Literature Review

Master Thesis

Benchmarking machine learning performances with compositional

Jennifer Neumaier

INSTITUTE OF GENOMICS, UNIVERSITY OF TARTU

TABLE OF CONTENTS

INTRODUCTION

GOAL

Machine learning in microbiome studies is widely used and the interest is growing. However, there is no universal understanding of the algorithmic approaches that can best utilize the information present in the microbiome data. Thus, this is an interesting and widely discussed topic that can have a great impact on the potential applications leveraging microbiome data. A key topic in microbiome research is the sample space of the input data. The sequencing data appears as count data, but, only relative abundance of the microbial features can be observed, commonly called “compositional data”. Thus, transforming the read counts to relative abundances is usually the first step and machine learning methods are usually applied on relative abundances. However, relative abundances raise several limitations, which can have an impact on the performance of the prediction models. Therefore, log-ratio transformations are a proposition made by several studies now, however their impact on machine learning performances has never been tested in large-scale studies. The goal of this benchmarking project is to rectify that and conduct several machine learning models under several log-ratio transformations in comparison to standard microbiome approaches like *selbal* or *CoDaCoRe*. This way it will become clearer if a scientist should make the effort in learning about machine learning methods, when automated algorithms perform well enough and no heavy prior machine learning knowledge is necessary.

BACKGROUND

MICROBIOME DATA IS COMPOSITIONAL

Microbiome data is achieved by taking a population of (total or fractionated) RNA, converting them to a library of cDNA fragments, optionally amplifying the fragments, and then sequencing those fragments in a ‘high-throughput manner’ (Quinn et al. 2018). This methodology is known as next generation sequencing. The result of NGS is a virtual library of many short sequence fragments that are converted to a numeric dataset through alignment (most often to a previously established reference genome or transcriptome) and quantification (Griffith et al. 2015). Thus, sequence abundances are not absolute abundances because the total number of sequences measured by sequencing machines ultimately depends on the chemistry of the assay, not the input material (Quinn et al. 2018).

Commented [OA1]: Having had a look on the structure of the “Background”, I personally would love to have a short summary/introduction paragraph of the problem in hand, even before the background. That makes the following sections easier to gasp: Followingly, I will describe… peculiarities with data, problems that arise, possible solutions, work so far etc

Commented [JN2R1]: I have used a lot of phrases from your draft, is that alright? :3

Commented [OA3]: I would recommend the introduction of this article to get a better overview: https://cdnsciencepub.com/doi/10.1139/cjm-2015-0821

Commented [OA4R3]: But all in all, Elin has a better understanding of the procedure and correct usage of the terms :D

Commented [JN5R3]: I am quite familiar with the procedures of NGS as I have worked in an NGS lab for over a year. But with this comment I gather you think I should add more explanations here? Since this topic focuses mostly on statistics I didn’t go into a lot of detail about the procedure

This leads to the illusion that sequencing data appears as count data, but in reality, only relative abundance of the microbial features can be observed (Gloor et al. 2017), since the abundances for each sample are constrained by an arbitrary total sum (Quinn et al. 2018). Thus, the individual values of the observed counts are irrelevant (Quinn et al. 2018). The following figure displays this problematic visually:

