VIE ET MILIEU - LIFE AND ENVIRONMENT, 2016, 66 (3-4): 333-342

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

D. BEAUCHESNE[1*], *P. DESJARDINS-PROULX*[2], *P. ARCHAMBAULT*[3], *D. GRAVEL*[2]

[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*

INTERACTIONS
MACHINE LEARNING
FOOD WEBS
K-NEAREST NEIGHBOR
TAXONOMY
ST. LAWRENCE

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.

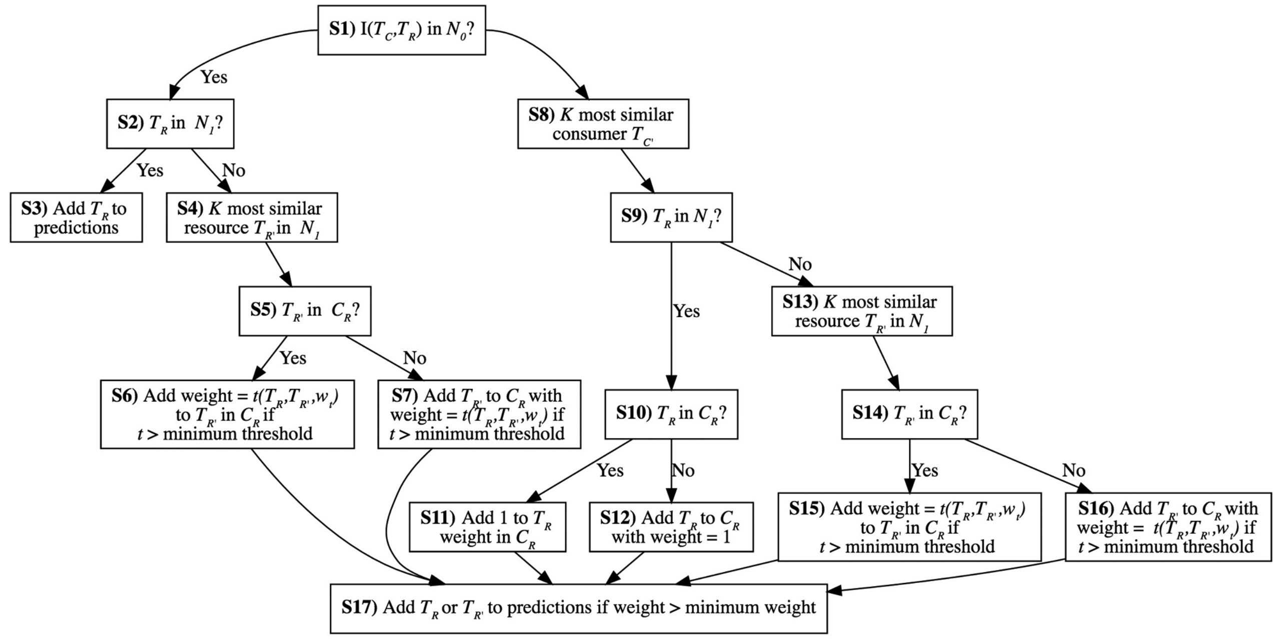

S1) I(T_C,T_R) in N_0?
Yes
S2) T_R in N_1?
Yes
No
S3) Add T_R to predictions
S4) K most similar resource $T_{R'}$ in N_1
S5) $T_{R'}$ in C_R?
Yes
No
S6) Add weight = $t(T_R,T_{R'},w_t)$ to $T_{R'}$ in C_R if t > minimum threshold
S7) Add $T_{R'}$ to C_R with weight = $t(T_R,T_{R'},w_t)$ if t > minimum threshold
S8) K most similar consumer $T_{C'}$
S9) T_R in N_1?
Yes
No
S10) T_R in C_R?
Yes
No
S11) Add 1 to T_R weight in C_R
S12) Add T_R to C_R with weight = 1
S13) K most similar resource $T_{R'}$ in N_1
S14) $T_{R'}$ in C_R?
Yes
No
S15) Add weight = $t(T_R,T_{R'},w_t)$ to $T_{R'}$ in C_R if t > minimum threshold
S16) Add $T_{R'}$ to C_R with weight = $t(T_R,T_{R'},w_t)$ if t > minimum threshold
S17) Add T_R or $T_{R'}$ to predictions if weight > minimum weight

