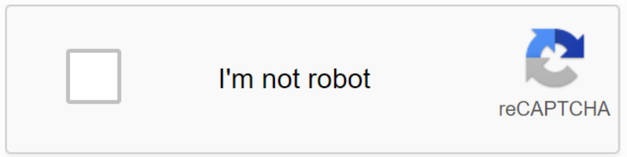

Continue

Ashp midyear clinical meeting 2017

In fact, both PCoA and PCA charts are used scatterpics to show the results of PCoA and PCA, PCoA and PCA are accurate data-dimensional analysis methods. By the way, this holiday season, I wish you and scientific research can have a happy time and happy ending. Main component analysis (main component analysis, PCA) is a method of statistical analysis that simplifies data sets. It uses orthographic transformations to linearly transform observations of a number of potentially relevant variables that project the values of a series of linearly irrelevant variables called main components. The main component can be specifically seen as a linear equation containing a series of linear coefficients to indicate the direction of the projection (see figure). The PCA is sensitive to normalization or preprocessing (relative scaling) of the original data. PCA is the simplest way to analyze multiple statistical distributions by function volume. Often, this operation can be seen as a way to expose the internal structure of the data and thus better interpret the data variables. PCA Schematic Principal Coordinates Analysis, PCoA, or classic multidimensional scaling, is used to study similarities between data. Both PCoA and PCA are ways to reduce the data dimension, but the difference is that PCA is based on the original matrix, while PCoA is based on the distance matrix calculated from the original matrix. Pca therefore tries to preserve the variation in the data so that the position of the points does not change, while pcoa tries to ensure that the original distance ratio does not change, that is, to make the distance between the original data points as relevant as possible to the distance between the points in the projection, that is, the points in the result (see figure). PCoA diagram How to perform PCA and PCoA analysis R There are many packages that provide PCA and PCoA, such as the commonly used ADE4 package. This article will be based on the package for PCA and PCoA analysis, the data is brought with deug, this data gives 104 students 9 courses in performance (see screenshot) and comprehensive assessment. The overall rating is as follows: A, A, B, B-, C-, D. Let's look at whether such a comprehensive assessment is reasonable through PCA and PCoA, and whether 104 students are reasonably assigned to different groups (one group per level) based on these nine courses. Deug's 9 courses (1) PCA analysis and graphing PCA are based on raw data so PCA analysis can be done directly. Since the previously described scatterpicle drawing methods, here is no longer detailed, PCA analysis after we directly plot the results. library (ade4) library (ggplot2) library (RColor Brewer) data (deug) #PCA分析 pca<-dudi.pca (deug $, tab's FALSE, center s deug $cent, scan's fake) #坐标轴解释量 (first two axes) pca_eig <- (pca $eig) s 1:2 s/sum (pca $eig) #提取样本点坐标 (first two axes) sample_site <data.frame $li) 1:2) sample_site$names <-row names (sample_site) names (sample_site) (1:2) <-c ('PCA1', 'PCA2') #以最终成绩作为分组 sample_site$ library with level<factor (deug$result, levels c ('A', 'A', 'B', 'B', 'C-', 'D')) (ggplot2) pca_plot <-ggplot (sample_site, aes (PCA1, PCA2, color=level)) and theme_classic() #去掉背景框 geom_vline (xintercept s 0, color s 0, color s 0, color s 0, color s 0, color s 0, color s 0, color s 0, color s 0, color s 0 0 size s 0.4) s geom_hline (yintercept s 0, color', color', size 0.4) s geom_point (size s 1.5) s #可在这里修改点的透明度 scale_color_manual (values s brewer.pal (6, Set2) s #可在这里修改点的颜色 theme (panel.grid's element_line ('color', linetype s 2, size s 0.1), panel.background's element_rect (color's 'black', filled 'transparent'), legend.title's element_blank() s labs (x's pasta ('PCA1:', round (100 x pca_eig, 1.2), '%'), y's pasta ('PCA2:', round (100 x pca_eig s2,2), '%')) pca_plot generally looks good, that is, B and C students seem unreacted. (2) PCoA Analysis and Graphing library (ade4) library (ggplot2) library (RColor Brewer) library (vegan) #用于计算距离 data (deug) tab<-deug $tab.dist<-vegist (tab, method,'euclidean') #基于euclidean距离 <-dudi.pco scan s FALSE, nf = 3) #坐标轴解释量 (first two axes) pcoa_eig <- (pcoa $eig) s 1:2? / sum (pcoa$ eig) #提取样本点坐标 (the first two axes) sample_site <- data.frame ((pcoa$li) s1:2 s sample_site$names < - row names (sample_site) names (sample_site) 1:2) <-c ('PCoA1', 'PCoA2') #以最终成绩作为分组 sample_site$level<factor (deug$result, levels?c('A', 'A', 'B', 'B', 'C-', 'D'))) library (ggplot2) pcoa_plot <-ggplot (sample_site, aes (PCoA1, PCoA2, color=level)) theme_classic () #去掉背景框 geom_vline (xintercept s 0, color', size 0.4) s geom_hline (yceptinters 0, color's 'gray', size s 0.4) s geom_point (size s 1.0,0, color', size s 0.4) s geom_point (size s 1,1,0 5) s #可在这里修改点的透明度, size scale_color_manual (values s brewer.pal (6, Set2) s #可在这里修改点的颜色 theme (panel.grid's element_line (color's 'gray', linetype s 2, size s 0.1), panel.background's element_rect (color's 'black' , fill's 'transparent'), legend.title's element_blank()) s labs ('PCoA1:', The pcoa_eig results of PC pcoa_plot pcoa_eig A and PCoA are similar and sometimes one method can effectively separate the samples and the other cannot function well depending on the characteristics of the sample data. Because no microbiome data is available for sharing, a dataset was used for this performance. Typically, in microbiome studies, we perform PCA or PCoA analysis of data based on species abundance documents, or what we call beta-diveristy analysis, based on PCA or PCoA results to see if the disease group and control group can be separated to understand the overall changes in the microbiome. The distance matrix obtained by Unifrac analysis can be used for a number of analytical methods, and the evolutionary similarities and differences in microorganisms in different environmental samples can be visually displayed using the multivariable statistical method PCoA analysis. Primary co-images (PCoA) is a visualization method for surveying data equality or differentiation, and after sorting through a series of function values and function vectors, select the top 可以观察个体或群体间的差异。 