Paul English
BIO 1615-002
Larry Jones
July 14, 2013

Summary of GEOGRAPHIC AND ECOLOGIC DISTRIBUTIONS OF THE *ANOPHELES GAMBIAE* COMPLEX PREDICTED USING A GENETIC ALGORITHM

1 Introduction

Anopheles gambiae is a complex of mosquitoes that are a vector for human malaria.[1,6] A complex represents a group of species that are close relatives, or sibling species.[4] A biological vector is a species or organism that helps to spread specific pathogens amongst hosts.[9] Understanding and predicting the geographic distribution of this complex helps us to recognize, treat, contain and prevent possible outbreaks of malaria in human populations. We can use predictive modeling techniques to estimate probable geographic distributions of this complex. Previous models have used non-linear statistical models along with spatial mapping to help understand what attributes of an area or ecological niche are beneficial to the Anopheles gambiae complex.[10]

A genetic algorithm is a method of parameter optimization for statistical modeling or similarly machine learning problems. It's composed of a series of successive simulations on a problem in an attempt to naturally select the best fitting model inputs using ideas of evolution, including inheritance, and mutation.[5] Use of a genetic algorithm can help build robust probability distributions for a species in a geographic area. Use of this kind of model can help to define rules that the A. gambiae complex will follow, as well as predict where it could grow or live.[10]

2 Materials and Methods

Using data collected from past research and studies a combined data set was created. Out of this data set of 14 common features, 12 were selected to generate the predictive model of this experiment. These attributes represent characteristics of an environment that can either help or hinder the development of Anopheles gambiae. The selected 12 attributes include the mean annual temperature, mean annual maximum temperature, mean annual minimum temperature, daily temperature range, occurrence of frost days, topographic aspect, flow accumulation, topographic index, mean annual precipitation, occurrence of wet days, tree cover, and land use/cover.[10]

This data set was partitioned by country, and a random selection of half of the countries was used to build, allowing the remaining half to help validate the accuracy of the predictions. A tool called the Genetic Algorithm for Rule-set Prediction, or GARP, was used to build the statistical models that help predict geographic distribution. 100 models were generated, allowing final predictions to be made by averaging the top 5 most accurate models. Model accuracy was tested by comparing model results to a chi-squared distribution.[10] The chi-squared distribution $\chi^2(x|v)$, is a commonly used probability distribution in assessing statistical fit.[1,3,7,11] It is special case of the Gamma distribution $Ga(T|\text{shape} = a, \text{rate} = b)$, where rate is fixed at $\frac{1}{2}$ and the shape is given by dividing a parameter, degrees of freedom v, by 2, $\chi^2(x|v) = Ga(x|\frac{v}{2}, \frac{1}{2})$.[12] The Gamma distribution has been used in life testing, and waiting time until death in other disciplines.[7] Using test data reserved from building the model produced results that appear to be significant using this test, represented by a low P-value.[10]

3 Results

Final predictions were created using all available data. It showed to be generally consistent, though discrepancies were found in the distribution of A. arabiensis between the predicted results, that of Lindsay and others, and the generalized distribution map in the central Africa region. This would mean that either the predicted maps for A. arabiensis are incorrect, the generalized map is incorrect, or both maps correctly identify certain potential and realized niches, but other unidentified factors lead to a discrepancy.

In predictions for A. gambiae, other differences can be seen in Northeast Africa, lower ethiopian elevations, North Kenya, South Coastal Somalia, and Southeast Sudan. Previous predictive efforts have used similar datasets, but have had very different modeling approaches, which offer contrasting results. Some differences of the modeling approaches can be seen in sparsely sampled areas like central Africa, where the GARP model predicted broad areas of presence for A. gambiae and A. arabiensis.

