

Metagenome Ordination in IMG

Contents

1	Principal Component Analysis	2
2	Principal Coordinates Analysis	4
3	Non-metric Multidimensional Scaling	6

1 Principal Component Analysis

PCA Input: a $m \times n$ matrix of m metagenomes and n functions/taxa

PCA Output: a 2 or 3-D plot of metagenomes

Principal Components Analysis is an ordination tool for exploratory data analysis which reduces the dimension of a data set such that it can be visualized in a 2 or 3-D plot [1][2]. The method creates synthetic variables which are linear combinations of the original variables, and can be plotted on their corresponding orthogonal principal axes. The principal components can explain much of the variation found in the data set. The most useful feature of PCA, compared to other ordination techniques, is that it can show what variables drive the separation of objects in the plot.

When comparing metagenomes, we define the objects in ordination as metagenomes, and the variables as functional or taxonomic classifications. We refer to the list of gene counts in either a functional (Pfam, COG, etc) or taxonomic (class, family, etc.) classification as a metagenome's profile. We are interested in seeing what metagenomes cluster together, which suggests profile similarity. We are also interested in viewing gradients, which are varying aspects of the environment related to the profiles.

It is difficult to prove that metagenomic data sets meet the assumptions needed for appropriate use of PCA. Functions/taxa need to be linearly related. For the principal axes to be an informative summary of the data, a change in one function/taxon must result in a linear change of another function/taxon. Also, the variables need to be normally distributed. With the prevalence of zeros in metagenome profiles, this assumption is not met in most cases.

However, PCA is a geometric technique and not a statistical test. So, even though the assumptions are not strictly met, it is possible to find meaningful information in the ordination. If the first few principal components explain much of the variance, PCA is an informative representation of the objects. As an exploratory data tool only, PCA in IMG can be used to form hypotheses that can be later tested in properly designed statistical studies.

In order to account for the large variance in metagenome profiles, every metagenome is normalized by dividing each gene abundance by the total number of genes in the metagenome. The normalized abundances for each metagenome sum to 1. For ex-

ample, if metagenome A has profile 0,100,50,150, and metagenome B has profile 0,50,25,75, both profiles will be normalized such that they have equal profiles of 0,0.33,0.17,0.5.

Although the normalization reduces the Euclidean distance between metagenomes that have similar profiles but vastly different gene sizes, it does not solve the double zeros issue [3]. If two metagenomes have similar absence profiles, meaning they have no genes from the same functions/taxa, they will possibly be close together in the PCA ordination. However, it is likely that an absence of a function/taxon in a metagenome is due to lack of read depth and/or coverage [4]. We are more interested in viewing what metagenomes are more similar based on their present functions/taxa, in which case PCA is not the most appropriate tool for visualizing metagenomic data sets.

2 Principal Coordinates Analysis

PCoA Input: a $n \times n$ matrix of metagenome dissimilarities calculated from a $m \times n$ matrix of m metagenomes and n functions/taxa

PCoA Output: a 2 or 3-D plot of metagenomes

Principal Coordinates Analysis (PCoA) [5][6] is an eigenanalysis algorithm like Principal Component Analysis (PCA). Whereas PCA only uses the Euclidean distances between objects to perform an ordination in reduced space, PCoA performs an ordination on any user-selected dissimilarity measure. If Euclidean distance is chosen, PCoA gives the same solution as PCA. If an appropriate measure is selected, no data assumptions need to be verified before analysis.

One of the best measures to use with raw abundances is the Bray-Curtis dissimilarity coefficient [3][7], which we use to measure the compositional dissimilarity between two metagenomes. If we have a $m \times n$ abundance matrix X , with m metagenomes and n functions/taxa, we calculate the Bray-Curtis dissimilarity between metagenome j and metagenome k as [8]:

$$d_{jk} = \frac{\sum_{i=1}^n |x_{ij} - x_{ik}|}{\sum_{i=1}^n (x_{ij} + x_{ik})}$$

If two metagenomes have exactly the same profile, their index is 0, which is the smallest (most similar) Bray-Curtis index. If two metagenomes share no functions or taxa in their profile, their index is 1, representing the largest measure of Bray-Curtis dissimilarity.

