THINKING OUTSIDE THE BOX - PREDICTING BIOTIC INTERACTIONS IN DATA-POOR ENVIRONMENTS

DAVID BEAUCHESNE[1*], *PHILIPPE DESJARDINS-PROULX*[2], *PHILIPPE ARCHAMBAULT*[1], and *DOMINIQUE GRAVEL*[2]

[*] *email: david.beauchesne@uqar.ca*
[1] *Université du Québec à Rimouski*
[2] *Université de Sherbrooke*

September 8, 2016

0 Abstract

Large networks of ecological interactions, such as food webs, are complex to characterize, be it empirically or theoretically. The former requires exhaustive observations, while the latter generally requires ample data to be validated. Although large empirical datasets describing pairwise consumer-resource interactions are increasingly available, full descriptions of entire ecosystems are still few and far between. We therefore wondered whether available data, albeit originating from different ecosystems, could be used to predict species interactions in data deficient ecosystems. *A potential resolution for this issue lies in the taxonomic proximity between species, which increases their likelihood of sharing both consumers and resources.* To test this, we built a biotic interaction catalogue from a collection of *94* empirical food webs, detailed predator-prey interaction databases and interactions from the Global Biotic Interaction (GloBI) database. We use an unsupervised machine learning method to predict interactions between any given set of taxa, given pairwise taxonomic proximity and known consumer and resource sets found in the interaction catalogue. Initial results suggest that pairwise interactions can be predicted with high accuracy, *opening promising avenues for further research and for management initiatives*. The results are however dependent on the comprehensiveness of the catalogue rather than taxonomic proximity between taxa, with prediction accuracy decreasing concomitantly with information available through the catalogue. Given it's high accuracy, this methodology could democratize the use of food webs and network level descriptors in remote location where empirical data is hard to gather. Network characteristics could then be efficiently evaluated and correlated to levels of environmental stressors in order to improve vulnerability assessments of ecosystems to global changes.

1 Introduction

To include in introduction: interactions can also be predicted by functional traits such as body size and metabolism type (e.g. Gravel et al 2013;), yet even those can be difficult to extensively characterize for multiple taxa. We therefore sought more readily available information for all taxa of interest.

We therefore turned our attention towards phylogenies, taxonomies, to get more readily available data for each taxon of interest

The study of ecosystem structure and function is increasingly focusing on communities of interacting species that form the biological backbone of ecosystems. Fully describing consumer-resource interactions can however be a daunting task due tho the sheer number of potential interactions, even for a limited number of species (Dodds and Nelson 2006). The empirical description of interactions at the community scale can thus be limited by logistical constraints (Dunne 2006; Morales-Castilla et al. 2015). In the context of environmental impact evaluation, the time required to gather empirical data can further diminish the applicability of such methodologies due to time constraints (ref).

Alternative approaches to predicting interactions among community members have thus been explored in order to efficiently resolve interaction webs using incomplete, imperfect or even unavailable data (e.g. Gravel et al., 2013; Morales-Castilla et al., 2015). These approaches rely on the use of proxies such as functional traits to infer interactions consumer-resource relationships (Morales-Castilla et al. 2015). Multiple traits can play a significant role in community dynamics and influence the presence and intensity of biotic interactions, like the influence of body size on predator-prey interactions, a literal take on *big fish eats small fish* (Cohen et al., 2003; Brose et al., 2006; Gravel et al. 2013).

Evolutionary processes causes closely related species to share similar traits. This proximity increases the likelihood of closely related species sharing consumer and resources. Furthermore, trait differentiation arises through evolution, dictated by the coevolutionary dynamics and trait matching of consumers and resources. It is therefore likely that phylogenies could serve as proxies of important traits and, by extention, help predict pairwise interactions.

These methodologies require large datasets from which models must be trained. There exist a large collection of empirical data that could be used to train such models. However, this collection comes from a wide variety of species and ecosystems.

Could we make use of such data to predict interactions in a given community in which knowledge is lacking, based on the phylogenetic proximity of species. Is it a truism to say that a species spatially distant from each other are likely to consume other species that are closely related phylogenetically?

