Eliciting Characteristics of H5N1 in High-Risk Regions Using Phylogeography and Phylodynamic Simulations

Neil Giridharan | William Mason High School, Mason, Ohio
Dhananjai M. Rao | Miami University, Oxford, Ohio

To design vaccines and mitigate epidemics, evolutionary characteristics of avian influenza viruses are typically studied through longitudinal surveillance and serological assays. However, such in vivo analysis is often reactive and limited due to the complexities and costs of long-term, multinational surveillance. A novel in silico approach combines two different agent-based simulation methods to help inform vaccine design.

Avian influenza is caused by several serotypes of the influenza A viruses (including the dominant H5N1 strain) that are endemic in migratory waterfowl, the natural intercontinental vectors of these viruses.[1] The avian influenza virus (AIV) transmits to poultry via contaminated water and feed, causing widespread mortality and resulting in severe economic losses, including the 2015 US epidemic. AIVs cause recurring epidemics because they undergo continuous change in the haemmagglutinin (HA) surface protein. Changes to HA cause antigenic drift, which enables new strains to escape host immunity, ultimately causing new infections. Infected hosts shed viruses with changes to HA, giving rise to more diverse strains. As Figure 1 illustrates, the cycle continues with the establishment of new viral lineages; strain prevalence varies both temporally and spatially throughout the world.

Vaccination is the most prevalent prophylactic method for containing avian influenza epidemics.[1] Vaccines are designed from several ancestors of the prevalent viral strains circulating in the host population.

 Copublished by the IEEE CS and the AIP 1521-9615/16/$33.00 © 2016 IEEE

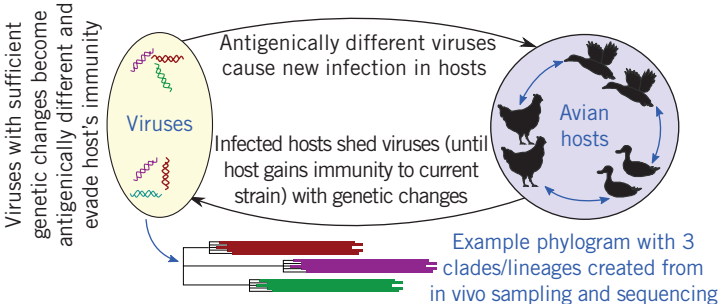

Figure 1. Overview of the genetic and antigenic diversity in avian influenza viruses (AIVs) that arise due to repeated epidemics. Prevalent strains are identified in vivo through surveillance, sampling, sequencing, and phylogenetic analysis.

These prevalent strains are identified in vivo through surveillance, sampling, sequencing, and phylogenetic analysis. As of December 2014, the World Health Organization (WHO) recommends 27 different H5N1 vaccine candidate strains for different parts of the world to immunize against prevalent strains from 14 different clades in the WHO H5N1 nomenclature phylogram.[1]

However, continuous antigenic drift in viruses requires candidate strains and vaccines to be updated every six to eight months,[2] posing significant national and international challenges to surveillance, sequencing, and analysis efforts. Moreover, the in vivo vaccine candidate strain selection method is reactive and typically lags an outbreak by 10 to 18 months, degrading vaccine efficacy and impeding effective containment of emergent epidemics. Furthermore, in vivo sampling and analysis doesn't provide a comprehensive understanding of the ecological factors influencing evolutionary processes and epidemic progressions, thereby hindering design and administration of prophylaxis or containment strategies.

These aforementioned shortcomings of the traditional in vivo approach can be alleviated by enhancing them using in silico or computational methods, thereby informing vaccine design for rapidly evolving lineages to improve vaccine efficacy, using antigenic drifts and environmental factors to determine frequency of vaccine redesign in different geographic regions, identifying high-risk regions for antigenic shift and increased surveillance, and identifying influential ecological and epidemiological factors to mitigate their impacts and contain epidemics. Specifically, our investigation proposes a novel multidisciplinary approach that combines two different agent-based simulation methods, namely, epidemiological simulation and phylogeographic annotations to identify high-risk countries, and phylodynamic simulations to elicit AIVs evolutionary characteristics in high-risk countries.

Background

Computational epidemiology integrates computer science and classical surveillance-based epidemiology to enable comprehensive understanding of diseases, epidemic forecasting, and administration of prophylactic strategies. It's fundamentally based on modeling and simulation (M&S) with agent-based, discrete modeling methods being widely used. An agent models epidemic progression in a single individual or a collection of individuals along with pertinent ecological processes and interactions. The de facto standard for modeling epidemics is the compartmental model, in which the population is subdivided into independent subsets or compartments based on their epidemiological states, such as susceptible (*S*), exposed (*E*), infective (*I*), and recovered (*R*), or *SEIR*. Disease progression is modeled via probabilistic transitions between the compartments.

SEARUMS, the M&S environment used in this study, utilizes agent-based modeling and parallel discrete event simulation (PDES) to enable rapid epidemiological analysis.[3,4] The agents and migratory flyways are generated by using surveillance data available in GIS format. The agents in the SEARUMS model the migratory life cycle of a flock of waterfowl of the same species and implement the *SEIR* compartmental model to characterize epidemic progression. Agents interact with each other when flocks logically overlap to propagate epidemics during and after migration. Aggregation of many birds into a single agent has been adopted to reduce computational resources and simulation runtime, which can be significant for larger models even when PDES techniques are utilized.[4] The waterfowl model used in this study has been generated, verified, and validated as discussed in our earlier publications[3,5] and supplements. This investigation utilizes a validated model from our earlier investigations[5] to identify highly connected countries involved in the spread of H5N1 and uses phylodynamics to elicit evolutionary characteristics of H5N1 viruses in these high-risk countries.

Phylodynamics is a special case of computational epidemiology that combines phylogenetics with epidemiological modeling and simulation to simultaneously analyze the phenotypic and genotypic evolution (see Figure 2).[6] Specifically, agents model both epidemic progression as well as abstract

changes occurring in viral genotypes, which are sampled during simulation to construct a phylogenetic tree or phylogram (see Figure 2). Ecological parameters in the simulation are set such that the in vivo and in silico phylograms are similar, thereby validating the phylodynamic model. The calibrated model provides information about unobservable or unknown in vivo ecological parameters. The model is also used to analyze the effect of influencing ecological parameters via vaccination or other prophylactic strategies via varying parameter settings. Extending the simulation time to logically simulate into the future provides forecasts about anticipated evolutionary changes to assist planning.

As summarized by Erik Volz and colleagues,[6] Bryan Grenfell and coauthors[6] postulated the theory and general approach for phylodynamic modeling, whereas our research focuses on its application to AIVs. Neil Ferguson and colleagues[6] as well as Katia Koelle and coauthors[6] discuss the use of phylodynamic simulations and sensitivity analysis to identify the role of ecological factors in the spread of influenza A infections, particularly in humans. Trevor Bedford and colleagues[7] use a phylodynamic model of influenza A evolution to show that few antigenic dimensions are sufficient to account for the paradoxically limited diversity of H3N2 human influenza strains, whereas Benjamin Roche and coauthors[8] discuss extensive validation of their individual-based phylodynamic model. In a recent article,[9] they present the use of phylodynamic analyses to assess the impact of antigenic, epidemiological, and ecological factors to explain higher genetic diversity and weaker immune escape in avian influenza viruses when compared to human strains.

