Machine learning of microbial interactions using Abductive ILP and Hypothesis Frequency/Compression Estimation

Didac Barroso-Bergada[1], Alireza Tamaddoni-Nezhad[2], Stephen H. Muggleton[3], Corinne Vacher[4], Nika Galic[5], and David A. Bohan[1]

[1] Agroécologie, AgroSup Dijon, INRAE, Université de Bourgogne Franche-Comté, Dijon, France
[2] University of Surrey, Guildford GU2 7XH, UK
[3] Imperial College London, South Kensington, London SW7 2AZ, UK
[4] INRAE, Univ. Bordeaux, BIOGECO, Pessac, France
[5] Syngenta Crop Protection LLC, Greensboro, NC, USA

Abstract. Interaction between species in microbial communities plays an important role in the functioning of all ecosystems, from cropland soils to human gut microbiota. Many statistical approaches have been proposed to infer these interactions from microbial abundance information. However, these statistical approaches have no general mechanisms for incorporating existing ecological knowledge in the inference process. We propose an Abductive / Inductive Logic Programming (A/ILP) framework to infer microbial interactions from microbial abundance data, by including logical descriptions of different types of interaction as background knowledge in the learning. This framework also includes a new mechanism for estimating the probability of each interaction based on the frequency and compression of hypotheses computed during the abduction process. This is then used to identify real interactions using a bootstrapping, re-sampling procedure. We evaluate our proposed framework on simulated data previously used to benchmark statistical interaction inference tools. Our approach has comparable accuracy to SparCC, which is one of the state-of-the-art statistical interaction inference algorithms, but with the the advantage of including ecological background knowledge. Our proposed framework opens up the opportunity of inferring ecological interaction information from diverse ecosystems that currently cannot be studied using other methods.

Keywords: Abductive/Inductive Logic Programming (A/ILP) · Interaction Network Inference · Machine learning of ecological networks · Hypothesis Frequency Estimation (HFE)

1 Introduction

Networks of interactions between species of microbes are believed to drive many of the biological functions that determine effects as diverse as soil health, crop

growth, and plant and human disease. Next generation sequencing of DNA samples taken from microbial communities can produce lists of those species present and metrics for their abundance, by treating the number of each sequence type in the sample either as absolute or relative counts. Inferring networks from these data could yield important results, improving our ability to manage these systems and issues [19]. For example, learning interactions of competition or predation of a disease-causing microbial agent could be used to identify species for biological control, and the chemistry that is involved could lead to the development of new drugs [10]. Current approaches to reconstructing ecological networks of interaction between microbial species use statistical learning to infer the presence of an interaction via correlation. Human experts subsequently interpret whether the correlation indicates an interaction between the two correlated microbial species, such as competition or predation.

Abductive/Inductive Logic Programming (A/ILP) was previously used to automatically generate plausible and testable food webs from ecological census data [17]. The approach in Tamaddoni-Nezhad et al. (2012) [17] also included a probabilistic approach, called Hypothesis Frequency Estimation (HFE) for estimating probabilities of hypothetical trophic links based on their frequency of occurrence when randomly sampling the hypothesis space. Through a review of the literature, it was found that many of the learned trophic links are corroborated by the literature. In particular, links ascribed with high probability by machine learning are shown to correspond well with those having multiple references in the literature. In some cases novel, high probability links were suggested, some of which were subsequently tested and confirmed in empirical studies [18].

In this paper we extend the A/ILP and HFE approaches in Tamaddoni-Nezhad et al. (2012) [17] for the purpose of learning microbial interactions. We will describe the existing context on interaction inference, detail the A/ILP based inference method and evaluate this method with a benchmark dataset. We also compare our results with SparCC, which is a state-of-the-art statistical interaction inference algorithm.

2 Background and related work

Microbial ecologists have clear criteria for interactions between species that can readily be transcribed into logical statements. In effect, past or ongoing interactions between two microbial species will have led to changes in the abundance of one or both species. Conceptually, therefore, two species might have undergone or might be undergoing an interaction if there is some pattern to the changes of the two species across a data-set. Thus, if one of the species always increases or decreases in abundance in the presence of the other, microbial ecologists might hypothesize an interaction between the two species. The ecological mechanisms of these interactions, along with their expected changes in abundance of the two species, have previously been described in Derocles et al. (2018) [3] as shown in Table 1.

Table 1. Type of interactions in function of the changes in abundance [3].

Type of interaction	**Effect on Specie1 abundance**	**Effect on Specie2 abundance**	**Nature of interaction**
Mutualism	Up	Up	Mutual benefits of the species
Competition	Down	Down	Species have negative effect on each other
Predation/Parasitism	Up	Down	Parasite develops at the expense of the host
Commensalism	Up	Null	Specie1 benefits while Specie2 is not affected
Amensalism	Down	Null	Specie2 has a negative effect on Specie1, but Specie2 is not affected

In this paper we extend the A/ILP approach in Tamaddoni-Nezhad et al. (2012) [17] with logical statements for putative microbial interactions included as background knowledge, to infer ecological interactions directly, with less or even without the intervention of humans at the interpretation step. This direct approach would be particularly valuable for reconstructing microbial networks in previously unstudied ecosystems where human knowledge for interpretation may effectively be non-existent (Figure 1).

