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

D. BEAUCHESNE1*, P. DESJARDINS-PROULX2, P. ARCHAMBAULT3, D. GRAVEL2

1 Université du Québec à Rimouski, 300 Allée des Ursulines, Rimouski, Québec, Canada G5L 3A1
2 Université de Sherbrooke, 2500, de l’Université, Sherbrooke, Québec, Canada J1K 2R1
3 Université Laval, 2325 de l’Université, Québec City, Québec, Canada G1V 0A6

*Corresponding author: david.beauchesne@uqar.ca

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. We therefore wondered whether readily available data, namely empirically described interactions in a variety of ecosystems, could be combined to predict species interactions in data deficient ecosystems. To test this, we built a catalogue of biotic interactions from a collection of 94 empirical food webs, detailed predator-prey interaction databases and interactions from the Global Biotic Interactions (GloBI) database. We used 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. Results suggest that pairwise interactions can be predicted with high accuracy. While conclusions are seemingly dependent on the comprehensiveness of the catalogue, knowledge of taxonomy was found to complement well the catalogue and improve predictions, especially when empirical information available is scarce. Given its high accuracy, this methodology could promote the use of food webs and network level descriptors in remote and frontier location where empirical data is hard to collect. 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, opening promising avenues for further research and for management initiatives.

INTRODUCTION

Large networks of ecological interactions, such as food webs, are complex to characterize (Polis 1991, Martinez 1992, Pascual & Dunne 2006). Empirical descriptions require exhaustive observations, while theoretical inference generally requires ample data to be validated. For this reason, studies focusing on communities of interacting species remain understudied, even though we acknowledge the importance of considering the reticulated nature of complex networks (Ings et al. 2009, Tylianakis et al. 2008). When time is of the essence, the long term studies required quickly become impractical and the use of network level approaches relegated to the sideline.

Alternatively, an approach currently gaining in popularity is to predict interactions using proxies such as functional traits, phylogenies and species distribution data (e.g. Morales-Castilla et al. 2015, Bartomeus et al. 2016). For example, 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, Séguin et al. 2014). However, the time required to gather the necessary data to apply those methods may still be restrictive, or the data be unavailable altogether, so much so that other methods such as imputation techniques applied to phylogenies and traits have been developed to fill gaps in knowledge (e.g. Penone et al. 2014, Schrodt et al. 2015).

We therefore wondered whether more readily available data could be used to infer interactions in data deficient ecosystems. There is an increasing amount of data describing worldwide species interactions, some freely available through the Global Biotic Interactions (GloBI) database (Poelen et al. 2014). Similarly, while detailed and calibrated phylogenies can be challenging to construct and require ample data, a taxonomical description of species is easily accessible through initiatives like the World Register of Marine Species (WoRMS; Bailly et al. 2016). Evolutionary processes are hypothesized to influence and shape consumer-resource interactions through trait matching (Mouquet et al. 2012, Rohr & Bascompte 2014), so that taxonomically related species would be more likely to share similar types of both consumers and resources because they share similar traits (Eklof et al. 2012, Gray et al. 2015, Morales-Castilla et al. 2015). Based on that hypothesis, taxonomy might be a useful surrogate to predict interactions when trait data are unavailable.

The objective of this work is thus to combine empirical biotic interactions originating from a variety of ecosystems with taxonomic relatedness to predict interactions for data deficient ecosystems. The concept underlying our methodology is that instead of constraining ourselves to
a specific locality, we would look to other environments – outside the box – to glean insights as to the inner workings of an area of interest. As an example, we compare the observed interactions in the southern Gulf of St. Lawrence in Canada (SGSL; Savenkoff et al. 2004) with predictions made using our approach.

METHODS

The objective of our methodology is to predict the interactions between all pairs of taxa within an arbitrary set \( N_1 \), using a set of taxa \( N_0 \) with empirically described interactions from which we can extract pairs of consumers and resources and their taxonomy. We couple the use of empirical data with an unsupervised machine learning method to achieve this.