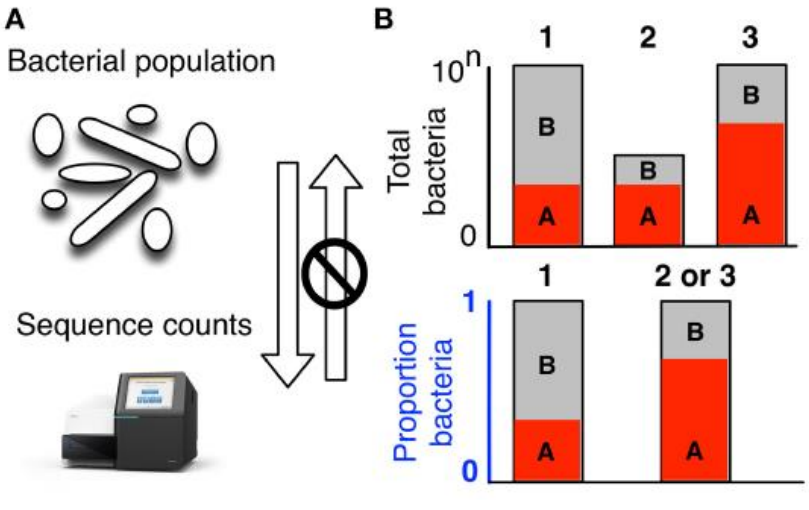

Figure 1: Characteristics of compositional data

Taken from (Gloor et al. 2017). (A) After sequencing the data observed from a bacterial population cannot inform on the absolute abundance of molecules. The number of counts in a high throughput sequencing dataset reflect the proportion of counts per feature per sample, multiplied by the sequencing depth. Therefore, only relative abundances are available. The consequences are portrayed in (B). The bar plots show the difference between the count of molecules and the proportion of molecules for two features, A (red) and B (gray) in three samples. The top bar graphs show the total counts for three samples, and the height of the color illustrates the total count of the feature. When the three samples are sequenced, we lose the absolute count information and only have relative abundances, proportions, or "normalized counts" as shown in the bottom bar graph. Note that features A and B in samples 2 and 3 appear with the same relative abundances, even though the counts in the environment are different.

Thus, relative abundance data - and microbiome data - are mathematically considered "compositional data", with its own mathematical theory and properties. Compositional data lives in the positive simplex space and not in real Euclidean space, which is assumed by commonly used data analysis (Quinn et al. 2018). Thus, compositional data is very awkward to handle due to its scarcity of meaningful definitions of independence (Aitchison 1982). Luckily, the relative abundances of microbial features still carry meaning. Several propositions have been made in the

Commented [OA6]: Microbial features/species/…

last few years to acknowledge compositional data in statistical analysis (Aitchison 1982) and increase its interpretability.

CURRENT SOLUTIONS FOR COMPOSITIONAL DATA IN STATISTICAL ANALYSIS

Gloor et al. (2017) pointed out the importance of an alternative tool kit for compositional data. One of the first analysis steps in traditional analysis is the calculation of a distance or dissimilarity (DD) matrix from the data after rarefaction or count normalization. Figure 2 shows a standard

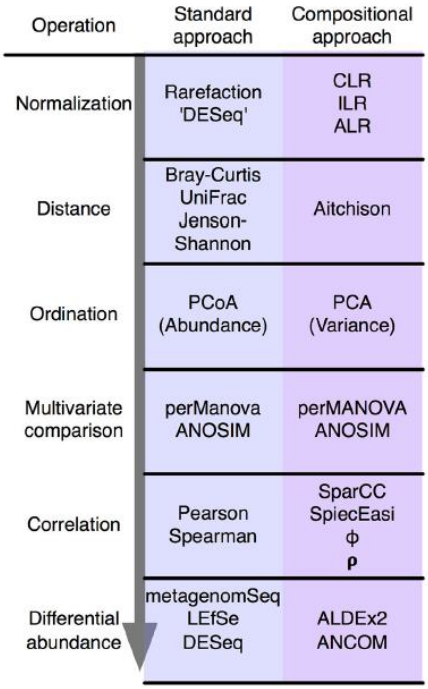

Figure 2: Standard microbiome analysis tool kit and compositional replacements

Figure was taken from Gloor et al. (2017). It depicts a simplified standard microbiome computational workflow.

microbiome toolkit and its alternatives for compositional data.

Common in microbiome analysis are UniFrac, Bray-Curtis and Jensen-Shannon divergence. Inherently, DD methods are sensitive to the total read depth of a sample. Thus, they do not account for the compositional nature of the data and since they largely discriminate between samples based on the most relative abundant features, instead the most variable, this can lead to drastic

changes when different features are included or excluded in the dataset (Gloor et al. 2017). Therefore, Aitchison proposed the so called “Aitchison distance”. It is more stable to sub setting and aggregating of the data, and being a true linear distance (Gloor et al. 2017).

The major uses for DD matrices are ordination and clustering (Gloor et al. 2017). Using the Aitchison distance solves the problem of sensitivity in ordination. However, it has been shown that differential abundance tools are sensitive to sparsity and correlation is not reliable or a reproducible indicator when dealing with compositional data (Gloor et al. 2017). The replacement for β-diversity exploration of microbiome data is the variance-based compositional principal component (PCA) biplot (Gloor et al. 2017). It has the advantage that exploratory data analysis is not driven by the presence-absence relationships in the data nor by excessive sparsity. Also, it also does not rely on an underlying phylogenetic tree.

Severe problems with correlation in compositional data were noted early (Gloor et al. 2017), as compositional data have a negative correlation bias and a different correlation structure than the underlying count data (Gloor et al. 2017). This is a severe problem in compositional data analysis. Possible approaches to analyze correlation are SPARCC and SpiecEasi, which both assume a sparse data matrix, as well as two metrices which require a non-sparse matrix (Gloor et al. 2017). Finally, differential abundance of OTUs in compositional data can be analyzed by ANCOM, which performs statistical tests on point estimates of data transformed by an ALR. ALDEx2 performs statistical tests on log-transformed values from a modelled probability distribution of the data set (Gloor et al. 2017).

