Improving interpretation of microbiome studies by incorporating phylogenetic information

Cédric Arisdakessian *

University of Hawai'i at Mānoa

December 2020

Abstract

To understand the intricate link between microbes and their environment, metagenomic analyses need to account for dependencies between related organisms as described by a phylogenetic tree. The induced correlation structure, alongside the high diversity of organisms in metagenomic data, represents a challenge when identifying microbial drivers of an environmental factor. Recently, metagenomic analysis tools have been developed to conjointly analyze microbial counts in multiple environments with their phylogeny, by either transforming the data into a unified data structure or by including the phylogeny into a comprehensive modeling framework. These methods naturally handle dependencies between organisms and increase interpretability by organizing features into homogeneous groups, which can generate biologically meaningful hypotheses. In this review, we describe and compare these methods as well as discuss opportunities for improvements.

*Electronic address: carisdak@hawaii.edu

Introduction

Dynamics of microbial populations are driven by the constraints of the environments they inhabit. Thus, through reverse-engineering, we can learn about an environment's properties by studying its microorganisms and their abundance. There are multiple ways to measure microbial composition and abundance experimentally. The most popular approach is to collect a biological sample (earth, water, dust,...), and sequence one "marker" gene present in every organism. Like a barcode, the specific DNA sequence of this marker gene can uniquely identify each species. The abundance of each organism is then derived by counting the number of copies corresponding to each unique sequence. Because those sequences do not always exactly match species definitions, they are referred to as Operational Taxonomic Units (OTUs) in the literature, a term that we will use throughout this paper.

Metagenomic data contains two main types of information (see Figure 1). The first is the taxonomy of an organism (its identity), which is typically broken down into multiple hierarchical ranks ranging in scale from 3 domains (Bacteria, Archaea, Eukarya) to millions of species. While some organisms (like Escherichia coli) have been extensively studied and have a completely resolved taxonomy, many others lack taxonomic information, which complicates taxonomy-based analysis. Alternatively, the relationship between OTUs can also be described with a phylogenetic tree (also called a phylogeny), which models the evolution process that resulted in the observed microbial population. A phylogenetic tree is a binary tree where the leaves correspond to OTUs and each node corresponds to a putative common ancestor of the leaves it contains. From a biological perspective, a node represents an ancestral lineage (a clade) where the children share similar functions. Edges in the tree are typically weighted to provide a measure of evolutionary distance between clades. Unlike the taxonomy, the tree inference does not usually depend on databases nor is it restricted to a fixed number of taxonomic ranks. Its inference usually relies on evolutionary mutation models between the input DNA sequences. Therefore, it is a more accurate representation of relationships between organisms as compared to the taxonomy. To facilitate its integration in analyses, the phylogenetic tree is sometimes converted into a patristic distance matrix, where the entry (i, j) is the length of the shortest path between OTUs i and j in the tree. Although this transformation does not preserve all the information in the tree, it is a convenient way to capture evolutionary distance between OTUs.

The second type of metagenomic data is the abundance table, which counts the number

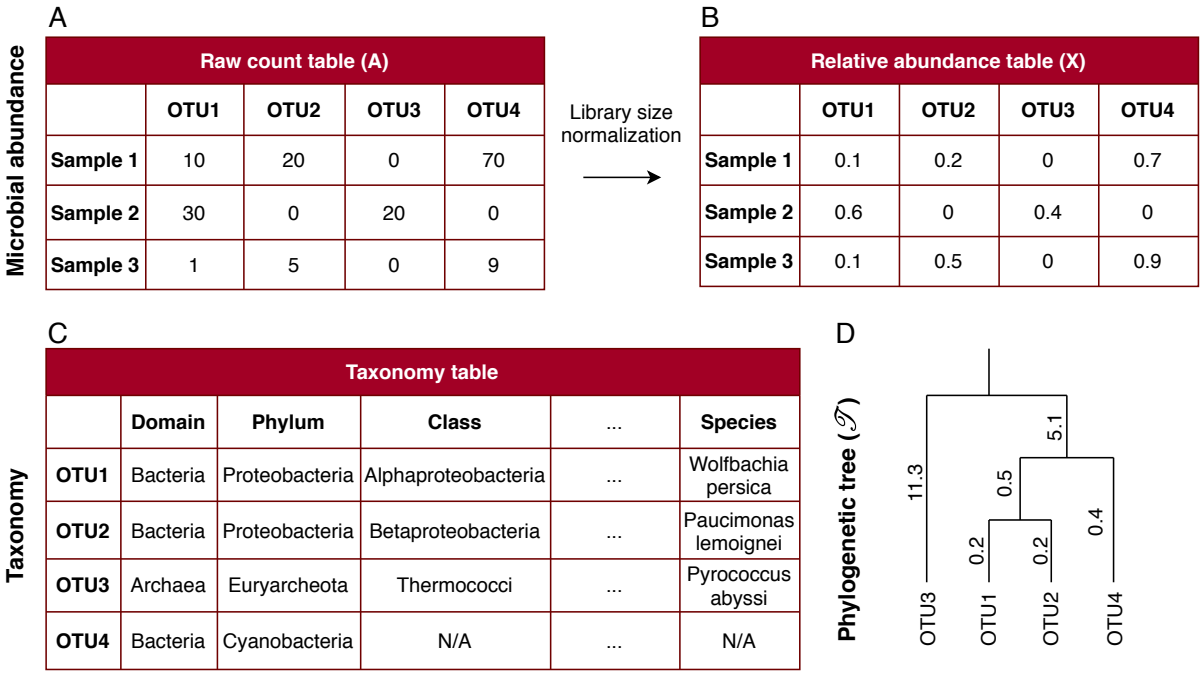


Figure 1: Metagenomic data types. When provided, letters associated with a data structure correspond to the corresponding notation in the review. (A) Raw abundance data. Each row is a different sample and each column is an OTU (proxy for species) (B) Corresponding relative abundance tables where each row is normalized by its sum. (C) Taxonomic assignment table. Each OTU is assigned to a category for each of seven taxonomic ranks. N/A represent missing information. (D) Phylogenetic tree. The edges are labeled with the phylogenetic distance.

of occurrences of each OTU in each sample. The total count for each sample (the library size) varies due to systematic biases, such as the number of sequences sampled. To prevent making erroneous assumptions when comparing samples, the values are routinely converted to frequencies by dividing each sample by its library size. Thus, if we have p OTUs across n samples, the abundance data consists of n points on the simplex of dimension $p-1$, which is defined as:

$$S_p = \left\{ x_1, x_2 ..., x_p \in [0,1] \;\; | \;\; \sum_{i=1}^{p} x_i = 1 \right\}$$

In biological terms, S_p is the set of all possible relative abundances over p OTUs — each element of S_p is a potential sample.

Most analyses also involve one or multiple environmental variables that are being studied in the light of microbes. These variables can be either continuous or categorical. For example, one can study the differences between microbial communities in volcanic rift zones on the island of Hawaii compared to non-volcanic areas (categorical variable) or study the communities in an environment composed of a gradient of sulfur (continuous variable).

The analysis of microbiome data poses a number of challenges. First, the analysis of compositional data (i.e. points on the simplex) is inherently complex: the components of each point compete against the others to sum up to 1. It constrains the variance-covariance structure of the variables to lie in a narrow range which can create spurious correlations between variables [1]. Standard statistical tools such as Pearson's correlations and t-tests are known to be invalid in the simplex simply because the variable independence hypothesis is violated. In addition, the distribution of bacterial counts are highly skewed and are usually modeled using negative binomial distributions. The count table generally contains a significant amount of zeros, which is either due to the fact that some OTUs are missing from some environments, or that the OTUs were undetected due to an insufficient library size. Those constraints require more complex models, making the analysis more challenging.

Another challenge in the analysis of metagenomes is the integrated analysis of the OTUs abundance table and the phylogenetic tree, two data structures that are not routinely modeled together. Phylogeny is a source of dependence between OTUs since the abundance of closely related organisms will likely be correlated. The phylogeny provides additional information about the OTUs relatedness and thus helps identify the relevant OTU groups correlated with an outcome. Moreover, the inclusion of phylogenetic information makes it possible to analyse the abundance data at multiple phylogenetic scales by comparing the aggregated nodes' abundance in the tree [2].