Unlike earlier investigations[6–9] that focus on or involve human influenza strains, our research focuses on avian influenza, specifically H5N1. This investigation distinguishes itself from prior work, including our own,[3,5] by combining phylodynamics with results from phylogeography. This research utilizes a validated model of migratory waterfowl to identify key countries with many direct infection pathways for H5N1 to other countries. Because highly connected countries influence global diversity of viruses, they're deemed high-risk regions; this study uses phylodynamic simulation to quantify the ecological characteristics underlying the evolution of H5N1 strains in these high-risk countries. The phylodynamic modeling and simulation extends validated models proposed by Bedford and colleagues[7] by explicitly modeling multiple high-risk waterfowl species native to each country. The results from phylogeographic and phylodynamic analyses are then used to infer epidemiological, ecological, and temporogeospatial characteristics to inform vaccine design and prophylaxis in each country.

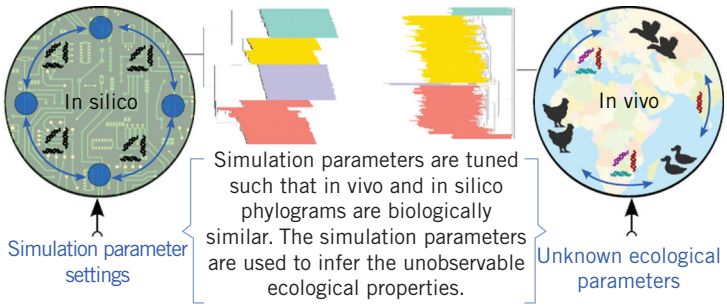

Figure 2. Overview of using phylodynamic simulation to infer ecological properties. agents model both epidemic progression as well as abstract changes occurring in viral genotypes, which are sampled during simulation to construct a phylogenetic tree or phylogram.

Methodology

The proposed method for identifying high risk countries and eliciting H5N1 evolutionary characteristics in them consists of two distinct phases (see Figure 3).

Phase 1: Identifying High-Risk Countries

The high-risk countries used in this study have been identified via epidemiological simulations of migratory waterfowl conducted via SEARUMS,[3] an ecological and epidemiological modeling and analysis system developed in Java.[5] SEARUMS includes tools for generating epidemiological models of migratory birds from GIS data obtained from the Global Register of Migratory Species (GROMS) database.[10,11] The process of generating the model from GROMS GIS data is discussed in our earlier publication.[11] Figure 4a illustrates the generated model for one waterfowl species, but SEARUMS has been used to generate a model that includes all 22 high-risk waterfowl species involved in the global dispersion of H5N1 as shown in Figure 4b.[12] Each agent in the model, shown as a circle in Figure 4, represents a flock of collocated birds that migrate as a unit. Lifecycle activities, migratory behaviors, interagent interactions, and epidemic progressions are accomplished via the exchange of discrete events.

Generation of infection graph. The epidemic spread is simulated for a period of five years by introducing an initial infection seeded in Guangdong, China, corresponding to the H5N1 nomenclature phylogram.[1]

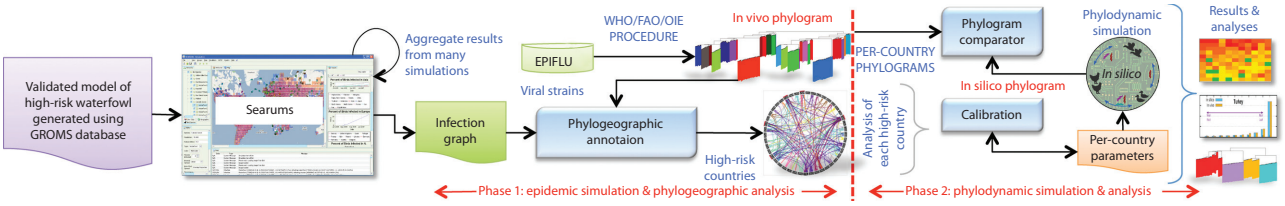

Figure 3. Overview of the key steps in the proposed methodology that combines epidemiological, phylogeographic, and phylodynamic analysis (high-resolution image in supplementary materials at http://pc2lab.cec.miamioh.edu/documents/cise16_suppl.pdf).

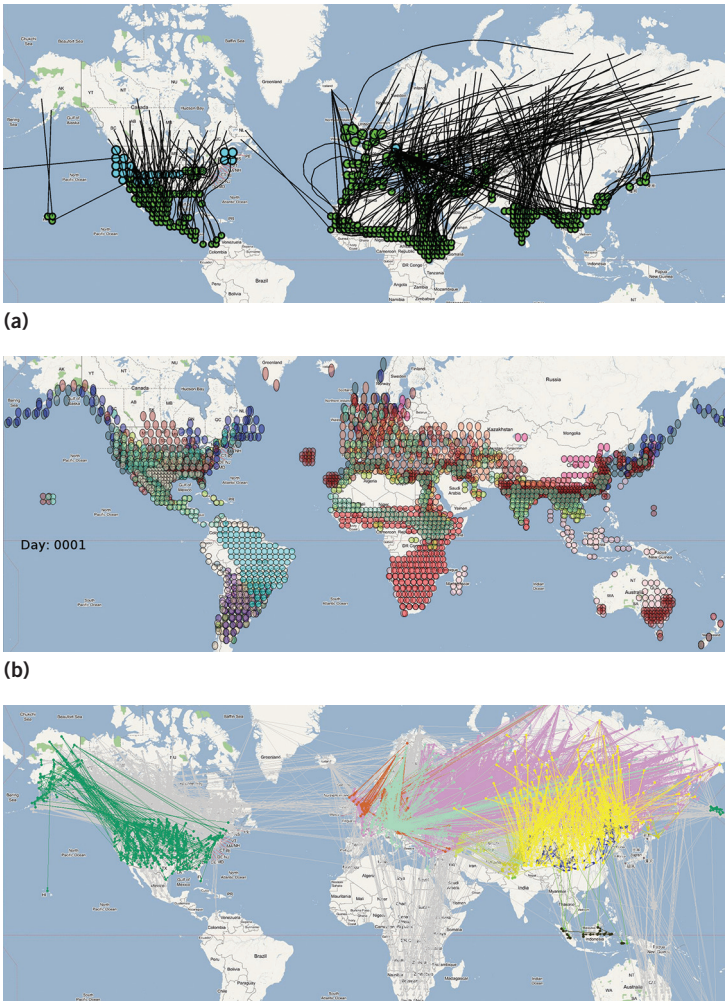

Figure 4. Epidemiological model of migratory waterfowl and resulting infection pathways: (a) single waterfowl species, (b) model with 22 high-risk species, and (c) annotated infection graph. (See http://pc2lab.cec.miamioh.edu/documents/cise16_suppl.pdf for videos, high-resolution images, and additional images of full phylogenetic trees.)

SEARUMS tracks and annotates infections occurring between pairs of agents during simulations to generate an infection graph at the end of the simulation. The resulting directed acyclic graph (DAG) shows locations where the flocks originate as nodes with edges, indicating direct infection between pairs of locations. The stochastic nature of simulations requires 25 runs; edges that occur in the majority of simulations are retained as the dominant infection pathways. Each simulation run requires about 5.5 hours when using eight threads on an Intel Core i7-3770K CPU at 3.5 GHz and about 4 Gbytes of RAM.