The described methods show clearly the problems when trying to analyze compositional data in Euclidean space. They successfully work around the characteristics of compositional data however their interpretability and practicality leave much to wish for. Additionally, these methods are also not feasible for machine learning purposes, as it would increase the computational complexity dramatically. Thankfully, there is a very elegant way of solving this predicament: log-ratio transformations.

LOG-RATIO TRANSFORMATIONS MAKE COMPOSITIONAL DATA ANALYSIS A LOT EASIER

The difficulty of confined data points has already been commented on by Pearson (1897) in the context of spurious correlations and has been taken up by Aitchison 1982 in an attempt to overcome the "bounded sum problem". Although compositional data exist in the simplex, Aitchison first documented that these data could get mapped into real space by use of the log-ratio transformation (Quinn et al. 2018; Aitchison 1982). This does not normalize the data (does not "open it"), but makes the interpretation of the transformed data dependent on the reference used and aim for a straight-forward univariate interpretation of the data (Quinn et al. 2018). Furthermore, it allows to employ standard analysis methods instead of the more complex alternatives introduced in the chapter before.

For all log-ratio transformations, relationships between the features in the data set are captured and taking the logarithm of these ratios makes the data symmetric and linearly related. It moves the simplex into real space and imparts key properties to the data set: scale invariance (performance does not change with e.g., sequencing depth), perturbation invariance (i.e., converting a composition between equivalent units will not change the results), and permutation invariance (i.e., changing the order of the components within a composition will not change the results).

Two more important properties exist that are transformation-specific: sub-compositional coherence (i.e., identical results are enforced when components are included in compositions or sub-compositions), and sub-compositional dominance (i.e., using a subset of a complete composition carries less information than using the whole) (Quinn et al., 2018). It has been shown recently that quasi-coherence is sufficient in practice, as well as quasi-isometry (Greenacre et al. 2022), as true isometry is difficult to interpret. However, keeping these characteristics in mind when choosing log-ratio transformations is important, as not every log-ratio transformation inherently incorporates all traits from Euclidean space.

Typical transformation techniques in compositional data consist of CLR (centered log-ratio). Here, the geometric mean of the composition is used in place of the reference feature (Gloor et al. 2017). It has the advantage of being scale invariant and a good interpretability which makes it very practical. However, is not very useful in sparse data containing a lot of 0s. Later on, methodologies will be discussed to deal with 0 count values (Gloor et al. 2017).

Commented [OA7]: This is a property used for distances usually, although used in different contexts. I would say subcompositional coherence is a more important property (especially regarding the initial filtering done). The difference has been made clear here: https://arxiv.org/abs/2201.05197

Commented [OA8]: A detail, but the following formulas definitely needs an explanation of the elements in the formula: x_i, g(x) etc. Seemingly trivial, but for example, it remains unclear if the transformations are row-wise or column-wise, which in turn is vital for ML: do we need to wrap it within CV or not as it is necessary for standardization or PCA etc.

Commented [JN9R8]: I added a few more information below the equations but since the transformation procedure is something I am not sure myself at this point, I would like to include it later on in the review. Does this work?

$$\mathrm{clr}(\mathbf{x}_j) = \left[\ln\frac{x_{1j}}{g(\mathbf{x}_j)}, \ldots, \ln\frac{x_{Dj}}{g(\mathbf{x}_j)}\right].$$

Figure 2: Equation for CLR

Equation describes calculation of CLR, with x_j as vector of sample features, D_j the total number of features, and g(x) the geometric mean of sample vector x. Log-ratio transformations are applied within a sample (i.e., in a patient).

More complex is ILR (isometric log-ratio), which transforms the data with respect to an orthonormal coordinate system that is constructed from sequential binary partitions of features (Quinn et al. 2018). The ILR-transform maps a composition in the D-part Aitchison-simplex isometrically to a D-1 dimensional Euclidian vector, with clr(x) the centered log-ratio transform and V a matrix which columns form an orthonormal basis of the clr-plane (Greenacre et al. 2022).

$$ilr(x) := V^t\, clr(x)$$

Figure 3: Equation of ILR

Equation describes calculation of ILR, with x as vector of sample features, V a matrix which columns form on orthonormal basis of the clr-plane. Log-ratio transformations are applied within a sample (i.e., in a patient).

Isometric log-ratios are the “gold standard” of log-transformations, as they engender exactly the same multivariate geometric structure of the sample points as that of all the pairwise log-ratios, called the “log-ratio geometry” or also “Aitchison geometry” (Greenacre et al. 2021). Unfortunately, isometric log-ratios are particularly problematic when the numbers of components in the geometric means are high and thus lack interpretability (Greenacre et al. 2021).