Biotic interactions catalogue: We built a biotic interaction catalogue to serve as a set of taxa \( N_0 \) with empirically described interactions. The empirical data used to construct the interactions catalogue was assembled in two successive steps. The first consisted of gathering data from a collection of 94 empirical food webs from which we extracted pairwise taxa interactions (see Brose et al. 2005, Kortsch et al. 2015, University of Canberra 2016 for more information). We also used a detailed predator-prey interaction database describing trophic relationships between marine fishes 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. Data used came exclusively from marine and coastal ecosystems and encompassed a wide variety of organisms: fungi, algae, parasites, phytoplankton, zooplankton, benthic and pelagic invertebrates, demersal and pelagic fishes, marine birds and marine mammals. As empirical food webs are vastly dominated (96% in our datasets) by unobserved or absent interactions ("0", hereafter referred to 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 (Poelen et al. 2014), which describes binary interactions for a wide range of taxa worldwide. We extracted all trophic 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 (Poelen et al. 2015). As per 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 (Chamberlain & Szöcs 2013, Chamberlain et al. 2014) 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 to build the catalogue is available at https://github.com/david-beauchesne/Interaction_catalog.

Fig. 1. – Description of logical steps used by the algorithm to suggest a list of candidate resources \( (C_r) \) for each consumer taxa \( (C_c) \) in a set of \( N_1 \) for which interactions are predicted, using a set of taxa \( N_0 \) with empirically described interactions. Interactions between consumer and resource taxa are denoted as \( I(T_c, T_r) \). \( K \) is the number of most similar neighbors selected for the KNN algorithm; \( t \) stands for tanimoto in equation 1; \( w_i \) is the weight given to sets of resources and consumers in equation 2; the minimum threshold is a value setting the minimal similarity value accepted for taxa to be considered as close neighbors in the KNN algorithm; the weight is the value added to a candidate resource each time it is added to \( C_r \); the minimum weight is the minimal weight value accepted for candidate resources to be selected as predicted sources in the algorithm.
Unsupervised machine learning: We use the K-nearest neighbor (KNN) algorithm (Murphy 2012) to predict pairwise interactions for a set of taxa \( N_f \). The KNN algorithm predicts missing entries or proposes additional entries by a majority vote based on the K nearest (i.e., most similar) entries (see Box I for an example). 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, which compares two vectors \( x \) and \( y \) with \( n = |x| = |y| \) elements, and is defined as the size of the intersection (\( \cap \)) of two sets divided by their union (\( \cup \)):

\[
tanimoto(x, y) = \frac{|x \cap y|}{|x \cup y|} \quad (1)
\]

Adding a weighting scheme, we can measure the similarity using two different sets of vectors \( (x, y) \) and \( (u,v) \):

\[
tanimoto(x,y,u,v) = w_t \cdot tanimoto(x, y) + (1-w_t) \cdot tanimoto(u, v) \quad (2)
\]

where \( w_t \) is the weight in [0; 1]. For our analyses, the first element on the right-hand side of equation 2 is the Tanimoto similarity measured using the taxonomy of two taxa. The second is the Tanimoto similarity between the sets of resources (or consumers) of the same taxa. Hence, when \( w_t = 0 \) only resource or consumer sets are used to compute similarity, while \( w_t = 1 \) solely uses taxonomy. This approach to consider the relative contribution of two sets of vectors to the Tanimoto similarity was developed by Desjardins-Proulx et al. (2016).

Predicting interactions: The algorithm consists of a series of logical steps that ultimately predicts a candidate resources list \( C_b \) for each taxon in \( N_f \) based on empirical data available and the similarity among consumers and among resources (Fig. 1). For all consumer taxa \( T_c \) in \( N_f \), the algorithm first verifies, for all resources in resource set \( T_r \), if they are found in \( N_f \) (Step S1, Fig. 1). When it does, all \( T_r \) taxa that also are in \( N_f \) are added as predicted resources for \( T_c \) (Steps S2 and S3). This corresponds to what we refer to as the catalogue contribution to resource predictions. Essentially, two taxa in \( N_f \) that are known to interact through empirical data in the catalogue are assumed to interact in \( N_f \).

Otherwise, the algorithm passes to what we refer to as the predictive contribution to resource predictions (Steps S4 to S16), with candidate resources for \( T_{cr} \) (focal taxon for explanation) identified with the KNN algorithm. For each resource in \( T_r \) that was not in \( N_f \) (Step S2), K most similar resources \( T_{ro} \) are identified from \( N_f \) (Step S4). If similar resources \( T_{ro} \) have a similarity value above a minimal similarity threshold set to 0.3 in our analysis, they are added to \( C_{cr} \) as candidate resources. If not, they are automatically discarded (Steps S5 to S7). This minimal threshold is an arbitrary parameter used to avoid predicting resources that have very small and insignificant similarity values and hence unlikely to share consumers and resources with target taxon.