Phylogeographic annotation of infection graph. The phylogeographic method for validation discussed in our earlier publication[5] has been adapted to annotate edges in the infection graph using clades constituting the complete H5N1 nomenclature phylogram from World Health Organization (WHO)/World Organization for Animal Health (OIE)/Food and Agriculture Organization of the United Nations (FAO).[1] Each clade contains in vivo viral samples with less than 1.5 percent difference in nucleotides from various countries. An edge is assigned to a clade if the clade contains strains from the two countries connected by an edge. The annotation is based on the inference that genetically similar viruses from two different countries require a direct vector and pathway to enable their dispersion. For example, the infection pathway between ⟨30°52′12″N, 28°22′14.5194″E⟩ and ⟨36°41′4.9194″N, 36°41′4.9194″E⟩ is geocoded to Matruh, Egypt, and Mugla, Turkey, respectively, with annotation to clade #2.2.1, which contains H5N1 HA sequences from Egypt and Turkey with a less than 1.5 percent difference: `A/duck/Egypt/08355S-NLQP/2008` and `A/chicken/Turkey/Ipsala563/2008`. Figure 4c shows the resulting annotated infection graph.

Identification of high-risk countries. Phylogeographically annotated edges are strong evidence supporting infection pathways and are used to assess influence of countries. The diagram in Figure 5 summarizes the number of annotated edges (from Figure 4c) between pairs of countries. The countries involved

in the highest number of intercountry infections are deemed "high risk" because outbreaks can result in maximum dispersion of novel viral strains. The high-risk countries are CHN (China, 14), TUR (Turkey, 7), VNM (Vietnam, 6), RUS (Russia, 6), and NGA (Nigeria, 4), where the value in parentheses is the number of other countries to which viruses are dispersed. Most of these countries had two or just one other directly connected country. China and Russia weren't used for further analysis due to lack of sufficient in vivo samples and low geospatial resolution of annotation when compared to the size and geographic diversity reported elsewhere. Consequently, the phylodynamic analysis to elicit H5N1 characteristics are focused on the other three countries, namely, Turkey, Vietnam, and Nigeria.

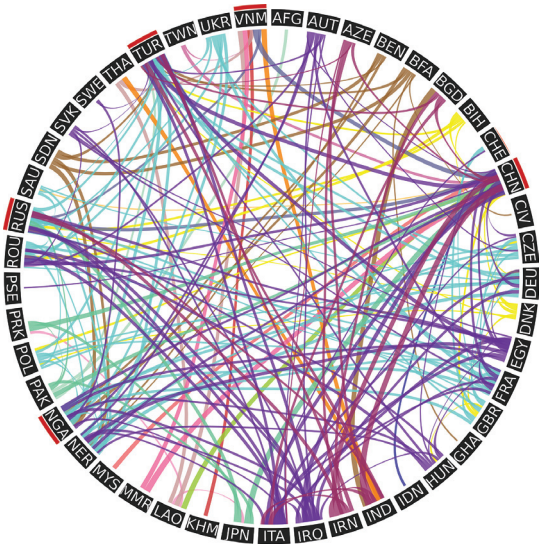

Figure 5. Infection pathways between pairs of countries. High-risk countries are in red.

Phase 2: Phylodynamic Simulation

The second phase commences with construction of the reference in vivo phylograms shown in Figure 6 for Turkey, Vietnam, and Nigeria. The in vivo phylograms are used to compare with the in silico phylograms and validate the phylodynamic simulation and the parameters settings. The phylogenetic trees for the three countries were generated using the same procedure used by WHO/OIE/FAO.[1] Full-length HA segments (more than 1,600 nucleotides) were obtained from the GISAID EpiFlu database[13] and were collected between 2005 and 2009. The HA sequences were aligned using MUSCLE[14] and the phylogenetic trees were constructed with PAUP*[15] using neighbor-joining and the same standard GTR+I+Γ model as WHO/OIE/FAO.[1,5] The newick form of the phylogram generated by PAUP* was used to categorize leaves into clades such that percentage pairwise nucleotide distances between and within clades are more than 1.5 percent and less than 1.5 percent, respectively, concordant with WHO/OIE/FAO clade definition criteria.[1] The clades in the trees are shown in different colors (colors aren't significant) in Figure 6, with Turkey, Vietnam, and Nigeria having 2, 26, and 4 clades, respectively—that is, Turkey has the lowest genetic diversity of H5N1 strains while Vietnam has the highest diversity among the three countries. The depth of the branches between each pair of sequences is proportional to the number of nucleotide differences between them.

Phylodynamic model and assumptions. The phylodynamic model and simulator used in this study have been developed in Java by enhancing the Antigen[7] simulator. Antigen has been developed for phylodynamic analysis of human influenza, with the following features added to extend it for avian influenza analysis: simulation of multiple species with different birth and death rates; births occurring only during specific brooding seasons rather than throughout the year; genetic and antigenic properties of viruses modeled independently; antigenic distances between simulated HA strains estimated by using a cross-immunity approach; phylogenetic trees constructed based on genetic differences rather than difference in emergence times; and infection rates and infective periods accounting for seasonal variations in the countries.

Unlike SEARUMS, the phylodynamic simulator uses an individual-based approach. The epidemiological properties of each individual is characterized using the standard SIS (susceptible → infected → susceptible) compartmental model summarized in Figure 7. The SIS model is used due to the endemic nature of H5N1 in waterfowl, resulting in very low mortality rates. The model doesn't include human interactions because infections in humans are very sporadic, and human-human transmission is unsustained.[5,9] Consequently, humans don't play a role in antigenic diversity in viruses[9] and aren't included in our model.

Hosts in each species are added during their respective brooding season to model births (at rate μ_b) and removed throughout the year to model deaths (at rate μ_d). Average lifespan of different species (see column LS in Table 1) has been used to determine birth and death rates so as to maintain bird