Therefore, transformations such as ALR (additive log-ratio) are re-evaluated in their effectiveness. In ALR, the logarithm is taken of each measurement within a composition and divided by a reference feature.

$$\mathrm{alr}(\mathbf{x}_j) = \left[\ln\frac{x_{1j}}{x_{Dj}}, \ldots, \ln\frac{x_{D-1j}}{x_{Dj}}\right].$$

Figure 4: Equation ALR

Equation describes calculation of ALR, with x_j as vector of sample features, D_j the total number of features, and x_{Dj} the reference feature. Log-ratio transformations are applied within a sample (i.e., in a patient).

Here, a small loss of isometry is traded off in favor of the benefit of a simpler and clearer interpretation of the log-ratio variables, as the interpretation of the results is always according to the chosen reference. When choosing a reference, Greenacre et al. 2021 propose to use three

criteria to find a good reference for the denominator: (i) the reference component should maximize the Procrustes correlation between the additive log-ratio geometry and the exact log-ratio geometry, (ii) the reference should minimize the variance the relative abundances of log-transformed components, and (iii) it should be a well populated component. Using these guidelines, produces additive log-ratios close to being isometric, which would make them a favorable log-transformation. For machine learning purposes however, it is still unclear if isometric log-ratio transformations improve the performance in a prediction task. This will be one of the core goals of this benchmarking project.

Further discussable log-ratios are IQLR (inter-quartile log-ratio), PWLR (pair-wise log-ratio), and SLR (summed log-ratio). As those log-ratios come with higher complexity in terms of computational power and interpretability, they will be expanded to if the size of the project allows it.

Commented [OA10]: "However, it remains unclear, whether this can improve the performance of the ML models in a prediction task" I would add such a comment or discussion somewhere, because this analysis for isometry deals with distances which might not be necessary.

Commented [OA11]: Can be expanded, but a reasoning for selecting the log-ratios to expand is necessary, I think (simplicity/popularity/interpretability…).

COMPOSITIONAL DATA IN MACHINE LEARNING

In terms of statistical analysis, Machine Learning models are of great interest for microbiome analysis, as they allow predictions of biomarkers, phenotypes or microbial taxa, as well as other interesting tasks, that are not possible with the standard microbiome tool kit (Marcos-Zambrano et al. 2021). Therefore, a correct application of Machine Learning models is key to reproducible and interpretable research results. Several studies (Zhang and Shi 2019; Coenders and Greenacre 2021) showed that log-ratio transformations improve the performance of machine learning models, but no large scale benchmarking has been conducted so far. Furthermore, employing log-ratio transformations leads to an increase in complexity in the correct application of machine learning models. Thus, it is of increasing importance to create a practical guide for all scientists who need to employ such analysis.

Predictive methods such as random forests (RF), artificial neural networks (ANN), deep learning (DL) or support-vector machines (SVM) and other methods have become in the last years increasingly popular (Tolosana-Delgado et al. 2019). Traditional machine learning methods can provide added predictive power with the price of limited explainability. Thus, balancing the predictive power with explainability becomes important for the conclusions.

Several papers showed some interesting observations which will be useful for the purpose of this benchmarking project. In 2019, Zhang and Shi compared several machine learning algorithms on geological compositional data and showed that overall, RF was the best performing model and

that ILR and CLR were superior to ALR (Zhang and Shi 2019). Tolosana-Delgado et al. (2019) showed that ridge regression and SVM both need ILR. More observations were also made by Quinn et al. 2020. They performed linear discriminant analysis (LDR) on ILR-transformed data and partial least squares (PLS) to CLR-transformed data and showed good predictive results (Quinn and Erb 2020). Neural Networks require further research, but does not seem to be equivariant (Tolosana-Delgado et al. 2019).

These observations demonstrate the core problematic of compositional data. Log-ratio selection in linear and generalized linear models is not easily chosen and depends heavily on the observations at hand. The reason why ALR was outperformed by Zhang and Shis study (2019) for example, was probably due to a badly chosen reference and this makes the direct comparison of several studies difficult. In general, log-ratio transformations seem to outperform raw proportions for classification tasks, but it is not clear how log-ratio transformations relate to the changes in predictive performance.

