BIORANKINGS®
TECHNICAL REPORT SERIES

Finding Distinct Subgroups of Samples Using Microbiome Taxa Count Data

A recent paper by Shaikh and Beyene raised our interest in revisiting having an objective way to decide how many subgroups exist in a microbiome dataset (Shaikh & Beyene, 2017). Many researchers ask if there are clinically or biologically important distinct subgroups of subjects defined by taxa communities in the data they collected. Perhaps the best-known example of this is enterotypes (Arumugam et al., 2011).

Shaikh and Beyene proposed using ***finite mixture models*** combined with the ***Dirichlet-multinomial distribution*** for this problem. In this first *BioRankings' Technical Report* we show how cluster analysis is highly subjective with results changing for different inputs, why arguments against the Dirichlet-multinomial distribution for microbiome data are wrong, how the finite mixture models operate, and finally an example of this analysis using HMP stool samples.

The Value of Objectivity in Data Analysis

Many researchers use nonparametric cluster analysis to find sample subgroups (Shannon, Culverhouse, & Duncan, 2003; Zhou et al., 2013). While a good exploratory tool, cluster analysis does not provide objective rigor that is desired in advanced statistical analysis, nor will it, or other exploratory methods, be acceptable by the FDA as statistical evidence. In **Figure 1** HMP stool samples are clustered based on count or proportion data and complete or single linkage clustering algorithms. Visually, these unlabeled dendrograms indicate different results suggesting a researcher's conclusion will be based on which subjective decisions are made concerning the analyses.

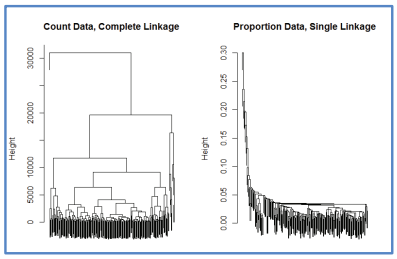

Figure 1 Two hierarchical clustering analyses of the same data produces different results.

Parametric models are an example of a class of objective statistical methods that should be used in microbiome data analysis. An example of a commonly used parametric model from introductory statistics is the *Normal Distribution* defined by two parameters – the mean and standard deviation – to describe 'bell-shaped' data. Parametric models let the analyst summarize data by estimating the parameters (known as *sufficient statistics*), calculate probabilities on the data (e.g., probability a student is 6 feet tall or taller), and make use of known mathematical properties of parametric models such *uniformly most powerful (UMP)* hypothesis tests which have the greatest power compared to any other test that might be used, convergence, and unbiasedness.

The objectivity and statistical rigor provided by parametric methods in data analysis is required by the FDA for biomarker validation and drug approval. Incorporating parametric analytics into your R&D pipeline early will save you time and frustration of having to go back and retrospectively apply these methods to the data in your NDA application. (FDA, 2017)

Parametric Model for Microbiome Data

Shaikh and Beyene start with a review of the literature on parametric models for microbiome data. They write in their introduction:

> *Processing microbiome data results in a long list of sequences for every sample processed. As mentioned, these sequences can be aggregated to create a table tallying how many of each OTU was found in each sample. This OTU table forms the basis of many analyses. One approach is to analyze the relative proportions of each OTU within the sample. A natural distribution to model proportions is the multinomial distribution, a generalization of the binomial distribution. Unfortunately, microbiome data exhibit dramatic overdispersion for such a simple distribution, making any results using just the multinomial questionable. One approach to handle this overdispersion is through a more flexible variant of the distribution, the Dirichlet-Multinomial (DM) distribution(Moismann, 1962). Several uses and extensions to the DM have been introduced in the microbiome literature for a variety of useful applications. A hypothesis test to see if the data are indeed suitably modelled by a DM was previously outlined (La Rosa et al., 2012). The DM has been used in sparse variable selection (Chen & Li, 2013), and been*

advocated as a realistic means of simulating microbiome data (Chen et al., 2012). The DM has also been used in clustering algorithms to find structure using finite mixture models (FMMs) (Holmes, Harris, & Quince, 2012).

Many researchers have resisted the use of the Dirichlet-multinomial distribution for microbiome data based on fundamental errors in their statistical thinking. In one paper the authors mistakenly claim it is inappropriate since positive correlations between taxa were observed when correlations between taxa must be negative in this type of model (Mandal et al., 2015). *Observing positive correlations is not a proof that the DM is wrong for microbiome data.* Positive correlations could easily be seen when the DM model is correct due to the statistical ideas of *convergence, data representation,* and *independent and identically distributed (iid)* data.

Convergence says that as a sample size gets larger the estimate of statistics (e.g., correlation) becomes more accurate. For a small sample size random variability in the count data will likely produce positive correlations which dissappear as the sample size increases. **Figure 2** shows this where positive correlations are seen for N = 100 DM simulated data points, but fewer as N = 1,000 as expected due to convergence. ***The presence of positive correlations does not negate the appropriateness or use of the Dirichlet-multinomial.***

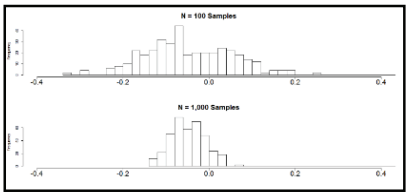

Figure 2 Increasing sample sizes shows convergence onto true correlation coefficients.

