PLOS ONE

RESEARCH ARTICLE

Inferring interactions in complex microbial communities from nucleotide sequence data and environmental parameters

Yu Shang[1]*, Johannes Sikorski[1,6], Michael Bonkowski[2], Anna-Maria Fiore-Donno[2], Ellen Kandeler[3], Sven Marhan[3], Runa S. Boeddinghaus[3], Emily F. Solly[4], Marion Schrumpf[4], Ingo Schöning[4], Tesfaye Wubet[5,6], Francois Buscot[5,6], Jörg Overmann[1,6]

1 Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures, Inhoffenstraße 7B, D-38124, Braunschweig, Deutschland, **2** Department of Terrestrial Ecology, Institute of Zoology, University of Cologne, Zülpicher Straße 47b, D-50674 Köln, Deutschland, **3** Institute of Soil Science and Land Evaluation, Soil Biology Section (310b), University of Hohenheim, Emil-Wolff-Straße. 27, D-70593 Stuttgart, Deutschland, **4** Max-Planck-Institut für Biogeochemie, Hans-Knöll-Straße 10, D-07745 Jena, Deutschland, **5** Department of Soil Ecology, Helmholtz Centre for Environmental Research - UFZ, Theodor-Lieser-Straße 4, D-06120, Halle/Saale, Deutschland, **6** German Center for Integrative Biodiversity Research (iDiv) Jena Halle Leipzig, Deutscher Platz 5e, 04103 Leipzig, Deutschland

* yus14@dsmz.de

Citation: Shang Y, Sikorski J, Bonkowski M, Fiore-Donno A-M, Kandeler E, Marhan S, et al. (2017) Inferring interactions in complex microbial communities from nucleotide sequence data and environmental parameters. PLoS ONE 12(3): e0173765. https://doi.org/10.1371/journal.pone.0173765

Editor: Hauke Smidt, Wageningen University, NETHERLANDS

Received: May 30, 2016

Accepted: February 27, 2017

Published: March 13, 2017

Data Availability Statement: My code and test data are in the github website https://github.com/amssshangyu/interaction_analysis.

Funding: The work has been (partly) funded by the Deutsche Forschungsgemeinschaft (DFG) Priority Program 1374 "Infrastructure-Biodiversity-Exploratories" (Grants No. OV 20/21-1, 22-1 http://gepris.dfg.de/gepris/projekt/252263987). J.O. received the funding. Field work permits were issued by the responsible state environmental

Abstract

Interactions occur between two or more organisms affecting each other. Interactions are decisive for the ecology of the organisms. Without direct experimental evidence the analysis of interactions is difficult. Correlation analyses that are based on co-occurrences are often used to approximate interaction. Here, we present a new mathematical model to estimate the interaction strengths between taxa, based on changes in their relative abundances across environmental gradients.

Introduction

The composition of microbial communities is a key driver of ecological processes [1–3]. Changes in the abundances of species can occur in response to abiotic selection pressures, neutral assembly processes and be affected by organismal interactions. Biotic interactions are multifarious (competition, mutualism, commensalism, amensalism, and antagonism including parasitism and predation) and may have positive, negative, or neutral consequences for either one or both interacting partners. In most of the cases, outcomes of interactions between two species may be asymmetric in terms of abundance, e.g. predators will negatively affect the abundance of prey, but prey consumption will increase the abundance of the predator. As another example, a strong competitor will decrease the abundance of a neighboring species, but the latter may have no net effect on the former. Moreover, the strength of the influence can also differ, e.g. a preferred prey receives stronger top-down control than a general prey. Conversely, the prey could exert only a weak positive influence on the predator if it is of minor food quality.

offices of Baden-Württemberg, Thüringen, and Brandenburg (according to §72 BbgNatSchG).

Competing interests: The authors have declared that no competing interests exist.

While biodiversity studies of higher organisms have yielded large sets of multitrophic data and enabled a comprehensive analysis of interactions [4], the peculiarities of microbial communities render the detection and study of interactions very challenging. Whereas potential pairwise interactions among microorganisms can be studied directly in the laboratory, the determination of interactions in large and complex biotic communities under natural conditions is limited. Soil ecosystems in particular harbor extremely diverse microbial communities. The diversity of bacteria may reach 10^4 species [5] and average cell numbers of 10^{10} per gram of soil [6]. This results in highly complex networks of coexisting microbes [7]. Secondly, only a minority (0.1%–0.001%) of the microbial diversity has been cultivated to date [8], which precludes experimental analysis of the majority of interactions in the laboratory. Thirdly, the heterogeneous structure of the soil habitat at the microscale represents an additional methodological and conceptual challenge. Due to the complexity of soil structure, microbial cells typically occur in a non-random fashion in clusters with cell-to-cell distances of only 1-10 μm, which is the distance at which interactions between microbes is assumed to take place by either direct cell to cell contact or by diffusion limit of chemical substances. However, at this spatial scale, it is impossible to study organismal interactions simply by passive observations as in macroorganisms. In contrast, most studies of microbes in complex habitats typically are conducted through destructive sampling methodology, precluding the consecutive analysis of the same microbial community over time, which would be a prerequisite to applying the established discrete-time Lotka-Volterra models [9]. As a result, the lack of knowledge on species interactions has remained one of the major shortfalls in understanding the drivers of ecosystem functions [10]

A major step forward in analyzing the composition of microbial communities in the environment has been the advent of high-throughput next-generation sequencing technology for microbial DNA or RNA extracted from the environment. This enables the determination of the relative abundance of many taxa per sample and the analysis of co-occurrence and correlation patterns, which have been suggested as proxies for species interactions [7, 11]. However, the latter approach may only reflect similar responses of different species towards environmental pressures rather than direct interaction and cannot resolve interactions that are asymmetric. Also, several important properties of interactions can not be captured by correlations. Firstly, interactions may be asymmetric such that taxon A may influence B negatively but B influences A positively. In contrast, the correlation of A with B is the same as B with A, hence symmetric. Secondly, the direction of interaction may be different, e.g., A influences B positively, but under some circumstances, A may influence B negatively. In contrast, the concept of correlation analysis requires that the abundance relationship of two taxa remains consistent throughout: if A increases, B always either increases too (positive correlation) or decreases (negative correlation).

Fisher et al. [9] established the LIMITS algorithm to infer the interaction among microbial species from the time series data based on the discrete time Lotka-Volterra Model. Bucci et al. [12] also suggested utilizing the generalized Lotka-Volterra equation with time-dependent perturbation to analyse the interaction from the time series data. However, both of these approaches can not be used for series of cross-sectional data which originate from samples along gradients in environmental parameters but do not have a temporal dimension. Therefore, here, we present a novel method which is based on generalized Lotka-Volterra models and allows to quantify interactions from high-throughput microbial sequence data derived from cross-sectional samples for which also larger datasets on environmental parameters are available. This method does not rely on data obtained during different time points, since the strength and direction of interactions between partners are also a function of gradients in abiotic environmental parameters, for example of temperature [13], nutrient conditions [14] or

other factors such as soil moisture, mineral content, pH or osmolarity [15–17]. Biswas et al. [18] developed a Poisson-multivariate normal hierarchical model to analyse interactions, taking into account also the environmental gradients. In this work, the environmental parameters are included into the parameters of the Poisson function which is used to express the species abundance distribution. But, this calculation still represents a type of correlation analysis. Hence, the interaction matrix is symmetric and can not account for asymmetric interactions. On the contrary, our method is fundamentally different from correlation analysis and can analyse asymmetric interactions. Sugihara et al. [19] developed a causality test (CCM) to infer the causal link between the time-series variables which are not correlated. CCM analysed locally the historical time trajectory of the different variables and measured the extent to which the historical record of one variable can reliably estimate states of another variable. CCM can infer whether these variables are strongly linked to each other without consideration of environmental gradients. However, CCM can not account for the asymmetry property in the interaction relationship. In contrast, our approach can address this issue in the sense of changing environmental gradients but does not take into account the time series data. The basic assumptions of our model are that (i) interactions lead to changes of abundances within individual species, and that (ii) the abundances of a species are a function of the species abundances of the remaining community as well as of environmental parameters. We first describe the theoretical foundation of our framework and then suggest a numerical implementation calculating the direction and strength of interactions between two pairs of taxa, which depend on the quality criteria assigned to the determination of environmental parameter values and relative taxon abundances retrieved from high-throughput sequences. We further present methods to calculate single representative interaction values and to determine their robustness by appropriate statistical tools such as random sampling. Finally, we discuss the potentials of our approach.

Methods

The aim of our approach is to determine the interaction coefficient β_{ij}, which quantifies the interaction strength of species j on species i. We firstly present the theoretical basis for our approach. Secondly, depending on the precision quality of the input data, we suggest numerical calculation approaches estimating $\beta_{ij\alpha}^{k}$ for a given environmental parameter α and a given sample k. The result is a two-dimensional matrix of numerical $\beta_{ij\alpha}^{k}$ values with m (m is the number of environmental parameters) rows and N (N is the number of samples) columns for each pair of species j having interaction influence on species i. Thirdly, we addressed the issue about data structure and the precision of the calculation. We fourthly present several strategies to summarize the $\beta_{ij\alpha}^{k}$ values into a global β_{ij} value. Finally, we present an estimate for the robustness of β_{ij} based on random sampling and the addition of numerical noise to the input data.

Theoretical deductions

Species abundances change as a response to changing environmental conditions and abundances of interacting organisms. Therefore the information on the direction and strength of biotic interactions must be stored in the change of species abundances and hence can be extracted from that.

The basic idea of this methods is to analyze the influence from other interacting species abundance on the rate of change of specific species abundance in the sense of environmental gradients.