Then for all consumer taxa \( T_c \) in \( N_f \), K most similar consumers \( T_{ci} \) are identified from \( N_f \). This step aims at extracting sets of potential resources \( T_r \) from similar types of consumers found in the catalogue (Step S8). Resources \( T_r \) are added to candidate resources \( C_b \) for \( T_c \) if they are also found in \( N_f \) (Steps S10 to S12). Otherwise, Steps S4 to S7 are duplicated to identify potential similar resources for \( T_c \) in \( N_f \) from the set of resources \( T_b \) of similar consumers \( T_{ci} \) (Steps S13 to S16). A simple working example is presented at Box 1. A comprehensive mathematical description of the algorithm and the parameters used is available through Fig. 1 and the complete R code and data used for the algorithm are available at https://github.com/david-beauchesne/Predict_interactions.

Algorithm prediction accuracy: We used datasets including more than 50 taxa (Christian & Luczkovich 1999, Link 2002; Thompson et al. 2004, Brose et al. 2005, Barnes et al. 2008, Kortsch et al. 2015) to assess the prediction accuracy of the algorithm. For each dataset, we first removed all the interactions it contributed to the interactions catalogue and then used the algorithm to predict the structure of interactions among all taxa included in the dataset. We then compared the predicted and observed networks to evaluate the accuracy of the predictions using four different parameters: \( a \) is the number of interactions correctly predicted (i.e., true positives), \( b \) is the number of non-interactions predicted as interactions (i.e., false positives), \( c \) is the number of observed interactions predicted as non-interactions (i.e., false negatives) and \( d \) is the number of non-interactions correctly predicted (i.e., true negatives). These parameters are used in three different statistics:

1. Score\(_r\), the fraction of interactions correctly predicted:

\[
Score_r = \frac{a}{a + c}
\]

2. Score\(_{y}\), the fraction of non-interactions correctly predicted

\[
Score_{y} = \frac{d}{b + d}
\]

3. TSS, The True Skilled Statistics is the evaluated prediction success by considering both true and false predictions, returning a value ranging from 1 (perfect predictions) to −1 (inverted predictions; Allouche et al. 2006):

\[
TSS = \frac{(ad - bc)}{(a + c) \cdot (b + d)}
\]

These three statistics give a different perspective on prediction accuracy, focusing on true interactions, non-interactions, and on both true and false predictions respectively. It is however important to note that false positives and true negatives are solely representative of the datasets used rather than the environment itself, as even exhaustively described food webs may not fully describe interactions in a given environment.

For each statistic, we evaluated prediction accuracy 1) for the complete algorithm, 2) for predictions made with the predictive portion of the algorithm (Steps S4–S16; Fig. 1) and 3) for the catalogue contribution of the algorithm (Steps S1–S3; Fig. 1). We evaluated these steps separately in order to partition the relative contribution of the catalogue and of the predictions made using the KNN algorithm to the overall predictive accuracy of the algorithm. 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 marine food web from Kortsch et al. (2015) as a test. This food web was selected because it is highly detailed taxonomically. Furthermore, once its data were removed from the catalogue, almost 100 % of its taxa still had information available on sets of consumers and resources, which was necessary for testing the effect of catalogue comprehensiveness on prediction accuracy. We iteratively and randomly (n = 50 randomizations) removed a percentage of empirical data describing the food web taxa from the catalogue before generating new predictions with the algorithm. We also tested w, values of 0.5 and 1 to evaluate whether taxonomic similarity could support predictive accuracy in cases when empirical data for species in N were unavailable in the catalogue.

RESULTS

Biotic interactions catalogue

The data compilation process allowed us to build an interactions catalogue composed of 276 708 pairwise interactions (interactions = 72 110; non-interactions = 204 598). A total of 9 712 taxa (Superfamily = 15; Family = 591; Subfamily = 29; Tribe = 8; Genus = 1 972; Species = 7 097) are included in the catalogue, 4 159 of which have data as consumers and 4 375 as resources.

Algorithm predictive accuracy

The overall predictive accuracy of the algorithm ranged between 80 % to almost 100 % in certain cases (Fig. 2). Both interactions and non-interactions were well predicted by the algorithm. TSS scores were lower than predicted by the algorithm. TSS scores were lower than (Fig. 2). Both interactions and non-interactions were well ranged between 80 % to almost 100 % in certain cases. While prediction accuracy of the predictive portion of the algorithm was somewhat lower, it nonetheless supported high prediction accuracy when the use of taxonomy for similarity computation was more efficient when empirical data was scarcer and no different than resource/consumer sets for the complete algorithm when ample data was available (Fig. 3).