分析软件:R 语言PCoA 分析啦作PCoA 图。 unifrac.pcoa.tiff :样品PCoA 分析图 参考文献: Xiao-Tao Jiang, Xin Peng, et al. Illumina Sequencing of 16S rRNA Roof Revealed spatial variations of bacterial communities in a Mangrove Wetland. Microb Ecol (2013) 66:96-104.DOI10.1007/s00248-013-0238-8. PCoA主坐标分析 注:PC1 啦PC2 是两个主坐标成分,PC1 表示尽可能最大解释数据变化的主坐标成分,PC2 为解释余下的变化度中占比例最大的主坐标成分,PC3 等依次类推。 Hal Whitehead, Susan Dufault, in Advances in the Study of Behavior, 1999In nonmetric multidimensional scaling and main coordinates analysis (sometimes called metric scaling and related to the main components analysis), animals are also represented at points in space, but in these show the distance between points is inversely related to the strength of the relationship. In the main coordinate analysis, the distances between the points are inversely proportional to the strength of the relationship measure (perhaps transformed in some way). In non-commercial multidimensional scaling, the correspondence is only ordinal: more closely related animals should be closer. Multidimensional scaling is most useful when the mapping matrix can be scaled in one or two dimensions with little stress (indicating the degree of discrepancy between the sequence of distances between the points on the display and the values in the union matrix; see Kruskal and Wish, 1978). In such cases, the arrangement of animals in the plot mimics the order of their associations, so that carefully depicted animals have high attachment indices. However, reducing associations to rows can remove some important elements of an association matrix. The parcels produced by main coordinate analysis are usually similar to those from multidimensional scaling. The main coordinates have the advantage of not reducing the data to rank, but the disadvantage that multiple dimensions are generally required to represent a given mapping matrix satisfactorily (see Digby and Kempton, 1987). Multidimensional scaling areas are inferred iteratively, while the main coordinates analysis results from an eigen vector analysis of the mapping matrix. This gives main coordinates analyzing the benefits of being computationally faster, possible with multiple individuals and producing a unique display. José A. Navas-Molina, ... Rob Knight, in Methods in Enzymology, 2013Ayear we want to compare samples in PCoA spaces that were treated in different ways, such as different ribosomal RNA subunits, primer sets, or algorithmic choices for treatment, we can use procrustes analysis (Gower, 1966; Muegge et al., 2011; Vinten et al., 2011). Analysis of procrustes is a statistical form algorithm that allows To compare different distributions by converting and using a rotational matrix, i.e. a rotation matrix. As an example, we present the results of the comparison of the different OTU picking algorithms, see section 4.2.2, where we can see that even as the number of OTU clusters changes the distribution described, the same is with a trust of MC p value: 0.00 and M2: 0.097 for closed reference versus de novo and MC p-value: 0.00 and M2: 0.035 for closed reference versus open reference. Both cases used the first three axes (i.e. the axes shown in the plot) and 100 iterations, Fig. 19.18. To generate these plots, we ran these commands:Figure 19.18. Procruste's analysis of different picking algorithms, where we can see that different OTU cluster methods provide similar PCoA distributions. PCoA plots are colored by BODY_HABITAT. (A) Comparison of samples with clusters picked using the novo picking protocol with the closed reference. (B) Comparison of samples with clusters picked using open-reference picking protocol with closed-reference.transform_coordinate_matrices.py -in $PWD/diversity_analysis/closed_ref/bdiv_even7205/unweighted_unifrac_pc.txt,$PWD/diversity_analysis/denovo/bdiv_even7205/unweighted_unifrac_pc.txt -r 100 -o $PWD/procrustes/closed_ref-denovocompare_3d_plots.py -i $PWD/procrustes/closed_ref-denovo/pc1_transformed.txt.$PWD/procrustes/closed_ref-denovo/pc2_transformed.txt -o $PWD/procrustes/closed_ref-denovo/plot -m $PWD/IQ_Bio_16sV4_L001_map.txttransform_coordinate_matrices.py -i $PWD/diversity_analysis/closed_ref/bdiv_even7205/unweighted_unifrac_pc.txt,$PWD/diversity_analysis/open_ref/bdiv_even7205/unweighted_unifrac_pc.txt -r 100 -o $PWD/procrustes/closed_ref-open_refcompare_3d_plots.py -i $PWD/procruste s/closed_ref-open_ref/pc1_transformed.txt,$PWD/procrustes/closed_ref-open_ref/pc2_transformed.txt -o $PWD/procrustes/closed_ref-open_ref/plot -m $PWD/IQ_Bio_16sV4_L001_map.txtHuilu Caoo , ... Ji-Dong Gu, in Methods in Enzymology, 2011Two multivariate statistical tools, Jackknife environment clustering and main coordinates analysis (PCoA), are provided within the UniFrac for microbial community classification using gene sequence data. Both statistics can work on qualitative data (unweighted, sequences only) and quantitative data (weighted with sequence density) (Lozupone et al., 2007). Qualitative analyses have been shown to detect the long-term development or adaptation of the microbial community with a strong genetic background, while quantitative analyses can detect microbial societal adaptation to the transient properties of the environment (Lozupone et al., 2007). Thus, the unweighted and weighted UniFrac analyses can produce different results on the microbial community depending on the specific state of the environment and/or environmental history. Sarah