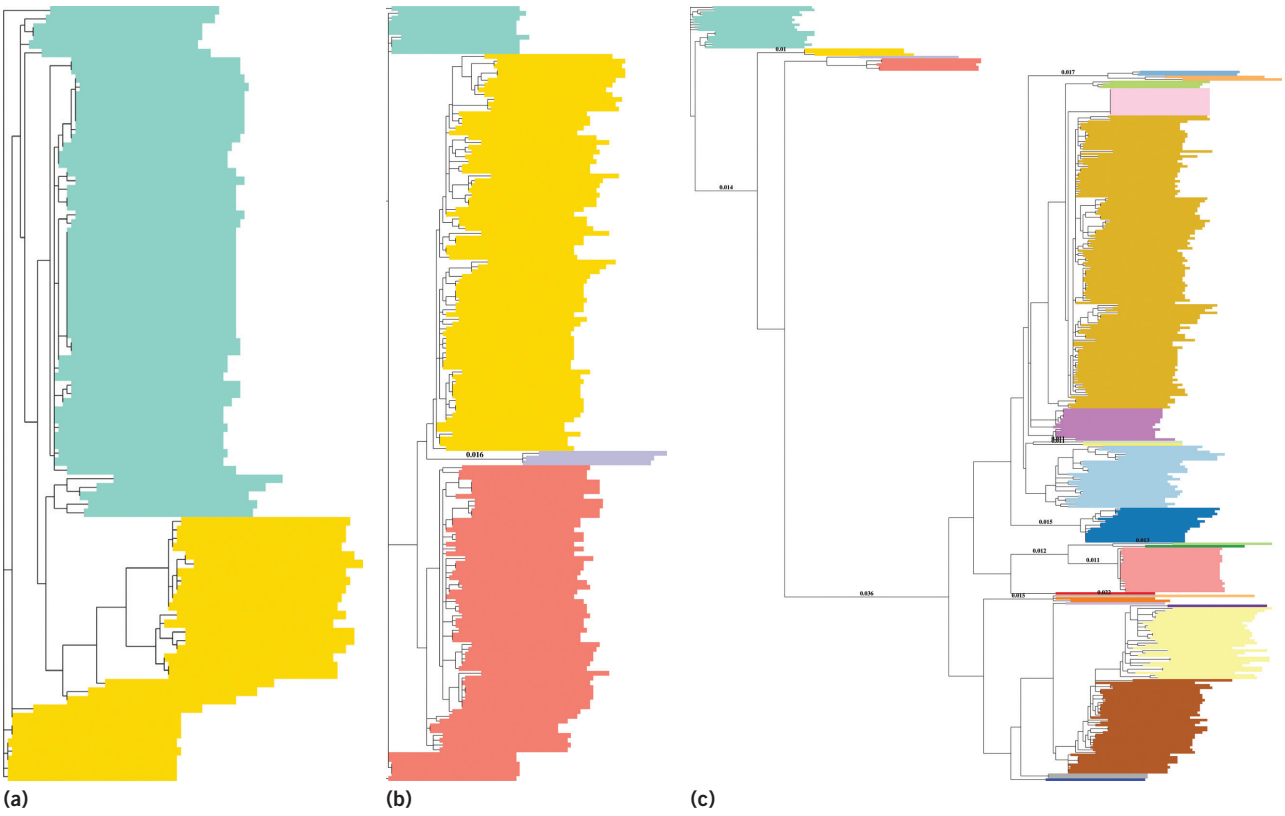

Figure 6. Reference in vivo phylogenetic trees generated using the procedure recommended by the WHO/OIE/FAO H5N1 evolution group.[1] The clades in the trees are shown in different colors (colors aren't significant), with (a) Turkey, (b) Vietnam, and (c) Nigeria having 2, 26, and 4 clades, respectively—that is, Turkey has the lowest genetic diversity of H5N1 strains while Vietnam has the highest diversity among the three countries.

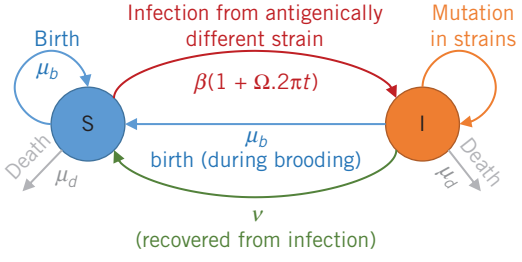

Figure 7. Overview of conceptual SIS (susceptible → infected → susceptible) model. The model doesn't include human interactions because infections in humans are very sporadic, and human-human transmission is unsustained.

populations at the end of each year. Infection spreads from an infective host to a susceptible host based on the contact rate β and seasonal sinusoidal modulation with amplitude Ω as shown in Figure 7. Seasonal modulation[9] has been used to account for influence of temperature on viral durability on surfaces where birds come in contact with each other.

The viral phenotypes are modeled using abstract 2D vectors similar to the approach used by Bedford and colleagues.[7] Mutations in the viruses shed by hosts are uniformly distributed throughout the strain at the rate of ψ. Mutations essentially change the coordinates of the viruses in the abstract 2D space. This enables Euclidean distances to be mapped to antigenic change using the method proposed by Julia Gog and colleagues, which is used by other investigators as well.[7,9] The antigenic parameters proposed by Roche and colleagues[9] for H5N1 have been used in the simulations so that an antigenically different[1] strain that isn't present in the host's immune history causes new infections. Host immune history drives the viral selection pressure, causing specific serotypes to establish in the waterfowl population. The viruses are periodically sampled during simulation (similar to the in vivo sampling process) to yield the in silico phylogenetic tree at the end of the simulation. The phylogenetic

Table 1. Population distribution (N = 20,000 for all three countries) and brooding period of high-risk waterfowl species identified using GIS data from the GROMS database.

Waterfowl species	Life span (years)	Brood time	Population fraction		
			Vietnam	Turkey	Nigeria
Eurasian wigeon (*A. Penelope*)	2.02	Feb.–Apr.	0.75	0.74	0.15
Common teal (*A. Crecca*)	2.5	Dec.–Feb.		0.02	0.02
Northern pintail (*A. Acuta*)	3	Feb.–July	0.05	0.04	0.06
Tufted duck (*A. Fuligula*)	3.5	Feb.–Apr.		0.02	0.02
Northern lapwing (*V. Vanellus*)	3.5	Apr.–July		0.15	
Black godwit (*L. Limosa*)	18	Feb.–Apr.	0.15	0	0.13
Ruff (*P. Pugnax*)	4.4	Mar.–Jun.			0.62
Common gull (*L. Canus*)	10	Mar.–Apr.	0.05	0.03	

tree construction uses genetic distances between viruses along with common ancestor information to produce the phylogram. The simulation uses Gillespie's stochastic simulation algorithm (SSA) along with Tau-Leap optimization and a time step of 0.1 days[8] to ensure accurate simulations.

The design of the phylodynamic model involves the following domain-specific assumptions and limitations:

- It doesn't explicitly model the influence of seasonal migration on contact rates (β) and uses β as an aggregate parameter.
- Birth (μ_b) and death (μ_d) are set such that the overall population of each species (and consequently their relative fractions; see Table 1) is maintained in each year of simulation. However, over 15-year periods, the relative fraction of bird species can vary, which isn't modeled.
- The role of temperature and environmental uptake is only implicitly modeled as modulation on contact rate (β) and therefore their influence can't be separately assessed.

Phylodynamic model validation. We used the same validation procedure proposed by Roche and colleagues[8] to validate our multispecies phylodynamic model by comparing the results against the extensively validated phylodynamic model from Bedford and colleagues[7] (referred to as the "reference model"). For validation, we created a multispecies model, but with all the species having identical parameter settings to compare against the original single species model. The chart in Figure 8a shows a comparison of the 95 percent confidence interval (CI) from 250 replications of the stochastic simulation over a period of the years. The red line in the chart shows the percentage difference between the average number of infective individuals. Figure 8b compares antigenic diversity in the reference model versus the multispecies models. Antigenic diversity directly determines the number of clades in the resulting phylogram. Note that the 95 percent CI of the reference and multispecies models are very close and overlap each other in Figure 8.

The box plots in Figure 9 show the results from statistical comparison of key epidemiological and antigenic attributes, namely, peak infection, peak infection day, and antigenic diversity. The p-values from Kolmogorov-Smirnov (KS) two-sample tests conducted on the reference and multispecies model were greater than 0.05, establishing that the validated reference and proposed multispecies models are statistically indistinguishable. Similar comparisons were conducted for different settings of parameters to ensure that the epidemiological and antigenic results from the reference and proposed multispecies models were statistically the same, thereby establishing the validity of salient aspects of the multispecies model.

