Theor Ecol
DOI 10.1007/s12080-015-0281-9

ORIGINAL PAPER

A theory for species co-occurrence in interaction networks

Kévin Cazelles[1,2] · Miguel B. Araújo[3,4,5] · Nicolas Mouquet[1] · Dominique Gravel[2]

Received: 1 October 2014 / Accepted: 13 October 2015

Abstract The study of species co-occurrences has been central in community ecology since the foundation of the discipline. Co-occurrence data are, nevertheless, a neglected source of information to model species distributions and biogeographers are still debating about the impact of biotic interactions on species distributions across geographical scales. We argue that a theory of species co-occurrence in ecological networks is needed to better inform interpretation of co-occurrence data, to formulate hypotheses for different community assembly mechanisms, and to extend the analysis of species distributions currently focused on the relationship between occurrences and abiotic factors. The main objective of this paper is to provide the first building blocks of a general theory for species co-occurrences. We formalize the problem with definitions of the different probabilities that are studied in the context of co-occurrence analyses. We analyze three species interactions modules and conduct multi-species simulations in order to document five principles influencing the associations between species within an ecological network: (i) direct interactions impact pairwise co-occurrence, (ii) indirect interactions impact pairwise co-occurrence, (iii) pairwise co-occurrence rarely are symmetric, (iv) the strength of an association decreases with the length of the shortest path between two species, and (v) the strength of an association decreases with the number of interactions a species is experiencing. Our analyses reveal the difficulty of the interpretation of species interactions from co-occurrence data. We discuss whether the inference of the structure of interaction networks is feasible from co-occurrence data. We also argue that species distributions models could benefit from incorporating conditional probabilities of interactions within the models as an attempt to take into account the contribution of biotic interactions to shaping individual distributions of species.

Keywords Co-occurrence · Ecological networks · Biogeography · Indirect interactions · Null models

Kévin Cazelles and Dominique Gravel contributed equally to this work.

✉ Kévin Cazelles
kevin.cazelles@univ-montp2.fr

[1] Institut des Sciences de l'Evolution, CNRS UMR 5554, Université de Montpellier, Place Eugène Bataillon, CC 065, 32095, Montpellier Cedex 5, France

[2] Département de Biologie, Chimie et Géographie, Université du Québec à Rimouski, 300 Allée des Ursulines, Québec G5L 3A1, Canada

[3] Imperial College London, Silwood Park Campus, Buckhurst Road, Ascot SL5 7PY, Berks, UK

[4] Department of Biogeography Global Change, National Museum of Natural Sciences, CSIC, Calle José Gutiérrez Abascal, 2, ES-28006, Madrid, Spain

[5] Center for Macroecology, Evolution and Climate, University of Copenhagen, Universitetsparken 15, Copenhagen 2100, Denmark

Introduction

Understanding of the processes driving the assembly of communities has been a central theme of ecology since the foundation of the discipline. How do we start from a

Published online: 27 October 2015

regional species pool to assemble a structured community? Why are some species associated with each other? The work of Diamond (1975) pioneered the analysis of species co-occurrence in geographical space and, together with the controversy triggered by Connor and Simberloff (1979), it stimulated the development of a new field of research in numerical ecology (Stone and Roberts 1990; Gotelli and Graves 1996; Legendre and Legendre 2012). The foundational work on species co-occurrences also led to the development of a rich array of methodological tools designed to test null hypotheses in ecology. Even if null models could be achieved numerically (e.g., Aráujo et al. 2011), typically they are based on permutations of distribution data. Null models have been used to infer the role of biotic interactions between pairs of species on their individual distributions. Studying the different drivers of species co-occurrence is not only of theoretical interest for improving understanding of the mechanisms of community assembly. It is also instrumental in predictive ecology, because a considerable amount of information is contained in species distributions data.

Despite its historical importance for community ecology, co-occurrence data remain a neglected source of information in models of species distributions. Biogeographers are still debating the impact of biotic interactions on species distributions (Guisan and Thuiller 2005; Gotelli et al. 2010; Kissling et al. 2012; Pellissier et al. 2013). The distribution of a species is thought to be first influenced by its physiological tolerance to environmental conditions, but also by interactions with other species (Hutchinson 1957; MacArthur 1972; Peterson 2011; Boulangeat et al. 2012). The question of whether such interactions leave imprints in the distributions of individual species at biogeographical scales is still open to debate (e.g., Davis et al. 1998), but recent empirical (Gotelli et al. 2010), modeling (e.g., Aráujo and Luoto 2007), and theoretical (Aráujo et al. 2011; Jabot and Bascompte 2012) evidence invites the interpretation that this might indeed be the case.

The overwhelming majority of species distributions modeling applications, nonetheless, neglect information contained in joint distributions. Even multivariate analysis of community data (e.g., redundancy analysis—Legendre and Legendre 2012) do not use co-occurrence in geographical space to condition individual species response to environmental variation. There has been a recent rise of interest, however, in joint species distribution modeling (Clark et al. 2014; Harris 2015; Pollock et al. 2014). These methods estimate the distribution of all species from a pool simultaneously and allow to condition the presence of a species on all other ones. However, estimated relationships are inferred from co-occurrence in environmental space rather than geographical space. That is, joint responses to the environment are inferred rather than biotic interactions themselves (Baselga and Aráujo 2009). JSDMs are, nonetheless, a first step towards developing a next generation of models accounting for the impact of biotic interactions on the distributions of species. They are, however, purely empirically driven and carry no specific hypotheses about how interactions can affect distributions. An exception is the recent attempt to model the effects of predator-prey dynamics on distributions and abundances using a metacommunity framework coupled with phenomenological species distributions models (Fordham et al. 2013). The problem with such approaches is that data to parameterize interactions mechanistically are generally lacking (Morales-Castilla et al. 2015); therefore, they are hardly applied in most circumstances. It follows that we are faced with at least two major problems: (i) understanding of the ecological interactions underlying the distributions of species is limited, and (ii) knowledge of interactions is typically limited to net interactions, mixing both direct and indirect interactions. A theory of species co-occurrences in ecological networks is, therefore, needed to help interpret co-occurrence data, to formulate hypotheses for different community assembly mechanisms, and to extend the analysis of species distributions currently focused on the relationship between occurrences and abiotic factors.