Here, we focus on methods that use jointly the abundance table and the phylogenetic tree to quantify microbial contribution in a microbiome study. We organize this review in two sections. First, we present methods that combine abundance and the phylogeny to define new metagenomic features. Second, we present the machine learning models incorporating the tree structure in the analysis either by using regularization in an optimization problem or by using hierarchical models that explicitly describe the phylogenetic dependencies. There are no available results that compare these methods directly. This is principally due to the fact that the ground truth about the relation between microbes and the environment is still predominantly unknown. Thus, each method uses validation strategies that cannot be easily extrapolated to other methods. As such, we will focus in this review on comparing these methods from a conceptual perspective i.e. contrasting which issues they address or ignore. Each method's characteristics are summarized at the end of this paper in table **??**.

Notations

In the rest of this review, we use the following naming conventions:

- $A \in \mathbb{N}^{n \times p}$ is the raw count matrix with n samples and p OTUs
- $X \in \mathbb{R}^{n \times p}$ is the relative abundance matrix where each row sums up to 1
- $\mathcal{T}$ is the phylogenetic tree corresponding to the OTUs

In addition we use the following mathematical notations:

- $\mathbf{1}_p$ is the unit column vector of length p
- $\forall x \in \mathbb{R}^k, g(x) = \left(\prod_{i=1}^{k} x^k\right)^{1/k}$ (g is the geometric mean function)
- $||.||_k$ is the k-norm in $\mathbb{R}^p$ defined as $||x||_k = \sqrt[k]{\sum_{i=1}^{p} x_i^k}$
- For a given vector x, we note $x^+ = \sum_i x_i$
- $S_p = \{x_1, x_2..., x_p \in [0,1] \ | \ \sum_{i=1}^{p} x_i = 1\}$ is the $p-1$ dimensional simplex

1 Phylogenetic-aware transformation of the OTU abundances

One possible approach to integrate phylogenetic structure in the analysis is to derive new features that aggregate both abundance and phylogenetic information. The methods surveyed in this section naturally solve compositionality issues by shifting the problem from the simplex to the real space and expanding the abundance table definition from an OTU perspective — the leaves of the tree — to the full tree. These methods facilitate subsequent analyses by allowing the use of existing statistical methods while enhancing interpretability since the derived features directly refer to placements of biologically meaningful OTU grouping along the phylogenetic tree. A node-centered data transformation assumes that the variations in a response variable can be explained by competing sister sub-trees at the same phylogenetic scale. In other words, it assumes that differences arise from one or multiple branching events where the descendents in each branch capture the variation in the response. In contrast, an edge-centered data transformation assumes that differences arise along an edge, and therefore, a given clade differentially expresses the response compared to the rest of the tree. These approaches are described in what follows.

1.1 Log-ratio based approaches

Two methods to solve the compositionality issue were proposed by Aitchison [3] in 1982. These methods introduced algebraic notions in the simplex and used log-ratios to transfer the problem from the simplex to the real space. These methods, termed additive log-ratio (ALR) and the centered log-ratio (CLR), are defined as follows:

$$\begin{array}{rll} & S_p & \to \mathbb{R}^{p-1} \\ \texttt{alr}: & & \\ & x_1, ..., x_p & \to \ln\left(\frac{x_1}{x_p}\right), ..., \ln\left(\frac{x_{p-1}}{x_p}\right) \end{array} \tag{1}$$

$$\begin{array}{rll} & S_p & \to \mathbb{R}^{p} \\ \texttt{clr}: & & \\ & x_1, ..., x_p & \to \ln\left(\frac{x_1}{g(x)}\right), ..., \ln\left(\frac{x_p}{g(x)}\right) \end{array} \tag{2}$$

Aitchison proved that these simple transformations render possible the use of standard statistical analyses. His work was later expanded by [4] where the authors define an inner product in the simplex and prove its hilbertian structure — a vector space with an inner-product, allowing for distance and angle measurements. However, those transformations suffer some shortcomings. First, the expression of ALR depends on an arbitrary reference (x_p in equation 1) which makes the transformation less robust to outliers in the data compared to CLR. Second, these transformations both rely on a compositional vector expressed in the standard basis of $\mathbb{R}^p$, which is not orthonormal in the simplex geometry (using Aitchison distance). This causes the angles to be distorted when the data is transformed from the simplex to the real space and hinders the interpretability of the results [5]. The Isometric Log-Ratio transform (ILR) [5] addresses those issues by re-expressing the compositional vector in an Aitchison-orthonormal basis of S_p. The new coordinates of a vector x in this basis (a.k.a. the ILR coordinates) are obtained by recursively partitioning the space and calculating the projection of x with each partition's basis element (see [5] for more details about the full expression of the basis). In practice, the k^{th} ILR coordinate of x is the log-ratio between the geometric means of the coordinates of x on partitions I and J for the k^{th} iteration (see Figure 2)

$$[\texttt{ilr}(x)]_k = \sqrt{\frac{|I| + |J|}{|I| \times |J|}} \cdot \log\left(\frac{g(x_I)}{g(x_J)}\right) \tag{3}$$

Since the ILR components contrast values between two partitions, they are usually referred

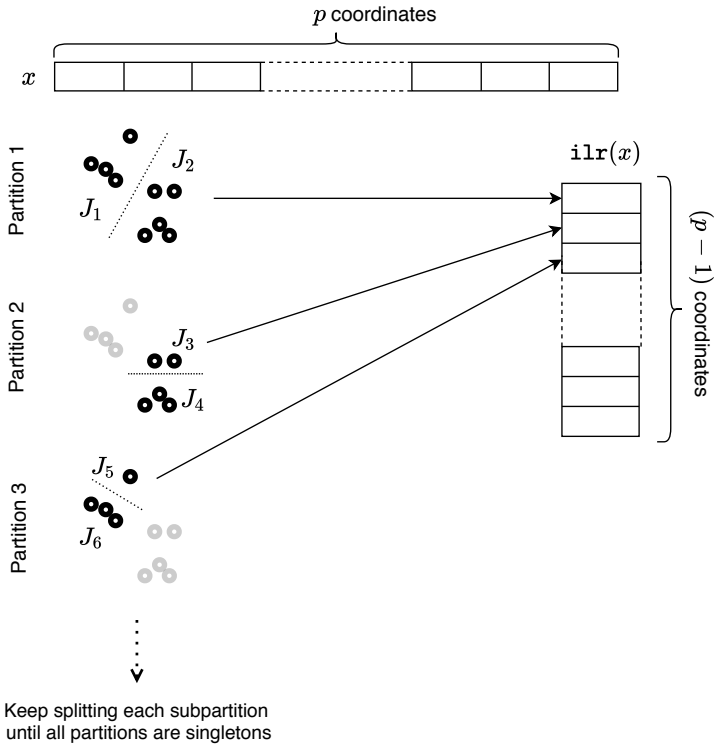


Figure 2: Computation of ILR coordinates of a sample x based on a sequential binary partition $J_1, .., J_{p-1}$. Each circle represents an OTU (a coordinate of x) and the dotted lines separating points represent the partition boundaries. Faded points are data points set aside for the calculation of a given ILR coordinate. The dataset is recursively partitioned in two, and for each partition, a new coordinate of ILR is computed. The procedure stops when all partitions are singletons.

to as "balances". In the rest of this section, we will use interchangeably the terms "ILR components" and "balances" to designate the coordinates of a compositional vector in the ILR basis. Although the geometric properties of ILR are more intuitive compared to the ALR and CLR, the choice of the binary partition is crucial for interpretability. Hence, methods relying on ILR need to partition the variable space in a biologically meaningful way.