In this paper we also extend the Hypothesis Frequency Estimation (HFE) approach introduced in Tamaddoni-Nezhad et al. (2012) [17]. Microbial ecologists rely on statistical probability estimates, typically at the conventional 5% significance level, to evaluate the importance of a correlational link between any two microbial species [15]. Most ILP approaches, including HFE rely on coverage based measures such as 'compression' for selecting hypotheses.

The problem we try to address here is whether compression can be evaluated within a statistical framework that meets the needs of microbial ecologists, to a degree that might be sufficient to convince them of the statistical importance and veracity of any learned interaction. In particular, we explore an extension of HFE where both the frequency and the compression of the hypotheses are considered within an statistical framework.

Benchmarking statistical learning approaches for inferring correlational links have used simulated data-sets. For example, Weiss et al. (2016) [21] produced simulated microbial data-sets to benchmark the ability of different statistical methods, such as SparCC [7] and CoNet [6], to detect different interaction types via correlation. In this paper we use the method of Weiss et al. [21] to simulate ecological-like replicated data-sets of interactions, of given interaction strengths. We then use ILP to evaluate the presence of the simulated interactions, as a known set of expectations. Our specific goals are to: determine the most sensitive parameter of compression for recovering an interaction, given a discrete number of permutations; and, evaluate the probabilistic significance of the compression parameter using a form of bootstrapping.

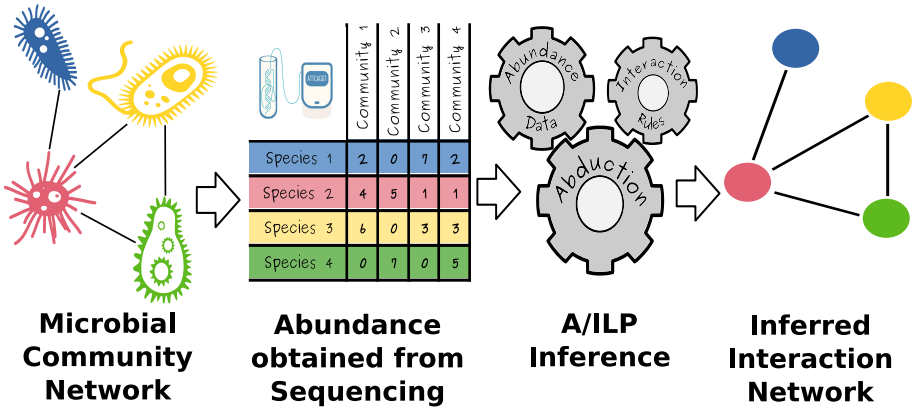

Fig. 1. Description of the interaction inference process. Microbial communities are shaped by the interaction between their members. DNA sequencing together with bioinformatic processes allow to estimate the abundance of the different microbes present in the communities. Using the abundance information from different communities as training examples, and the rules of interaction as background knowledge, it is possible to infer an interaction network that generalizes the interactions between microbes.

3 Methods

3.1 Logical description of microbial interactions

A microbial interaction can be defined as a conserved effect on the abundance of one microbial species caused by the presence of another microbial species. Thus, the aim of the abductive procedure is to infer interactions, following ecological theory to explain the observed changes in the abundance of the species. To do this, the first step is to reflect the abundance changes between communities of each species using logical statements, following the form: abundance(C1, C2, S1, Dir). Here C1 and C2 symbolize two different community samples where species S1 is present and Dir the change in direction of abundance. To calculate the change in direction, the abundances of a species in the two different samples are compared using a Pearson Chi-square test. The test uses the total, summed abundance of all species in a community as the total population and checks the independence of the abundances of the species between the two samples. Where the species counts are found to be independent an abundance change is deemed to exist. An increase is symbolized as an up (↑) and a decrease as a down (↓). Where the species abundances are not independent between the two samples, a no abundance change condition is symbolized as zero (0). The presence of each species is also converted to logical clauses with the structure: presence(C1, S2, yes/no) where C1 refers to a sample community, S2 to a species and yes/no describes if S2 is present in C1 or not.

