Generalized Linear Models for Identifying Predictors of the Evolutionary Diffusion of Viruses

Rachel Beard[1], Daniel Magee[1,], Marc A. Suchard MD, PhD[2], Philippe Lemey PhD[3], Matthew Scotch PhD, MPH[1]
[1]Arizona State University, Tempe, AZ, USA
[2]University of California, Los Angeles, CA, USA
[3]KU Leuven, Leuven, Belgium

Abstract

Bioinformatics and phylogeography models use viral sequence data to analyze spread of epidemics and pandemics. However, few of these models have included analytical methods for testing whether certain predictors such as population density, rates of disease migration, and climate are drivers of spatial spread. Understanding the specific factors that drive spatial diffusion of viruses is critical for targeting public health interventions and curbing spread. In this paper we describe the application and evaluation of a model that integrates demographic and environmental predictors with molecular sequence data. The approach parameterizes evolutionary spread of RNA viruses as a generalized linear model (GLM) within a Bayesian inference framework using Markov chain Monte Carlo (MCMC). We evaluate this approach by reconstructing the spread of H5N1 in Egypt while assessing the impact of individual predictors on evolutionary diffusion of the virus.

Introduction

Bioinformatics and phylogeography models use viral sequence data to analyze spread of epidemics and pandemics. However, few of these models have included analytical methods for testing whether certain predictors such as population density, rates of disease migration, and climate are drivers of spatial spread. While spatial epidemiology has successfully developed models of environmental predictors such as global mobility and air travel, these models remain disconnected to molecular sequence data that are analyzed through bioinformatics and phylogeography applications to unlock information about virus coalescence, spatial spread, and gene flow.[1] Combining spatial epidemiology and molecular sequence data can lead to discoveries about risk of transmission between animals and humans as well as the relationship between geography and genetic evolution of the virus. In addition, understanding the specific factors that influence spatial diffusion of viruses is critical for targeting public health interventions and limiting spread. In this study, we describe the application and evaluation of a phylogeographic model that integrates demographic and environmental factors. Here we focus on a variant clade of H5N1 viruses in Egypt and its countrywide diffusion among avian and human hosts. This approach is generalizable to other RNA viruses and may enhance both public health prevention and response by identifying the drivers that are most vital to viral spread.

Background

Many emerging or re-emerging infectious diseases are zoonotic in origin, and pose significant threats to human and animal health.[2] There are many potential drivers of transmission between animals and humans and many of these drivers likely vary between countries. This variation could be caused by climate differences, population sizes, and living conditions, as well as cultural practices related to food preparation and distribution. In response to these complexities, many epidemiologic models have studied potential contributors such as human and avian population densities, or precipitation.[3] For example Van Boeckel *et al.* examined anthropogenic and ecological variables relating to avian species within developed regions in Asian farming communities following flood conditions,[4] while Tamerius *et al.* observed the effects of temperature, humidity, and precipitation on H5N1 spread in tropical climates.[5] While this research has resulted in valuable epidemiologic insights, it has traditionally ignored the information about the evolutionary processes occurring within the viral genome. Phylodynamic analysis of RNA viruses can lead to crucial information regarding transmission, genetic diversity and selection, as well as epidemiologic characteristics.[6] Bioinformatics and phylogeography techniques have enabled researchers to depict local and global virus spread, providing valuable information to the public health community as to the origin and epidemic patterns of spread. For instance, Lam *et al.* determined that the spread of influenza A subtype H5N1 was likely introduced into Indonesia by a single introduction in East Java in approximately 2002, followed by both an east and westward migration throughout the country.[7] Bioinformatics approaches such as these are informative; though few incorporate demographic and environmental factors often used in epidemiology. Ypma *et al.* demonstrate this concept by including geographic and temporal elements as well as genetic data to estimate the

migration patterns of influenza A subtype H7N7 in the Netherlands.[8] By taking an integrated approach, this work highlighted the estimates of certain drivers on evolutionary transmission with greater accuracy.[8] The same group also demonstrated that using within-host dynamics and genetic data of pathogens to simultaneously generate both the phylogenetic tree and transmission route leads to more accurate models and plausible estimation of connecting variables.[9] Thus, epidemiologic and viral phylogenetic approaches have been incorporated into a rough framework which join evolutionary and ecologic dynamics to explain spatial diffusion.[10] Phylogeography naturally compliments models based on observed epidemiologic data, as the genomic data can provide a record by which to confirm or reject hypothesized patterns of viral spread. Our aim is to demonstrate the utility of combining epidemiologic and phylogeographic approaches to identify drivers of virus diffusion. We evaluate this approach by reconstructing the spread of H5N1 in Egypt while assessing the impact of individual predictors on evolutionary diffusion of the virus.

