Divergence as a Diversity Measure. Application to Gut Microbiome Analysis

Karim T. Abou–Moustafa*
Dept. of Computing Science
University of Alberta
Edmonton, AB T6G 2E8, Canada
aboumous@ualberta.ca

Yutaka Yasui
School of Public Health
University of Alberta
Edmonton, AB T6G 1C9, Canada
yyasui@ualberta.ca

David S. Guttman
Dept. of Cell and Systems Biology
University of Toronto
Toronto, ON M5S 3B2, Canada
david.guttman@utoronto.ca

James A. Scott
Dalla Lana School of Public Health
University of Toronto
Toronto, ON M5T 1R4, Canada
james.scott@utoronto.ca

Anita L. Kozyrskyj
Dept. of Pediatrics
University of Alberta
Edmonton, AB T6G 1C9, Canada
kozyrsky@ualberta.ca

Friday, July 25th 2014.

This is the first draft of our manuscript entitled above, and made ready on Thursday, August 22nd 2013. The draft is registered as a Tech. Report for the Dept. of Computing Science, University of Alberta, with No. TR13–05, on Sept. 4th 2013. Revised by Yutaka on Oct. 6th 2013. Revised by Karim on Oct. 27th 2013. Revised by Yutaka on Oct. 28th 2013. Revised by Karim on Dec. 9th 2013 (based on comments form Anita and Yutaka). Revised by Karim on Jan. 1st 2014 (based on comments from David). Last revised by Karim on July 24th 2014 (based on comments from James, Anita and David).

*Corresponding Author

Abstract

Entropy measures of probability distributions are widely used measures in ecology, biology, genetics, and in other fields, to quantify species diversity of a community. Unfortunately, entropy–based diversity indices, or diversity indices for short, suffer from three problems. First, when computing the diversity for samples withdrawn from communities with different structures, diversity indices can easily yield non-comparable and hard to interpret results. Second, diversity indices impose weighting schemes on the species distributions that unnecessarily emphasize low abundant rare species, or erroneously identified ones. Third, diversity indices do not allow for comparing distributions against each other, which is necessary when a community has a well-known species' distribution. In this paper we propose a new methodology based on divergence measures to quantify the species diversity of a community. Our two–step approach naturally overcomes the previous mentioned problems, and can be used as an efficient biomarker for health risks. We validate our proposed approach in the diversity analysis of infants' gut microbiota according to mode of delivery and diet. Unlike entropy–based indices, divergence–based measures yield sharp and significantly different diversity results between the groups of each mode, which is consistent with recently reported taxa profiles for these cases.

Keywords

Divergence measures; Diversity measures; Entropy measures; Human microbiome; Species diversity.

1 Introduction

With advances in biotechnologies, characterizations of microbiome diversity has become increasingly important. Recently, various studies in humans and animal model have shown that disturbed acquisition and composition of microbiota during early infancy, for instance subsequent to caesarean section delivery, antibiotic use and formula–feeding, is linked to a greater risk of developing diseases later in life such as allergy, asthma, obesity, metabolic syndrome, necrotizing enterocolitis, diabetes, cancer, infantile colic, and inflammatory bowel diseases [32]. Species diversity is a key component in the portfolio of studying infants' gut microbiota, and in order to proceed to the problem addressed here, we shall begin our discussion with a formal definition of diversity.

Let $\mathcal{C}$ be a community of living organisms where each member of this community (called an individual) has the label of a species. Let s be the number of different species (or individual categories) in $\mathcal{C}$, where the species are labelled from 1 to s. Denote the probabilities of species discovery, or relative abundance, by $\boldsymbol{\pi} = [\pi_1, \pi_2, \ldots, \pi_s]^\top$, where $\sum_{j=1}^{s} \pi_j = 1$, and $\pi_j \geq 0$. Suppose a random sample of m individuals is taken from $\mathcal{C}$ and each individual is correctly classified according to its species identity. If x_j is the number of individuals of the jth species observed in the sample, then $\mathbf{x} = [x_1, x_2, \ldots, x_s]^\top$, where $\sum_{j=1}^{s} x_j = m$, is a multinomial distribution $\mathcal{M}$ with parameters $(m, \boldsymbol{\pi})$; or $\mathbf{x} \sim \mathcal{M}(\boldsymbol{\pi})$ for short.

The diversity of community $\mathcal{C}$ is a key concept in ecological studies. The main difficulty in measuring the (self) diversity of a community (or α–diversity) is compressing the complexity of a distribution, with a multidimensional representation of species relative abundance, into a single scalar statistic [15]. In its simplest definition, a diversity index is a function of two properties that characterize the species in $\mathcal{C}$: (i) the number of species present in the community (species *richness* or *abundance*), and (ii) the evenness with which the individuals are distributed among these species (species *relative evenness* or *equitability*). If s is the number of species in $\mathcal{C}$, then the diversity is higher whenever s is increasing, *and/or* $\mathcal{M}(\boldsymbol{\pi})$ approaches the uniform distribution $\mathcal{U}$; i.e., $\pi_i \approx \pi_j$ for $1 \leq i, j \leq s$ and $i \neq j$.

The previous verbal definition of diversity, although based on *"ecological"* concepts, naturally coincides with the definition of entropy in information theory [26]. Indeed, plant, animal, and microbial ecologists have heavily relied on entropy measures as diversity indices. Further, each research community has proposed its own variants of diversity measures, each exhibiting different sensitivity to one of the aspects characterizing the community (richness, evenness, etc.). Despite the plethora of these diversity indices, the ubiquitous Shannon (or Shannon–Wiener) entropy [26] seems to be the index of choice for various ecology researchers[1]. A widely used estimator for Shannon's entropy H is the maximum likelihood estimate (MLE) given by:

$$\widehat{H} \;=\; -\sum_{j=1}^{s} \widehat{\pi}_j \log_2(\widehat{\pi}_j) = -\sum_{j=1}^{s} \tfrac{x_j}{m} \log_2(\tfrac{x_j}{m}), \qquad (1)$$

where $\widehat{\pi}_j$ is the MLE of π_j. Note that H, like any other entropy measure, is a function defined on the space of distribution functions satisfying some postulates: (i) non negativity, (ii) attains a maximum for the uniform distribution, and (iii) has a minimum when the distribution is degenerate.

[1] Other diversity indices will be discussed in the following sections.

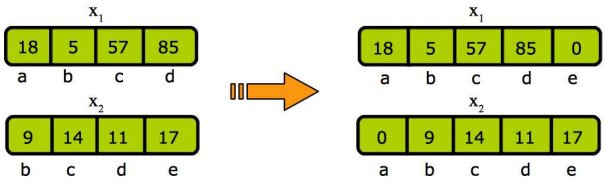
Figure 1: In this example, and using Equation (1), the entropies for samples $\mathbf{x}_1$ and $\mathbf{x}_2$ are: 1.52 and 1.96, respectively. Although it is possible to conclude that $\mathbf{x}_2$ is more diverse than $\mathbf{x}_1$, one should note that these two samples are not comparable since the common species between both samples are only 'b', 'c', and 'd' but not 'a' nor 'e'.

Thus a measure of entropy is in fact, an index of similarity of a distribution function with the uniform distribution $\mathcal{U}$.