Many approaches make use of the phylogenetic tree as a natural way to partition the compositional space. The transformation usually results in an alternative view of the abundance table, from a samples-by-OTUs matrix to a samples-by-tree-features matrix, where the features are either the phylogenetic tree's nodes or edges. For example, Gneiss [6] adopts a node-centered approach to define a sequential binary partition. A binary tree naturally defines such a partition by considering the two sets of leaves (OTUs) associated with each node's two subtrees. For each sample, the balance of a node v can be computed as described in Equation 4:

$$\texttt{ilr}(v)_s = \sqrt{\frac{|v_L| \cdot |v_R|}{|v_L| + |v_R|}} \cdot \log\left(\frac{g(x_{v_L}(s))}{g(x_{v_R}(s))}\right) \tag{4}$$

where: $\begin{cases} |v_L| \text{ and } |v_R| \text{ are the number of leaves in the left and right subtrees of node } v \\ x_{v_L}(s) \text{ and } x_{v_R}(s) \text{ are the relative abundances in each subtree of } v \text{ in sample } s \end{cases}$

Because the ILR is not defined when null values are present, the authors added a pseudo-count of 1 to all abundance before normalization. This results in $p-1$ ILR components (or balances) for each node in the tree. PhILR [7] is also a node-centered approach with the same partition scheme. However, it extends the expression of the ILR transform in order to account for two characteristics of metagenomic data. First, low abundance OTUs are usually less reliable because they could be the result of sequencing errors or correspond to rare organisms that were identified in samples with large total counts. Therefore, PhILR adopts a heuristic soft-thresholding approach by downweighting low abundance OTUs. The OTU weight is the combination of 2 factors: 1) the euclidean norm of its relative abundances and 2) the geometric mean of its raw abundance counts. Second, although the ILR takes the tree topology into account, it does not incorporate evolutionary distance (i.e. branch length in the tree). To address this issue, PhILR incorporates branch length information by adjusting the ILR value of each node, which it does by multiplying its value by the square root of the sum of the branches length to its direct descendents.

The two previous methods provide a very interpretable view of the data since significant features — the ILR components of the nodes — are directly linked to locations in the tree. However, both methods rely on two limiting assumptions. First, the ILR can only be computed for a binary partition, and not for nodes with polytomies (i.e. more than two direct descendents). Second, a node-centered approach captures variations in the response variable between clades at the same phylogenetic scale, which limits the potential of the analysis.

To overcome the limitation described above, an edge-based approach was proposed [8]. This approach assumes differences arise along an edge, and communities on each side of an edge are maximally different with regards to a response variable. Phylofactorization [8] adopts this edge-based partitioning scheme and computes the balances using the OTUs on each side of an edge. Unlike the node-centered approach, the recursive selection of edges to partition on is arbitrary. Phylofactorization's approach consists of recursively splitting the tree along edges that maximize an objective function. The choice of an appropriate objective function is left to the user, but Washburne et al. provide an example when predicting a set of response variables. These variables are used as features of a generalized linear model (GLM) to predict the balances

of a node such that:

$$\texttt{ilr}_\texttt{e} \sim \texttt{GLM}(y_1, ..., y_q), \text{ where: } \begin{cases} \texttt{ilr}_\texttt{e} \in \mathbb{R}^n \text{ are the balances for edge } \texttt{e} \\ y = [y_1, ..., y_q] \in \mathbb{R}^{n \times q} \text{ are response variables} \end{cases}$$

For each iteration, the edge that maximizes the amount of variance explained by the model is selected to partition the variables. Phylofactorization is expanded in [9] by generalizing it to take arbitrary objective functions, other operators for aggregating the values of node in each of the subsets (equivalent to geometric mean in equation 3) and operators for contrasting the values aggregated independently in each subset (this is equivalent to the log-ratio in eq. 3). Washburne et al. show how phylofactorization can be used for a wide range of data types and applications through an adaptation of one or multiple of these three operators. One example is their phylogenetic component analysis, PhyCA, which is an unsupervised version of Phylofactorization. Its implementation relies on the choice of an alternative objective function, where edges are selected based on the variance of their ILR value. Gappa [10] goes a step further by generalizing PhILR [7] and Phylofactorization [8, 9] to take into account uncertainties in OTU positions in the tree. This method is best suited when the phylogenetic tree is constructed using a phylogenetic placement approach, meaning that the OTUs of interest are placed on an existing high-quality tree as opposed to being used to build a phylogenetic tree. These types of approaches not only position the OTU at the most likely position, but usually provide a probability distribution for all possible placements in the tree. The ILR transform is applied as described in the previous works, except that the abundance values are weighted by their respective probabilities.

All the methods described above address multiple challenges posed by metagenomic analyses transforming OTU abundance from the simplex to the real space. ILR has been criticized for its lack of interpretability due to its reliance on a partitioning scheme. However, the phylogenetic tree induces a natural partition of the variable space in a biologically meaningful way. Each feature can be directly mapped to a tree feature which actually improves interpretability. In addition, ILR solves the problems of data compositionality and diversity at the same time by integrating both feature types — the abundance table and the phylogeny — in a single data structure. As a result, a user can use any statistical tool to compare sample groups, and directly identify the group of organisms responsible for observed differences in some response variable.

One caveat of log-ratio approaches is that they are not defined when zeros are present. The above methods address this issue by adding a pseudocount of 1 to the raw abundances. However, in [11], the authors point out the fact that in this case, dividing by the geometric mean (e.g. in the ILR) of highly sparse data is equivalent to not normalizing at all. The value of the pseudocount itself is subject to controversies and its value is known to affect analyses [12].

1.2 PCA based approaches

The Principal Components Analysis (PCA) is a very popular method to summarize high dimensionality data. PCA consists in finding an orthonormal basis of the variable space representing the direction of maximum variability ordered from highest to lowest. This transformation is linear, which simplifies its interpretation. However, PCA suffers some shortcomings when applied to metagenomic data. Indeed, the variation in the data is often driven by multiple interacting features, which causes the principal components to be the combinations of hundreds of OTUs. This hinders its interpretability since there is no biologically meaningful group of OTUs being selected. This can be solved by leveraging the hierarchical organization of features (the OTUs), which is not handled by regular PCA. This section surveys methods that integrate phylogenetic constraints into the PCA optimization problem to find more interpretable principal axes, which can be achieved by enforcing sparsity of the principal axes and constraining their nonzero values on contiguous parts of the tree.

Similar to Phylofactorization [8, 9], EdgePCA [13] is a decomposition method based on the tree edges. Instead of using a recursive binary partition [8, 9], edgePCA defines independent binary partitions for each edge in the tree as the OTUs on each side of an edge. Edge features are computed for each partition as the difference between the total abundance of OTUs on each side of an edge, starting with the root side of the tree. EdgePCA was developed in the context of phylogenetic placement for which uncertainties about OTU location in the tree are available. In this case, edge features are computed by weighing the abundance of each OTU by its placement probability. This transformation generates a samples-by-edges matrix that is then decomposed using standard PCA. The advantage of this method is that it provides highly interpretable results since components are a linear combination of the tree edges. However, Matsen et al. do not address the problem of compositionality, which can sometimes generate artifacts on PCA as discussed in [14].

In [15], Purdom takes a different approach, and integrates the phylogenetic tree into the

optimization problem defined by the PCA. Purdom's work builds on a previous generalization of PCA (gPCA) for non-standard metrics in $\mathbb{R}^p$. Like PCA, gPCA finds a set of p orthonormal axes that optimally explain the variance in the data. However, gPCA differs from PCA because it relies on alternative measurements of distance, which affects the dispersion of the data points. Purdom's idea is to incorporate external knowledge (the phylogenetic relationships between OTUs) by altering the distances between points so that closely related OTUs would appear closer than they would be if they were considered independent (as done in PCA). The chosen metric can be described with a pairwise distance matrix between each pair of OTUs. Using the identity matrix as a distance matrix results in a PCA (all points in the basis are equidistant). In the case of Purdom's gPCA, distances between OTUs are altered by their phylogeny, causing the principal axes to share little phylogenetic similarity with each other. The entry (i, j) of the phylogenetic distance matrix Φ is defined in [15] as the distance between the root and OTUs i and j's immediate parent node.

The algebraic structure of matrix Φ is critical for interpreting the principal components. For example, if Φ's eigenvectors are sparse, then the principal axes correspond to small sections of the tree, therefore enhancing interpretability. Moreover, different levels of sparsity in the principal axes correspond to different phylogenetic scales — sparse axes are small scale features while dense axes are global features. Although these properties have not yet been fully characterized, empirical observations shows that Φ has a block structure, suggesting it could provide a multiscale view of the data.

Both edgePCA and gPCA integrate the phylogenetic dependencies in the PCA algorithm in a different way. In the case of edgePCA, the data is transformed to yield a sample-by-edge matrix that can be subsequently used with the PCA. For the gPCA, the distance metric is modified so that PCA can adapt itself to the unique structure of the data. One assumption when using PCA is that linear combination between features is sufficient to describe the correlation structure of the data, which is not necessarily the case for metagenomes where nonlinear patterns are present [16].

2 Phylogeny-aware machine learning models

In contrast to the previously described models where the main focus was to find an alternative representation of the data, this section describes phylogeny-aware machine learning models

where the objective is to identify a set of features correlated with one or multiple outcomes. For instance, one could be interested in studying how microbial communities are affected by both seasonality and geographical location. Throughout this section, we will refer to the q observed outcomes in each sample as the matrix Y of dimension (n, q) (or the vector y of dimension n if a single outcome variable is considered).

