BMC Bioinformatics

RESEARCH Open Access

A comparative analysis of mutual information methods for pairwise relationship detection in metagenomic data

Dallace Francis[1*] and Fengzhu Sun[1]

*Correspondence: dallacef@usc.edu

[1] Quantitative and Computational Biology Department, University of Southern California, Los Angeles, CA 90089, USA

Abstract

Background: Construction of co-occurrence networks in metagenomic data often employs correlation to infer pairwise relationships between microbes. However, biological systems are complex and often display qualities non-linear in nature. Therefore, the reliance on correlation alone may overlook important relationships and fail to capture the full breadth of intricacies presented in underlying interaction networks. It is of interest to incorporate metrics that are not only robust in detecting linear relationships, but non-linear ones as well.

Results: In this paper, we explore the use of various mutual information (MI) estimation approaches for quantifying pairwise relationships in biological data and compare their performances against two traditional measures–Pearson's correlation coefficient, *r*, and Spearman's rank correlation coefficient, ρ. Metrics are tested on both simulated data designed to mimic pairwise relationships that may be found in ecological systems and real data from a previous study on *C. diff* infection. The results demonstrate that, in the case of asymmetric relationships, mutual information estimators can provide better detection ability than Pearson's or Spearman's correlation coefficients. Specifically, we find that these estimators have elevated performances in the detection of exploitative relationships, demonstrating the potential benefit of including them in future metagenomic studies.

Conclusions: Mutual information (MI) can uncover complex pairwise relationships in biological data that may be missed by traditional measures of association. The inclusion of such relationships when constructing co-occurrence networks can result in a more comprehensive analysis than the use of correlation alone.

Keywords: Mutual information, Co-occurrence networks, Non-linear relationships, Asymmetrical relationships

Background

From as early on as in utero development, the microbiome serves as an important mediating agent for human health [1, 2]. However, the mechanisms of its impact are not fully understood as the human microbiome is rather complex. Each person's microbiota contains hundreds to thousands of bacterial and other species represented by a population of trillions of cells; these populations are in constant flux and rely on a multitude of host factors including diet, geography, lifestyle, and medication use [3–5]. Microbiomes are not only unique from individual to individual, but also show high levels of internal variability across the landscape of the human body [6, 7]. One consequence of this is that different body habitats have their own unique implications on host health. For example, the state of the *gut* microbiome and its composition has been associated with Alzheimer's disease [8], Parkinson's disease [9], diabetes [10], chronic kidney disease [11], and inflammatory bowel diseases [12], amongst many others.

The task remains to elucidate the connection between the makeup and processes of the microbiome and host health outcomes. It is well known that the microbiome is a complex ecosystem of many different microbes potentially interacting with one another. Currently, the most common way of approaching this task is through the analysis of co-occurrence networks [13]. In these analyses, microbial networks are inferred from biological data using some quantification of association or connectivity (primarily correlation) and can provide insight of underlying biological processes. For example, in the case of microbiome analysis, a strong correlation between the abundance levels of two microbes may indicate that one of the microbes participates in the regulation of the other [14]. Further analysis of microbial networks could provide novel understandings in microbiota community structure, bacterial niche preferences and resiliency to perturbations, and the identities of key species in bacterial communities [15].

A key aspect in the construction of network representations is the choice of measure used to represent relationships between graph entities. As previously mentioned, correlation is the de facto measure used to quantify connections in microbial networks [16–18]. However, due to the properties of ecological environments and their characteristic interaction types, a simple correlation analysis may not always suffice. For example, in [19] it was concluded that correlation-based methods could consistently identify symmetric relationships, where microbes qualitatively affected each other in a similar way, but failed to provide meaningful conclusions when the relationship of interest was asymmetric in nature, e.g., microbe A increases in abundance while microbe B decreases. Weiss et al. [20] came to similar conclusions–while correlation-based methods are without a doubt useful and important tools in microbial analysis, they are not perfect and perform poorly when the monotonic assumption is violated. Given that the microbiome harbors potentially important instances of asymmetrical, non-linearly associated interactions, it is important that methodologies used in its analysis be sensitive to such cases. It should also be noted that non-linearity is not the only challenge pertaining to the study of the microbiome. As with data from other ecological environments, microbiome data suffers from the curse of dimensionality, sparsity, and compositionality amongst other complications [21]. If the microbiome is to be understood with enough depth to develop effective health interventions in the future, each of these complications must be addressed.

Entropy and mutual information

There is evidence that the use of information theoretics measures in studying biological data can provide an interesting and useful alternative to more traditional approaches [14, 22]. Information theory provides a framework of identifying *general* dependencies between variables; this aspect of generality is particularly promising in the context of biological data where non-linear relationships have been observed experimentally [23, 24]. In this study, we use this framework to analyze pairwise relationships in simulated count data, as well as real-world data from a previously published study on *C. diff* infection.

Consider a pair of discrete random variables X and Y with probability mass functions $p(x)$ and $p(y)$, respectively. The entropies (or uncertainties) of X and Y can be expressed as:

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \geq 0 \tag{1}$$

$$H(Y) = -\sum_{y \in Y} p(y) \log p(y) \geq 0 \tag{2}$$

Given a joint probability mass function, $p(x,y)$, the joint entropy of the variables, $H(X,Y)$, can be expressed as:

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x,y) \tag{3}$$

The conditional entropies of each variable with respect to the other follows as:

$$H(X|Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x|Y = y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(y)} \tag{4}$$

$$H(Y|X) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(y|X = y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)} \tag{5}$$

The conditional entropy is the entropy of a random variable given knowledge of another. Given that X and Y are independent, $H(X|Y)$ reduces to $H(X)$ and likewise for $H(Y|X)$ reducing to $H(Y)$.

Mutual information (MI) extends entropy to quantify how much information one variable contains about another. In other words, it measures the reduction in uncertainty of one variable given knowledge of another [22]. Mutual information can be calculated as a Kullback–Leibler (KL) divergence, representing the distance between a joint distribution, $p(x,y)$, and a product of marginal distributions, $p(x)p(y)$.

$$I_2(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{6}$$

From the definition, one can show that $I_2(X;Y)$ is symmetric, non-negative, and zero if and only if X and Y are independent random variables (e.g., X and Y contain no information on one another). A large mutual information implies a strong dependency between two random variables, with decreasing magnitude corresponding to decreasing dependency.

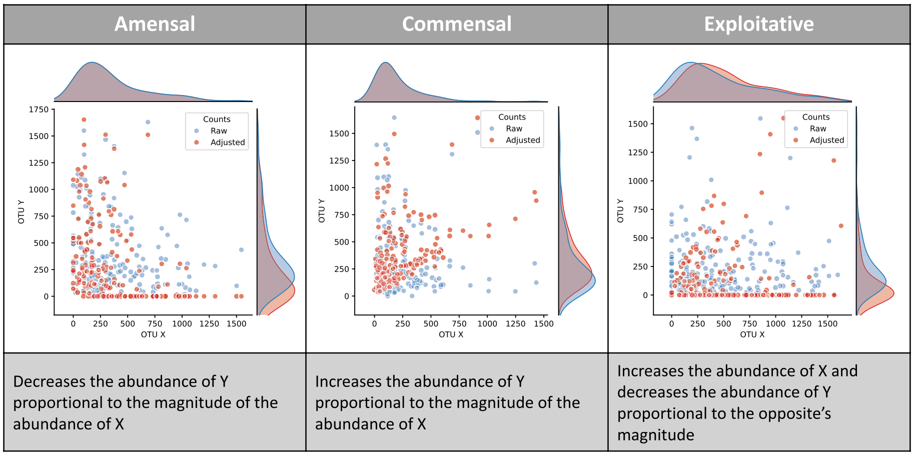
Calculation of the mutual information is trivial when the distribution of a variable is known. However, when working with sampled data, it must be estimated empirically. Given sampled realizations of joint random variables $(X, Y) \sim P_{XY}$, the naïve approach is to approximate the marginal probability distributions of X and Y by imposing fixed, equidistant (or equiprobable) partitions on their respective sample spaces. Subsequently, by using the product of marginal partitions the joint probability density can be approximated. Given that (6) can be rearranged to give the following,

$$I_2(X; Y) = H(X) + H(Y) - H(X, Y), \tag{7}$$

mutual information can be estimated. While computationally favorable, this approach results in unreliable estimates of probability distributions due to the fixing of partitions [25]. Algorithms presented in [25] and [26] are similar to the naïve approach but employ adaptive partitioning methods. Another class of entropy-related estimators rely on kernel density estimation (KDE) [27, 28]. In contrast with the partitioning approach where bins are restricted to rigid, non-overlapping intervals, KDE uses a generalized kernel function to model underlying distributions. This property produces "bins" that overlap and vary in shape and size. Kernel density estimation is a well-studied topic in mathematics; extensive work has been done from describing asymptotic behaviors [29] to determining optimal smoothing parameters [30]. K Nearest Neighbors (kNN) is another entropy estimation technique that focuses on the local distributions of samples rather than their distribution as a whole [31, 32]. In this class of method, each sample is assigned a probability based on the density of its k nearest neighbors. There are many techniques to define the surrounding density, and the reader is referred to [33–35] for further explanation of estimation procedures. Recently, machine learning-based estimation techniques have received significant attention in regard to information theoretics. The motivation being that by using neural networks to implicitly model distributions and their entropy, more robust models that handle non-linearity can be created. The reader is referred to [36] for an extensive review of these types of methods along with their derivations, proofs and specific use cases.