Methods

A Bayesian generalized linear model (GLM) approach was adopted which was developed by Lemey *et al.*, in which the spatiotemporal patterns of viral diffusion are reconstructed while potential contributing factors are simultaneously assessed.[11] We use the work of Scotch *et al.*[12] as a basis by which to analyze the potential environmental drivers of highly pathogenic avian influenza (HPAI) H5N1 movement among multiple hosts by considering discrete geographic locations within Egypt. We chose to focus on Egypt because it has recently emerged as an epicenter for H5N1, with 173 human cases reported to the World Health Organization (WHO) as of June 2013.[12] In addition, the local cultures prefer to obtain their poultry via live bird markets which create an atmosphere of high human-avian transmissibility.

Sequence data

We used the same dataset described by Scotch *et al.*[12] that included 226 H5N1 hemagglutinin (HA) sequences previously isolated, however we excluded two sequences for which the host was recorded as environmental. Sequences collected from avian (n=210) and human (n=14) hosts in Egypt spanning 2007-2012. The sequences were selected based on their Egyptian origin and classification within the recently defined variant subclade 2.2.1.1. published by WHO.[13]

We reconstructed the spread of H5N1 in Egypt using a discrete phylogeography approach while estimating the effect of a diverse set of variables on phylogeographic diffusion within a GLM. This process was implemented using the development version of the BEAST software package, available at http://code.google.com/p/beast-mcmc/, which uses a Bayesian Markov Chain Monte Carlo (MCMC) analysis.[14] We modeled sequence evolution using the generalized time-reversible (GTR) model of nucleotide substitution, while using a relaxed molecular clock. Multiple chain lengths were tested using Tracer,[15] with the final run set at 20 million.

Generalized linear model

We tested the effect of predictors on spatial spread while reconstructing the spatiotemporal history. Here, we used modeling techniques described in Lemey *et al.*,[11] and innovative methods for Bayesian phylogeographic inference of phylogenetic history and discretized diffusion processes.[16] We utilized a GLM model by integrating diffusion of viral spread as a non-reversible continuous time Markov chain processes expressed as a K x K infinitesimal rate matrix of location change (Λ) among K discrete locations.[11] We represented all rates of movement Λ_{ij} using a log linear function to incorporate a set of n predictors on the log-scale.

$$\log\Lambda_{ij} = \beta_1\delta_1\log(p_1) + \beta_2\delta_2\log(p_2) + \ldots + \beta_n\delta_n\log(p_n)$$ [11]

Here, β signifies the contribution of a given predictor to the model, and δ is a binary indicator (0, 1) variable that oversees whether a particular predictor is to be incorporated in the model.[17] This allows for Bayesian stochastic search variable selection (BSSVS),[16-18] in which posterior probabilities of all possible models that may or may not include a given predictor are estimated, as discussed in Lemey *et al*, 2009.[17, 18] We utilized a Bernoulli prior probability distribution for δ as in Lemey *et al.* 2012, to place equal probability of inclusion or exclusion of predictors.[11]

We selected local predictors based on feedback from experts who study H5N1 in Egypt.[19] These predictors were

chosen to represent genomic, geographical, demographical, and numerical indicators to develop a preliminary model and include:

***Avian and human population density**:* We incorporated population density for all possible origins and destinations for both humans and chickens from City Population, an online resource for worldwide population statistics, and the Food and Agriculture Organization of the United Nations (FAO).[20, 21]

Latitude: We obtained the latitude of the centroid location for each governorate in order to reflect diverse climatic conditions within the country by using GeoNames.[22] While this likely does not reflect the true locations of where sequences were collected, this method was adopted to impose uniformity across the model.

