host species barriers to Influenza virus infections**. *Science* 2006, **312**:394–397.

[2] Jakiela B, Brockman-Schneider R, Amineva S, Lee WM, Gern JE: **Basal cells of differentiated bronchial epithelium are more susceptible to Rhinovirus infection**. *American Journal of Respiratory Cell and Molecular Biology* 2008, **38**(5):517–523.

[3] Baron S (Ed): *Medical Microbiology*. University of Texas Medical Branch, 4th edition 1996.

[4] Koonin EV, Senkevich TG, Dolja VV: **The ancient virus world and evolution of cells**. *Biology Direct* 2006, **1**:29.

[5] Moya A, Holmes EC, González-Candelas F: **The population genetics and evolutionary epidemiology of RNA viruses**. *Nature Reviews Microbiology* 2004, **2**(4):279–288.

[6] Ghedin E, Claverie JM: **Mimivirus relatives in the Sargasso sea.** *Virology Journal* 2005, **2**(62).

[7] Faurez F, Dory D, Grasland B: **Replication of porcine circoviruses**. *Virology Journal* 2009, **6**(60).

[8] Taylor T, Brockman M, McNanmee E: **Herpes simplex virus**. *Frontiers in Bioscience* 2002, **7**:752–764.

[9] Desnues C, Raoult D: **Inside the lifestyle of the virophage.** *Intervirology* 2010, **53**(5):293–303.

[10] Belser JA, Blixt O, Chen LM, Pappas C, Maines TR, Hoeven NV, Donis R, Busch J, McBride R, Paulson JC, Katz JM, Tumpey TM: **Contemporary North American influenza H7 viruses possess human receptor specificity: Implications for virus transmissibility**. *PNAS* 2008, **105**(21):7558–7563.

[11] Enserink M: **Controversial studies give a deadly flu virus wings**. *Science* 2011, **334**:1192–1193.

[12] Ghedin E, Sengamalay NA, Shumway M, Zaborsky J, Feldblyum T, Subbu V, Spiro DJ, Sitz J, Koo H, Bolotov P, Dernovoy D, Tatusova T, Bao Y, George KS, Taylor J, Lipman DJ, Fraser CM, Taubenberger JK, Salzberg SL: **Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution**. *Nature* 2005, **437**:1162–1166.

[13] Das K, Aramini J, Ma LC, Krug R, Arnold E: **Structures of influenza A proteins and insights into antiviral drug targets**. *Nature Structural and Molecular Biology* 2010, **17**(5):530–538.

[14] Nelson MI, Holmes EC: **The evolution of epidemic influenza**. *Nature Reviews Genetics* 2007, **8**(3):196–205.

[15] Liu N, Wang G, Lee KC, Guan Y, Chen H, Cai Z: **Mutations in influenza virus replication and transcription: detection of amino acid substitutions in hemagglutinin of an avian influenza virus (H1N1)**. *FASEB Journal* 2009, **23**(10):3377–3382.

[16] Wise H, Foeglein A, Sun J, Dalton R, Patel S, Howard W, Anderson E, Barclay W, Digard P: **A complicated message: Identification of a novel PB1-related protein translated from Influenza A virus segment 2 mRNA**. *Journal of Virology* 2009, **83**(16):8021–8031.

[17] Bouvier NM, Palese P: **The biology of influenza viruses**. *Vaccine* 2008, **26**(Supplement 4):D49–D53.

[18] Chen R, Holmes EC: **The evolutionary dynamics of human Influenza B virus**. *Journal of Molecular Evolution* 2008, **66**(6):655–663.

[19] Taubenberger JK, Morens DM: **The pathology of influenza virus infections.** *Annual Review of Pathology* 2008, **3**:499–522.

[20] 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: **A distinct lineage of influenza A virus from bats**. *PNAS* 2012, **109**(11):4269–4274.

[21] Orr P: **Statement on Influenza vaccination for the 2004-2005 season**. *Canada Communicable Disease Report* 2004, **30**.

[22] Owen R, Barr I, Pengilley A, Liu C, Paterson B: **Annual report of the national Influenza surveillance scheme, 2007**. *Annual Reports* 2008.