The abundance change and presence logical statements are used as observations in an abduction process conducted using the A/ILP system Progol 5.0 [12]. The effect on species abundances, either up or down, is described as the change in abundance of one species, S1, due to a second species, S2, when they co-occur in a community, C2. To ensure that the change is caused by S2 it is necessary to evaluate the abundance changes observed in communities where only S2 is present, C1, to communities where both co-occur, C2.

$$\begin{aligned} abundance(C1, C2, S1, up) :- \\ presence(C2, S2, yes), \\ presence(C1, S2, no), \\ effect\_up(S2, S1). \end{aligned} \tag{1}$$

$$\begin{aligned} abundance(C1, C2, S1, down) :- \\ presence(C2, S2, yes), \\ presence(C1, S2, no), \\ effect\_down(S2, S1). \end{aligned}$$

Progol5.0 uses a standard covering algorithm to conduct the abduction process where each observation is generalised using a multi-predicate search. This search is carried out over all the predicates associated with 'modeh' declarations, or abducible predicates, effect_up and effect_down. These two abducible predicates limit the possible variations in abundance that a species can experience due to the effect of another species. The search for the best hypotheses is guided by an evaluation function called 'compression' which is defined as follows:

$$f = p - (c + n) \tag{2}$$

where p is the number of observations (training examples) correctly explained by the hypothesis (positive examples), n is the number incorrectly explained (negative examples) and c is the length of the hypothesis (in this study, always 1 because the hypothesis is a single fact).

At the end of the abduction process, a list of ground hypotheses with the form effect_up/down(S2,S1) is returned, each hypothesis being supported by a compression value f. Implementations of A/ILP usually consider hypotheses with positive compression values. However, compression also offers a quantitative measure of information that can be used to discriminate between true and false interactions. For this purpose, first it is necessary to normalize compression values to a common scale. This is because while some species may not be present in all communities due their different random distribution. It is also possible that negative interactions reduce the abundance of a species to zero. Hence, each species will experience uneven combinations of abundance change mechanisms that require normalization. The normalization is performed using the logarithmic co-occurrence/occurrence ratio of the interacting species.

For an interaction between S1 and S2 to exist, there must be a consistent and constant effect, either up or down, on at least one of the species over all communities. Hence, we use the probabilistic estimator I supporting the interaction between S2 and S1 as defined below:

$$I_{S2,S1} = |f_{up}(S2, S1) - f_{down}(S2, S1)| \tag{3}$$

Compression is dependent on the order in which abundance clauses used as observations are supplied, due to the predicate search process that uses observations as seeds. To obtain reliable compression values, it is necessary to perform the inference several times, permuting randomly the order of examples to obtain different sampling of the hypothesis space. The permutation process will produce a set of possible effects and a corresponding compression value for each pair of species. Effects can be present in all the samples of the hypothesis space or just one part. Thus, it is necessary to define an approach to use the output of the abduction process, after sampling the hypothesis space, as a probabilistic measure. The HFE approach [17] estimates probabilities for hypothetical links in ecological networks, based on their frequency of occurrence when randomly sampling the hypothesis space. In this approach, the compression value was not taken into account to obtain a probabilistic measure of interaction. We propose a different method here that extends HFE to compute a probabilistic estimator I from the values of compression values f. In place of using the frequency of hypotheses with positive compression over all re-samples, here a function $func$ is applied to the f values to obtain an estimator I which summarizes the information contained in all the samples.

$$I_{S2,S1} = |func(f_{up}(S2, S1)_{1,\ldots,n}) - func(f_{down}(S2, S1)_{1,\ldots,m})| \tag{4}$$

In the experiments in this paper, we examined the following $func$ functions to obtain the estimator, I:

- **Frequency** = HFE is computed for each effect.
- **Independent permutations** = Where there is more than one compression value in a permutation, the sum is computed. Maximum values for each interaction among all permutations are retained.
- **Maximum** = Compression values from all permutations are pooled. Then, maximum compression is selected for each effect.
- **Sum** = Compression values from all permutations are pooled. Then, compression is summed for each effect.

3.2 Bootstrapping

Having a probabilistic measure of the likelihood of an interaction is critical for ecologists to interpret the networks resulting from interaction inference. This should allow the selection of those interactions that are realistic and might then be tested in cost- and time-expensive laboratory experiments. The most intuitive selection method would establish a threshold for the estimator value. However,

most of the ecological systems where A/ILP interaction inference could help are poorly described and there are no references to guide selection of such a threshold [15].

It is a common assumption that the interaction networks shaping microbial communities are sparse. This means that the number of interactions of each species is only a small fraction of the total set of interactions that are possible. Thus, where the estimator value of the observed interactions, I, is greater than the values of non existing interactions, it is possible to assess the statistical significance of an interaction using a bootstrapping procedure. Statistical bootstrapping is a method of re-sampling a dataset to create new simulated datasets [4]. Let d be all the compression values for effect up and effect down, involving at least one of the potential interacting species S1 and S2, the real final estimator value, I_0, is obtained by applying equation 4 to n compression values that support an effect up of S2 on S1 and m compression values that supports and effect down of S2 on S1. The bootstrapping procedure re-samples compression values in d to obtain two new sets of values d^{up*} and d^{down*} of n and m lengths, respectively. Then, an alternative estimator value I_a is obtained applying equation 4 to d^{up*} and d^{down*}. If the re-sampling process is repeated B times, a pseudo p-value can be computed for the potential interaction between S1 and S2 averaging the simulated values I_a that are bigger than I_0 [11].

