Machine Learning to Identify Likely Reservoir Hosts of *Leishmania* Species

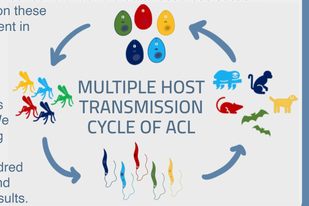
Aisling Murran, CKG, BH, BC, EM I Department of Biology I amurran@stanford.edu
Mordecai Lab, Stanford University

Introduction

Two of the biggest threats to human health, zoonoses and global change, are not separate issues. Land-use and climate change shifts the range of reservoir hosts, altering contact patterns and the likelihood of disease spillover. Consequently, global change is spurring an increase in the number of new cases and ranges of many dangerous diseases. One such disease is American Cutaneous Leishmaniasis, or ACL, a tropical disease which continues to spread throughout Central and South America, in urban, rural and forested areas. But, despite the widespread threat to human health, ACL is ill-contained and under-researched. This is because ACL transmission is highly complex: It involves over 15 species of intracellular protists and a diverse group of reservoir hosts - everything from rodents, to sloths. Though we've identified a handful of reservoirs, the full range of hosts for each species of *Leishmania* is poorly resolved. Without a clear map of the transmission cycle, effectively pinpointing what factors affect the risk of outbreaks is incredibly challenging. We need all the puzzle pieces, or at least most of them, if we are to have a hope of understanding the full picture. This is the issue our research addresses.

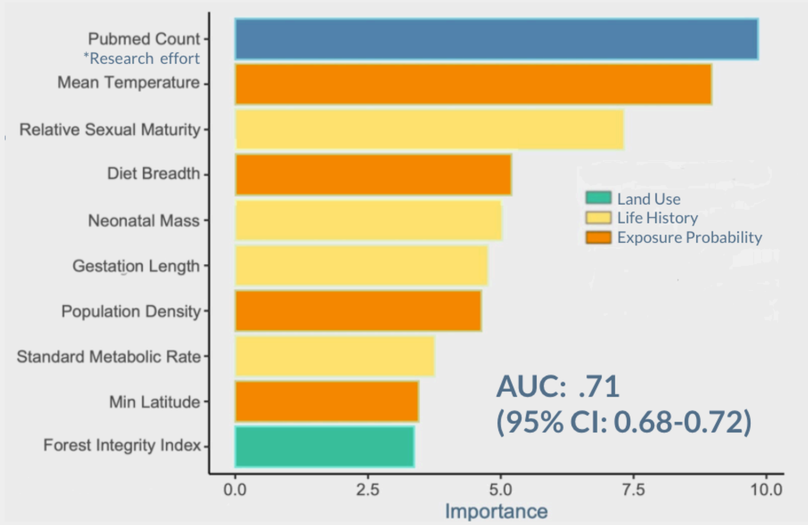
Methodology

The goal is to identify overlooked links in the ACL transmission cycle. To do so, we used a boosted tree to identify the most probable hosts. The idea is to predict which mammals could *potentially* harbor the pathogen based on the traits they have in common with species we *know* harbor the disease. As a bonus, we are simultaneously identifying the traits which have made these known reservoirs successful. Both kinds of data can be invaluable to directing further research and predicting future outbreaks. For our preliminary study we chose a boosted tree because the predictions are easier to work with and it's resilient against over-fitting. But before we could get to the fun part of programming, we first compiled a comprehensive list of known and potential reservoir hosts for each *Leishmania* species that causes ACL. Then, we created a matrix of common traits related to reservoir host capacity such as life history traits, or environmental range. More information about our sources can be found under the data resources section. Finally, we collected information on these traits for as many mammals present in Central and South America as possible. After processing, this came out to 947 mammals and around 40 traits. Having cleaned, reconfigured and transformed this data, we applied the algorithm. We split our data up with 85% making up our training set and 15% reserved for testing. Then, a hundred models were generated, tuned and averaged to calculate our final results.

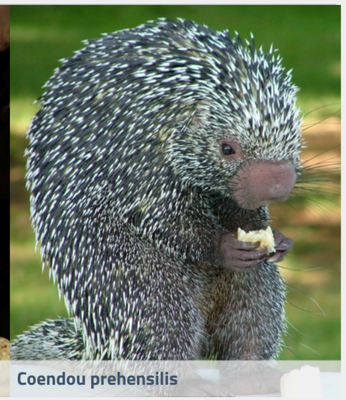


Results

Our first results relate to the strength of our model. As you can see in the bottom corner of the chart on the right, the AUC from our train and test set is nicely situated at .70 and the sensitivity is 95%. AUC measures our models ability to differentiate between true and false positives with .5 being randomly assigning. Our model wasn't designed to prove, only to predict, so these values are important in order to instill confidence in the accuracy of the model's predictions. Satisfied our model is performing well, we can dig into the more interesting results. First up are the animals our model predicted to be potential reservoirs. The two mammals depicted below were in the top 5% of ours models predictions. Neither of these mammals have been shown to be a reservoir for Leishmaniasis, but overlap in numerous ways with other reservoirs. Because these mammals could be contributing to the spread of ACL, we'd suggest further research into both mammals, among others identified by our model.




Akodon azarae


Coendou prehensilis

On the flip side is our significant traits. The top ten most significant traits identified by the model are depicted above with their corresponding importance. Many of the top traits are indicators of similar life histories, which is consistent with findings in other papers - for example, relative age of sexual maturity. Other factors, can be related to ease of disease transmission. For example, the mean temperature of the mammal's environment is significant because a mammal is more likely to host ACL if it lives in a zone where the temperature is ideal for *Leishmania* species. To me, though, the most interesting result was this last one, a trait representing forest integrity in the area the mammal occupied. It made it into our top ten of more than forty traits. Although there are many possible explanations, this is very compelling evidence that land-use change could be an integral factor in the continued propagation of ACL.

Data Resources

The data used for our model was drawn from multiple sources. Before collecting any of our data we needed to identify and narrow down what traits were of interest. Reading through research papers on the topic of machine learning in disease transmission allowed us to identify two major categories of data: traits which are traditionally used to describe and distinguish different mammals, and traits which are directly related to disease transmission. With these two categories in mind, we moved to data collection. For traits intrinsic to mammals, we primarily used the panTHERA database. For traits related to geographical location, we used spatial maps from the IUCN Red List. For diet information we used the MammalDIET database. Lastly, we used GIDEON for data related to diseases in mammals. It was difficult to find a database for other information we needed, such as whether or not a mammal was a reservoir host. Instead, this kind of data was found by compiling information from research papers in the field.

Moving Forward

Ultimately, our model predictions can enable more targeted field research, facilitate better preventative management, and improved hypotheses regarding the effect of global change on disease transmission. I'm excited to continue to unpack the potential for machine learning as a predictor of reservoir hosts. Moving forward we'll be trying to reproduce our results for each individual species of *Leishmania* using another algorithm outlined in *Han et al. 2019* which is better suited to handle unbalanced and sparse datasets.

Acknowledgements

A big thank you to Dr. Caroline Glidden, Dr. Barbara Han, Dr. Bruno Carvalho, Professor Erin Mordecai, Gowri Vadmal and all of my colleagues at the Stanford University Mordecai Lab for their invaluable support.

I am also very grateful for funding from the National Science Foundation, National Institute of Health, and Stanford Small Grants.

Stanford