The natural representation of microbiome data is as ***compositional data (Aitchison, 2003)***. However, researchers who use an incorrect **data representation** can also be misled to see positive correlations between taxa and incorrectly conclude the DM model is wrong. In an experiment the total number of reads varies across samples. **Figure 3** shows how this can occur. In the left plot the total read count in each sample is not taken into account and a plot of the reads of taxa A and B show a correlation of 1. The appropriate representation is the ratio of taxa A to B which is shown in the right plot where the X axis is the total count and the Y axis the reads for taxa A and B. For each sample the ratio is 2:1 indicating a correlation of 0. ***The DM model uses the read counts as the natural representation of the microbiome data avoiding the need to use a subjective transformation of the data which can lead to different conclusions depending on which transformation is selected.***

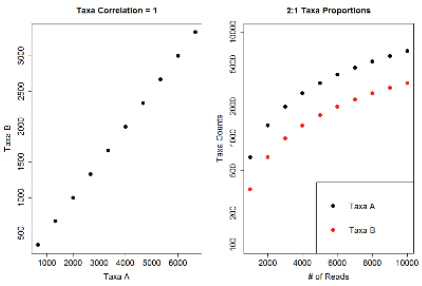

Figure 3 Incorrect data representation can easily lead to incorrect conclusions.

Independent and Identically Distributed

The **iid** assumption makes two claims about data. The first, ***independence***, is that the values of one sample's data are not influenced by nor influence the values of another sample. The second, ***identically distributed***, assumes that the data are sampled from the same probability model. This second issue is perhaps less intuitive than independence so is illustrated here by a clinical example. Suppose a treatment is given which has a 50% cure rate and a group of random patients have been selected to receive the treatment. If the probability of cure is 50% for each patient, we say they are identically distributed. If it is found that males have a lower cure rate of 40% versus females 50% then males and females are not identically distributed – males come from a Bernoulli distribution with a probability of 0.4 of cure, and females from a different Bernoulli distribution with a probability of 0.5 of cure. However, males are identically distributed and females are identically distributed.

Assuming i.i.d. on datasets is routinely done in most statistical hypothesis testing and is appropriate for the microbiome and Dirichlet-multinomial. In cases where samples are not identically distributed but may come from 2 or more distributions then positive correlations might be observed. The following section discusses how this can be tested.

Finite Mixture Models

Figure 4 illustrates ***finite mixture models*** with an example of two Normal Distributions centered at X = -1 and X = 1, shown by the colored dashed lines. If the data are combined and analyzed together the density curve, represented by the histogram and black line would suggest one distribution centered at X = 0 exists.

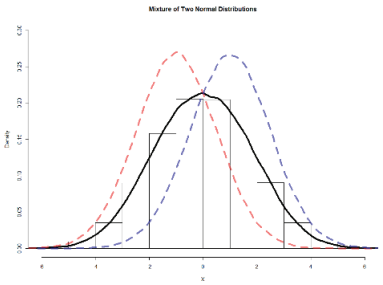

Figure 4 Illustration of the mixture of two Normal distributions.

Finite mixture model methods provide the formal statistical machinery and algorithms to test for mixtures of distributions in data and identify which samples belong to which subgroup (McLachlan & Peel, 2000; Schlattmann, 2009). Using this approach combined with the Dirichlet-multinomial model, the analyst can partition the microbiome samples into subgroups where the data within each group are iid. Since it is automatic and objective, the results will not depend on the input.

Finding Subgroups in Microbiome Data Using Objective Finite Mixture Models

We fit ***finite mixture models using the Dirichlet-multinomial distribution*** to 205 HMP stool samples at the phylum level. Since counts are the natural representation for the DM model, no transformation of the data was required.

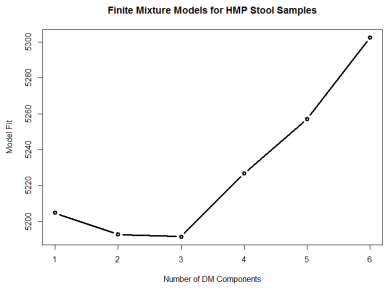

Figure 5 Model goodness-of-fit indicates the HMP stool samples consist of 3 distinct and non-overlapping subgroups.

To decide how many distributions are in the data, the ***fmm-dm*** was run for 1-6 possible subgroups. Each of the six runs was scored based on the model fit with the minimum score indicating the number of subgroups. **Figure 5** shows there are 3 distinct subgroups of stool samples, where each of the subgroups is modeled by a separate DM distribution.

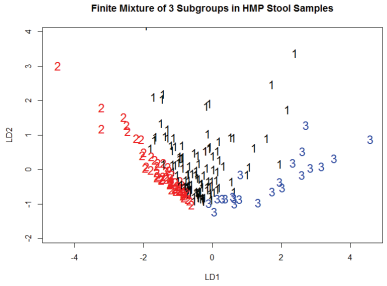

Figure 6 Projection of the HMP stool samples labeled by subgroup membership shows clear separation.