2.1 Tree-penalized linear regression models

Linear regression models were developed to explain a response variable based on a set of features. The regression coefficients can be uniquely determined using linear algebra when the number of features is lower than the number of instances. However, this is rarely the case in metagenomics studies where the OTUs generally outnumber the samples. Penalized regressions solve this issue by converting the linear algebra problem into an optimization problem with additional constraints to enforce sparsity of the coefficients. Amongst the most popular methods, the Lasso regression penalizes the $||.||_1$-norm of the coefficients to shrink them to zero while the Ridge regression (using the $||.||_2$-norm) tends to make them small. The approaches in this section use different versions of penalized regression to constrain the solution space in a way that leverages the phylogeny.

ssCCA [17] is an adaptation of the popular canonical correspondence analysis (CCA [18]) decomposition and includes prior knowledge about phylogenetic structure. CCA can be seen as a supervised version of PCA: Instead of finding linear combinations of p variables that maximize the explained variance, CCA finds linear combinations of p variables that are best correlated to linear combinations of q response variables. The computation of the k^{th} pair of canonical vector (u_k, v_k) is the solution to the following problem:

$$(u_k, v_k) = \text{argmax}_{u,v} \left(\texttt{corr}(u^T x, v^T y)\right) \text{ s.t. } \begin{cases} \text{(Unit norm)} & u^T \Sigma_X u = 1 \text{ and } v^T \Sigma_Y v = 1 \\ \text{(Orthogonality)} & \forall i < k, u_i \perp u_k \text{ and } v_i \perp v_k \end{cases}$$

with $\Sigma_X = \texttt{var}(X)$ and $\Sigma_Y = \texttt{var}(Y)$

For high-dimensionality data, CCA lacks interpretability since it does not perform any variable selection and can therefore result in canonical axes being a combination of hundreds of variables. As a result, extensions of CCA (such as the sparse CCA, a.k.a. sCCA) include a Lasso penalty on u and v to shrink some of the coefficients to 0. ssCCA further expands these

methods with yet another penalty to constrain the variable selection.

$$\text{(Sparsity penalty)} \qquad ||u||_1 < \mathtt{c}_1 \text{ and } ||v||_1 < \mathtt{c}_2$$

$$\text{(Smoothness penalty)} \qquad \sum_{1 \leq j < k \leq p} \frac{1}{d_{j,k}^2} |u_j - u_k|^2 < \mathtt{c}_3$$

where:
- $\mathtt{c}_1$ and $\mathtt{c}_2$ are regularization constants
- $d_{j,k}$ is the patristic distance between OTU j and k

The smoothness penalty shares some similarity with a fused-Lasso [19] ridge regularization, which makes the coefficients smooth in regard to phylogenetic distance. The assumption is that closely related OTUs will likely share a similar effect and should therefore have similar coefficients. Chen et al. mention multiple options for the measure of phylogenetic distance between OTUs, but eventually argue that any sensible choice should provide similar results.

In [20], Wang and Zhao develop a penalized linear model to predict a continuous outcome y. Rather than smoothing the regression coefficients of closely related OTUs, the regularization is used at multiple phylogenetic scales — one for each node in the phylogenetic tree. Each regularization term smoothes the difference between the sum of the coefficients for the two subtrees of a node:

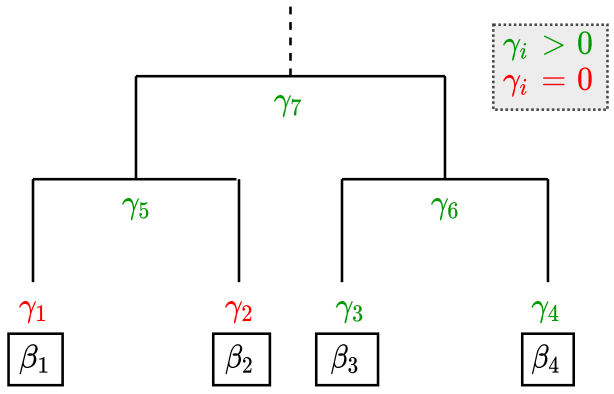
$$\text{(Regression model)} \qquad \beta_1, ..., \beta_p = \text{argmin}_\beta \left(y_i - \beta_0 - \sum x_{i,j} \beta_j \right)$$

$$\text{(Smoothness penalty)} \qquad \sum_{v \in \text{nodes}} \frac{1}{d(v_1, v_2)} \left| \beta_{v_1} - \beta_{v_2} \right| < \mathtt{c}$$

where:
- β_{v_i} is the mean regression coefficients of OTUs in the i^{th}subtree of node v $(i = 1..2)$
- $d(v_1, v_2)$ is the phylogenetic distance between nodes v_1 and v_2
- $\mathtt{c}$ is a regularization constant

Therefore, the regression coefficients of the two direct descendents of a node will tend to be close. In biological terms, it means that two clades descending from a common ancestor should have a similar effect.

Although Trac [21] shares some similarity with the previous methods, the intent is different. The objective consists in aggregating parts of the phylogenetic tree at different depths in a way that is meaningful in regard to some continuous response variable y. In other words, it collapses sub-trees by summing the contribution of its OTUs if its variations are not associated with y. Following the results of Aitchison [3], trac uses a log-ratio model to free itself from



$$\begin{aligned}\beta_1 &= \gamma_5 + \gamma_7\\ \beta_2 &= \beta_1\\ \beta_3 &= \gamma_3 + \gamma_6 + \gamma_7\\ \beta_4 &= \gamma_4 + \gamma_6 + \gamma_4\end{aligned}$$

Figure 3: Trac's latent parameters. β_i is the coefficient of OTU i and γ_j is the additive effect at node j. Color indicates whether the coefficient is positive or null (no effect).

the compositionality constraints. A linear model based on the CLR-transformed abundances is equivalent to a linear model on the log relative abundances with a sum constraint on the coefficients:

$$\begin{aligned}\forall i \in \{1..n\}, y_i &= \sum_{j=1}^{p-1} \log\left(\frac{x_{i,j}}{x_{i,p}}\right) \cdot \beta_j \\ &= \sum_{j=1}^{p-1} \log(x_{i,j}) \cdot \beta_j - \log(x_{i,p}) \cdot \sum_{j=1}^{p-1} \beta_j \\ &= \sum_{j=1}^{p} \log(x_{i,j}) \cdot \beta_j \text{ with } \sum_{j=1}^{p} \beta_j = 0 \\ &\quad \left(\text{i.e. we extend } \beta \in \mathbb{R}^{p-1} \text{ with } \beta_p = -\sum_{j=1}^{p-1} \beta_j\right)\end{aligned} \tag{5}$$

Trac first transforms the model by re-expressing the coefficients β_i for each OTU (leaves) as a sum of node coefficients γ_i. A subtree is then collapsed if the β coefficients for all of its OTUs are the same, or equivalently, if all of its γ values are null (see Figure 3).

Trac enforces sparsity of the γ coefficients with a Lasso penalization on their value. The regularization parameter can control the level of aggregation in the tree. A high regularization parameter will result in a very condensed tree and provide a high-level view of the important clades affecting the response while a small regularization parameter will provide more fine-

grained differences. Therefore, the entire solution path can provide a multiscale view of the problem through multiple aggregation schemes.

Statistics in the simplex are complex due to the strong dependencies between variables that render linear regression invalid. The present penalized regressions circumvent this issue, by either using fused-lasso penalties [20] or log-transforming the data [21]. However, most penalized approaches require setting regularization parameters that do not provide any probabilistic intuition. Similarly to the PCA-based approaches, they also rely on linear combinations of the OTUs, which limits the power of the analyses when handling complex metagenomes.

2.2 Bayesian models

The strength of Bayesian approaches is to provide a full probabilistic framework to describe the parameter's distributions. They usually rely on prior distributions which reflect our knowledge of the problem. BAZE [22] is a bayesian model that takes advantage of the CLR transform introduced by Aitchison [3]. Similar to trac [21], the regression model is expressed as a linear combination of the log abundances, with the log-ratio constraint expressed as a condition on the regression coefficients' sum (equation 5). BAZE encodes this constraint in a z-prior on the regression coefficients. The z-prior enforces the coefficients to sum up to 0 by modeling them as a multivariate gaussian of mean 0 and a covariance matrix that sums up to 0 (Figure 4).