P. Preheim, ... Eric Alm, in Methods in Enzymology, 2013Multivariate statistical techniques have been developed specifically to handle high-dimensional data, including Principal Coordinate Analysis. These techniques work by first reducing the dimensionality of the data by identifying the main axes of differentiation between communities, representing different combinations of bacterial species. These tools are particularly effective when the variable of interest (e.g. pH, disease condition) is associated with major changes in the structure of society, but are less effective in detecting subtle variations in society's structure. In addition, they have problems in identifying the specific bacterial species that drive these associations. Statistical learning techniques have recently been used to detect correlations between bacterial species and environmental metadata. Statistical learning has many advantages over multivariate statistical approaches, including the ability to detect both larger and minor variations in microbial community structure, the ability to detect non-linear associations involving combinations of bacterial species, and the ability to pinpoint the specific bacterial species that underlie these associations. A statistical learning technique suitable for microbial community analysis is the Random Forest classification, which is implemented in the SLIME software package. For example, Slime was recently used to discriminate against patients with inflammatory bowel disease from sound control and to identify the specific bacterial taxis underlying this association (Papa et al., 2012). SLIME can detect mappings using categorical or quantitative metadata and runs permutation tests to assess statistical significance. It is best used for relatively large data sets where the total number of samples exceeds 50. The input for this analysis is a table of OTU enumerations and corresponding metadata. It will go on to test the meaning of each column of metadata, reporting the P-value of each association, the main bacterial taxi, and advanced metrics such as the area below the curve and other classification results.J.T. Hefner, in Biological Distance Analysis, 2016Legendre and Legendre (1998) first suggested a canonical discriminant analysis performed on the transformed values of the main coordinates. In short, a cap uses a main coordinate analysis using one of several distance measures (Anderson and Willis, 2003), which is essentially a conversion of categorical variables into continuous, normally distributed variables. In this way, the CAP method is highly effective in processing morphoscopic data and enables the classification and visualisation of the groups in a way that approximates craniometric The following Following description of the CAP method is a summary of Anderson and Willis's (2003) treatment of the same. Consider an N × p matrix (Y) that includes morphostic tensile data for a population where N = number of individuals and p = the number of morphoscopic traits. Also consider an N × q matrix (X) that represents the populations. CAP performs a non-parametric canonical analysis for the effect of X, if any, on Y using a rangefinder. Because X contains predictor variables (explanatory) variables, the result is a general discriminant analysis. The first step in an analysis of the CAP is to calculate a distance matrix. So let D = (dij) be a (N × N) distance array calculated from tensile values from Y, using rangefinder, J. Then we allow A=(aij)=(−1/2dij2). Now we calculate Gower's (1966) centered matrix (G) by centering the elements of A:(16.1)G =(I−(1N)1·1′)A(I−(1N)1·1′), where 1 is a column of 1's length of N and I am an identity matrix in form (N × N). Decomposition matrix G in component eigenvalues and corresponding orthoreical eigenvectors, we get Q. A subset of these eigenvectors (m) is then selected to form a matrix (Qm) for the canonical analysis. To avoid overfitting the model, m kept small compared to N. To avoid this problem and questions of subjectivity in the selection of m, the derived diagnostic information for appropriate dimensionality recommended by Anderson (2002) was used for the successive canonical analyses. Next, we calculate the influence, or hat, matrix, which is idempotent; Thus by definition H2 = HH = H. So for our purpose hat matrix is [which Anderson (2002) attributes to Johnson and Wichern (1992)]. We can now calculate the canonical test statistics as tracks:Two problems need to be solved: (1) selecting the number of relevant axes m and (2) selecting the most appropriate distance/dissensity target for the analysis. Anderson and Willis (2003) consider the selection of the number of primary coordinates (PCO) axes to remain one of the most fundamental issues in an analysis of the CAP. If the number of axes is too small, relevant information in the data may be lost. if the number is too large, the data may overfit the model overfitting. To remedy this, they propose to increase the value of m until the remaining error or misclassification rate (for generalized discriminatory features) is minimized. In addition, Anderson and Willis (2003) suggest that the selected m axes should explain at least 60% of the variation in the dataset, but should never reach 100%, as would be the case for an overfit for the model. The second question is only briefly discussed by Anderson and Willis (2003), which propose the use of a methodological approach similar to that used for the selection of PCOs. They analyse their data sets using multiple distances and then plot the proportion of against the number of axes (1 error rating). The measure that gives the highest proportion correctly using the lowest value m provides the best customization model. In any event, the authors consider that several different measures should