Next, the influence of varying brooding periods and seasonal modulation was verified using a metamorphic validation approach.[16] Specifically, the brooding period for one species was modified to be 10 days earlier when compared to the reference model. Results from 250 stochastic simulations

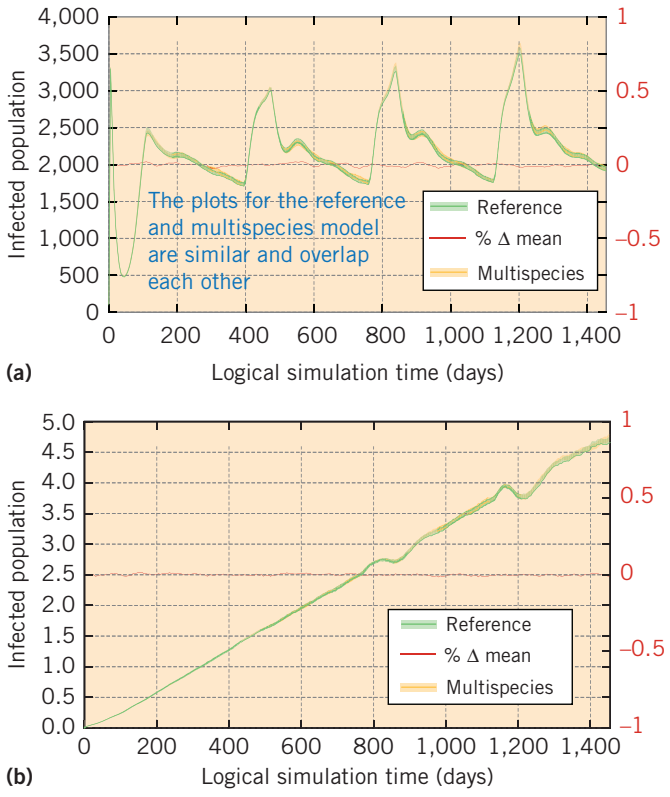

Figure 8. Comparison of infective population and antigenic diversity in the reference model[7] versus multispecies models from 250 simulations: (a) infective population and (b) antigenic diversity. The red line in the chart shows the percentage difference between the average number of infective individuals.

were used to verify that the peak infection day correctly changed on an average by 5.9 ± 4.41 days as per expectation. Similarly, the influence of 0.1 sinusoidal modulation on contact rates was verified to introduce the expected 10 percent sinusoidal change in infection patterns as shown in Figure 10, thereby establishing the proposed model's validity.

Per-country models. The three independent phylodynamic models for Turkey, Vietnam, and Nigeria involve different subsets of high-risk waterfowl species found in those countries (see Table 1). The waterfowl species in the different countries and their relative populations were extracted from the GROMS GIS dataset.[10] The brooding seasons and average life spans of the waterfowl have been obtained from various data sources aggregated in the animal diversity database maintained by the University of Michigan.[17] The average infective duration, that is, the time during which an infected bird can spread infection to other birds, is five days.[9]

The ecological parameters in the per-country phylodynamic model whose values aren't known and can vary between countries include

- Contact rate β, the average number of birds that can get infected each day by another infective host, has been estimated at 2.847 contacts/day[18] on average in small lakes. This value is used as the reference to determine contact rate via calibration.
- Mutation rate ψ is the average number of mutations per day in the HA segment of viruses shed by infective hosts. The diversity in ecosystems, climatic conditions, and bird species give raise to different mutation rates for the viruses in different countries.

Several investigations[19] have analyzed the evolutionary rate of the HA segment of H5N1 viruses and have consistently reported the following per-site/year (τ) mutation rates: Turkey (17.8 to 2.06), Nigeria (1.62 to 4.05), and Vietnam (15.81 to 18.65). The per-site/year values are converted to ψ, the daily mutation rate using the relationship $\psi = 1{,}800\tau \div 365$, where 1,800 is the average number of nucleotides in HA.

The actual values for β and ψ for each country are determined via calibration, which requires a systematic search of the parameter space based on their estimated values. Note that additional combinations of β and ψ (beyond estimated ranges) could yield in silico phylograms with the same number of clades and properties as in vivo phylograms. However, such combinations of β and ψ that are beyond the estimated range of values can't be substantiated with observed data and aren't biologically meaningful. The seasonal temperature in Turkey and Vietnam, for example, swings considerably (from 16 °C to 29 °C), which influences environmental viral durability and consequently the spread of epidemics. Accordingly, for these two countries, the contact rate β is scaled by a seasonal sinusoidal modulation of $1 + 0.1 \cos(2\pi t \div 365)$ (see Figure 7), where simulation time t is day of year ($0 < t < 365$); the value of $\Omega = 0.1$ is proposed by Roche and colleagues,[9] but sinusoidal modulation isn't used for Nigeria as the temperature is very uniform throughout the year (with less than a 2 °C change).

Model calibration. Calibration is an empirical process through which selected parameters in the model are fine-tuned such that the model characterizes

observed data with sufficient accuracy. Calibrating the two parameters, namely, contact rate β and mutation rate ψ, for a given country is accomplished by comparing the in silico phylogram generated via simulation to the corresponding in vivo phylogram (see Figure 6). The range of values for the two parameters has been explored using the following initial estimates: 2.847 contacts/day for β and different ranges of mutation rates listed earlier.[18]

The in silico phylograms don't have taxa (identifiers such as `A/duck/Egypt`) similar to the in vivo phylogram. Consequently, comparison of the two phylograms is accomplished by using the following standard phylogram topology (or shape) metrics[20] in decreasing order of importance: number of clades (primary determinant of antigenic diversity), interclade distance (estimates evolutionary distance between clades), average number of child nodes (leaf nodes have zero children, which reflects speciation), and average depth (number of intermediate nodes to the root, which reflects establishment/extinction rates of strains). Note that both in vivo and in silico clades contain sequences that have a less than 1.5 percent difference between each other. Consequently, intraclade distances aren't a useful distinguishing factor and aren't included in the set of metrics. The ETE toolkit[21] has been used to develop Python scripts to analyze the phylograms and generate comparative metrics. Because the parameter space is reasonably constrained, calibration has been performed by exhaustively searching the solution space in small increments (to avoid potential pitfalls with heuristic searches).

The simulations for calibration and analysis were conducted for a period of 18 years, with the first 15 serving as "burn-in" time. The burn-in period accounts for the difference between the putative root of the WHO/OIE/FAO reference tree (`A/turkey/England/5092/1991`) versus the actual viral isolates in the nomenclature phylogram that start from 2006 for the three countries. The in silico viral sampling commences only after the burn-in period and continues until end of simulation. The simulations for Turkey, Vietnam, and Nigeria were conducted for three, four, and three years, respectively, corresponding to the range of years used in the WHO/OIE/FAO reference phylogram. Each simulation requires 4 Gbytes of RAM and approximately 522 ± 42 seconds of runtime on an Intel Xeon X5550 CPU at 2.67 GHz.

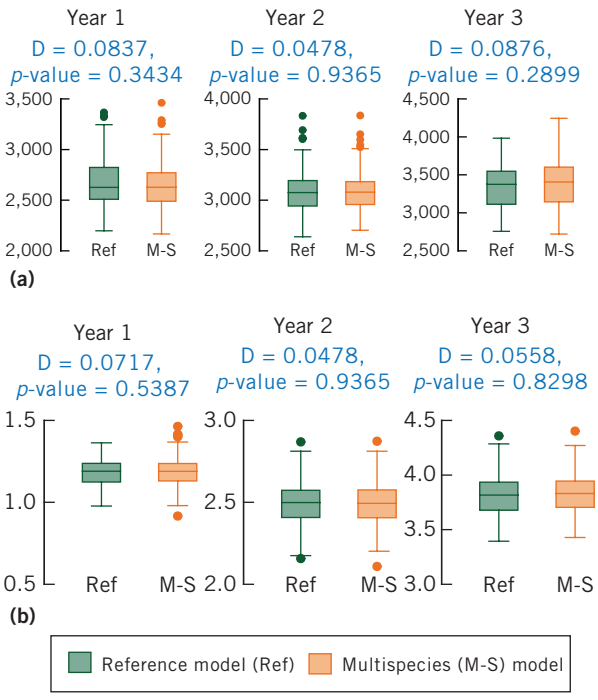

Figure 9. Statistical comparison of key epidemiological and antigenic characteristics of the reference model (green) and the multispecies model (orange): (a) peak infection and (b) antigenic diversity.