In this study, we test the abilities of various measures of pairwise association to detect different ecological dependencies in simulated data. Loosely following the data simulation protocol in [20], we create synthetic OTU (operational taxonomic unit) count tables of various distributions and induce asymmetric pairwise dependencies that mimic some of those encountered in ecological settings–amensal, commensal, and exploitative relationships (Fig. 1). Given a pair (X, Y) of OTUs, amensal relationships are characterized by the presence of X depressing the abundance of Y, while Y has no effect on X. In contrast, commensal relationships occur when the presence of X increases the abundance of Y, with Y having no effect on X. Exploitative relationships are defined by the presence of X increasing the abundance of Y, while Y simultaneously reduces the abundance of X. While not the exhaustive set of all ecological relationships, the three dependencies represented in this study provide a useful framework for assessing the performance of different association measures in ecological settings.

In total, we apply seven different mutual information (MI) estimators to pairs within each relationship category: Mutual Information Neural Estimation (MINE) [37], an estimator proposed by Nguyen et al. (NWJ) [38], Difference of Entropies (DoE)



Amensal	Commensal	Exploitative
Decreases the abundance of Y proportional to the magnitude of the abundance of X	Increases the abundance of Y proportional to the magnitude of the abundance of X	Increases the abundance of X and decreases the abundance of Y proportional to the opposite's magnitude

Fig. 1 Illustrative examples of the asymmetric ecological relationships explored in this study

[39], Maximal Information Coefficient (MIC) [40], the KSG estimator [33], Local Nonuniformity Correction (LNC) [34], and a naïve grid-based partitioning estimator (see Methods). These MI estimators are compared against the Pearson correlation and Spearman rank correlation coefficients. The performance of each metric in identifying associated pairs is assessed and presented in the following sections.

Figure 1 Each graph displays a pair of simulated OTUs (operational taxonomic units) under two scenarios: the pair is non-interacting/independent (blue) or the pair is subject to a specified ecological relationship (red). Density estimations for each OTU pair are shown along their respective axes

Results

Mutual information provides meaningful results on identifying associated pairs

Each mutual information estimator in this study has the property that its magnitude (or *score*) is proportional to the strength of the dependency it measures–i.e., a higher score corresponds to a stronger pairwise association. Therefore, the task of detecting relationships can be viewed as a classification problem where pairs with null relationships are assigned a score of low magnitude while pairs with dependent relationships are assigned a score of high magnitude. Here, we treat each metric as a binary classifier and assess its classification ability on single-actor (amensal, commensal) and dual-actor (exploitative) asymmetric relationships. Receiver operating characteristic (ROC) curves were constructed for each metric using an equal number of dependent relationship pairs and null relationship pairs. Table 1 shows the AUCs (and corresponding 95% confidence intervals) of all mutual information estimators when detecting each relationship type, based on count table data generated from various prior distributions.

Table 1 AUC results shown are for tools tested on TMM [41] normalized data (n = 50) across several prior distributional assumptions under exploitative, commensal, and amensal relationships. Best results for each prior distribution by relationship type are indicated by bolded font. Bootstrapping was used to produce 95% confidence intervals. It was ensured that a 1:1 parity existed between positive and negative (dependent and null) examples – in particular, each bootstrapped test consisted of 100 dependent and

Table 1 Areas under the ROC curves (AUCs) for each metric under exploitative, commensal, and amensal relationships

		Log normal	Exponential	Negative binomial	Gamma	Beta negative binomial	Avg
Exploitative	MINE	**0.941 ± 0.028**	**0.936 ± 0.025**	0.731 ± 0.059	**0.884 ± 0.041**	**0.826 ± 0.050**	0.864
	NWJ	0.925 ± 0.032	0.879 ± 0.040	**0.745 ± 0.057**	0.856 ± 0.044	0.738 ± 0.058	0.823
	DOE	0.800 ± 0.050	0.772 ± 0.057	0.727 ± 0.062	0.691 ± 0.062	0.721 ± 0.060	0.742
	MIC	0.823 ± 0.051	0.724 ± 0.059	0.663 ± 0.062	0.656 ± 0.062	0.606 ± 0.064	0.694
	KSG (5)	0.917 ± 0.033	0.866 ± 0.041	0.699 ± 0.062	0.804 ± 0.051	0.700 ± 0.066	0.797
	LNC (7)	0.901 ± 0.033	0.865 ± 0.042	0.723 ± 0.058	0.874 ± 0.040	0.700 ± 0.059	0.813
	Partitioning (3)	0.781 ± 0.055	0.786 ± 0.053	0.677 ± 0.062	0.739 ± 0.058	0.720 ± 0.057	0.741
Commensal	MINE	**0.971 ± 0.017**	0.966 ± 0.021	0.900 ± 0.038	**0.974 ± 0.015**	0.922 ± 0.031	0.946
	NWJ	0.886 ± 0.044	0.879 ± 0.043	0.820 ± 0.056	0.913 ± 0.036	0.773 ± 0.056	0.854
	DOE	0.921 ± 0.031	0.943 ± 0.024	**0.925 ± 0.029**	0.933 ± 0.027	0.933 ± 0.028	0.931
	MIC	0.895 ± 0.035	0.871 ± 0.041	0.868 ± 0.039	0.861 ± 0.042	0.806 ± 0.051	0.860
	KSG (9)	0.945 ± 0.024	0.959 ± 0.022	0.905 ± 0.034	0.927 ± 0.030	0.937 ± 0.025	0.935
	LNC (9)	0.955 ± 0.025	**0.977 ± 0.012**	0.884 ± 0.036	0.919 ± 0.031	**0.948 ± 0.022**	0.937
	Partitioning (5)	0.902 ± 0.035	0.897 ± 0.035	0.838 ± 0.049	0.894 ± 0.037	0.846 ± 0.046	0.875
Amensal	MINE	0.833 ± 0.048	0.760 ± 0.054	0.800 ± 0.051	0.824 ± 0.046	0.831 ± 0.049	0.801
	NWJ	**0.842 ± 0.044**	**0.840 ± 0.047**	**0.854 ± 0.043**	**0.870 ± 0.044**	**0.854 ± 0.043**	0.852
	DOE	0.678 ± 0.060	0.732 ± 0.059	0.731 ± 0.058	0.679 ± 0.062	0.772 ± 0.054	0.718
	MIC	0.798 ± 0.050	0.691 ± 0.064	0.723 ± 0.058	0.726 ± 0.057	0.717 ± 0.059	0.731
	KSG (5)	0.815 ± 0.048	0.729 ± 0.059	0.726 ± 0.057	0.787 ± 0.050	0.770 ± 0.055	0.765
	LNC (7)	0.763 ± 0.052	0.705 ± 0.061	0.820 ± 0.050	0.756 ± 0.056	0.749 ± 0.056	0.759
	Partitioning (3)	0.467 ± 0.066	0.543 ± 0.070	0.708 ± 0.058	0.616 ± 0.062	0.719 ± 0.057	0.611

Best results for each prior distribution by relationship type are indicated by bolded font

100 null pairs. For the KSG, LNC, and naïve partitioning methods, only the results from the best performing parameters are included, whereby performance was assessed by the average AUC across all distributions (best parameters are provided in parentheses).

Table 1 shows that the highest performance was observed in commensal interactions, with an average AUC of 0.905 across all methods and prior distributions. This was followed by exploitative interactions (average AUC of 0.782) and amensal interactions (average AUC of 0.748). For commensal interactions, AUCs ranged from 0.806 to 0.977, indicating robust discriminatory power across all distributions. This high performance is expected, as commensal relationships exhibit the least deviation from linearity among all tested relationships. Conversely, all methods showed a decline in performance for exploitative and amensal interactions. AUCs for exploitative relationships ranged from 0.606 to 0.941, while for amensal relationships, they ranged from 0.467 to 0.870, indicating a significant decrease in performance compared to commensal interactions.

