Modeling rodent-virus interactions in North America from different biodiversity dimensions through machine learning

Angel Robles
Nathan Upham

April 2022

Introduction

Understanding the eco-geographic ’rules’ of host-pathogen interactions is needed for developing fine-grained tools for predicting the risk of different wildlife pathogens being transmitted to humans (i.e., spillover risk; [6]). Biodiversity conservation efforts similarly benefit from understanding these general rules of community assembly, which predict expected host-host, host-environment, and host-pathogen interactions under different amounts of human disturbance.

Recent studies spanning a diverse array of host-parasite interaction types have demonstrated that a large amount of variation in host breadth or host specificity is explained by host phylogenetic relationships (e.g., beetle-plants, [12]; bird-malaria, [2]; mammal-flaviviruses, [13]; mammal-ectoparasites [4]). Sometimes environmental variables such as temperature and precipitation also explain a good amount of variation in host breadth (e.g., 30-50 % in bird-malaria, [5]). Thus, the use of analytical statistical tools can take advantage of these relationships in phylogenetic, environmental, and geographic traits to predict interactions for specific host-parasite systems and generate geographic assessments of infection risk of a focal species by the pathogen of interest (e.g.[12]).

Information currently available regarding the interaction of hosts with their pathogens is generally scarce and pertains to a few geographic areas. This in essence imposes a limitation on our capacity to predict such interactions and have forecasts about regions where they may be happening now or in the future. Such limitations may, however, be solved by the use of information about known hosts and their phylogenetic, geographic, and environmental distances with other (target) species of interest to make inferences regarding host-pathogen interactions at different levels.

In the context of large volumes of data about different dimensions of biodiversity, patterns within these volumes can be discovered using a suite of tools based on statistics, machine learning, and artificial intelligence [1]. Using classification algorithms such as random forest or support vector machines, it is possible to find deep relationships among biodiversity dimensions of known hosts, which can then help identify additional likely susceptible hosts. By keeping track of the performance of the different algorithms and selecting the best model to evaluate predictions about new data sets, it is possible to identify the assembly of possible pathogens on different assemblages of hosts [9][8]. Thus, this analytical approach provides an effective means of leveraging current scientific knowledge of the relationships between pathogenic organisms and their hosts to geographically evaluate the suitability of this assembly to other conditions, which can then inform decision-makers on the potential hazards of different systems.

In conducting this research, which will be targeted on North American rodents and Hantaviridae, I propose to develop an analytical approach whereby it is possible to: (1) evaluate the probability of new potential hosts to different pathogens given the phylogenetic, geographic, and environmental distances of known host-pathogen interactions according to different artificial intelligence (AI) algorithms (e.g., random forests, convolutional neural networks, support vector machines), and (2) provide geographically explicit scenarios of host susceptibility according to different taxonomic and phylogenetic assemblages, which can be interpreted as a spatial index of potential interactions between hosts and pathogens. This research will provide both an analytical method to quantify the probability of host-pathogen interactions at coarse spatial resolutions and give results that can be incorporated in decision-making. These outputs will also provide hypotheses for validation in future theoretical and field research.

Study scope

This study aims at fostering innovative research and discovery in biodiversity informatics, where the overarching question is to provide of robust statistical and analytical information to have a broad perspective on host-pathogen interactions. To reach the goal in this research I propose to build a taxonomically harmonized database of multidimensional biodiversity information, which will be the basis for analysis of pairwise distances among species and predictions of potential host-pathogen interactions. This database will contain summarized geographic, phylogenetic, and environmental information between pairs of species. The design of this database involves the use of expert range maps for 300 mammal species, as modified from the IUCN 2017 [https://www.iucnredlist.org/resources/spatial-data-download] database to match the phylogeny of Upham et al. 2019 [14] [https://data.vertlife.org]. This coarse-grain geographic information will be used to extract environmental correlates from the CHELSA project [https://chelsa-climate.org/] (which is more fine-grain than the data of WorldClim [7]).