[23] Hay AJ, Gregory V, Douglas AR, Lin YP: **The evolution of human influenza viruses**. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2001, **356**:1861–1870.

[24] Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC: **The genomic and epidemiological dynamics of human influenza A virus**. *Nature* 2008, **453**:615–619.

[25] Sullivan SJ, Jacobson RM, Dowdle WR, Poland GA: **2009 H1N1 influenza.** *Mayo Clinic Proceedings* 2010, **85**:64–76.

[26] Liu S, Ji K, Chen J, Tai D, Jiang W, Hou G, Chen J, Li J, Huang B: **Panorama phylogenetic diversity and distribution of type A Influenza virus**. *PLoS ONE* 2009, **4**(3):e5022.

[27] Bragstad K, Nielsen LP, Fomsgaard A: **The evolution of human influenza A viruses from 1999 to 2006 - a complete genome study**. *Virology Journal* 2008, **5**(40).

[28] Liao YC, Lee MS, Ko CY, Hsiung CA: **Bioinformatics models for predicting antigenic variants of influenza A/H3N2 virus**. *Bioinformatics* 2007, **24**(4):505–512.

[29] Worobey M, Holmes EC: **Evolutionary aspects of recombination in RNA viruses**. *Journal of General Virology* 1999, **80**:2535–2543.

[30] Kingsford C, Nagarajan N, Salzberg S, Dov J: **Swine-origin Influenza A (H1N1) resembles previous influenza isolates**. *PLoS ONE* 2009, **4**(7):e6402.

[31] He CQ, Han GZ, Wang D, Liu W, Li GR, Liu XP, Ding NZ: **Homologous recombination evidence in human and swine influenza A viruses**. *Virology* 2008, **380**:12–20.

[32] Gibbs M, Armstrong J, Gibbs A: **Recombination in the hemagglutinin gene of the 1918 "Spanish Flu"**. *Science* 2001, **293**:1842–1845.

[33] Robertson D, Gibbs M, Armstrong J, Gibbs A: **Questioning the evidence for genetic recombination in the 1918 "Spanish Flu" virus**. *Science* 2002, **296**:211a.

[34] Boni M, Zhou Y, Taubenberger J, Holmes E: **Homologous recombination is very rare or absent in human Influenza A virus.** *Journal of Virology* 2008, **82**(10):4807–4811.

[35] Boni M, Posada D, Feldman M: **An exact nonparametric method for inferring mosaic structure in sequence triplets**. *Genetics* 2007, **176**:1035–1047.

[36] Holmes E: **Error thresholds and the constraints to RNA virus evolution**. *Trends in Microbiology* 2003, **11**(12):543–546.

[37] Boni M: **Vaccination and antigenic drift in influenza**. *Vaccine* 2008, **26**(Supplement 3):C8–C14.

[38] Salzberg S: **The contents of the syringe**. *Nature* 2008, **454**:160–161.

[39] Plotkin JB, Dushoff J, Levin SA: **Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus**. *PNAS* 2002, **99**(9):6263–6268.

[40] Smith DJ: **Mapping the antigenic and genetic evolution of Influenza virus**. *Science* 2004, **305**(5682):371–376.

[41] Saks M: **Was this a bad flu season or what?** *Emergency Medicine News* 2008, **30**(7):14.

[42] Holmes EC, Ghedin E, Miller N, Taylor J, Bao Y, St George K, Grenfell BT, Salzberg SL, Fraser CM, Lipman DJ, Taubenberger JK: **Whole-genome analysis of human Influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses**. *PLoS Biology* 2005, **3**(9):e300.

[43] Bush R, Bender C, Subbarao K, Cox N, Fitch W: **Predicting the evolution of human Influenza A**. *Science* 1999, **286**:1921–1925.

[44] Koelle K, Kamradt M, Pascual M: **Understanding the dynamics of rapidly evolving pathogens through modeling the tempo of antigenic change: Influenza as a case study**. *Epidemics* 2009, **1**:129–137.

[45] Gog J, Grenfell B: **Dynamics and selection of many-strain pathogens**. *PNAS* 2002, **99**(26):17209–17214.

[46] Heiny A, Miotto O, Srinivasan K, Khan A: **Evolutionarily conserved protein sequences of Influenza A viruses, avian and human, as vaccine targets**. *PLoS ONE* 2007, **11**:e1190.