The Bray-Curtis index is a semi-metric distance, meaning it does not exhibit the properties of the triangle inequality. Thus, negative eigenvalues may result from the PCoA. However, if the negative eigenvalues do not occur in the first few principal coordinates, the ordination may be meaningful in some cases [3].

Although performing PCoA with the Bray-Curtis index gives a much more appropriate representation of the relationship between metagenome profiles than PCA, it is difficult to recover the functions/taxa contributing to the principal coordinates. Unlike PCA, the new variables are complex functions of the original variables, not linear combinations.

PCoA in IMG is a data exploratory tool only, and may be used to form hypotheses that can later be tested in properly designed statistical studies.

3 Non-metric Multidimensional Scaling

NMDS Input: a $n \times n$ matrix of metagenome dissimilarities calculated from a $m \times n$ matrix of m metagenomes and n functions/taxa

NMDS Output: a 2 or 3-D plot of metagenomes

Non-metric Multidimensional Scaling (NMDS) [9][10][11][12] is an iterative ordination technique that preserves the rank order correlation between objects, rather than their linear correlation. It is better than PCoA (Principal Coordinates Analysis) at representing relationships between objects because of this model flexibility [3]. However, since it is an iterative algorithm two problems may arise. First, there is no guarantee that the solution found is best, as the algorithm could return on a local minimum. Second, many iterations on large matrices is computationally intensive, so NMDS can take an extended amount of time to return a solution.

Before running NMDS, both the dissimilarity metric and the number of dimensions of the solution need to be chosen. Unlike PCA and PCoA, which find solutions in n and $n-1$ dimensions respectively, NMDS can find a solution in 1 to $n-1$ dimensions. NMDS in IMG is displayed in 3 dimensions because the solution is easily viewed in a 3-D plot, not because 3 dimensions necessarily gives the best solution. Like PCoA, NMDS has no data assumptions, and can be performed on any dissimilarity measure.

One of the best measures to use with raw abundances is the Bray-Curtis dissimilarity coefficient [3][7], which we use to measure the compositional dissimilarity between two metagenomes. If we have a $m \times n$ abundance matrix X , with m metagenomes and n functions/taxa, we calculate the Bray-Curtis dissimilarity between metagenome j and metagenome k as[8]:

$$d_{jk} = \frac{\sum_{i=1}^n |x_{ij} - x_{ik}|}{\sum_{i=1}^n (x_{ij} + x_{ik})}$$

If two metagenomes have exactly the same profile, their index is 0, which is the smallest (most similar) Bray-Curtis index. If two metagenomes share no functions or taxa in their profile, their index is 1, representing the largest measure of Bray-Curtis dissimilarity.

Because the Bray-Curtis index is used in the ordination, rather than the abundances, it is difficult to recover the functions/taxa contributing to the separation

of the metagenomes. Unlike PCA, the new variables are complex functions of the original variables, not linear combinations.

NMDS is the most appropriate tool in IMG for reducing the dimensions of metagenomic functional/taxonomic profiles for comparative analysis. However, both PCA and PCoA find a solution more quickly on some datasets.

NMDS in IMG is a data exploratory tool only, and may be used to form hypotheses that can later be tested in properly designed statistical studies.

References

- [1] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.
- [2] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441,498–520, 1933.
- [3] P. Legendre and L. Legendre. *Numerical Ecology*. Developments in Environmental Modelling. Elsevier, 1998.
- [4] V. Kunin, A. Copeland, A. Lapidus, K. Mavromatis, and P. Hugenholtz. A bioinformaticians guide to metagenomics. *Microbiology and molecular biology reviews MMBR*, 72(4):557–78, 2008.
- [5] J. C. Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53:325–338, 1966.
- [6] W.S. Torgerson. *Theory and methods of scaling*. Wiley, 1958.
- [7] A.E. Magurran. *Measuring Biological Diversity*. Blackwell, 2004.
- [8] J.T. Curtis J. Roger Bray. An ordination of the upland forest communities of southern wisconsin. *Ecological Monographs*, 27(4):325–349, 1957.
- [9] R.N. Shepard. The analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika*, 27:125–139, 1962.
- [10] J.B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. *Psychometrika*, 29:1–27, 1964.
- [11] J.B. Kruskal. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29:115–129, 1964.
- [12] R.N. Shepard. Metric structures in ordinal data. *Mathematical Psychology*, 3:287–315, 1966.