Fig. 1. – Description of logical steps used by the algorithm to suggest a list of candidate resources (C_R) for each consumer taxa (T_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_t 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_1. 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 ($\cup$):

$$\text{tanimoto}\,(x, y) = \frac{|x \cap y|}{|x \cup y|} \qquad (1)$$

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

$$\text{tanimoto}_t(x, y, u, v, w_t) = w_t\text{tanimoto}(x, y) + (1 - w_t)\text{tanimoto}(u, v) \qquad (2)$$

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_R for each taxon in N_1 based on empirical data available and the similarity among consumers and among resources (Fig. 1). For all consumer taxa T_C in N_1, the algorithm first verifies, for all resources in resource set T_R, if they are found in N_0 (Step S1, Fig. 1). When it does, all T_R taxa that are also in N_1 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_1 that are known to interact through empirical data in the catalogue are assumed to interact in N_1.

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_{Ci} (focal taxon for explanation) identified with the KNN algorithm. For each resource in T_R that was not in N_1 (Step S2), K most similar resources T_{R0} are identified from N_1 (Step S4). If similar resources T_{R0} have a similarity value above a minimal similarity threshold set to 0.3 in our analysis, they are added to C_R 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_1, K most similar consumers T_{C0} are identified from N_0. 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_R for T_{Ci} if they are also found in N_1 (Steps S10 to S12). Otherwise, Steps S4 to S7 are duplicated to identify potential similar resources for T_{Ci} in N_1 from the set of resources T_R of similar consumers T_{C0} (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$_y$* is the fraction of interactions correctly predicted:

$$\text{Score}_y = \frac{a}{a + c} \qquad (3)$$

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

$$\text{Score}_{\neg y} = \frac{d}{b + d} \qquad (4)$$

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

$$\text{TSS} = \frac{(ad - bc)}{(a + c)\,(b + d)} \qquad (5)$$

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_t values of 0.5 and 1 to evaluate whether taxonomic similarity could support predictive accuracy in cases when empirical data for species in N_1 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 $Score_y$ and $Score_{-y}$ due to misclassified interactions and non-interactions. This can also be observed through the effect of varying K values, which increased the number of potential candidate resources for each taxon in the predictive portion of the algorithm. Prediction accuracy increased for interactions, while it decreased for non-interactions, as K values increase.

Similarity being predominantly measured with resource/consumer sets (w_t closer to 0) yielded better predictions than when measured with taxonomy (w_t closer to 1; Fig. 2). Resource/consumer sets therefore appear to serve as a better measure of similarity between taxa. Note that, although the predictive contribution of the algorithm decreases as w_t increases, an increased mean and decreased variability values for the TSS and $Score_y$ statistics is also observed (Fig. 2). This suggests that while resource/consumer similarity yielded higher predictive accuracy, taxonomy better complements the catalogue contribution by predicting interactions not captured through empirical data, effectively increasing the predictive accuracy of the complete algorithm.

The partitioning of the catalogue and predictive portions of the algorithm revealed the importance of the comprehensiveness of the catalogue in prediction accuracy (Figs 2, 3). As the amount of empirical data available in the catalogue increased so did the overall accuracy of the algorithm (Fig. 3). While prediction accuracy of the predictive portion of the algorithm was somewhat lower, it nonetheless supported high prediction accuracy when the catalogue comprehensiveness was lower (Fig. 3). Prediction accuracy still remained around 75 % with only 40 % of N_1 taxa found in the catalogue (Fig. 3). Furthermore, 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