Distance: We calculated the distance between governorates using the centroid latitude and longitude obtained from GeoNames.[22]

Case and Sequence counts: We obtained estimates of human and avian H5N1 cases for each governorate from the FAO for the years of 2006-2012.[19] We averaged these to obtain the final predictor values for our model. The sequences incorporated into the phylogeographic analysis were differentiated by the location from which they were isolated for both human and avian sequences. We included these variables not to explain diffusion, but rather to minimize bias on predictors being tested by indicating the sample sizes at particular locations throughout viral spread.

We log transformed and standardized all predictors before their incorporation into the model.

Evaluation of predictor inclusion

Following Lemey *et al.*[11, 16] we determined the support for predictors within the model using Bayes factors (BFs). To calculate the BFs, the posterior odds of predictor inclusion were divided by their prior odds:

$$\text{BF} = \left(\frac{pi}{1-pi}\right)/\left(\frac{qi}{1-qi}\right) \quad [11]$$

Here p_i represents an estimate of the posterior probability that a given predictor is included while q_i represents the prior probability. For this study, the BF cutoff for support within the model was set at 3. We implemented a technique for adjusting β to a fixed correlation $X'X$ in order to account for possible high correlation between predictors. Finally, we evaluated δ under a bit flip operator as discussed by Drummond *et al*. in greater detail.[23]

Results

The BF results suggest the importance of avian populations to the viral diffusion of H5N1 clade 2.2.1.1 in Egypt (figure 1). Most notably, avian population density at the origin had a strong support for inclusion within the model of viral spread with a BF score of 22.3. Additionally, we derived the 95% Bayesian credible interval for the coefficient of each predictor which indicates the level of uncertainty of a particular variable. The inclusion of avian densities at the origin within the model was also supported in this respect, with a credible interval which did not span zero. However, the credible interval for distance, latitude of origin, and human density at the origin did span zero. Compared to avian densities at the origin, human population density did not indicate nearly the degree of support. For both populations the origin achieved a higher probability of inclusion compared to the destination of spread during the observed time period. Other predictors included in the model such as distance between the origin and destination of spread and latitude within Egypt achieved negligible BF scores and inclusion probability. Human density, avian density and latitude at the destination were not supported within model as BF values dropped to approximately 1 or below. Finally, while the variables relating to sample size of sequences and case counts do not directly contribute to the model, their inclusion lends increased credibility for the predictors relating to the avian host data, in particular the avian sequence data which received a BF score of 61.5 and variables associated with the human host obtained unsupportive BF values.

Discussion

Mitigation and prevention of infectious disease is essential to population health, and to achieve these goals we must first understand the processes that drive the spread of viruses such as influenza. Our preliminary work indicates the potential to uncover variables of interest for a particular virus and region, which highlight the integration of

epidemiologic and phylogenetic approaches. Of the tested predictors for H5N1 spread within Egypt, we have found host population densities within the region to be strong indicators for viral dispersal and highly supported for inclusion within the model by BF values. These results are consistent with the nature of close proximity within large populations, and with other findings related to H5N1 risk factors. For instance, Martin *et al*. found that chicken and