The analysis of species co-occurrences starts with a matrix representing the presence and absence of each species over a set of sites. There are two aspects to the quantitative study of co-occurrence. The first is the choice of the metric used to quantify the strength of associations (relationships between species occurrences) between pairs of species. The simplest measure of species co-occurrence is the number of species combinations, as defined by Pielou and Pielou (1968). A second index is the count of checkerboards Diamond (1975): "In such a pattern, two or more ecologically similar species have mutually exclusive but interdigitating distributions in an archipelago, each island supporting only one species" (p. 32). Another popular index of co-occurrence is the C-score (Stone and Roberts 1990). This index is similar to the count of checkerboards; it measures the average association or repulsion between pairs of species.

The second aspect of the analysis of species co-occurrence is the formulation of a null model. The controversy generated in Connor and Simberloff (1979) was partly (and rightly) based on the absence of a valid null hypothesis in Diamond's analysis. Subsequent debates were mostly concerned with the formulation of the null hypothesis (e.g., Diamond and Gilpin 1982). Thanks to the theoretical work of Gotelli and Graves (1996), there is now a clear understanding of the different null models that can be constructed

from the community matrix. New indices are constantly proposed, such as in Boulangeat et al. (2012) and Veech (2013); see also Table 2 in Ulrich and Gotelli (2013) for a description of 15 indices for co-occurrence analysis. A promising avenue is the one proposed by Aráujo et al. (2011) for the study of the matrix of species co-occurrence with tools borrowed from network theory.

Surprisingly, there is currently no theory for co-occurrence in multi-species communities. The basic hypotheses are that pairwise negative interactions result in repulsion, while pairwise positive interactions result in attraction. Attraction and repulsion are assessed by a comparison of the number of co-occurrence events to the number expected under a totally independent distribution. Similar environmental requirements between species could also result in attraction, even in the absence of interactions, if the sampling is conducted across heterogeneous environmental conditions. This theory is limited to pairwise and symmetric interactions; there is nothing for antagonistic and indirect interactions. Food web ecologists were among the first to recognize the important effect of indirect interactions on abundance (Wootton 1994). For instance, plant and carnivore abundances are expected to correlate across a productivity gradient (Hairston et al. 1960; Oksanen et al. 1981) because of top-down control on the herbivore population. Similarly, the propagation of indirect interactions has been studied in more complex interaction networks (Yodzis 1988). Indirect interactions could reverse the net interaction in a surprising way, such that predator-prey abundances could be positively related (Montoya et al. 2009). Empirical analysis of co-occurrence for several taxa has shown that they are usually asymmetric (Aráujo et al. 2011), such that a species distribution tended to be nested within the distribution of other (e.g., predator-prey distributions; Holt and Barfield 2009; Gravel et al. 2011). In such a case, even if the co-distribution signature is quite understood, available methods will likely fail at using this piece of information to improve forecasts.

The main objective of this paper is to provide the first building blocks of a general theory of species co-occurrences. We formalize the proposed theory with definitions of different quantities that are studied in the context of co-occurrence analyses. Herewith, we analyze three species interactions modules in order to document five principles influencing the association between pairs of species from an ecological network: (i) direct interactions impact pairwise co-occurrence; (ii) indirect interactions impact pairwise co-occurrence; (iii) pairwise co-occurrence does not have to be symmetric; (iv) the strength of an association decreases with the length of the shortest path between two species; and (v) strength of an association decreases with the number of interactions a species is experiencing. We base our mathematical argument on a general model of species distributions that is free of any assumption about how the ecological interactions operate. Finally, we extend our analysis with simulations of multi-species networks in order to analyze how these mechanisms scale up in species-rich communities.

Definitions

We start with definitions to formalize the quantities that can be computed from species distribution data and be used in the context of co-occurrence analyses. Let X_i be the random variable representing the presence of species i. $X_i = 1$ when species i is present, $X_i = 0$ otherwise. Then $X_{i,t>0}$ is the random process associated, giving the value that $X_{i,t}$ takes at any time t. Let $p_{i,t}$ standing for the probability $\mathbb{P}(X_{i,t} = 1)$. Also, to illustrate the definitions, we derive the quantities for a simple presence/absence dataset (see Table 1).

The **marginal occurrence probability** $\mathbb{P}(X_{i,\infty} = 1) = p_i^*$ represents the occurrence probability of species i when the system is at equilibrium, in the sense of the classical theory of island Biogeography MacArthur and Wilson (1967). As we assume so for all species, we drop the $*$ and the ∞ for the sake of clarity. The marginal occurrence probability is the sum of the occurrence of the species across all possible set of species in the data. In other words, it corresponds to the sum of the column of the site $\times$ species table, divided by the total number of sites N. Marginal occurrence probabilities for species in Table 1 are $p_1 = 0.6$, $p_2 = 0.6$, and $p_3 = 0.4$.

The **observed co-occurrence** between species i and j is the joint probability $p_{i,j} = \mathbb{P}(X_i = 1 \cap X_j = 1)$. It represents the number of sites where the two species are found together, across all possible set of species in the data (in other words, it is a marginal probability with respect to other species), divided by N. In our dataset, for instance, we have $p_{1,2} = 0.3$ and $p_{1,3} = 0.2$.