$$\begin{aligned} I_0 &= |func(f_{up}(S2,S1)_{1,\dots,n}) - func(f_{down}(S2,S1)_{1,\dots,m})| \\ I_a &= |func(d^{*up}_{1,\dots,n}) - func(d^{*down}_{1,\dots,m})| \\ p-value &= \sum_{b=1}^{B}\{(I_{a_b} \geq I_0)\}/B \end{aligned} \tag{5}$$

3.3 Simulated data-sets

The aim of ILP based network inference is to use logical descriptions of interactions to detect and classify those interactions between species as a function of the ecological mechanism that drives them. Hence, a simulation model to generate test-datasets should follow the different ecological mechanisms that they are simulating [5]. Information required for network inference is structured in tables, where each row contains the information for a species and each column contains the information for a microbial community. Each cell summarizes the count of individuals of each species in each community (abundance). Weiss et al. (2016) [21] proposed a simulation model to create computer-generated tables including the effects of ecological-like, linear interactions. The model uses the log-normal distribution to simulate the abundance of non-interacting species in a set of microbial communities or samples. The log-normal distribution has been shown to appropriately model the abundance distributions of microbial communities [16]. Interactions are then introduced by modifying the abundance of species in accordance with the different ecological mechanisms [5]. For any two species, say S1 and S2, the abundance modifications only happen in communities where the

species co-occur. The abundance modification is based on the effect that S2 has on S1. If the effect is positive, the abundance of S1 increases as a function of the abundance of S2, modulated by a strength of the interaction. If the effect is negative, the abundance of S1 is decreased following a similar mechanism. In the case that the interaction affects both species, their abundance is modified in parallel.

Using the method proposed in Weiss et al. (2016) [21] we generated three tables containing the abundances of 16 pairs of interacting species in 100 communities. The tables were simulated using interactions of different strength values (2, 3 and 5), and four different ecological mechanisms: amensalism, commensalism, competition and mutualism [3].

3.4 Compositionality and bias

Modern sequencing technologies allow us to recover information about microbial communities from samples of environmental DNA. As noted in Section 1, the number of times that a DNA sequence from a species is 'read' in a sample can be used as a measure of abundance or count. A sequencer can only read a limited number of sequences in a sample, and these are shared amongst species, imposing a compositional bias on the data [9]. Thus, to generate ecological-like microbial tables it is necessary to re-introduce compositionality into the simulated data-sets. To do this, we normalized the sequencing depth as probabilities in a multinomial distribution and then sampled the distribution to obtain the simulated counts across a common sequencing depth.

4 Experimental evaluation

The performance of the A/ILP based microbial inference (Figure 2) is evaluated using the computer-generated datasets. First, it is tested the number of samples of the hypothesis space and the different functions used to obtain the I statistic. Then the best setting found in the first experiment is used to asses the performance of the bootstrapping procedure compared with a threshold for I and SparCC. The simulated data and the code used to perform the experimental evaluation have been included in a public repository[1].

4.1 Experiment 1

Null Hypothesis 1: Using the estimator I as defined in (4) using different functions does not lead to higher accuracy over the frequency-based approach HFE for predicting microbial interactions.