Southern Gulf of St. Lawrence

As an example, we predicted interactions in the southern Gulf of St. Lawrence (SGSL) in eastern Canada. The empirical data and taxa list come from Savenkoff et al. (2004). They presented a list of 29 functional groups for a total of 80 taxa presented at least at the taxonomical scale of the family. Other coarser functional groups were not used for this example (see Table S1 in Supplementary information (SI) and Savenkoff et al. (2004) for a complete description of documented groups). We used the algorithm to predict interactions between all 80 taxa selected. As the interactions are reported for functional groups rather than taxa, we then aggregated them back to their original functional groups for comparison with interactions presented in Savenkoff et al. (2004). In total, there was empirical data available in the catalogue for 78 % of SGSL taxa (62/80). The algorithm correctly predicted close to 80 % of interactions (a = 135/170) and non-interactions (d = 354/455) extracted from Savenkoff et al. (2004). It also predicted 101 additional interactions that were not noted in Savenkoff et al. (2004) (Table S2) and failed to predict 36 observed interactions that were (Table S3), resulting in a TSS score of 0.57. A visual comparison of results obtained from the algorithm with interactions noted in Savenkoff et al. (2004) is provided at Fig. 4. The network presented is centered on the observed and predicted interactions of the capelin (Mallotus villosus) and piscivorous small pelagic feeders (e.g. Scomber scombrus and Illex illecebrosus).

DISCUSSION

Algorithm accuracy

We show that out of the box interaction inference for a set of taxa with incomplete or unavailable preexisting
Fig. 2. – Representation of the three statistics (i.e. Score, Score_−y and TSS) used to evaluate the accuracy of the algorithm as a function of K values tested (i.e. 2, 4, 6 and 8 most similar neighbors, top x-axis) and weight for taxonomy (bottom x-axis), which varies between 0 and 1. A weight of 0 means that similarity is measured only using set of resources/consumers for each taxon, while a weight of 1 means that similarity is based solely on taxonomy. For each statistic, the topmost panel presents prediction accuracy for the complete algorithm, the middle panel corresponds to predictions made through the predictive portion of the algorithm (Steps S4-S16; Fig. 1) and the bottom panel presents the catalogue contribution for the algorithm (Steps S1-S3; Fig. 1). Note that the sum of the predictive and catalogue contributions can be over 100% as there is overlap between predictions made through both. The 7 datasets used for this analysis contained over 50 taxa (Christian & Luczkovich 1999, Link 2002, Brose et al. 2005, Thompson et al. 2004, Barnes et al. 2008, Kortsch et al. 2015).
information can be achieved with high accuracy using a combination of empirical data describing biotic interactions and taxonomic relatedness. Although the efficiency of the algorithm is dependent on the comprehensiveness of the interactions catalogue, taxonomic proximity acts as a complement to increase predictive capability, especially when the catalogue is less comprehensive.

**Taxonomic similarity**

We found that taxonomy can be highly useful in complementing predictions made using empirical data. Much like the findings from Eklöf & Stouffer (2016), evolutionary history serves as an efficient surrogate for traits from which inferences on consumer-resource interactions and network structure can be made (Mouquet et al. 2012, Rohr & Bascompte 2014). It can also capture traits that we are unable to measure or use through a comparative approach, such as stoichiometric constraints or dietary quality requirements. Nonetheless, phylogenetic relationships do not necessarily capture certain traits that significantly drive interactions, in particular when traits are not conserved in phylogenies (Losos 2008, Wiens et al. 2010). Complementing our methodology with additional, higher resolution information such as functional traits (e.g. metabolism and body size) and species co-occurrence could thus yield even higher predictive efficiency. Similarly, considering time calibrated phylogenies rather than taxonomy could enhance the resolution at which evolutionary history is considered. This could be achieved through recent efforts to extensively describe all-encompassing phylogenies (e.g. Hedges et al. 2015). Even though our methodology was designed for data-poor situations, such data could and should be used if available.

**Interactions classification**

That Score$_y$ and Score$-y$ are inversely proportional to each other.