The **conditional co-occurrence** between species i and j is $p_{i|j} = \mathbb{P}(X_i = 1|X_j = 1)$. It represents the probability of observing species i, knowing that species j is already present. This quantity is close to the measure of association between two species because it is independent of the marginal occurrence probability of both species. The problem is that, as soon as there are other species present, the conditional co-occurrence as expressed here is marginalized over the set of all other species from the community K. For instance, for three species, we have $p_{1|2} = \mathbb{P}(X_1 = 1|X_2 = 1, X_3 = 1) + \mathbb{P}(X_1 = 1|X_2 = 1, X_3 = 0)$. It, therefore, includes both the effect of *direct* and*indirect*

Table 1 Presence/absence dataset for three species and 10 sites

Sites	Species 1	Species 2	Species3
1	0	1	1
2	0	1	1
3	1	1	0
4	1	0	1
5	0	0	0
6	1	1	1
7	0	1	0
8	1	0	0
9	1	0	0
10	1	1	0

associations between species, e.g., the direct association of species 1 with species 2 or the indirect association of species 3 with 1 via its effect on 2. Consequently, the measure of pairwise association should be $p_{i|j,\overline{K}} = \mathbb{P}(X_i = 1|X_j = 1, X_K = 0)$, where the horizontal bar over K denotes absence of all other species. We name this the **fundamental conditional co-occurrence**. For instance, in Table 1, we get $p_{1|2} = \frac{p_{1,2}}{p_2} = 0.5$ and $p_{1|2,\overline{3}} = \frac{p_{1,2,\overline{3}}}{p_{2,\overline{3}}} = \frac{0.2}{0.3} = 0.67$.

Following the same logic, we define the **fundamental occurrence** as $p_{i|\overline{K}} = \mathbb{P}(X_i = 1|X_K = 0)$. The fundamental occurrence is conceptually equivalent to the fundamental niche of Hutchinson (1957) and represents the probability of observing a species in the absence of biotic interactions, i.e., when all other species are absent. By analogy, the marginal occurrence should be interpreted as the realized distribution. For species 1 in Table 1, we calculate $p_{1|\overline{23}} = \frac{p_{1,\overline{2},\overline{3}}}{p_{\overline{2},\overline{3}}} = \frac{0.2}{0.3} = 0.67$. Finally, we define the **independent co-occurrence** as $p_{i,j;IND} = \mathbb{P}(X_i = 1)\mathbb{P}(X_j = 1)$. It represents the co-occurrence between any pairs of species expected in the absence of any association between them. In ecological terms, it would represent the co-occurrence when ecological interactions and habitat filtering do not impact species distribution. It also represents the null model to which observed co-occurrence is usually compared. Note that the independent co-occurrence is different from the one expected under a neutral model (Hubbell 2001). Firstly because strong competitive interactions in the neutral model forces repulsion and, secondly, because dispersal limitation also causes spatial aggregation and thus a non-random distribution of co-occurrence (Bell 2005). In our example, we obtain, for instance, $p_{1,2;IND} = 0.36$ and $p_{2,3;IND} = 0.24$.

Direct association between two species

We start with the analysis of a two species situation, labeled species 1 and species 2, in order to understand direct associations between species pairs. A third species, 3, will be introduced in the next section to study indirect associations. The model we develop is general, as we do not specify the type of ecological interactions involved. It therefore accounts for all possible mechanisms from which an association between a pair of species could arise, such as trophic interactions involving energy fluxes, non-consumptive interactions, parasitism, direct interference, territoriality, space pre-emption, niche construction, etc. The impact of predator-prey interactions in a metapopulation setting with colonization and extinction dynamics will be considered for the multi-species simulations.

As we are willing to understand the role played by interactions in co-occurrence, we start by defining marginal co-occurrence probabilities of our two species by a decomposition into conditional co-occurrences. By the formula of total probability and Bayes's theorem, we have:

$$\begin{aligned} p_1 &= \mathbb{P}(X_1 = 1 \cap X_2 = 1) + \mathbb{P}(X_1 = 1 \cap X_2 = 0) \\ &= \mathbb{P}(X_1 = 1|X_2 = 1)\mathbb{P}(X_2 = 1) \\ &\quad + \mathbb{P}(X_1 = 1|X_2 = 0)\mathbb{P}(X_2 = 0) \end{aligned} \tag{1}$$

We do the same for species 2. Using the notation described above, Eq. 1 could be rewritten as:

$$\begin{cases} p_1 = p_{1|2}p_2 + p_{1|\overline{2}}(1 - p_2) \\ p_2 = p_{2|1}p_1 + p_{2|\overline{1}}(1 - p_1) \end{cases} \tag{2}$$

where the vertical bar denotes the absence of a species. By solving the latter system, we get:

$$\begin{cases} p_1 = \dfrac{p_{1|\overline{2}} + p_{2|\overline{1}}(p_{1|2} - p_{1|\overline{2}})}{1 - (p_{2|1} - p_{2|\overline{1}})(p_{1|2} - p_{1|\overline{2}})} \\ p_2 = \dfrac{p_{2|\overline{1}} + p_{1|\overline{2}}(p_{2|1} - p_{2|\overline{1}})}{1 - (p_{2|1} - p_{2|\overline{1}})(p_{1|2} - p_{1|\overline{2}})} \end{cases} \tag{3}$$

When species are independent, we have $p_{1|\overline{2}} = p_{1|2} = p_1$ and $p_{2|\overline{1}} = p_{2|1} = p_2$, then we logically find Eq. 1 again. Then, we can deduce the following interpretation of the impact of **direct interactions** on co-occurrence:

i. If species 1 cannot persist in absence of 2 (e.g., a parasite requiring its host), then $p_{1|\overline{2}} \to 0$, therefore $p_1 \to p_{1|2}p_2$
ii. If species 1 depends strongly on 2 thereby perfectly tracking its distribution 2, the $p_{1|\overline{2}} \to 0$ and $p_{1|2} \to 1$, and therefore $p_1 \to p_2$
iii. If species 2 excludes 1, then $p_{1|2} \to 0$ and $p_{1|\overline{2}} \to p_1$ together with $p_{2|1} \to 0$ and $p_{2|\overline{1}} \to p_2$. Hence, for strong exclusion, we get $p_1 = \frac{p_1 - p_2 p_1}{1 - p_2 p_1}$ and $p_2 = \frac{p_2 - p_2 p_1}{1 - p_2 p_1}$. Therefore, if $p_2 \to 1$, then $p_1 \to 0$.