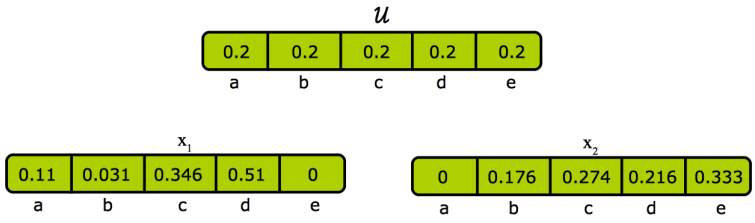
In this paper, we consider three problems of entropy–based diversity measures, exemplified by Shannon's entropy, when used to compare the diversity between two or more communities. Although we faced these problems in the context of gut microbiome analysis, we will show that these problems are independent from the community type under consideration.

The first problem that affects the comparison of multiple communities is due to the convex weight $\widehat{\pi}_j$ assigned to the log term in Equation (1), thereby assigning a larger weight per individual to rare than common species. Such a weighting scheme will increase the influence of rare species while decrease the influence of common species, thereby creating a balance between rare and common species. While such a weighting scheme might be useful in some cases, we argue whether it is always desirable. For instance, if some of the rare species are not the usual habitants of a community, i.e., noisy samples, or some individuals were not correctly classified to their true species identity, then H will unnecessarily emphasize the importance of such samples. More importantly, the reader should note that this weighting scheme alters the true distribution of the species. Thus, it would be desirable to have the flexibility of computing the diversity of $\mathcal{C}$ without relying on such weights.

The second problem arises when comparing two or more values of the Shannon index. That is, when comparing the diversity of two samples, and each collected from a different community, if the two samples do not contain the same species categories and all their relative abundances are non–zeros, Shannon's entropy will be a misleading index of the diversity of both communities. The reason for that is that Shannon's entropy positively correlates with species richness (the number of species categories) and evenness. To see this, consider the example depicted in Figure (1). In this example, $\mathbf{x}_1$ and $\mathbf{x}_2$ are two samples withdrawn from communities $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively. Using



(a) Step 1



(b) Step 2

Figure 2: **(a)** The first step of our proposed approach creates a set of species that is the union of all species from $\mathbf{x}_1$ and $\mathbf{x}_2$. Then, $\mathbf{x}_1$ and $\mathbf{x}_2$ are represented using the unified set of species. Note that this will introduce zero counts in the new representation. Since $0 \log 0 = 0$, using Shannon's entropy for the new $\mathbf{x}_1$ and $\mathbf{x}_2$ will be identical to the situation in Figure (1). However, using the divergence naturally overcomes this problem. **(b)** Samples $\mathbf{x}_1$ and $\mathbf{x}_2$ are represented using their distributions (or relative normalized abundances), and $\mathcal{U}$ is the uniform distribution over the unified set of species. Here we have used $\mathcal{U}$ as the reference distribution to illustrate the main idea. In the second step, the proposed framework measures the diversity of $\mathbf{x}_1$ and $\mathbf{x}_2$ as the divergence between $\mathbf{x}_1$ and $\mathcal{U}$, and between $\mathbf{x}_2$ and $\mathcal{U}$, respectively.

Equation (1), the entropies for $\mathbf{x}_1$ and $\mathbf{x}_2$ are 1.52 and 1.96, respectively. Although, at first glance, it is possible to conclude that $\mathcal{C}_2$ is more diverse than $\mathcal{C}_1$, one should note that these two values are not comparable since the common species between both samples are only 'b', 'c', and 'd' but not 'a' nor 'e'. In fact, it is enough to have one different species in both samples to render the values not comparable. Note that the value of H in the examples above will be more perplexing if the number of species in both samples are not equal, and the situation becomes worse when there are tens or hundreds of samples to compare, each with hundreds or thousands of species.

The third problem is due to the definition of entropy itself which turns to limit the scope of diversity. First, based on the definition of entropy, note that computing the diversity of $\mathcal{C}$ is equivalent to measuring the similarity

between the distribution of species in $\mathcal{C}$ and the uniform distribution $\mathcal{U}$ over the same set of species. Second, note also that $\mathcal{U}$ has the highest entropy (or diversity) among all other possible distributions defined over the s species of $\mathcal{C}$. These two remarks imply that $\mathcal{U}$ is the ultimate reference distribution for comparisons for any community $\mathcal{C}$. However, in nature surrounding us, it is less probable that a community of any living species can have such a uniform distribution. It is more reasonable to believe that each community will have a latent distribution $\mathcal{M}(\boldsymbol{\pi}^*)$ that is not necessarily uniform. Biologists, after a fair amount of research, may provide a reasonable estimate or a model $\mathcal{M}(\widehat{\boldsymbol{\pi}}^*)$ for the latent distribution, which makes it the new reference distribution for a given type of communities instead of $\mathcal{U}$. For instance, in macroecology and community ecology, this is know as the occupancy frequency distribution (OFD) and there has been many advances in that regard since it was first introduced by Raunkiær in 1918 [19, 10]. In such cases, using $\mathcal{M}(\widehat{\boldsymbol{\pi}}^*)$ as a reference distribution will be preferable over using $\mathcal{U}$. Further, if $\mathcal{M}(\widehat{\boldsymbol{\pi}}_1)$ and $\mathcal{M}(\widehat{\boldsymbol{\pi}}_2)$ are empirically estimated from two other communities $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively, an interesting question then is, how to measure the pairwise similarity/dissimilarity directly between $\mathcal{M}(\widehat{\boldsymbol{\pi}}_1)$, $\mathcal{M}(\widehat{\boldsymbol{\pi}}_2)$, and $\mathcal{M}(\widehat{\boldsymbol{\pi}}^*)$ without relying on their entropies?

To overcome the aforementioned problems, we propose a new methodology for assessing and comparing the diversity of multiple communities. Our approach, which is also grounded on information theoretic principles, has two steps. In the first step, depicted in Figure (2a), we overcome the problem of communities with different species by first defining a new set of species that is the union of all species from all communities under consideration. Next, each community is re-represented using the new unified set of species, thereby creating a common ground for comparisons for all communities under study. Note that, for Figure (2b), using the new representation will introduce species with zero counts in the sample. If entropy is used to assess the diversity of these communities, then zero count species will be neglected by H since $0 \log 0$ is 0, which reduces to the problem depicted in Figure (1). We overcome this problem, however, using the second step of our proposed framework.

In the second step, we generalize entropy–based diversity indices to divergence–based indices. That is, instead of measuring the entropy for each community given its new representation based on the unified set of species, we compute the divergence between the distribution of each community and the reference model $\mathcal{M}(\widehat{\boldsymbol{\pi}}^*)$. When $\mathcal{M}(\widehat{\boldsymbol{\pi}}^*)$ is not known for the community under consideration, then one has no other option but to use the ultimate diverse distribution which is the uniform distribution $\mathcal{U}$ defined

over the unified set of species, as depicted in Figure (2b). Unlike entropy-based measures, the divergence measures the dissimilarity (or difference) between any two probability distribution functions defined over the same set of outcomes. In other words, the divergence between two distribution functions is analogous to the distance between two points in an Euclidean space. As will be explained in § 3, zero count species are not neglected by divergence measures, and they increase the dissimilarity between the two distributions. Hence, by definition, the divergence overcomes the second and third problems of entropy-based measures mentioned above. Further, divergence measures do not impose any weights that alter the original sample distribution under consideration, and therefore they also overcome the first problem we discussed above of entropy-based measures.