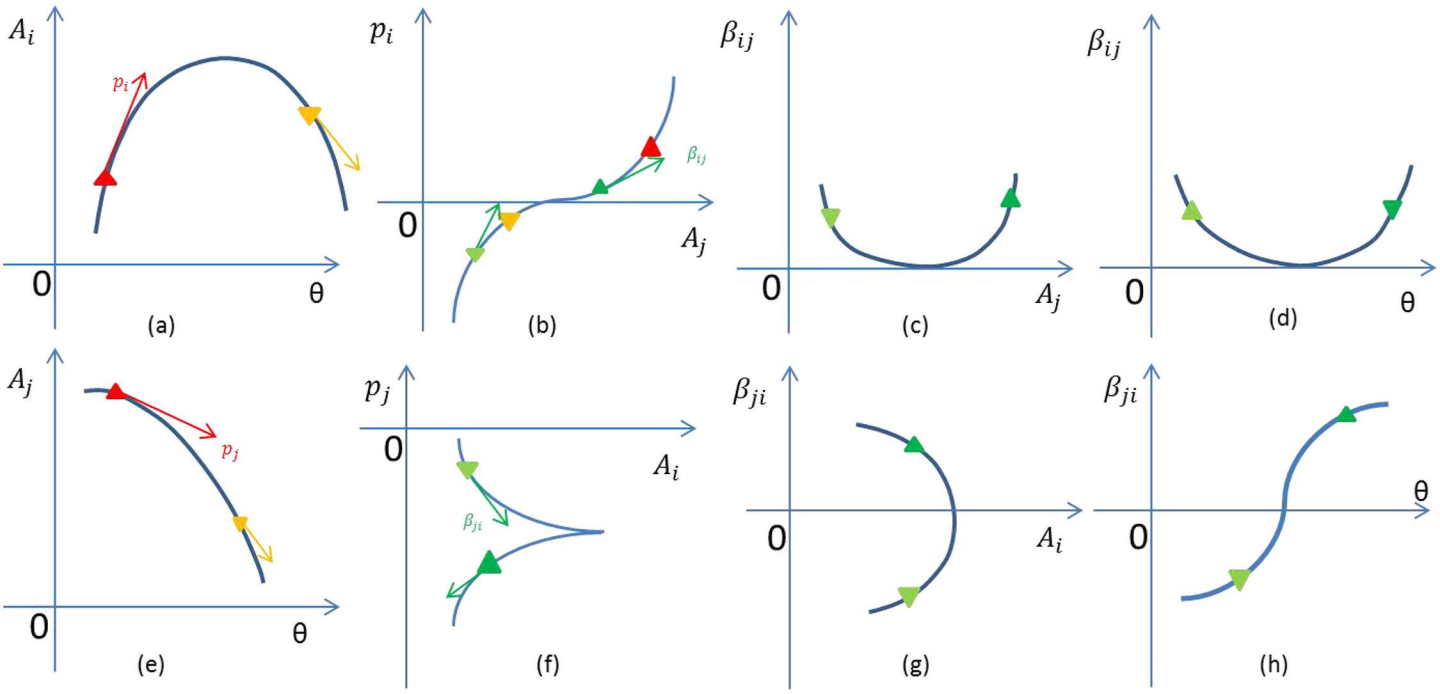

Fig 1. The sketch on the analysis workflow.

https://doi.org/10.1371/journal.pone.0173765.g001

The Fig 1 shows the detailed conceptual idea of this approach and the sketch on the analysis workflow of obtaining β_{ij} interaction values from values of relative abundances of taxa and environmental parameters obtained from a set of sampling sites. The change of species abundance A_i and A_j with respect to one environmental parameter θ are presented as a curve in subplots (a) and (e), respectively. In subplot(a), the rate of change p_i of A_i with respect to Θ was determined by calculating the slope of the tangent vector on each curve point. It reflects the influence of environmental parameter Θ on the change of abundance A_i across the environmental gradients. On the red triangular point, the slope is positive, means the rate change at this point (corresponding to one θ value) is positive, the abundance A_i has the increasing tendency at that gradient value of Θ. Similarly, the orange triangular point has a negative rate of change reflecting the decreasing tendency of A_i under these (larger) values of the gradient of Θ. Based on the calculated p_i for each values of θ and the abundance curve of A_j, the relationship between p_i and A_j is shown in subplot (b) (the red triangular has higher abundance values of A_j than the orange triangular). Then, the influence of A_j on the change of p_i can be denoted by the tangent vector (light green triangular and dark green triangular are the two example points). The slope of this new tangent vector stand for the rate of change of p_i with respect to A_j, and it is denoted as β_{ij}. β_{ij} is the interaction influence from species j on species i in the sense of θ. The relationship of β_{ij} and A_j is presented in subplot (c). Due to the relationship of A_j and θ in subplot(b), the relationship between β_{ij} and θ is shown in subplot (d). The subplot (d) inform us the change of interaction influence β_{ij} in different environmental conditions which are corresponding to the different value of θ. Similarly, we can do the same rate change analysis for A_j the corresponding results are shown in subplots (f), (g) and (h).

By analyzing the rate of change of species abundance, p_i, and the change of p_i with respect to other species, we can deduce the interaction influence between them. From this figure, β_{ij} is positive, suggesting that the species j has a positive interaction influence on species i. Conversely, β_{ji} is negative, suggesting that the species i has a negative interaction influence on species j. Although one pair of species shares the same interaction relationship, the effect of the interaction on both of them could be different, not only in the direction but also on the strength. Moreover, based on the subplot (a) and (e), there is no clear correlation between A_i and A_j. This indicates that the correlation is not equal to interaction, hence, the analysis of correlation between the species abundances is not suitable to infer the interaction relationship between them. The parameter β_{ij} is chosen in analogy to the Lotka-Volterra equation. The details are explained as follows.

Assume that the abundance A_i of species i is the smooth function of interacting species $A_j (j \neq i)$ and the environmental parameters Θ_α, $\alpha = 1, 2, \cdots, m$ (in case of soil, for example, soil moisture, pH, nutrient contents; where changes in time are available, even time t can be used). In a two species interaction system, the change in abundance of both species in response to the change of environmental parameters and biotic interactions are

$$\begin{aligned} \frac{dA_i}{d\Theta} &= S_i(A_i) + I_{ij}(A_i A_j) \\ \frac{dA_j}{d\Theta} &= S_j(A_j) + I_{ji}(A_j A_i) \end{aligned} \tag{1}$$

where S_i can be treated as the solitary part of species i, i.e. change of A_i independent of any influence from other species, it is also influenced by the environmental parameter. The derivative $\frac{dA_i}{d\Theta}$ is the rate of change of A_i with respect to the change of values of Θ. I_{ij} is the influence from species j on species i, which is a function of A_i, A_j, and also Θ. In different environmental conditions, I_{ij} will be different. Accordingly, the interaction can be analysed based on the gradient of Θ and will demonstrate how the interaction levels change across the different environmental conditions. Note that this can be asymmetric, i.e., $I_{ij} \neq I_{ji}$. Thus, the effects of the interaction between species i and j could be different with respect to the rate of change of A_i and of A_j. Note also that the exact mathematical form of S_i, S_j is often unknown due to lack of suitable experimental data, but can be approximated to follow the Monod equation or logistic equation [20–22]. To analyze the interaction, I_{ij} and I_{ji} need to be resolved. In order to calculate the change of $\frac{dA_i}{d\Theta}$ with respect to A_j we remove the unknown part S_i, as follows

$$\frac{d\left(\frac{A_i}{d\Theta}\right)}{dA_j} = \frac{dI_{ij}(A_i, A_j)}{dA_j} \tag{2}$$

The interaction information is contained on the right-hand side of the equation. For calculating the interaction numerically, we need the concrete mathematical expression of I_{ij}. For simplification, we assume:

$$I_{ij} = \beta_{ij} A_i A_j \tag{3}$$

Because A_i is a multivariate function of Θ_α, the rate of change of A_i with respect to Θ_α which is also a multivariate function of Θ_α can be expressed by using the partial derivative:

$$p_{i\alpha} = \frac{\partial A_i}{\partial \Theta_\alpha} \tag{4}$$

As the change of $p_{i\alpha}$ is affected by β_{ij}, the information of β_{ij} is stored in the change of $p_{i\alpha}$. With the approximation of Eq (3), β_{ij} can then be estimated as:

$$\beta_{ij\alpha} = \frac{\partial p_{i\alpha}}{\partial A_j} \frac{1}{A_i} \tag{5}$$

We define the interaction level β_{ij} as the rate of change of $p_{i\alpha}$ with respect to the abundance A_j of species j. Thus, the interaction level β_{ij} will be the smooth functions of species abundances A_i, A_j and the environmental gradients stored in Θ_α.

The above concept of using changes in species abundance for the calculation of interaction values is analogous to time-dependent generalized Lotka-Volterra equations (predator-prey equations):

$$\begin{aligned} \frac{dA_1}{dt} &= r_1 A_1 \left(1 - \frac{A_1}{K_1} + \beta_{12} \frac{A_2}{K_1}\right) \\ \frac{dA_2}{dt} &= r_2 A_2 \left(1 - \frac{A_2}{K_2} + \beta_{21} \frac{A_1}{K_2}\right) \end{aligned} \tag{6}$$

Here, the parameters r_1, r_2 are the growth rate, K_1, K_2 are the carrying capacity of the system [22]. Comparison of Eq (6) to Eq (1) demonstrates that: Θ is equivalent to the time parameter t, and $I_{12} = \frac{r_1}{K_1}\beta_{12}A_1A_2$. Incorporation of $\frac{r_1}{K_1}$ into β_{12} yields the:

$$I_{12} = \beta^*_{12} A_1 A_2 \tag{7}$$

By using Eq (5), we can estimate $\beta^*_{12} = \frac{d\left(\frac{A_1}{dt}\right)}{dA_2}\frac{1}{A_1}$, which represents the estimation of the interaction level from the species abundance change in the Lotka-Valterra equation.

Numerical determination of $\beta^k_{ij\alpha}$ values

In microbial ecology, absolute abundances of individual cells can usually not be determined for all taxa at all taxonomic hierarchy levels. With high-throughput sequence data, the abundance of a given taxon in sample k is actually given as a relative abundance value, which is the number of sequences reads assigned to that taxon among all sequence reads in the respective sample k. The determination of the relative abundance value of a specific taxon by high-throughput sequencing is not error-free. Small but uncontrollable variations in nucleic acid extractions, cDNA synthesis (in case RNA is extracted), amplicon primer ligation, and sequencing runs on high-throughput sequencers add uncertainty to the estimated relative abundance value. In the case of abundant taxa, typically at class or phylum level, the uncertainty may encompass just a 1% to 10% error level [23]. However, for less abundant taxa at the level of genera or species (defined by 97% similarity of the 16S rRNA gene [24]), the error could be much larger (two–fold, own unpublished data).

Similarly, the determination of physicochemical environmental parameters from soil such as pH, soil moisture, carbon and nitrogen content, is accompanied by uncertainty errors mostly due to soil heterogeneity which may also be in the range of 1% to 15% (own unpublished data).

We refer to data with assumed low (1-10%) experimental error in the estimation of numerical input data, and describe how to numerically calculate $\frac{\partial A_i}{\partial \Theta_\alpha}$ and $\frac{\partial p_{i\alpha}}{\partial A_j}$ from a data set derived from different samples using the Taylor expansion [25].

If the samples are denoted by using the index $k = 1, 2, \cdots N$, we denote A^k_i, Θ^k_α as the abundance of species A_i and environmental parameter Θ_α in sample k, respectively. The rate of

change of A_i with respect to Θ_α in sample k is defined as the partial derivative:

$$p_{i\alpha}^k = \frac{\partial A_i^k}{\partial \Theta_\alpha^k} \tag{8}$$

and the interaction level β_{ij} as the rate of change of $p_{i\alpha}^k$ with respect to species j abundance A_j^k according to

$$\beta_{ij\alpha}^k = \frac{\partial p_{i\alpha}^k}{\partial A_j^k} \frac{1}{A_i^k} \tag{9}$$

$\beta_{ij\alpha}^k$ represents the interaction value characterizing the interaction influence of species j on species i accompanying the change in environmental parameter Θ_α in sample k. This allows analyzing the interaction of species j on species i for different environmental parameters, and its change across different environmental conditions.