The rodent-virus assemblages will be considered for this study:

- North American Rodentia (including members of the families Aplodontiidae, Castoridae, Cricetidae, Dasyproctidae, Erethizontidae, Geomyidae, Heteromyidae, Muridae [invasive *Rattus*], Sciuridae, and Zapodidae

[3][https://www.mammaldiversity.org]); and

- Hantaviridae (including members of the subfamilies Actantavirinae, Agantavirinae, Mammantavirinae, and Repantavirinae [10] urlhttps://talk.ictvonline.org/taxonomy/]).

Plan

The stages of the project are specified in the following diagram. The results of this study will be summarized in two peer-reviewed articles. The first seeks at showing the power of models to predict different sets of interactions in a common framework. The second aims at unifying the framework and methods used in a package developed in the R language so that the interested community can apply these methods more broadly.

Methods

Data preparation and model calibration. To model each host-pathogen system, a count of pathogen incidences will be obtained for all host species according to Global Interactions database [https://www.globalbioticinteractions.org] [11] . Here, statistical considerations about pathogens' incidences and prevalence will be taken in order to select the initial set of data needed for model calibration. In order to avoid biases in the sampling effort and to generate a common framework between databases, the optimum of this empirical probability distribution is selected to concentrate the host-pathogen assembly in the least amount of hosts.

The previous step will provide classifications of either "susceptible" or "unknown" species. These data are the input of the machine learning algorithms (i.e., the dependent variables) together with the phylogenetic, geographic, and environmental distances between the pairs of species (i.e., independent variables). In this way, the susceptibility of a species to an infection is being modeled based on its interaction with an assembly of other species.

Comparisons and evaluation of models. A common framework is proposed to compare different models that allow a better classification of this information and thus always obtain the best possible classification. Each algorithm will be re-sampled with repeated cross-validation and all the models will be optimized towards the same metric to guarantee an ideal comparison between them. Finally, validation metrics of the resulting classification will be used to select the best model among the various algorithms. We propose to use state-of-the-art classification algorithms such as general additive models, extreme gradient boosting, support vector machines, and neural networks.

Distribution of host-pathogen susceptibility. Finally, with the selected model, the probability of susceptibility for the pairs of species will be collapsed to provide an average susceptibility for the assembly of that species with respect to the others. The end result is the probability of pathogen susceptibility for each species. Taking this measure of susceptibility, we can then project the information on geographic space using the known (or potential) distribution of the species and their susceptibility values. Subsequent treatment to this data can be done in order to show different output scenarios, including patterns of the richness of these species' assemblages or their associated probabilities of interaction.

Preliminary results

In the figure 1 we show the power law cummulative density function from Hantavirus counts over rodents. In the figure 5 we can observe the probability density functions for the three distance in our model given the class assigned to host that has at least one record and that hsot that doesnt has a record in Globi. On the figure 3 we can observe the variable importance boxplot, where the blackline is the median value. On figure 4 we can observe the probability density function for the ROC values for each model. The median value is 0.816. This mean that the median classification for all models is acceptable

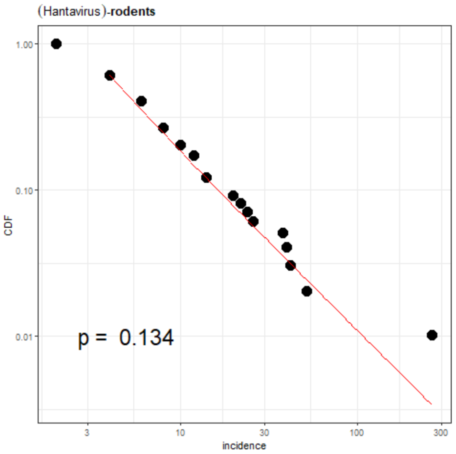

Figure 1: Hantavirus - rodents system model under this protocol

References

[1] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2020.