be used to explore different aspects of the dataset. One example can clarify this problem: The Chi-square measurement highlights composite rather than abundance information in the dataset, while Bray-Curtis differences highlight relative abundance information. In a sense, both types of information can be important, especially when dealing with morphophoscopic traits across populations. What variables include Europeans, or what is the proportional distribution of an inferior nasal morphology score of 2? When creating a model to answer questions like these, each diversity measure can provide information that is not captured by someone else. Underlying assumptions about the data should always be taken into account when selecting a distance/unlikeality target, but one is not automatically chosen over another because of possible violations. Euclidean rangefinders are usually used with continuous variables and are defined as the shortest distance between two points. The calculation of the Euclidean distance corresponds to the Pythagorean distance (a straight line connecting A to B). Manhattan Block distance is suitable for use with discrete data (although it is not limited to this particular type of data) and is calculated as whole units (e.g. three units can be equal to three people, etc.). The Canberra distance is usually reserved only for non-supportive variables. This distance measurement sums up the relationship between the corresponding values and takes into account not only the distance between the two points, but also their relationship to an origin (0,0). The mesrual Chi-square measurement calculates distances based on composite data obtained from synthesized profiles (the calculated percentages of the response variables and each category) and is extremely useful for visualizing patterns, relative dispersions, and total variance in a dataset (Hefner and Ousley, 2006). The Chi square range is probably the most appropriate measure for use with categorical, non-metric cranial features. Most importantly, the rangefinder chosen for analysis should be robust and able to highlight patterns in the dataset. There are many ways to calculate the distance between two points in multidimensional space, and of the many distance measures to choose from, one or more measures can provide classification rates that meet the criterion necessary for a particular research paradigm. Secondly, the CAP methodology for variable correlations, as opposed to other statistics (e.g. naïve Bavarian), assumes that all response variables are independent. This is particularly important in view of the research from Hefner (2003), which showed that correlations between morphoscopic variables. By taking into account the correlated structure of the response matrix, significant patterns in the dataset can be illuminated, in particular with regard to a priori labels. The complex relationship between morphoscopic traits, environment, genetics and head shape should and can be investigated more thoroughly if the relationship between the data is properly understood. The final advantage of the CAP method is that unknown persons can be assigned to a group through a generalized discriminant analysis (Anderson and Willis, 2003). This is especially useful for forensic anthropologists who seek to identify the lineage of an unknown skeleton using morphophoscopy traits. Although traditional discriminant function analysis can produce admirable results using categorical variables (Hefner and Ousley, 2014), the underlying assumptions of multivariate normality and equal variance covariance may not be met. Krzanowski (2002) reviews the use of categorical data in multidimensional discriminatory analysis and acknowledges that underlying assumptions will be violated, but he also points out that the method may be appropriate for descriptive purposes and as an exploratory method. However, the CAP method circumvents this issue by transforming data values from categorical variables to main coordinat scores. In a sense, the data change from categorical responses to continuous variables (Anderson and Willis, 2003). Norman NH Mason, David Mouillot, in the Encyclopedia of Biodiversity (Second Edition), 2013The niche complementarity hypothesis states that differences between species in their functional traits will improve ecosystem function by allowing greater complementarity in resource use between species, which in turn will increase the total amount of resources available to society. In plant communities, niche complementarity predicts that primary productivity should be increased by increasing FD. Increased primary productivity by increasing the energy and nutrients available to consumers is likely to increase consumer diversity, which can increase the rate of function (especially littering) in consumer communities. Mouillot et al. (2011) tested these ideas by comparing the relative influence of species richness, evenness in species density, functional identity (societal traits mean weighted by relative density) and FD measures on a measure of ecosystem multifunctionality on an experimental grassland near Bayreuth, Germany. This site was part of the pan-European project BIODiversity and Ecological Processes in Terrestrial Herbaceous Ecosystems (BIODEPTH), where societies that differed in plant diversity were established from seeds. The multifunctionality of the ecosystem was calculated as the average performance of local communities over four processes (above-ground biomass production, litter and nitrogen in live aboveterrestrial vegetation) after standardising each social process (average of 0 and standard deviation of 1) to give them the same weight. The FD measures used were CHV for functional richness, FEve for functional evenness and FDiv for functional divergence, all of