Can we make use of existing heterogeneous data to predict interactions in a given community

The goal is to obtain a fully or optimally resolved interaction matrix for all taxa in a given community, i.e. constructing the community metaweb, or its topology (Dunne 2006).

Consumer-resource trait matching dictate the interactions in a given community Trait differentiation arise through evolution, dictated by the coevolutionary dynamics of consumer-resource

Thus, can we take advantage of available data to predict biotic interactions in data-poor environments? The objective of this paper is to use the link between taxonomic proximity and shared interactions to build models predicting biotic interactions in data poor ecosystems.

2 Methods

The objective of the methods is to predict the interactions for all taxa within an arbitrary set *S1* using a set of taxa *S0* with empirically described interactions and their similarity measured from taxa consumer/resource sets and their taxonomy. We couple the use of empirical data with an unsupervised machine learning method to achieve this.

2.1 Biotic interaction catalogue

We built a biotic interaction catalogue to serve as a set of taxa *S0* with empirically described interactions. The empirical data used to construct the interaction catalogue was gathered in two successive steps. The first consisted of gathering data from a collection of 94 empirical food webs in marine and coastal ecosystems from which we extracted pairwise taxa interactions (Brose et al. 2005; Kortsch et al. 2015; GlobalWeb database; see Appendix 1 for a description of food webs). We also used a detailed predator-prey interaction database describing trophic relationships between XX predators and their prey (Barnes et al. 2008). From these datasets, only interactions between taxa at the taxonomic scale of the family or higher were selected for inclusion in the catalogue.

As empirical food webs are vastly dominated by non-interactions, these datasets yielded a highly skewed distribution of interactions vs non-interactions. To counterbalance this, the second step of data compilation consisted of extracting observed interactions from the Global Biotic Interaction (GloBI) database (ref), which describes binary interactions for a wide range of taxa worldwide. We extracted all interactions available on GloBI for species belonging to the families of taxa identified through step 1. Interactions were extracted using the rGloBI package in R (ref). As for step 1, only interactions between taxa at the taxonomic scale of the family or higher were retained

The nomenclature used between datasets and food webs varied substantially. Taxa names thus had to be verified, modified according to the scientific nomenclature and validated. This process was performed using the Taxize package in R (ref) and manually verified for errors. The same package was used to extract the taxonomy of all taxa for which interactions were obtained in previous steps. The complete R code and data used for the catalogue is available at https://github.com/david-beauchesne/Interaction_catalog.

2.2 Unsupervised machine learning

We use the K-nearest neighbor (KNN) algorithm (**ref**) to predict pairwise interactions for a set of taxa S. The KNN algorithm predicts missing entries or proposes additional entries by a majority vote based on the K nearest (*i.e.* most similar) entries. In this case, taxa are described by a set of resources when considered as a consumer, a set of consumers when considered as a resource and their taxonomy (*i.e.* kingdom, phylum, class, order, family, genus, species). Similarity between taxa was evaluated using the Tanimoto similarity measure (**ref**), which compares two vectors with i elements based on the number of elements they share and contain:

$$tanimoto(\mathbf{x}, \mathbf{y}) = \frac{\sum_i x_i \wedge y_i}{\sum_i x_i \vee y_i}, \tag{1}$$

where $\wedge$ is bitwise *and*, while $\vee$ is the bitwise *or* operators. Adding a weighing scheme, we can measure the similarity using two different sets of vectors

with i and j elements, respectively.

$$tanimoto_t(\mathbf{x}, \mathbf{y}, w_t) = w_t tanimoto(\mathbf{x_i}, \mathbf{y_i}) + (1 - w_t) tanimoto(\mathbf{x_j}, \mathbf{y_j}), \quad (2)$$

where w_t is the weight given to vector i, $\mathbf{x_i}$, $\mathbf{y_i}$ are the resource or consumer sets of the two taxa and $\mathbf{x_j}$ and $\mathbf{y_j}$ are the vectors for taxonomy of two taxa. When $w_t = 0$ only resource or consumer sets are used to compute similarity, while $w_t = 1$ solely uses taxonomy.