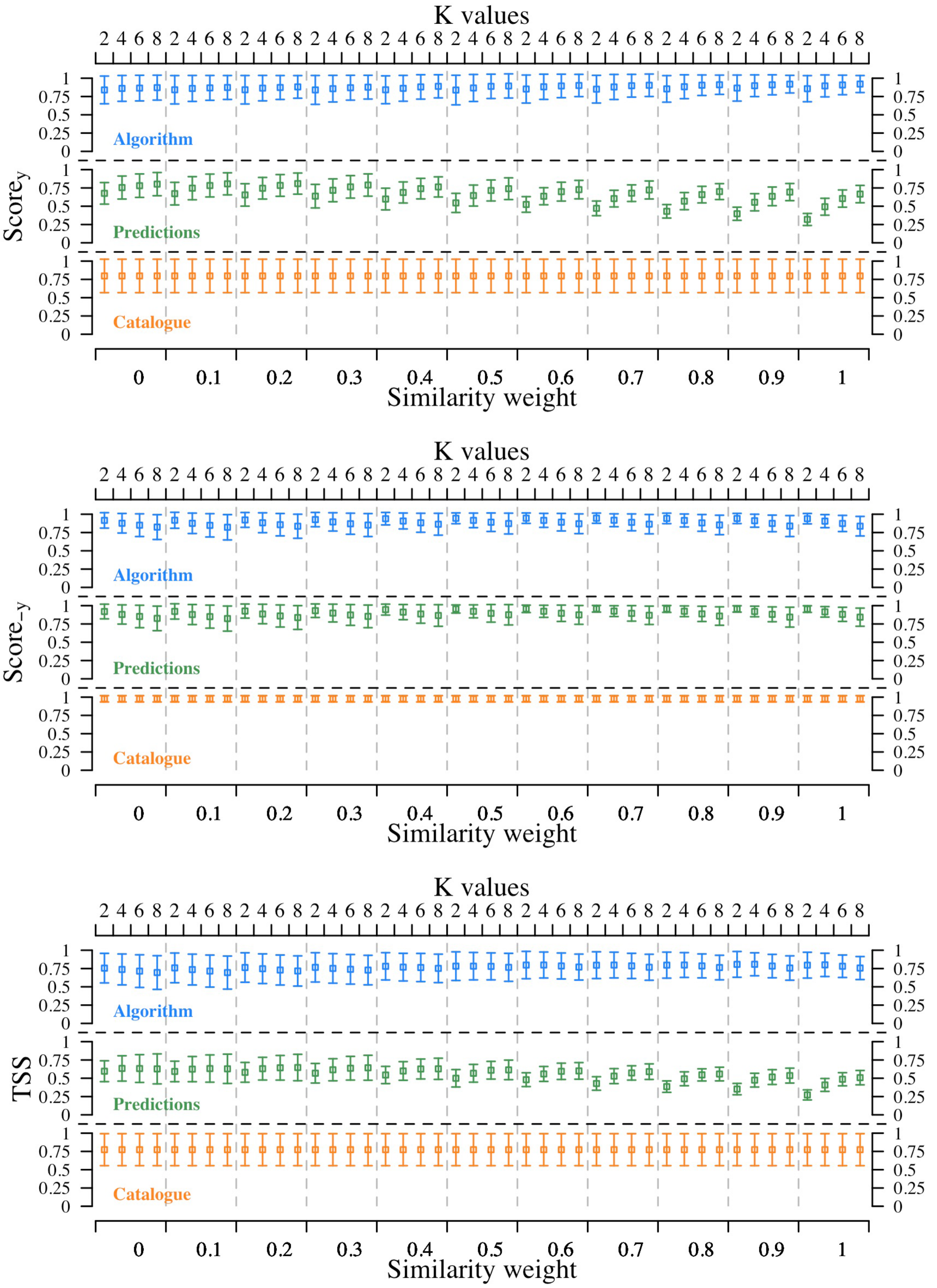


Fig. 2. – Representation of the three statistics (*i.e. Score*$_y$, *Score*$_{-y}$ and TSS) used to evaluate the accuracy of the algorithm as a function of K values tested (*i.e.* 2, 4, 6 and 8 most similar neighbors, top *x*-axis) and weight for taxonomy (bottom *x*-axis), which varies between 0 and 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).