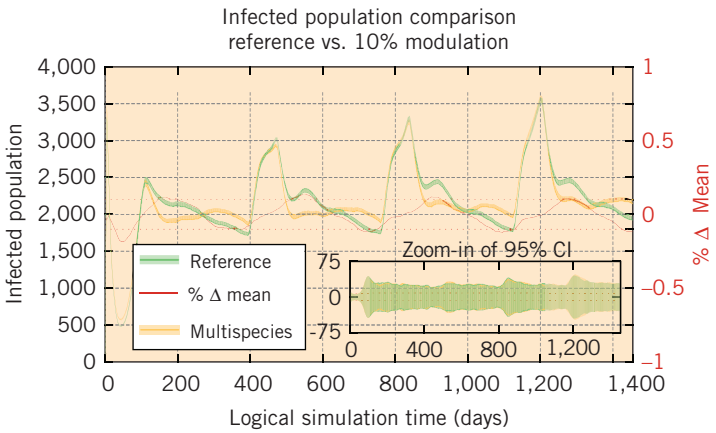

Figure 10. Comparison between reference and multispecies models with 0.1 sinusoidal modulation ($\Omega = 0.1$) on contact rate (β).

Results and Discussions

The phylogeographic and phylodynamic analysis methods we just discussed were utilized to analyze the evolution of H5N1 in three high-risk countries, namely, Turkey, Vietnam, and Nigeria. The charts in Figure 11 show success rates, that is, the fraction of stochastic simulations that yield the same number

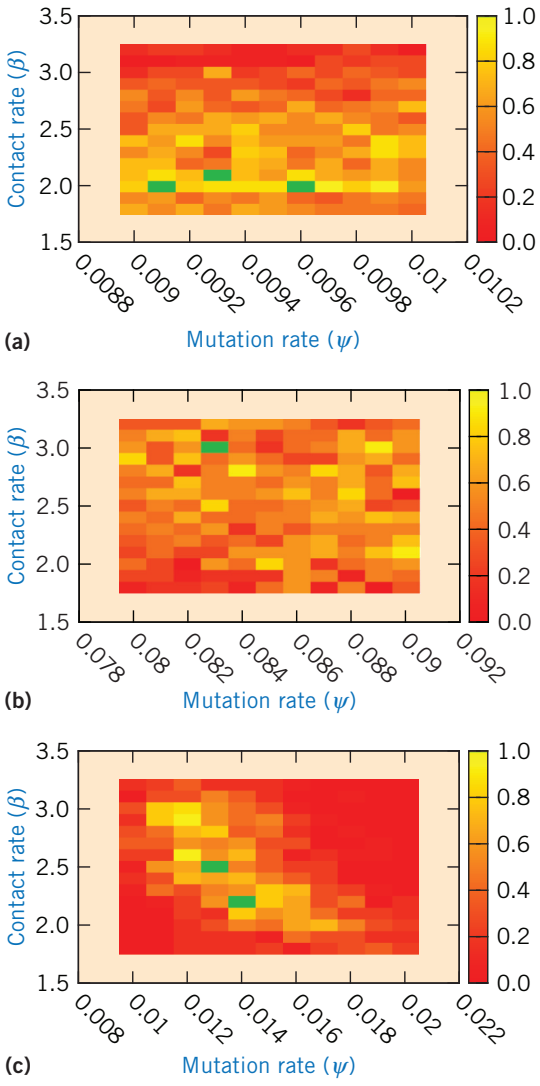

Figure 11. Comparison of success rates (fraction of in silico phylograms that yield same number of clades as in vivo phylograms) for different parameter settings for β and ψ: (a) Turkey, (b) Vietnam, and (c) Nigeria. The highest success rates are highlighted in green.

of clades (the primary metric) between in silico and in vivo phylograms. The values plotted are from 50 stochastic runs for each pair of parameter settings, for a total of 7,500 simulation replications per country. Table 2 tabulates the other metrics corresponding to parameter values for contact rate β and mutation rate ψ that yield the same number of clades. The entries in the table have been ordered with the best configuration for each country being listed first.

The data in Figure 11a shows that different pairs of values for β and ψ can explain the diversity of viruses in Turkey and Nigeria after 15 years of burn-in time. However, the secondary metric in Table 2, namely, interclade distance, provides a strong arbitration to identify the best ecological settings for these two countries. Note that, for Vietnam, there is a much higher viral diversity, so only one set of parameter settings repeatably yields the observed viral diversity.

Analysis of Overfitting

To estimate issues with overfitting, we looked at the impact of changing calibrated parameter settings by ±10 percent on the mean number of clades. Statistical analysis using F-test (for variance) and T-test (for means) used results from 20 independent simulations for each parameter settings. In all cases, the variance was statistically similar, with F-tests yielding $p\text{-value}_{\alpha = 0.05} \gg 0.05$. The results from the T-tests also showed that in almost all cases, a ± 10 percent change in parameters didn't result in a statistically significant ($p\text{-value}_{\alpha = 0.05} > 0.05$) change in the mean number of clades from the simulation. The overall statistically significant consistency in mean and variance suggests that the model and parameter settings haven't been overfitted to the observed number of clades for Turkey, Nigeria, and Vietnam.

Assessment of Critical Parameters

We used the GSA method[22] to rank and identify critical parameters in the model. The GSA method utilizes success/failure rates from simulations conducted using a range of values for a given parameter to yield a $d_{m,n}$ statistic. This statistic is sensitive to differences in both central tendency and any difference in the distribution functions. The magnitude of the $d_{m,n}$ statistic indicates that parameter's importance in a model (larger values indicate that the parameter has a higher influence).

Table 3 shows the $d_{m,n}$ statistic measured by varying the parameter by ± 50 percent (in steps of 10 percent) around its calibrated settings and from 20 simulation runs for each setting. The parameters in Table 3 are ordered from most influential to least. Overall, contact rate β is the most critical parameter that influences genetic diversity as it plays a critical role in the spread of epidemics. However, the range for β is small (from 2.0 to 3.0 in Table 2, a 50 percent change), consistent with the expectation that behaviors and interactions of wild waterfowl species don't significantly vary between the countries. The next influential parameter is mutation rate ψ, which introduces antigenic variations enabling viral lineages to become endemic in the population. Collectively, as indicated by Table 3, lower contact and mutation rates reduce the evolutionary rate and

Table 2. Metrics used for comparing in vivo and in silico phylograms for contract rate (β) and mutation rate (ψ) values that yield the same number of clades.

Country	β	ψ	Interclade distances	No. child nodes (%)	Node depth (%)
Turkey	2.1	0.0093	20	1.92	10
	2.0	0.0091	21	0.31	16
	2.0	0.0096	23	0.30	16
Vietnam	3.0	0.083	32	0.18	7
Nigeria	2.5	0.013	10	0.18	21
	2.2	0.014	16	0.23	22

genetic diversity of viruses, resulting in fewer distinct clades in the phylogram.