2.3 Predicting interactions, Biotic predictor algorithm, Two-way Tanimoto algorithm, Feng shui name algorithm, Find a name for the algorithm

The XXX algorithm is built on a series of logical steps in order to suggest a series of candidate resources for each taxon in *S1* (Figure 1). For all consumer taxa T_C in *S1*, the algorithm first verify whether it has empirical resources listed in the catalogue. When it does, if resource taxa T_R are also in *S1*, they are added as predicted resources for T_C. This corresponds to what we refer to as the catalogue contribution to resource predictions. Two taxa in *S1* that are known to interact through the catalogue are automatically assumed to interact in *S1*.

Otherwise, the algorithm passes to what we refer to as the predictive contribution to resource predictions, with candidate resources for T_C identified with the KNN algorithm. If resource taxa T_R are also in *S1*, K most similar resource T'_R are identified in *S1* to add to a candidate resource list C_R for T_C. Then for all T_C in *S1*, the algorithm identifies K most similar consumer T'_C in *S0* and extracts their resource sets. As before, if those resources are found in *S1* they are added to C_R, otherwise K most similar resources T'_R are identified in *S1* to add to C_R. A more comprehensive mathematical description of the algorithm is available through Figure 1 and the complete R code and data used for the algorithm is available at https://github.com/david-beauchesne/Predict_interactions.

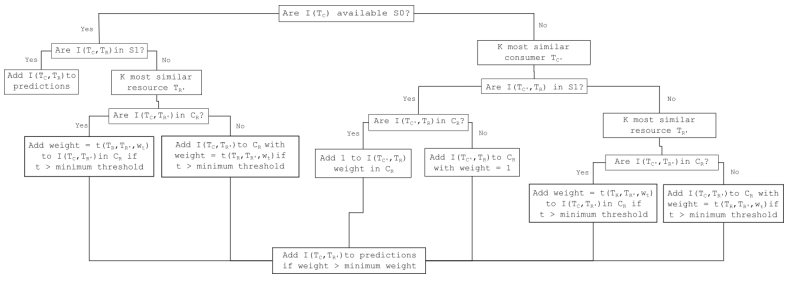

Figure 1: Description of the logical steps used by the algorithm to suggest a list of candidate resources (C_R) for each taxa (T_C) in *S1*

2.4 Algorithm prediction accuracy

We used the most extensive and taxonomically detailed datasets included in the catalogue (**ref**) to assess the prediction accuracy of the algorithm. Testing accuracy of a particular dataset was done by first removing all pairwise interacting originating from that dataset. Accuracy was evaluated using four different statistics:

1. $Score_y$ is the fraction of interactions correctly predicted

$$Score_y = \frac{a}{a+c} \tag{3}$$

2. $Score_{\neg y}$ is the fraction of non-interactions correctly predicted

$$Score_{\neg y} = \frac{d}{b+d} \tag{4}$$

3. Accuracy score is the normalized sum of correctly predicted interactions and non-interactions

$$Accuracy = \frac{a+d}{a+b+c+d}, \tag{5}$$

4. TSS, The True Skilled Statistics (TSS) evaluated prediction success by considering both true and false predictions, returning a value ranging from 1 (prefect predictions) to 0 (inverted predictions; **ref**).

$$TSS = \frac{(ad-bc)}{(a+c)(b+d)} \tag{6}$$

where a is the number of links predicted (1) and observed (1), b is the number predicted (1) but not observed (0), c is the number predicted absent (0) but observed (1) and d is the number of predicted absent (0) and observed absent (0). These four statistics give a different perspective on prediction accuracy, focusing in turn on true interactions and non-interactions, on true events an on both true and false predictions.

We evaluated the four statistics for the full algorithm, but also for the catalogue and the predictions individually to evaluate their respective contribution to the algorithm predictive accuracy. Multiple w_t values were also tested to evaluate whether taxa similarity measured as a function of resource/consumer sets or taxonomy contributed more significantly towards increased predictive accuracy. The same was done with multiple K values.

Finally, we evaluated the influence of the comprehensiveness of the catalogue on prediction accuracy. We selected the arctic food web from Kortsch et al. (2015) to evaluate this, as it is highly detailed and because almost all taxa found in it had information in the catalogue even when not included in the catalogue. We evaluated this by iteratively ($n = 50$) removing a random

percentage of information available in the catalogue for the food web and generating predictions from the algorithm as before. We tested with w_t values = 0.5 and 1 in order to evaluate whether taxonomic similarity could support predictive accuracy in cases when empirical data for species in *S1* in the catalogue is unavailable.