**Fig. 3.** Representation of Score$_y$ as a function of catalogue comprehensiveness, i.e. the amount of information on sets of consumer and resources available in the catalogue. The sensitivity of the algorithm to data accuracy was evaluated with the arctic food web from Kortch et al. (2015). This food web is highly detailed taxonomically. Almost 100% of its taxa are documented in the interactions catalogue, which was necessary to test the effect of full range of catalogue comprehensiveness on prediction accuracy. A random percentage of data available in the catalogue for taxa in the food web (i.e. 0 to 100%) was iteratively removed (n = 50 randomizations) before generating new predictions with the algorithm. $w_y$ values of 0.5 and 1 were evaluated to verify the usefulness of taxonomy in supporting predictive accuracy. The topmost panel presents prediction accuracy for the complete algorithm, the middle panel corresponds to predictions made through the predictive portion of the algorithm (Steps S4-S16; Fig. 1) and the bottom panel presents the catalogue contribution for the algorithm (Steps S1-S3; Fig. 1). Note that the sum of the predictive and catalogue contributions can be over 100% as there is overlap between predictions made through both.
means that non-interactions are misclassified as interactions in the process of increasing Score_, consequently decreasing Score_. This could either stem from the algorithm poorly predicting non-interactions or from the empirical data itself. Accuracy evaluation assumes that non-interactions from empirical food web are observed data, yet it is usually not the case. Most empirical webs have a strong focus attributed to higher order consumer species and often-uneven effort made to thoroughly detail species interactions (Dunne 2006). Furthermore, the methodologies used to obtain consumer-resource data, often relying on gut content analyses or stable isotopes, while efficient at observing interactions, may be inefficient to detect absence of interactions in natural systems (Dunne 2006). This is especially true with our methodology, where we predict interactions between species whose co-occurrence may have been observed in the other ecosystems we are using to predict interactions. Misclassified interactions could thus be real, albeit unobserved through empirical data available.

**Southern Gulf of St. Lawrence**

The St Lawrence example (Fig. 4 and SI) provides adequate material to discuss predictions in greater detail. The algorithm failed to predict 20% of interactions presented in Savenkoff et al. (2004). Interactions that failed to be predicted were mainly centered on invertebrate species (e.g. polychaetes and mollusks) and taxonomically diverse functional groups described by coarse taxonomic categories (e.g. diatoms) alongside few species in Savenkoff et al. (2004) (e.g. piscivorous small pelagic feeders; Table S3). As we focused on the taxa at least at the scale of family, it is likely that their functional groups had a broader range of possible interactions included than what the algorithm could predict using only a few taxa. Furthermore, the efficiency of the algorithm greatly depends on the underlying empirical data that defines the catalogue. If the empirical data used to build the catalogue focuses on higher order consumers, it should come as no surprise that the algorithm would be afflicted by the same limitations.

On the other hand the algorithm also predicted substantially more interactions than those presented in Savenkoff et al. (2004) (Fig. 4; Table S2). For instance, an important number of additional consumer interactions were predicted for small piscivorous pelagic feeders, whereas the empirical data suggest that this species group has very few preys (Fig. 4). This situation could be explained by ontogenetic shifts in diet that are captured by the wide spectrum of interactions covered in the catalogue, such as small piscivorous pelagic feeders consuming eggs and/or juvenile cod. This exemplifies the point we made in the previous section with regards to misclassified interactions being real rather than false positives. The resulting TSS score for the St. Lawrence analysis is thus greatly diminished by classifying additional interactions as false positives and, as such, we believe it to be an underestimation of the efficiency of our methodology to predict interactions.

**Perspectives**

We show that out of the box interactions inference can be achieved with high accuracy using readily available data.
data, suggesting that ecological networks are characterized by a degree of predictability and that this predictive value can be recovered through learning (see Tamaddoni-Nezhad et al. 2013, Gray et al. 2015 for other examples). This adds weight to claims that regularities can be observed and predicted in network structure (Eklof et al. 2013).

We believe that our methodology offers promising avenues for further applied research and management initiatives. The flexibility of our methodology allows it to take advantage of multiple types of data. Complementing and testing it with additional ecological information such as functional traits and phylogenies would therefore be highly valuable. Interaction strength and species co-occurrence are additional major attributes affecting the probability of observing interactions and the resulting network structure. Interaction strength is instrumental to understand community dynamics, stability and robustness (Laska & Wootton 1998, Morales-Castilla et al. 2015), while the co-occurrence of species encloses valuable information on interactions and is obviously a pre-requisite for interactions to exist (Cazelles et al. 2016). Considering them in our methodology would be highly valuable to correctly assess interactions in a given ecosystem and predict the spatial distribution of interaction networks.

