Analyzing Evolution of Avian Influenza using detailed Genotypic and Antigenic Models and Phylodynamic Simulation

Liam Mosley[1] and Dhananjai M. Rao[1]

[1]Miami University, Oxford, OHIO 45056. USA.
raodm@miamiOH.edu

Abstract

Avian Influenza Viruses (AIV), specifically H5N1, are highly adaptive and mutate continuously throughout their life-cycle. The accumulation of constant mutations causes antigenic drift, leading to the spread of epidemics which result in billions of dollars in socioeconomic losses each year. Consequently, the containment of AIV epidemics is of vital importance. Computational approaches to the study of epidemiology, such as phylodynamic simulations, enhance in vivo analysis by examining the impact of ecological parameters and evolutionary traits, as well as forecasting the rise of future variants. We propose an improvement on existing phylodynamic simulation models through the introduction of: ❶ actual Hemagglutinin (HA) protein sequences, ❷ simulating mutations, ❸ and implementing an amino-acid level antigenic analysis algorithm to model natural selection pressure. In contrast to prior approaches that use abstract antigenic models, our method uses and yields actual HA strains enabling robust validation and direct application of results to inform vaccine design. We assess the validity of our method against the current World Health Organization (WHO) H5N1 nomenclature phylogram for 3 countries. Our calibration and validation experiments use $>$ 10,000 simulations with 1,000s of different parameter settings requiring over 2,500 hours of computing time. Our results show that our calibrated models yield the expected evolutionary characteristics but with a compromise of ~10× longer simulation times.

Introduction

Avian Influenza Viruses (AIVs), specifically H5N1 serotype, cause billions of dollars of socio-economic losses every year. Endemic in multiple species of waterfowl, H5N1 transmits both directly between hosts as well as indirectly via environmental contamination. Influenza strains that fall under the subtype H5N1 are able to spread to poultry, in turn causing widespread devastation. One of the more prominent examples of its impact was between the years of 2014-2015 where over 45 million chickens and turkeys were culled in order to stop the spread of a major epidemic (Giridharan and Rao, 2016).

AIV epidemics are perpetuated by continuous change to the nucleotide structure of the protein haemagglutinin (HA) which defines the receptor shape on the surface of influenza viruses. Small changes to the protein structure are introduced over time, accumulating into larger changes that drastically morph the shape of the receptors on AIVs. The accumulation of mutations in phylogenetic code is called antigenic drift, and is the primary source of epidemics.

Challenges with current *in vivo* methods

There are a variety of approaches used in the containment of AIVs such as livestock isolation, vaccination of at-risk populations, and culling of infected hosts. Vaccinations are the primary method used to prevent epidemics, allowing recipients to gain immunity against the most common influenza strains in their region (WHO, 2012). Vaccine design in respect to in vivo analysis involves the collection of viral data from infected hosts, tracking host migration patterns, and sequencing collected viral data in order to make informed decisions on the prevalence of different AIV strains. This process can be lengthy in regards to the evolutionary time line of AIV epidemics, requiring 10 to 18 months to finalize analysis. The demand for new strain selection dictates vaccine candidates to be identified every 6 to 8 months. Another drawback to this methodology is that the analysis is reactionary and does not allow analysts to predict future epidemics. Limitations also arise due to the spatial and temporal locality of surveillance and sampling. Hence, in vivo analysis is ineffective alone when it comes to informing H5N1 containment efforts.

In silico approaches & shortcomings

Computational analysis methods enhance in vivo efforts by providing a platform for predictive modeling with results delivered in the span of days or weeks. Of particular interest are phylodynamic simulations (discussed in detail in Section Background and related works) which enable exploration of the effects of selection pressure, ecological parameters, and regional factors on the spread of viruses to forecast future epidemics. Future forecasts are of particular importance as they can be used to inform time lines for design of new vaccines and validate containment measures (Bedford et al., 2012; Giridharan and Rao, 2016; Volz et al., 2013).

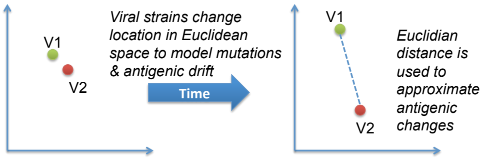

V1
V2
Viral strains change location in Euclidean space to model mutations & antigenic drift
Time
V1
V2
Euclidian distance is used to approximate antigenic changes

Figure 1: Overview of the current state-of-the-art antigenic modeling approach

The evolutionary and antigenic models used in phylodynamic simulations play a critical role in enabling effective modeling and analysis. Current state-of-the-art phylodynamic simulations merely use an abstract 2-dimensional or multidimensional space to represent evolutionary changes as summarized in Figure 1. Mutations in viral strains are modeled by changing their associated coordinate values. The euclidean distances are then used to approximate phylogenetic differences and their ensuing antigenic differences.