Co-occurrence in three-species modules

Now, we consider the co-occurrence between three species. We start with a general derivation of co-occurrence and then interpret the results for particular modules in order to reveal fundamental principles underling co-occurrence in ecological networks. Our solution provides insights to decipher the solution of species-rich networks since the three-node connected subgraphs are fundamental building blocks of larger networks (Milo et al. 2002; Stouffer et al. 2007; Stouffer and Bascompte 2010). We use the same approach as in Eq. 1 and get the subsequent equation:

$$\begin{aligned} p_1 &= \mathbb{P}(X_1 = 1 \cap X_2 = 1 \cap X_3 = 1) \\ &\quad + \mathbb{P}(X_1 = 1 \cap X_2 = 0 \cap X_3 = 1) \\ &\quad + \mathbb{P}(X_1 = 1 \cap X_2 = 1 \cap X_3 = 0) \\ &\quad + \mathbb{P}(X_1 = 1 \cap X_2 = 0 \cap X_3 = 0) \end{aligned} \tag{4}$$

As $\{X_3 = 1, X_3 = 0\}$ forms a partition, irrespective of the value for X_2, we get:

$$p_1 = \mathbb{P}(X_1 = 1|X_3 = 1)p_3 + \mathbb{P}(X_1 = 1|X_3 = 0)(1 - p_3) \tag{5}$$

This equation is analogous to the two-species interactions equation but enables the study of networks involving three species interactions, with species 2 being hidden by marginalization. We split the three species problem in two distinct two-interactions species problems. Firstly, we solve the equation for sites without species 3 and get:

$$\begin{aligned} p_{1|\bar{3}} &= \mathbb{P}(X_1 = 1|X_3 = 0) \\ &= \frac{p_{1|\bar{2}\bar{3}} + p_{2|\bar{1}\bar{3}}(p_{1|\bar{2}\bar{3}} - p_{1|\bar{2}\bar{3}})}{1 - (p_{2|1\bar{3}} - p_{2|\bar{1}\bar{3}})(p_{1|\bar{2}\bar{3}} - p_{1|\bar{2}\bar{3}})} \end{aligned} \tag{6}$$

which is similar to Eq. 3 but with an explicit absence of species 3. We do similarly for the conditional occurrence of 1 on species 3 present:

$$\begin{aligned} p_{1|3} &= \mathbb{P}(X_1 = 1|X_3 = 1) \\ &= \frac{p_{1|\bar{2}3} + p_{2|\bar{1}3}(p_{1|23} - p_{1|\bar{2}3})}{1 - (p_{2|13} - p_{2|\bar{1}3})(p_{1|23} - p_{1|\bar{2}3})} \end{aligned} \tag{7}$$

Doing so, we get the following set of equations describing the marginal occurrence probabilities for the three species:

$$\begin{cases} p_1 = p_{1|3}p_3 + p_{1|\bar{3}}(1 - p_3) \\ p_2 = p_{2|3}p_3 + p_{2|\bar{3}}(1 - p_3) \\ p_3 = p_{3|2}p_2 + p_{3|\bar{2}}(1 - p_2) \end{cases} \tag{8}$$

Note that we could have chosen a different set of equations depending on the way we split the problem, for instance, we could have started by considering the occurrence of species 1 given the occurrence of species 2 instead of species 3. Now, we solve the above linear system of three equations with three unknowns and find that:

$$\begin{cases} p_1 = \dfrac{p_{1|\bar{3}} + p_{3|\bar{2}}(p_{1|3} - p_{1|\bar{3}}) + (p_{3|2} - p_{3|\bar{2}})(p_{1|3}p_{2|\bar{3}} - p_{1|\bar{3}}p_{2|3})}{1 - (p_{2|3} - p_{2|\bar{3}})(p_{3|2} - p_{3|\bar{2}})} \\ p_2 = \dfrac{p_{2|\bar{3}} + p_{3|\bar{2}}(p_{2|3} - p_{3|\bar{2}})}{1 - (p_{2|3} - p_{2|\bar{3}})(p_{3|2} - p_{3|\bar{2}})} \\ p_3 = \dfrac{p_{3|\bar{2}} + p_{2|\bar{3}}(p_{3|2} - p_{3|\bar{2}})}{1 - (p_{2|3} - p_{2|\bar{3}})(p_{3|2} - p_{3|\bar{2}})} \end{cases} \tag{9}$$

Conditional probabilities of the right-hand sides can all be derived as we did for $p_{1|3}$ in Eq. 7.

Community modules

We now interpret these equations with examples of well-studied food web modules in community ecology: (1) linear food chain, (2) exploitative competition, and (3) apparent competition. To do so, we consider matrices of direct associations representing the conditional co-occurrence probabilities among all pairs of species (see Table 2).

We are interested by the *observed co-occurrence* because this is the quantity that is easily measurable from species distributions data, thus being the one that is typically studied. We consider that the marginal occurrence is also a known quantity and, therefore, we examine the effect of particular conditional co-occurrence arrangements on observed co-occurrences. We will not provide derivations for each module, but focus on particular pairs to illustrate two of the five principles.

Indirect interactions The comparison between the observed co-occurrence and the conditional co-occurrence reveals the role of indirect interactions on species associations. Based on Eqs. 9 and 6, we get the association between species i and k:

$$\begin{aligned} p_{i,k} &= p_i - p_{i,\bar{k}}(1 - p_k) \\ p_{i,k} &= p_i - \frac{p_{i|\overline{jk}} + p_{j|i\bar{k}}(p_{i|j\bar{k}} - p_{i|\overline{jk}})}{1 - (p_{j|i\bar{k}} - p_{j|\bar{i}\bar{k}})(p_{i|j\bar{k}} - p_{i|\overline{jk}})}(1 - p_k) \end{aligned} \tag{10}$$

Therefore, the observed co-occurrence between species i and k depends on their respective interaction with species j ($p_{j|i\bar{k}}$, $p_{j|i\bar{k}}$ and $p_{j|\bar{i}\bar{k}}$). The conditional co-occurrence between two species could be null, but their observed co-occurrence be non-independent because of a shared interaction. This principle is best illustrated by the co-occurrence between a carnivore and a plant (species 3 and 1, respectively) in a linear food chain. In this situation, according to Table 2, we find that the observed co-occurrence between