Note that for this part of the analysis, the numerical calculation of the partial derivative $\frac{\partial A_i^k}{\partial \Theta_\alpha^k}$ and $\frac{\partial p_{i\alpha}^k}{\partial A_j^k}$ normally requires to fix the values of the environmental parameters Θ_α to be the same in the other samples as in sample k. This mathematical requirement can not be fulfilled as real world samples differ typically at the same time in both species abundances and values of environmental parameters. We, therefore, make use of the Taylor expansion of multivariate functions to obtain the accurate numerical calculation when both environmental parameter values α and species abundances A change across samples k simultaneously. As addressed above, A_i^k is a multivariate function of environmental parameters $\{\Theta_\alpha\}$ and other interacting species A_j^k.

Between two different samples, the abundance of species i can be expressed as $A_i^k(\Theta^k)$ and $A_i^l(\Theta^l)$. Here, Θ^k and Θ^l are the corresponding environmental parameters in sample k and sample l. Using the Taylor expansion, the difference between $A_i^k(\Theta^k)$ and $A_i^l(\Theta^l)$ can be expressed as

$$\begin{aligned} A_i^l \quad = & \sum_{r_1=0}^{\infty} \cdots \sum_{r_d=0}^{\infty} \sum_{s_1=0}^{\infty} \cdots \sum_{s_m=0}^{\infty} \frac{(A_1^l - A_1^k)^{r_1} \cdots (A_d^l - A_d^k)^{r_d} (\Theta_1^l - \Theta_1^k)^{s_1} \cdots (\Theta_m^l - \Theta_m^k)^{s_m}}{r_1! \cdots r_d! s_1! \cdots s_m!} \\ & \cdot \left(\frac{\partial^{r_1+\cdots+r_d+s_1+\cdots+s_m} A_i^l}{\partial A_1^{r_1} \cdots \partial A_d^{r_d} \partial \Theta_1^{s_1} \cdots \partial \Theta_m^{s_m}} \right) (A_1^k, \cdots, A_d^k, \Theta_1^k, \cdots, \Theta_m^k) \end{aligned} \tag{10}$$

Using only the linear part of the approximation simplifies the formula as follows:

$$\begin{aligned} A_i^l(\Theta^l) - A_i^k(\Theta^k) &= \sum_\alpha (\Theta_\alpha^l - \Theta_\alpha^k) \frac{\partial A_i^k(\Theta)}{\partial \Theta_\alpha^k} + \sum_{j \neq i} (A_j^l - A_j^k) \frac{\partial A_i^k(\Theta)}{\partial A_j^k} \\ &= \sum_\alpha (\Theta_\alpha^l - \Theta_\alpha^k) p_{i\alpha}^k + \sum_{j \neq i} (A_j^l - A_j^k) p_{ij}^k \end{aligned} \tag{11}$$

The first part addresses the influence from the change of environmental parameters whereas the second part addresses the influence from the change of other species abundances. We fix the sample k, and let the sample l run across all remaining samples, in order to then estimate $p_{i\alpha}^k$ from linear regression. Actually, the term p_{ij}^k, which addresses the rate of change of A_i with respect to A_j in sample k, can also be estimated as the by-product.

Because p_{ij}^k is not necessary in the following $\beta_{ij\alpha}^k$ calculation, we reduce the model in Eq (11) to

$$\begin{aligned} A_i^l \quad &= \sum_{s_1=0}^{\infty} \cdots \sum_{s_m=0}^{\infty} \frac{(\Theta_1^l - \Theta_1^k)^{s_1} \cdots (\Theta_m^l - \Theta_m^k)^{s_m}}{s_1! \cdots s_m!} \\ &\cdot \left(\frac{\partial^{s_1+\cdots+s_m} A_i^l}{\partial \Theta_1^{s_1} \cdots \partial \Theta_m^{s_m}} \right) (\Theta_1^k, \cdots, \Theta_m^k) \end{aligned} \tag{12}$$

and then address the linear part of the approximation by

$$\begin{aligned} A_i^l(\Theta^l) - A_i^k(\Theta^k) \quad &= \quad \sum_{\alpha} (\Theta_\alpha^l - \Theta_\alpha^k) \frac{\partial A_i^k(\Theta)}{\partial \Theta_\alpha^k} \\ &= \quad \sum_{\alpha} (\Theta_\alpha^l - \Theta_\alpha^k) p_{i\alpha}^k \end{aligned} \tag{13}$$

in order to estimate $p_{i\alpha}^k$ from the linear regression using Eq (13).

After fixing the value of $p_{i\alpha}^k$, $\beta_{ij\alpha}^k$ is calculated using the same strategy. Because $p_{i\alpha}^k$ is the multivariate function of $A_j^k, j = 1, 2, \cdots, n$ and the environmental parameters Θ_α, the difference between $p_i^k(\Theta^k)$ and $p_i^l(\Theta^l)$ can be also expressed using the full version of Taylor expansion.

$$\begin{aligned} p_{i\alpha}^l \quad &= \sum_{r_1=0}^{\infty} \cdots \sum_{r_d=0}^{\infty} \sum_{s_1=0}^{\infty} \cdots \sum_{s_m=0}^{\infty} \frac{(A_1^l - A_1^k)^{r_1} \cdots (A_d^l - A_d^k)^{r_d} (\Theta_1^l - \Theta_1^k)^{s_1} \cdots (\Theta_m^l - \Theta_m^k)^{s_m}}{r_1! \cdots r_d! s_1! \cdots s_m!} \\ &\cdot \left(\frac{\partial^{r_1+\cdots+r_d+s_1+\cdots+s_m} p_{i\alpha}^l}{\partial A_1^{r_1} \cdots \partial A_d^{r_d} \partial \Theta_1^{s_1} \cdots \partial \Theta_m^{s_m}} \right) (A_1^k, \cdots, A_d^k, \Theta_1^k, \cdots, \Theta_m^k) \end{aligned} \tag{14}$$

Hence, analogous to Eq (11), the linear part to describe $\beta_{ij\alpha}^k$ based on two samples k and l is given by:

$$p_{i\alpha}^l - p_{i\alpha}^k \quad = \quad \sum_j (A_j^l - A_j^k) \frac{\partial p_{i\alpha}^k}{\partial A_j^k} + \sum_\alpha (\Theta_\alpha^l - \Theta_\alpha^k) \frac{\partial p_{i\alpha}^k}{\partial \Theta_\alpha^k} \tag{15}$$

here, the first part in Eq (14) addresses the change of other species abundances, whereas the second part addresses the change of environmental parameters. Analogous to Eq (13), the linear regression can be used to estimate $\frac{\partial p_{i\alpha}^k}{\partial A_j^k}$. Based on the Eq (9), the values of $\beta_{ij\alpha}^k$ are calculated as

$$\beta_{ij\alpha}^k = \frac{\partial p_{i\alpha}^k}{\partial A_j^k} \frac{1}{A_i^k} \tag{16}$$

The terms $\frac{\partial p_{i\alpha}^k}{\partial \Theta_\alpha^k}$ can be estimated also from Eq (15). These by-products address the influence from the environmental parameter change on the change of $p_{i\alpha}^k$. Although they do not provide information on the interaction between species i and j, they may provide valuable information of the interaction value $\beta_{ij\alpha}^k$.

In case that $\frac{\partial p_{i\alpha}^k}{\partial \Theta_\alpha^k}$ is of no interest, the model Eq (14) can be reduced to a simplified version by making use of only the first part in each Eqs (14) and (15) to then calculate $\beta_{ij\alpha}^k$ using Eq (16). This simplified version represents a different model of interaction calculation.

All the numerical calculations above are based on the linear part of the Taylor expansion. In order to cope with potential nonlinear properties of the data, it is also possible to include the higher order terms in the linear regression model. For example, when the second order terms are added into the Eq (13)

$$A_i^l(\Theta^l) - A_i^k(\Theta^k) = \sum_\alpha (\Theta_\alpha^l - \Theta_\alpha^k)\frac{\partial A_i^k(\Theta)}{\partial \Theta_\alpha^k} + \frac{1}{2}\sum_{\alpha\beta}(\Theta_\alpha^l - \Theta_\alpha^k)(\Theta_\beta^l - \Theta_\beta^k)\frac{\partial^2 A_i^k(\Theta)}{\partial \Theta_\alpha^k \partial \Theta_\beta^k} \quad (17)$$

the resulting Eq (17) will allow performing numerical calculations by additionally including the nonlinear part. However, the numerical calculations will substantially increase in complexity.

The models introduced allow different levels of precisions (Eqs (10)–(17)). The user can choose any of these based on the defined preferences, e.g., numerical precision of the input data or also the availability of computational power.

Data structure and precision of calculation

Data sparsity is the first issue which needs to be taken into account. In Eq (16), term A_i^k is in the denominator. If the species has not been observed and hence has the abundance value 0 in one sample, this species cannot be included in the mathematical treatment. Therefore, abundance values of 0 have to be removed before interaction analysis.

The second important issue is the data type of abundance A_i. For several organismic groups such as most plants and animals absolute abundance values A_i can be determined for each taxon. In contrast, a taxon-wise determination of absolute abundances is technically hardly feasible for bacterial microbes. Microbial communities are typically assessed by metagenomic high-throughput sequence data which yield relative abundances of taxa. However, since the overall cell numbers of microorganism in many cases do not change to a large extent, changes in absolute abundances are expected to be less pronounced than those in relative composition.

The structure of relative abundance data is characterized by an intrinsic compositional effect. This could produce misleading results in correlation analysis and would not reflect the true correlation as would have been the case for absolute abundance values [26–31]. As the sum of relative species abundance values by definition is constrained to 1, these values are not independent of each other. Fluctuations in the relative abundance of one species have an effect on the relative abundance of the rest of the community without that the rest of the community may actually have changed in absolute abundance. For example, if the abundance of a dominant species (e.g. 95% of all species) changes, relative abundances of all other species vary in the opposite direction. This creates artificial negative correlations with the dominant species which would not be the case with absolute abundances. Compositional effects may be severe in some data sets but mild in others. For example, compositional effects are most pronounced in communities with low species richness and/or pronounced dominance structure. The α diversity (of the samples in question is, therefore, a good predictor of the strength of compositional effects [28]. To decrease the influence of compositional effects, a series of methods based on the log-transformed techniques was developed [26–29].

Our interaction approach is not only conceptually but also mathematically fundamentally different from a correlation analysis. For example, whereas correlation aims to maximize the recovery of covariance $cov(A_i, A_j)$, our interaction approach aims to derive the correct partial derivative of abundance A_j with respect to the environmental parameters Θ and other species A_j. Thus, compositional effects have to be treated differently, as we show in detail below.

First, we discuss the relationship between the absolute abundance and the relative abundance, and the difference in results when we did the interaction analysis on both types of abundance data.

In the formalism of the interaction analysis, the terms $\frac{\partial A_i^k(\Theta)}{\partial \Theta_\alpha^k}$, $\frac{\partial A_i^k(\Theta)}{\partial A_j^k}$ play the very important parts in the whole calculation. For generality, we suppose y_i as the absolute abundance of species i, x_i as the relative abundance. We need to deduce the relationship between $\frac{\partial y_i}{\partial \Theta_\alpha}\frac{\partial y_i}{\partial y_j}$ and $\frac{\partial x_i}{\partial \Theta_\alpha}$ $\frac{\partial x_i}{\partial x_j}$.

