BI390: Bioinformatics Workshop

Lab13: Microbiome Data Analysis

Maoying,Wu (ricket.woo@gmail.com)

Dept. of Bioinformatics & Biostatistics
Shanghai Jiao Tong University

Learning objectives

- Understanding the basics of microbiome studies.
  - Microbiota, metagenomics, microbiome
  - Amplicon sequencing: 16S rRNA, 18S rRNA, ITS
- Grasping the basic pipeline for microbiome studies.
  - Demultiplexing - Barcode
  - Quality control
  - Read trimming, chimera detection and adapter removal
  - Clustering and OTUs (operating taxonomic units)/ASV (amplicon sequencing variants) table
  - Taxonomic assignment
  - Phylogenetic construction
- Knowledge of concepts and metrics in molecular microbial ecology (分子微生物生态学).
  - Alpha diversity
  - Beta diversity
- Understanding the mathematical background of ordination analyses and how to conduct them.
  - Unstrained ordination (非约束排序分析)
  - Constrained ordination (约束排序分析)

Next Topic ...

1. Introduction
2. Qiime2 Pipeline
3. Ordination
4. Unconstrained ordination analysis
   - Eigenanalysis-based approaches
     - PCA
     - CA
   - Distance-based approaches
     - PCoA
     - NMDS
5. Constrained ordination
   - RDA
   - CCA
   - Extension

What is a microbiome?

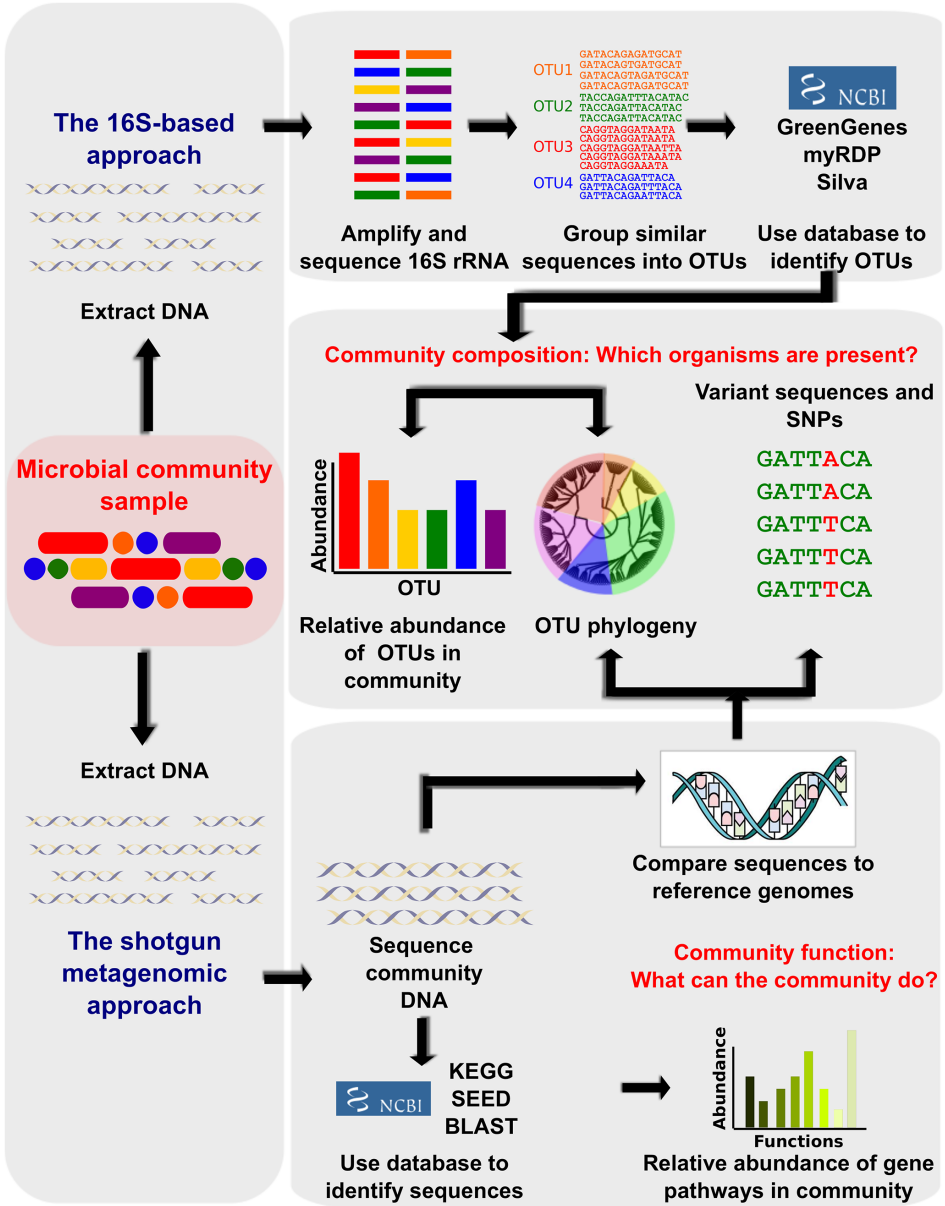
The totality of microbes in a defined environment, especially their genomes and interactions with each other, and also the surrounding environment.