Variable inclusion in the model is controlled with the binary indicator variable γ, equal to 1 if the variable is kept in the model and 0 otherwise. Variable inclusion (i.e. γ) should be performed in phylogenetically-aware way so that relevant OTUs are selected alongside closely related neighbors. This is achieved through an Ising prior [23] on γ, which, given a OTU similarity matrix Q, models γ as $P(\gamma) = \exp(a^T\gamma + \gamma^T Q\gamma - \psi(a, Q))$, where:

$$\begin{cases} a \in \mathbb{R}^p \text{ is a shrinkage parameter} \\ \psi(a, Q) \text{ is a normalizing constant} \\ Q = \left(\frac{l_{i,j}}{\sqrt{l_{i,i}}\sqrt{l_{j,j}}} \right) \in \mathbb{R}^{p \times p} \text{ where } l_{i,j} \text{ is the shared branch length between } i \text{ and } j \end{cases}$$

Using this prior, closely related OTUs will have a higher chance of being selected together. The parameter `a` controls the sparsity of the solution: the lower the value of $\mathtt{a}_i$ and the less likely the variable i will be included. In [24], Wang et al. take a radically different approach by directly modeling the raw OTU abundances which lends themselves to Dirichlet-multinomial (DM) distributions: a sample of size N consists of N independent draws among p OTUs with

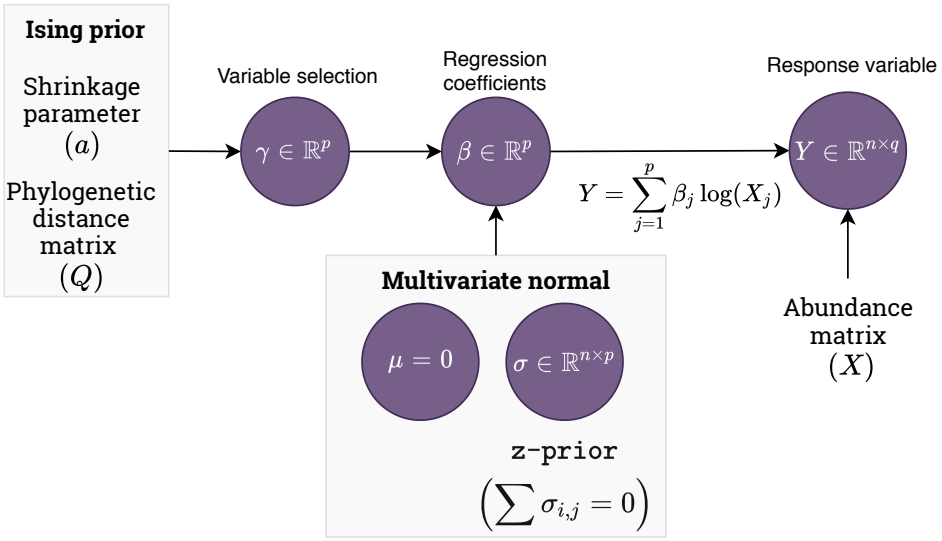


Figure 4: Graphical representation of BAZE model. The response variable Y is modeled as a linear combination of the relative abundance for each OTU. The regression coefficients are assumed to follow a multivariate normal distribution with mean 0 and standard deviation *sigma* following a z-prior (to enforce the 0-sum constraint). Variable selection is controlled by the parameter γ which follows an Ising prior parametrized by a shrinkage parameter (a) controlling the sparsity of the solution and an OTU distance matrix (Q)

probabilities $\pi = (\pi_1, ..., \pi_p)$. Using these priors and after integrating on the mixing probabilities (π) the posterior distribution of the abundances is dirichlet distributed:

$$\mathcal{P}_{DM}(a_1, ..., a_p | \alpha_1, ..., \alpha_p) = \frac{\Gamma(a^+ + 1)\Gamma(\alpha^+)}{\Gamma(a^+ + \alpha^+)} \prod_{i=1}^{p} \frac{\Gamma(a_i + \alpha_i)}{\Gamma(a_i + 1)\Gamma(\alpha_i)}$$

where $a_1, .., a_p$ are the raw abundance counts of the p OTUs, $\alpha_1, ..., \alpha_p$ are the dirichlet distribution concentration parameters and $\Gamma(\cdot)$ is the gamma function defined as $\Gamma(z) = \int_0^{+\infty} x^{z-1} e^{-x} dx$

However, such a model does not take tree structure into account. The dirichlet-tree multinomial (DTM) distribution is a more faithful representation of the data. The model consists of a mixture of DM components across the internal nodes of the tree. The component corresponding to a node describes the distribution of the accumulated counts of its children (Figure 5).

The estimation of the model's parameter is done by maximum likelihood in the regression problem: Given q target variables $Y \in \mathbb{R}^{n \times q}$, a log linear model is built using the DTM

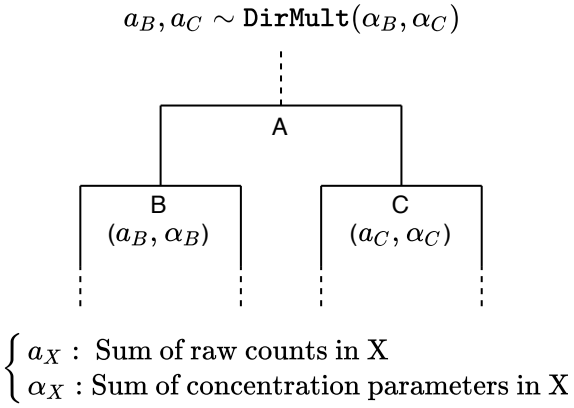


Figure 5: Dirichlet-Tree Multinomial component for a tree node — which corresponds to a local dirichlet-multinomial model. The distribution of the total abundance of the two children nodes of A is modeled with a dirichlet-multinomial model, where the two parameters are the sum of the parameters of each sub-tree

parameters as features:

$\forall v \in \texttt{nodes}(\mathcal{T}), \forall c \in \texttt{children}(v), \log(\alpha_c) = Y^T \beta_c$, where

β_c are the regression coefficients and α_c are the concentration parameters for the DTM

Instead of defining a prior on the regression coefficients, the authors opt for a maximum likelihood approach to get point estimates. In order to promote sparse results, the optimization is complemented with penalized approach with both a Lasso and Ridge regularization as follows:

$$\hat{\beta} = \texttt{argmax}_\beta \left[-\log \left(\mathcal{L}_{DTM}(\beta) \right) + \lambda_1 ||\beta||_1 + \lambda_2 ||\beta||_2 \right]$$

where $\mathcal{L}_{DTM}$ is the maximum likelihood operator for the DTM model.

In contrast, eBay [25] is not a prediction model and mainly focuses on the normalization of the raw counts to solve the problem of inflated zero counts. Similar to BAZE, the data is modeled with a Dirichlet-Tree Multinomial and the concentration parameters $(\alpha_1, ..., \alpha_p)$ are directly estimated from the data through maximum likelihood. The posterior mean of OTU i in sample j is a weighted average of the OTU raw proportion and the relative contribution of concentration parameter α_j :

$$\texttt{E}(\pi_{i,j} \mid a_i, \alpha) = \frac{a_i^+}{a_i^+ + \alpha^+} \frac{a_{i,j}}{a_i} + \frac{\alpha^+}{a_i^+ + \alpha^+} \frac{\alpha_j}{\alpha^+}$$

where $\begin{cases} a_{i,j} \text{ is the raw abundance count of OTU } j \text{ in sample } i \\ \pi_{i,j} \text{ is the proportion of OTU } j \text{ in sample } i \\ \alpha_j \text{ is the concentration parameter of OTU } j \end{cases}$