Readers familiar with Whittaker's *beta diversity* [30] should note the difference between this type of diversity on one hand, and the methodology proposed here on the other hand. Beta diversity [30, p. 320] measures the extent of change in community composition, or degree of community differentiation, in relation to a complex-gradient of environment, or a pattern of environments. Note that this description covers two different aspects for a community: (*i*) the change in the composition of the community itself, and (*ii*) the degrees of differences in diversity between the community itself (as a subgroup), its surrounding communities (other subgroups), and the species diversity at the regional or landscape scale. See [29] for a clear overview of beta diversity. Our proposed methodology as described above, is not addressing the extent of compositional change in one community, nor is addressing the relation and structural differences between a community and its surrounding communities, or its surrounding region at large.

To the best of our knowledge, we are unaware of any research in the literature that has addressed the above issues together with a proposed solution. We validate the proposed divergence–based diversity measures in the analysis of 24 infants' gut microbiota according to infant diet (breast-fed vs. formula-fed), and mode of delivery (vaginal vs. caesarean section). Note that in this context, community sequencing of hypervariable portions of the bacterial 16S rRNA gene yields a series of unique motifs considered individually as "operational taxonomic units" (OTUs). Each OTU corresponds to a unique group of individuals sharing a common taxonomic affiliation (e.g., at the level of subspecies, species, genus, family or higher). In this analysis, we show that diversity indices such as Shannon's entropy index, Gini–Simpson's index, and Hill–Jost numbers, fail to detect any significant difference in the microbiota diversity among these groups. However, divergence–based diversity measures are more sensitive in detecting dysbio-

sis of infant gut microbiota secondary to the delivery method and infant diet, and yield findings consistent with observed differences in the relative abundance of individual microbiota species, as we and others have reported [25, 2]. Last, we show how a precise statement using our proposed measures can turn into an efficient biomarker for health risks.

2 Background and Literature Review

In this section we cover two different aspects for the research presented in this paper: (*i*) a brief literature review of diversity indices, and (*ii*) the specific context of our study which assesses the diversity of infants' microbiota according delivery mode and diet.

2.1 Overview of Diversity Indices

Since its introduction in 1943 [7, 17, 18], the concept of species diversity has been defined in various and disparate ways leading to a plethora of diversity measures with different and rather "conflicting" characteristics [13]. This has led some researchers in the 70's, such as Hurlbert [11], to conclude that species diversity is meaningless. More recently, this debate has evolved to the need for a consistent terminology for quantifying species diversity [20, 28]. The first effort to disambiguate the term is due to Whittaker [31], followed by Hill [8], and more recently by Jost [13]. Most researchers, including Hurlbert, have agreed that the definition of a community's diversity within itself (α–diversity) should, at best, be restricted to the one introduced in § 1. Jost [13] made a further distinction between a diversity index, such as H, and a diversity number. In his argument: "*A diversity index is not necessarily a diversity. The radius of a sphere is an index of its volume but is not itself the volume, and using the radius in place of volume in engineering equations will give catastrophic misleading results*". Based on his argument, the diversity of a community reduces to finding a community that is composed of equally common species. Using simple algebra, he devises an algorithm for recovering the diversity number given the value of a diversity index. For instance, the expression for the diversity number based on Shannon's index is $\exp(-H)$.

In the literature, there are two other well known indices, the Simpson's index [27]: $Sp = \sum_{j=1}^{s} \widehat{\pi}_j^2$, and the Chao-1 index [4]: $Ch = s + \frac{a^2}{2b}$, where a is the number of singletons (species with a single occurrence), and b is the number of doubletons (species with a double occurrences) in $\mathcal{C}$. Simpson's index is sensitive to the abundance of the more plentiful species in a sample

and therefore can be regarded as a measure of dominance concentration. Similar to H, Simpson's index is a weighted mean of the relative abundances, and both measures were shown to be special cases from Rényi's entropy. Hill [8] and Jost [13], however, advised to use the reciprocal of Simpson's index, $1/Sp$, or the generalized entropy, $\ln(Sp)$, as diversity numbers, while Whittaker [31] and Pielou [22] favoured the Gini–Simpson index: $1 - Sp$.

Shannon's and Simpson's indices perform as expected when approximating the diversity of common species, however each may fall short as a single complete measure when examining numerous low abundant organisms that dominate the composition of a community [15]. Both indices have been shown by Hill, through Renyi's definition of generalized entropy [24], to have similar characteristics, but differing only in the contribution of low abundant species to the magnitude of the calculated statistic. Renyi's entropy unifies Shannon's and Simpson's diversity indices as entropies with a parmeter q, the power to which the contribution of taxonomic abundances is raised:

$$D_q \quad = \quad \left(\sum_{i=1}^{s} \pi_i^q \right)^{\frac{1}{1-q}} . \tag{2}$$

Hence, q values of 2, 1, and 0, are associated with Simpson's index, Shannon's index, and the total number of species detected, respectively. While these are known as Hill numbers, surprisingly, Jost's interpretation and algorithm for recovering the diversity number from any entropy–based diversity index yields exactly the expression in Equation (2).

Chao-1 index, in fact, is a richness estimator – i.e., an estimator for s – although various studies have used it as a diversity measure. Chao-1 relies on the existence of singletons and doubletons in the sample. If no singletons nor doubletons in the sample, Chao-1 equals the number of observed species in the sample. Note that Chao-1 does not strictly follow our chosen definition of diversity introduced in § 1 since it does not address the equitability of relative abundances in the sample.

Despite the differences between all the above indices, it is worth noting that various researchers consider that the number of species, Simpson's index, and Shannon's index are in some sense, similar evaluations for the number of species present in the sample, and they only differ in their propensity to include or exclude the relatively rare species [8].

In a different research path, Chao and Shen [5] consider three shortcomings of the MLE for H in Equation (1): (*i*) Equation (1) is derived under the assumptions that s is known, (*ii*) it is assumed that $m > s$, and (*iii*) the fact that the MLE $\widehat{\pi}_j$ is negatively biased; i.e., $\widehat{H}$ is an underestimate

for H. In practice, the true value of s is unknown, and rare species may not be discovered in a sample due the existence of numerous low abundant species. Further, due to negative bias of $\widehat{\pi}_j$, $\widehat{H}$ yields an estimation error that will differ between samples, depending on the diversity and evenness in each, and will be large for small samples [9]. Hence the authors proposed a nonparametric estimator for H for the particular case when s is unknown, while taking into account the possibility of having unseen species. Note that the motivations for the Chao and Chen estimator are different from our motivations discussed in § 1. Further, their estimator relies on the concept of sample coverage to adjust the sample fraction for unseen species which relies on the presence of singletons and doubletons as in the Chao-1 index. Such assumptions on singletons and doubletons might not be applicable in some domains. For instance, due to the current bioinformatics approaches for translating raw sequenced output into OTU abundance, it is rare, if not impossible to enumerate OTUs with single or double occurrences since they are usually filtered out prior to the determination of OTU abundances in sequenced samples.

