PREDICTION OF ECOLOGICAL FUNCTION IN THE MICROBIOME USING MACHINE LEARNING ON THE GRAPH SPECTRA OF COEVOLVING SUBNETWORKS

Russell Y. Neches
Matthew D. McGee
Peter C. Wainwright
Jonathan A. Eisen

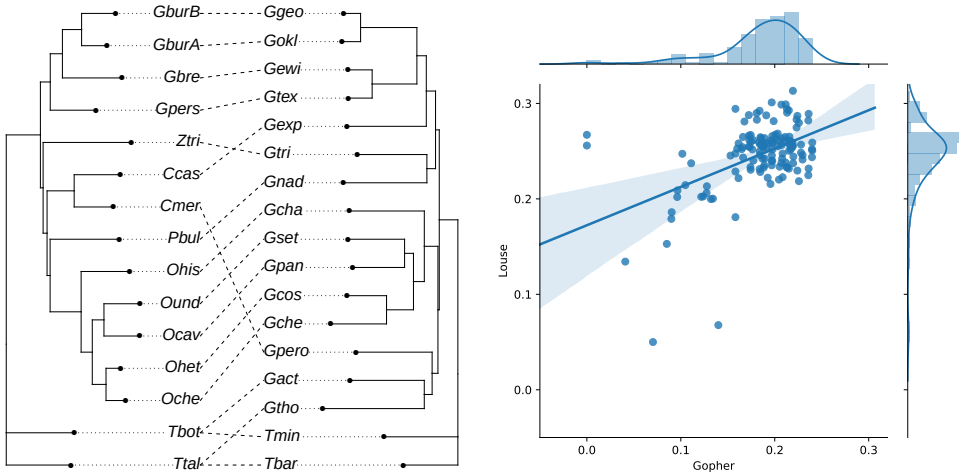
March 12, 2018

Biology has a major problem with charisma. Humans relate to some organisms more easily than others. Fuzzy creatures get more attention than scaly or slimy ones. Creatures with faces get more attention than creatures without. Animals get more attention than plans. By abundance, diversity, biomass, age or metabolic wattage, eukaryotes make up a tiny fraction of life on Earth, but they occupy almost all of our attention. The same phenomenon happens in ecology.

Ecology is the study of how organisms interact with one another and with their environment. Some relationships are more charismatic than others, and those relationships dominate our attention. There are trophic strategies in myriad diversity, but predator-prey interactions make the best television. Parasites give us the creeps as much as they fascinate us. Mutualisms speak powerfully to aspirations and anxieties regarding our own societies. Disease has shaped and reshaped nations, and drives a large part of the moral imperative behind biological research. Nevertheless, the great majority of ecological relationships are none of the above, or cannot neatly fit into any one category.

Charisma is one of many heuristics that help us identify things that are likely to be relevant to our own experience. Heuristics are useful, but usefulness should not be mistaken for accuracy. The fact that charisma cannot be separated from the observer means that it generalizes poorly. The fact that charisma is a heuristic measure of importance means that it can still be wrong more often than right. It is a form of bias, and distorts the model of reality we use to understand how things work.

Biologists address charisma bias by applying other metrics for importance, and often look to ecology for parameters to include in these metrics. Importance is contextual, after all, and ecology is the study of the relationships that comprise the context in which organisms live. How, then, can ecology correct for its own charisma bias? Relationships can be categorized by their effects or their dynamics, but unless they exhibit a charismatic property – often a symmetry in structure, a simplicity of concept or an analogy to human experience – they can defy categorization or escape



(a) Cophylogeny of gophers and lice

(b) Paired patristic distances

Figure 1: The relationship between pocket gophers and their chewing lice parasites has served as a benchmark case in the literature on coevolution since its appearance in Hafner *et al.* [1]. However, despite the strong case for coevolution from multiple lines of evidence, the agreement between the two trees (as measured by the correlation of pairwise patristic distances through the two trees [2]) is modest, with a Pierson's r of 0.49. If one were to exclude the relationships between the outgroups as outliers, the correlation would collapse. Without other forms of evidence, the detection of such relationships is challenging.

notice altogether. The symmetry of the Red Queen's Race, the straightforwardness of a predator's relationship with its prey, or the (projected) virtue of cooperation are somewhat unusual properties to find in the natural world. Most relationships are too complicated or ambiguous to clearly exhibit them. They are not charismatic.

This is especially evident in microbial interactions, where high-throughput sequencing has made it possible to generate spectacular quantities of complex, nuanced and mostly inconclusive data. This is frustrating, but the scale of the problem represents a unique opportunity to address charisma bias in ecology.

Correcting a bias always begins with the same task : find a way to collect data in a way this isn't subject to the bias. Microbiome surveys are vulnerable to a variety of technical biases – there is *always* another bias – but sequencing machines are not impressed by charisma. Based on this data, it is possible to construct and test new metrics of importance, and to see which relationships merit further attention.

Unfortunately, we do not have very many theoretical tools that address the question of how to categorize ecological relationship gathered using an non-targeted, unsupervised process.

Fortunately, ecology is not the only field that cares about the problem of detecting and assigning categories to complex relationships from non-targeted data. Vast resources have been devoted to this problem in the form of social networks among human beings. The theoretical approaches and software developed for this purpose can be directly applied to microbiome data.