[1] https://github.com/didacb/Machine-learning-of-microbial-interaction

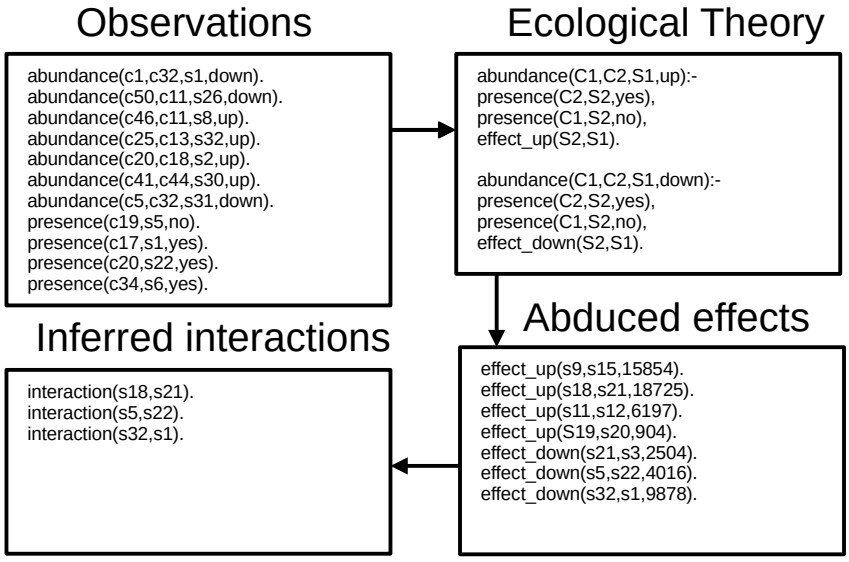

Fig. 2. Summary of the inference process of microbial interactions using A/ILP

Materials and Methods: The three computer-generated tables described in section 2.1, computed using the methodology of Weiss et al. (2016) [21], are used to test the performance of functions used to obtain the estimators. 100 abductions of possible effects are performed for each table. The observations produced from the tables are randomly permuted at each execution. The logical description of effect is used as background knowledge. Then the estimators are obtained using the different functions described previously.

Since the interactions that drive the abundances of the computer-generated tables are known, it is possible to treat interaction inference as a classification problem. Interactions can be classified between existing and non existing and the estimator values obtained using the different functions are the classification accuracy. Thus, the area under the curve (AUC) of the true positive rate against the false positive rate (ROC curve) can be used as a measure of performance. AUC is computed for all functions at n permutations = 1, 5, 10, 25 and 50. An ANOVA test is performed together with a Tuckey's range test to asses the significance of differences of AUC values between all functions.

Results and Discussion: AUC values for the different methodologies to obtain estimators and number of permutations are displayed in Figure 3. As expected, values of AUC increase as the number of permutations used for the inference increases. These stabilize at around n = 50 permutations. AUC values are similar where the strength of interaction is reduced, being significantly lower at the highest strength. This can be explained by the low performance of the logical model