2.2 Infant Gut Microbiome Profile According to Infant Diet and Mode of Delivery

In this section we discuss the specific context of our study where we applied our proposed divergence–based measures. Recently, [2] have profiled the gut microbiome (using fecal samples) of 24 healthy Canadian infants selected from a national birth cohort, according to mode of delivery – vaginally (V) vs. cesarean section (CS), and infant diet – breast-fed (BF) vs. formula-fed (FF). In their study[2], the authors found that, unlike vaginally born infants, CS–delivered infants had bacterial communities with significantly lower relative abundances of genus *Escherichia-Shigella* and an absence of *Bacteroides*. Further, compared with infants who were breast-fed, those who were not breast-fed had bacterial communities with significantly higher abundances of the family Peptostreptococcaceae and the family Verrucomicrobiaceae (genus *Akkermansia*). Colonization with *Clostridium difficile* was least likely significantly lower among exclusively breastfed infants than among infants receiving formula; the prevalence did not differ by mode of delivery.

To measure richness and diversity of the samples, the authors used the Chao-1 estimator (see § 2.1) and Shannon's entropy MLE in Equation (1).

[2]For complete details on this study, please refer to [2]

Table 1: Richness and diversity (with standard deviation) of fecal microbiota in infants, by early-life exposures [2].

Exposure	**No. infants (n)**	**Chao-1 (Richness)**	**p value**	**Shannon (Diversity)**	**p value**
Mode of delivery					
Vaginal	18	11.2 (4.4)	0.007	1.33 (0.49)	0.06
Emergency CS	3	19.7 (3.2)		2.02 (0.48)	
Elective CS	3	9.3 (1.5)		1.09 (0.47)	
Diet at 4 months					
Exclusive BF	10	9.0 (4.1)	0.006	1.19 (0.51)	0.1
Partially BF	5	12.6 (5.3)		1.42 (0.64)	
Not BF	9	15.0 (4.0)		1.58 (0.47)	

Table (1), taken from Table (4) in [2], shows the values of Chao-1 and Shannon's entropy (with standard deviation), and p values for the significance tests (two-tailed Student t–test, with trend test for diet group). It can be seen that the values of the Chao-1 estimator are consistent with the general profile characteristics described above. Indeed, there is a statistical significance ($p = 0.007$) between the richness of vaginally delivered infants, emergency CS infants, and elective CS infants. Similarly, there is a statistical significance ($p = 0.006$) between exclusively BF infants, partially BF infants, and only FF infants.

On the other hand, Shannon's index values did not show any statistical significance between any of the groups, which is not consistent with the richness estimator, nor with the microbiota's profile described above. For instance, for infant diet, $p = 0.1$, not BF infants had the highest diversity, followed by partially BF infants, and finally exclusively BF infants. According to Shannon's index, no significant difference can be found between the different groups. The same interpretation follows for delivery mode, $p = 0.06$.

Note that when we carried our own analysis using the same dataset (§ 4) no statistically significant differences in diversity were found among these groups when using the Gini-Simpson index and Jost's diversity numbers [13] (or Hill numbers [8]). In § 4, we will show that our proposed approach for measuring diversity will yield sharper results, with statistically significant differences among all these group.

3 Divergence–based Diversity Measures

In this section we introduce our two-step framework for measuring the diversity using divergence measures. We begin our discussion with the necessary notations. Let $\{\mathcal{C}_i\}_{i=1}^n$ be the set of communities under study, and $\mathbf{x}_i = [x_i^1, \ldots, x_i^j, \ldots, x_i^{s_i}]^\top$ be the sample withdrawn from $\mathcal{C}_i$, where s_i is the number of observed species (or OTUs) in $\mathcal{C}_i$. Accordingly, $\mathbf{x}_i \sim \mathcal{M}(m_i, \boldsymbol{\pi}_i)$, where $\sum_{j=1}^{s_i} x_i^j = m_i$ is the total number of individuals in the sample $\mathbf{x}_i$. Let $\Omega_i = \{o_1, \ldots, o_j, \ldots, o_{s_i}\}$ be the set of species' labels (or OTUs) found in $\mathcal{C}_i$. To avoid any reliance on the order of species labels in Ω_i, for any label o, we use the following notation to index the elements of sample $\mathbf{x}_i$:

$$\mathbf{x}_i(o) \;=\; \begin{cases} x_i^j & \text{if } o = o_j \text{ and } o_j \in \Omega_i, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

The first step of our proposed framework is to have a unified representation for all samples. To achieve this, let Ω^* be the union set of species collected from all the samples under consideration:

$$\Omega^* \;=\; \bigcup_{i=1}^{n} \Omega_i \;\equiv\; \{o_1, \ldots, o_s\}, \tag{4}$$

where the cardinality of Ω^* is s. The set Ω^* includes all $\{\Omega_i\}_{i=1}^n$, and hence all samples $\{\mathbf{x}_i\}_{i=1}^n$ need to be represented in terms of its elements. This can be obtained using our notation for indexing the elements of $\mathbf{x}_i$ in Equation (3):

$$\bar{\mathbf{x}}_i \;=\; [\mathbf{x}_i(o_1), \mathbf{x}_i(o_2), \ldots, \mathbf{x}_i(o_s)]^\top,\; 1 \le i \le n, \tag{5}$$

where $\bar{\mathbf{x}}_i$ is the new sample representing $\mathcal{C}_i$ using Ω^*. Further, we define the empirical discrete distribution $\mathcal{X}_i$ from $\bar{\mathbf{x}}_i$ as:

$$\mathcal{X}_i \;=\; \left[\widehat{\pi}_i^1, \ldots, \widehat{\pi}_i^s\right]^\top \;\equiv\; \left[\tfrac{\mathbf{x}_i(o_1)}{m_i}, \cdots, \tfrac{\mathbf{x}_i(o_s)}{m_i}\right]^\top,\; 1 \le i \le n. \tag{6}$$

The rational for using Ω^* instead of $\{\Omega_i\}_{i=1}^n$ is that it provides a common ground for comparing all samples from different communities. That is, it reduces the comparison between communities to the differences in the distribution of relatives abundances. The problem, however, is that the new representation $\bar{\mathbf{x}}_i$, and consequently the discrete distribution $\mathcal{X}_i$, is sparse; i.e., it contains a considerable number of zero elements since not all species in Ω^* are present in all $\mathcal{C}_i$'s. Recall that entropy–based diversity measures

correlate with the number of (nonzero) species in the sample, and with the evenness (or equitability) of the relative abundances (or the individuals' distribution in a sample). When using entropy–based diversity measures on such representations, it is enough to have one zero element per sample (in any location) to render the entropy values meaningless and not comparable. This is exactly the scenario depicted in Figure (2a), and since $0 \log 0 = 0$, it reduces to the problem in Figure (1). Even if $\mathcal{X}_i$ does not have any zero elements, entropy–based measures will alter the original distribution to create a balance between rare and abundant species. In addition, entropy–based measures are not flexible in terms of the reference distribution, nor they allow for pairwise comparisons between all samples. We overcome these problem, however, using the second step of our proposed framework.