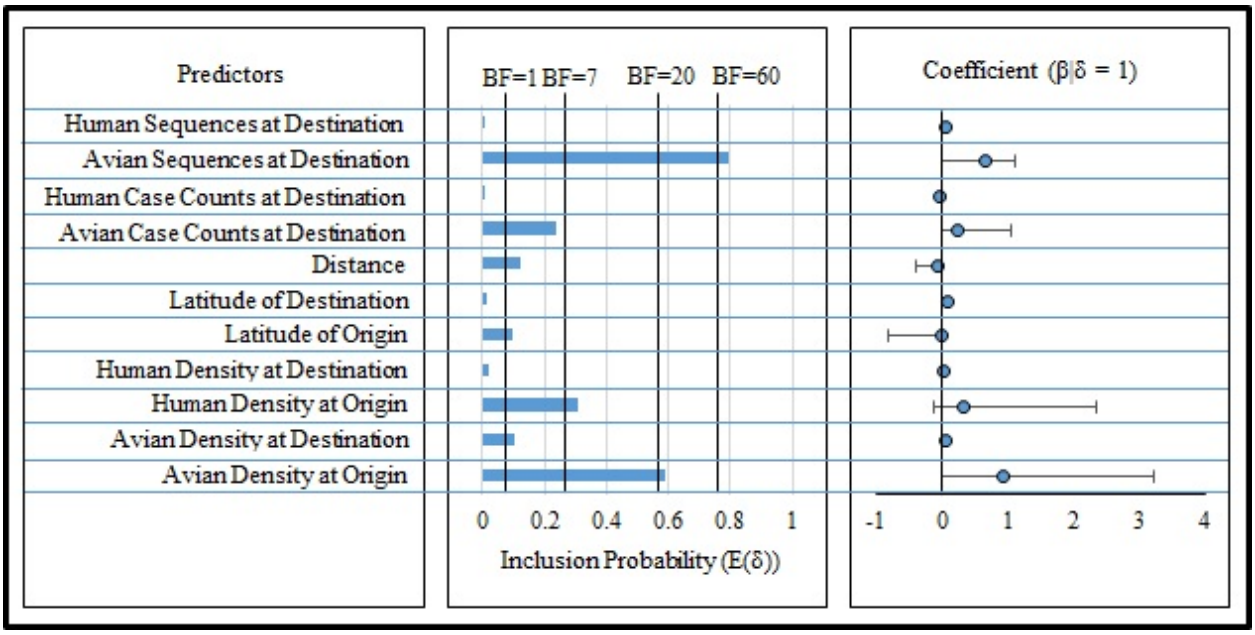

Figure 1. Predictors of H5N1 diffusion in Egypt. Inclusion probability defined by indicator expectations E(δ), which reflects the likelihood of meaningful impact of the predictor on viral diffusion. Bayes Factor (BF) support values shown at the top of the figure and are indicated by vertical lines. Coefficient (β|δ=1) represents the contribution of each predictor, with the 95% credible interval represented by brackets.

human density in China was a leading contributor to risk of infection.[24] However, we do not preclude the possibility of other potential underlying dynamics driving influenza H5N1 in Egypt. While our case study involved influenza, this approach can be applied to other RNA viruses as they have shorter genomes and more rapid nucleotide substitutions compared to other pathogens.[25]

Limitations

There are several limitations of this work, largely related to incomplete or outdated data sources. Our assignment of the centroid of each governorate as the latitude for discrete locations can only approximate the geographic distribution of viral spread. In addition, it is nearly certain the actual number of case counts observed in human and avian populations was not represented as mild cases may go unrecognized. Case counts can also vary year-to-year, possibly indicating the influence of another predictor. This possibility is overlooked using our current method of averaging a range of years. Additional sequencing of collected viruses from known cases would also aid our depiction of the spatial distribution, particularly human sequences as this data is sparse. Finally, estimates of avian population densities used here were collected in 2005, which may over or underestimate actual densities throughout our study period.

Conclusion

We demonstrate the potential of phylogeography and bioinformatics techniques to incorporate traditional epidemiologic data for understanding the evolutionary diffusion of viruses. Future work will involve testing additional variables that are indicated in viral proliferation within Egypt. Predictors of interest include domestic avian population ranges with migratory bird habitat overlap, cross species spill over migration rates, as well as the recent discovery of an important shift in amino acid composition of the hemagglutinin cleavage site to viral pathogenicity within Egyptian strains.[26]

Acknowledgments

The project described was supported by award number R00LM009825 from the National Library of Medicine to MS and by the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement no. 278433-PREDEMICS and ERC Grant agreement no. 260864 to PL. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine or the National Institutes of Health. The authors would like to thank the Arizona State University Advanced Computing Center (A2C2) for the use of the Saguaro supercomputer.

