**Distance and Dissimilarity Measures**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Formula** | **Values** | **Description** | **Notes** |
| Bray-Curtis dissimilarity | Also written as: | weighted | For each species in common, sum the minimum abundance over total abundances in both samples. | Bounded between 0 and 1. Quantitatively related to the Canberra distance and Sørensen and Kulczynski dissimilarity indices. Emphasizes species in common.  Not sensitive to large values or outlier values.  Could also use an unweighted generalization of Bray-Curtis. |
| Kulczynski dissimilarity |  | weighted | For each species in common, sum the minimum over each sample total abundance. | Bounded between 0 and 1.  Could also use an unweighted generalization of Kulczynski. |
| Manhattan distance |  | weighted | Sum of the lengths of n perpendicular line segments connecting two points in n-dimensional space. | Non-squared version of Euclidean distance, making it less susceptible to outlier species and/or samples.  Manhattan, Kulczynski, and Bray-Curtis are functionally equivalent with standardization. |
| Cosine dissimilarity |  | weighted | Measures the angle between two points from the origin of the n-dimensional feature (OTU/taxa) space. | Bounded from -1 to 1, such that subtracting the cosine dissimilarity from 1 yields the cosine similarity measure.  Popular measure in ML applications such as data mining. |
| Jaccard dissimilarity |  | unweighted | Measures overlap of two communities by the ratio of shared species over all species found. | Bounded between 0 and 1. Can be likened to an unweighted Bray-Curtis dissimilarity. |
| Unweighted UniFrac distance |  | unweighted | Measures the amount of shared evolutionary history between two communities. | Estimates relatedness of samples by the phylogenetic relatedness of species found in the samples. |
| Weighted UniFrac distance |  | weighted | Measures the amount of shared evolutionary history weighted by relative abundance of these species in communities. | Assumes samples are provided in terms of relative abundance (rarefied or normalized data).  Also of note is Generalized UniFrac, deemphasizing more abundant species in the weighted UniFrac formula. |
| Euclidean distance |  | weighted | Geometric distance between two points in n-dimensional space. | Sensitive to outliers (Supp Fig 1), squared term heavily penalizes large distances. |
| Aitchison distance | \*CLR-transformed data | weighted | Euclidean distance of CLR-transformed data. | Aitchison and robust Aitchison differ in their attention to the non-zero values of the community matrix. **Robust Aitchison** uses a robust form of CLR and later robust PCA to approximate the geometric mean without the zero values (missing species) and ordinate based on the non-zero values using matrix completion. |
| Chi-Squared distance | \*Chi-Sq-transformed data | weighted | Euclidean distance between Chi-square transformed communities. | Sensitive to outliers in individual species.  This description is based on the vegan package (R) implementation, and one should note that all distances will change with the addition of new samples. Python implementations may differ from this approach. |

**Table 1.** An overview of popular distance metrics and dissimilarity measures, including the defined mathematical functions and a written description of the functions. Symbols in the functions are defined as follows. A, B: values of taxa for samples A and B respectively; a, b: unweighted, presence/absence values for samples A and B respectively, N: number of taxa in the dataset; *l*: branch lengths for UniFrac.

**Visualization Methods**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Algorithm Description** | **Assumptions** | **Notes** |
| Principal Coordinates Analysis (PCoA) | Eigen decomposition of the pairwise sample dissimilarity/distance matrix.  Orders the eigenvectors by decreasing eigenvalues, so that the most inter-sample variation is described by the first dimension and least variation in the last dimension. | PCoA assumes linear combinations of pairwise distances, due to the nature of eigen decomposition. | PCoA and PCA (Principal components analysis) are equivalent when PCoA is performed on Euclidean distances.  PCoA and CA are equivalent when PCoA is performed on Chi-sq distances.  PCoA is a form of Multidimensional Scaling (MDS). |
| Non-metric Multidimensional Scaling (NMDS) | Converts dissimilarity/distance matrix into ranks of dissimilarity.  Iteratively constructs configurations of all samples in *m* dimensions (specified by user), reshuffling until the stress (difference between the dissimilarity rank and the configured dissimilarity) is minimized. | Sample relationships can be adequately represented in a few dimensions. | NMDS is not very easily reproducible, may change on re-analysis due to the iterative algorithm.  Since NMDS converts distances to ranks, it is robust to non-linearity and may be less sensitive to outliers. |
| t-distributed Stochastic Neighbor Embedding (t-SNE) | Obtains probabilities of each sample according to the t-distribution of other samples around it.  Using these probabilities, gradient descent is used to place the points in the reduced *m-*dimensional space (usually 2-3 dimensions), using Kullback-Leibler (KL divergence) as the cost function between probabilities. | Assumes an underlying manifold, a flexible plane over which the data falls. Trusts local relationships to approximate global trends. | Manifold learning is best suited to uncovering gradients, due to manifold assumption.  t-SNE may reveal true group clusters but visually exaggerate these clusters due to the sensitivity to local distances/dissimilarities.  Perplexity parameter essentially tunes the neighborhood size, changing the degree of locality to trust and thereby changing clusters and/or proximity of clusters. |
| Correspondence Analysis (CA) | Eigen decomposition performed on Chi-squared transformed distance matrix. | Assumes samples can be represented as differences from expected proportions of taxa given the rest of the table. | CA maximizes the similarity of taxon frequencies between samples, which may be helpful in studies with categorical variables.  CA has been known to produce the “arch effect” visual artifact.  Detrended CA uses reciprocal averaging to resolve the “arch” or “horseshoe” effect, but has been criticized for distorting the data. |
| Constrained Ordination | Regress the taxa abundances on environmental variables to determine predicted values. Find residuals between true taxa abundances and environmentally-influenced predictions.  Perform PCoA on predicted values and residuals separately to obtain environmental variable and sample ordinations, respectively. | Assumes environmental variables and samples can share an ordinal space. | Redundancy analysis (RDA) is linear and implicitly Euclidean-based (uses PCA).  Canonical Correspondence Analysis uses CA in place of PCoA, may also be prone to the “arch.” |

**Table 2.** Description of commonly used visualization methods in microbiome studies. Notes on the considerations when using each algorithm.