3.1 From Entropy to Divergence

To overcome the above problem, we rely on the basic definition of entropy (which coincides with our definition of diversity). That is, an entropy measure is a function defined on the space of distribution functions satisfying some postulates: (i) non negativity, (ii) attains a maximum for the uniform distribution (i.e., maximum diversity), and (iii) has a minimum when the distribution is degenerate. Thus a measure of entropy is in fact, an index of similarity of a distribution function with the uniform distribution $\mathcal{U}$. Let us define the uniform discrete distribution over Ω^*:

$$\mathcal{U} \;=\; [u_1, u_2, \ldots, u_s]^\top \;=\; [\tfrac{1}{s}, \tfrac{1}{s}, \cdots, \tfrac{1}{s}]^\top. \tag{7}$$

The second step of our proposed framework is to replace the entropy of a distribution with a surrogate function that measures the dissimilarity between the given distribution, say $\mathcal{X}_i$, and the reference distribution $\mathcal{M}(\widehat{\boldsymbol{\pi}}^*)$. When $\mathcal{M}(\widehat{\boldsymbol{\pi}}^*)$ is not known, then one has no other option but to use the uniform distribution $\mathcal{U}$ defined over Ω^* as a reference distribution.

The natural function that measures the dissimilarity between any two probability distributions is the divergence, Ali–Silvey distance [1], or f-divergence according to Csiszar [6, 14]. If $\mathfrak{D}$ is the space of probability distributions, and $\mathcal{P}, \mathcal{Q} \in \mathfrak{D}$ are two distributions defined over the same set of outcomes $\mathfrak{E}$, then the divergence quantifies how $\mathcal{P}$ diverges from $\mathcal{Q}$ over all the elements of $\mathfrak{E}$. For simplicity, the divergence between two probability distributions is analogous, for instance, to the Euclidean distance between two points in an Euclidean space. The smaller the divergence between two distributions, the more similar these two distributions are, and vice versa.

The divergence between $\mathcal{P}$ and $\mathcal{Q}$, denoted by $\text{div}(\mathcal{P}, \mathcal{Q})$, has to satisfy some conditions. One of the conditions relevant to our discussion is that div should be zero when $\mathcal{P} = \mathcal{Q}$, and as large as possible when $\mathcal{P}$ and $\mathcal{Q}$ are completely different. The divergence by definition does not need to be symmetric, nor does it need to satisfy the triangle inequality, and hence it is different from distance metrics in that regard. However, in this research work, we will consider symmetric divergence measures, and some will satisfy the triangle inequality. That is, for $\mathcal{P}, \mathcal{Q}, \mathcal{Z} \in \mathfrak{D}$, all defined over $\mathfrak{E}$, then $\text{div}(\mathcal{P}, \mathcal{Q}) = \text{div}(\mathcal{Q}, \mathcal{P})$, and $\text{div}(\mathcal{P}, \mathcal{Z}) \leq \text{div}(\mathcal{P}, \mathcal{Q}) + \text{div}(\mathcal{Q}, \mathcal{Z})$.

Since we are interested in discrete probability distributions, let $\mathcal{P} = [p_1, \ldots, p_s]^\top$, and $\mathcal{Q} = [q_1, \ldots, q_s]^\top$, where for $1 \leq j \leq s$, $p_j \geq 0$, $q_j \geq 0$, $\sum_{j=1}^s p_j = 1$, and $\sum_{j=1}^s q_j = 1$. For the purpose of measuring the diversity of a distribution, we shall consider the following divergence measures:

1. The total variational distance (or the L_1 distance) [1, 6]:

$$D_V(\mathcal{P}, \mathcal{Q}) \quad = \quad \frac{1}{2} \sum_{j=1}^{s} |p_j - q_j|. \tag{8}$$

2. The Hellinger distance [23]:

$$D_H(\mathcal{P}, \mathcal{Q}) \quad = \quad \frac{1}{\sqrt{2}} \sum_{j=1}^{s} \left(\sqrt{p_j} - \sqrt{q_j}\right)^2. \tag{9}$$

3. The symmetric Kullback-Leibler (KL) divergence [14]:

$$D_{\text{SKL}}(\mathcal{P}, \mathcal{Q}) \quad = \quad \sum_{j=1}^{s} (p_j - q_j) \log_2 \frac{p_j}{q_j}. \tag{10}$$

4. The Bhattacharyya distance [3]:

$$D_B(\mathcal{P}, \mathcal{Q}) \quad = \quad -\log \left(\sum_{j=1}^{s} \sqrt{p_j q_j} \right). \tag{11}$$

5. The square root of Jensen-Shannon divergence [16]:

$$\begin{aligned} D_{\text{JS}}(\mathcal{P}, \mathcal{Q}) \quad &= \quad \sqrt{\tfrac{1}{2}\text{div}_{\text{KL}}(\mathcal{P}, \mathcal{Z}) + \tfrac{1}{2}\text{div}_{\text{KL}}(\mathcal{Q}, \mathcal{Z})}, \qquad (12) \\ \text{div}_{\text{KL}}(\mathcal{P}, \mathcal{Z}) \quad &= \quad \sum_{j=1}^{s} p_j \log \frac{p_j}{z_j}, \\ \text{div}_{\text{KL}}(\mathcal{Q}, \mathcal{Z}) \quad &= \quad \sum_{j=1}^{s} q_j \log \frac{q_j}{z_j}, \end{aligned}$$

where $\mathcal{Z} = \frac{1}{2}(\mathcal{P} + \mathcal{Q}) = \frac{1}{2}[p_1 + q_1, \ldots, p_s + q_s]^\top$ is the middle distribution for $\mathcal{P}$ and $\mathcal{Q}$, and div_{KL} is the directed KL divergence [14] between two distributions. All measures in Equations (8) – (12) have the following properties: (i) $\text{div}(\mathcal{P}, \mathcal{Q}) \geq 0$, (ii) $\text{div}(\mathcal{P}, \mathcal{P}) = 0$, (iii) $\text{div}(\mathcal{P}, \mathcal{Q}) = 0$ iff $\mathcal{P} = \mathcal{Q}$, and (iv) symmetry. Only D_H and D_{JS} satisfy the triangle inequality. Note that both D_H and D_B are derived from the Bhattacharyya coefficient $\Gamma(\mathcal{P}, \mathcal{Q}) = \sum_{j=1}^{s} \sqrt{(p_j q_j)}$, where $D_H = 1 - \Gamma(\mathcal{P}, \mathcal{Q})$, and $D_B = -\log \Gamma(\mathcal{P}, \mathcal{Q})$.

Given all the divergence measures in Equations (8) – (12), the diversity of any discrete distribution from $\{\mathcal{X}_i\}_{i=1}^{n}$ can be measured as follows:

1. Replace $\mathcal{P}$ in Equations (8) – (12) with $\mathcal{X}_i$.

2. Replace $\mathcal{Q}$ in Equations (8) – (12) with the reference distribution, whether it be $\mathcal{M}(\widehat{\boldsymbol{\pi}}^*)$, or $\mathcal{U}$ from Equation (7) if $\mathcal{M}(\widehat{\boldsymbol{\pi}}^*)$ is not available.