In an attempt to alleviate the subjectivity in deciding which log-transformations to use, Rivera-Pinto et al. (2018) proposed *selbal*. It is a greedy stepwise algorithm that searches for a sparse model that adequately explains the response variable of interest. In multiple regression a new taxon is added to the model each time and assessed if its relative abundance (or balance) is predictive of the outcome (Rivera-Pinto et al. 2018). It has been developed specifically for microbiome data and has been shown to work effectively. Another algorithm was implemented by (Gordon-Rodriguez et al. 2021) in CoDaCoRe (**C**ompositional **D**ata via **C**ontinuous **R**elaxations). It approximates a combinatorial optimization problem over the set of log-ratios by using gradient descent and thus dramatically reduces the runtime without sacrificing interpretability nor predictive accuracy (Gordon-Rodriguez et al. 2021).

In summary, a lot of the provided information show promises in terms of predictive performances of log-ratio transformations for machine learning models compared to no transformation. However, considering the small number of studies and its benchmarking aspect, it context for this project they should be taken more as a guideline instead of face-value. Therefore, re-validating their results could prove to be beneficial for the scientific community. Furthermore, the observations from all these papers show, that the selection and performance of the best algorithm is heavily dependent on the dataset, its research hypotheses, and models. It is therefore difficult to understand and handle for non-experts, but unfortunately vital to the scientific community. Thus, this benchmarking project will focus on establishing a pipeline, as well as recommendations and

Commented [OA12]: What do you mean by that? Selection of the transformation?

Commented [OA13R12]: I would elaborate on the differences of selbal and codacore from the other ML methods more thoroughly. The first have been developed keeping microbiome data in mind, the latter are general algorithms etc

Commented [JN14R12]: If I am a scientist unfamiliar with transformations and I am reading this papers I would be like: “but which transformation do I now choose”? that’s what I mean by “subjectivity” where there shouldn’t be one.

Commented [OA15]: When compared to no transformation (see comment above)

Commented [OA16]: Possibility with a potential? As we don’t have a guideline in my optinion, which you can provide 😉

Commented [OA17]: Performance/selection of the best algorithm?

Commented [OA18R17]: As the transformations themselves are not dependent on the dataset – they have a formula

guidelines that reduce human error and hopefully improves quality management in machine learning methodology.

Commented [OA19]: And a recommendation/decision guideline.

METHODOLOGY

PIPELINE

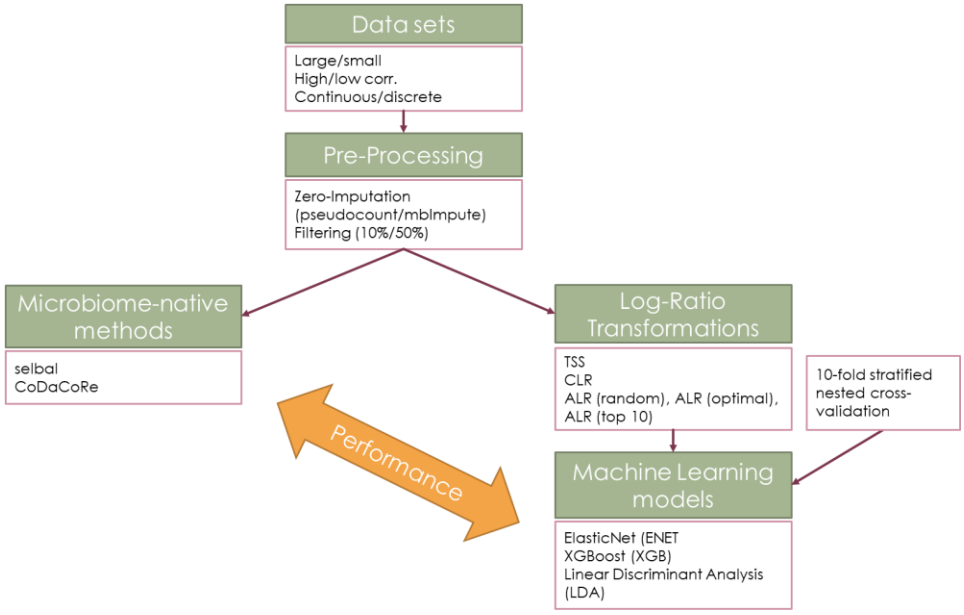

Figure 5: Proposed Pipeline

The graph shows the proposed pipeline for the benchmarking project. Data sets will be collected by their characteristics large/small, high/low correlations and continuous/discrete variables. Afterwards, data sets will be pre-processed by zero-imputation methods and filtering. Microbiome-native methods will be employed and compared to the data being log-transformed and used in machine learning models.

The goal of this pipeline is to compare the performance of classic machine learning models after different log-ratio transformations with microbiome-native and automated algorithms such as *selbal* and *CoDaCoRe*. This way it will become clearer if a scientist should make the effort in learning about machine learning methods, when automated algorithms perform well and no prior machine learning knowledge is necessary.