which are described in Villeger et al. (2008). Six functional traits were measured in total, containing a mixture of continuous and categorical variables. To allow the categorical variables to be included in FD calculations, PCoA was performed. This resulted in three independent functional tensile axes. The distribution of species in functional draughts were thus described by species positions along each of these axes. For each community, FD was calculated using PCoA coordinates instead of raw functional draught measurements with species above ground biomass as a measure of abundance. The analyses were limited to communities with four or more species, as CHV and FDiv can only be calculated when species richness is greater than the number of dimensions in functional draughts. The Community weighted mean values (CWM) were calculated as the density-weighted average coordinate value for each PCoA axis. Species richness was determined by the number of species sown in the indigenous community, while the evenness of species above ground biomass was estimated using the Pielou Index. These eight social structure indices were used as predictor variables in a Structural Equation Model (SEM) to predict ecosystem multifunctionality. SEM showed that FDiv had a strong positive effect on multifunctionality, regardless of the other social structure variables (Figure 9). It also showed that the CWM values for PCoA axis 1 had a positive and, for PCoA axis 3, a negative relationship to multifunctionality. Overall, SEM revealed a synergistic effect of FDiv and functional identity (CWM values) on ecosystem function, indicating that highest functional rates were achieved by societies that had special CWM values and where the most abundant species were complementary in their functional traits. This example shows how SMEs in combination with field experimental data can reveal FD's impact on ecosystem function. Several regression analyses show that the FDiv had a highly significant positive effect on all four ecosystem processes studied, suggesting that increased functional divergence in plant communities can not only increase primary productivity, but also degradation and nutrient cycling in grassland ecosystems. Figure 9. The relationship between the multifunctionality of the ecosystem and functional differences (as estimated by the FDiv). Multifunctionality was estimated as the mean of standardised values taken across four measures for ecosystem function (production of above-ground biomass, bedding and cotton composition and assimilation of nitrogen alive). vegetation). This explains why some plotts have negative multifunctionality values. A complication in this example is that PCoA axes are not immediately functionally interpreted. Subsequent analyses show that the effect of FDiv was largely due to differentiation along PCoA axes 1 and 3. Axis 1 is largely a contrast between plants with vertical leaves and plants with horizontal or oblique leaves. Axis 3 is a contrast between plants with evergreen and non-utter green leaves. Thus, the results of Mouillot et al. (2011) suggest that the functioning of the ecosystem is enhanced by complementarity in the use of resources between species that differ in leaf ingenet and hiltology. Differences in leaf inntanning provide complementarity in the light-catching between species, which can increase the total amount of light absorbed in photosynthesis for society as a whole, while differences in phenotyology allow for temporal differentiation in resource use between species, increasing annual biomass and litter production. This reinforcing effect of FDiv on productivity may be responsible for its relationship to degradation rates, by increasing the flux of nutrients available to the decomposer community. Thus, although the use of prescribing to reduce the dimensionality of drag space introduces some complexity into the interpretation FD ecosystem function relationships ecologically, it is still possible to find an ecological basis for the observed patterns. Jean-Luc Bouchot, ... Gail Rosen, in Computational Systems Biology (Second Edition), 2014Ecologists uses prescribing methods when a visual relationship is desired based on similarity between a set of multivariate objects (Jongman et al. 1987; TerBraak and Prentice 2004). Typically, each object represents a test site and is a table that represents instances of organisms in the community. This composition generally varies between sites and can be structured according to environmental variables, which are termed increases. An ideal prescribing technique would display all test sites in the same order as they exist along the environmental gradient, with inter-sample distances proportional to their separation along the gradient (Legends and Legends 2012). Different prescribing methods are available for both direct (limited) and indirect (unlimited) analyses. The difference between the method classes depends on whether environmental gradient measurements are included (direct) or omitted (indirectly). Popular methods discussed below include PCA, PCoA, NMDS, CA, CCA, RDA, and DCA (Jongman et al. 1987; TerBraak and Prentice 2004; Legends and Legends 2012; Fasham 1977; Kuczynski et al. 2010; Gauch et al. 1977). Metagenome annotations are a class of high-dimensional data that can be explored and examined using dimension reduction. In such cases, we will receive a set of n samples (e.g. samples collected from different body). Each sample, which we call x, is described by a certain number of K of functions (e.g. frequencies or abundance of annotated microbial species), and all of them are gathered in what is called a data matrix: X = [x(1),x (2) ,...,x (n)]∈RK×n. The number of functions can be between 10 and 107, making visualization impossible (as displaying more than three dimensions is a challenge at best). Picking just three variables to show randomly is not a good choice, it throws out most of the data without trying to identify and preserve aspects of the data that may prove interesting. For concreteness, consider the case where the features are abundance of each of K = 150 types of known bacteria. If we decide to choose a subset of these dimensions of size m = 3 for visualization or further study that is K!m! (K-m)!