Since these particular divergences are analogous to distance measures, the smaller the divergence, the more diverse is the discrete distribution $\mathcal{X}_i$ with respect to the reference distribution of choice.

3.2 Properties of Divergence–based Diversity Measures

Consider now how the proposed approach for measuring diversity differs from entropy measures with regards to the three problems introduced in § 1 for comparing the diversity of multiple communities.

First, using the set Ω^*, we have a fixed unified set of species (or OTUs) for comparing all the samples. This eliminates one source of variation among all the samples, and renders the difference between samples to be based only on the difference between their distributions.

Second, it can be noticed from all the divergence measures in Equations (8) – (12) that, zero elements in any distribution $\mathcal{X}_i$ penalizes the divergence between $\mathcal{X}_i$ and the reference distribution (whether it be $\mathcal{M}(\widehat{\boldsymbol{\pi}}^*)$ or $\mathcal{U}$), and hence increases the divergence. This is unlike entropy measures which ignores these zero elements.

Third, except for D_{SKL} and D_{JS}, all other divergence measures do not impose any weighting scheme on the distribution $\mathcal{X}_i$. For D_{SKL} in Equation (10), the imposed weights $(p_j - q_j)$, are the differences between the probabilities for each outcome, which is maximized when the distributions are in complete disagreement, and zero when the distributions match. This weighting scheme penalizes the difference (or disagreement) between the two distributions. For D_{JS} in Equation (12), both distributions $\mathcal{P}$ and $\mathcal{Q}$

are compared against the middle distribution $\mathcal{Z}$. If $\mathcal{P}$ completely disagrees with $\mathcal{Z}$, the difference $\log(p_j/z_j) = \log p_j - \log z_j$ is maximum, and it penalizes the final divergence D_{JS}. A similar interpretation follows for $\mathcal{Q}$ and $\mathcal{Z}$. Here, it is important to note the difference between the weighting scheme for $\widehat{H}$ in Equation (1) on one hand, and that for D_{SKL} and D_{JS} on the other. In $\widehat{H}$, the weights are set to create a balance between rare and common species, and hence they alter the original distribution of the sample. However, in D_{SKL} and D_{JS} the weights penalize the disagreement (or the difference) between $\mathcal{X}_i$ and the reference distribution without altering any of them.

Divergence measures in general can be seen as distances between probability distributions. However, unlike distance metrics which have measurement units, in information theory, divergence measures do not have such units. Nevertheless, one cannot compare two different divergence values measured using two different divergence measures. At this point, one may ask whether there is a biological interpretation for the divergence measures presented here. Currently, from a statistical and information theoretic perspective, we cannot claim whether such an interpretation exist or not. If such an interpretation exists, it can be established by domain experts from each field through extensive analysis of these measures on their communities of interest.

Throughout the previous discussion we have always considered two reference distributions: (i) the latent species distribution $\mathcal{M}(\boldsymbol{\pi}^*)$, and (ii) the discrete uniform distribution $\mathcal{U}$. In principle, we believe that any community $\mathcal{C}$ has its own latent species distribution $\mathcal{M}(\boldsymbol{\pi}^*)$. If an estimate for this distribution is available, say $\mathcal{M}(\widehat{\boldsymbol{\pi}}^*)$, then one can use it as the reference distribution to measure the diversity of a community. Due to their definition, entropy-based measures do not enjoy such a flexibility. When $\mathcal{M}(\boldsymbol{\pi}^*)$ is not known, and hence $\mathcal{M}(\widehat{\boldsymbol{\pi}}^*)$ is not available, one has no other option but to use $\mathcal{U}$ as the reference distribution. Still, divergence-based measures will be better to use for the three reasons mentioned above. Another advantage of divergence-based measures is that they allow direct pairwise comparisons between all communities, which is not possible to compute using entropy-based measures.

Invariance of ranking among groups. When comparing the diversity of two communities, the ranking of the two communities should not be changed when a third community is added to the comparison. This is known as the invariance of ranking among groups. This property holds as well for divergence-based diversity measures under the condition that all groups have the same reference distribution. If the reference distributions changes for one

community, or for all communities, then the ranking among communities can change. Note that this is a natural consequence of changing the reference distribution for one or all communities, and hence it should not be considered a disadvantage of divergence-based diversity measures. Also note that it is not possible to compare the diversity of two or more communities with different reference distributions.

Monotonicity and principle of transfer. For a community $\mathcal{C}$ with multinomial distribution $\mathcal{M}(m, \boldsymbol{\pi})$, Patil and Taille [21] define the diversity of $\mathcal{C}$ as the average rarity $\delta(\mathcal{C}) = \sum_{i=1}^{m} \pi_i R(\pi_i)$, where $R(\pi_i)$ is the rarity of species i. For instance, for Shannon's index $R(\pi_i) = -\log(\pi_i)$, while for Simpson's index $R(\pi_i) = (1 - \pi_i)$. The rarity coefficient R should satisfy two requirements: *(i)* R is a nonnegative monotonic function, and *(ii)* R satisfies the principle of transfer; i.e. diversity increases if a new species is introduced to the community, and/or by making the distribution more even.

Monotonicity is satisfied by the definition of divergence according to [1], albeit in a different sense that suits the nature of probability distributions. Too see this, let div denote any of the previously mentioned divergence measures. Then, by definition of divergence [1], $\text{div}(\mathcal{P}, \mathcal{Q})$ is minimum when $\mathcal{P} = \mathcal{Q}$, and maximum when $\mathcal{P}$ and $\mathcal{Q}$ are orthogonal. Further, let θ be a real parameter, and $\{\mathcal{P}_\theta \; s.t. \; \theta \in (a, b)\}$ be a family of mutually continuous distributions on the real line, such that $\mathcal{P}_\theta$ has a monotone likelihood ratio[3]. Then, if $a < \theta_1 < \theta_2 < \theta_3 < b$, we have that $\text{div}(\mathcal{P}_{\theta_1}, \mathcal{P}_{\theta_2}) \leq \text{div}(\mathcal{P}_{\theta_1}, \mathcal{P}_{\theta_3})$. This property says that as the distance between the parameters (defining the distributions) increases, the divergence will increase as well. This property immediately applies to our multinomial distributions parameterized with $(m, \boldsymbol{\pi})$.

The principle of transfer, as explained above, has two aspects. The first is that adding a new species to the community should increase the diversity. This property holds for all the proposed divergence since they are sums of individual coefficients, each representing one species. The second is that increasing evenness should increase the diversity. This property also holds for the proposed divergence measures when the reference distribution is the uniform distribution. Increasing the evenness of a distribution makes it more similar to the uniform distribution, and hence decreases the divergence; i.e. increases diversity.

[3]Any two probability distributions $\mathcal{P}(x)$ and $\mathcal{Q}(x)$ have the monotone likelihood ratio property if for any $x_1 > x_2$, we have that $\mathcal{P}(x_1)/\mathcal{Q}(x_1) \geq \mathcal{P}(x_2)/\mathcal{Q}(x_2)$.