When reviewing method-specific performances, the two machine learning estimators using KL-divergence in their formulation, MINE and NWJ, outperformed the other approaches. Across all prior distributions and relationship types, MINE achieved an average AUC of 0.870, while NWJ closely followed with 0.843. Following them in performance were LNC, KSG, DoE, MIC, and the partitioning approach with average AUCs of 0.836, 0.832, 0.797, 0.762, and 0.742, respectively, across all settings. For individual relationship types, MINE was the best approach for both exploitative and commensal interactions, achieving average AUCs of 0.864 and 0.946, respectively. NWJ was the best approach for amensal interactions with an average AUC of 0.852 across all prior distributions. In contrast, DoE, MIC and the naïve partitioning were the worst performing methods, all yielding average AUCs below 0.8 across all relationship types. Outside of the commensal setting, DoE and MIC achieved AUCs above 0.8 only in the case of exploitative relationships with a log-normal prior, while the naïve partitioning approach did not surpass an AUC of 0.8 in any exploitative or amensal scenario. Furthermore, DoE registered sub-0.7 AUCs in three of the fifteen possible combinations of relationship types and prior count distributions, while MIC and naïve partitioning each produced four of such instances. This suggests that purely distributional or grid-based mutual information estimators have limitations and may not be ideal in the context of studying ecological relationships.

It should be noted that results shown for the machine learning-based tools (MINE, NWJ, DoE) were based on scores generated by neural networks that were not tuned for hyperparameters, so their performance is likely sub-optimal. Because each unique interaction pair requires its own independent network, tuning for hyperparameters would result in a drastic increase of the already significant computational time and resources. However, despite the trade-offs between model optimization and performance, MINE and NWJ outperformed all other metrics in over half of distribution × relationship type settings, with 12 of the 15 best results coming from one of these two approaches. When DoE is included, the machine learning-based tools produced 13 of the 15 best results.

Mutual information estimators can reveal associations not detected by conventional correlation measures

In the following sections, we analyze the performance of each metric as a statistical test where the null hypothesis (H_0) is that a pair of variables is independent, and the alternative hypothesis (H_1) is that a pair of variables share a dependent relationship. This is done to enable the direct comparison amongst the MI estimators and with their traditional correlation counterparts. The significance of each pair's interaction (i.e., its p-value) was determined empirically by permutation and subsequently corrected for multiple testing using the Benjamini–Hochberg procedure [42] (Methods).

Figure 2 displays the true positive rates (TPRs) of each metric when applied to simulated count tables; box and whiskers are constructed from 1,000 bootstrapped runs where true ecological relationships are tested against an equal number of null relationships. Immediately, it can be seen that all metrics are much more consistent at detecting commensal relationships than exploitative or amensal ones–in every case, commensal relationships were consistently assigned the highest TPRs regardless of prior distribution. For the correlation measures (Pearson's correlation coefficient and Spearman's rank correlation coefficient), amensal relationships were detected at a higher rate than exploitative ones, and for the MI estimators, exploitative relationships are generally better detected than amensal ones. It can also be observed that methods generally perform worse when analyzing data from discrete prior count distributions. This trend is broken in the case of the traditional correlation methods where an increase in performance for discrete distributions over continuous ones is observed (Fig. 2B, C). Interestingly, in the case of commensal relationships where conditions are closest to linearity, conventional correlation measures were only able to produce results on par

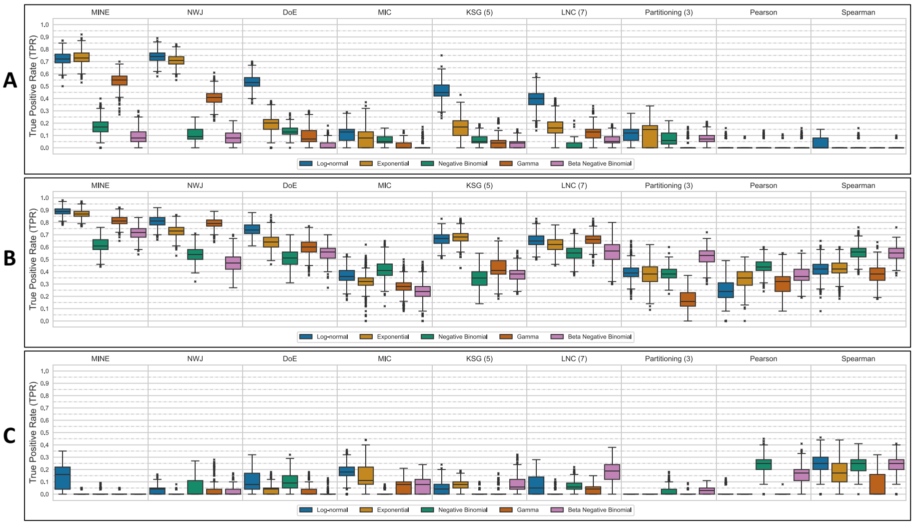


Fig. 2 The true positive rate (TPR) of different methods for detecting (**A**) exploitative, (**B**) commensal, and (**C**) amensal relationships based on different prior distributions. Results for log-normal, exponential, negative binomial, gamma, and beta negative binomial distributed data are distinguished by blue, light orange, green, dark orange, and pink boxplots respectively. TPR values are collected from 1,000 bootstrapped samples of true and null pairwise interactions. Results are separated on the x-axis by method. Boxplots were constructed using results from 1,000 bootstrapped iterations where the TPR was calculated after randomly sampling (with replacement) 100 true positive pairwise relationships and 100 null relationships. Results are shown for data that was TMM normalized and p-values that were corrected using the Benjamini–Hochberg procedure

with the worst performing mutual information estimators (Fig. 2B). However, in the case of the most difficult to detect interactions, amensal relationships, the conventional measures, namely the Spearman rank correlation coefficient, produced the best results (Fig. 2C) although the TPRs are low for all methods tested.

Effects of normalization techniques and choice of multiple test correction on performance

When analyzing metagenomic data, choice of prior count distribution and normalization technique is imperative to the quality of results [43]. All data used for analyses up to this point have been TMM normalized and resulting p-values have been corrected using the Benjamini–Hochberg false discovery rate correction. While these approaches are common in the field, they are not unilaterally considered the "best" normalization method or the "best" way to perform multiple testing correction. To this end, we explored the effects of different combinations of normalization approaches and multiple testing correction methods on the task of interaction detection.

Figure 3 displays how TMM [41], RLE [44], and TSS (total sum scaling) normalizations impacted the ability of each method to (A) detect true pairwise interactions (TPR–true positive rate) and (B) avoid false detection of null ones (FDR–false discovery rate) for exploitative relationships. The results suggest that underlying count distribution plays a much more significant role in the detection ability of each method than choice of normalization (Fig. 3A). All approaches (with the exception of MINE and NWJ) showed stability in TPR across the three normalization procedures–the differences in performance seen in the figures can be primarily attributed to varying distributions. Figure 3B shows that the effect of normalization type and

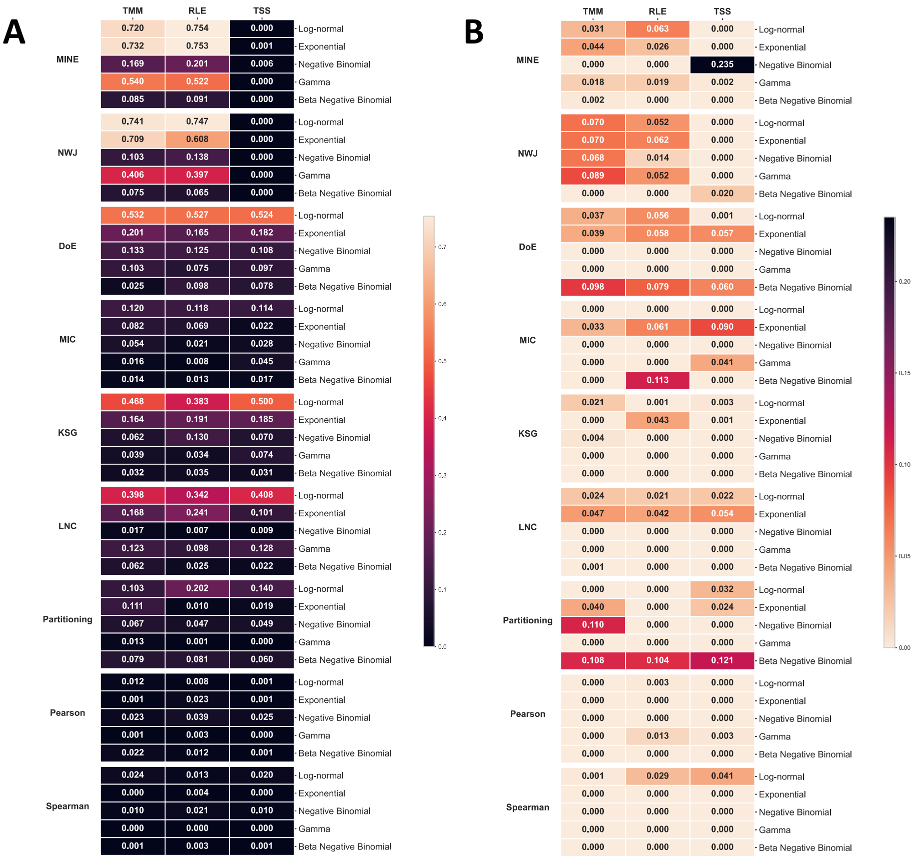


Fig. 3 Effects of normalization and distribution for each method on (**A**) TPR and (**B**) FDR for exploitative relationships. Generally, normalization does not impact results as much as data distribution. Two of the machine learning methods (MINE and NWJ) are exceptions to this, as restricting their input to TSS normalized data renders them uninformative

count distribution on the FDR of each method generally reflected those seen with the TPR. These results are also reflected in commensal and amensal relationship cases (Additional file 2: Figure S1, S2).