- A population of a single species/strain is a culture, extremely rare outside of lab, some infections.
- A micorbiome is a a mixed population of different microbial species (microbial ecosystem)

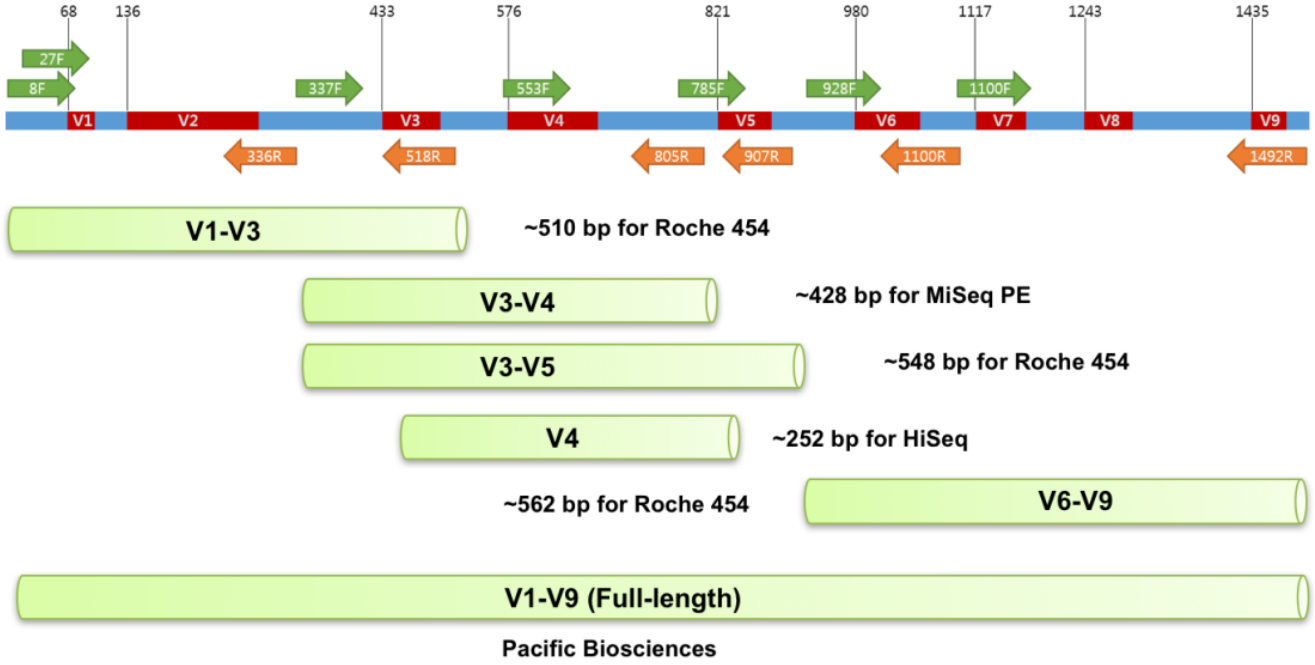
Methods for characterizing microbiomes

- Universal Gene census (amplicon)
- Shotgun Metagenome Sequencing
- Transcriptome (shotgun mRNA)
- Proteome (protein fragments)
- Metabolomics (excreted chemicals)