Each sample is assigned to one of the 3 distributions. **Figure 6** shows the first and second linear discriminant analysis scores plot with the 3 groups color coded. Clear boundaries separating the groups are evident suggesting possible clinical or biological importance distinguishing these samples. Since each group is represented by a DM distribution, the estimate of the taxa proportions within each group could be estimated using maximum likelihood to identify how these groups differ. **Figure 7** shows Group 1 having moderate Bacteroidetes and Firmicutes, Group 2 having high Bacteroidetes and low Firmicutes, Group 3 having low Bacteroidetes and high Firmicutes. The proportion of the 205 HMP stool samples falling into groups 1, 2, and 3 are 55%, 33%, and 12%, respectively.

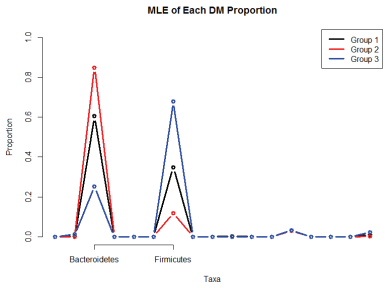

Figure 7 Plot of taxa proportion differences defining the 3 subgroups. (The non-differentiating taxa are not labeled.)

Conclusion

Objective statistical tools based on parametric models will almost surely be required for FDA approval of microbiome biomarkers and drugs. Developing your R&D pipeline using statistical models such as presented here should make future interactions with the FDA more seamless. An example is ***finite mixture models using the Dirichlet-multinomia***l as an objective and automatic way for finding subgroups in contrast to cluster analysis where the analysts' choices on methods can significantly impact the results they obtain.

BioRankings' Technical Report Series

BioRankings' mission is to help biomedical researchers move their technology from the lab to clinical applications using statistically valid analytical tools for efficient study designs, correct data analyses and conclusions, and rigorous and objective decision making for designing follow-up studies and eventual FDA approval.

To help achieve its mission, BioRankings publishes a Technical Report series focused on applying various statistical methods to real data analyses. Written for understanding by scientists and administrators, these reports will provide an intuitive understanding of the analyses leaving the statistical details to other publications.

For more information, contact BioRankings at 314-704-8725 or bill@biorankings.com.

Aitchison, J. (2003). The statistical analysis of compositional data. Caldwell, N.J.: Blackburn Press.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., . . . Bork, P. (2011). Enterotypes of the human gut microbiome. Nature, 473(7346), 174-180. doi:10.1038/nature09944

Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., . . . Li, H. (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics, 28(16), 2106-2113. doi:10.1093/bioinformatics/bts342

Chen, J., & Li, H. (2013). Variable Selection for Sparse Dirichlet-Multinomial Regression with an Application to Microbiome Data Analysis. Ann Appl Stat, 7(1). doi:10.1214/12-AOAS592

FDA. (2017). Drug Development and Review Definitions. Retrieved from https://www.fda.gov/drugs/developmentapprovalprocess/howdrugsaredevelopedandapproved/approvalapplications/investigationalnewdrugindapplication/ucm176522.htm

Holmes, I., Harris, K., & Quince, C. (2012). Dirichlet multinomial mixtures: generative models for microbial metagenomics. PLoS One, 7(2), e30126. doi:10.1371/journal.pone.0030126

La Rosa, P. S., Brooks, J. P., Deych, E., Boone, E. L., Edwards, D. J., Wang, Q., . . . Shannon, W. D. (2012). Hypothesis testing and power calculations for taxonomic-based human microbiome data. PLoS One, 7(12), e52078. doi:10.1371/journal.pone.0052078

Mandal, S., Van Treuren, W., White, R. A., Eggesbo, M., Knight, R., & Peddada, S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb Ecol Health Dis, 26, 27663. doi:10.3402/mehd.v26.27663

McLachlan, G. J., & Peel, D. (2000). Finite mixture models. New York: Wiley.

Moismann, J. (1962). On the Compound Multinomial Distribution, the Multivariate β-Distribution, and Correlations Among Proportions. Biometrika, 49(1/2), 65-82.

Schlattmann, P. (2009). Medical applications of finite mixture models. Berlin: Springer.

Shaikh, M. R., & Beyene, J. (2017). Statistical models and computational algorithms for discovering relationships in microbiome data. Stat Appl Genet Mol Biol, 16(1), 1-12. doi:10.1515/sagmb-2015-0096

Shannon, W., Culverhouse, R., & Duncan, J. (2003). Analyzing microarray data using cluster analysis. Pharmacogenomics, 4(1), 41-52. doi:10.1517/phgs.4.1.41.22581

Zhou, Y., Gao, H., Mihindukulasuriya, K. A., La Rosa, P. S., Wylie, K. M., Vishnivetskaya, T., . . . Weinstock, G. M. (2013). Biogeography of the ecosystems of the healthy human body. Genome Biol, 14(1), R1. doi:10.1186/gb-2013-14-1-r1