The diversity in bird species, which influences brooding season, birth (μ_b), and death (μ_d), ranks third, indicating it's a key parameter. In other words, diversity of bird species and differences in their life cycle is an important ecological characteristics to be modeled. Seasonal modulation plays some role in explaining the antigenic diversity and can't be completely ignored. In contrast, the net number of individuals in the model (N) and the number of initially infected birds (Io) don't have significant influence on the results from simulations.

Table 3. Ranking of parameters from generalized sensitivity analysis (GSA).[22]

Parameter	Score ($d_{m,n}$ statistic)		
	Turkey	Nigeria	Vietnam
Contact rate (β)	0.526	0.447	0.216
Mutation rate (ψ)	0.446	0.385	0.354
Bird species	0.191	0.138	0.343
Seasonal modulation (Ω)	0.174	N/A	0.123
Population (N)	0.156	0.133	0.094
Initial infection (Io)	0.109	0.091	0.0878

Phylogram Comparisons

Figure 12 shows an in silico phylogenetic tree for the three countries using the best parameter settings tabulated in Table 2. Comparing to the in vivo phylograms in Figure 6, the in silico phylograms also exhibit similar and consistent structure with the same number of clades. As summarized by the column titled "No. child nodes" in Table 2, the degree of nodes in the in vivo and in silico phylograms were almost identical, establishing that the evolutionary characteristics are consistently modeled.

The depth of nodes is a measure of long-term lineage establishment and overall genetic diversity among the strains in each country. Figure 13 shows a comparison of the fraction of nodes at different depths, with nodes closer to the root having a lower depth. The column titled "Node depth" in Table 2 summarizes the average node depths. The data in Figure 13 shows that the evolutionary shape of the in vivo and in silico phylograms are similar. Combined with the intercluster distances, consistency in node depths establishes that the two phylograms show similar evolutionary characteristics. Collectively, the phylodynamic analysis highlights that the in vivo and in silico phylograms are quantitatively, structurally, and visually accurate, thereby establishing validity of the calibrated parameter settings.

Inferences

The parameter settings used to generate the in silico phylogram now provide additional information about the evolution of H5N1 in the three countries, information that can't be directly obtained from in vivo data. The charts in Figure 11 along with data from Tables 2 and 3 show that even modest changes in raw contact rates can have noticeable impacts on H5N1 evolution and immune escape. Reduction of contact rates is typically accomplished by culling infected birds, particularly livestock. The inferences from the proposed method support the current practice of large-scale culling, similar to

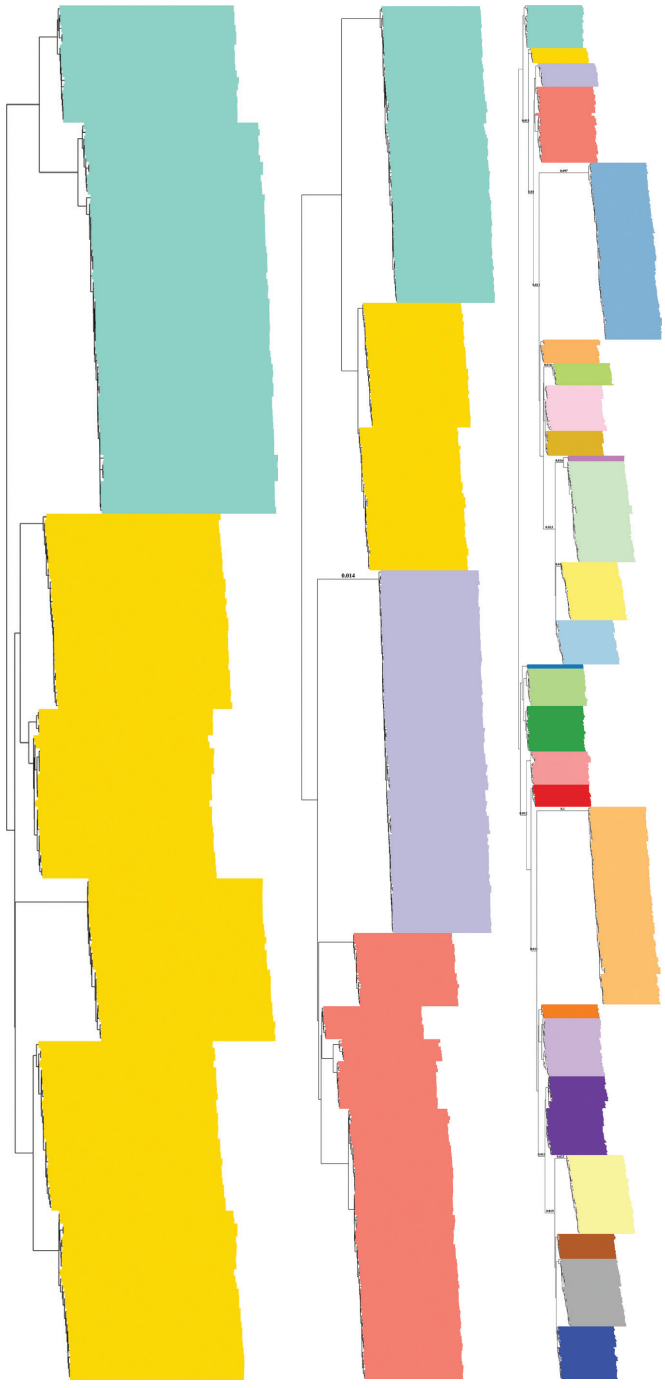

Figure 12. Sample in silico phylogenetic trees generated using the best parameter settings shown in Table 2 identified via calibration. The clades in the tree are shown in different colors.

the containment efforts pursued during the 2015 H5N1 outbreak in the American Midwest. Furthermore, the data also supports CDC/FAO efforts to encourage isolation of water and feed of domestic birds, which further reduces contact rates.

Influencing the mutation rate (ψ) of viruses, particularly in livestock, is indirectly accomplished through vaccination. Vaccinated birds are immune to viruses in several clades, thereby preventing epidemics and consequently improving further diversification in those lineages. The charts in Figure 11 along with data in Table 3 suggest that extensive vaccination can decrease viral diversification rates in the different countries. However, unlike contact rates, a more extensive vaccination effort is necessary to influence viral diversification rates. The data also suggests that economic constrains, particularly in poorer countries, drive culling as primary containment strategy over vaccination. The mutation rates for H5N1 in Vietnam (0.083) is significantly higher when compared to Turkey (0.0093) and Nigeria (0.013). The data suggests that vi-ral strains are experiencing a more rapid evolution and have the potential to give rise to novel strains through reassortments in Vietnam. Consequently, Vietnam requires greater emphasis on surveillance and containment efforts to mitigate emerging epidemics.