Table 2 Direct associations between pairs of species for different modules

General case

$$\begin{pmatrix} p_{1|\overline{23}} & p_{1|2\overline{3}} & p_{1|\overline{2}3} \\ p_{2|1\overline{3}} & p_{2|\overline{13}} & p_{2|\overline{1}3} \\ p_{3|1\overline{2}} & p_{3|\overline{12}} & p_{3|\overline{12}} \end{pmatrix}$$

Linear chain

$$\begin{pmatrix} p_{1|\overline{23}} & p_{1|2\overline{3}} & p_{1|\overline{2}3} \\ p_{2|1\overline{3}} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Exploitative competition

$$\begin{pmatrix} p_{1|\overline{23}} & p_{1|2\overline{3}} & p_{1|\overline{2}3} \\ p_{2|1\overline{3}} & 0 & 0 \\ p_{3|1\overline{2}} & 0 & 0 \end{pmatrix}$$

Apparent competition

$$\begin{pmatrix} p_{1|\overline{23}} & p_{1|\overline{23}} & p_{1|\overline{2}3} \\ p_{2|\overline{13}} & p_{2|\overline{13}} & p_{2|\overline{1}3} \\ p_{3|1\overline{2}} & p_{3|\overline{12}} & 0 \end{pmatrix}$$

Entries indicate the fundamental conditional probabilities of occurence of species i given the presence of species j and the absence of species k. *Linear chain*: 1 is the resource, 3 the top predator ; *Exploitative competition*: 2 and 3 are the consumers; *Apparent competition*: 1 and 2 are the resources. When $p_{i|j\overline{k}} = 0$, it means that species i cannot be found without k. When $p_{i|j\overline{k}}$=$p_{i|\overline{jk}}$ then species j does not impact species i survival. For apparent competition, if species 1 and 2 are interchangeable for species 3 then : $p_{3|1\overline{2}} = p_{3|\overline{1}2}$

the plant and the carnivore is:

$$p_{1,3} = p_1 - \frac{p_{1|\overline{23}}}{1 - p_{2|1\overline{3}}(p_{1|2\overline{3}} - p_{1|\overline{23}})}(1 - p_3) \tag{11}$$

It is clear from this equation that there is a significant association between the carnivore and the plant, despite the conditional co-occurrence of the two species being totally independent. The indirect association gets stronger with the strength of both conditional co-occurrence. Similar observations could be made by studying the observed co-occurrence between consumers (species 2 and 3) in the exploitative competition module:

$$p_{2,3} = p_2 - \frac{p_{1|\overline{23}}p_{2|1\overline{3}}}{1 - (p_{1|2\overline{3}} - p_{1|\overline{23}})p_{2|1\overline{3}}}(1 - p_3) \tag{12}$$

And between resources in the apparent competition module (species 1 and 2):

$$p_{1,2} = p_1 - \frac{p_{1|\overline{23}}}{1 - p_{3|1\overline{2}}(p_{1|\overline{2}3} - p_{1|\overline{23}})}(1 - p_2) \tag{13}$$

Associations do not have to be symmetrical Many studies of co-occurrence assume pairwise associations to be symmetrical (but see Aráujo et al. 2011; Boulangeat et al. 2012). The reason is simple, usually the observed co-occurrence is compared to the independent co-occurrence. These two metrics of association are perfectly symmetrical. This information is providing us an inappropriate interpretation of the effect of interactions on species distribution. If we consider for instance the association between the two consumers (species 2 and 3) competing for a single resource (species 1), we have the observed co- occurrence at Eq. 12, which is symmetrical by definition. The proportion of the area occupied by species 2 where species 3 is also present is not, however, equivalent to the proportion of the areas occupied by species 3. Rephrasing the problem,we find that using Eqs. 7 and 12, $p_{2,3}/p_2$ is not equal to $p_{2,3}/p_3$. One species could have a stronger impact on the distribution of the other one. Predator distribution for instance tends to be nested within the distribution of the prey (Gravel et al. 2011), and consequently the predator has a high conditional co-occurrence with the prey, and alternatively the prey has a lower conditional co-occurrence with the predator.

Multi-species simulations

Now, we move to multi-species simulations of more complex networks to reveal the last two principles of our theory. To do so, we run simulations of the model of trophic island biogeography developped by Gravel et al. (2011). The model describes the occurrence of a S species regional network. Species stochastically colonize islands with probability c and go extinct with probability e, as in the original model of MacArthur and Wilson (1967). Interactions are introduced with three additional assumptions: (i) a consumer species could colonize an island only if it has at least one prey present (for simplicity, we consider producers to be resident permanently on the island); (ii) a consumer species goes extinct if it loses its last prey species; and (iii) the presence of at least one predator species increases the extinction probability by e_d. The consequence of these assumptions is a sequential build-up of the food web on the island, starting with low trophic level species with a general diet. Small and isolated islands promote selection in favor of the most generalist species. The predictions converge to the classic island biogeography theory for highly connected regional food webs and large and connected islands (details in Gravel et al. 2011).

As mentioned above, there is a strong dependence of the predator occurrence on the presence of its preys. Alternatively, when e_d is sufficiently large, the preys will tend to avoid locations with the predator present. We consequently expect a strong signature of the network of interactions on the co-occurrence matrix. We are, however, concerned that indirect associations could emerge, as exemplified with the analysis of three species modules above, and thereby mask the signal of conditional co-occurrences.

We simulated complex networks from 5 to 100 species using the niche model of food web structure (Williams and Martinez 2000). The diversity of primary producers was fixed at 2, and their niche position was drawn randomly between 0 and 1 according to a uniform distribution. We fixed connectance at $C = 0.1$ to get comparable and realistic numbers of interactions for our simulations. Colonization probability was set at $c = 0.1$, baseline extinction probability at $e = 0.2$, and predator-dependent additional extinction

probability at $e_d = 0.2$. Simulations were run for 10^7 time steps to evaluate the conditional occurrence probabilities, and 100 replicated networks were simulated for each level of species richness.