Microbiome data yields two pieces of information : the evolutionary relationships among the organisms, and the observed interactions among those organisms. The first pertains to how the organisms have evolved, and the second pertains to how they interact. Evolutionary relationships tells us about what happened over deep time, and ecological relationships tell us about what was happening at the moment the samples were collected. There are many ways structure this information, and each carries an implicit opinion about what properties of these relationships are important.

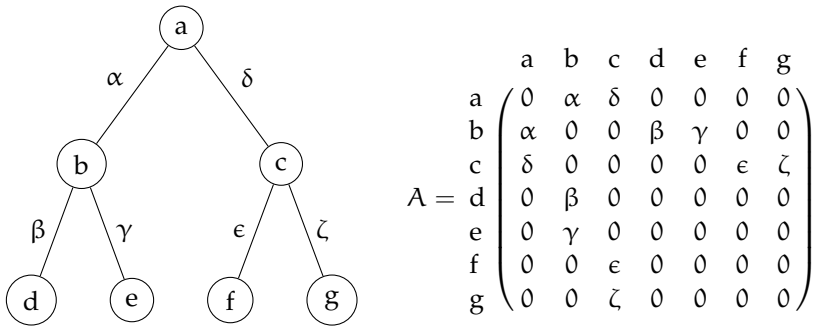
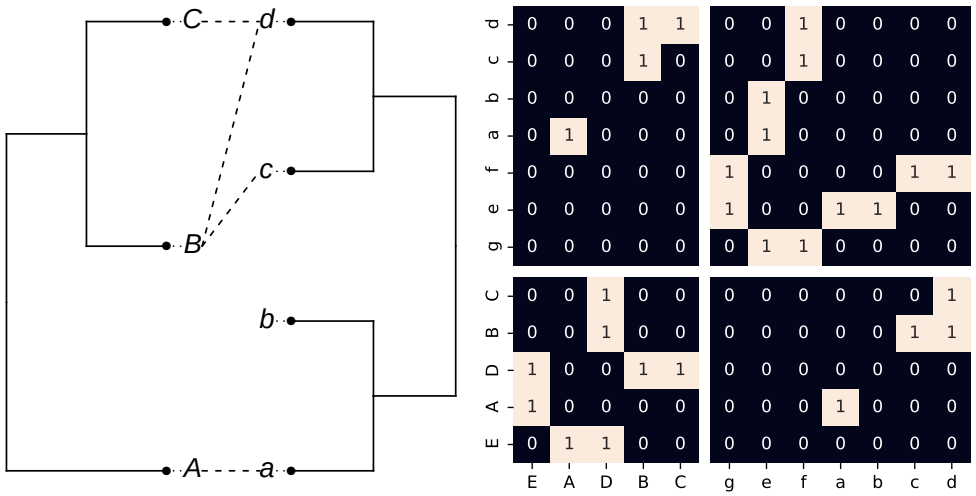


Figure 2: Construction of the graph adjacency matrix for a phylogenetic tree.



(a) Cophylogeny graph **(b)** Adjacency matrix

Figure 3: The construction of the graph adjacency matrix of a cophylogeny.

One way to structure this information to build phylogenetic trees for each group of organisms involved in the interaction, and note which organisms at the tips of these trees were observed to interact with which other organisms. This is called a co-phylogeny, and is a model found in the literature on host-parasite interactions and on the reconciliation of gene trees into species trees. Usually, there are two trees linked by interactions, as with host-parasite interactions.

This structure contains an implicit opinion that relationships that are "important" now were "important" in the past, and as such, may have influenced the evolution of the organisms in involved. Exactly what is meant by "important" depends on what calculations one performs, but it should be acknowledged that simply posing the problem this way is not a neutral choice. There are other valid ways to think about this information.

Although it is not often done this way, there is no reason to restrict co-phylogeny to two trees. For example, *Leucochloridium* is a genus of flatworms that parasitize land snails, causing the eyestalks of the snail to resemble caterpillars. The snails are then eaten by birds, and the bird completes the parasite's life-cycle by dispersing its eggs. A complete co-phylogeny of this relationship would have phylogenetic trees of the parasites, the snails and the birds linked by observations of their respective interactions. This approach can be generalized to systems like hot-springs, with members of each major clade of organisms linked by the sites where they were observed to co-occur.

This is a complex way to structure this data, but it does lend itself to a formalized mathematical treatment as a labeled graph. Once placed into this form, a whole new

stack of tools becomes available. Dimension reduction and feature extraction methods like graph spectral analysis, graph kernels, extreme sets and random walks can be used to project high-dimensional data about ecological and evolutionary relationships into a common feature space suitable for supervised learning. Simulation data can be projected into the same feature space for unsupervised learning.

There are downsides to working in these abstract spaces. Particularly, the intermediate steps can be difficult to interpret intuitively. Nevertheless, if the goal is to address charisma bias, placing *all* of the available data into the same context is a good start.

Machine learning does not eliminate bias; it formalizes it. It provides a framework within which one can select, quantify, explore and hopefully understand bias. Perhaps a formal approach will simply confirm the categories we already use to understand ecological relationships. Perhaps it will suggest new categories.

1 FUNDING

RYN was funded by a grant from the Alfred P. Sloan Foundation to Jonathan A. Eisen.

References

[1] Mark S Hafner et al. "Disparate rates of molecular evolution in cospeciating hosts and parasites". In: *Science* 265.5175 (1994), p. 1087.

[2] Kerstin Hommola et al. "A permutation test of host–parasite cospeciation". In: *Molecular biology and evolution* 26.7 (2009), pp. 1457–1468.