=551,300 possible candidates to choose from. We will present two examples of data reduction techniques that attempt to select interesting subsets of the function space commonly used for visualization purposes: Principal Component Analysis (PCA) and Principal Coordinate Analysis (PCoA). Both rely on projections on orthogonal axes that represent the most variance of the data as possible. PCA is a method that projects the data on to axes that contain most of the variance. This is done by calculating the eigenvalue the degradation of the covariance matrix and keeping the projections of the data on the few main eigenvectors. This procedure is called spectral degradation, and the eigenvectors, which correspond to the largest eigenvalues, are called the main components. They point in directions with the highest variation of the data. This method reduces the data to a sum of typical mixtures of the different bacteria in a sample, but it will not give the most relevant bacteria given an external parameter. The entire PCA degradation can be seen as a combination of an average microbiome mixture over all samples plus some small variations according to some calculated mixtures. The large dimensionality of the function vectors makes spectral degradation of the kovarian matrix unwieldy in most cases. Therefore, PCA is used as a dimensionality reduction technique for the analysis of (n×n) equality or distance meshdom. PCoA is another dimension reduction method useful for visualization (Legends and Legends 2012). First introduced (Gower 1966) as a method that would preserve the distances between objects, even when representing them using fewer components, it performs PCA on a modified version of the distance matrix. A major advantage of using PCoA over PCA is that it allows the user to tune the similarity between the distance metric used for comparison; this fact provides a more flexible tool in terms of the dynamics or range of the different variables. A is built by each element of the data matrix to each other using the selected comparison metric d:∀i,j∈{1,...,n},Dij=dx(i),x(j)This array is then squared and centered as follows: where Ai¨;Aj¨andA¨specifies the funds for each row, columns, and the entire array. This matrix transformation does not affect the distance relationships between the samples and thus preserves the structure of the data. But on the other hand, the centered data on the centroid of the samples. Finally, a spectral degradation is applied to the B matrix and its eigenvectors are scaled by the eigenvalues the square root. Normally, only a few eigenvectors cover most variabilitys of the data. Note that when you use Euclidean spacing as a metric for the likeness matrix, PCoA gives a similar result to the PCA approach based on the spectral degradation of the covariance matrix. Some care should be taken with this method. It works well when used with a metric distance measures (i.e. a binary positive particular shape meets triangle inequality). As some beloved inconsistencies do not meet the inequality of the triangle, they should not be used within this framework if one wishes to transform the data into another room. This should not be a major problem if the sole purpose of the transformation is visualization of the data. Figure 14.4 shows some PCoA plots for the human microbiome study comparing visualizations based on Hellinger distance (left side) and Euclidean distance (right). Note that PCA would give similar results to PCoA with Euclidean distance. Figure 14.4. PCoA plots of the 15-month study of the human microbiome(G. Caporaso et al., 2011) left Hellinger and right Euclidean.Pierre Legendre, in the Encyclopedia of Biodiversity (Second Edition), 2013Dufrêne and Legendre (1997) analyzed a large data set of Carabid beetle distributions in open habitats in Belgium (189 species collected in the pitfalls of 69 sites). The classification of the sites was obtained by distance-based K-means-breakdown calculated as follows: first, a distance matrix (percentage difference, also called Bray-Curtis distance) was calculated from the logtranstransformed species density data; this distance matrix was subjected to the main coordinate analysis (PCoA), also known as multidimensional metric scaling all prescribing axes (i.e. the main coordinates) produced by PCoA were used as input data for K-means division. Although clusters produced by K-means were not forced to be hierarchically embedded, they showed a strong hierarchical structure for K= two to ten groups. This allowed the authors to represent the relationship between partitions as a dendrogram with K= 10 groups, which corresponded to the main types of habitats recognized a priori by examining the sites. Indicator values were calculated for each species and level of breakdown. Some species were to be stenotope (narrow niches), while others were eurytopic (species with wide niches, present in a variety of habitats). Other species characterized intermediate levels in the hierarchy. The best indicator categories (IndVal>0.25) were grouped together in a two-way indicator table. this table representation showed hierarchical relationships between the species. Indval results were compared to Twinspan. The