Distance decay of observed co-occurrence The distribution of observed co-occurrence is illustrated for pairs of species separated by different path lengths at Fig. 1a. The observed co-occurrence is presented as a function of the expected co-occurrence under the hypothesis of independent distributions. The strongest associations (given by the distance between the observed and the independent co-occurrence) are observed among pairs of species directly interacting with each other. The variance of the distribution reduces from direct to first-order indirect interactions, and from first-order to higher interactions. We conclude that indirect non-independent co-occurrences are possible in complex networks, but their magnitude decreases as the number of links between two nodes decreases. This result is similar to the observation of a distance decay of indirect interactions in food webs (Berlow et al. 2009).

Strength of co-occurrence decreases with degree and species richness We performed simulations with a gradient of species richness and observed that the variance of observed co-occurrence also decreases with the degree of a species, i.e., the number of direct interactions a species is experiencing (Fig. 1b). We illustrated the relationship between the degree of a species and the observed co-occurrence for pairs of species with a direct association (Fig. 1c). This phenomenon has the consequence that the strength of observed co-occurrence reduces with species richness. The niche model has a constant connectance (Williams and Martinez 2000), which has for consequence an increase of the degree with species richness. We find that the strength of co-occurrence decreases with the degree. This result is straightforward to interpret: the more diverse are the interactions, the weaker the impact of each pairwise direct interaction on the species distribution. Again, this result is similar to the observation of a scaling relationship between pairwise interactions and food web diversity (Berlow et al. 2009).

Discussion

We first develop a probabilistic species distribution model constrained by biotic interactions using conditional probabilities of co-occurrence. We then illustrate five general principles underlying the impact of ecological interactions on co-occurrence and that should be considered for the formulation of a general theory of species co-occurrence. Two of them have been widely noted before: (i) direct interactions affect species distributions and generate deviations in co-occurrences from that expected if distributions of species were independent from each other; (ii) the effect of direct associations is often asymmetric, as envisioned in trophic metacommunity ecology (Holt and Barfield 2009). We also illustrate principles that have been overlooked in most studies of co-occurrence (Aráujo et al. 2011); (iii) indirect interactions generate deviate co-occurrence from expectation under independence assumption; (iv) the strength of indirect associations decreases with the length of the shortest path distance between species pairs in a network; while (v) also decreasing with the number of interactions a species is experiencing. We started with the analysis of three species modules to document these principles and then showed their applicability in multi-species networks. We find that the above principles also apply in larger networks, but that the strength of pairwise associations weakens as the number of species increases.

Our results have considerable implications for interpretation of co-occurrence data. Firstly, they demonstrate the considerable variety of mechanisms causing pairwise associations. Such variety of mechanisms makes interpretation of aggregated indices of co-occurrence, such as the C-score, very difficult (see also Aráujo and Luoto 2014). Previous studies already made the argument that positive and negative interactions could balance each other (Boulangeat et al. 2012) and consequently associations should be studied on a pairwise basis (Veech 2013). At least, some measure of the variability of the associations is required, and at best metrics such as network analyses (Aráujo et al. 2011) should be used to characterize their complex structure. But most importantly, our analyses reveal the difficulty to infer species interactions from co-occurrence matrices. Associations are not symmetric and, therefore, indices that are capable of dealing with them are required. Null model testing is not sufficient; significance is assessed from the difference between observed co-occurrence and co-occurrence expected under independent distributions and is, consequently, symmetric. In addition, statistically significant associations cannot be interpreted as evidence of direct interactions. Our results also show that indirect interactions, and not only second-order interactions, contribute to generate apparent non-independent co-occurrence. These indirect associations could be of any kind and are impossible to detect solely based on knowledge of direct interactions.

Null models of species associations should, thus, be used only to reveal the structure of co-occurrence data. The lack of an association between a pair of species is no unequivocal evidence of absence of direct interactions. It must be interpreted as the absence of a net effect in the spatial co-occurrence arising from pairwise interaction alone. For instance, in the case species A is competing with species

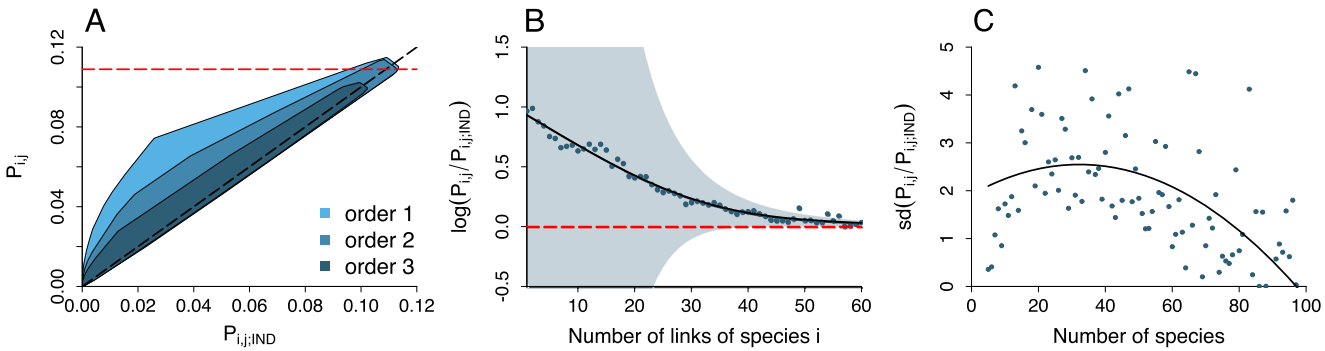

Fig. 1 Co-occurrence in multi-species networks. **a** The disparity between observed co-occurrence ($P_{i,j}$) and independent co-occurrence ($P_{i,j;\mathrm{IND}}$) decreases with the path length between nodes (species). The envelopes are drawn around the 5 and 95 % quantiles of all of the data, from 100 replicated simulations for every species richness value (5 to 100 species). **b** The strength of co-occurrence ($log(P_{i,j}/P_{i,j;\mathrm{IND}})$) decreases with the number of interactions of a species i (i.e., the degree of a node). Points represent the mean for a particular degree of node value (1 to 60). The *solid line* represents the overall trends and the *grey envelop* reflects the variance associated. At least 3000 values were used for each point. **c** The standard deviation of the strength of association ($sd(P_{i,j}/P_{i,j;\mathrm{IND}})$) and thus the variance decreases with species richness. Taken together, (**b**) and (**c**) imply that species distributions converge to independence with increasing species richness

B and species C, and B with C, it is possible that A and C could be independently co-occurring if there is a strong indirect positive interaction A-C arising from the A-B and B-C direct interactions. Null model testing is consequently subject to important type I (false interpretation of a significant association) and type II errors (false interpretation of an absence of association). The problem itself does not come from the statistical method per se; the description of co-occurrence in the data will be right provided that the technique is adequate, but from the interpretation of the null model analysis.