Addressing the multiple testing problem is an important step in supporting the statistical validity of any p-value. Due to its general acceptance as a standard, the bulk of this analysis has relied on the Benjamini–Hochberg procedure's false discovery rate correction. However, there is no hard rule that requires this to be the correction method of choice. Figures 4 and 5 display the TPRs and FDRs, respectively, of each method for various significance thresholds in the context of exploitative relationships when using Benjamini–Hochberg procedure corrected p-values, Bonferroni corrected p-values, and q-values [62, 63]. Of the correction methods that produced viable results, the Benjamini–Hochberg procedure proved to be the more conservative approach (Fig. 4, blue). Replacing p-values with q-values consistently improved TPR albeit a small to moderate increase in the FDR; this held true for all metrics and for both empirical and parametric approaches of determining q-values (Fig. 5, green and red). On the other side of the spectrum, an overly conservative approach like the Bonferroni correction (Figs. 4, 5, orange) can lead to a complete loss of detection ability (TPR of 0.0 across all methods).

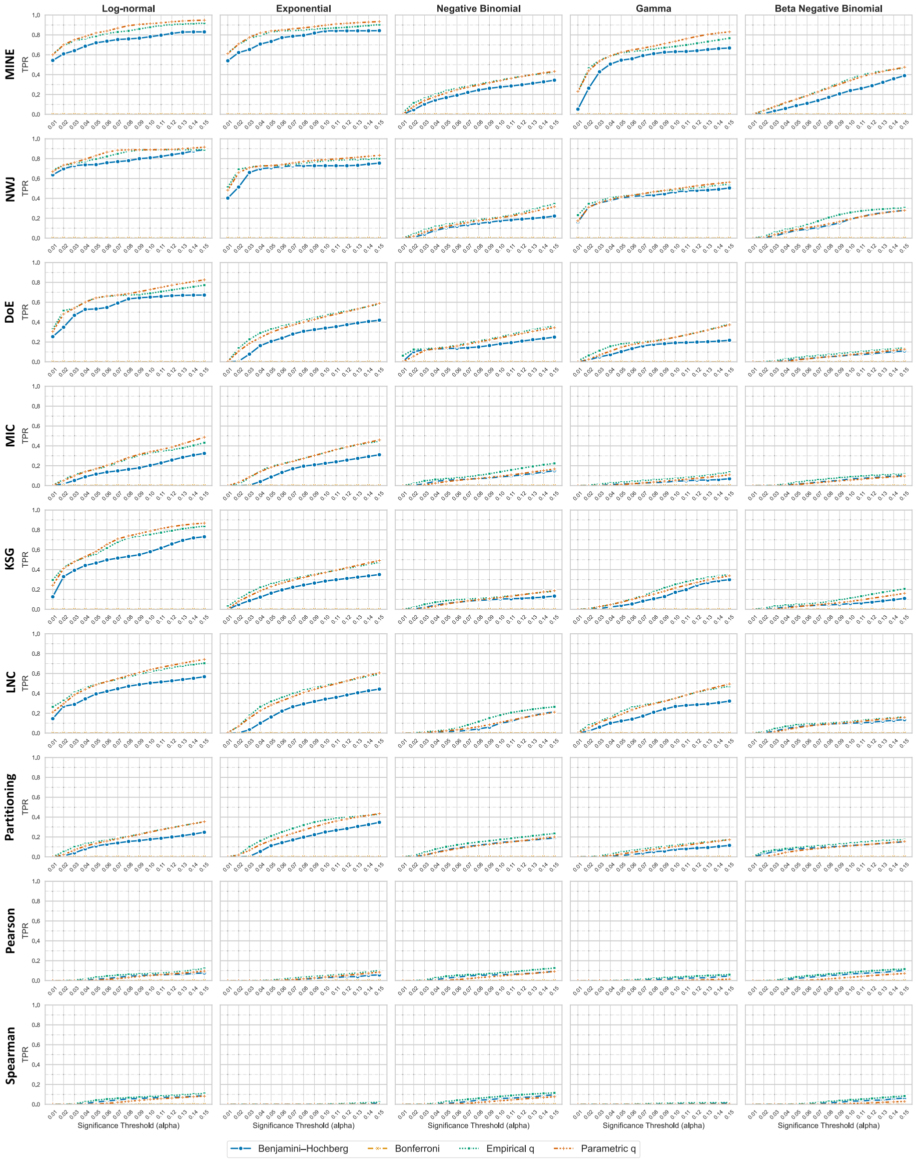


Fig. 4 True positive rates (TPRs) for varying significance thresholds using the Benjamini–Hochberg procedure (blue), Bonferroni (orange), empirical q-values (green), and parametric q-values (red). Both empirical and parametric q-value approaches produce a higher TPR for the same significance threshold than the Benjamini–Hochberg procedure

Data sparsity reduces performance, but the effects are mitigated by increasing sample size

A final analysis on simulated data was interested in the effect of zero-inflated counts on detection ability. It is well established that biological count data is often of this form–approaches have been developed to address the issues sparsity introduces as they have been shown to have a profound impact on model performance [45–47]. We model zero-inflated data by subtracting the mean entry of each generated count table from itself, setting any negative values to zero. Here we return to treating the detection of pairwise relationships as a classification problem and assess the AUC of each metric as a binary classifier. As expected, performance drops for all tools when zeros are inflated (Table 2, top). However, increasing the number of samples used in estimation

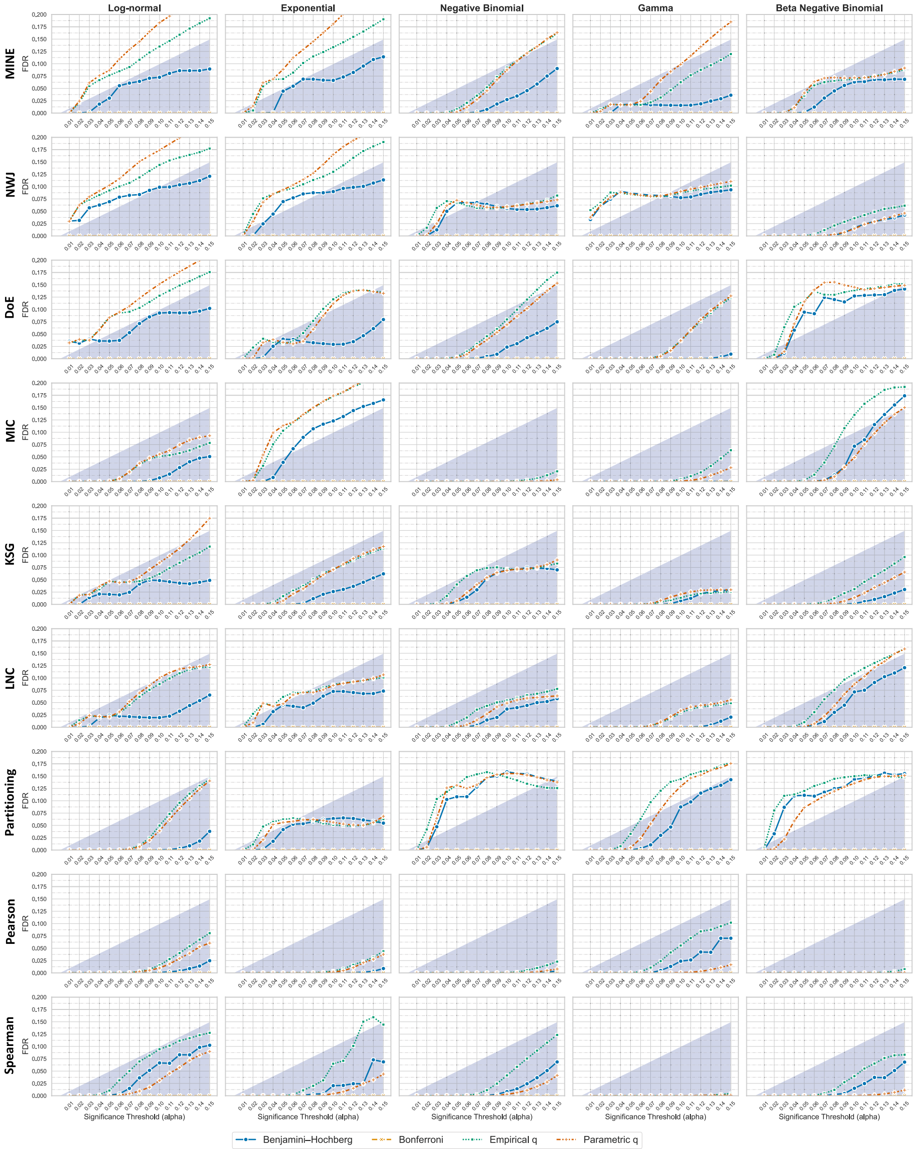


Fig. 5 Respective false discovery rates for the data presented in Fig. 4. Both empirical (green) and parametric (dark orange) q-value approaches usually result in a slight increase in FDR for the same significance threshold than the Benjamini–Hochberg procedure (blue). The shaded blue regions in each plot correspond to FDR values at or below each significance threshold