partitions of sites used in the two methods were not the same; The Twinspan typology was achieved by dividing CA prescribing axes. Twinspan identified, as indicators, pseudo species regarding very low cut-off levels. These species were not very useful for prediction because they were simply known to be present in all places in a group. Several species identified by Twinspan as indicators also received a high indicator value from the IndVal procedure for the same or a closely related habitat class. IndVal identified several other indicator species with fairly high indicator values that also contributed to the specificity of the groups of sites but that Twinspan had missed. So the IndVal method seemed to be more sensitive than Twinspan to the fidelity and specificity of the species. Joseph T. Hefner, Kandus C. Linde, in Atlas of Human Cranial Macromorphoscopic Traits, 2018One of the more robust analytical methods for macromorphoscopic draught data is a canonical analysis of the main coordinates (CAP) method. The CAP method deals very effectively with macromorphoscopic data, enabling the classification and visualisation of the groups in a way that approximates craniometric analysis. Legends and Legends (1998) proposed a robust statistical approach to numerical ecology data that incorporated a canonical discriminant analysis performed on the transformed values of the main coordinates. In short, the CAP calculates the common cooperation organisations (inter-individual measures) using one of several distance measures (Anderson and Willis, 2003). Using these transformed values, the categorical responses to macromorphoscopic traits are essentially converted into continuous, usually distributed variables that match a canonical analysis. The following mathematical description of the cap method is a summary of Anderson's and Willis' (2003) treatment of the same thing and is drawn predominantly from Hefner (2016). Consider an N × p matrix (Y) that includes macromorphostic tensile data for a population where N=number of individuals and p=number of macromorphoscopic traits. An N × q matrix (X) represents the populations. CAP performs a nonparametric canonical analysis that tests the effect of any Y on X using an a priori distance measurement. Since Y contains predictor variables (explanatory) variables, the result is a general discriminatory analysis for classification purposes. The first step in a cap is to create an interindividual distance matrix. distance matrix. D=(dij) be a (N × N) space matrix calculated from tensile values from Y, using rangefinder, J. Allow A=(aij)=(−1/2dij2), calculate Gower's (1966) centered matrix (G) of A:G=(I−(1N)1·1′1′s(I−(1N)1·1′), where 1 is a column of 1's length of N and I am an identity matrix. Decomposition matrix G in component eigenvalues and corresponding orthoreical eigenvectors, we get Q. A subset of these eigenvectors (m) is then selected to form a matrix (Qm) for the canonical analysis. To avoid overfitting the model, m kept small compared to N. To avoid overfitting and subjective selection of m, the derived diagnostic information for appropriate dimensionality recommended by Anderson (2002) should be used for the canonical analyses. Next, we calculate the influence, or hat, matrix, which is idempotent; Thus by definition H2=HH=H. So the hat matrix is [as Anderson (2002) attributes to Johnson and Wichern (1992)]. The canonical test statistics (tracking statistics) are calculated, as two questions to be considered during a CAP analysis include 1) the choice of the number of relevant axes m and 2) the choice of the most appropriate distance/nonconformance measurement. Anderson and Willis (2003) regard the selection of the number of PCO axes retained as a fundamental issue in an analysis of the CAP. If the number of axes is too small, relevant information in the data may be lost, and if the number is too large, the data may overfitting. They propose to increase the value of m until the remaining error or misclassification rate (for generalized discriminant function) is minimized. In addition, Anderson and Willis (2003) suggest that the selected m axes should explain at least 60% of the variation in the dataset, but should never reach 100% (i.e. overfitting). Anderson and Willis (2003) briefly discuss the second question and propose to take a methodological approach. They analyse their datasets using multiple rangefinders and then plot the proportion of correct alloces in relation to the number of axes (1 error classification) to identify which distance measurement provides the highest proportion of correct alloces using the lowest value m. Hefner (2007) performed a similar analysis using several distance measures on a sample of macromorphoscopic data. He identified chi-square rangefinders as the most likely appropriate measure for use with categorical, non-metric cranial features (Table 20.2). Table 20.2. Classification accuracy in a two-way cap analysis. Eight best rangefinders includedDistance measureVariablesmBlackWhite%CorrectChi-square53819387Chi-square (metric)1310888084Euclidean53838484Bray-Curtis53838484Manhattan53828483Orloci's53808482Canberra53738881Jaccard53638574 Cap method accounts for variable correlations as opposed to other statistics (e.g. Naive Bayesian). This is particularly important in light of previous research by Hefner Hefner (2003a) Hefner (2003b), which showed intertrait correlations between macromorphoscopic variables. By taking into account the correlated structure of the response matrix, significant patterns in the dataset can be illuminated, in particular with regard to a priori labels. R-code for an analysis of the CAP is found in Appendix A.Arpita Ghosh, ... Asif M. Khan, in the Encyclopedia of Bioinformatics and Computational Biology, 2019Studie of microbial diversity in natural environments can be performed by targeting specific genes. For diversity analysis of bacteria and archaea