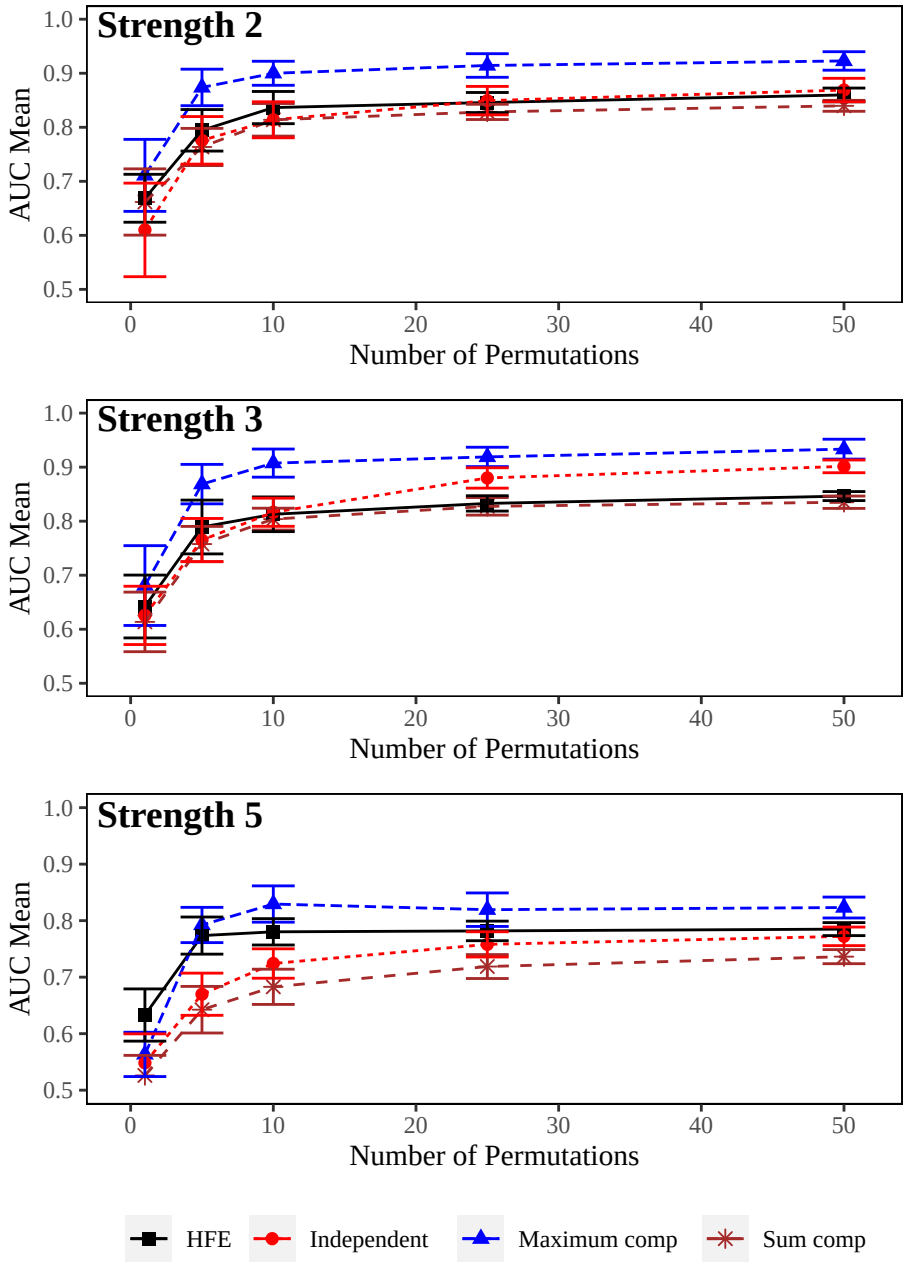

Fig. 3. Area under the ROC curve values (AUC) obtained using different number of permutations. Each plot shows the AUCs obtained inferring interactions of different strengths. Each line represents a method used to obtain the estimators. Error bars show the standard deviation of the means.

in describing the specific case of a negative interaction reducing the abundance of a given species to 0, the likelihood of which increases with stronger interactions. This ecological process, called exclusion, greatly reduces the co-occurrence between species and, as a consequence, the information available to infer an interaction. Maximum compression used to obtain I is the metric that gives the highest values of AUC, for any given number of permutations and interaction strengths. HFE, Sum and independent permutations have similar AUC values at strength 2 and 5 while the independent permutation method performs best at strength 3. The ANOVA shows that all functions have significantly different AUC values, except independent permutations and HFE at strength 2. Consequently, the null hypothesis can be rejected because the method using the maximum compression values to obtain I performs better than HFE in all cases.

4.2 Experiment 2

Null Hypothesis 2: The bootstrapping procedure described in Sec 3.2 leads to lower accuracy compared to the optimal threshold and other statistical methods for interaction inference.

Materials and Methods: The Bootstrapping procedure is conducted using the three computer generated tables used in the preceding experiment. The procedure uses the maximum compression to obtain the I estimator. Two different bootstrapping techniques are evaluated: ordinary and strata. Ordinary bootstrapping performs the bootstrapping independently on all compression values while the strata method constrains the bootstrapping to compression values by effect. Interactions with p value < 0.05 are considered to exist. Bootstrapping accuracy is compared with the accuracy of prediction using an optimal threshold for I estimator. The optimal threshold metric is obtained automatically from the ROC curves of the preceding test using the pROC R package best threshold method [14]. To have a reference for comparing the performance of A/ILP inferring interactions, the interaction inference was also performed using SparCC [7], a widely used statistical inference tool. The process was performed using the FastSpar 1.0 implementation [20] with default parameters.

Results and Discussion: Accuracy measures are displayed in Table 2. Ordinary bootstrap presents better accuracy than strata bootstrap at strength = 2 and 3, while strata performs better at strength = 5. However, ordinary bootstrap allows the detection of a larger number of true positives in contrast to strata. We believe this to be the better option to use for detecting interactions, therefore. In all cases bootstrap accuracy is higher than the optimal threshold accuracy. Bootstrapped A/ILP sensitivity values are significantly lower than SparCC at all strengths. However, the specificity values are slightly higher. Thus, SparCC has a greater number of false positives, while A/ILP generates a higher number of false negatives. This produces similar accuracy measures for SparCC and bootstrapped A/ILP, independent of the interaction strength. We therefore reject the null hypothesis.

Table 2. Performance of estimator bootstrapping compared with optimal threshold obtained from the ROC curve and SparCC. The three datsets used for the interaction inference have 16 real interactions over 496 possible interactions.

Strength 2				
	Optimal threshold	Ordinary Bootstrap	Strata Bootstrap	SparCC
Total	40	13	3	26
TP	13	9	2	12
FP	27	4	1	14
TN	453	476	479	466
FN	3	7	14	4
Sensitivity	0.812	0.562	0.125	0.75
Specificity	0.944	0.992	0.998	0.971
Accuracy	0.94	0.978	0.97	0.964
Strength 3				
	Optimal threshold	Ordinary Bootstrap	Strata Bootstrap	SparCC
Total	69	7	2	31
TP	14	6	1	11
FP	55	10	1	20
TN	425	479	479	460
FN	2	10	15	5
Sensitivity	0.875	0.375	0.062	0.688
Specificity	0.885	0.998	0.998	0.958
Accuracy	0.885	0.978	0.968	0.95
Strength 5				
	Optimal threshold	Ordinary Bootstrap	Strata Bootstrap	SparCC
Total	50	27	4	40
TP	12	10	3	13
FP	38	17	1	27
TN	442	463	479	453
FN	4	6	13	3
Sensitivity	0.75	0.625	0.188	0.812
Specificity	0.921	0.965	0.998	0.944
Accuracy	0.915	0.954	0.972	0.94

5 Discussion and conclusion

This work proposes a framework to infer ecological-like microbial interactions that allows us to use abundance information and descriptions of interactions as logic statements for obtaining a probabilistic measure of the significance of a given interaction, based on compression values. By using logical descriptions of interactions, microbial ecologists can apply their knowledge not only to the interpretation of results but also to the inference process itself.