(Table 2, bottom) restores the effectiveness of nearly every approach along with some gains in performance. When augmenting the sample size of zero inflated tables from 50 to 200, MINE, NWJ, KSG, and naïve partitioning saw the largest increases in AUC at 0.155, 0.147, 0.153, and 0.163, respectively, DoE and LNC saw more moderate increases in AUC at 0.112 and 0.130, respectively, and MIC saw the smallest increase in AUC at 0.098 across prior distributions.

Table 2 AUC results for tools tested on zero inflated, TMM normalized data across several count prior distributions with n = 50 samples (top) and n = 200 samples (bottom) for the exploitative relationship. Similar results for the commensal and amensal cases are provided in Additional file 1: Tables S1 and S2.

Table 2 AUCs for each metric under exploitative relationships with zero inflated counts

		Log normal	Exponential	Negative binomial	Gamma	Beta negative binomial
n = 50	MINE	**0.888 ± 0.041**	**0.866 ± 0.040**	0.590 ± 0.063	0.743 ± 0.057	**0.620 ± 0.066**
	NWJ	0.821 ± 0.051	0.784 ± 0.055	**0.646 ± 0.062**	**0.744 ± 0.055**	0.582 ± 0.067
	DOE	0.459 ± 0.070	0.530 ± 0.070	0.531 ± 0.064	0.479 ± 0.068	0.486 ± 0.066
	MIC	0.835 ± 0.046	0.817 ± 0.050	0.558 ± 0.068	0.660 ± 0.059	0.526 ± 0.068
	KSG (3)	0.648 ± 0.063	0.692 ± 0.060	0.572 ± 0.070	0.658 ± 0.062	0.563 ± 0.066
	LNC (7)	0.689 ± 0.063	0.668 ± 0.062	0.592 ± 0.064	0.668 ± 0.060	0.611 ± 0.066
	Partitioning (5)	0.754 ± 0.057	0.671 ± 0.063	0.556 ± 0.067	0.662 ± 0.061	0.580 ± 0.071
n = 200	MINE	**0.991 ± 0.006**	**0.989 ± 0.007**	0.754 ± 0.056	**0.938 ± 0.023**	**0.810 ± 0.050**
	NWJ	0.945 ± 0.027	0.961 ± 0.021	**0.762 ± 0.058**	0.908 ± 0.035	0.737 ± 0.060
	DOE	0.708 ± 0.060	0.646 ± 0.066	0.552 ± 0.067	0.540 ± 0.069	0.599 ± 0.065
	MIC	0.892 ± 0.038	0.920 ± 0.030	0.715 ± 0.058	0.805 ± 0.050	0.556 ± 0.066
	KSG (9)	0.881 ± 0.039	0.881 ± 0.039	0.713 ± 0.059	0.737 ± 0.059	0.688 ± 0.059
	LNC (7)	0.839 ± 0.044	0.902 ± 0.036	0.659 ± 0.063	0.770 ± 0.055	0.709 ± 0.062
	Partitioning (5)	0.853 ± 0.045	0.905 ± 0.033	0.714 ± 0.059	0.885 ± 0.038	0.682 ± 0.062

Best results for each prior distribution by relationship type are indicated by bolded font

Application of mutual information estimators in the study of *C. diff* infection

To explore mutual information estimators in the real data setting, we applied several of the aforementioned metrics to a publicly available dataset originating from a study on the dynamics of the microbiome following treatment of recurrent *C. diff* infection (CDI) [48]. This dataset contains 16S rRNA profiles of the microbiomes of 38 CDI patients and their respective treatment donors. Sequencing data was retrieved from NCBI using the accession PRJEB19232 and processed to yield genus-level counts (Methods).

MINE, MIC, and KSG were chosen as representative MI estimators for further analysis due to their performances on simulated data. Using only CDI sample data as input, we restricted analyses to pairwise combinations amongst the 30 most abundant genera on average across those samples. Significance for each pair of genera was determined independently of other pairs using permutation (Additional file 3). Figure 6A shows the concurrence among the 20 most significant interaction pairs from each MI estimator for CDI patients. We observe that a large majority of each method's top-ranking pairs are unique, with no pairs being shared amongst all three approaches. When compared against the correlation approaches, we find that MINE shares only 3 of its 20 highly ranked pairs with each of the Pearson and Spearman correlation coefficients while MIC (sharing 9/20 and 11/20 with Pearson's and Spearman's coefficients, respectively) and KSG (sharing 7/20 and 10/20 with Pearson's and Spearman's coefficients, respectively) displays much more overlap. Taking the top pairs from all three MI estimators, we see that their superset covers two-thirds (19/28) of the pairs identified by either the Pearson or Spearman correlation coefficients (Figs. 6B, 7).

While these results speak to the utility MI estimators have in linear settings, we are primarily concerned with their application in non-linear settings. For this, we searched for instances where interaction pairs were highly ranked by MI estimators but deemed

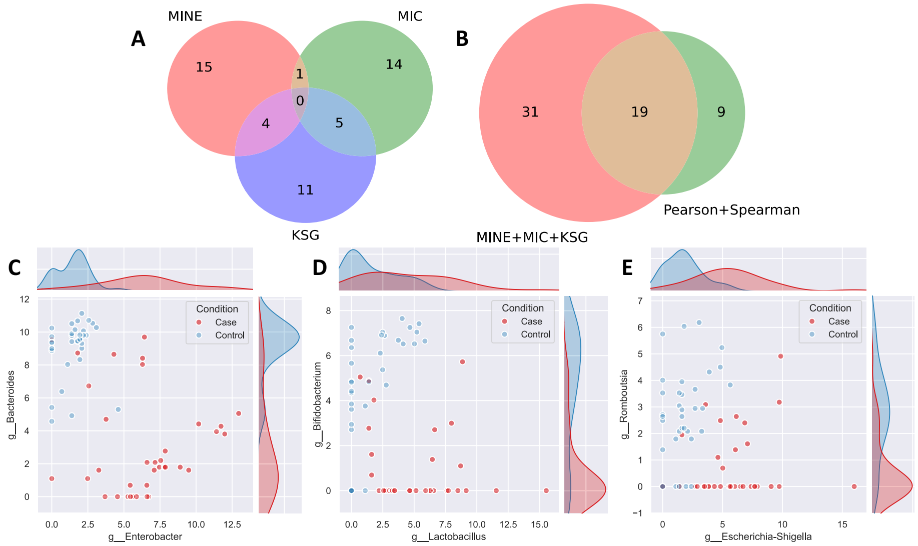
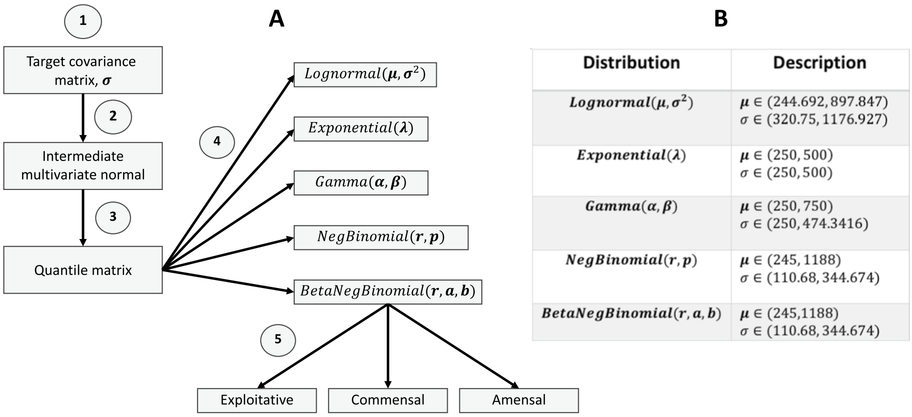