This method normalizes the abundances and accounts for library size differences using phylogenetic correlations between OTUs. Moreover, the expression of the posterior proportions are non-zero, paving the way for log-ratio approaches. To complement their approach, Liu et al. propose to detect significant features (i.e. tree nodes) using an iterative approach. For each node, a t-test (or alternatively, wilcoxon test) is performed using the posterior means across all samples of the two subtrees and significant nodes are reported. In summary, eBay can be used as a preprocessing step for all the other approaches in this review as a zero imputation procedure. Their complementary approach provides some interesting results, but has multiple drawbacks. First, large trees will increase the number of tests, and will therefore significantly increase the false discovery rate. If accounted for, this will in turn affect the statistical power for feature detection. Second, because of the hierarchical tree structure, a significant test on a given node might be also significant for its parent nodes, making it hard to identify real relevant features. BGCR [26] expands the previous method to address those issues. It adopts a Bayesian approach to testing for association with the response variable while adjusting for covariates. This choice is motivated by the fact that the hierarchical dependencies between tests are more easily captured by a Bayesian approach, which also integrates better with the DTM model aforementioned. The DTM parameters are modeled as a logit-linear combination of the covariates y and stratified over the response variable of interest, z:

$$\texttt{logit}(\alpha_{i,j}(v)) = y_{i,j} \cdot \beta(v) + z_{i,j} \cdot \gamma(v)$$

where: $\begin{cases} \alpha_{i,j}(v) \text{ are the DTM parameter at node } v \\ z_{i,j} \text{ is the indicator variable defining the sample groups} \\ y_{i,j} \text{ are the covariates} \\ \beta(v) \text{ and } \gamma(v) \text{ are the coefficients for the local model at node } v \end{cases}$

The testing relies on the estimation of parameter γ for each node v (0: no association between v and z, positive otherwise). The prior on γ enforces dependencies between the tests of two children nodes and their ancestor. This results in an auto-regressive model that can be computed

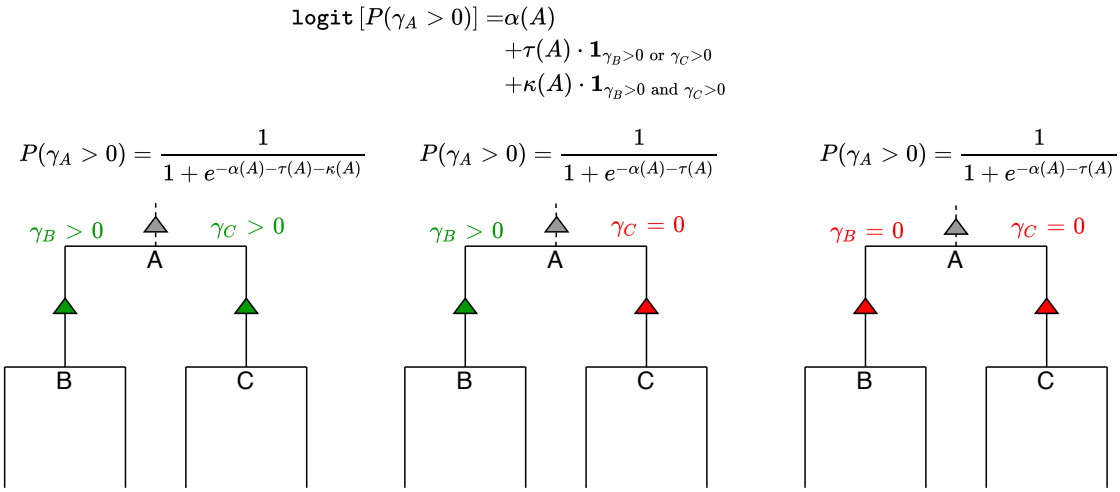


Figure 6: Hierarchical Bayes testing. Arrows show signals propagating through the model. Green are significant association tests at the given node, and red is non significant. Each subfigure shows a different scenario on how the test on the daughter nodes impacts the model of the parent node. The general prior for the association parameter γ is written at the top of the figure.

in a bottom-up fashion (Figure 6). The shared information between a node and its children results in an increased statistical power. Rather than relying on Monte-Carlo simulations, the inference is accelerated by converting the model into a Bayesian network and posterior inference is carried out through an exact message passing algorithm.

The Bayesian methods in this section provide a rigorous framework to model the sampling process and include phylogenetic prior to describe the correlation structure in the counts. The chosen models naturally handle the compositionality and dispersion of the data, either using a model based on log-ratios [22] or using a dirichlet-tree multinomial model. However, most models take a node-centered approach. As discussed in part 1, this assumes that variations in the response variable is best explained from two competing sister sub-tree, and not that an evolutionary trait was acquired along an edge.

2.3 Deep learning models

Various deep learning architectures have been proposed to model microbiome data. These architectures use as input the OTU abundance and attempt to model how it can predict a continuous or categorical response variable of interest (e.g. medical condition, salinity, water quality,...). Convolutional Neural Networks (CNN) are known to achieve great success for classification when there is a strong spatial dependence in the data [27, 28], a property that

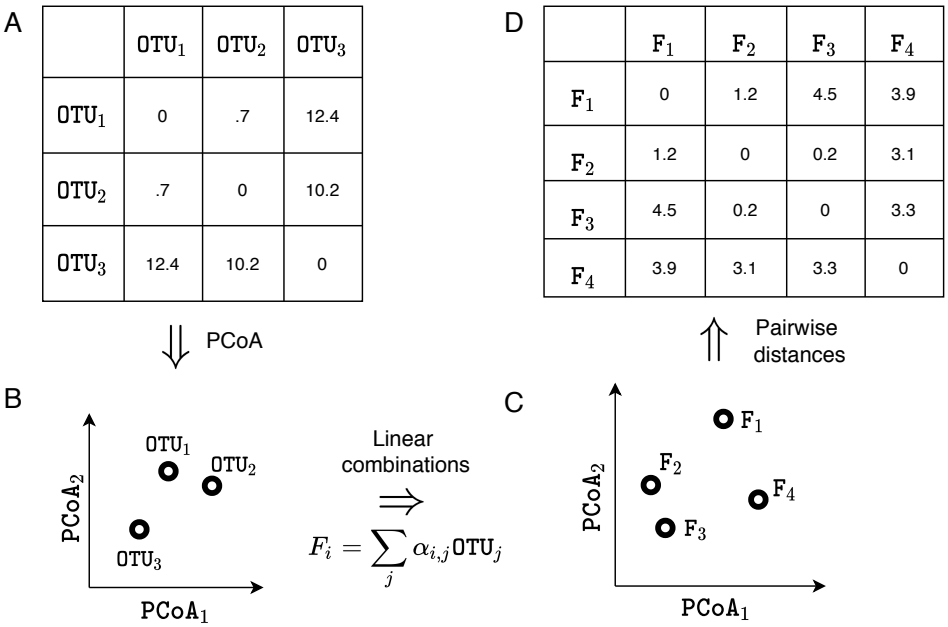


Figure 7: Ph-CNN: Computation of distance matrix between linear combinations of OTUs. (A) Patristic distance matrix between OTUs. (B) PCoA components of each OTU based on patristic distance matrix. (C) Linear combination of OTUs after first phylo-conv layer. (D) New distance matrix based on pairwise distance between features from the first phylo-conv layer (C).

metagenomic abundance does not naturally exhibit. Ph-CNN [29] is a neural network based on "phylo-conv" layers. To compute the output of a feature, a phylo-conv layer transforms the 1D OTU abundance column vector into a matrix by stacking each OTU's k-nearest neighbors (k=16) and applying a 1D convolution filter on the rows. In order to extract the neighbors, the layer uses pairwise distances between features, chosen as the patristic distance matrix. The network's architecture (Figure 8) involves two phylo-conv layers. The second phylo conv layer takes as input a convolved first layer, and therefore finding neighbors is not as straightforward since the input features are a linear combination of the original OTUs. To derive new pairwise distance information, Fioravanti et al. convert the patristic distance matrix into a set of m-dimensional points using PCoA [30], update the point positions, and compute the pairwise feature distance matrix (Figure 7). The phylo-conv computation is followed by a max-pooling layer, a fully connected layer, a dropout layer and a fully connected layer. Although Ph-CNN reports good classification results, its lack of interpretability might hinder its adoption in practice.

The approach developed in [31] combines the abundance table with the phylogenetic tree into a 2D matrix with spatial structure to predict a categorical outcome y. For each sample, the tree nodes are labeled with the sum of the abundance of their leaves. The tree is then

converted into a matrix, where each row represents a depth in the tree. The values are filled in order from left to right, and the matrix is zero-padded if needed. For example, the first row of the matrix consists of a single value equal to the sum of the OTUs abundance in a sample. The deep neural network architecture consists of three convolutional layers (including convolution and max pooling layers) followed by a fully-connected layer and an output softmax layer. The architecture is updated in a later version of the tool [32] with two convolutional layers without any max pooling (Figure 8) since the authors report improved performance when removing it. While the first convolutional layer scans the input with a rectangular filter to detect local features, the second layer uses a 1x1 filter to collapse the feature maps from the first layer and reduce the number of parameters.