The implications of studying AIVs extends to human and swine influenza viruses because studies have established that all currently known influenza A viruses originated in aquatic birds. Although this study focused on three high-risk countries, the methods and analysis can be readily extended to all 12 distinct global geographical regions included in the WHO vaccine design and recommendation process.[2] The analysis can be used to guide time frames for redesigning vaccines for different geographic regions, thereby extending applicability of vaccines without compromising their efficacy. Similarly, the results can also be used to inform prophylaxis of poultry and livestock to prevent and contain emergent outbreaks from novel viral strains such as H5N3. Furthermore, the models can be used to guide and focus sampling and surveillance efforts of migratory waterfowl in areas with high antigenic drift. Moreover, the assumptions in the model, particularly the influence of migration and impact of environmental uptake, can be relaxed to provide a more comprehensive model. The proposed methodology can guide surveillance efforts to assess risk of novel strains emerging through reassortments at the human-animal interface. ■

References

1. WHO/OIE/FAO H5N1 Evolution Working Group, "Continued Evolution of Highly Pathogenic Avian Influenza A (H5N1): Updated Nomenclature," *Influenza and Other Respiratory Viruses*, vol. 6, no. 1, 2012, pp. 1–5.
2. "Candidate Vaccine Viruses and Potency Testing Reagents for Influenza A (H5N1)," World Health Org., 2014; www.who.int/influenza/vaccines/virus/candidates_reagents/a_h5n1.
3. D.M. Rao, A. Chernyakhovsky, and V. Rao, "Modeling and Analysis of Global Epidemiology of Avian Influenza," *Environmental Modelling & Software*, vol. 24, no. 1, 2009, pp. 124–134.
4. D.M. Rao, "Accelerating Parallel Agent-Based Epidemiological Simulations," *Proc. ACM SIGSIM PADS Conf.*, 2014, pp. 127–138.
5. D.M. Rao, "Enhancing Epidemiological Analysis of Intercontinental Dispersion of H5N1 Viral Strains by Migratory Waterfowl Using Phylogeography," *BMC Proc.*, vol. 8, no. 6, 2014, p. S1.
6. E.M. Volz, K. Koelle, and T. Bedford, "Viral Phylodynamics," *PLoS Computational Biology*, vol. 9, no. 3, 2013, p. e1002947.
7. T. Bedford, A. Rambaut, and M. Pascual, "Canalization of the Evolutionary Trajectory of the Human Influenza Virus," *BMC Biology*, vol. 10, no. 1, 2012, pp. 1– 12.
8. B. Roche, J.M. Drake, and P. Rohani, "An Agent-Based Model to Study the Epidemiological and Evolutionary Dynamics of Influenza Viruses," *BMC Bioinformatics*, vol. 12, no. 1, 2011, pp. 1–10.
9. B. Roche et al., "Adaptive Evolution and Environmental Durability Jointly Structure Phylodynamic Patterns in Avian Influenza Viruses," *PLoS Biology*, vol. 12, no. 8, 2014, p. e1001931.
10. "Global Register of Migratory Species (GROMS): Summarising Knowledge about Migratory Species for Conservation," 2013; www.groms.de.
11. D.M. Rao and A. Chernyakhovsky, "Automatic Generation of Global Agent-Based Model of Migratory Waterfowl for Epidemiological Analysis," *Proc. 27th European Simulation and Modelling Conf.*, 2013, pp. 21– 28.
12. W. Hagemeijer and T. Mundkur, "Migratory Flyways in Europe, Africa, and Asia and the Spread of HPAI H5N1," *Proc. Int'l Scientific Conf. Avian Influenza and Wild Birds*, 2006; www.fao.org/avianflu/conferences/rome_avian/documents/hagemeijer-mundkur.pdf.
13. "Action Stations: The Time for Sitting on Flu Data Is Over," *Nature*, vol. 441, no. 7097, 2006, p. 1028.
14. R.C. Edgar, "MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput," *Nucleic Acids Research*, vol. 32, no. 5, 2004, pp. 1792–1797.

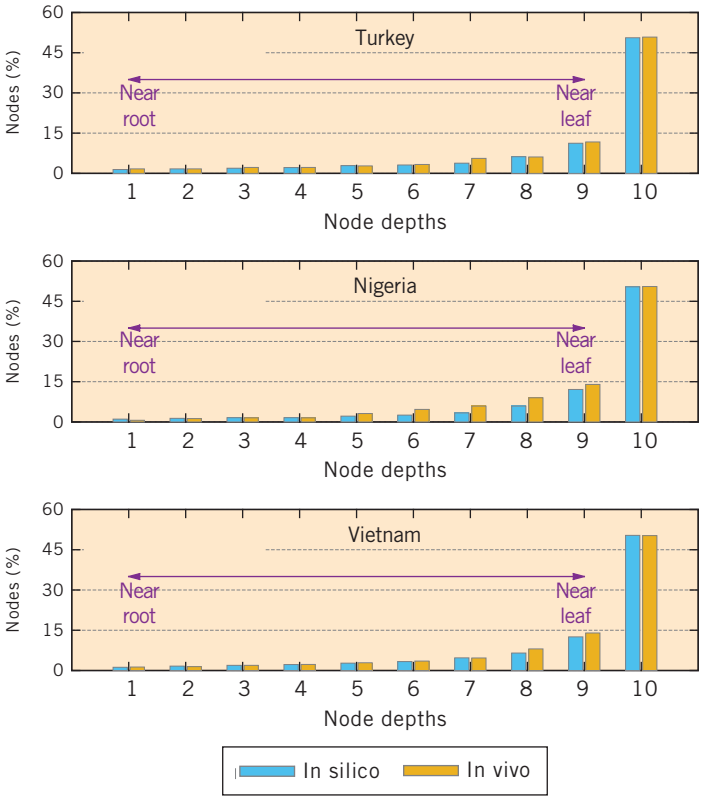

Figure 13. Comparison of node depths (other depths with less than 1 percent of nodes not shown for brevity). Combined with the intercluster distances, consistency in node depths establishes that the two phylograms show similar evolutionary characteristics.

15. D.L. Swofford, "PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods) 4.0 Beta," 2003; http://paup.csit.fsu.edu.
16. A. Ramanathan, C.A. Steed, and L.L. Pullum, "Verification of Compartmental Epidemiological Models Using Metamorphic Testing, Model Checking and Visual Analytics," *Proc. ASE/IEEE Int'l Conf. BioMedical Computing*, 2012, pp. 68–73.
17. P. Myers et al., "Animal Diversity Database," 2015; http://animaldiversity.org.
18. P. Rohani et al., "Environmental Transmission of Low Pathogenicity Avian Influenza Viruses and Its Implications for Pathogen Invasion," *Proc. Nat'l Academy of Sciences*, vol. 106, no. 25, 2009, pp. 10365–10369.
19. G. Cattolia et al., "Evidence for Differing Evolutionary Dynamics of A/H5N1 Viruses among Countries Applying or Not Applying Avian Influenza Vaccination in Poultry," *Vaccine*, vol. 29, no. 11, 2011, pp. 9368– 9375.
20. P. Puigbo, Y.I. Wolf, and E.V. Koonin, "Genome-Wide Comparative Analysis Of Phylogenetic Trees:

The Prokaryotic Forest of Life," *Methods in Molecular Biology*, vol. 856, 2013, pp. 1792–1797.
21. J. Huerta-Cepas, J. Dopazo, and T. Gabaldn, "ETE: A Python Environment for Tree Exploration," *BMC Bioinformatics*, vol. 11, no. 24, 2010; http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-24.
22. B. Guven and A. Howard, "Identifying the Critical Parameters of a Cyanobacterial Growth and Movement Model by Using Generalised Sensitivity Analysis," *Ecological Modelling*, vol. 207, no. 1, 2007, pp. 11–21.

Neil Giridharan is a student at William Mason High School in Mason, Ohio. His research interests include computational biology, disease forecasting, and viral phylodynamics.

Dhananjai M. Rao is an assistant professor in the Department of Computer Science and Software Engineering (CSE), Miami University, Oxford, Ohio. His research interests include computational epidemiology, disease forecasting, and parallel simulation. Rao has a PhD in computer science and computer engineering from the University of Cincinnati. Contact him at raodm@miamiOH.edu.