Tamaddoni-Nezhad et al. (2013) [18] showed how the introduction of ecological expertise (in the form of background knowledge) to the interaction inference can lead to interesting results for invertebrate food webs, and we believe that the work presented here will facilitate similar results for microbial networks.

Interactions between species can be driven by different mechanisms, thus it is necessary to obtain a common quantitative measure of these interactions for appropriate ecological interpretation. Statistical relation learning (SRL) has been used to obtain quantitative measures using ILP-like representations and inference [8]. However, most of the cases where SRL has been used requires probabilistic data, where each observation has an associated probability. However, the observations obtained from NGS data are purely deterministic; a species is either present or not in a given community, and its abundance in this community is also an invariable number. Other authors have proposed different methods to perform probabilistic approaches to deterministic data, such as using a binary matrix obtained from a deterministic process to obtain a support vector machine [1]. The idea of using compression as a probabilistic estimation was also used by Bryant et al. (2001) [2] in their implementation of ASE-Progol. ASE-Progol uses compression to select between contradictory candidate hypothesis. Tamaddoni-Nezhad et al. (2012) [17] developed the Hypothesis Frequency Estimation approach for sampling and estimating the probability of abductive hypotheses. We extended this idea to use the value of compression as a measure for estimating the likelihood of any given interaction. To do this, it is necessary to sample the hypothesis space enough times to ensure that the distribution of compression values obtained for each interaction is representative of all the possible values. Our first experiment showed that a re-sampling of 50 times is enough in a setting involving 32 species and 50 communities, given that the AUC values obtained using a larger sampling were not significantly different. This experiment also showed that retaining the maximum compression values among all hypothesis space samples has greater accuracy than using the HFE, or the other numeric metrics of compression tested, independent of the strength of interactions. This is consistent with the predicate search algorithm of Progol5.0 which selects the hypothesis with the maximum compression from all possible hypotheses [13]. Lastly, it is important to note that the AUC values decreased in all cases where the interactions were strong enough to cause exclusion [3]. In future applications of A/ILP-based interaction inference, it will be important to incorporate logical rules describing exclusion in the learning.

Bootstrapping is a statistic technique used in many areas of knowledge discovery. It has also been applied in statistical inference of interactions [7]. We showed that the bootstrapping procedure has better accuracy values than the optimal thresholds obtained using ROC curves. Thus, it is possible to use this procedure for real data, where the interactions are unknown and the ROC curves cannot be used. Even though bootstrapping offers good accuracy and specificity measures, the sensitivity of inference is insufficient to detect all true interactions. As noted previously, this is in part related to the effect of interaction-derived exclusion of one or both species. It is also due to a restrictive effect of the bootstrapping procedure. Where the bootstrapping is constrained by the effect of abundance, leading to a low number of examples, the sensitivity is low. It is expected that, in real cases where each species interacts with more than one species providing more high compression values for the bootstrapping, the sensitivity will increase.

Weiss et al. (2016) [21] used their method of generate ecological-like datasets, as described in section 3.3, to benchmark many of these interaction inference tools. Comparing the results obtained by SparCC in Weiss et al. (2016) [21] and in this work, a reduction in the number of interacting species reduced specificity and increased sensitivity. However the accuracy values remained similar. A/ILP inference using bootstrap obtained accuracy measures in the same range as SparCC, using the same computer-generated data. This accuracy can be further improved by expanding the range of logical descriptions to other ecological effects and interactions, such as exclusion. Also, it makes it possible to include other sources of biological and ecological information from existing databases as background knowledge.

Our work shows that A/ILP can be used to infer ecological interactions accurately from computer-generated datasets, using an estimator obtained from compression as a numeric measure of interaction and a bootstrap procedure to detect true interactions. Hence, A/ILP interaction inference has the potential to become a valuable tool for microbial ecologists for the inference of ecological interactions.

6 Acknowledgements

This work was supported by the Agence Nationale de la Recherche, Grant/ Award Number: ANR-17-CE32-0011, and SYNGENTA CROP PROTECTION AG. Corinne Vacher and David A. Bohan acknowledge the support of the Learn-Biocontrol project, funded by the INRAE MEM metaprogramme, and the BCMicrobiome project funded by the Consortium Biocontrôle. Alireza Tamaddoni-Nezhad and Stephen Muggleton were supported by the EPSRC Network Plus grant on Human-Like Computing (HLC).