Should we, therefore, abandon joint species distributions modeling (JSDM) and all of the information contained in co-distribution data? While our results might lead to such an interpretation, there is still some value in species co-occurrence data that could be used in distribution models. The appropriate use of JSDMs is to remove biases in the evaluation of species-specific relationship with the environment. Accounting for joint distribution will contribute to the evaluation of the conditional distribution of a species when all other species are absent. In other words, they should be used to improve the evaluation of the fundamental niche. The JSDMs will, however, fail to predict the right occurrence probability of a species for communities that have no analogue to the training dataset. JSDMs are using only the net associations between pairs of species and are not meant to recover the direct pairwise conditional co-occurrences. For instance, a JSDM evaluated for a plant, a herbivore, and a carnivore will provide the correct description of the joint distribution of all three species, but will be of limited use to predict the distribution of the plant and the herbivore if the carnivore disappears from the system. Further developments are, consequently, required to solve the issue and account for both direct and indirect interactions. One possible solution would be to constrain JSDMs with a prior expectation of the underlying structure of direct interactions.

It is also valuable to ask whether the inference of the structure of interaction networks is feasible from the observation of co-occurrences (as they result from many ecological processes). There is growing interest in inferring ecological network structure from alternative sources of information (Gravel et al. 2013; Morales-Castilla et al. 2015). This problem is challenging because of the multiple influences on co-occurrence. Our analysis of three species modules with conditional probabilities revealed it is feasible numerically, to obtain an estimate of all pairwise conditional probabilities when accounting for higher order interactions. Known quantities are the marginal probabilities and observed co-occurrence. The parameters to be evaluated are all fundamental conditional probabilities, representing the direct associations between pairs of species (the $p_{i|j,\overline{K}}$). This is a $S \times S$ problem to solve and thus requires a significant amount of data. It might, however, be solved with large datasets where the number of sites N is much larger than S. There might also be methods to reduce the dimension of the problem because usually only a small fraction of potential interactions are met in a network (corresponding to the connectance C). While a net interaction network (i.e., a network that takes direct and indirect interactions into account) is likely to be fully connected ($S \times S$ links), the direct interaction network has still only a fraction C of these links realized. Bayesian approach with latent variables could even further help reducing the dimension of the problem (e.g., Rohr et al. 2010; Ovaskainen et al. 2010). In such methods, latent variables are evaluated for

each species to represent the underlying structure of the ecological network. It was found that between two and four parameters per species would be required to successfully represent more than 80 % of interactions in a predator-prey network (Rohr et al. 2010). This approach could, therefore, be used to represent the underlying structure of direct interactions and to evaluate numerically the non-null conditional probabilities. Note that these pairwise direct interactions should be interpreted specifically with reference to spatial dynamics because they would still represent phenomenologically the consequences of interactions, not the mechanisms of interactions.

To apply our theory, we need occurrence data along with information on ecological interactions. Although such data require additional sampling efforts, they provide the adequate material to test the five principles we develop above. However, before doing so, we should expand the theory of species co-occurrence (and of species distribution) to include environmental constraints. Our approach assumed a homogeneous environment, mainly for tractability of equations. We acknowledge that non-independent co-occurrence could also arise because of shared environmental requirements. The addition of environmental constraints would be easy to implement in our framework by simply making the conditional probability in absence of interactions a function of the environment. Every quantity we derive thereafter would be conditional on the environment. What would be more challenging but, nonetheless, feasible numerically would be to make the direct interaction itself a function of the environment. There is now growing evidence that ecological interactions are context dependent (Chamberlain et al. 2014; Poisot et al. 2012). We view this integration as the next step to the derivation of a theory-driven species distribution model taking into account biotic interactions (Thuiller et al. 2013).

Acknowledgments This work was inspired by discussions with T. Poisot, D. Stouffer, A. Cyrtwill, and A. Rozenfield. Thanks to Matt Talluto and Isabelle Boulangeat for helpful comments on a previous version of the manuscript. We are also thankful to Lia Hemerik for providing advice that greatly improved the manuscript. Financial support was provided by the Canada Research Chair program and a NSERC-Discovery grant to D. Gravel. M. Araújo acknowledges support from Imperial Colege's Grand Challenges in Ecosystems and Environment Initiative.

References

Aráujo MB, Luoto M (2007) The importance of biotic interactions for modelling species distributions under climate change. Glob Ecol Biogeogr 16(6):743–753

Aráujo MB, Rozenfeld A (2014) The geographic scaling of biotic interactions. Ecography 37(5):406–415

Aráujo MB, Rozenfeld A, Rahbek C, Marquet PA (2011) Using species co-occurrence networks to assess the impacts of climate change. Ecography 34(6):897–908

Baselga A, Aráujo MB (2009) Individualistic vs community modelling of species distributions under climate change. Ecography 32(1):55–65

Bell G (2005) The co-distribution of species in relation to the neutral theory of community ecology. Ecology 86(7):1757–1770

Berlow EL, Dunne JA, Martinez ND, Stark PB, Williams RJ, Brose U (2009) Simple prediction of interaction strengths in complex food webs. Proc Natl Acad Sci U S A 106(1):187–191

Boulangeat I, Gravel D, Thuiller W (2012) Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. Ecol Lett 15(6):584–593