Shortcomings of state-of-the-art: The contemporary genetic and antigenic modeling approaches shown in Figure 1 have several shortcomings, including:

1. They do not model the actual viral strains. Consequently, mutations are grossly approximated.
2. Due to abstract nature, the mutation rates in the simulation do not directly reflect mutation rates observed in nature, such as those reported by Dang et al. (2010).
3. The antigenic characteristics are also approximated. Consequently, all mutations are deemed equally significant, a stark contrast to actual antigenic behaviors that primarily arise from mutations to epitope regions.
4. Since the phylodynamic simulations cannot output actual viral strains, forecasting to inform vaccine design is not straightforward.

The aforementioned shortcomings significantly limit the applicability of current phylodynamic methods.

Proposed enhancements: Our contributions

In this study we propose and assess an alternative antigenic model. It addresses the aforementioned shortcomings of the Euclidean model (shown in Figure 1) via the following three improvements, namely:

1. We propose to use actual HA sequence(s) instead of the abstract Euclidean model, starting with the root HA sequence (`A/turkey/England/5092/1991`) corresponding to the root of the WHO H5N1 nomenclature (WHO, 2012).
2. We simulate realistic mutations based on observed mutation rates in nature as reported by Dang et al. (2010). However, the mutation rates are further calibrated to characterize phylogenetic diversity in a given region.
3. Antigenic diversity is measured using an amino-acid level comparison algorithm, called P-Epitope proposed by Gupta et al. (2006).

This paper presents a detailed overview of our proposed enhancements in Methods section. Experiments & Validation section presents results from experiments conducted to calibrate and verify our model enhancements. In addition, results from sensitivity analysis are also discussed to identify influential parameters in the model. Conclusions presents concluding remarks along with a summary of our envisioned future work.

Background and related works

Phylodynamic models are used to characterize the epidemiological and evolutionary characteristics of viruses (Volz et al., 2013). Computational phylodynamic simulations typically use agent-based models in conjunction with discrete time simulation. Simulations enable analysis of the interplay between ecological processes and viral phylogenies. Figure 2 represents an abstract view of the ecological process that our phylodynamic simulations recreate. Waterfowl hosts are seeded with an initial viral strain and the virus then begins to mutate. For up to 8 days the virus will be shed from infected individuals, with the potential to infect not only other waterfowl but also the environment the host has contact with (Wibawa et al., 2014). Water sources are particularly vulnerable and can harbor infections for up to 20 days (Roche et al., 2014). Host immunity prevents hosts from acquiring a new infection if the virus strain is antigenically similar to a recent previous infection. The mutations that occur within individual hosts accumulate over time, causing antigenic drift. Antigenic drift causes new viral strains to diverge from their ancestral lineage, enabling them to escape host immunity and cause new infections. Figure 2 summarizes this process and exemplifies how new virus lineages diverge into new clades, or groupings, of viruses.

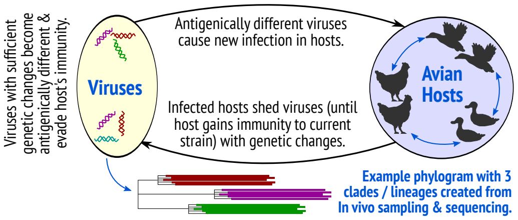

Viruses with sufficient genetic changes become antigenically different & evade host's immunity.
Viruses
Antigenically different viruses cause new infection in hosts.
Infected hosts shed viruses (until host gains immunity to current strain) with genetic changes.
Avian Hosts
Example phylogram with 3 clades / lineages created from In vivo sampling & sequencing.

Figure 2: Ecological model of the influenza life cycle

The current leading practices for phylodynamic modeling were introduced by Gog and Grenfell, who utilize the classical model for epidemiological studies (Gog and Grenfell, 2002). This model is based on classic compartmental epidemic models, such as: susceptible (S), exposed (E), infective (I), and recovered (R). However, for avian influenza which is endemic in waterfowl, we use a S-I-S model. Hosts that are susceptible have not been in contact with a specific viral strain but can potentially be infected. A viral strain can infect a host only if an antigenically similar strain is not present in the immune history of a host. A host remains in the infective category until the infection has run its course. While in the infective compartment, the host can spread the virus to susceptible hosts that it comes into contact with. In addition, an infective host also contaminates its environment by shedding the virus. After the host acquires immunity against a viral strain they transition back to the susceptible state and the cycle repeats.

PhySim: A phylodynamic simulator

PhySim is the computational implementation of the ecology of avian influenza shown in Figure 2. PhySim is an adaptation of a general simulation tool named Antigen (Bedford et al., 2012). As discussed by Giridharan et al. (Giridharan and Rao, 2016), PhySim's enhancements include: ① enabling simulation of avian influenza strains, ② simulation of multiple species with different birth and death rates; ③ births occurring only during specific brooding seasons rather than throughout the year; ④ genetic and antigenic properties of viruses are independently modeled; ⑤ antigenic distances between simulated HA strains are estimated using the cross-immunity approach proposed by Gog et al.; ⑥ phylogenetic trees are constructed based on genetic differences rather than difference in emergence times; and ⑦ infection rates and infective periods account for seasonal variations in the countries.