Fig. 6 Venn diagrams detailing overlap of significant relationships found in the CDI dataset (**A**) between MI estimators and (**B**) between MI estimators and correlation measures for the case group. Only the top 20 most significant pairs of each metric are used in the construction of each diagram. (**C**, **D**, **E**) Scatter plots and accompanying density estimations for various relationships found by MI estimators. In each case, there is evidence of an exploitative interaction type, supported by the simultaneous shift of one genus to larger abundances (*Enterobacter, Lactobacillus, Escherichia-Shigella*) and the other to smaller abundances (*Bacteroides, Bifidobacterium, Romboutsia*) when comparing controls (blue) to cases (red). Abundance data is plotted after a $\log(x + 1)$ transform



Distribution	Description
$\boldsymbol{Lognormal(\mu, \sigma^2)}$	$\mu \in (244.692, 897.847)$ $\sigma \in (320.75, 1176.927)$
$\boldsymbol{Exponential(\lambda)}$	$\mu \in (250, 500)$ $\sigma \in (250, 500)$
$\boldsymbol{Gamma(\alpha, \beta)}$	$\mu \in (250, 750)$ $\sigma \in (250, 474.3416)$
$\boldsymbol{NegBinomial(r, p)}$	$\mu \in (245, 1188)$ $\sigma \in (110.68, 344.674)$
$\boldsymbol{BetaNegBinomial(r, a, b)}$	$\mu \in (245,1188)$ $\sigma \in (110.68, 344.674)$

Fig. 7 **A** Flowchart of the data simulation technique. (1) A $d \times d$ target covariance matrix σ with diagonal elements equal to one and off-diagonal elements equal to zero is generated. (2) Using the target covariance matrix, n d-dimensional multivariate normal vectors with mean zero and covariance matrix σ are drawn resulting in an $n \times d$ matrix. (3) Their values transformed into quantiles using the standard normal cumulative distribution function. (4) One of five marginal distributions are imparted on each of the d columns by applying the chosen distribution's inverse cumulative distribution function. (5) Various interaction relationships (exploitative, commensal, and amensal) are introduced between random pairs of columns (representing microbes), producing a final table that simulates an ecological environment in the context of this study. **B** Description of each marginal distribution used in this study. The parameters of each distribution were randomly selected from ranges that resulted in each distribution having a comparable mean, μ, and standard deviation, σ

insignificant by both correlation methods. Scatter plots and accompanying density estimation curves are presented for three of such cases (Fig. 6C-E), all of which can be described by the exploitative ecological relationship. For example, Fig. 6C displays the relationship between *Bacteroides* and *Enterobacter*. Multiple species of *Bacteroides* have been implicated in contributing to a healthy microbiome [49]. Specifically, bacteria from this genus have bile salt hydrolase, allowing them to hydrolyze bile acids (BAs) resulting in host benefits through BA signaling [50]. Proliferation of *Enterobacteriaceae* has been linked to increased gut inflammation [51] and more aggressive diagnoses of ulcerative colitis [52] as well as Crohn's disease [53]. Past research has demonstrated that members of *Enterobacteriaceae* fare particularly well in microbiotas subject to BA dysmetabolism [54]. When examining the plot of CDI patient data alone, there is no obvious interaction taking place between the two genera. However, when control data is included in the plot, one can see a clear directional shift in the distributions of each genus. Reliance on correlation alone would result in this relationship not being detected (p-values of 0.5788 and 0.4431 for Pearson's and Spearman's correlations, respectively); however, all three MI estimators denoted this as a significant relationship (p-values of 0.01607, 0.01198, and 0.001996 for MINE, MIC, and KSG, respectively.

Discussion

In this paper, we explored the use of mutual information in the detection of pairwise, asymmetrical ecological relationships. Several estimators were assessed in performance and compared against the Pearson correlation coefficient and the Spearman rank correlation coefficient; two measures often considered as gold standards in quantifying pairwise relationships.

The results suggest that for exploitative and commensal relationships, mutual information estimators work just as well or better than correlation measures in identifying pairwise dependencies–this conclusion held regardless of normalization approach or count distribution. The advantages of MI estimators, specifically the machine learning-based ones, were clear in the case of exploitative relationships. Correlation alone was insufficient in identifying exploitative relationships but showed some ability in identifying commensal and amensal interactions. This is unsurprising, as in the case of single-actor asymmetrical relationships, shifts from a baseline joint distribution to an ecologically adjusted joint distribution only occur along one axis. In the context of how we simulated data, this type of change is akin to imparting a monotonic function on the joint distribution of a pair of variables–a scenario in which a straightforward measure such as correlation is expected to produce satisfactory results. A potentially more rigorous ecological data simulation technique may very well produce different conclusions.

Entropy-based approaches may be very useful in the task of identifying dual-actor asymmetrical relationships. While they are not perfect identifiers for this class of relationships, entropy-based approaches regularly outperformed correlation approaches under similar simulated conditions and were the only methods able to identify the exploitative relationships presented in the CDI analysis. This phenomenon is likely due to the lack of assumptions made on the type of interaction by entropy-based methods (with the exception of DoE)–both Pearson's and Spearman's correlation coefficients are designed to describe a monotonic relationship between two variables (specifically linear relationships in the case of the Pearson correlation coefficient). As demonstrated in the CDI analysis, pairwise relationships in ecological-type data can manifest themselves in nonmonotonic fashions. The reliance on classical correlation analysis as a "catch all" for every type of ecological relationships may inadvertently exclude very important/interesting organizations of ecological communities. This is an area where entropy-based methods may be of use in future studies.

Due to the realities of working with metagenomic or ecological data, it is often impossible to accurately identify a prior distribution of sampled data. DoE (Difference of Entropies) while having theoretical guarantees of estimation, consistently performed worse than the other machine learning methods. The best results for DoE were seen with Log-Normal data, this is expected given the assumption DoE makes that a variable's conditional distribution shares the same form as its marginal. While this holds for Gaussian distributions, assuming it with other distributions results in ill-specified entropy equations that produce unreliable results. In comparison with other approaches, even without the additional computational stress of hyperparameter tuning, all machine learning methods required significantly longer runtimes to produce, in some cases, comparable results. If computational resources and time are not of concern, they could potentially provide much better results if optimized. MIC, KSG, and LNC were less powerful than the machine learning methods but showed a higher level of stability and consistency with the benefit of a much shorter runtime. One advantage of using MIC is its rigorous theoretical proofs [40] and interpretability. If interactions between groups of variables (rather than between pairs) were the focus of study, kNN based methods could provide a useful approach in quantifying those higher-order relationships as they are applicable to any number of dimensions [33, 34]. As network-based analyses of similar communities continue to grow in scale and importance, the ability to measure groupwise relationships will be a crucial task. Though outside the scope of this particular study, it is of great interest to benchmark mutual information approaches in the context of network analysis.

Conclusions

When studying ecological communities, it may be of great use to incorporate entropy-based metrics alongside traditional correlation measures. While the traditional methods have provided great benefit in the study of these communities, we show in this set of analyses that in the case of asymmetrical relationships (particularly, exploitative relationships), alternative metrics of association can provide higher power. To that end, we encourage future studies to utilize an ensemble approach where multiple measures are used in a complementary fashion. This way, the shortcomings of each can be complemented by the strengths of the other.

Methods

Data simulation

Simulated data generation loosely follows the procedure outlined by Weiss et al. [20]. First, a $d \times d$ target covariance matrixσ, representing the underlying correlations of variables (e.g., microbes) in the count tables, is generated with diagonal elements equal to one and off-diagonal elements equal to zero. Using this covariance matrix, n d-dimensional multivariate normal vectors with mean zero and covariance matrix σ are drawn resulting in an $n \times d$ matrix. The cumulative distribution function (CDF) of the standard normal distribution is then used to transform each element of the matrix into quantiles. From here, one of five marginal distributions (log-normal, exponential, gamma, negative binomial, or beta negative binomial) are imparted on each of the d vectors by applying the chosen distribution's inverse cumulative distribution function. The parameters of each distribution were randomly selected from ranges that resulted in each distribution having a comparable mean and standard deviation (Fig. 6B). Finally, random subsets of variable pairs are adjusted to reflect amensal, commensal, or exploitative relationships using the following non-linear heuristic. Given unadjusted vectors $X = (x_1, x_2, \ldots, x_n)$ and $Y = \left(y_1, y_2, \ldots, y_n\right)$, the pair (x_i, y_i) are adjusted (depending on the modeled interaction) by

$$x_i \pm y_i \times s \times \ln\left(1 + e + e^{\frac{x_i}{x_i + y_i}}\right) \tag{8}$$

$$y_i \pm x_i \times s \times \ln\left(1 + e + e^{\frac{y_i}{x_i + y_i}}\right) \tag{9}$$

In the case of amensal relationships, y_i is depressed by (9) and x is left unaltered. In the case of commensal relationships, y_i is increased by (9) and x is left unaltered. Finally, for exploitative relationships y_i is depressed by (9) and x_i is increased by (8). By modeling pairwise interactions in this fashion, x_i and y_i are adjusted by a factor that: (i) is a function of the other, (ii) depends on the relative magnitudes between the two, and (iii) has non-linear components. We use the variable s as a way to control the strength of relationship between X and Y and set $s = 3$ for the analyses performed as the adjustments at this level provided interactions with enough signal to be detected, but not enough to make detection of pairs trivial. It was ensured that each variable could only participate in one pairwise interaction. This was done to ensure that *only pairwise* relationships were present during analysis. We find that this heuristic provides a non-linear relationship between X and Y without affecting their relative marginal distributions too much and works well for scope of this study. Zero-inflated count data was modeled by subtracting the mean entry of each adjusted count table from itself, then setting any negative values to zero. Prior to any analysis, count tables were subject to either TMM normalization [41], RLE normalization [44], or total sum scaling. Unless stated otherwise, count tables are designed to yield n = 50 samples of d = 1200 variables containing 100 unique examples of each ecological relationship.