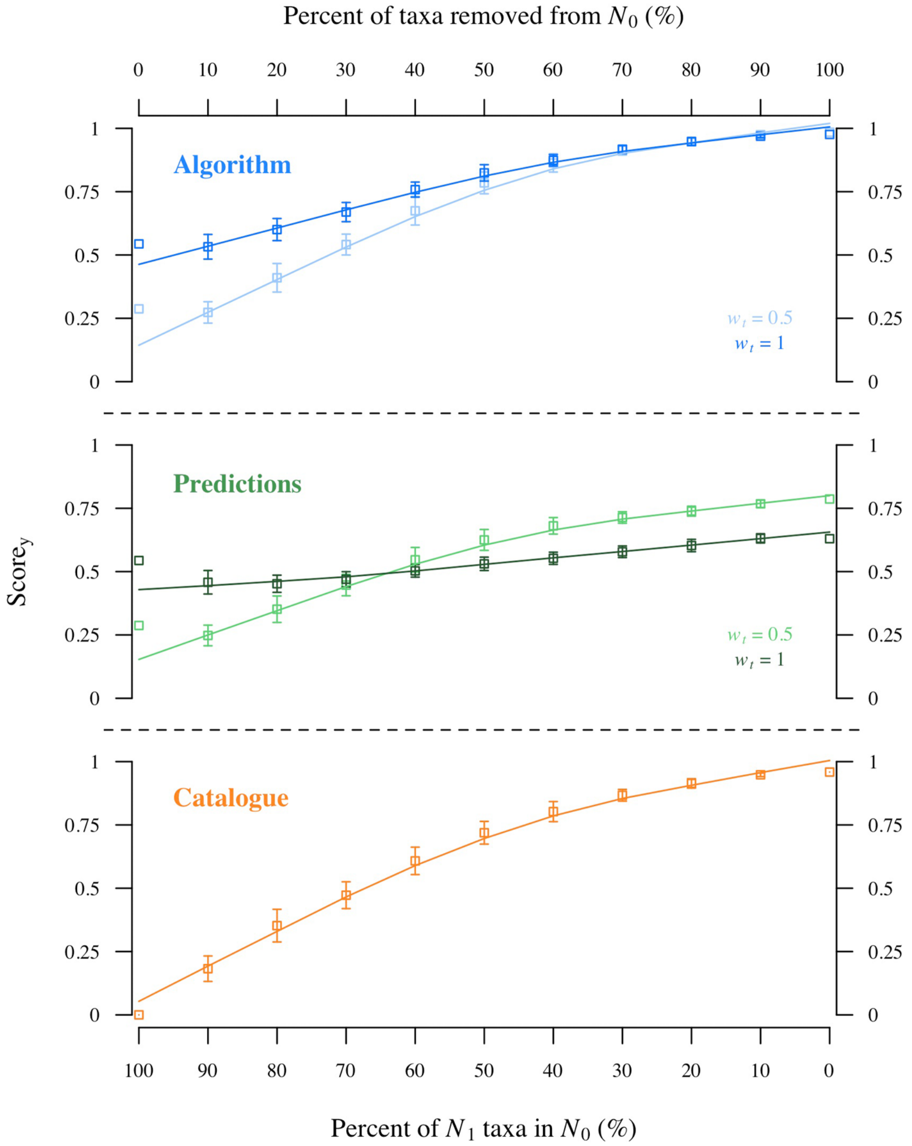


Fig. 3. – Representation of *Score$_y$* as a function of catalogue comprehensiveness, *i.e.* the amount of information on sets of consumer and resources available in the catalogue. The sensitivity of the algorithm to data accuracy was evaluated with the arctic food web from Kortsch *et al.* (2015). This 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_t 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.