[2] Nicholas J Clark and Sonya M Clegg. Integrating phylogenetic and ecological distances reveals new insights into parasite host specificity. *Molecular ecology*, 26(11):3074–3086, 2017.

[3] Mammal Diversity Database. Mammal diversity database, February 2022. This is a real-time upload of the MDD v.1.8 taxonomy published 1 February 2022 on the mammaldiversity.org website.

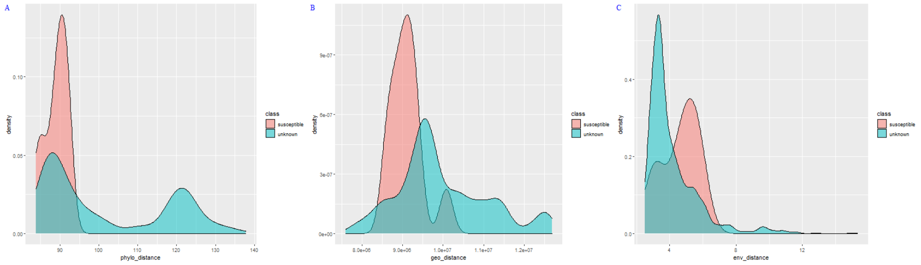

Figure 2: Density functions of distance for rodents that has been labeled as susceptible (i. e. one incidence recorded) and unknown (i. e. random sampled from the rodent species pool with no records)