Written in Java, PhySim uses Gillespie's Stochastic Simulation Algorithm (SSA) with Tau-Leap optimization. In order to simulate epidemic progression with sufficient accuracy, PhySim uses a time step of 0.1. PhySim uses an an Individual-Based Model (IBM) for modeling epidemic progression. PhySim only moves hosts between the susceptible and infective compartments to form an S-I-S model where the exposure of infections is simulated by probability at the point of contact and hosts move directly to infective, once the infection dies off in the host they are returned to the susceptible compartment. An S-I-S model is used due to the endemic nature of H5N1 in waterfowl, the Exposed compartment is modeled as a transition from S to I as infections immediately take hold in the host. Hosts become immediately susceptible to new infections after recovery as the R compartment becomes the transition from I back to S. Figure 3 represents a broadened view of how the S-I-S and ecological models for our simulation interact.

We are able to analyze the impact of a variety of parameters summarized in Table 1. Hosts represent a group of waterfowl from a specific species, where multiple species can be present in each simulation. We can target specific countries and model the spread and mutation of an influenza virus for that country by setting parameter values specific for the waterfowl found in the region. Nigeria and Turkey were identified as high risk countries using a combination of Phylodynamic and Phylogeographic analysis and will be the focus of our attention (Giridharan and Rao, 2016). A point of interest in our model is that we take seasonal temperature fluctuation into consideration when determining transmission of potential infections. By using a sinusoidal curve as a modulation factor we can increase the chance of infection in colder seasons and decrease the chance of transmission during warmer months.

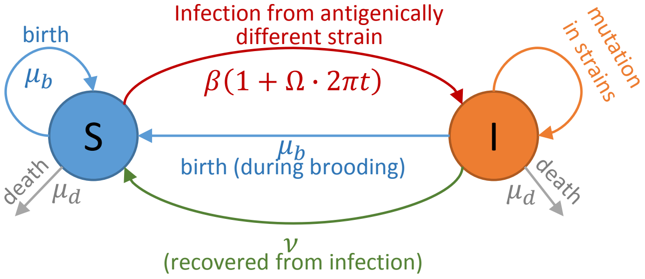

birth
μ_b
Infection from antigenically different strain
$\beta(1+\Omega\cdot 2\pi t)$
mutation in strains
S
I
μ_b
birth (during brooding)
death
μ_d
death
μ_d
ν
(recovered from infection)

Figure 3: Ecological and SIS Modeled with example parameters from PhySim

μ_b	Species specific daily birth rate during brooding season
μ_d	Species specific daily death rate derived from lifespan
β	Contact rate (direct) between hosts for a region
Ω	Sinusoidal seasonal modulation factor
ψ	Average daily phenotypic mutation rate
ν	Inverse of infectuous period

Table 1: A sample of PhySim ecological parameters for a multi-species simulation model

The models for phylodynamic simulations can be validated by setting ecological parameters such as those in table 1 to produce phylograms that mirror those constructed from in vivo analysis. By examining the parameters that produce the best matching phylograms we are able to deduce what factors play the most impact on the spread of AIVs and can succinctly inform vaccine design. Simulations are run with a burn-in period of 15 years to simulate the time leading up to current day, by stepping past the burn-in period simulations are able to then effectively predict what the evolutionary landscape will look like in the future. Parameters such as contact rate can be abstractly represented as livestock isolation, and features such as average mutation rate can be mapped to vaccination efforts.

Comparisons with recent related works

Current phylodynamic simulations (Bedford et al., 2012; Giridharan and Rao, 2016) model AIV strains as abstract 2D-vectors representing evolutionary data. Euclidean distance between two vectors represents phylogenetic distance between two viral strains, we will call this the geometric approach (Gog and Grenfell, 2002). Antigenic similarity is merely approximated using Euclidean distances as summarized in Figure 1.

Our research proposes to improve upon the current modeling standards used in phylodynamic simulations by relying on a new measure of antigenic distance called P-Epitope (Gupta et al., 2006). P-Epitope has been shown to have a higher correlation to vaccine efficacy when compared to other measures of antigenic distance such as P-Sequence which is the current measure used by the WHO (Gupta et al., 2006).

In order to utilize P-Epitope as a measure of antigenic distance our representation of AIV strains in PhySim has been enhanced. We propose to implement amino acid sequencing and use an amino acid substitution model to represent competing viral strains and adjust our mutation model to reflect the current 2D-vector approach. Our work is distinguishable from recent state-of-the-art in a variety of ways. Instead of distancing ourselves from biological functions we instead embrace the computational complexity of working with protein sequence data in order to derive a simulation model that better reflects vaccine efficacy and natural mutation. The amino acid substitution model we chose to implement in our approach has been shown to be a top performer in regards to predicting future phylograms using machine learning approaches (Dang et al., 2010). This work is similar to ours in that we are also striving to predict future phylograms, but our work distinguishes itself from the machine learning approaches explore by (Dang et al., 2010) in that we are producing HA protein sequences in a more organic way through selection pressure, environmental influence, and relying on strain comparisons using a method that is more closely related to vaccine efficacy.