4 Application to Gut Microbiome Analysis

In this section we validate our propposed divergence–based diversity measures in a real world test case. To this end, we consider the application introduced in § 2.2, in which it is desired to assess the diversity of microbial consortia sampled from the feces of 24 infants stratified according to mode of delivery and diet, and see how it can deliver results that are consistent with the taxa profile reported in recent studies [25, 2]. The final data set that is used here is a table with 24 rows and 188 columns for the different species categories (or taxa in this case). Each entry in the table is the abundance (or counts) of a specific taxon (OTU) in each infant's gut. This table is sparse since relatively few taxa were recovered from each infant. Details of this study and the dataset can be found in [2].

Tables (2) and (3) report bacterial diversity of fecal samples according to mode of delivery and infant diet, respectively, with p values of significance tests. Similar to [2], all diversity measures used a rarefied dataset of 10,000 sequences (or individuals) per sample[4]. For statistical significance, we used the Kruskal–Wallis (KW) test (at $\alpha = 0.05$ level) since we do not assume normality of the data. All our implementation and experimental analysis were carried out using MATLAB from Mathworks. Statistical significance tests were carried out using R (`www.r-project.org`). Note that in the upper part of Tables (2) and (3), higher index values indicate higher diversity, while in the lower part, smaller divergence values indicate higher diversity.

4.1 Microbiome Diversity According to Delivery Mode

The upper part of Table (2) shows the diversity values (with standard deviation) and p values for the different delivery mode groups. All three entropy-based indices show a consistent trend; emergency CS–delivered infants have the highest diversity, followed by vaginally delivered infants, and the lowest diversity is for elective CS–delivered infants. In the lower part of Table (2), divergence based indices show a similar trend, however the KW test shows significant difference between emergency CS–delivered infants and vaginally delivered infants.

The results in Table (2) are worth careful consideration. Entropy-based

[4]When the total number of individuals in each sample $\mathbf{x}_i$ are not equal, stratified sampling (or rarefaction) is used to make all samples $\mathbf{x}_i$ have equal number of individuals. We used stratified sampling to generate 10,000 sequences (or individuals) per sample, and this process was repeated for 1000 trials. All reported diversity measures are averages over the 1000 trials.

Table 2: Diversity of fecal microbiota in infants (with standard deviation) according to mode of delivery. Bold numbers indicate statistically significant at $\alpha < 0.05$ level from Vaginal.

Index	**Elective CS** $n = 3$	**Emergency CS** $n = 3$	**Vaginal** $n = 18$	p **value**
Shannon's Entropy	1.09 (0.47)	2.01 (0.48)	1.33 (0.48)	0.088
Hill–Jost No.	3.22 (1.61)	8.03 (3.28)	4.17 (1.86)	0.088
Gini–Simpson	0.49 (0.23)	0.77 (0.15)	0.59 (0.21)	0.125
D_V	0.96 (0.01)	**0.9 (0.02)**	0.95 (0.02)	0.022
D_H	0.84 (0.02)	**0.75 (0.049)**	0.82 (0.04)	0.031
D_{SKL}	20.98 (0.37)	**19.03 (0.81)**	20.56 (0.83)	0.033
D_B	1.85 (0.12)	**1.39 (0.16)**	1.76 (0.24)	0.031
D_{JS}	0.79 (0.01)	**0.75 (0.02)**	0.78 (0.02)	0.023

measures show indeed differing diversity values for the three birth mode groups. However, due to the problems depicted in Figures (1) and (2a), these values cannot be compared to each other since we know from our data that the gut microbiota of individual infants is composed of few taxa with abundances greater than zero. This will result in communities with different taxa (or attributes) as depicted in Figures (1), and hence comparisons of entropy-based values do not permit unambiguous conclusions. In our proposed framework, divergence-based measures also differ by birth mode, however these values are able to be compared with each other. Recall that our framework, first, uses a unified set of taxa for all the samples, and second, divergence-based measures by definition, penalize taxa with zero abundances and not discarded[5] as in the case of entropy-based measures. That is, even if one uses a unified set of taxa for all samples, entropy-based measures cannot handle zero count taxa as depicted in Figure (2a), and this where divergence-based measures have an edge over entropy-based measures. From the definition of divergence, it suffices to have a difference in one taxon's relative abundance to change the value of divergence. Whether statistically significant or not, differences in divergence-based measures indicate that there are differences in taxonimic distributions. The fact that this difference is statistically significant is a strong evidence that at least two groups have different taxonomic distributions.

The key advantage of using a unified set of taxa and divergence measures (instead of entropy), is that it creates a direct relation between the relative abundance of any taxon and the value of divergence. This is particularly

[5]Due to $0 \log 0$ as depicted in Figure (2a)

important when computing the average diversity over multiple samples, or infants, as is the case studied here. Any change in the relative abundance of any taxon in any sample, is detected and quantified by any of the above divergence measures. Unfortunately, this is not the case for entropy-based measures, again due to the problems in Figures (1) and (2a). That is, while one can observe differences in the taxon's relative abundance across samples, these differences might not be uniformly detected in comparisons of entropy-based measures of these samples. For instance, while [2] found no correlation between the Shannon's index of total diversity and the relative abundance of Bacteroidetes taxa in CS–delivered infants, [12] observed lower total gut microbiota diversity, as well as lower Bacteroidetes diversity in their CS–delivered infants.

4.2 Microbiome Diversity According to Infant Diet

The upper part of Table (3) shows the diversity values (with standard deviation) and p values for the different diet groups. All three entropy-based indices show a consistent trend; exclusively formula-fed infants have the highest diversity, followed by partially breast-fed infants, and the lowest diversity is for exclusively breast-fed infants. In the lower part of Table (3), divergence-based indices show a similar trend, however the KW test shows significant difference between exclusively breast-fed infants and exclusively formula-fed infants. The results' interpretation for Table (3) is similar to the discussion above for Table (2). The values for divergence-based indices can be compared to each other, and they show that there are differences in the taxonomic distributions among these groups. The statistical significant difference is a strong evidence that at least two groups have different taxonomic distributions.

4.3 Divergence–based Diversity Measures as Biomarkers

In this work, diversity indices have been viewed as measures for quantifying the difference (or discrepancy) between two probability distributions. Entropy–based indices measure this difference in terms of similarity between a given distribution and the uniform distribution, while divergence–based indices measure the difference between any two given distributions in a similar fashion to distances between points. We also introduced the notion of a reference distribution; we believe that any community of organisms has a specific species distribution. If a sample of species is randomly drawn from this community, then diversity measures how similar is the sample's species

Table 3: Diversity of fecal microbiota in infants (with standard deviation) according to infant diet. Bold numbers indicate statistically significant at $\alpha < 0.05$ level from Exclusive FF.

Index	**Exclusive BF** $n = 10$	**Partially BF** $n = 5$	**Exclusive FF** $n = 9$	p **value**
Shannon's Entropy	1.19 (0.51)	1.42 (0.64)	1.57 (0.47)	0.369
Hill–Jost No.	3.64 (1.56)	4.9 (3.39)	5.32 (2.45)	0.369
Gini–Simpson	0.56 (0.24)	0.6 (0.24)	0.64 (0.17)	0.719
D_V	**0.96 (0.02)**	0.94 (0.03)	0.93 (0.02)	0.009
D_H	**0.84 (0.03)**	0.81 (0.05)	0.79 (0.04)	0.049
D_{SKL}	20.91 (0.79)	20.32 (1.13)	19.93 (0.79)	0.084
D_B	**1.86 (0.25)**	1.69 (0.25)	1.59 (0.18)	0.049
D_{JS}	**0.79 (0.01)**	0.78 (0.02)	0.77 (0.01)	0.018

distribution to the species distribution of the original community. Based on this view, it is possible that diversity indices can act as biomarkers for health risks. To do this, diversity needs to be put in a precise statement that includes the following information. Here, we shall only consider divergence–based diversity measures.