There are a number of caveats to the proposed approach. First, the matrix representation does not always faithfully represent the topology of the tree. Misalignments between rows in the matrix can occur when the tree is wide, causing descendants to be far away from their ancestral nodes. Second, the padding procedure yields a matrix with many zeros on the top right section of the matrix which might affect the convolution of non-zero values close to this section. Finally, the lack of interpretability of deep learning models is also an issue for this model. PopPhy-CNN [32] is an extension of this model that addresses this last issue. To identify meaningful features associated with each outcome, PopPhy-CNN traces the signal generated by each sample class in the first convolutional layer back to specific positions in the input. The feature importance computation relies on the velocities, which are the result of the convolution between a kernel and a reference window from the input. The feature importance of a reference position is simply its contribution to the total velocity in the kernel, averaged across all samples in the class. The total feature importance $I_c(i,j)$ is then its maximum value across all windows spanning (i,j) and all kernels. Finally, a score is derived for this feature as the difference between the feature importance in class c and the feature importance in the other classes: $S_c(i,j) = I_c(i,j) - I_{\bar{c}}(i,j)$

DeepBiome [33] is a feed-forward neural network for predicting a continuous or categorical outcome. The approach is radically different than in PopPhy-CNN. The network is trained to minimize the mean squared error for a continuous outcome or the cross-entropy for a categorical outcome. DeepBiome's specificity is its architecture, which consists of a succession of layers that mimics the hierarchical taxonomic levels (species $<$ genus $<$ family $<$ ...). The phylogenetic structure is enforced on the network through weight decay, a regularization procedure that penalizes large weights to prevent overfitting. DeepBiome alters the weight decay procedure to

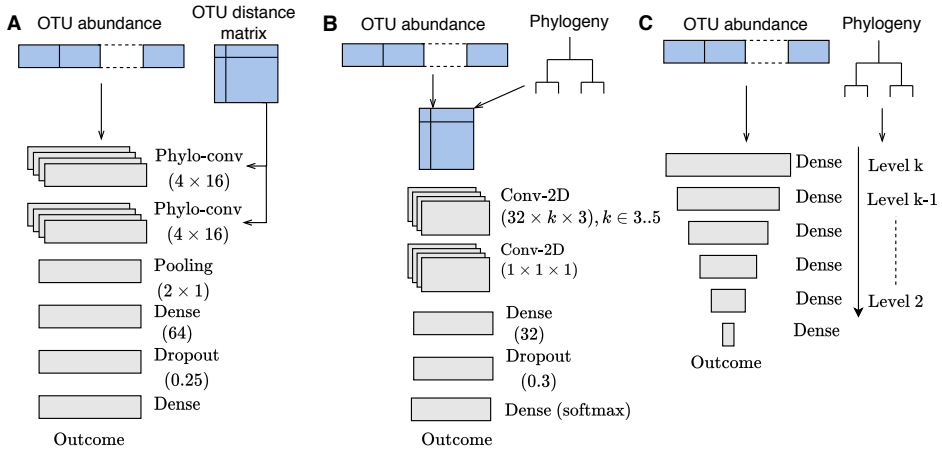


Figure 8: Neural network architectures. (A) Ph-CNN architecture. The patristic distance matrix is used at each phylo-conv layer to identify each OTU's closest neighbors. Each entry of the abundance vector is stacked with its closest neighbors in order to perform a 1D convolution. The phylo-conv output is subsequently processed by fully-connected layers (B) PopPhy-CNN architecture. The abundance values are combined with the phylogenetic tree to provide a node-level representation of the abundances. A classical CNN is subsequently applied to the transformed input. (C) DeepBiome architecture. Multi-layer perceptron with one fully-connected layer per level in the tree, starting with the deepest.

incorporate phylogenetic information in the gradient descent. More specifically, at each update, the network weights are multiplied by a small value (0.01) if they link two unrelated taxonomic levels. For example, the neuron corresponding to the *proteobacteria* phylum would have a strong link to the preceding layer neurons of each of its descendent classes (e.g. *alphaproteobacteria*, *betaproteobacteria*, ...) but would be downweighted if the link involves a non-related class (e.g. *bacilli*). Therefore, it promotes strong links between closely related organisms. Because of its unique architecture, DeepBiome would benefit from feature importance approaches to trace the signal propagation in each layer. Although the authors report feature selection results, they do not provide extensive details on the procedure.

Discussion and conclusion

Metagenomic data analysis poses a number of challenges for identifying meaningful groups of OTUs driving the microbial communities. In this review, we present multiple approaches to solving those challenges. First, compositional variables (OTU abundances) have a biased covariance structure which can invalidate their analysis if not accounted for. The methods presented here address this issue by either using log-ratios [6, 7, 8, 9, 10, 22], regularization [17, 24, 20], alternative inner-products [15], or modeling the counts, explicitly with Bayesian

approaches [22, 25, 26] or implicitly with neural networks [29, 31, 32, 33]. Second, most analyses are challenged by the high number of OTUs and their correlation structure since it decreases the power of statistical tests. Instead of testing individual OTUs, the phylogeny can guide the analysis to organize OTUs into homogeneous groups. A phylogenetic tree contains both topological and geometrical information [34]. Topological information is the relative positions of the OTUs in the tree, which is conveyed by the branching patterns in the tree, while geometrical information are the evolutionary distances between OTUs, which is conveyed by the branch weights. The methods in this paper all use one of these two types of information, but rarely both. Methods based on topology [6, 8, 9, 10, 13, 24, 29, 31, 32, 33] rely on the hierarchical grouping defined by the phylogenetic tree. Because these methods discard the evolutionary distance and simply use the nodes hierarchy, OTU groups are fully coerced by the tree which makes them more robust and interpretable features. In contrast, methods based on the geometry of the tree [15, 17, 22, 29] discard any groups information and instead fully use the evolutionary distance. As a result, these methods are more flexible since they can define their own groups for the problem at hand. However, they might also define groups that are less biologically meaningful since these are not coerced by the evolutionary model. PhILR [7] and the tree-guided fused lasso approach used in [20] are the only methods that use both types of information. DeepBiome [33] does not seem to take the phylogeny into account in a strict sense, since it relies on the taxonomy. However, the approach would be easily generalizable for an arbitrary phylogeny, where each layer in the neural network could match a specific depth in the tree. Moreover, DeepBiome currently uses the tree topology but disregards its geometry. This could also be implemented using a more complex weight regularization process.

The unique distribution of the counts is rarely modeled explicitly. PCA and log-ratio based approaches create a mixed data structure including phylogenetic information, making the modeling of abundance distribution more complex for the subsequent analysis as compared to raw counts, where a dirichlet-multinomial naturally describes the count generation process. As such, Bayesian models based on Dirichlet-Tree Multinomial distributions [25, 26] are the most faithful representation of the problem. Deep learning models [29, 31, 32, 33] also represent an interesting option since they implicitly model the data distribution. Given the complexity of metagenomics data (hierarchical structure, high proportion of zeros, overdispersion), these models are very promising for regression and classification purposes. However, this strength can also be a weakness, since the intricacies of the model are, in general, not fully understood.

This can result in counterintuitive predictions errors [35] or, in the worst case, learning features related to a representation bias in the dataset [36]. The proposed deep learning approaches try to lift the veil on the models through regularization [33] or the use complementary heuristics to provide insights about the most meaningful features [32]. However, feature extraction remains a computationally intensive and an experimental procedure that lacks validation. In spite of those issues, rapid advances are currently made in neural network interpretability and will undoubtedly affect their use in metagenomics in the future.

Although some methods take the OTU placement uncertainty into account by weighting the abundance values [10, 13], none of the methods explicitly handle and include this uncertainty in their model. For example, one could expect that a phylogenetic-aware method would rely less on the tree when the probabilities are too uniform. As a result, the methods described in this paper would likely perform worse than a method based simply on the abundances if the phylogenetic tree is very noisy. This is especially the case for large metagenomes, for which approximations are required to compute the phylogenetic tree in reasonable time. Furthermore, trees are inferred based on specific properties of organisms (called traits), which are not always related to the response variable being studied. For example, a tree based on antibiotic resistance genes would not be helpful to guide an analysis focused on the salinity of an environment. In summary, an erroneous tree or a tree based on the wrong traits could negatively affect phylogenetic approaches, which is why present methods could benefit from levers controlling the amount of phylogenetic prior in the model. Although this is very experimental at its current stage, new methods are being developed to combine the information from multiple phylogenetic trees into a phylogenetic network [37]. None of the surveyed methods have yet considered this option and methods based on graph neural networks or Bayesian networks could be naturally adapted to this problem. Indeed, as their name suggests, those approaches can predict an outcome based on features organized as a graph and are not restricted to binary trees. However, the optimal choice of the traits to use for the phylogenetic networks would represent another challenge to address when using this kind of approaches.