CDI data processing

Sequencing data was retrieved from NCBI using the accession PRJEB19232 [48]. The software fastp with default parameters was used to perform an initial round of quality filtering [55]. Following this, reads were imported into QIIME 2 [56] where further quality filtering and denoising was performed using the "deblur denoise-16S" command with the parameter "–p-trim-length" set to 250. Resulting amplicon sequence variants

(ASVs) were then aligned to the SILVA 138–99 rRNA database [57] using the command "feature-classifier classify-consensus-vsearch" with default parameters. The resulting feature table was filtered to remove samples with less than 1000 counts and collapsed to the genus level. Furthermore, only genera present in at least 15% of samples were considered for further analysis.

MINE/NWJ

Without knowledge of underlying distributions, one cannot directly calculate mutual information using its KL-divergence representation (6). However, if a bound on the true value can be established, then optimization techniques can be used to arrive at an approximation. Mutual Information Neural Estimation (MINE) is an approach outlined by Belghazi et al. (2018) that utilizes the Donsker-Varadhan (8) representation of KL divergence and estimates its lower bound by gradient descent over a neural network [37, 58]. Let $\mathcal{F}$ be the family of functions $T_\theta\left(x, y\right)$ parameterized by a deep neural network with parameters $\theta \in \Theta$. Then the lower bound of KL-divergence can be estimated by the following:

$$D_{KL}(\mathbb{P}||\mathbb{Q}) \geq \sup\left(\mathbb{E}_{\mathbb{P}}[T_\theta] - \log \mathbb{E}_{\mathbb{Q}}\left[e^{\tilde{T}_\theta}\right]\right) \tag{8}$$

Here, T_θ and $\tilde{T}_\theta$ refer to the same, identical neural network. The only difference being that T_θ uses realizations of the joint (X, Y) as input, while $\tilde{T}_\theta$ uses independent realizations of the marginals, X and Y, as input. When mutual information is considered, $\mathbb{P}$ is replaced with P_{XY} and $\mathbb{Q}$ with $P_X \otimes P_Y$. Nguyen, Wainwright, and Jordan (2010) opted for an f-divergence representation of KL divergence for estimating mutual information [38]. Their approach (henceforth referred to as NWJ) is very similar to the one detailed in Belghazi et al. (2018), the only difference being an adjustment to the objective function (9).

$$D_{KL}(\mathbb{P}||\mathbb{Q}) \geq \sup\left(\mathbb{E}_{\mathbb{P}}[T_\theta] - \mathbb{E}_{\mathbb{Q}}\left[e^{\tilde{T}_\theta - 1}\right]\right) \tag{9}$$

In both cases, the estimators provide a lower bound for mutual information. The neural networks used for MINE and NWJ were built using the PyTorch library [59] and consisted of one hidden layer of 12 nodes, RELU activation function, and a learning rate of 1e-3.

DoE

McAllester and Stratos (2020) devised an approach that calculates mutual information using estimates of marginal and conditional entropies [39]. The Difference of Entropies (DoE) method uses neural networks to fit prior marginal and conditional distributions to sampled data. It then uses the fitted priors to directly calculate marginal and conditional entropies. Finally, it takes advantage of the fact that mutual information can be formulated by the difference of these quantities:

$$I_2(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{7}$$

By approaching the estimation task in this manner, DoE attempts to avoid a major limitation of lower-bound estimators in that they cannot reliably estimate large values of mutual information. PyTorch was used to build corresponding neural networks with architectures remaining identical to those used for MINE and NWJ.

MIC

The Maximal Information Coefficient (MIC) is a non-parametric, grid-based approach of quantifying the strength of association between a pair of variables [35]. MIC does not set restrictions on the type of association (linear/non-linear) and attempts to assign scores close in magnitude to different relationships with similar noise levels. MIC was calculated using the MICtools software [60] with default inputs.

KSG/LNC

KSG [34] and LNC [34] are two modifications of the kNN algorithm that address different aspects of the estimation procedure. All calculations were carried out using the accompanying software from [35] with k ranging from the default of 3 to 12.

Grid partitioning

The previously mentioned tools are compared against well-established measurements of association–the Pearson and Spearman rank correlation coefficients. Additionally, a simple grid based partitioning approach is also tested. Each variable is divided into equidistant bins and entropies are empirically calculated by counting.

ROC curves and the AUC

The receiver operating characteristic (ROC) curve of a binary classifier is a plot of the classifier's true positive rate (TPR) vs. its false positive rate (FPR) at various classification thresholds. The TPR is defined as the fraction of truly related pairs that are declared as related and the FPR is defined the fraction of non-related pairs that are declared as related. The area under the ROC curve (AUC) of a classifier provides a measure of its discrimination ability. An AUC close to 1 is indicative of high power (i.e., perfect discrimination between classes) while an AUC close to 0.5 is indicative of low power (i.e., random guessing). The Scikit-learn package in Python was used for all ROC and AUC calculations [61]. Confidence intervals were empirically estimated by bootstrapping (with replacement) each set of scores 5000 times, reconstructing ROC curves, and recalculating AUCs.

Hypothesis testing procedure

The performance of each method is assessed using a statistical test where the null hypothesis (H_0) is that a pair of variables is independent, and the alternative hypothesis (H_1) is that a pair of variables share a dependent relationship. Because of their differences in estimation techniques, different methods can assign a wide range of values to the same pairwise relationship. To address this, p-values for each interaction's score were calculated independently of other scores using permutation. Multiple approaches were taken to address the multiple comparisons problem and consisted of the Bonferroni correction, the Benjamini–Hochberg false discovery rate correction [42], an empirical q-value approach [62] and a parametric q-value approach suggested in [63]. Unless specified otherwise, a significance level of 0.05 after multiple test adjustment was used to declare relationships significant for consistency across all tests.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05883-7.

Additional file 1. Supplementary Tables.

Additional file 2. Supplementary Figures.

Additional file 3. Results for CDI analysis.

Acknowledgements

Not applicable

Author contributions

DF designed the study, implemented all relevant code, analyzed the results, and wrote the manuscript. FS supervised the study and assisted in editing drafts of the manuscript. All authors reviewed and approved the final manuscript.

Funding

This research is partially supported by US National Science Foundation (NSF) Grant EF-2125142.

Availability of data and materials

Code for the generation of count tables, implementation of each tested metric, and creation of figures utilized in the manuscript can be found at https://github.com/dallacef/MI-benchmark

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 2 August 2023 Accepted: 29 July 2024
Published online: 14 August 2024

References

1. Robertson RC, Manges AR, Finlay BB, Prendergast AJ. The human microbiome and child growth–first 1000 days and beyond. Trends Microbiol. 2019;27(2):131–47.
2. Mohammadkhah AI, Simpson EB, Patterson SG, Ferguson JF. Development of the gut microbiome in children, and lifetime implications for obesity and cardiometabolic disease. Children. 2018;5(12):160.
3. Sekirov I, Finlay BB. The role of the intestinal microbiota in enteric infection: intestinal microbiota and enteric infections. J Physiol. 2009;587(17):4159–67.
4. Coyte KZ, Schluter J, Foster KR. The ecology of the microbiome: Networks, competition, and stability. Science. 2015;350(6261):663–6.
5. Jandhyala SM. Role of the normal gut microbiota. WJG. 2015;21(29):8787.
6. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. Bacterial community variation in human body habitats across space and time. Science. 2009;326(5960):1694–7.
7. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. Nat Rev Genet. 2012;13(4):260–70.
8. Vogt NM, Kerby RL, Dill-McFarland KA, Harding SJ, Merluzzi AP, Johnson SC, et al. Gut microbiome alterations in Alzheimer's disease. Sci Rep. 2017;7(1):13537.
9. Baldini F, Hertel J, Sandt E, Thinnes CC, Neuberger-Castillo L, Pavelka L, et al. Parkinson's disease-associated alterations of the gut microbiome predict disease-relevant changes in metabolic functions. BMC Biol. 2020;18(1):62.
10. Vallianou NG, Stratigou T, Tsagarakis S. Microbiome and diabetes: Where are we now? Diabetes Res Clin Pract. 2018;146:111–8.
11. Wing MR, Patel SS, Ramezani A, Raj DS. Gut microbiome in chronic kidney disease: Gut microbiome in chronic kidney disease. Exp Physiol. 2016;101(4):471–7.
12. Ferreira CM, Vieira AT, Vinolo MAR, Oliveira FA, Curi R, Martins FDS. The central role of the gut microbiota in chronic inflammatory diseases. J Immunol Res. 2014;2014:1–12.
13. Berry D, Widder S. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. Front Microbiol. 2014;20:5.