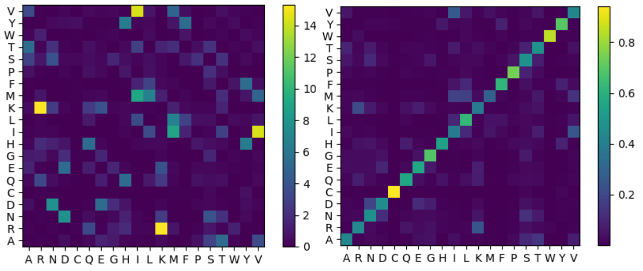
Methods

PhySim is progressed on a daily basis, and actions are controlled used a time step value (e.g. delta=0.1 means a day is divided into 10 time steps). The daily rate of contact and mutation are defined through input parameters. Simulation runs are conducted to match with WHO H5N1 nomenclature – *i.e.,* starting with 1991, with 15 years of burn-in time to produce strains for 2006–2010 for constructing phylograms.

Hosts are introduced and removed from both the susceptible and infective compartments at the same rate on a daily basis in order to maintain a stable population for each waterfowl species. The birth and death rates account for abundance and lifespans of different high-risk waterfowl species (that are endemic to a given region), including – *A.Acuta* (Northern Pintail), *A.Crecca* (Common Teal), *A.Fuligula* (Tufted Duck), *A.Penelope* (Eurasian Wigeon), *L.Canus* (Common Gull), *L.Limosa* (Black-tailed Godwit), *P.Pugnax* (Ruff), and *V.Vanellus* (Northern Lapwing).

Antigenic model enhancements

PhySim was originally adapted to simulate changes in the nucleotide structure of H5N1 HA protein sequences. One of the major assumptions of this model is that changes in protein sequences are uniform and random, this is known to not be the case under in vivo analysis and is a limitation to the original geometric model. FLUModel, an amino acid substitution matrix derived from a database of currently spreading H5N1 strains, is our solution to this limitation (Dang et al., 2010). Given a parameter value for *t* a substituion matrix can be calculated from an instantaneous rate of change matrix and steady state vector for amino acids in HA proteins. FLUModel was derived from the same set of H5N1 protein sequences that we are looking to recreate.



(a) Rate Matrix (b) Substitution Matrix

Figure 4: Using matrix exponentiation figure (a) is transformed into figure (b), this example is done using a large value for *t* to show the differences in substitution rates

Figure 4 illustrates the transformation from a rate of change matrix to a substitution matrix. Given the rate of change matrix *q* and a steady state matrix π we can get the ensuing substitution matrix *P(t)* using the following calculations:

$$q_{xy} = \pi r_{xy}, q_{xx} = -\Sigma_{x \neq y} q_{xy}$$
$$P(t) = e^{tQ}$$

A value of $t = 1.0$ represents the substitution rates for all amino acids over the course of an entire branch. The substitution rates have been shown to be exceptionally accurate for small values of t.

Algorithm's 1 and 2 exemplify how the two methods of mutation differentiate between the geometric and P-Epitope simulation models. Algorithm 1 is the current geometric substitution model, a percentage of the infective population has their associated infection mutated approximately once per day based on the time step delta value. The mutation rate ψ was computed from in vivo sequence analysis, the direction of the mutations are controlled using sin and cosine waves (Cattoli et al., 2011).

Algorithm 2 is our proposed substitution model. Its assumed that amino acid substitutions are independent and generally time reversible. That is to say in one example mutation there could be multiple amino acid substitutions that occur, and their substitution rates are mutually exclusive.

Algorithm 1 2dMutate(delta)

```
$I = \Sigma species.I$, for all species in model
$I' = poisson(I * delta)$
while $I' > 0$ do
    $i = getInfected(uniform(I))$
    $v = i.getVirus()$
    $\theta = uniform(2\pi)$
    $v.trait_X = v.trait_X + \psi * cos(\theta)$
    $v.trait_Y = v.trait_Y + \psi * sin(\theta)$
    $I' = I' - 1$
end while
```

Algorithm 2 subMatrixMutate(delta)

```
$I = \Sigma species.I$, for all species in model
$I' = poisson(I * delta)$
while $I' > 0$ do
    $i = getInfected(uniform(I))$
    $v = i.getVirus()$
    for $aminoAcid \ \epsilon \ v.HASequence$ do
        $substitute(aminoAcid, uniform(1))$
    end for
end while
```

This mutation model allows us to represent actual changes in protein structure over the course of the simulation. More importantly it enables the use of P-Epitope to measure antigenic distance between competing viral strains.