3 Results

3.1 Biotic interaction catalogue

The data compilation process allowed us to build an interaction catalogue composed of 276708 pairwise interactions (interactions = 72110; non-interactions = 204598). A total of 9712 taxa (Superfamily = 15; Family = 591; Subfamily = 29; Tribe = 8; Genus = 1972; Species = 7097) are included in the catalogue, 4159 of which have data as consumers and 4375 as resources.

3.2 Algorithm predictive accuracy

The overall predictive accuracy of the algorithm ranges between 80% to almost 100% in certain cases (Figure 2), except for one food web (*i.e.* **ref**). Both interactions and non-interactions are well predicted by the algorithm, but TSS scores are lower than $Score_y$ and $Score_{\neg y}$ due to wrongly classified interactions and non-interactions. This can also be easily observed through the effect of varying K values, which increases the number of potential candidate resources for each taxa in the predictive portion of the algorithm. Increased K values increases prediction accuracy for interactions, but decreases for non-interactions. This means that both interactions and non-interactions are classified as interactions in the process of increasing $Score_y$, accordingly decreasing $Score_{\neg y}$.

Prediction accuracy varies with w_t values, with predictions closer to 0 yielding better predictions than those closer to 1. This means that resource/consumer sets serve as better predictor of similarity between taxa than does taxonomy. It is however interesting to note that also the predictive contribution of the algorithm decreases as w_t increases, the TSS and $Score_y$ means and variability for the complete algorithm increases and decreases, respectively. This suggests that while using taxonomy for similarity measurements yields lower predictive accuracy, it may also complement the catalogue contribution by predicting interactions not captured through empirical data, effectively increasing the predictive accuracy of the algorithm.

The catalogue and the predictive contributions is also unequal, with the algorithm predictive prowess seemingly originated mainly from the empirical data used rather than the KNN algorithm. This suggest that the efficiency of the algorithm is dependent on the comprehensiveness of the catalogue. Figure 3 shows that it is indeed the case, with the predictive contribution and complete algorithm decreasing quickly as the proportion of taxa in *S1* found in the catalogue decreases. Predictions however buffers the loss of data in the cata-

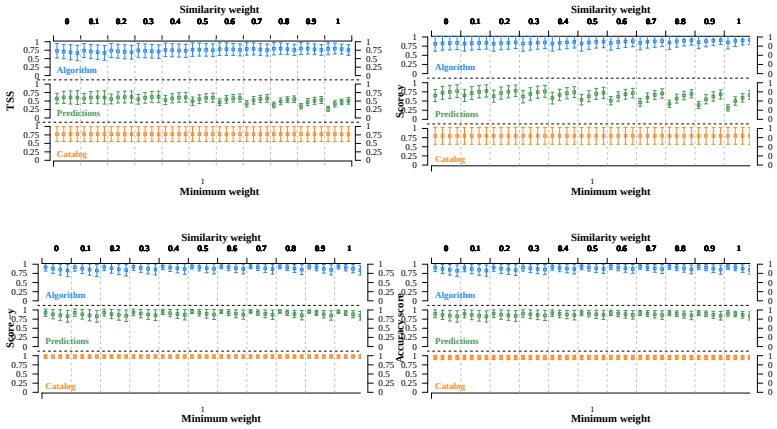

Figure 2: The graph presents the four statistics as a function of trait weight, which varies between 0 and 1. A weight of 0 means that similarity is measured only using set of resources for each taxa, while a weight equal to 1 means that similarity is based solely on taxonomy. We present 6 food webs with over 50 taxa each and the Barnes et al. (2008) dataset.

logue and still allows the complete algorithm to accurately described at least XX% of interactions even when more that XX% of taxa are missing in the catalogue. Interestingly, as the proportion of *S1* taxa in the catalogue decreases, the similarity measured with taxonomy only performs better, with predictions performed with $w_t = 1$ surpassing those made with $w_t = 0.5$ as percent of taxa removed from catalogue increases. This suggests that resource/consumer sets are better predictor of taxa similarity, but that as information gets more scarce, using taxonomy is a valid replacement to obtain a rather high predictive power.