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

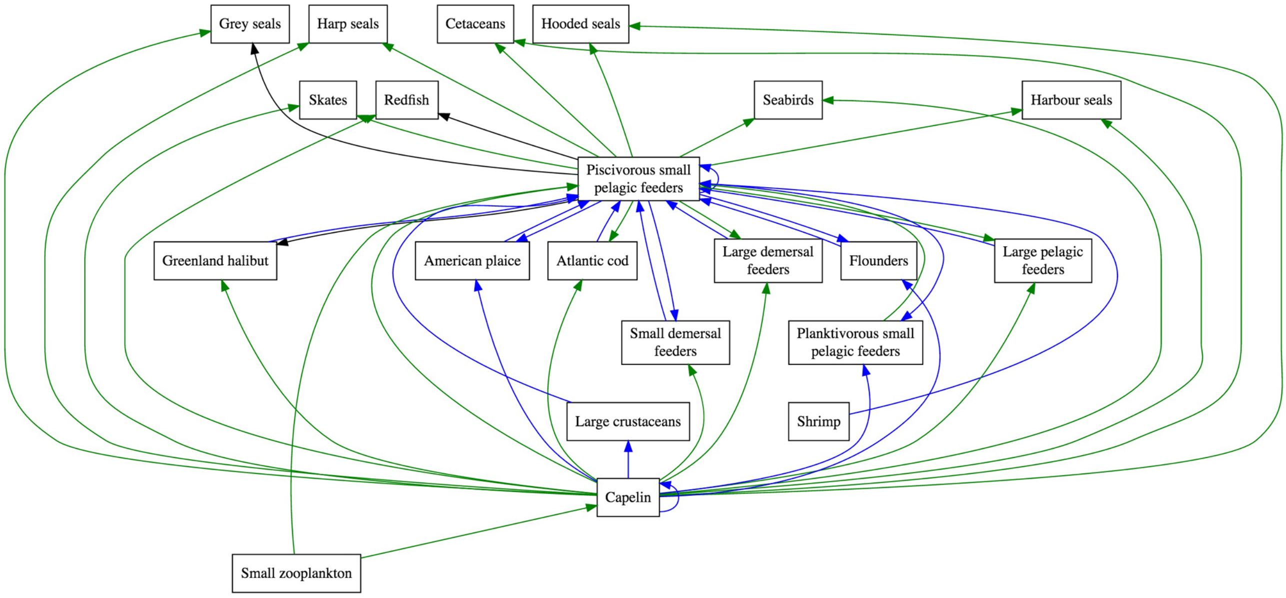


Fig. 4. – Figure 4: Example of predicted interactions with the network of the southern Gulf of St. Lawrence (Savenkoff *et al.* 2004) centered on the interactions of the capelin (*Mallotus villosus*) and piscivorous small pelagic feeders (*e.g. Scomber scombrus* and *Illex illecebrosus*). Edge with colors green (dashed) were both predicted and observed (26), black (solid) were observed only (3) and blue (dotted) were predicted only (19). Arrows are pointed towards consumers.

means that non-interactions are misclassified as interactions in the process of increasing $Score_y$, consequently decreasing $Score_{-y}$. 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 ontogenic 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, 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 (Eklöf *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 CHONeII 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

Albouy C, Velez L, Coll M, Colloca F, Le Loc'h D, Mouillot D, Gravel D 2014. From projected species distribution to food-web structure under climate change. *Glob Change Biol* 20 (3): 730-741.

Allouche O, Tsoar A, Kadmon R 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *J Appl Ecol* 43(6): 1223-1232.

Bailly N *et al*. 2016. World Register of Marine Species (WoRMS). http://www.marinespecies.org.

Barnes C, Bethea DM, Brodeur RD, Spitz J, Ridoux V, Pusineri C, Chase BC, Hunsicker ME, Juanes F, Kellermann A, Lancaster J, Ménard F, Bard FX, Munk P, Pinnegar JK, Scharf FS, Rountree RA, Stergiou KI, Sassa C, Sabates A, Jennings S 2008. Predator and prey body sizes in marine food webs. *Ecology* 89(3): 881.

Bartomeus I, Gravel D, Tylianakis JM, Aizen MA, Dickie IA, Bernard-Verdier M 2016. A common framework for identifying linkage rules across different types of interactions. *Funct Ecol*. doi:10.1111/1365-2435.12666.

Beauchesne D, Grant C, Gravel D, Archambault P 2016. L'évaluation des impacts cumulés dans l'estuaire et le golfe du Saint-Laurent : vers une planification systémique de l'exploitation des ressources. *Nat Can* 140(2): 45-55.

Brose U, Cushing L, Berlow EL, Jonsson T, Banasek-Richter C, Bersier LF, Blanchard JL, Brey T, Carpenter SR, Cattin Blandenier MF, Cohen JE, Dawah HA, Dell T, Edwards F, Harper-Smith S, Jacob U, Knapp RA, Ledger ME, Memmott J, Mintenbeck K, Pinnegar JK, Rall BC, Rayner T, Ruess L, Ulrich W, Warren P, Williams RJ, Woodward G, Yodzis P, Martinez ND 2005. Body sizes of consumers and their resources. *Ecology* 86(9): 2545-2545.

Brose U, Jonsson T, Berlow EL, Warren P, Banasek-Richter C, Bersier LF, Blanchard JL, Brey T, Carpenter SR, Cattin Blandenier M-, Cushing L, Dawah HA, Dell T, Edwards F, Harper-Smith S, Jacob U, Ledger ME, Martinez ND, Memmott J, Mintenbeck K, Pinnegar JK, Rall BC, Rayner TS, Reuman DC, Ruess L, Ulrich W, Williams RJ, Woodward G, Cohen JE 2006. Consumer-resource body-size relationships in natural food webs. *Ecology* 87(10): 2411-2417.