The relationship between x_i and y_i is

$$x_i = \frac{y_i}{\sum_\alpha y_\alpha} \tag{18}$$

calculate the derivative on both sides of Eq (18), we have

$$\partial x_i = \frac{\partial y_i \sum_\alpha y_\alpha - y_i \sum_\alpha \partial y_\alpha}{\left(\sum_\alpha y_\alpha\right)^2} \tag{19}$$

Therefore, we have

$$\begin{aligned} \frac{\partial x_i}{\partial \Theta} &= \frac{\frac{\partial y_i}{\partial \Theta}\sum_\alpha y_\alpha - y_i \sum_\alpha \frac{\partial y_\alpha}{\partial \Theta}}{\left(\sum_\alpha y_\alpha\right)^2} \\ &= \frac{\frac{\partial y_i}{\partial \Theta}}{\sum_\alpha y_\alpha} - \frac{y_i \sum_\alpha \frac{\partial y_\alpha}{\partial \Theta}}{\left(\sum_\alpha y_\alpha\right)^2} \end{aligned} \tag{20}$$

$$\begin{aligned} \frac{\partial x_i}{\partial x_j} &= \frac{\partial y_i \sum_\alpha y_\alpha - y_i \sum_\alpha \partial y_\alpha}{\partial y_j \sum_\alpha y_\alpha - y_j \sum_\alpha \partial y_\alpha} \\ &= \frac{\left(\sum_\alpha y_\alpha - y_i\right)\frac{\partial y_i}{\partial y_j} - y_j \sum_{\alpha \neq i} \frac{\partial y_\alpha}{\partial y_j}}{\sum_\alpha y_\alpha - y_j \frac{\partial y_i}{\partial y_j} - y_j \sum_{\alpha \neq i} \frac{\partial y_\alpha}{\partial y_j}} \end{aligned} \tag{21}$$

When $\Sigma_\alpha y_\alpha >> y_\alpha$, diversity of the species is high. The term $\frac{y_i \sum_\alpha \frac{\partial y_\alpha}{\partial \Theta}}{\sum_\alpha y_\alpha}$ approaches zero. Eq (19) reduce to $\frac{\partial x_i}{\partial \Theta} \approx \frac{\frac{\partial y_i}{\partial \Theta}}{\sum_\alpha y_\alpha}$. This approximation reveals that the effect of the environmental parameter Θ on the rate of change of relative abundance is different from the corresponding rate of change of absolute abundance by a factor $\frac{1}{\sum_\alpha y_\alpha}$. This factor has a positive value, and will keep the sign of $\frac{\partial x_i}{\partial x_j}\frac{\partial y_i}{\partial y_j}$ the same. Similarly, Eq (21) reduce to $\frac{\partial x_i}{\partial x_j} \approx \frac{\partial y_i}{\partial y_j}$. This approximation means that the effect of the species j on the rate of change of relative abundance of species i is roughly the same as the corresponding rate of change of species absolute abundance of the species. Therefore, the interaction analysis based on the relative abundance can be roughly approximate to the corresponding analysis based on the absolute abundance. However, in the case of a lower species diversity, the upper approximation will not exist anymore, and the relation between the relative abundance and absolute abundance will become complex. Since in most cases microbial communities are highly diverse, our interaction analysis is expected to

yield reasonable results. Even for lower diversity, interaction analysis of relative abundance data can yield useful data for understanding ecosystem properties.

Actually, the compositional effect has its basis in the non-independence of the relative abundance. In order to decrease the effect of non-independence, a robust algorithm is needed for the numerical calculations. The precision and robustness of our numerical calculation depend on the linear regression estimates (Eqs (11)–(13), etc.). The least squares estimation (function stats::*lm*() in the R language) is not a robust algorithm since it is very sensitive to the initial input data. Therefore, we applied the more robust maximal likelihood estimation instead (R function MASS::*rlm*()). However, this algorithm has some requirements: input data should not have singularity, i.e. no linear relationship (collinearity) among the columns of the input data matrix [32, 33]. Therefore, the test for singularity on both relative abundance data and environmental parameters data needs to be performed before regression analysis. If the input data have collinearity, the algorithm will remove one species or one environmental parameter randomly, and repeat the test for singularity in the new data sets until all the collinearity relationships are removed. This pretreatment not only improves the robust numerical calculation but also decreases the compositional effects.

Another issue related to the precision of calculation is the relationship between the samples number and the number of variables. Sample number should be larger than the number of variables to avoid indeterminate equations or overfitting. Other suggestions to avoid overfitting which are not used in our methods are discussed in [34–36].

Summarizing $\beta^k_{ij\alpha}$ into the global interaction level β_{ij}

The interaction level $\beta^k_{ij\alpha}$ has four indexes i, j, α, k, which refer to a specific pair of taxa i, j, a specific environmental parameter α and a specific sample k. For each pair of species j with interaction influence on species i, there is a two-dimensional matrix of numerical $\beta^k_{ij\alpha}$ values with α rows and k columns. In either row or column, the values can be either positive, negative or zero. Positive values indicate a positive influence, negative values indicate a negative influence. Values may be non-normal distributed including extreme outlier values. To summarize these results into a more global interaction level value β_{ij} between species i and species j, we suggest the following different methods which can be chosen based on user preference.

Prior to any summarizing approach, users may decide to give different weight to $\beta^k_{ij\alpha}$ values for different environmental parameters, based on some prior knowledge about Θ_α. We estimate β^k_{ij} by performing a linear combination of $\beta^k_{ij\alpha}$ across all the environmental parameters:

$$\beta^k_{ij} = \sum_{\alpha} (C_\alpha \times \beta^k_{ij\alpha}) \tag{22}$$

where C_α is the applied weight of a given environmental parameter α. Prior knowledge on C_α can be obtained from, e.g. multivariate statistics. A Redundancy Analysis (RDA) allows determining those environmental parameters which significantly contribute to an observed community composition. The eigenvalue for each Θ_α in RDA analysis can be used as C_α weight. The derived weighted $\beta^k_{ij\alpha}$ values can be summarized in the same way as the original $\beta^k_{ij\alpha}$ values using the methods suggested below.

The most straightforward way is to summarize $\beta^k_{ij\alpha}$ estimates by standard summarizing statistics (mean, median, maximum, minimum). This approach retains the strength and direction of the interaction.

It is not recommendable, to sum up to $\beta_{ij\alpha}^{k}$ values, as positive and negative values could equal out each other resulting in a rather low β_{ij} value. In case the user is interested in a sum value, the standard norm definition could be applied. Note that this is possible only at the expense of losing information about the direction of interaction, as only positive values will be obtained. For example,

$$\beta_{ij}^{k} = \| (\beta_{ij\alpha}^{k}) \| = \sqrt{\sum_{\alpha} (\beta_{ij\alpha}^{k})^{2}} \tag{23}$$

represents a summed $\beta_{ij\alpha}^{k}$ interaction level between species i and species j across all environmental parameters α at sample k. Using the same formula as in Eq (23), several other quantities could be determined, e.g., β_{ij} would represent a summed β_{ij}^{k} across all samples k. Similarly, β_{i}^{k} represents the summed β_{ij}^{k} across all cases where $j \neq i$. Finally, β_{i} represents the summed β_{i}^{k}, which is the global influence of interaction from all the other species on species i across all the samples k.

Neither of the summarizing statistics addressed above captures the center of $\beta_{ij\alpha}^{k}$ values for a given environmental parameter α across all samples k appropriately and might therefore not yield the necessary insight into the interaction structure. A curve fitting approach including bootstrapping on $\beta_{ij\alpha}^{k}$ values with subsequent peak value extraction, as implemented in the eHOF R package [37] would be appropriate but could be computationally demanding with increasing taxa and environmental parameter numbers. As a compromise, we extract the median of those values which are represented in the peak from a $\beta_{ij\alpha}^{k}$ density distribution. The peak values, one per environmental parameter α for each species pair ij or ji and denoted therefore as M_{α}, could be summarized using the above standard summarizing statistics but would also suffer from the same shortcomings.

We, therefore, propose a custom approach which focuses on the dominant patterns of direction and strength of $\beta_{ij\alpha}^{k}$ values. The result will be a conservative estimate of direction and strength of β_{ij} values reflecting the dominant interactions between species i and species j. This procedure is based on two steps and may involve several user-based definitions of applied threshold values.

Firstly, the direction of interaction by categorizing $\beta_{ij\alpha}^{k}$ values is determined for each α across all samples k as being either positive or negative. In case the majority (we use 80%) of all $\beta_{ij\alpha}^{k}$ of a given α belongs to either category, the direction of interaction is classified either as positive or negative, respectively. In case that no preponderance can be identified, the given α does not contribute to a global β_{ij} determination and hence is ignored in the further analysis. The above peak determination approach yields a set of M_{α} values along with robust assignment of direction (either positive or negative).

Secondly, the set of M_{α} values per each taxon pair ij or ji is used to yield a global interaction value β_{ij} or β_{ji}, respectively. Note that M_{α} values are characterized by a direction (positive or negative) and by a certain strength (magnitude of the numerical value). Depending on the type of distribution of both direction and strength of value, two different ways for further evaluation can be taken into account. In case that the majority (we use 80%) of M_{α} values can be assigned to either direction, the respective M_{α} values are summarized by determining the median value, which represents then the global β_{ij} across all α parameter and k samples and is additionally characterized by a specific direction (positive or negative). Note that based on users interest, any other majority threshold value and summarizing statistic such as mean, minimum, or maximum can also be taken into account. However, in case that no

preponderance can be identified, a decision based on the above majority rule on direction can not be taken and will be replaced by a decision based on the magnitude of M_α values. We determine for each group of positive or negative M_α values the respective median values $M_{\alpha+}$ and $M_{\alpha-}$ values. In case the ratio of absolute values of $M_{\alpha+}$ or $M_{\alpha-}$ value is larger than two, the direction of the interaction is assumed to be represented by the larger M value (either $M_{\alpha+}$ or $M_{\alpha-}$), with the respective M value being the global β_{ij} across all α parameter and k samples. If $M_{\alpha+}$ and $M_{\alpha-}$ have comparable absolute values and also have equal proportions of either positive or negative direction, it is concluded that it is not possible to determine a global β_{ij} across all α parameter and k samples for the specific species pair of i and j.

In sum, we have suggested several workflows summarizing $\beta^k_{ij\alpha}$ values into a global interaction value β_{ij} that also contains the information on the direction (positive or negative interaction). Note that the choice of methods and choice of settings of several threshold values for a decision on intermediate steps is dependent on user preferences.