References

1. Amini, A., Muggleton, S.H., Lodhi, H., Sternberg, M.J.E.: A Novel Logic-Based Approach for Quantitative Toxicology Prediction. J. Chem. Inf. Model. **47**(3), 998–1006 (May 2007). https://doi.org/10.1021/ci600223d
2. Bryant, C.H., Muggleton, S.H., Oliver, S.G., Kell, D.B., , Reiser, P., King, R.D.: Combining inductive logic programming, active learning and robotics to discover the function of genes. Electronic Transactions in Artificial Intelligence **6**, 1–36 (2001)
3. Derocles, S.A., Bohan, D.A., Dumbrell, A.J., Kitson, J.J., Massol, F., Pauvert, C., Plantegenest, M., Vacher, C., Evans, D.M.: Chapter One - Biomonitoring for the 21st Century: Integrating Next-Generation Sequencing Into Ecological Network Analysis, Advances in Ecological Research, vol. 58. Academic Press (2018). https://doi.org/10.1016/bs.aecr.2017.12.001
4. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. No. 57 in Monographs on Statistics and Applied Probability, Chapman & Hall/CRC, Boca Raton, Florida, USA (1993)

5. Faust, K., Raes, J.: Microbial interactions: from networks to models. Nat. Rev. Microbiol. **10**, 538–550 (Aug 2012). https://doi.org/10.1038/nrmicro2832
6. Faust, K., Raes, J.: CoNet app: inference of biological association networks using Cytoscape. F1000Research **5** (2016). https://doi.org/10.12688/f1000research.9050.2
7. Friedman, J., Alm, E.J.: Inferring Correlation Networks from Genomic Survey Data. PLoS Comput. Biol. **8**(9), e1002687 (Sep 2012). https://doi.org/10.1371/journal.pcbi.1002687
8. Getoor, L., Taskar, B.: Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning). The MIT Press (2007)
9. Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., Egozcue, J.J.: Microbiome datasets are compositional: And this is not optional. Frontiers in Microbiology **8**, 2224 (2017). https://doi.org/10.3389/fmicb.2017.02224
10. Golubev, W.: Antagonistic Interactions Among Yeasts. Springer Berlin Heidelberg, Berlin, Heidelberg (2006). https://doi.org/10.1007/3-540-30985-3_10
11. Li, J., Tai, B.C., Nott, D.J.: Confidence interval for the bootstrap p-value and sample size calculation of the bootstrap test. Journal of Nonparametric Statistics **21**(5), 649–661 (2009). https://doi.org/10.1080/10485250902770035
12. Muggleton, S.: Inverse entailment and progol. NGCO **13**(3), 245–286 (Dec 1995). https://doi.org/10.1007/BF03037227
13. Muggleton, S.H., Bryant, C.H.: Theory Completion Using Inverse Entailment. Springer Berlin Heidelberg, Berlin, Heidelberg (2000)
14. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., Müller, M.: proc: an open-source package for r and s+ to analyze and compare roc curves. BMC Bioinformatics **12**, 77 (2011)
15. Röttjers, L., Faust, K.: From hairballs to hypotheses–biological insights from microbial networks. FEMS Microbiol. Rev. **42**(6), 761–780 (Nov 2018). https://doi.org/10.1093/femsre/fuy030
16. Shoemaker, W.R., Locey, K.J., Lennon, J.T.: A macroecological theory of microbial biodiversity. Nat. Ecol. Evol. **1**(0107), 1–6 (Apr 2017). https://doi.org/10.1038/s41559-017-0107
17. Tamaddoni-Nezhad, A., Bohan, D., Raybould, A., Muggleton, S.: Machine learning a probabilistic network of ecological interactions. In: Proceedings of the 21st International Conference on Inductive Logic Programming. pp. 332–346. LNAI 7207 (2012)
18. Tamaddoni-Nezhad, A., Milani, G., Raybould, A., Muggleton, S., Bohan, D.: Construction and validation of food-webs using logic-based machine learning and text-mining. Advances in Ecological Research **49**, 225–289 (2013)
19. Vacher, C., Tamaddoni-Nezhad, A., Kamenova, S., Peyrard, N., Moalic, Y., Sabbadin, R., Schwaller, L., Chiquet, J., Smith, M.A., Vallance, J., Fievet, V., Jakuschkin, B., Bohan, D.A.: Chapter One - Learning Ecological Networks from Next-Generation Sequencing Data, Advances in Ecological Research, vol. 54. Academic Press (2016). https://doi.org/10.1016/bs.aecr.2015.10.004
20. Watts, S.C., Ritchie, S.C., Inouye, M., Holt, K.E.: FastSpar: rapid and scalable correlation estimation for compositional data. Bioinformatics **35**(6), 1064–1066 (08 2018). https://doi.org/10.1093/bioinformatics/bty734
21. Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., Xia, L.C., Xu, Z.Z., Ursell, L., Alm, E.J., Birmingham, A., Cram, J.A., Fuhrman, J.A., Raes, J., Sun, F., Zhou, J., Knight, R.: Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. ISME J. **10**, 1669–1681 (Jul 2016). https://doi.org/10.1038/ismej.2015.235