Cazelles, K, Araújo MB, Mouquet N, Gravel D 2016. A theory for species co-occurrence in interaction networks. *Theor Ecol* 9(1): 39-48.

Chamberlain SA, Szöcs E 2013. Taxize: taxonomic search and retrieval in R. F1000Research 2.

Chamberlain SA, Szöcs E, Boettiger C, Ram K, Bartomeus I, Baumgartner J 2014. Taxize: Taxonomic information from around the web. https://github.com/ropensci/taxize.

Christian RR, Luczkovich JJ 1999. Organizing and understanding a winter's seagrass foodweb network through effective trophic levels. *Ecol Modell* 117(1): 99-124.

Cohen JE, Jonsson T, Carpenter SR 2003. Ecological community description using the food web, species abundance, and body size. *PNAS USA* 100(4): 1781-1786.

Desjardins-Proulx P, Laigle I, Poisot T, Gravel D 2016. Ecological Interactions and the Netflix Problem. bioRxiv 089771, doi: 10.1101/089771.

Dunne JA 2006. The network structure of food webs. *In* Pascual M, Dunne JA eds, Ecological Networks: Linking Structure to Dynamics in Food Webs. Oxford University Press: 27-86.

Eklöf A, Stouffer DB 2016. The phylogenetic component of food web structure and intervality. *Theor Ecol* 9(1): 107-115.

Eklöf A, Helmus MR, Moore M, Allesina S 2012. Relevance of evolutionary history for food web structure. *Proc R Soc Lond B Biol* 279(1733). DOI: 10.1098/rspb.2011.2149.

Eklöf A, Jacob U, Kopp J, Bosch J, Castro-Urgal R, Chacoff NP, Chacoff NP, Dalsgaard B, de Sassi C, Galetti M, Guimarães PR, Lomáscolo SB, Martín González AM, Pizo MA, Rader R, Rodrigo A, Tylianakis JM, Vázquez DP, Allesina S 2013. The dimensionality of ecological networks. *Ecol Lett* 16(5): 577-583.

Giakoumi S, Halpern BS, Michel LN, Gobert S, Sini M, Boudouresque CF, Gambi MC, Katsanevakis S, Lejeune P, Montefalcone M, Pergent G, Pergent-Martini C, Sanchez-Jerez P, Velimirov B, Vizzini S, Abadie A, Coll M, Guidetti P, Micheli F, Possingham HP 2015. Towards a framework for assessment and management of cumulative human impacts on marine food webs. *Conserv Biol* 29(4): 1228-1234.

Gravel D, Poisot T, Albouy C, Velez L, Mouillot D 2013. Inferring food web structure from predator-prey body size relationships. *Method Ecol Evol* 4(11): 1083-1090.

Gray C, Figueroa DH, Hudson LN, Ma A, Perkins D, Woodward G 2015. Joining the dots: an automated method for constructing food webs from compendia of published interactions. *Food Webs* 5: 11-20.

Hedges SB, Marin J, Suleski M, Paymer M, Kumar S 2015. Tree of life reveals clock-like speciation and diversification. *Mol Biol Evol* 32(4): 835-845.

Ings TC, Montoya JM, Bascompte J, Blüthgen N, Brown L, Dormann CF, Edwards F, Figueroa D, Jacob U, Jones JI, Lauridsen RB, Ledger ME, Lewis HM, Olesen JM, van Veen FJF, Warren PH, Woodward G 2009. Review: ecological networks – beyond food webs. *J Anim Ecol* 78(1): 253-269.

Kortsch S, Primicerio R, Fossheim M, Dolgov AV, Aschan M 2015. Climate change alters the structure of arctic marine food webs due to poleward shifts of boreal generalists. *Proc R Soc Lond B Bio* 282(1814): 1-9.

Laska MS, Wootton JT 1998. Theoretical concepts and empirical approaches to measuring interaction strength. *Ecology* 79(2): 461-476.

Link J 2002. Does food web theory work for marine ecosystems? *Mar Ecol Prog Ser* 230: 1-9.

Losos JB 2008. Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. *Ecol Lett* 11(10): 995-1003.

Martinez ND 1992. Constant connectance in community food webs. *Am Nat* 139(6): 1208-1218.

Morales-Castilla I, Matias MG, Gravel D, Araújo MB 2015. Inferring biotic interactions from proxies. *Trends Ecol Evol* 30(6): 347-356.