However, several steps are necessary before and after to assess performances and biases and to find out in which way log-ratio transformations influence the performance. Unfortunately, procedures and methodology varied greatly throughout all introduced papers with only a small number of common features. The general pipeline will be constructed of the following building

Commented [OA20]: I guess, a graphical abstract-like figure would be good and would ease understanding the following steps.

Commented [OA21]: And microbiome-native methods such as selbal and codacore

blocks: Pre-processing, Transformation, Standard Microbiome Approaches, and Machine Learning Models. For each building block methodology will be proposed and discussed and statistical indicators introduced to assess the performance of each building block. The introduced pipeline is an initial proposal and can be updated throughout the project.

DATA SET

The data sets will be compiled out of the Estonian Biobank microbiome cohort (EstMB) data. It includes 2509 individuals with several phenotypical markers collected over time. Additional possible considerations are the PCOS data set from Kreete et al. (2020) and the colorectal cancer data set, which is quite well-known in the microbiome community.

There are three common types of analyses conducted in microbiomics with machine learning approaches (Marcos-Zambrano et al. 2021): (i) classification and prediction of microbial taxa, (ii) prediction of host phenotype, and (iii) usage of microbial communities for understanding disease mechanisms (i.e., biomarker-finding).

To keep the size of this project manageable, the focus will be on prediction and classification tasks. This partially includes feature selection, e.g., in ElasticNet (ENET) and RF models. Additionally, as several authors pointed out (Quinn and Erb 2020; Gloor et al. 2017), machine learning performance is also influenced by data size. Therefore, data sets should be chosen accordingly to include direct comparison of performances of small and large data sets. Furthermore, phenotype variables with high and low known correlations between microbiome and host will be chosen, as well as continuous and discrete predictive variables.

PRE-PROCESSING

Important data pre-processing steps are potential filtering steps. As microbiome data usually has a lot of features, the computational work can be taxing. Therefore, filters will be applied before zero-imputation. In this benchmarking project, taxa with 0 abundance in in ≥10% of samples will be discarded. Additionally, a filter of ≥50% of 0 abundance in samples will be applied.

One of the main problems of microbiome data is its sparse nature. When working with relative abundances this is annoying but doesn't have to any mathematical consequences. On the other hand in log-ratio transformation zeros lead to problems, as log(0) is undefined. Therefore, one of the first steps after filtering and before log-ratio transformation is zero-imputation. Introduced by

Commented [OA22]: The choice of datasets used definitely needs discussion, but hare are a few others to consider:
1) Kreete's PCOS data from this paper: https://www.researchgate.net/publication/347017762_The_Gut_Microbiome_in_Polycystic_Ovary_Syndrome_and_Its_Association_with_Metabolic_Traits
2) Well-known colorectal cancer paper: https://www.nature.com/articles/s41591-019-0406-6

Commented [OA23]: Phenotypes/target variables

Commented [OA24]: This applies to log-ratio transformations. The standard, mentioned above, is to use relative abundances, where zeros are not replaced.

Commented [JN25R24]: Why are they not also replaced in relative abundances?

(Palarea-Albaladejo and Martín-Fernández 2015) is pseudocount. It has been frequently used for statistical analysis of microbiome data. It adds a pseudo-count of 1E-05 to avoid non-finite values resulting from log(0). An alternative to zCompositions will be mbImpute. The mbImpute method identifies and corrects the zeros and low counts that are unlikely biological in microbiome taxon count data. The goal of mbImpute is to provide a principled data-driven approach to relieve the microbiome data sparsity issue due to prevalent non-biological zeros (Jiang et al. 2021). As it hasn't been compared to pseudocount yet, it will be an interesting addition to this benchmarking project.

Commented [OA26]: Have you tested it, does it leave some zeros in the end?

Commented [JN27R26]: No, I have never used either of those methods and just used explanations given from the respective papers, but how they are working is actually still a mystery to me

STANDARD MICROBIOME APPROACHES

Using tested and published libraries for microbiome analysis is the easiest way to reduce human error and improve quality management. Two approaches are used frequently, and they will present the baseline comparison if one should use those packages or a machine learning model. One is called *selbal* and was proposed by Rivera-Pinto et al. (2018). It is based on standard generalized linear models. The second one will be *CoDaCoRe* proposed by (Gordon-Rodriguez et al. 2021), which is based on random forest analysis.

TRANSFORMATIONS

As mentioned in the introduction, choosing a log-ratio is not an easy decision. In order to stay with the goal of improving quality management and reducing human error, ILR will left out, as it is the most difficult one to work with and interpret. Similarly, pair-wise log-ratio transformations will also not be tested, as they are very computationally taxing. It has been decided to use TSS (total sum scaling transformation), which is standard relative abundance data, and compare it to CLR and ALR transformed data.