Chamberlain SA, Bronstein JL, Rudgers JA (2014) How context dependent are species interactions? Ecol Lett 17:881–890

Clark JS, Gelfand AE, Woodall CW, Zhu K (2014) More than the sum of the parts: forest climate response from joint species distribution models. Ecol Appl: A publication of the Ecological Society of America 24(5):990–999

Connor EF, Simberloff D (1979) The assembly of species communities: chance or competition? Ecology 60(6):1132

Davis AJ, Jenkinson LS, Lawton JH, Shorrocks B, Wood S (1998) Making mistakes when predicting shifts in species range in response to global warming. Nature 391(6669):783–786

Diamond JM (1975) Assembly of species communities. In: Ecology and evolution of communities, pp 342–444

Diamond JM, Gilpin ME (1982) Examination of the null model of connor and simberloff for species co-occurrences on islands. Oecologia 52(1):64–74

Fordham DA, Akakaya HR, Brook BW, Rodríguez A, Alves PC, Civantos E, Triviño M, Watts MJ, Aráujo MB (2013) Adapted conservation measures are required to save the iberian lynx in a changing climate. Nat Clim Chang 3(10):899–903

Gotelli NJ, Graves GR (1996) Null models in ecology, vol 14, p 368

Gotelli NJ, Graves GR, Rahbek C (2010) Macroecological signals of species interactions in the danish avifauna. Proc Natl Acad Sci U S A 107(11):5030–5035

Gravel D, Massol F, Canard E, Mouillot D, Mouquet N (2011) Trophic theory of island biogeography. Ecol Lett 14(10):1010–1016

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

Guisan A, Thuiller W (2005) Predicting species distribution: offering more than simple habitat models. Ecol Lett 8(9):993–1009

Hairston NG, Smith FE, Slobodkin LB (1960) Community structure, population control, and competition. Am Nat 94(879):421

Harris DJ (2015) Estimating species interactions from observational data with markov networks. Biorix preprint. doi:10.1101/018861

Holt RD, Barfield M (2009) Trophic interactions and range limits: the diverse roles of predation. Proceedings. Biol Sci/The Royal Society 276(1661):1435–1442

Hubbell SP (2001) The unified neutral theory of biodiversity and biogeography (MPB-32). Monographs in population biology. Princeton University Press, Princeton

Hutchinson GE (1957) Concluding remarks. Cold Spring Harb Symp Quant Biol 22:415–427

Jabot F, Bascompte J (2012) Bitrophic interactions shape biodiversity in space. Proc Natl Acad Sci U S A 109(12):4521–4526

Kissling WD, Dormann CF, Groeneveld J, Hickler T, Kühn I, McInerny GJ, Montoya JM, Römermann C, Schiffers K, Schurr FM, Singer A, Svenning J-C, Zimmermann NE, O'Hara RB (2012) Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. J Biogeogr 39:1–16

Legendre P, Legendre L (2012) Numerical ecology, 3rd edn. Elsevier, Amsterdam

MacArthur RH (1972) Geographical ecology, vol 54, chap 6. Princeton University Press, Princeton, pp 460–461

MacArthur RH, Wilson EO (1967) Theory of island biogeography. Princeton landmarks in biology, vol 1. Princeton University Press, Princeton, p 203

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. Science 298(5594):824–827

Montoya JM, Woodward G, Emmerson MC, Solé RV (2009) Press perturbations and indirect effects in real food webs. Ecology 90(9):2426–2433

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

Oksanen L, Fretwell SD, Arruda J, Niemela P (1981) Exploitation ecosystems in gradients of primary productivity. Am Nat 118(2):240

Ovaskainen O, Hottola J, Shtonen J (2010) Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. Ecology 91(9):2514–2521

Pellissier L, Rohr RP, Ndiribe C, Pradervand J-N, Salamin N, Guisan A, Wisz M (2013) Combining food web and species distribution models for improved community projections. Ecol Evol 3(13):4572–4583

Peterson AT (2011) Ecological niche conservatism: a time-structured review of evidence. J Biogeogr 38(5):817–827

Pielou DP, Pielou EC (1968) Association among species of infrequent occurrence: the insect and spider fauna of polyporus betulinus (bulliard) fries. J Theor Biol 21(2):202–216

Poisot T, Canard E, Mouillot D, Mouquet N, Gravel D, Jordan F (2012) The dissimilarity of species interaction networks. Ecol Lett 15(12):1353–1361

Pollock LJ, Tingley R, Morris WK, Golding N, O'Hara RB, Parris KM, Vesk PA, McCarthy MA (2014) Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (jsdm). Methods Ecol Evol 5(5):397–406

Rohr RP, Scherer H, Kehrli P, Mazza C, Bersier L-F (2010) Modeling food webs: exploring unexplained structure using latent traits. Am Nat 176(2):170–177

Stone L, Roberts A (1990) The checkerboard score and species distributions. Oecologia 85(1):74–79

Stouffer DB, Bascompte J (2010) Understanding food-web persistence from local to global scales. Ecol Lett 13(2):154–161

Stouffer DB, Camacho J, Jiang W, Amaral LAN (2007) Evidence for the existence of a robust pattern of prey selection in food webs. Proceedings. Biol Sci/The Royal Society 274(1621):1931–1940

Thuiller W, Mnkemller T, Lavergne S, Mouillot D, Mouquet N, Schiffers K, Gravel D (2013) A road map for integrating eco-evolutionary processes into biodiversity models. Ecol Lett 16:94–105

Ulrich W, Gotelli NJ (2013) Pattern detection in null model analysis. Oikos 122(1):2–18

Veech JA (2013) A probabilistic model for analysing species co-occurrence. Glob Ecol Biogeogr 22(2):252–260

Williams RJ, Martinez ND (2000) Simple rules yield complex food webs. Nature 404(6774):180–183

Wootton JT (1994) The nature and consequences of indirect effects in ecological communities. Annu Rev Ecol Syst 25(1):443–466

Yodzis P (1988) The indeterminacy of ecological interactions as perceived through perturbation experiments. Ecology 69(2):508–515