Mouquet N, Devictor V, Meynard CN, Munoz F, Bersier LF, Chave J, Couteron P, Dalecky A, Fontaine C, Gravel D, Hardy OJ, Jabot F, Lavergne S, Leibold M, Mouillot D, Münkemüller R, Pavoine S, Prinzing A, Rodrigues ASL, Rohr RP, Thébault E, Thuiller W 2012. Ecophylogenetics: advances and perspectives. *Biol Rev* 87(4): 769-785.

Murphy KP 2012. Machine Learning: A Probabilistic Perspective. MIT Press: 1067 p.

Pascual M, Dunne JA 2006. Ecological Networks: Linking Structure to Dynamics in Food Webs. Oxford University Press: 415 p.

Penone C, Davidson AD, Shoemaker KT, Di Marco M, Rondinini C, Brooks TM, Young BE, Graham CH, Costa GC 2014. Imputation of missing data in life-history trait datasets: which approach performs the best? *Method Ecol Evol* 5(9): 961-970.

Poelen JH, Gosnell S, Slyusarev S 2015. rglobi: R Interface to Global Biotic Interactions. https://cran.r-project.org/package=rglobi.

Poelen JH, Simons JD, Mungall CJ 2014. Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecol Inform* 24: 148-159.

Polis GA 1991. Complex trophic interactions in deserts: an empirical critique of food-web theory. *Am Nat* 138(1): 123-155.

Rohr RP, Bascompte J 2014. Components of Phylogenetic Signal in Antagonistic and Mutualistic Networks. *Am Nat* 184(5): 556-564.

Savenkoff C, Bourdages H, Swain DP, Despatie SP, Hanson JM, Méthot R, Morissette L, Hammil MO 2004. Input data and parameter estimates for ecosystem models of the southern Gulf of St. Lawrence (mid-1980s and mid-1990s). Tech rep Mont-Joli, Québec, Canada: Canadian Technical Report of Fisheries, Aquatic Sciences 2529, Department of Fisheries, and Oceans.

Schrodt F, Kattge J, Shan H, Fazayeli F, Joswig J, Banerjee A, Reichstein M, Bönisch G, Díaz S, Dickie J, Gillison A, Karpatne A, Lavorel S, Leadley P, Wirth CB, Wright IJ, Wright SJ, Reich PB 2015. BHPMF – a hierarchical Bayesian approach to gap-filling and trait prediction for macroecology and functional biogeography. *Global Ecol Biogeogr* 24(12): 1510-1521.

Séguin A, Harvey E, Archambault P, Nozais C, Gravel D 2014. Body size as a predictor of species loss effect on ecosystem functioning. *Sci Rep* 4: 4616

Tamaddoni-Nezhad A, Milani GA, Raybould A, Muggleton S 2013. Chapter Four – Construction and validation of food webs using logic-based machine learning and text mining. *Adv Ecol Res*. 49: 225-289.

Thompson RM, Mouritsen KN, Poulin R 2004. Importance of parasites and their life cycle characteristics in determining the structure of a large marine food web. *J Anim Ecol* 74(1): 77-85.

Tylianakis JM, Didham RK, Bascompte J, Wardle DA 2008. Global change and species interactions in terrestrial ecosystems. *Ecol Lett* 11(12): 1351-1363.

University of Canberra 2016. Food Web Database. University of Canberra. http://globalwebdb.com/.

Wiens JJ, Ackerly DD, Allen AP, Anacker BL, Buckley LB, Cornell HV, Damschen EI, Davies JT, Grytnes JA, Harrison SP, Hawkins BA, Holt RD, McCain CM, Stephens PR 2010. Niche conservatism as an emerging principle in ecology and conservation biology. *Ecol Lett* 13(10): 1310-1324.

Yeakel JD, Guimarães PR, Bocherens H, Koch PL 2013. The impact of climate change on the structure of Pleistocene food webs across the mammoth steppe. *Proc R Soc Lond B Biol* 280(1762). DOI: 10.1098/rspb.2013.0239.

Yeakel JD, Pires MM, Rudolf L, Dominy NJ, Koch PL, Guimarães PR, Gross T 2014. Collapse of an ecological network in Ancient Egypt. *PNAS USA* 111(40): 14472-14477.

Received on September 20, 2016
Accepted on October 21, 2016
Associate editor: F Charles

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.