The significance of this approach also extends to other areas of ecological research where gathering data can be highly difficult, such as the reconstruction of interaction networks of palaeocommunities (e.g. Yeakel et al. 2013, Yeakel et al. 2014). Predicted networks of taxa known to co-occur could be used in hindsight to evaluate the influence of major events such as biodiversity collapse or significant climatic regime shifts on the structure of past ecological communities. Another example is in applying the methodology to identify knowledge gaps to guide targeted survey efforts, especially for ecosystems that are hard to document empirically. As an example, additional interactions predicted by the algorithm could be used as hypotheses to test through targeted surveys.

Ultimately, given its high efficiency and simplicity, our methodology could help in promoting the use and the accessibility of food webs and network level descriptors for integrative management initiatives such as cumulative impacts assessments and systematic planning (Giakoumi et al. 2015, Beauchesne et al. 2016), especially for remote locations and frontier areas where empirical data is hard to gather. Network characteristics could be efficiently evaluated and correlated to levels of multiple environmental stressors to assess the vulnerability of ecosystems to global changes (Albouy et al. 2014). We believe that the development of such predictive approaches could represent the first much-needed steps towards the use of ecological networks in systematic impacts assessments.

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 CHONeI2 networks. We also wish to thank K Cazelles for the help, constructive comments and suggestions. We also thank D Bohan and an anonymous reviewer for their constructive comments and suggestions.

REFERENCES


BOX 1

The algorithm follows a series of logical steps to predict resources for all taxa in an arbitrary set of taxa \( N_1 \) using a set of taxa \( N_0 \) with empirically described interactions from which we can extract sets of consumers and resources and their taxonomy. In this example, we are predicting interactions for a fictitious \( n_1 = \{ t_1, t_9, t_{10}, t_{11}, t_{12} \} \) using \( n_0 \) with information on 12 taxa. This catalogue holds information on consumer or resource for 10 taxa and the taxonomy for all 12 taxa in the list.

Similarity between all pairs of taxa in \( n_0 \) is measured using equation 1. The upper triangular matrix represents similarity measured with taxa sets of resources/consumers, while the lower triangular represents taxonomic similarities. For consumer/resource set similarities, values of 0 mean that similarity equals 0 for both similarity measurements.

From these, the algorithm goes through logical steps (Fig. 1) to identify a candidate resource list \( C_k \) for each taxon in \( n_1 \) using either empirical data directly or \( K \) most similar taxa with equation 2. Going through the process for \( t_1 \), using \( K = 1 \) and \( w_i = 1 \):

<table>
<thead>
<tr>
<th>Steps</th>
<th>Catalogue</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( {t_i, t_9} ) in ( n_1 )?</td>
<td>{} {}</td>
</tr>
<tr>
<td>2</td>
<td>( t_9 ) in ( n_1 )?</td>
<td>{} {}</td>
</tr>
<tr>
<td>3</td>
<td>( t_9 = yes )</td>
<td>( {t_9} )</td>
</tr>
<tr>
<td>4-7</td>
<td>( t_9 = no ) ( t(t_9, t_{12}, w_i) = NA )</td>
<td>{} {}</td>
</tr>
<tr>
<td>8</td>
<td>( t(t_1, t_{12}, w_i) = 0.5 )</td>
<td>( {t_{10}} )</td>
</tr>
<tr>
<td>9</td>
<td>( t(t_9, t_{10}) ) in ( N_1 )?</td>
<td>{} {}</td>
</tr>
<tr>
<td>10-12</td>
<td>( t_9 = yes )</td>
<td>( {t_9, t_{10}} )</td>
</tr>
<tr>
<td>13-16</td>
<td>( t_{12} = no ) ( t(t_{12}, t_{10}, w_i) = NA )</td>
<td>{} {}</td>
</tr>
<tr>
<td>13-16</td>
<td>( t_{12} = no ) ( t(t_{12}, t_{9}, w_i) = T_9 = 0.5 )</td>
<td>{} {}</td>
</tr>
<tr>
<td>10-12</td>
<td>( t_9 = yes )</td>
<td>( {t_9, t_{10}} )</td>
</tr>
</tbody>
</table>

The logical steps allow us to predict a set of resources for \( T_1 = \{ T_9, T_{10}, T_{12}\} \). Doing it for all taxa in \( N_1 \) with \( w_i = 0 \) and 1 predicts the following networks:

\[
wt = 0 \quad \text{and} \quad wt = 1
\]