1. The set of species (or OTUs) based on which the diversity is measured, or Ω^* in our case.

2. The reference distribution $\mathcal{M}(\widehat{\boldsymbol{\pi}}^*)$, or $\mathcal{U}$ defined over Ω^* when the earlier is not available.

3. A symmetric divergence measures, div for instance.

4. A specified range $\widehat{\mathcal{R}}_{\text{div}} = [t_1, t_2]$ that is dependent on the particular choice of divergence used, where $0 \leq t_1 < t_2$. Without loss of generality, we have assumed a reasonable sample size to estimate $\widehat{\mathcal{R}}_{\text{div}}$.

The estimate $\widehat{\mathcal{R}}_{\text{div}}$ defines the range for div values of normal and healthy cases. If the estimated diversity of a community under study in not within the range of $\widehat{\mathcal{R}}_{\text{div}}$, then this indicates a significant disruption in the species (or OTUs) distribution and signals a possible health risk. Any statement that includes this information gives a precise context for the measured diversity, and makes it possible to use it as a biomarker for health risks.

5 Concluding Remarks

We have proposed a new framework for measuring communities' diversity based on divergence measures. The framework overcomes various shortcomings in entropy-based measures, and is more accurate in detecting changes in species' distribution. Although we have used the framework in the analysis of infants gut's microbiome, the framework is not restricted to microbiome analysis and can be used for comparing diversities of any types of communities.

References

[1] S. Ali and S. Silvey. A general class of coefficients of divergence of one distribution from another. *J. of the Royal Statistical Society.* Series B, 28(1):131–142, 1966.

[2] Meghan Azad, Theodore Konya, Heather Maughan, David Guttman, Catherine Field, Radha Chari, Malcolm Sears, Allan Becker, James Scott, and Anita Kozyrskyj. Gut microbiota of healthy Canadian infants: profiles by mode of delivery and infant diet at 4 months. *Canadian Medical Association J.*, 185(5):385–394, 2013.

[3] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–109, 1943.

[4] Anne Chao. Nonparametric estimation of the number of classes in a population. *Scandinavian J. of Statistics*, 11(4):265–270, 1984.

[5] Anne Chao and Tsung-Jen Shen. Nonparametric estimation of shannons index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics*, 10(4):429–443, 2003.

[6] I. Csiszár. Information–type measures of difference of probability distributions and indirect observations. *Studia Scientiarium Mathematicarum Hungarica*, 2:299–318, 1967.

[7] R. Fisher, A. Corbet, and C. Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *J. of Animal Ecology*, 12(1):42–58, 1943.

[8] M. Hill. Diversity and evenness: A unifying notation and its consequences. *Ecology*, 54(2):427–432, 1973.

[9] Tom Hill, Kerry Walsh, James Harris, and Bruce Moffett. Using ecological diversity measures with bacterial communities. *FEMS Microbiology Ecology*, 43(1):1 – 11, 2003.

[10] Cang Hui, Ruan Veldtman, and Melodie A. McGeoch. Measures, perceptions and scaling patterns of aggregated species distributions. *Ecography*, 33(1):95–102, 2010.

[11] S. Hurlbert. The nonconcept of species diversity: A critique and alternative parameters. *Ecology*, 52(4):577–586, 1971.

[12] Hedvig Jakobsson, Thomas Abrahamsson, Maria Jenmalm, Keith Harris, Christopher Quince, Cecilia Jernberg, Bengt Bjrkstn, Lars Engstrand, and Anders Andersson. Decreased gut microbiota diversity, delayed bacteroidetes colonisation and reduced th1 responses in infants delivered by caesarean section. *Gut*, 63(4):559–566, 2014.

[13] Lou Jost. Entropy and diversity. *Oikos*, 113(2):363–375, 2006.

[14] S. Kullback. *Information Theory and Statistics – Dover Edition*. Dover, 1997.

[15] Kelvin Li, Monika Bihan, Shibu Yooseph, and Barbara Meth. Analyses of the microbial diversity across the human microbiome. *PLoS ONE*, 7(6), 2012.

[16] J. Lin. Divergence measures based on the Shannon entropy. *IEEE Trans. on Information Theory*, 37(1):145 – 151, Jan 1991.

[17] R. MacArthur. Fluctuations of animal populations, and a measure of community stability. *Ecology*, 36(3):533–536, 1955.

[18] D. Margalef. Information theory in ecology. *Gen. Sys.*, 3:36–71, 1958.

[19] Melodie McGeoch and Kevin Gaston. Occupancy frequency distributions: patterns, artefacts and mechanisms. *Biological Reviews*, 77:311–331, 8 2002.

[20] Claudia Moreno and Pilar Rodriguez. A consistent terminology for quantifying species diversity? *Oecologia*, 163(2):279–282, 2010.

[21] G. Patil and C. Taillie. Diversity as a concept and its measurement. *J. of the American Statistical Assoc.*, 77(379):548–561, 1982.

[22] E. Pielou. The use of information theory in the study of the diversity of biological populations. In *Proc. of 5th Berkeley Symp. Math. Stat. Prob.*, volume 4, pages 163–177, 1967.

[23] C. Rao. Use of Hellinger distance in graphical displays. In *Multivariate Statistics and Matrices in Statistics*, pages 143–161, 1995.

[24] Alfréd Rényi. On measures of entropy and information. In *Proc. of the 4th Berkeley Sym. on Math., Stat. and Prob.*, pages 547–561, 1960.

[25] José Saavedra and Anne Dattilo. Early development of intestinal microbiota: implications for future health. *Gastroenterol Clinincal of North America*, 41(4):717–731, 2012.

[26] Claude Shannon. A mathematical theory of communication. *The Bell System Technical J.*, 27:379–423, 623–656, 1948.

[27] E. Simpson. Measurement of diversity. *Nature*, 163:688, 1949.

[28] Hanna Tuomisto. A consistent terminology for quantifying species diversity? yes, it does exist. *Oecologia*, 164(4):853–860, 2010.

[29] Hanna Tuomisto. A diversity of beta diversities: straightening up a concept gone away. part 1. *Ecography*, 33(1):2–22, 2010.

[30] R. Whittaker. Vegetation of the Siskiyou mountains, Oregon and California. *Ecological Monographs*, 30(4):p. 407, 1960.

[31] R. Whittaker. Evolution and measurement of species diversity. *Taxon*, 21(2/3):213–251, 1972.

[32] Harm Wopereis, Raish Oozeer, Karen Knipping, Clara Belzer, and Jan Knol. The first thousand days intestinal microbiology of early life: establishing a symbiosis. *Pediatric Allergy and Immunol.*, June 2014.