Algorithm 3 canInfect(virus v, susceptibleHost s)

```
$minRisk = 1 - homologousImmunity$
$maxRisk = homologousImmunity$
$risk = 0.0$
for $v_i \epsilon s.immuneHistory$ do
    $distance(v, v_i)$
    if $distance < risk$ then
        $risk = distance$
    end if
end for
$risk = min(maxRisk, risk)$
$infectFlag = uniform(1) < risk$
return $infectFlag$
```

The method of determining if an infection occurs in a host is described in Algorithm 3. After contact has been established the distance between the virus and every virus in the hosts immune history is calculated. If a uniform random number generated at the time of contact is less than the risk associated with the immune history then the host is infected.

It is here that we propose to use P-Epitope to determine the risk factor of a potential infection. As shown in the research conducted by Gupta et al. there are various examples of past vaccination regimes failing due to strain selection relying on P-Sequence (Gupta et al., 2006). Had P-Epitope been used to compare strains for vaccine selection a more successful vaccination regime could have been promoted in many of the examples cited. Gupta et al. showed that there is a higher correlation between P-Epitope and vaccine efficacy than other measures of antigenic distance when examining past vaccination regimes.

Algorithm 4 $pEpitope(virus \ \ v1, virus \ \ v2)$

```
$pEpitope = 0$
for $epitope \ \epsilon \ epitopeRegions$ do
    $localDifference = 0$
    for $residue \ \epsilon \ epitope$ do
        if $v1[residue] \neq v2[residue]$ then
            $localDifference = localDifference + 1$
        end if
    end for
    $difference = localDifference/epitope.size$
    if $difference > pEpitope$ then
        $pEpitope = difference$
    end if
end for
return $pEpitope * pEpConv$
```

P-Epitope is described in detail in Algorithm 4. The epitope regions to be compared can be fed into PhySim as a parameter, there are five major epitope regions in HA protein sequences A, B, C, D, & E where a recent survey identified which residues can be attributed to respective epitopes (Peng et al., 2014). A scalar value is attached to P-Epitope to obtain a parabolic risk function. The advantage of calculating antigenic distance using P-Epitope is that we are able to identify antigenic similarity between virus strains that other measures of antigenic distance would overlook. Figure 5 illustrates the area of overlap that other measures, such as P-Sequence, are unable to detect. As seen in the Figure there is significant overlap between the frequency distributions when using P-Epitope to compare strains from the same clade and strains from different clades which were grouped using P-Sequence. These are similarities that measures such as P-Sequence can not detect.

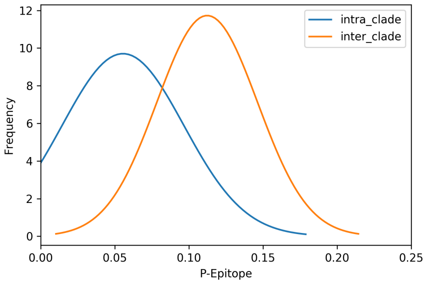
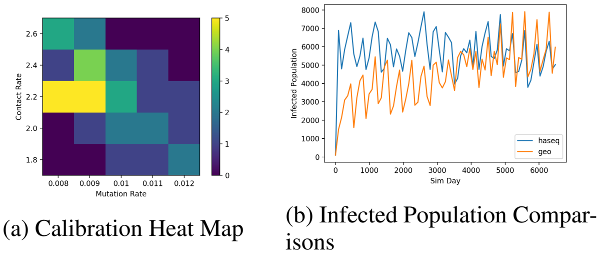


Figure 5: A sample of 100 H5N1 HA protein sequences from 10 different clades were compared using P-Epitope, frequency of P-Epitope values for inter and intra -clade distances is plotted

Experiments & Validation

We calibrated the model with respect to the following simulation outputs: ① Infective and susceptible populations for each time step, ② Number of clades in resulting phylogram, ③ antigenic diversity. The outputs of the simulation using the new model were compared against the previously calibrated geometric model.



(a) Calibration Heat Map

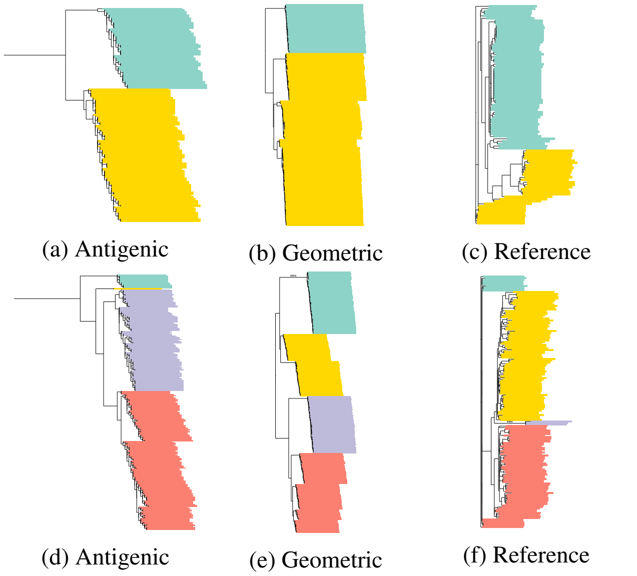
(b) Infected Population Comparisons

Figure 6: A window of the calibration efforts for PhySim with the antigenic model implemented