N_0 **taxa ID**	**Taxonomy**	**Resource**	**Consumer**
T_1	$\{a,b,c\}$	$\{T_2,T_3,T_{12}\}$	$\{T_4\}$
T_2	$\{e,f,g\}$		$\{T_1,T_5,T_6\}$
T_3	$\{i,j,k\}$		$\{T_1\}$
T_4	$\{m,n,o\}$	$\{T_1,T_5\}$	
T_5	$\{a,b,d\}$	$\{T_2,T_8,T_9\}$	$\{T_4\}$
T_6	$\{i,q,r\}$	$\{T_2,T_7,T_8\}$	
T_7	$\{e,f,h\}$		$\{T_1,T_6\}$
T_8	$\{s,t,u\}$		$\{T_5,T_6\}$
T_9	$\{s,t,v\}$		$\{T_5\}$
T_{10}	$\{i,j,l\}$		
T_{11}	$\{m,n,p\}$		
T_{12}	$\{q,r,s\}$		$\{T_1\}$

Similarity between all pairs of taxa in N_0 is measured for consumer, resource and taxonomic proximity 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.

tanimoto(T_Cx; T_Cy) / tanimoto(T_Rx; T_Ry)

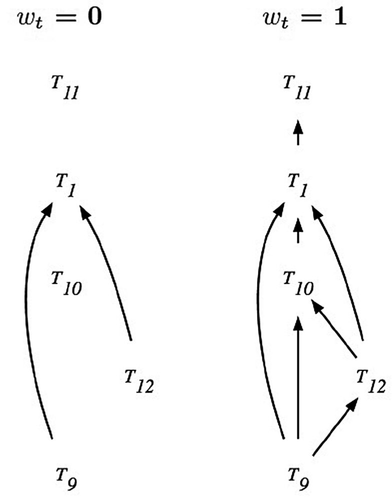
	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}	T_{11}	T_{12}
T_1	-	0	0	0	0.2/1	0	0	0	0	0	0	0
T_2	0	-	0/0.5	0	0	0	0/0.3	0/0.3	0/0.5	0	0	0/0.5
T_3	0	0	-	0	0	0	0/0.5	0	0	0	0	1
T_4	0	0	0	-	0	0	0	0	0	0	0	0
T_5	0.5	0	0	0	-	0.25/0	0	0	0	0	0	0
T_6	0	0	0.2	0	0	-	0	0	0	0	0	0
T_7	0	0.5	0	0	0	0	-	0/0.3	0	0	0	0/0.5
T_8	0	0		0	0	0	0	-	0/0.5	0	0	0
T_9	0	0	0	0	0	0	0	0.5	-	0	0	0
T_{10}	0	0	0.5	0	0	0.2	0	0	0	-	0	0
T_{11}	0	0	0	0.5	0	0	0	0	0	0	-	0
T_{12}	0	0	0	0	0	0.5	0	0.2	0.2	0	0	-

tanimoto(T_Tx; T_Ty)

From these, the algorithm goes through logical steps (Fig. 1) to identify a candidate resource list C_R 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 $w_t = 1$:

Steps		Catalogue	Prediction
1	I(T_1; T_R) in N_0?		
2	T_R in N_1?		
4-7	T_2 = no t(T_2,$T_{R'}$,w_t) = NA	{}	{}
4-7	T_3 = no = t(T_3,$T_{R'}$,w_t) = T_{10} = 0.5	{}	$\{T_{10}\}$
3	T_{12} = yes	$\{T_{12}\}$	$\{T_{10}\}$
8	t(T_1,$T_{C'}$,w_t) = T_5 = 0.5		
9	I(T_5,T_R) in N_1?		
13-16	T_2 = no = t(T_2,$T_{R'}$,w_t) = NA	$\{T_{12}\}$	$\{T_{10}\}$
13-16	T_8 = no = t(T_8,$T_{R'}$,w_t) = T9 = 0.5	$\{T_{12}\}$	$\{T_9,T_{10}\}$
10-12	T_9 = yes	$\{T_9,T_{12}\}$	$\{T_9,T_{10}\}$

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 $w_t = 0$ and 1 predicts the following networks:



$w_t = 0$
$w_t = 1$
T_{11}
T_1
T_{10}
T_{12}
T_9
T_{11}
T_1
T_{10}
T_{12}
T_9