communities, the 16S rRNA gene is commonly used as it contains one or more variable regions (Woese, 1987). For fungi and eukaryotes, the internal transcribed distancer (ITS) and 18S rRNA gene are used respectively. Commonly used tools for 16S rRNA analysis are QIIME (Quantitative Insights Into Microbial Ecology) (Caporaso et al., 2010), Mothur (Schloss et al., 2009), SILVAngs (Quast et al., 2013), RDPipeline (Ribosomal Database Project Pipeline) (Cole et al., 2013), MEGAN (Huson et al., 2007) and MG-RAST (Meomitagencs - Rapid Annotation using Subsystems Technology) (Meyer et al., 2008) as listed in Table 1. Despite the availability of so many tools and databases for 16S rRNA analysis, QIIME is considered to be the gold standard (Nilakanta et al., 2014). Widely used rRNA databases include Greengenes (16S) (DeSantis et al., 2006), Ribosomal Database Project (16S) (Cole et al., 2006, 2008; Maidak et al., 1996), Silva (16S & 18S) (Pruesse et al., 2007; Quast et al., 2012) and UNITE (ITS) (Kõljalg et al., 2013). The raw files generated from the next generation sequencing platform are exposed to demultiplexing, adapter trimming and quality filtering (Plummer et al., 2015). PCR product bell detection and removal is performed using the UCHIME algorithm (Sinclair et al., 2015). OTU plucking groups similar sequences by clustering or an equality-based method. OTU picking in the most popular tool QIIME is performed using the UCLUST program. The UCLUST program uses the USEARCH algorithm to assign the sequences to clusters (Edgar, 2010). Each OTU represents a cluster of similarities greater than a threshold, typically 97%-98%, which are then assigned to a corresponding taxonomy group. There are various OTU picking strategies: (1) De novo, where the reader is grouped without reference to known sequences. (2) Closed reference, where readers are grouped based on the adaptation to a reference database or (3) Open reference method, which clusters read in relation to a reference database and also unaligned clusters, reads using a de novo method. All these methods are incorporated into QIIME (Oulas et al., 2015). The taxonomy diversity can be represented as a density histogram. Further analysis can be carried out with degradation methods F.ex. F.ex. Correlation analysis (Johnson and Wichern, 2014). The taxonometric diversity can be represented using Krona, Fig. 2 shows a layer of a Krona plot at Proteobacteria level. The taxonomy wood in Newick format is available from QIIME, and can be visualized using any wood display tool, for example FigTree (FigTree is available at: . Alpha diversity measures variation within a single population, which measures wealth, dominance, and evenness. Other diversity metrics include, for example, Phylogenetic Diversity (PD), Chao (1984) Shannon entropy (Gorelick, 2006), among others. Rarefaction analysis is used to assess the coverage of the microbial community in the sample. Rarefaction curves plot the sample size relative to the estimated number of genera (Jaenicke et al., 2011). Beta diversity is the diversity across many populations or samples, which are calculated using different matrices, such as unweighted and weighted UniFrac (Lozupone et al., 2006) and PCoA (Principal Coordinating Analysis). It includes absolute or relative overlap between samples to estimate the proportion of taxis shared between them. Alpha and beta diversity calculation is supported by QIIME. The taxonomy methods can provide information on specific taxonomy hierarchy, such as phylum, class, order, family, genus and species (Darling et al., 2014), while phylogenetic approaches help identify species and new genera at taxonometric levels (Darling et al., 2014). Various tools used for phylogenetic analysis of metagenomes are AmphoraNet (Kerepesi et al., 2014), TIPP (taxonomic identification and phylogenetic profiling) (Nguyen et al., 2014), and Phylosift (Darling et al., 2014)), among others. The PhyloSift database contains a set of elite gene families of bacteria and Archaea, and also includes four additional sets of gene families, i.e. the geni. The methodology reader is searched against the defined set of gene database to predict taxonomy. The query sequences are adjusted according to reference genes in the database. The sequences are assigned on a phylogenic tree using Pplacer, and the taxonomy task is performed (Darling et al., 2014). This approach is based on statistical phylogenetic models, i.e. the phylogenetic models. The AmphoraNet web server uses the AMPHORA2 workflow for phylogenetic diversity analysis of metagenomic data (Wu and Eisen, 2008), where query sequences are adapted to bacterial and archaeal protein coding marker genes. Adaptation to marker genes is performed using a hidden Markov model trained in a reference database of fully sequenced bacterial genes (Wu M and Eisen J 2008). Metagenomic readings are used as input in AmphoraNet to search, adjust, and mask data against the HMM-trained model followed by phylogen analysis using RaxML on the masked metagenomics sequence (Kembel et al., 2011). Phylogenetic analysis is one of the critical steps in the analysis of metagenomics data, with applications in improved classification of sequences using phylogenetic methods, functional prediction of genes, and improved identification of OTUs, among others (Darling et al., 2014). To predict the functional composition of microbial communities from the 16S profile, PICRUSt (Langille et al., 2013) can be used. It uses an expanded ancestral state reconstruction algorithm, which predicts gene families and then combines gene families to assess the compound metagenome. The annotation of the expected family counts is obtained from ortologist groups of gene families, KOGs, COGs, NOGs or Pfam families (Langille et al., 2013). 2013).