References

1 Viboud C, Bjørnstad ON, Smith DL, Simonsen L, Miller MA, Grenfell BT. Synchrony, Waves, and Spatial Hierarchies in the Spread of Influenza. Science. 2006 April 21, 2006;**312**(5772):447-51.
2 Krauss H. Zoonoses: Infectious Diseases Transmissible from Animals to Humans: ASM Press; 2003.
3 Herrick K, Huettmann F, Lindgren M. A global model of avian influenza prediction in wild birds: the importance of northern regions. Veterinary Research. 2013;**44**(1):42.
4 Van Boeckel TP, Thanapongtharm W, Robinson T, Biradar CM, Xiao X, Gilbert M. Improving Risk Models for Avian Influenza: The Role of Intensive Poultry Farming and Flooded Land during the 2004 Thailand Epidemic. PloS one. 2012;**7**(11):e49528.
5 Tamerius JD, Shaman J, Alonso WJ, et al. Environmental Predictors of Seasonal Influenza Epidemics across Temperate and Tropical Climates. PLoS pathogens. 2013;**9**(3):e1003194.
6 Chu P-Y, Ke G-M, Chen P-C, Liu L-T, Tsai Y-C, Tsai J-J. Spatiotemporal Dynamics and Epistatic Interaction Sites in Dengue Virus Type 1: A Comprehensive Sequence-Based Analysis. PloS one. 2013;**8**(9):e74165.
7 Lam TT-Y, Hon C-C, Lemey P, et al. Phylodynamics of H5N1 avian influenza virus in Indonesia. Molecular ecology. 2012;**21**(12):3062-77.
8 Ypma RJF, Bataille AMA, Stegeman A, Koch G, Wallinga J, van Ballegooijen WM. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. Proceedings of the Royal Society B: Biological Sciences. 2012 February 7, 2012;**279**(1728):444-50.
9 Ypma RJF, van Ballegooijen WM, Wallinga J. Relating Phylogenetic Trees to Transmission Trees of Infectious Disease Outbreaks. Genetics. 2013 September 13, 2013.
10 Grenfell BT, Pybus OG, Gog JR, et al. Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. Science. 2004 January 16, 2004;**303**(5656):327-32.
11 Lemey P, Rambaut A, Bedford T, et al. The seasonal flight of influenza: a unified framework for spatiotemporal hypothesis testing. arXiv:12105877v1. 2012.
12 Scotch M, Mei C, Makoynnen Y, et al. Phylogeography of Influenza A H5N1 Clade 2.2.1.1 in Egypt. Unpublished. 2013.
13 WHO. H5N1 avian influenza: Timeline of major events. 2012.
14 Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular biology and evolution. 2012 Aug;**29**(8):1969-73.
15 Rambaut A. Tracer v1.5 [Internet]. c2009. [updated 2009 Nov 30; cited 2013 Sep 26] Available from http://treebioedacuk/software/tracer/
16 Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian Phylogeography Finds Its Roots. PLoS Comput Biol. 2009;**5**(9):e1000520.
17 Kuo L, Mallick B. Variable Selection for Regression Models. Sankhya. 1998;**60**(1):65-81.
18 Chipman H, George E, McCulloch R. BART: Bayesian Additive Regression Trees. The Annals of Applied Statistics. 2010;**4**(1):266-98.
19 Arafa A. Egypt Clade 2.2.1.1. [online]. E-mail to Matthew Scotch (matthew.scotch@asu.edu). 2013 Aug 12 [cited 2013 Sep 30].
20 FAO. Animal Production and Health Division [Internet]. Global Livestock Production and Health Atlas. c2011. [updated 2013 Mar; cited 2013 Sep 26] Available from http://kidsfaoorg/glipha/indexhtml

21 Egypt CAfPMaS. Arab Republic of Egypt [Internet]. c2012. [updated 2012 Jul 05; cited 2013 Sep 26]. Available from http://www.citypopulation.de/Egypt.html.
22 Geonames.org. [Internet]. Egypt. c2013. [updated 2013 Apr 30; cited 2013 Sep 26] Available from http://wwwgeonamesorg/EG/administrative-division-egypthtml
23 Drummond A, Suchard M. Bayesian random local clocks, or one rate to rule them all. BMC Biology. 2010;**8**(1):114.
24 Martin V, Pfeiffer DU, Zhou X, et al. Spatial Distribution and Risk Factors of Highly Pathogenic Avian Influenza (HPAI) H5N1 in China. PLoS pathogens. 2011;**7**(3):e1001308.
25 Holmes EC. The phylogeography of human viruses. Molecular ecology. 2004;**13**(4):745-56.
26 Yoon S-W, Kayali G, Ali MA, Webster RG, Webby RJ, Ducatez MF. A Single Amino Acid at the Hemagglutinin Cleavage Site Contributes to the Pathogenicity but Not the Transmission of Egyptian Highly Pathogenic H5N1 Influenza Virus in Chickens. Journal of Virology. 2013 April 15, 2013;**87**(8):4786-8.