Robustness estimation of β_{ij} values

The magnitude and sign of the β_{ij} value depend on variables of the experimental data, such as variability of sampling site choice and the reliability (degree of precision) with which numerical values such as relative abundances of taxa or environmental parameter were determined [38]. We, therefore, implemented several methods to explore the robustness of the β_{ij} value (strength and direction) with respect to sampling site choice and numerical precision uncertainties in determining relative abundances and values of environmental parameter. Firstly, the effect of samples, which include values for both the environmental parameter and the relative abundances of taxa, is accessed by random sampling on soil samples. Secondly, we add numerical noise to either the original data of relative abundances or the environmental parameter values by randomly adding or subtracting error terms using formula

$$\tilde{\Theta}^k_\alpha = \Theta^k_\alpha(1+\epsilon) \tag{24}$$

where Θ^k_α is the original environmental parameters matrix, $\tilde{\Theta}^k_\alpha$ is the perturbed data matrix, and ϵ is the error term. Values for ϵ are generated as follows:

$$\epsilon = U(-1,1)e \tag{25}$$

where $U(-1, 1)$ is the uniform distribution in the range $[-1, 1]$ and e is the error level. The height of the error level, given as the proportion of the original value, e.g. 0.01% to 50%, can be defined by the user.

Similarly, random perturbations can be added to the numerical values of the species abundances. Values of β_{ij} obtained in repeated runs of data perturbation are then summarized using monovariate statistics (mean, 95% confidence interval, null hypothesis testing).

Application and results

We tested our method on datasets obtained from grassland soils of the German Biodiversity Exploratories (http://www.biodiversity-exploratories.de; [39]). The sampling plots were located in three regions in Germany: Schorfheide-Chorin (Schorfheide Exploratory; SE) in Brandenburg, national park Hainich-Dün (HE) in Thuringia, and biosphere reserve Schwäbische Alb (AE) in Baden-Württemberg. In every region, 50 grassland sites with different land-use intensities were investigated in the year 2011, resulting in a total of 150 plots analyzed in this study. At each plot aboveground plant parts in grasslands were removed before fourteen soil cores (diameter, 5 *cm*) were taken from the upper 15 *cm* of the A horizon from a 20×20 *m*

subarea. The 14 samples were combined, homogenized and 10*g* of the homogenized soil were frozen immediately in liquid nitrogen and stored until nucleic acid extraction for the determination of relative abundances of prokaryotic (RNA extraction) and fungal and protist (DNA extraction) communities by high-throughput sequencing.

The original test data encompasses 14 environmental parameters which were determined for each sieved (2 *mm*) soil sample. These are pH, soil moisture (%), nitrate (NO_3^-), ammonium (NH_4^+), mineral nitrogen (N), microbial C [all values as $\mu \cdot g\ \text{soil}^{-1}$], organic and inorganic carbon (C) [all values as $mg \cdot g\ \text{soil}^{-1}$], fine root biomass [$g \cdot cm^{-3}$], total N and C in roots (%), C/N ratio in soil, microbial C/N ratio, and C/N ratio in roots. The test for singularity (QR decomposition [40]) finds a rank of the environmental parameters matrix of 13 due to the collinearity between NO_3^- and NH_4^+. This means that either NO_3^- or NH_4^+ should be removed. Therefore, the interaction analysis is done with only a set of 13 environmental parameters after removing NH_4^+. Details on nucleic acid extractions, high-throughput sequencing, taxonomic classification, and determination of physicochemical soil parameters have been published elsewhere [41–45].

Calculations were performed for 17 taxonomic groups at the level of phylum or classes which represent abundant groups. The estimation of their relative abundances is generally of high precision (typically less than 5% deviation [23]). $\beta_{ij\alpha}^k$ and $\beta_{ji\alpha}^k$ of each pair of species *i* and *j* were determined for each soil sample *k* and for each environmental parameter α.

Overall, 150 samples, 17 taxonomic groups, and 13 environmental parameters were included in the data analysis. This data structure satisfies the requirement of sample size which is discussed in section. We scaled the species abundance and environmental parameters data (variance equals 1, uncentralized) to avoid the problem of large difference scale in different variables.

The degree of interaction changes with the gradient of the environmental conditions

The Fig 2 exemplifies the change of $\beta_{ij\alpha}^k$ estimates across the environmental gradient of three soil parameters for the interaction influence of acidobacterial subgroup Gp3 on acidobacterial subgroup Gp1. Y-axis range is the same for all three subfigures.

Whereas for pH and organic carbon a strong positive interaction is predicted, soil moisture suggests a weak negative impact. Notably, the strength of interaction varies along the gradient of the environmental parameter and increases substantially above a pH value of 6.5. In contrast, the interaction strength along gradients of organic carbon and soil moisture appears to be larger at rather low values of organic carbon and soil moisture respectively. In sum, $\beta_{ij\alpha}^k$ values may vary substantially across the gradient of environmental parameters.

Comparison between the global interaction matrix and the correlation matrix

The $\beta_{ij\alpha}^k$ and $\beta_{ji\alpha}^k$ values were then summarized by global β_{ij} and β_{ji} values, respectively, to estimate (a) the direction of interaction, which can be either positive or negative, and (b) the strength of interaction. In addition, the global interaction values were compared to results of standard co-occurrence analyses calculated by means of a Spearman rank correlation matrix C_{ij} based on relative abundances of the taxa. Both sets of results were also visualized as networks which displayed the dominant patterns (for values $-0.1 > \beta_{ij} > 0.1$; $-0.4 > \rho_{Spearman} > 0.4$) (Fig 3).

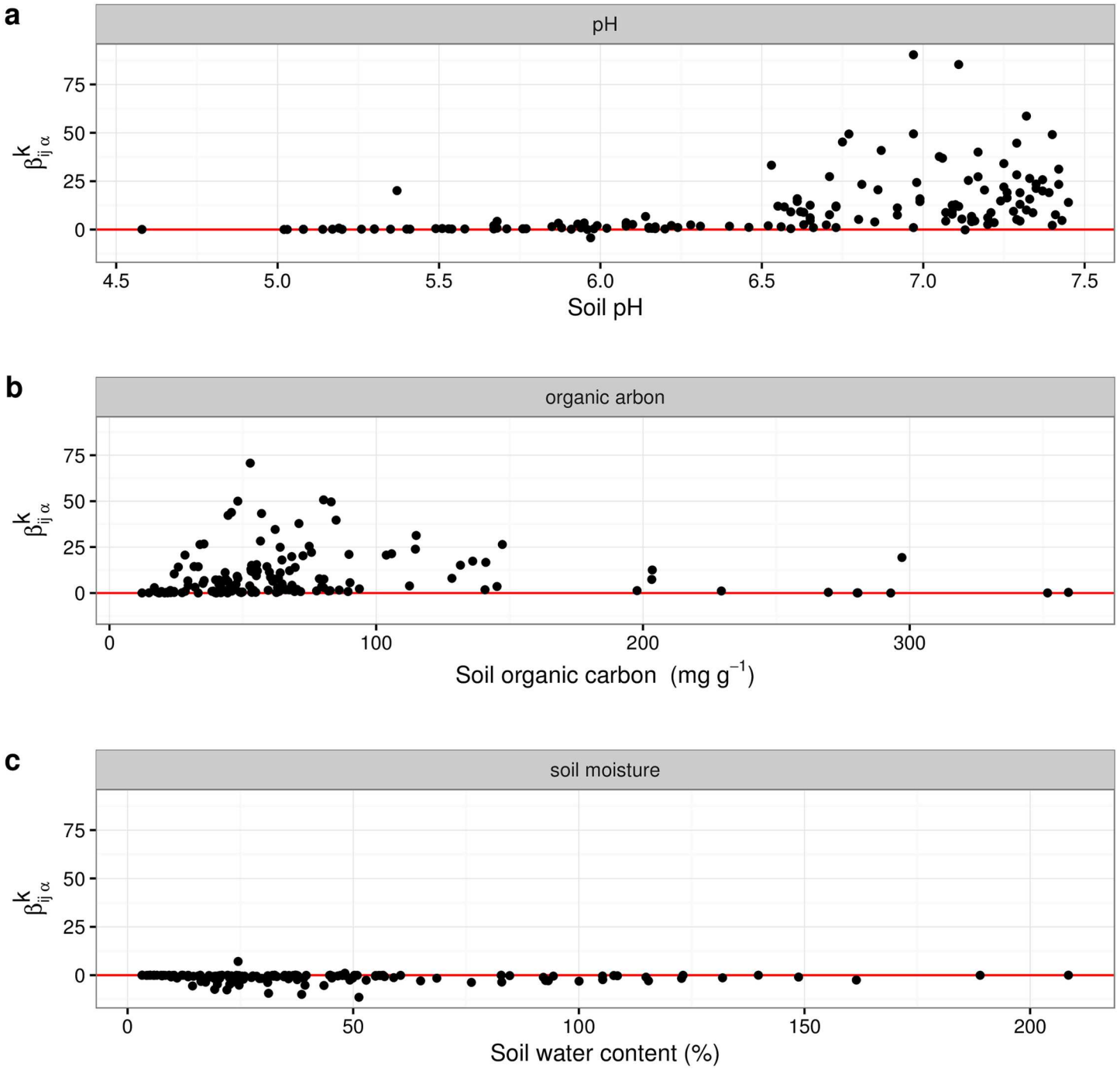

Fig 2. The distribution of $\beta^k_{ij\alpha}$ along the environmental gradient.