4 Discussion

The taxonomy alone does not seem enough to

Interactions that are wrongly classified as non-interactions merit greater attention, as the quality of non-interactions data is dubious at best.

Secondly, there is the case of non-interactions. Looking at $Score_y$ and $Score_{\neg y}$, you can see that the trend in TSS and Accuracy score values are closely matching those of $Score_{\neg y}$. In opposition, the $Score_y$ are instead increasing with taxonomy being more important in the similarity measurements. There could be multiple explanations for this.

First, the algorithm could simply perform poorly in predicting non-interactions, classifying them instead as interactions.

However, I rather believe that the original empirical food webs are the ones doing poorly at observing non-interactions. Indeed, we assumed that non-

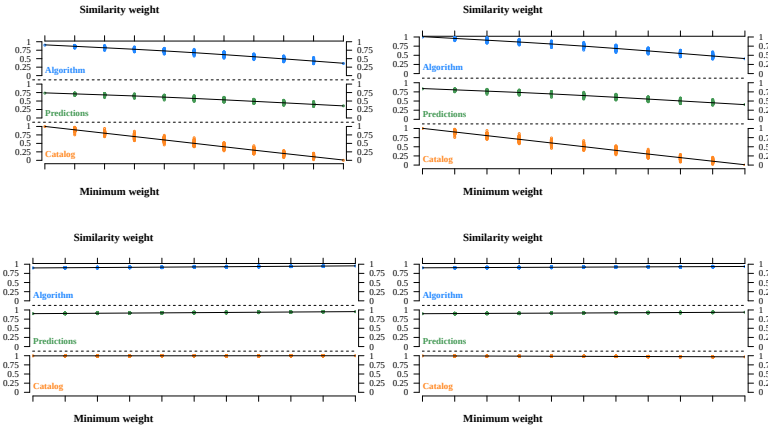

Figure 3: Graph presenting predictive accuracy as a function of the amount of information available in the catalogue. The arctic food web from Kortsch et al. (2015) was used for this, as it is highly detailed and because almost all taxa found in it had information in the catalogue even when not included in the catalogue. A random percentage of taxa in the web was iteratively removed from the catalogue (n = 50) before predicting interactions with the XXX algorithm.

interactions in empirical food webs meant that there were no interactions between taxa. However, most of the empirical webs had a strong focus attributed to higher order consumer species and very little attention given to other taxa (*e.g.* benthic invertebrates). The catalogue of interactions, on the other hand, has a much broader focus than the original empirical webs. It therefore very well may be that wrongly classified non-interactions could indeed be real but unobserved interactions in natural systems. This would need to be tested at some point.

The algorithm is currently coded to ignore non-interactions in the similarity measurements. Maybe it would make sense to do the same in the interpretation of the results itself. Or a third vector to the similarity measurements could be added that considers non-interactions as a set of resources not consumed by consumers (which we have in the data). While this could make sense, more thought needs to be given to non-interactions stemming from empirical food webs and their actual value as observed data. If we think them valuable, then extra logical steps could be added to the algorithm to decrease probabilities of taxa being proposed as candidate resources by the KNN algorithm when they were observed to not interact in food webs.

Interestingly, as the proportion of *S1* taxa in the catalogue decreases, the similarity measured with taxonomy only performs better, with predictions performed with $w_t = 1$ surpassing those made with $w_t = 0.5$ as percent of taxa removed from catalogue increases. Taxonomy could thus be highly useful in data poor environments that also lack data in the catalogue.

While taxonomy plays a significant role in influencing food web structure (Eklof et Stouffer 2015), it does not account for certain traits such as body size. Even though we advocate for as simple an approach as possible, we believe that in cases where the comprehensiveness of the catalogue does not encompass a significant portion of S1, traits could be added as a third set in the similarity measurement.

5 Acknowledgements

We thank the Fond de Recherche Québécois Nature et Technologie (FRQNT) and the Natural Science and Engineering Council of Canada (CRSNG) for financial support. This project is also supported by Québec Océan, the Quebec Centre for Biodiversity Science (QCBS), and the Notre Golfe and CHONeII networks. We also wish to thank K. Cazelles for constructive comments and suggestions.