14. Watkinson J, Liang KC, Wang X, Zheng T, Anastassiou D. Inference of regulatory gene interactions from expression data using three-way mutual information. Ann New York Acad Sci. 2009;1158(1):302–13.
15. Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, et al. Microbial co-occurrence relationships in the human microbiome. PLoS Comput Biol. 2012;8(7):e1002606.
16. Nusbaum DJ, Sun F, Ren J, Zhu Z, Ramsy N, Pervolarakis N, et al. Gut microbial and metabolomic profiles after fecal microbiota transplantation in pediatric ulcerative colitis patients. FEMS Microbiol Ecol. 2018;94(9):86.
17. Chaffron S, Rehrauer H, Pernthaler J, Von Mering C. A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Res. 2010;20(7):947–59.
18. Riera JL, Baldo L. Microbial co-occurrence networks of gut microbiota reveal community conservation and diet-associated shifts in cichlid fishes. Anim Microbiome. 2020;2(1):36.
19. Pinto S, Benincà E, Van Nes EH, Scheffer M, Bogaards JA. Species abundance correlations carry limited information about microbial network interactions. PLoS Comput Biol. 2022;18(9):e1010491.
20. Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. ISME J. 2016;10(7):1669–81.
21. Calgaro M, Romualdi C, Waldron L, Risso D, Vitulo N. Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. Genome Biol. 2020;21(1):191.
22. Villaverde A, Ross J, Banga J. Reverse engineering cellular networks with information theoretic methods. Cells. 2013;2(2):306–29.
23. Solvang HK, Lingjærde OC, Frigessi A, Børresen-Dale AL, Kristensen VN. Linear and non-linear dependencies between copy number aberrations and mRNA expression reveal distinct molecular pathways in breast cancer. BMC Bioinform. 2011;12(1):197.
24. Hou J, Ye X, Feng W, Zhang Q, Han Y, Liu Y, et al. Distance correlation application to gene co-expression network analysis. BMC Bioinform. 2022;23(1):81.
25. Darbellay GA, Vajda I. Estimation of the information by an adaptive partitioning of the observation space. IEEE Trans Inform Theory. 1999;45(4):1315–21.
26. Fraser AM, Swinney HL. Independent coordinates for strange attractors from mutual information. Phys Rev A. 1986;33(2):1134–40.
27. Moon YI, Rajagopalan B, Lall U. Estimation of mutual information using kernel density estimators. Phys Rev E. 1995;52(3):2318–21.
28. Steuer R, Kurths J, Daub CO, Weise J, Selbig J. The mutual information: Detecting and evaluating dependencies between variables. Bioinformatics. 2002;18:S231–40.
29. Parzen E. On estimation of a probability density function and mode. Ann Math Statist. 1962;33(3):1065–76.
30. Epanechnikov VA. Non-parametric estimation of a multivariate probability density. Theory Probab Appl. 1969;14(1):153–8.
31. Kozachenko LF, Leonenko NN. Sample estimate of the entropy of a random vector. Problemy Peredachi Inform. 1987;23(2):9–16.
32. Singh H, Misra N, Hnizdo V, Fedorowicz A, Demchuk E. Nearest neighbor estimates of entropy. Am J Math Manag Sci. 2003;23(3–4):301–21.
33. Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. Phys Rev E. 2004;69(6):066138.
34. Gao S, Steeg GV, Galstyan A. Efficient Estimation of Mutual Information for Strongly Dependent Variables. arXiv; 2015
35. Lombardi D, Pant S. Nonparametric k-nearest-neighbor entropy estimator. Phys Rev E. 2016;93(1):013310.
36. Poole B, Ozair S, Van Den Oord A, Alemi A, Tucker G. On Variational Bounds of Mutual Information. In: Proceedings of the 36th International Conference on Machine Learning. 2019. p. 5171–80. (PMLR; vol. 97).
37. Belghazi MI, Baratin A, Rajeswar S, Ozair S, Bengio Y, Courville A, et al. Mutual Information Neural Estimation. In: Proceedings of the 35th International Conference on Machine Learning. 2018. p. 531–40. (PMLR; vol. 80).
38. Nguyen X, Wainwright MJ, Jordan MI. Estimating divergence functionals and the likelihood ratio by convex risk minimization. IEEE Trans Inform Theory. 2010;56(11):5847–61.
39. McAllester D, Stratos K. Formal Limitations on the Measurement of Mutual Information. In: Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics. 2020. p. 875–84. (PMLR; vol. 108).
40. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, et al. Detecting novel associations in large data sets. Science. 2011;334(6062):1518–24.
41. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):R25.
42. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc Ser B (Methodol). 1995;57(1):289–300.
43. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol. 2014;10(4):81003531.
44. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
45. Hajihosseini M, Amini P, Saidi-Mehrabad A, Dinu I. Infants' gut microbiome data: a Bayesian Marginal Zero-inflated Negative Binomial regression model for multivariate analyses of count data. Comput Struct Biotechnol J. 2023;21:1621–9.
46. Hu T, Gallins P, Zhou YH. A zero-inflated beta-binomial model for microbiome data analysis: ZIBB. Stat. 2018;7(1):e185.
47. Zhang X, Guo B, Yi N. Zero-Inflated gaussian mixed models for analyzing longitudinal microbiome data. PLoS ONE. 2020;15(11):e0242073.
48. Khanna S, Yoshiki V-B, Antonio G, Sophie W, Bradley S, David AM-P, John FR, et al. Changes in microbial ecology after fecal microbiota transplantation for recurrent C. difficile infection affected by underlying inflammatory bowel disease. Microbiome. 2017;5(1):55.
49. Zafar H, Saier MH Jr. Gut Bacteroides species in health and disease. Gut Microbes. 2021;13(1):1–20.

50. Jia W, Rajani C, Xu H, Zheng X. Gut microbiota alterations are distinct for primary colorectal cancer and hepatocellular carcinoma. Protein Cell. 2021;12(5):374–93.
51. Baldelli V, Scaldaferri F, Putignani L, Del Chierico F. The role of enterobacteriaceae in gut microbiota dysbiosis in inflammatory bowel diseases. Microorganisms. 2021;9(4):697.
52. Walujkar SA, Dhotre DP, Marathe NP, Lawate PS, Bharadwaj RS, Shouche YS. Characterization of bacterial community shift in human ulcerative colitis patients revealed by illumina based 16S RRNA gene amplicon sequencing. Gut Pathog. 2014;6:22.
53. Olbjørn C, Cvancarova SM, Thiis-Evensen E, Nakstad B, Vatn MH, Jahnsen J, Ricanek P, Vatn S, Moen AE, Tannæs TM, et al. Fecal microbiota profiles in treatment-naïve pediatric inflammatory bowel disease—associations with disease phenotype, treatment, and outcome. Clin Exp Gastroenterol. 2019;12:37–49.
54. Kakiyama G, Pandak WM, Gillevet PM, et al. Modulation of the fecal bile acid profile by gut microbiota in cirrhosis. J Hepatol. 2013;58(5):949–55.
55. Chen S, Zhou Y, Chen Y, Jia Gu. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90.
56. Bolyen E, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019;37:852–7.
57. Quast C, Pruesse E, Yilmaz P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41:D590–6.
58. Donsker MD, Varadhan SRS. Asymptotic evaluation of certain markov process expectations for large time. IV. Commun Pure Appl Math. 1983;36(2):183–212.
59. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019.
60. Davide A, Samantha R, Claudio D, Pietro F. A practical tool for maximal information coefficient analysis. GigaScience. 2018;7(4):giy032.
61. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–30.
62. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA. 2003;100(16):9440–5.
63. Pounds S, Morris SW. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of *p* -values. Bioinformatics. 2003;19(10):1236–42.

Publisher's Note