https://doi.org/10.1371/journal.pone.0173765.g002

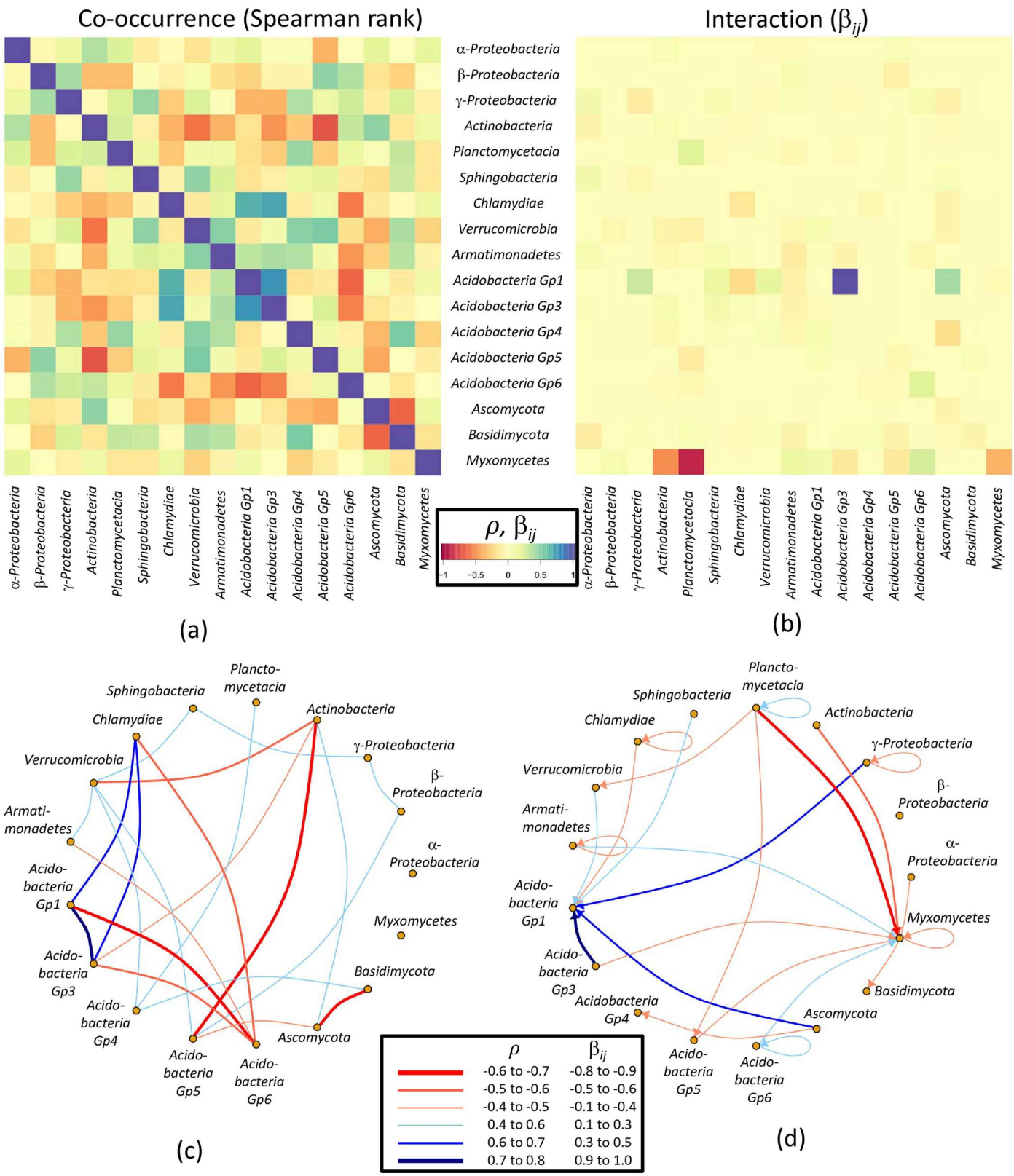

Fig 3. The comparison between the correlation and interaction analysis.

https://doi.org/10.1371/journal.pone.0173765.g003

Fig 3 provides the heatmaps (a,b) and network figures (c,d) of Spearman rank correlation matrix (a,c) and interaction matrix (b,d). Taxa depicted on the x-axis (columns) have interaction influence on taxa on the y-axis (rows). The red and blue colors indicate the negative and positive effects, respectively. For example, acidobacterial subgroup Gp3 has a high positive interaction influence on acidobacterial subgroup Gp1, whereas the bacterial *Planctomycetacia* have a strong negative interaction influence on protist group of *Myxomycetes*. Note that the heatmap color code is the same for both β_{ij} and Spearman ρ values. In the network figures, low Spearman ρ and low β_{ij} are not shown (but displayed in the heatmap), whereas the remaining values were artificially grouped into different categories (see color legend networks). Note that the interaction network displays also the direction of interaction by arrows, whereas correlation analysis does not enable any statement of directionality.

Important characteristics of interaction estimates and their difference to co-occurrence estimates are highlighted in Fig 3.

Firstly, taxa involved in multiple co-occurrences are not necessarily involved in corresponding interactions and vice versa. For example, acidobacterial subgroups Gp3, Gp5, Gp3 and also *Actinobacteria* share numerous co-occurrences with other taxa but are far less involved in interactions. Similarly, *Myxomycetes* and acidobacterial subgroup Gp1 both show numerous interactions to other taxa, but are less, if at all, involved in correlations.

Secondly, in case that two taxa are characterized by both strong correlation and interaction, it is not possible to predict from the type of correlation on the direction of interaction and vice versa. For example, both acidobacterial subgroup Gp3 and the *Chlamydiae* show strong positive correlations with each other and with acidobacterial subgroup Gp1. A strong positive interaction is observed only from Gp3 to Gp1, whereas the interactions of *Chlamydiae* on Gp1 are weakly negative and on Gp3 only very weak ($\beta_{ij} = 0.04$).

Thirdly, whereas co-occurrences within the same taxon are always positive at $\rho = 1$ (see heatmap, but not depicted in the network), overall interactions of taxa with itself can be both negative and positive. For example, *Myxomycetes* and *Chlamydiae* appear to have a negative interaction on themselves, whereas acidobacterial subgroup Gp6 and *Plancomycetacia* appear to have a positive influence on its own. Mathematically, this can be explained by analogy to the species self-effect in logistic equations, in which the rate of change of species abundance has also an influence in itself. In other words, this interaction value can be treated as the leading order of the solitary part in Eq (1). When the rate of change p_i decreases with the increase of its abundance A_i, the interaction influence from itself will be negative. In the converse situation, the influence will be positive. As a biological interpretation, taxa negatively interacting with each other (as implied here by negative β_{ij} values) have reached the carrying capacity within their ecological niche. Alternatively, these results could be a consequence of hierarchically nested taxa that are strongly interacting with each other, resulting in a cumulative positive or negative interaction of the higher level taxon on its self (e.g. *Chlamydia*).

Finally, it appears as if taxa are preferentially either exerting or experiencing interaction influence. For example, *Myxomycetes* share a lot of mostly negative interactions with other taxa, however, in all cases *Myxomycetes* are being influenced by others but are not exerting influence on others. Antibiotics production of bacteria could be a likely explanation [46]. The same is true for acidobacterial subgroup Gp1, which is, mostly positively, under interaction influence by other taxa. Only few taxa appear to both, experience as well as exert effects through influence (*Verrumicrobia*, *Acidobacteria* subgroup Gp5).

Estimation of robustness on the interaction influence calculation

In order to estimate the robustness of β_{ij} with respect to the numerical imprecision of the input data, we performed several perturbation assays. For this, we chose six examples of global β_{ij} interaction values from the Fig 3 which are representative of different strengths of interaction values with both a positive or a negative direction. Following the distribution of β_{ij} shown in the interaction heatmap of Fig 3, we tested larger, median, and low β_{ij} values for their robustness on data perturbations. The effect of variation in sample composition on β_{ij} is evaluated by 1000 iterations of randomly sampling 90% of the samples without replacement. We refrain from using the classical bootstrapping (sampling with replacement), as the deviation term (Eq (11)) will turn zero for twice or more of subsampled data and hence will be of no informative value in the downstream regression analysis (Eq (11)).

The effect of either numerical precision of environmental parameter values or relative abundances of taxa was evaluated by randomly adding or subtracting error terms (0.01%, 0.1%, 5%, 10%, 20% and 50%) to the original values. The effect of both numerical precision of environmental parameter values and relative abundances of taxa was evaluated by randomly adding or subtracting error terms (5%, 20%) to the original values.

Each 1000 iterations were performed for each error term and data type. We analyzed the data by means of comparison of 95% confidence interval, which provide information on effect sizes additionally to null hypothesis significance testing [47]. The robustness of β_{ij} estimations at different levels of data perturbation. are presented in Fig 4.

The plot (a) presents the distribution of β_{ij} as shown in the interaction heatmap in Fig 3. The plot (b) shows the robustness estimations on exemplary positive (upper row) and negative (lower row) β_{ij} values of decreasing strength (from left to right) taken from the interaction heatmap in Fig 3. The respective interactions from taxon *j* on *i* are listed as abbreviations in the panel header (Gp1 and Gp3: acidobacterial subgroups Gp1 and Gp3; Basid: *Basidiomycetes*; Myxomy: *Myxomycetes*; Planct: *Planctomycetacia*; gProt: *γ-Proteobacteria*; Sphing: *Sphingobacteria*). Only the strong interactions (left panels in (b) are depicted in the interaction network in Fig 3). The black horizontal line indicates the original β_{ij} values. Dots and vertical lines represent mean and 95% confidence interval bars from 1000 iterations of each type of data perturbation (see color legend). Horizontal dashed lines separate perturbations on environmental parameter values, relative taxon abundances, and sampling sites. Very small 95% CI values are are not visible as they are covered by the size of the point estimate dot (mean value of 1000 iterations).

Note, however, that in all cases where the 95% CI bar did not cross the zero line, p was < 0.01 in a two-sided one-sample t-test.

Typically, the 95% CIs are very small, suggesting the algorithm for numerical calculations to be robust. However, with increasing error level (from 0.01% to 10%) the 95% CIs become larger. This can be explained by the accumulated error in the numerical calculation and the nonlinear structure of the data. In our model, we use the linear part of the Taylor expansion as the approximation, and the numerical calculation is based on the linear regression. At small error level, the Taylor expansion can be reliably estimated by its linear part. At increasing error level, the potentially nonlinear structure of the data will become more relevant and therefore may generate increasing uncertainty in the estimation. Principally, this issue could be solved by extending the Taylor expansion to higher orders to take into account the nonlinear structure of the data.

The majority of β_{ij} values was very small for both positive and negative directions (plot a in Fig 4). This is the result of our conservative custom approach for β_{ij} summarizing, which is

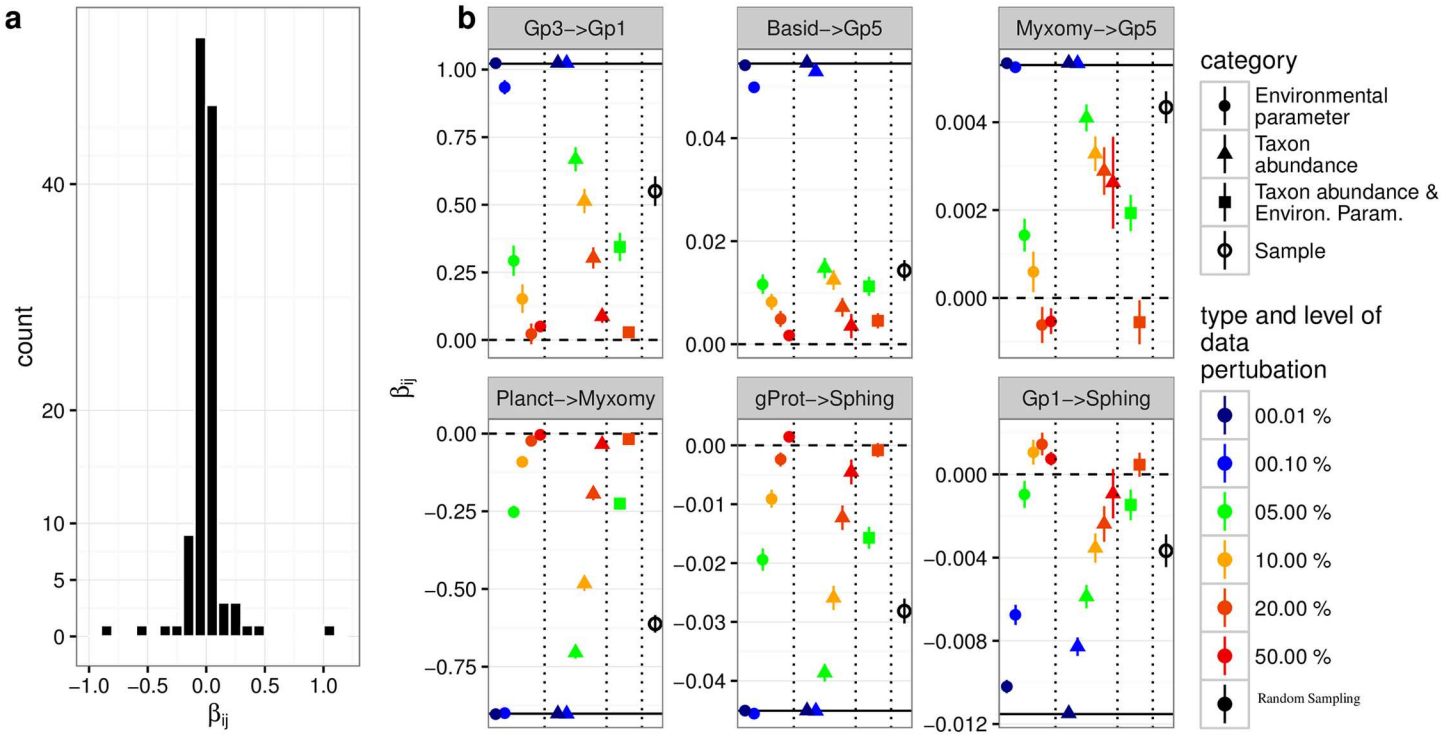

Fig 4. The test of robustness.