Authors	Tool name	Model type	Category
Morton et al.	Gneiss [6]	Node-based ILR	Transformation
Silverman et al.	PhILR [7]	Node-based ILR	Transformation
Washburn et al.	Phylofactorization [8, 9]	Edge-based ILR	Transformation
Czech and Stamatakis	Gappa [10]	Edge-based ILR	Transformation
Masten and Evans	edgePCA [13]	Edge-based ILR	Transformation
Purdom	gPCA [15]	PCA	Transformation
Chen et al.	ssCCA [17]	ssCCA	Feature selection
Wang and Zhao	Tree-guided fused Lasso [20]	Fused Lasso	Regression
Bien et al.	Trac [21]	Fused Lasso	Regression
Zhang et al.	BAZE [22]	Empirical Bayes	Regression
Wang and Zhao	Dirichlet-tree [24]	Empirical Bayes	Feature selection
Liu et al.	eBay [25]	Empirical Bayes	Feature selection
Mao et al.	BGCR [26]	Bayesian network	Feature selection
Fioravanti et al.	Ph-CNN [29]	CNN	Classification
Reiman et al.	PopPhy-CNN [31, 32]	CNN	Classification
Zhai et al.	DeepBiome [33]	MLP	Regression / Classification

Table 1: Summary of described methods

References

[1] Donald A Jackson. "COMPOSITIONAL DATA IN COMMUNITY ECOLOGY: THE PARADIGM OR PERIL OF PROPORTIONS?" In: *Ecology* 78.3 (Apr. 1997), pp. 929–940.

[2] Catherine H Graham, David Storch, and Antonin Machac. "Phylogenetic scale in ecology and evolution". In: *Glob. Ecol. Biogeogr.* 27.2 (Feb. 2018), pp. 175–187.

[3] J Aitchison. "The Statistical Analysis of Compositional Data". In: *J. R. Stat. Soc. Series B Stat. Methodol.* 44.2 (Jan. 1982), pp. 139–160.

[4] Dean Billheimer, Peter Guttorp, and William F Fagan. "Statistical interpretation of species composition". In: *Journal of the American statistical Association* 96.456 (2001), pp. 1205–1214.

[5] J J Egozcue et al. "Isometric Logratio Transformations for Compositional Data Analysis". In: *Math. Geol.* 35.3 (Apr. 2003), pp. 279–300.

[6] James T Morton et al. "Balance Trees Reveal Microbial Niche Differentiation". en. In: *mSystems* 2.1 (Jan. 2017).

[7] Justin D Silverman et al. "A phylogenetic transform enhances analysis of compositional microbiota data". en. In: *Elife* 6 (Feb. 2017).

[8] Alex D Washburne et al. “Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets”. en. In: *PeerJ* 5 (Feb. 2017), e2969.

[9] Alex D Washburne et al. “Phylofactorization: a graph partitioning algorithm to identify phylogenetic scales of ecological data”. In: *Ecol. Monogr.* 89.2 (May 2019), e01353.

[10] Lucas Czech and Alexandros Stamatakis. *Scalable methods for analyzing and visualizing phylogenetic placement of metagenomic samples*. 2019.

[11] Matthew C B Tsilimigras and Anthony A Fodor. “Compositional data analysis of the microbiome: fundamentals, tools, and challenges”. en. In: *Ann. Epidemiol.* 26.5 (May 2016), pp. 330–335.

[12] Justin D Silverman et al. “Naught all zeros in sequence count data are the same”. en. In: *Comput. Struct. Biotechnol. J.* 18 (Sept. 2020), pp. 2789–2798.

[13] Frederick A Matsen 4th and Steven N Evans. “Edge principal components and squash clustering: using the special structure of phylogenetic placement data for sample comparison”. en. In: *PLoS One* 8.3 (Mar. 2013), e56859.

[14] Pierre Legendre and Eugene D Gallagher. “Ecologically meaningful transformations for ordination of species data”. en. In: *Oecologia* 129.2 (Oct. 2001), pp. 271–280.

[15] Elizabeth Purdom. *Analysis of a data matrix and a graph: Metagenomic data and the phylogenetic tree*. 2011.

[16] X Jiang et al. “Manifold learning reveals nonlinear structure in metagenomic profiles”. In: *2012 IEEE International Conference on Bioinformatics and Biomedicine*. Oct. 2012, pp. 1–6.

[17] Jun Chen et al. “Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis”. en. In: *Biostatistics* 14.2 (Apr. 2013), pp. 244–258.

[18] Hotelling Harold. “Relations between two sets of variates”. In: *Biometrika* 28.3/4 (1936), pp. 321–377.

[19] Robert Tibshirani et al. “Sparsity and smoothness via the fused lasso”. In: *J. R. Stat. Soc. Series B Stat. Methodol.* 67.1 (Feb. 2005), pp. 91–108.

[20] Tao Wang and Hongyu Zhao. “Constructing Predictive Microbial Signatures at Multiple Taxonomic Levels”. In: *J. Am. Stat. Assoc.* 112.519 (July 2017), pp. 1022–1031.

[21] Jacob Bien et al. “Tree-Aggregated Predictive Modeling of Microbiome Data”. en. Sept. 2020.

[22] Liangliang Zhang et al. “Bayesian compositional regression with structured priors for microbiome feature selection”. en. In: *Biometrics* (July 2020).

[23] Fan Li and Nancy R Zhang. “Bayesian Variable Selection in Structured High-Dimensional Covariate Spaces With Applications in Genomics”. In: *J. Am. Stat. Assoc.* 105.491 (Sept. 2010), pp. 1202–1214.

[24] Tao Wang and Hongyu Zhao. “A Dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms”. en. In: *Biometrics* 73.3 (Sept. 2017), pp. 792–801.

[25] Tiantian Liu, Hongyu Zhao, and Tao Wang. “An empirical Bayes approach to normalization and differential abundance testing for microbiome data”. en. In: *BMC Bioinformatics* 21.1 (June 2020), p. 225.

[26] Jialiang Mao, Yuhan Chen, and Li Ma. “Bayesian Graphical Compositional Regression for Microbiome Data”. In: *J. Am. Stat. Assoc.* 115.530 (Apr. 2020), pp. 610–624.

[27] Kunihiko Fukushima. *Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position.* 1980.

[28] Peter W Battaglia et al. “Relational inductive biases, deep learning, and graph networks”. In: (June 2018). arXiv: `1806.01261 [cs.LG]`.

[29] Diego Fioravanti et al. “Phylogenetic convolutional neural networks in metagenomics”. en. In: *BMC Bioinformatics* 19.Suppl 2 (Mar. 2018), p. 49.

[30] Michael A A Cox and Trevor F Cox. “Multidimensional Scaling”. In: *Handbook of Data Visualization*. Ed. by Chun-Houh Chen, Wolfgang Härdle, and Antony Unwin. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 315–347.

[31] Derek Reiman, Ahmed Metwally, and Yang Dai. *Using convolutional neural networks to explore the microbiome.* 2017.

[32] Derek Reiman et al. *PopPhy-CNN: A Phylogenetic Tree Embedded Architecture for Convolutional Neural Networks to Predict Host Phenotype From Metagenomic Data.* 2020.

[33] Jing Zhai et al. “DeepBiome: a phylogenetic tree informed deep neural network for microbiome data analysis”. In: (2020).

[34] Tom M W Nye. “Principal components analysis in the space of phylogenetic trees”. en. In: *Ann. Stat.* 39.5 (Oct. 2011), pp. 2716–2739.

[35] Anh Nguyen, Jason Yosinski, and Jeff Clune. “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 427–436.

[36] Tolga Bolukbasi et al. “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings”. In: *Advances in Neural Information Processing Systems*. Ed. by D Lee et al. Vol. 29. Curran Associates, Inc., 2016, pp. 4349–4357.

[37] Leo van Iersel et al. “Phylogenetic networks do not need to be complex: using fewer reticulations to represent conflicting clusters”. en. In: *Bioinformatics* 26.12 (June 2010), pp. i124–31.