[47] Pagel M, Meade A: **Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo**. *The American Naturalist* 2006, **167**(6):808–825.

[48] Liu L, Pearl DK: **Species trees from gene trees: Reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions**. *Systematic Biology* 2007, **56**(3):504–514.

[49] Liu L, Pearl DK, Brumfield RT, Edwards SV: **Estimating species trees using multiple-allele DNA sequence data**. *Evolution* 2008, **62**(8):2080–2091.

[50] Wilson D, McVean G: **Estimating diversifying selection and functional constraint in the presence of recombination**. *Genetics* 2006, **172**:1411–1425.

[51] Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood**. *Molecular Biology and Evolution* 2007, **24**(8):1586–91.

[52] Goldman N, Yang Z: **A codon-based model of nucleotide substitution for protein-coding DNA sequences**. *Molecular Biology and Evolution* 1994, **11**(5):725–736.

[53] Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees**. *Molecular Biology and Evolution* 1987, **4**(4):406–425.

[54] Bruno WJ, Socci ND, Halpern AL: **Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction**. *Molecular Biology and Evolution* 2000, **17**:189–197.

[55] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *Journal of Molecular Biology* 1990, **215**(3):403–410.

[56] Bao Y, Bolotov P, Dernovoy D, Kiryutin B: **The Influenza virus resource at the National Center for Biotechnology Information**. *Journal of Virology* 2008, **82**(2):596–601.

[57] Edgar R: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity**. *BMC Bioinformatics* 2004, **5**:113.

[58] Swofford D: *PAUP Phylogenetic Analysis Using Parsimony (Version 4)* 2003.

[59] Schloss P, Handelsman J: **Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness**. *Applied Environmental Microbiology* 2005, **71**(3):1501–1506.

[60] Sokal RR, Rohlf FJ: *Biometry: the principles and practice of statistics in biological research*. New York: W.H. Freeman, 3rd edition 1995.

[61] Zhang GL, Khan AM, Srinivasan KN, August JT, Brusic V: **MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides**. *Nucleic Acids Research* 2005, **33**(Web Server issue):W172–179.

[62] Gribskov M, Robinson NL: **Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching**. *Computers & Chemistry* 1996, **20**:25–33.

[63] Sonego P, Kocsor A, Pongor S: **ROC analysis: applications to the classification of biological sequences and 3D structures**. *Briefings in Bioinformatics* 2008, **9**(3):198–209.

[64] Abdussamad J, Aris-Brosou S: **The nonadaptive nature of the H1N1 2009 Swine Flu pandemic contrasts with the adaptive facilitation of transmission to a new host**. *BMC Evolutionary Biology* 2011, **11**(6).

[65] Ferguson NM, Anderson RM: **Predicting evolutionary change in the influenza A virus**. *Nature Medicine* 2002, **8**(6):562–563.

[66] Loose C, Jensen K, Rigoutsos I, Stephanopoulos G: **A linguistic model for the rational design of antimicrobial peptides**. *Nature* 2006, **443**:867–869.

[67] Ferguson NM, Cummings DAT, Cauchemez S, Fraser C, Riley S, Meeyai A, Iamsirithaworn S, Burke DS: **Strategies for containing an emerging influenza pandemic in Southeast Asia**. *Nature* 2005, **437**:209–214.

[68] Ferguson N, Galvani A, Bush R: **Ecological and immunological determinants of influenza evolution**. *Nature* 2003, **422**:428–433.

[69] Wallace R, HoDac H, Lathrop R, Fitch W: **A statistical phylogeography of influenza A H5N1**. *PNAS* 2007, **104**(11):4473–4478.

[70] Parks DH, Porter M, Churcher S, Wang S, Blouin C, Whalley J, Brooks S, Beiko RG: **GenGIS: A geospatial information system for genomic data**. *Genome Research* 2009, **19**(10):1896–1904.

[71] Janies DA, Treseder T, Alexandrov B, Habib F, Chen J, Ferreira R, Catalyürek U, Varón A, Wheeler WC: **The Supramap project: linking pathogen genomes with geography to fight emergent infectious diseases**. *Cladistics* 2011, **27**:61–66.