https://doi.org/10.1371/journal.pone.0173765.g004

based on the peak values of $\beta^k_{ij\alpha}$ density distributions and which is typically close to zero. However, Fig 2 indicates that individual $\beta^k_{ij\alpha}$ values can be considerably larger than 1.

There is a substantial effect of increasing error term size on the reduction of the original β_{ij}, which appears to be much larger than the effect on the increase of 95% CI intervals with increasing error term (figure b in Fig 3). This finding is independent of direction and strength of the original β_{ij} value and suggests that conclusions on direction and strength of interactions, especially in comparison of different pairs of taxa to each other, appear to be stable in the light of moderate error rates (up to 10%). The overall effect of error term size on data perturbations is larger for environmental parameter values than for values for the relative abundances of taxa. As a result, at larger error rates of environmental parameter values, the direction of interaction may change, suggesting that biological interpretation of very low β_{ij} should be treated with caution.

β_{ij} resulting from random sampling on soil samples are at comparable levels to β_{ij} resulting from 5% to 10% error term data perturbations on relative abundance values. Obviously, variation in the composition of samples does not change the estimates and the algorithm remains robust.

Discussion

Our first application of the methods developed in the present study to real-world data allowed us to identify several biotic interactions that are likely to shape soil microbial communities but were previously not recognized. Notably, the novel approach can be used to resolve the full

spectrum from intra-specific, over intra-phylum, to inter-domain interactions in complex microbial communities.

In our approach, environmental parameters which are not quantified in a study but which are nevertheless relevant for the interaction of species *i* on *j* will affect the results and hence the inference of biotic interactions. By comparison, lack of quantitative information on environmental parameters does not affect the results of co-occurrence patterns, where the correlation value of species *i* and *j* will stay the same irrespectively of any environmental parameter that has been determined or not. Yet, the relevance of unknown parameters that might control the correlation instead of direct biotic interactions will not be revealed by co-occurrence analysis. Increasing the number of organismic groups and environmental parameters in our type of analysis ultimately will require some data reduction approaches. The choice of which taxa and which environmental parameter should be included requires independent knowledge about their potential relevance. Such knowledge could be retrieved from, e.g., multivariate statistics.

The novel approach to estimate the strength and direction of biological interactions among taxa provides several advantages.

Firstly, our approach does not require repeated measurements at different time points. This is in contrast to approaches based on the discrete-time Lotka–Volterra model which requires concrete differential equation models [9, 48–50]. Often, these interactions are analysed within a predator-prey framework. However, the sampling of microbial communities is typically conducted in a destructive manner, which renders reproducible sampling of heterogeneous soil environment rather difficult or even impossible [43]. Instead, many soil microbial studies use cross-sectional format by taking multiple samples from different sites in parallel at the same time [51–53]. Our approach is far more general than the discrete-time Lotka–Volterra models employs derivations from the multivariate Taylor expansion function, and therefore allows to assess interactions using comparative cross-sectional datasets.

Secondly, our approach allows analysing interactions between two species *i* and *j* in the presence of other taxa *x*, *y*, *z* which may affect the interaction of species *i* on *j* [54, 55]. Analysing interactions between species i and j within a more complex community has already been addressed in discrete-time Lotka-Volterra models [9]. In our model, the user can deliberately remove species *x*, *y*, *z* to analyse the effect of their presence or absence on the interaction of species *i* on *j* or identify whether interactions appear stable despite varying community compositions. As an example, we observed that Acidobacteria subgroup Gp1 is apparently influenced by several other taxa (Fig 3). Using slightly different community compositions and slightly different sets of the environmental parameter, we observed that (i) Acidobacteria subgroup Gp1 remains being influenced by numerous other groups and that (ii) the observation of a strong positive interaction of Acidobacteria subgroup Gp3 on Gp1 remains (data not shown). The ecological function of Acidobacteria subgroup Gp1 is still largely unknown, but its involvement in several rather strong interactions suggests that they represent a keystone group of bacteria.

Thirdly, we can address interaction in the light of qualitatively and quantitatively different environmental parameters. Analogous to studying the effect of species composition, the user will be able to study the effect of a specific environmental parameter by removing it from the data set or by testing different combinations. More detailed analyses could be undertaken by determining β_{ij} separately for different ranges of environmental parameter values, e.g. at low versus high values of either pH, soil moisture, or land use intensities.

Fourthly, the results from our interaction calculations serve not only for biological interpretations but can be further used in statistical or modeling approaches. Quantities such as $(p_{i\alpha}^{k})$,

(β^k_{ijx}), (β^k_{ij}), (β_{ij}), (β^k_i) and (β_i) could be used in multivariate statistics or to construct dynamical equations within the framework of system stability analysis [56–58].

Finally, whereas in co-occurrence analysis it is not possible to distinguish between effects of true interaction and effects of similar response to environmental parameters without actually interacting, our approach enables to address separately the effects of biotic interactions and the abiotic response to the environmental parameter by using Eq (14).

At present, our model does not incorporate specific assumption about the mathematical expression of S_i, I_{ij} in Eq (1). In the future work, the Monod equation or logistic functions could be included to update our model to a concrete form. The numerical calculation strategies to do the interaction estimation would not necessarily be affected by this.

Based on the quantification of the strength of interaction and the prediction of its direction that is provided by our new approach, the underlying mechanisms of interaction will have to be determined by complementary experimental approaches. For example, a strong negative interaction could be exerted via direct predation [55], antibiotics [59], or other types of chemical warfare such as volatiles [60]. Here, strong β_{ij} that link taxa which previously were not under suspect to interact under natural conditions could serve as models for future investigations of the interaction mechanisms. This will require the availability of cultured isolates, however.

Supporting information

S1 File. R code and data sets. "Abundscale.Rdata" is the original species relative abundance data. "Parascale.Rdata" is the original environmental parameters data. Due to the long computation time on the original data, two shorten data sets "Abundscale_short.RData" and "Parascale_short.RData" are also provided for the fast test. "test_git.R" is the code for the analysis workflow on both of original data and the shorten data. "InteractionAnalysis_git.R" contains all the developed functions which are used in "test_git.R".
(7Z)

Acknowledgments

We thank the managers of the three Exploratories, Kirsten Reichel-Jung, Swen Renner, Katrin Hartwich, Sonja Gockel, Kerstin Wiesner, and Martin Gorke for their work in maintaining the plot and project infrastructure; Christiane Fischer and Simone Pfeiffer for giving support through the central office, Michael Owonibi for managing the central data base, and Markus Fischer, Eduard Linsenmair, Dominik Hessenmöller, Jens Nieschulze, Daniel Prati, Ernst-Detlef Schulze, Wolfgang W. Weisser and the late Elisabeth Kalko for their role in setting up the Biodiversity Exploratories project. We thank Doreen Berner for her assistance during sampling and laboratory analyses.

The work has been (partly) funded by the DFG Priority Program 1374 "Infrastructure-Biodiversity-Exploratories" (Grants No. OV 20/21-1, 22-1). Field work permits were issued by the responsible state environmental offices of Baden-Württemberg, Thüringen, and Brandenburg (according to §72 BbgNatSchG).

Author Contributions

Conceptualization: YS JS JO.

Data curation: YS.

Formal analysis: YS JS JO.

Funding acquisition: JO.

Investigation: MB AD EK SM RB ES MS IS TW FB JO.

Methodology: YS JS JO.

Project administration: JO.

Resources: YS JS MB AD EK SM RB ES MS IS TW FB JO.

Software: YS JS.

Supervision: JO.

Validation: YS JS JO.

Visualization: YS JS JO.

Writing – original draft: YS JS JO.

Writing – review & editing: YS JS JO.

References

1. Bodelier P, Meima-Franke M, Hordijk C, et al. Microbial minorities modulate methane consumption through niche partitioning. ISME. 2013; J7:2214–2228. https://doi.org/10.1038/ismej.2013.99 PMID: 23788331
2. Levine U, Teal T, Robertson G, et al. Agriculture's impact on microbial diversity and associated fluxes of carbon dioxide and methane. ISME. 2011; J5:1683–1691. https://doi.org/10.1038/ismej.2011.40 PMID: 21490688
3. Strickland M, Lauber C, Fierer N, et al. Testing the functional significance of microbial community composition. Ecology. 2009; 90:441–451. https://doi.org/10.1890/08-0296.1 PMID: 19323228
4. Scherber C, Eisenhauer N, Weisser WW, et al. Bottom-up effects of plant diversity on multitrophic interactions in a biodiversity experiment. Nature. 2010; 468(7323):553–556. https://doi.org/10.1038/nature09492 PMID: 20981010
5. Roesch LFW, Fulthorpe RR, Riva A, et al. Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J. 2007; 1(4):283–290. http://www.nature.com/ismej/journal/v1/n4/suppinfo/ismej200753s1.html https://doi.org/10.1038/ismej.2007.53 PMID: 18043639
6. Torsvik V, Ovreas L. Microbial diversity and function in soil: from genes to ecosystems. Current Opinion in Microbiology. 2002; 5(3):240–245. https://doi.org/10.1016/S1369-5274(02)00324-7 PMID: 12057676
7. Chaffron S, Rehrauer H, Pernthaler J, et al. A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Research. 2010; 7(9):947–959. https://doi.org/10.1101/gr.104521.109 PMID: 20458099
8. Overmann J. 7. In: Principles of enrichment, isolation, cultivation, and preservation of prokaryotes. Springer Berlin Heidelberg; 2013. p. 149–207. Available from: http://dx.doi.org/10.1007/978-3-642-30194-0_7
9. K FC, Pankaj M. Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Time-series Using Sparse Linear Regression. PLoS ONE. 2014; 9(7):e102451. https://doi.org/10.1371/journal.pone.0102451
10. Hortal J, Bello Fd, Diniz-Filho JAF, et al. Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. Annual Review of Ecology, Evolution, and Systematics. 2015; 46(1):523–549. https://doi.org/10.1146/annurev-ecolsys-112414-054400
11. Faust K, Raes J. Microbial interactions: from networks to models. Nat Rev Microbiol. 2012; 10(8):538–550. https://doi.org/10.1038/nrmicro2832 PMID: 22796884
12. Bucci V, Xavier JB. Towards Predictive Models of the Human Gut Microbiome. Journal of Molecular Biology. 2014; 426(23):3907–3916. https://doi.org/10.1016/j.jmb.2014.03.017 PMID: 24727124
13. Tveit AT, Urich T, Frenzel P, et al. Metabolic and trophic interactions modulate methane production by Arctic peat microbiota in response to warming. Proceedings of the National Academy of Sciences. 2015; 112(19):E2507–E2516. https://doi.org/10.1073/pnas.1420797112
14. Kuzyakov Y, Blagodatskaya E. Microbial hotspots and hot moments in soil: Concept and review. Soil Biology and Biochemistry. 2015; 83(0):184–199. https://doi.org/10.1016/j.soilbio.2015.01.025