Suggested differences may be due to previous regression based attempts having the inability to infer in unsampled areas. Obvious errors can be seen in previous regression based predictions. Older regression attempts lack independent model validation, have uneven interpretation of kappa statistics, and over-reliance on assumptions of small samples of data.[10]

4 Discussion

We can recognize some genetic breaks within A. quadriannulatus, given discovery of a sibling species in Ethiopia. We can see that A. gambiae inhabits wetter and warmer environments than other species in the complex. We can measure the individual impact of specific features by selectively excluding features during assessment. This would suggest that the occurrence of frost days strongly influences the predictions for A. gambiae. We can see that both climatic and topographic factors influence A. quadriannulatus. We can also see that no unique attributes independently affect the distributions of A. arabiensis.

Using climate data from 1960-1990 Africa regions, projected back to 1930-1960 climate data from the Americas this modeling approach suggests that A. gambiae would have a widespread suitable habitat during the historical outbreak of Northeast Brazil in the 1930s. This outbreak occurred before it was understood that A. gambiae was a complex of multiple species. Efforts that helped stop the outbreak in Northeast Brazil were effective, and without which later eradication might have been impossible.

Discrepancies in this predicted distribution in the America's may occur for many reasons not related to modeling method or data, including local elimination, geographic climatic barriers, and biotic interactions not considered. Additionally a models precision may overrepresent one pixel as suitable when in reality only a small portion of that region is suitable. Model representations of standing bodies of water, rivers, streams, and areas subject to flooding result in over-prediction when modeling species that do not migrate.

Implications of new modeling methodologies are numerous including; species distributions in broader areas, inference of a species' distributional potential in other regions, how a species distribution may change given global climate change, and implications for potential bio-terrorism applications.[10]

5 Sources

1. Abramowitz, Milton, and Irene A. Stegun. "Chapter 26." *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. New York: Dover Publications, 1970. 940. Print.
2. "Anopheles." *Dictionary.com*. Dictionary.com, n.d. Web. 14 July 2013.
3. "Chi-Square Distribution." *Engineering Statistics Handbook*. NIST, 2006. Web. 14 July 2013. <http://www.itl.nist.gov/div898/handbook/eda/section3/eda3666.htm>.
4. "Cryptic Species Complex." *Wikipedia*. Wikimedia Foundation, 07 Apr. 2013. Web. 14 July 2013.
5. "Genetic Algorithm." *Wikipedia*. Wikimedia Foundation, 07 Sept. 2013. Web. 14 July 2013.
6. Giles, George Michael James. *A Handbook of the Gnats or Mosquitoes, Giving the Anatomy and Life History of the Culicidæ Together with Descriptions of All Species Noticed up to the Present Date*. London: J. Bale, Sons & Danielsson, 1902. 530. Print.
7. Hogg, Robert V., and Allen T. Craig. "Remark 3.3.1." *Introduction to Mathematical Statistics*. New York: Macmillan, 1978. N. pag. Print.
8. Johnson, Norman Lloyd, Samuel Kotz, N. Balakrishnan, Norman Lloyd Johnson, Norman Lloyd Johnson, Samuel Kotz, and Samuel Kotz. "Chapter 18." *Continuous Univariate Distributions* Vol. 1. New York: Wiley, 1994. N. pag. Print.
9. Last, John M., and Robert A. Spasoff. *A Dictionary of Epidemiology*. New York: Oxford UP, 2001. 185. Print.
10. Levine, Rebecca S., A. Townsend Peterson, and Mark Q. Benedict. "Geographic and Ecologic Distributions of the Anopheles Gambiae Complex Predicted Using a Genetic Algorithm." *The American Journal of Tropical Medecine and Hygiene* 70.2 (2004): 105-09. PubMed. Web. June 2013.
11. Mood, Alexander McFarlane, Franklin A. Graybill, and Duane C. Boes. *Introduction to the Theory of Statistics*. 3rd ed. New York: McGraw-Hill, 1973. 241-46. Print.
12. Murphy, Kevin P. "Probability." *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT, 2012. 41-42. Print.