As ALR would be the most promising log-ratio transformation in terms of interpretability and its closeness to ILR, we will compare ALR transformation in three ways: (i) a random reference will be picked as denominator, to create a worst possible performance for ALR, (ii) find the most optimal ALR via pair-wise log-ratio computations, (iii) use Greenacre et al. (2021) proposed way of finding a reference. The data matrix will be filtered for top 10 references that show the highest abundance and similarly lowest variance, with a following Procrustes analysis to assess their geometry. This should lead to a computationally balanced approach in finding a good ALR log-ratio transformation. It should be noted that the introduced log-ratio transformations will only be

Commented [OA28]: This selection should also depend on the work of Greenacre?

Commented [JN29R28]: I thought we deliberately wanted to pick just a random reference, to simulate the worst possible performance?

Commented [OA30]: I would go with more depending on the analysis. Can be clarified later

applied for machine learning models and their impact directly compared to the standard microbiome approaches of *selbal* and *CoDaCoRe*.

MACHINE LEARNING MODELS

Using Machine Learning models always includes some form of cross-validation to ensure a low bias in machine learning models. One of the recurring methodologies is nested cross-validation algorithm. This is an approach to model hyperparameter optimization and model selection that attempts to overcome the problem of overfitting the training data set which often happens in standard cross-validation procedures (Cawley and Talbot 2010). Typically, the k-fold cross-validation procedure involves fitting a model on all folds but one and evaluating the fit model on the holdout fold. Each training dataset is then provided to a hyperparameter optimized procedure that finds an optimal set of hyperparameters for the model (Cawley and Talbot 2010). Additionally, stratification will be included. In stratified nested cross-validation during splitting of data into folds it is ensured that each fold has the same proportion of observations to ensure balancing. Here, a 10-fold stratified nested cross-validation protocol will be implemented, as it is standard now in various microbiome analyses (Marcos-Zambrano et al. 2021; Wirbel et al. 2019).

Tsamardinos et al. (2015) showed that a stratified nested cross-validation algorithm shows the least bias compared to standard cross-validation algorithms. They also propose to always include repetitions of inner CV loop for small data sets to reduce variances (Tsamardinos et al. 2015). Their computation of bias could be implemented as a control before feeding the data into machine learning models. The bias is computed as $L_{(hold\text{-}out)} - L_{(estimation),}$ with $L_{(hold\text{-}out)}$ being the performance of 70% of the data set, whereas 30% of each data set were used for sub-sampling (here n = 30) and training of the model (Tsamardinos et al. 2015).

It is a general consensus in the statistical community that most problems can be described via classical machine learning models (Marcos-Zambrano et al. 2021). Therefore, this pipeline will only include standard and most frequently used models. In microbiome analyses, most applications for machine learning are classification tasks in supervised learning. Therefore, ElasticNet (ENET) will be used as regression model and XGBoost (XGB) as random forest approach, also to have a direct comparison to *selbal* and *CoDaCoRe.* Additionally, Linear Discriminant Analysis (LDA) will be conducted.

Commented [OA31]: I would restructure this paragraph so that the ML algorithms selected goes hand-to-hand with the transformations (because the transformations are not relevant to selbal and codacore). Then, this paragraph can describe the way models are built and evaluated: cross-validation schemes, metrics, model comparison, etc

Commented [OA32R31]: Regarding that, if you go with nested cross-vaidation (which is interesting to me as well, s I haven't used it and it makes a kind of paradigm shift for the traditional workflow), some schematic approach would be beneficial, because it is not a well used approach in microbiome sciences at least.

Commented [JN33R31]: I shifted the transformation paragraph down a bit but I wanted to keep it separated from the machine learning models because it is the core of this benchmarking project. But I see your point regardless

Commented [OA34]: Can have a reference to the COST paper as well that this is similarly the case for microbiome

Commented [OA35]: A non-linear tree-based approach?

Commented [JN36R35]: Did I confuse this? I thought XGBoost was linear

As most models will employ binary classification tasks, the following performance metrics will be proposed: steadily recurring performance metrics are of course AUROC, and Accuracy metrics, as well as MAE (mean absolute error).

To assess if the difference in model performances is statistically significant, Statnikov et al. (2013) employed Random Permutation testing. Additional methods mentioned in literature are McNemar's test, 5x2-fold cross-validation with modified paired students t-test and Wilcoxon signed-rank test.

Commented [OA37]: I would keep it simple and go with AUROC for classification and RMSE/MSE for regression as these are mostly supported by the ML packages etc.