15. Nadell CD, Xavier JB, Levin SA, et al. The evolution of quorum sensing in bacterial biofilms. PLoS Biology. 2008; 6(1):e14. https://doi.org/10.1371/journal.pbio.0060014 PMID: 18232735
16. West SA, Griffin AS, Gardner A, et al. Social evolution theory for microorganisms. Nature Reviews Microbiology. 2006; 4(8):597–607. https://doi.org/10.1038/nrmicro1461 PMID: 16845430
17. Williams P, Winzer K, Chan W, et al. Look who's talking: communication and quorum sensing in the bacterial world. Philos Trans R Soc London Ser B. 2007; 362(1483):1119–1134. https://doi.org/10.1098/rstb.2007.2039 PMID: 17360280
18. Biswas S, McDonald M, Lundberg DS, et al. In: Learning Microbial Interaction Networks from Metagenomic Count Data. Cham: Springer International Publishing; 2015. p. 32–43. Available from: http://dx.doi.org/10.1007/978-3-319-16706-0_6
19. Sugihara G, May R, Ye H, Hsieh Ch, Deyle E, Fogarty M, et al. Detecting Causality in Complex Ecosystems. Science. 2012; 338(6106):496–500. https://doi.org/10.1126/science.1227079 PMID: 22997134
20. Merchuk JC, Asenjo JA. The Monod equation and mass transfer. Biotechnology and Bioengineering. 1995; 45(1):91–94. https://doi.org/10.1002/bit.260450113 PMID: 18623056
21. Monod J. The Growth of Bacterial Cultures. Annual Review of Microbiology. 1949; 3(1):371–394. https://doi.org/10.1146/annurev.mi.03.100149.002103
22. Murray JD. Mathematical Biology: I. An introduction, Third Edition. Springer; 2002.
23. Uroz S, Buee M, Murat C, et al. Pyrosequencing reveals a contrasted bacterial diversity between oak rhizosphere and surrounding soil. Environ Microbiol Rep. 2010; 2(2):281–288. https://doi.org/10.1111/j.1758-2229.2009.00117.x PMID: 23766079
24. Rossello-Mora R, Amann R. The species concept for prokaryotes. FEMS Microbiol Rev. 2001; 25:39–67. https://doi.org/10.1111/j.1574-6976.2001.tb00571.x PMID: 11152940
25. Hazewinkel M. "Calculus", Encyclopedia of Mathematics. Springer; 2001.
26. Ban Y, An L, Jiang H. Investigating microbial co-occurrence patterns based on metagenomic compositional data. Bioinformatics. 2015; 31(20). https://doi.org/10.1093/bioinformatics/btv364 PMID: 26079350
27. Fang H, Huang C, Zhao H, Deng M. CCLasso: Correlation Inference for Compositional Data through Lasso. Bioinformatics. 2015; https://doi.org/10.1093/bioinformatics/btv349
28. Friedman J, Alm EJ. Inferring Correlation Networks from Genomic Survey Data. PLoS Comput Biol. 2012; 8(9):1–11. https://doi.org/10.1371/journal.pcbi.1002687
29. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. PLoS Comput Biol. 2015; 11(5):1–25. https://doi.org/10.1371/journal.pcbi.1004226 PMID: 25950956
30. Pearson K. On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proc R Soc London. 1897; 60:489–502. https://doi.org/10.1098/rspl.1896.0076
31. Weiss S, Treuren VW, Lozupone C, et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. The ISME Journal. 2016; 10:1669–1981. https://doi.org/10.1038/ismej.2015.235 PMID: 26905627
32. Ellis SP. Singularity and outliers in linear regression with application to least squares, least squares linear regression. Metron—International Journal of Statistics. 2000; LVIII(1-2):121–129.
33. Huber PJ, Ronchetti EM. Robust Statistics. Wiley; 2009.
34. Draper NR. Applied regression analysis. New York: Wiley; 1998.
35. Everitt BS. Cambridge Dictionary of Statistics. Cambridge University Press; 2002.
36. Hawkins DM. The problem of overfitting. Journal of chemical information and computer sciences. 2004; 44(1):1–12. https://doi.org/10.1021/ci0342472 PMID: 14741005
37. Jansen F, Oksanen J. How to model species responses along ecological gradients–Huisman–Olff–Fresco models revisited. Journal of Vegetation Science. 2013; 24(6):1108–1117. https://doi.org/10.1111/jvs.12050
38. Nayfach S, Pollard KS. Toward Accurate and Quantitative Comparative Metagenomicss. Cell. 2016; 166(5):1103–1116. https://doi.org/10.1016/j.cell.2016.08.007 PMID: 27565341
39. Fischer M, Bossdorf O, Gockel S, et al. Implementing large-scale and long-term functional biodiversity research: The Biodiversity Exploratories. Basic and Applied Ecology. 2010; 11(6):473–485. https://doi.org/10.1016/j.baae.2010.07.009
40. Golub GH, Van Loan CF. Matrix Computations (3rd Ed.). Baltimore, MD, USA: Johns Hopkins University Press; 1996.
41. Fiore-Donno AM, Weinert J, Wubet T, et al. Metacommunity analysis of amoeboid protists in grassland soils. Scientific Reports. 2016; 6:19068. https://doi.org/10.1038/srep19068 PMID: 26750872

42. Goldmann K, Schöning I, Buscot F, et al. Forest management type influences diversity and community composition of soil fungi across temperate forest ecosystems. Frontiers in Microbiology. 2015; 6. https://doi.org/10.3389/fmicb.2015.01300 PMID: 26635766

43. Regan KM, Nunan N, Boeddinghaus RS, et al. Seasonal controls on grassland microbial biogeography: Are they governed by plants, abiotic properties or both? Soil Biology and Biochemistry. 2014; 71(0):21–30. https://doi.org/10.1016/j.soilbio.2013.12.024

44. Solly EF, Schöning I, Boch S, et al. Factors controlling decomposition rates of fine root litter in temperate forests and grasslands. Plant and Soil. 2014; 382(1):203–218. https://doi.org/10.1007/s11104-014-2151-4

45. Wüst PK, Nacke H, Kaiser K, et al. Estimates of the bacterial ribosome content and diversity in soils are significantly affected by different nucleic acid extraction methods. Applied and Environmental Microbiology. 2016;

46. Jousset A, Scheu S, Bonkowski M. Secondary metabolite production facilitates establishment of rhizobacteria by reducing both protozoan predation and the competitive effects of indigenous bacteria. Functional Ecology. 2008; 22. https://doi.org/10.1111/j.1365-2435.2008.01411.x

47. Halsey LG, Curran-Everett D, Vowler SL, et al. The fickle P value generates irreproducible results. Nat Meth. 2015; 12:179–185. https://doi.org/10.1038/nmeth.3288 PMID: 25719825

48. Ives AR, Dennis B, Cottingham KL, et al. ESTIMATING COMMUNITY STABILITY AND ECOLOGICAL INTERACTIONS FROM TIME-SERIES DATA. Ecological Monographs. 2003; 73. https://doi.org/10.1890/0012-9615(2003)073%5B0301:ECSAEI%5D2.0.CO;2

49. J HD. Estimating species interactions from observational data with Markov networks. bioRxiv. 2015;

50. Power DA, Watson RA, Szathmáry E, et al. What can ecosystems learn? Expanding evolutionary ecology with learning theory. Biology direct. 2015; 10. https://doi.org/10.1186/s13062-015-0094-1 PMID: 26643685

51. Crowther TW, Maynard DS, Leff JW, et al. Predicting the responsiveness of soil biodiversity to deforestation: a cross-biome study. Global Change Biology. 2014; 20(9):2983–2994. https://doi.org/10.1111/gcb.12565 PMID: 24692253

52. Fierer N, Leff JW, Adams BJ, et al. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proc Natl Acad Sci U S A. 2012; 109(52):21390–21395. https://doi.org/10.1073/pnas.1215210110 PMID: 23236140

53. Lozupone CA, Knight R. Global patterns in bacterial diversity. Proceedings of the National Academy of Sciences of the United States of America. 2007; 104(27):11436–11440. https://doi.org/10.1073/pnas.0611525104 PMID: 17592124

54. Stenseth NC, Falck W, Bjørnstad ON, et al. Population regulation in snowshoe hare and Canadian lynx: Asymmetric food web configurations between hare and lynx. Proceedings of the National Academy of Sciences. 1997; 94(10):5147–5152. https://doi.org/10.1073/pnas.94.10.5147

55. Hoppener-Ogawa S, Leveau JHJ, van Veen JA, et al. Mycophagous growth of Collimonas bacteria in natural soils, impact on fungal biomass turnover and interactions with mycophagous Trichoderma fungi. ISME J. 2008; 3(2):190–198. http://www.nature.com/ismej/journal/v3/n2/suppinfo/ismej200897s1.html https://doi.org/10.1038/ismej.2008.97 PMID: 18923455

56. Legendre P, Legendre L. Numerical Ecology. ELSEVIER; 2012.

57. Daniel B, Francois G, Legendre P. Numerical Ecology with R. Springer; 2011.

58. Coyte KZ, Schluter J, Foster KR. The ecology of the microbiome: Networks, competition, and stability. Science. 2015; 350. https://doi.org/10.1126/science.aad2602 PMID: 26542567

59. Ling LL, Schneider T, Peoples AJ, et al. A new antibiotic kills pathogens without detectable resistance. Nature. 2015; 517:455–459. https://doi.org/10.1038/nature14098 PMID: 25561178

60. Schmidt R, Cordovez V, de Boer W, et al. Volatile affairs in microbial interactions. ISME J. 2015; 9(11):2329–2335. https://doi.org/10.1038/ismej.2015.42 PMID: 26023873