The starting point of the calibration effort consisted of estimating the average mutation rate for the new model based on the average nucleotide mutation rate used in the geometric model. The expected amino acid substitution rates were approximated to within 1% of the nucleotide substitution rates. This was done by adjusting the *t* parameter value for the substitution matrix, generating probabilities for each amino acid substitution and multiplying by the average number of each amino acid in a typical HA protein sequence.

Figure 6 illustrates an example of how the simulation model was validated for Turkey. The majority of parameter settings were kept consistent between the geometric model and our new model. In order to properly calibrate and test the new model only the mutation and contact rates were adjusted. Subfigure 6a shows a narrow window of calibration settings that produced consistent results, a more exhaustive calibration effort was conducted to find this window that required hundreds of different parameter combinations ranging the contact rate from 1.0 to 3.0 and the mutation rate from 0.002 to 0.10 in varying step sizes. The window illustrated represents the success rate of parameter combinations with a contact rate between 1.8 and 2.6 in steps of 0.2, and mutation rates between 0.008 and 0.012 in steps of 0.001.

Each simulation was seeded with slightly mutated root sequence variants equivalent to the number of initial infected individuals. Due to the discrete steps mutations take in the new model the initial propagation period is susceptible to low mutation rates, and can cause the number of infected individuals to zero out early. This was combated by spawning the slight variants, ensuring what is the equivalence of 100 simulation days of mutations. This results in the initial spike of infective individuals in subfigure 6b.



(a) Antigenic (b) Geometric (c) Reference

(d) Antigenic (e) Geometric (f) Reference

Figure 7: Antigenic phylograms were produced using our enhanced model (sucess is 2 clades for Turkey, 4 for Nigeria). Subfigures(a-c): Turkey. Subfigures(d-f): Nigeria

Parameter analyses

In this study we have used Generalized Sensitivity Analysis (GSA) (Guven and Howard, 2007) to assess the influence of parameter settings in our model. GSA is based on a two-sample Kolmogorov-Smirnov Test (KS-Test) and yields a $d_{m,n}$ statistic that is sensitive to differences in both central tendency and any differences in the distribution functions of parameters. The $d_{m,n}$ statistic ($0 \leq d_{m,n} \leq 1.0$) is the maximum separation between cumulative probability distributions observed in a two-sample KS-Test. The $d_{m,n}$ statistic is computed for each parameter by varying its value over a ±25% range, in steps of 10%, around its calibrated setting as shown in Figure 8. At each setting, 10 stochastic simulations are conducted and the number of successful (*i.e.,* simulation produces phylogram with same number of clades as reference *in vivo* phylogram) and unsuccessful outcomes are recorded. We have used the model for Turkey to conduct the sensitivity analysis.

The data is used to compute the cumulative probability of success and failure for each parameter as shown in Figure 8. The maximum difference between the cumulative success and failure probabilities is the $d_{m,n}$ statistic shown in red for each parameter. For example, from Figure 8, the $d_{m,n}$ statistic for `Contact` rate (β) is 0.256. Figure 9 shows a summary comparison of the influence of the parameters. The lightly shaded bands show the 95% Confidence Intervals (CI) computed using standard bootstrap approach using 5000 replications with 1000 samples in each.

As illustrated by the GSA $d_{m,n}$ statistic values, the following parameters do not have a strong influence on the model's characteristics – the initial population of birds (N),

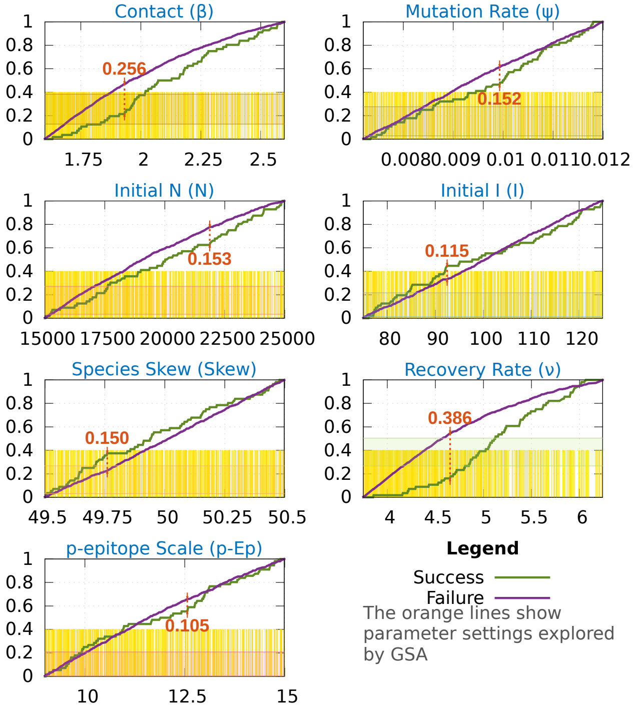


Figure 8: Detailed results from Generalized Sensitivity Analysis (GSA). The x-axis in each sub-chart indicates range of values for each parameter. In all sub-charts the y-axis is the $d_{m,n}$ statistic.