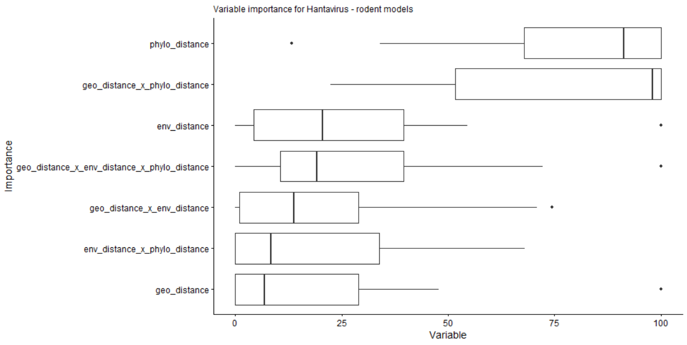

Figure 3: Variable importance from the models outputs after 1000 runs

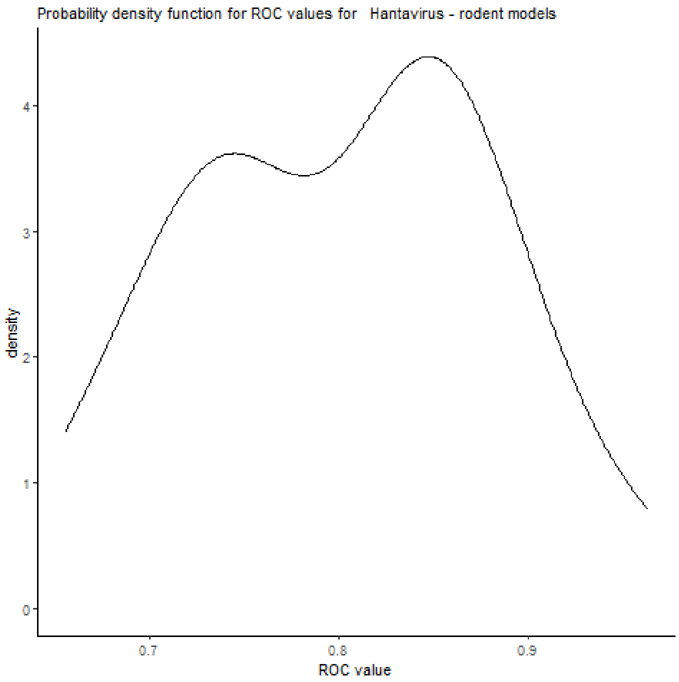

Figure 4: Probability density function of ROC values for each model

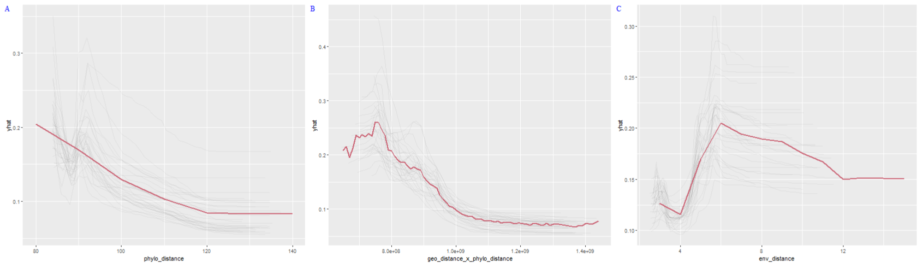

Figure 5: Partial dependent plots for the first most importan variables according to 3

[4] Wesley Dáttilo, Nathalia Barrozo-Chávez, Andrés Lira-Noriega, Roger Guevara, Fabricio Villalobos, Diego Santiago-Alarcon, Frederico Siqueira Neves, Thiago Izzo, and Sérvio Pontes Ribeiro. Species-level drivers of mammalian ectoparasite faunas. *Journal of Animal Ecology*, 2020.

[5] Alan Fecchio, Konstans Wells, Jeffrey A Bell, Vasyl V Tkach, Holly L Lutz, Jason D Weckstein, Sonya M Clegg, and Nicholas J Clark. Climate variation influences host specificity in avian malaria parasites. *Ecology letters*, 22(3):547–557, 2019.

[6] Gregory S Gilbert, Roger Magarey, Karl Suiter, and Campbell O Webb. Evolutionary tools for phytosanitary risk analysis: phylogenetic signal as a predictor of host range of plant pests and pathogens. *Evolutionary applications*, 5(8):869–878, 2012.

[7] Dirk Nikolaus Karger, Olaf Conrad, Jürgen Böhner, Tobias Kawohl, Holger Kreft, Rodrigo Wilber Soria-Auza, Niklaus E. Zimmermann, H. Peter Linder, and Michael Kessler. Climatologies at high resolution for the earth's land surface areas. *Scientific Data*, 4(1):170122, Sep 2017.

[8] Max Kuhn et al. Building predictive models in r using the caret package. *Journal of statistical software*, 28(5):1–26, 2008.

[9] Brett Lantz. *Machine learning with R: expert techniques for predictive modeling*. Packt Publishing Ltd, 2019.

[10] Elliot J Lefkowitz, Donald M Dempsey, Robert Curtis Hendrickson, Richard J Orton, Stuart G Siddell, and Donald B Smith. Virus taxonomy: the database of the international committee on taxonomy of viruses (ictv). *Nucleic acids research*, 46(D1):D708–D717, 2018.

[11] Jorrit H Poelen, James D Simons, and Chris J Mungall. Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecological Informatics*, 24:148–159, 2014.

[12] Ángel L Robles-Fernández and Andrés Lira-Noriega. Combining phylogenetic and occurrence information for risk assessment of pest and pathogen interactions with host plants. *Frontiers in Applied Mathematics and Statistics*, 3:17, 2017.

[13] Jesús Sotomayor-Bonilla, María José Tolsá-García, Gabriel E García-Peña, Diego Santiago-Alarcon, Hugo Mendoza, Paulina Alvarez-Mendizabal, Oscar Rico-Chávez, Rosa Elena Sarmiento-Silva, and Gerardo Suzán. Insights into the host specificity of mosquito-borne flaviviruses infecting wild mammals. *EcoHealth*, 16(4):726–733, 2019.

[14] Nathan S Upham, Jacob A Esselstyn, and Walter Jetz. Inferring the mammal tree: species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLoS biology*, 17(12):e3000494, 2019.