Commented [JN38]: MSE not good when comparing performances between different data sets. Rather MAE (mean absolute error)

Commented [JN39]: Discussions here?

PUBLICATION BIBLIOGRAPHY

Aitchison, J. (1982): The Statistical Analysis of Compositional Data. In *J. Royal Statistical Society* 44, pp. 139–177.

Cawley, Gavin C.; Talbot, Nicola L. C. (2010): On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. In *Journal of Machine Learning Research* 11, pp. 2079–2107.

Coenders, Germa; Greenacre, Michael (2021): Three approaches to supervised learning for compositional data with pairwise logratios. Available online at http://arxiv.org/pdf/2111.08953v1.

Gloor, Gregory B.; Macklaim, Jean M.; Pawlowsky-Glahn, Vera; Egozcue, Juan J. (2017): Microbiome Datasets Are Compositional: And This Is Not Optional. In *Frontiers in microbiology* 8, p. 2224. DOI: 10.3389/fmicb.2017.02224.

Gordon-Rodriguez, Elliott; Quinn, Thomas P.; Cunningham, John P. (2021): Learning Sparse Log-Ratios for High-Throughput Sequencing Data.

Greenacre, Michael; Grunsky, Eric; Bacon-Shone, John; Erb, Ionas; Quinn, Thomas (2022): Aitchison's Compositional Data Analysis 40 Years On: A Reappraisal. Available online at http://arxiv.org/pdf/2201.05197v1.

Greenacre, Michael; Martínez-Álvaro, Marina; Blasco, Agustín (2021): Compositional data analysis of microbiome and any-omics datasets: a revalidation of the additive logratio transformation.

Griffith, Malachi; Walker, Jason R.; Spies, Nicholas C.; Ainscough, Benjamin J.; Griffith, Obi L. (2015): Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. In *PLoS Comput Biol* 11 (8), e1004393. DOI: 10.1371/journal.pcbi.1004393.

Jiang, Ruochen; Li, Wei Vivian; Li, Jingyi Jessica (2021): mbImpute: an accurate and robust imputation method for microbiome data. In *Genome biology* 22 (1), p. 192. DOI: 10.1186/s13059-021-02400-4.

Marcos-Zambrano, Laura Judith; Karaduzovic-Hadziabdic, Kanita; Loncar Turukalo, Tatjana; Przymus, Piotr; Trajkovik, Vladimir; Aasmets, Oliver et al. (2021): Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment. In *Frontiers in microbiology* 12, p. 634511. DOI: 10.3389/fmicb.2021.634511.

Palarea-Albaladejo, Javier; Martín-Fernández, Josep Antoni (2015): zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. In *Chemometrics and Intelligent Laboratory Systems* 143, pp. 85–96. DOI: 10.1016/j.chemolab.2015.02.019.

Quinn, Thomas P.; Erb, Ionas (2020): Interpretable Log Contrasts for the Classification of Health Biomarkers: a New Approach to Balance Selection. In *mSystems* 5 (2). DOI: 10.1128/mSystems.00230-19.

Quinn, Thomas P.; Erb, Ionas; Richardson, Mark F.; Crowley, Tamsyn M. (2018): Understanding sequencing data as compositions: an outlook and review. In *Bioinformatics (Oxford, England)* 34 (16), pp. 2870–2878. DOI: 10.1093/bioinformatics/bty175.

Rivera-Pinto, J.; Egozcue, J. J.; Pawlowsky-Glahn, V.; Paredes, R.; Noguera-Julian, M.; Calle, M. L. (2018): Balances: a New Perspective for Microbiome Analysis. In *mSystems* 3 (4). DOI: 10.1128/mSystems.00053-18.

Tolosana-Delgado, Raimon; Khodadadzadeh, Mahdi; Talebi, Hassan (Eds.) (2019): On machine learning algorithms and compositional data.

Tsamardinos, Ioannis; Rakhshani, Amin; Lagani, Vincenzo (2015): Performance-Estimation Properties of Cross-Validation-Based Protocols with Simultaneous Hyper-Parameter Optimization. In *Int. J. Artif. Intell. Tools* 24 (05), p. 1540023. DOI: 10.1142/S0218213015400230.

Wirbel, Jakob; Pyl, Paul Theodor; Kartal, Ece; Zych, Konrad; Kashani, Alireza; Milanese, Alessio et al. (2019): Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. In *Nature medicine* 25 (4), pp. 679–689. DOI: 10.1038/s41591-019-0406-6.

Zhang, Mo; Shi, Wenjiao (2019): Systematic comparison of five machine-learning methods in classification and interpolation of soil particle size fractions using different transformed data.