initial number of infected birds (I), variance in the abundance of different species (Skew), and the antigenic scaling parameter used with p-epitope. In other words, assumptions made about the values of these parameters do not have a significant impact on the validity and outcome of our analyses. Insensitivity to these assumed parameter values is an important aspect of our model. It enables us to draw inferences with sufficient confidence without requiring to have a good estimate of waterfowl populations, waterfowl species abundance, initial infections etc.

On the other hand, the most influential factors that primarily drive diversity of viral strains are: recovery rate (ν) at 0.386, contact rate (β) at 0.256, and mutation rate at 0.152. The recovery rate for H5N1 has been set to the putative value of 5 days. Accordingly, the two key parameters whose values have been determined via calibration are β and ψ, which are specific to each region being analyzed. These three influential parameters are also the primary targets for containment and prophylaxis efforts.

Correlation analysis: Consistent with interrelationships in nature, the parameters in the model have inherent correlations as illustrated by the correlogram in Figure 10. The correlogram has been plotted using results from successful configurations, *i.e.,* parameter settings that yield the correct number of clades, *i.e.,*, the same number of clades as in the reference phylogram. The corellogram has been plotted using R and the PerformanceAnalytics package.

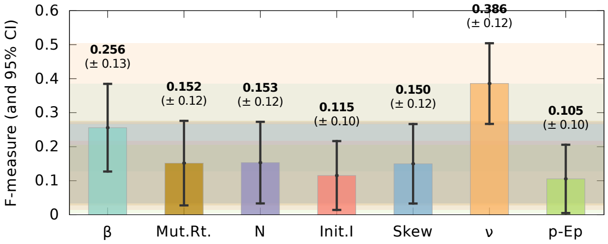


Figure 9: Parameter comparisons based on GSA

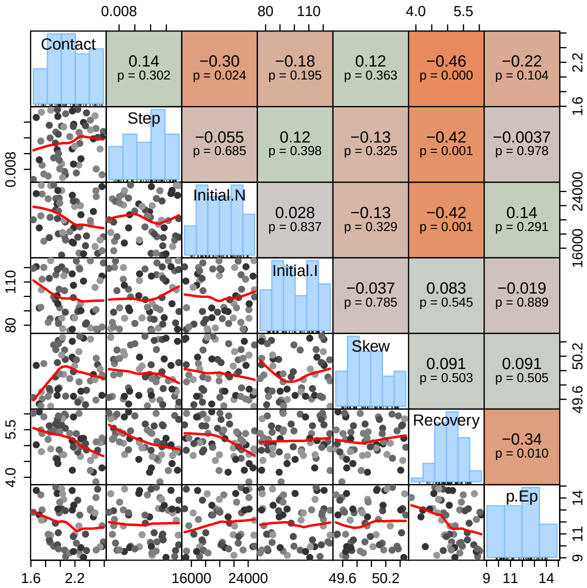


Figure 10: Correlation between parameters elicited by GSA

The correlogram shows that the recovery rate ($1/\nu$) is strongly, negatively correlated to contact rate (β), mutation rate (ψ), p-epitope scale (p-Ep), and initial population (N). These negative correlation are expected because of the nature of these parameters. For example, decreasing mutation rates (ψ) but increasing recovery time (ν) essentially maintains the antigenic diversity. Similarly, increasing contact rate (β) enables more infections to occur and hence, even with decreasing ν, overall antigenic diversity is maintained. The correlogram shows that the recovery rate ν plays a central role in anchoring other epidemiological and ecological parameters in the model. This observation also emphasizes the need for surveillance and assay-based identification of at least one of these four parameters in emergent epidemics and other parameters can be estimated via phylodynamic simulations.

Conclusions

Vaccinations are widely used to contain and mitigate epidemics caused by antigenic variants of Avian Influenza Viruses (AIVs), including the H5N1 serotype. However, vaccines need to be regularly updated to compensate for antigenic drift in AIVs. Currently, expensive *in vivo* assays are required to regularly update vaccines to compensate for antigenic drift. Furthermore, such *in vivo* assays and analyses do not provide insight into the underlying ecological processes that is necessary to inform containment and prophylaxis strategies. Consequently, *in vivo* methods are enhanced using computational or *in silico* approaches involving phylodynamic simulations. The antigenic models used for phylodynamic simulations play a critical role in overall effectiveness of *in silico* methods. Current state-of-the-art models merely use an abstract multidimensional space to approximate both genetic and antigenic changes.

This paper proposed and evaluated a novel antigenic model which is distinguished from current research by: ❶ use of actual Hemagglutinin (HA) protein sequences, ❷ simulating mutations occurring to the HA sequence(s) and further calibrating the mutation rates to mirror ecological niches, and ❸ and implementing an amino-acid level anitgenic analysis algorithm. The paper discussed the motivation for the aforementioned enhancements and presented an algorithmic overview. The models were verified and validated using over 10,000s of simulations with 1,000s of different parameter settings and requiring over 2,500 hours of compute time. We assess the validity of our method using the current World Health Organization (WHO) H5N1 nomenclature for Turkey and Nigeria.

The influence and impact of parameters in our model has been explored via Generalized Sensitivity Analysis (GSA). Our GSA analysis showed that recovery rate (ν), contact rate (β), and mutation rate (ψ) strongly influence the antigenic diversity. Correlation analysis revealed a strong, negative correlation between recovery rate (ν) and contact rate (β), mutation rate (ψ), p-epitope scale (p-Ep), and initial population (N). This correlation emphasizes the need for surveillance and assay-based identification of at least one of these four parameters in emergent epidemics. Once a putative value for one of the parameters is identified, the other parameters can be estimated via phylodynamic simulations.

This study lays the groundwork for using detailed antigenic models in phylodynamic simulations. A key issue that we encountered was the high computational times for the simulations. Currently, we are exploring solutions to reduce the computational times.

Nevertheless, we contend that the benefits accrued from our methods offset the higher computational times. The significance of this research is that not only are we able to inform containment efforts similar to the current state-of-the-art, we also produce actual HA protein sequences that can be used in different methods of analysis in the future. As an example, with a fine-tuned model there is the possibility to explore and monitor evolutionary characteristics and niches of avian influenza viruses. Unlike analysis done using current state-of-the-art models direct connections between clusters of viruses in our simulations to real world clades can be made, and the direct impact on containment efforts on the structure of real world avian influenza viruses will be able to be examined.

References

Bedford, T., Rambaut, A., and Pascual, M. (2012). Canalization of the evolutionary trajectory of the human influenza virus. *BMC Biology*, 10(1):1–12.

Cattoli, G., Fusaroa, A., Monnea, I., Covenb, F., Joannisc, T., El-Hamidd, H. S. A., Husseine, A. A., Corneliusf, C., Amaring, N. M., Mancina, M., Holmesh, E. C., and Capuaa, I. (2011). Evidence for differing evolutionary dynamics of A/H5N1 viruses among countries applying or not applying avian influenza vaccination in poultry. *Vaccine*, 29:9368–9375.

Dang, C. C., Le, Q. S., Gascuel, O., and Le, V. S. (2010). Flu, an amino acid substitution model for influenza proteins. *BMC Evolutionary Biology*, 10(1):99.

Giridharan, N. and Rao, D. M. (2016). Eliciting characteristics of h5n1 in high-risk regions using phylogeography and phylodynamic simulations. *Computing in Science Engineering*, 18(4):11–24.

Gog, J. R. and Grenfell, B. T. (2002). Dynamics and selection of many-strain pathogens. *Proceedings of the National Academy of Sciences*, 99(26):17209–17214.

Gupta, V., Earl, D. J., and Deem, M. W. (2006). Quantifying influenza vaccine efficacy and antigenic distance. *Vaccine*, 24(18):3881 – 3888. 3rd International Conference on Vaccines for Enteric Diseases.

Guven, B. and Howard, A. (2007). Identifying the critical parameters of a cyanobacterial growth and movement model by using generalised sensitivity analysis. *Ecological Modelling*, 207(1):11 – 21.

Peng, Y., Zou, Y., Li, H., Li, K., and Jiang, T. (2014). Inferring the antigenic epitopes for highly pathogenic avian influenza h5n1 viruses. *Vaccine*, 32(6):671 – 676.

Roche, B., Drake, J. M., Brown, J., Stallknecht, D. E., Bedford, T., and Rohani, P. (2014). Adaptive evolution and environmental durability jointly structure phylodynamic patterns in avian influenza viruses. *PLoS Biol*, 12(8):e1001931.

Volz, E. M., Koelle, K., and Bedford, T. (2013). Viral phylodynamics. *PLoS Computational Biology*, 9(3):e1002947.

WHO (2012). Continued evolution of highly pathogenic avian influenza A (H5N1): updated nomenclature. *Influenza and Other Respiratory Viruses*, 6(1):1–5.

Wibawa, H., Bingham, J., Nuradji, H., Lowther, S., Payne, J., Harper, J., Junaidi, A., Middleton, D., and Meers, J. (2014). Experimentally infected domestic ducks show efficient transmission of indonesian h5n1 highly pathogenic avian influenza virus, but lack persistent viral shedding. *PLoS ONE*, 9:e383417.