Two Main Approaches for Microbiome Study



16S rRNA Sequencing



- 16S rRNA gene is found in all bacterial species.
- Containing highly-conserved and highly-variable regions.
- Highly-variable can be thought of as molecular "fingerprint" for identification of bacterial genera and species:
  - V1-V2
  - (V3-)V4
- Illumina MiSeq, Illumina HiSeq
- Single-end (SE), paired-end (PE)

Reference 16S rRNA Database

Large public databases are available for comparison of 16S rRNA genes.

- Ribosomal Database Project (RDP): `http://rdp.msu.edu`
- ARB-SILVA: `https://www.arb-silva.de/`
- GreenGene: `http://qiime.org/home_static/dataFiles.html`
- NCBI 16SMicrobial: `ftp://ftp.ncbi.nlm.nih.gov/blast/db/16SMicrobial.tar.gz`

Next Topic ...

1. Introduction
2. Qiime2 Pipeline
3. Ordination
4. Unconstrained ordination analysis
   - Eigenanalysis-based approaches
     - PCA
     - CA
   - Distance-based approaches
     - PCoA
     - NMDS
5. Constrained ordination
   - RDA
   - CCA
   - Extension

Terminology

amplicon (扩增子) The product from targeted amplification of the genetic materials using non-random primers.

demultiplex (解复用) The preprocessing step used to assign each sequence to the sample via the barcode information.

barcode (序列条码) the unique sequences which are ligated to each individual sample's genetic material before the samples get mixed. Barcodes are unique to each sample.

OTU (操作分类单元) Operational taxonomic unit, an artificial, arbitrary construct useful for grouping sequences together into units to help summarize and analyze things.

ASV Amplicon sequence variant. Resulting sequences from newer processing methodologies that attempt to take into account sequencing error rates and are believed to represent true biological sequences.

Install and Run QIIME2

```
# https://docs.qiime2.org/2019.4/install/native
# wget https://data.qiime2.org/distro/core/qiime2-2019.4-
    py35-linux-conda.yml
# conda env create -n qiime2-2019.4 --file qiime2-2019.4-
    py35-linux-conda.yml

## Run Qiime2
source activate qiime2-2019.4

## Test your installation
qiime --help

## Deactivate
conda deactivate
```

Qiime2 Pipeline

1. Import data - from fastq to qza
2. Demultiplexing - Sample-separated file
3. Quality control - Determine parameters
4. Read trimming, chimera depletion - Clean sequences
5. OTU/ASV picking - OTU/ASV table
6. Taxonomic assignment - taxonomic table
7. Phylogenetic analysis - phylogenetic tree
8. Diversity analysis

1. Importing data

Once sequence file is ready, you need to import them to QIIME 2 "artifact":

- Artifact = data + metadata
- QIIME 2 artifact: *.qza
- Types of input data: single-end, paired-end
- Different formats of input data
- See "Importing data" tutorial in qiime docs.

Example

```
qiime tools import \
  --type 'SampleData[PairedEndSequencesWithQuality]' \
  --input-path pe-64-manifest \
  --output-path paired-end-demux.qza \
  --input-format PairedEndFastqManifestPhred64
```

Manifest file

sample manifest file (PE)

```
sample-id,absolute-filepath,direction
# Lines starting with '#' are ignored and can be used to cr
# "comments" or even "comment out" entries
sample-1,$PWD/some/filepath/sample1_R1.fastq.gz,forward
sample-2,$PWD/some/filepath/sample2_R1.fastq.gz,forward
...
sample-1,$PWD/some/filepath/sample1_R2.fastq.gz,reverse
sample-2,$PWD/some/filepath/sample2_R2.fastq.gz,reverse
...
```

2.Demultiplexing

- Must assign resulting sequences to samples for analyzing.
- The sequences may have already been demultiplexed.

example

```
qiime demux emp-single \
  --i-seqs emp-single-end-sequences.qza \
  --m-barcodes-file sample-metadata.tsv \
  --m-barcodes-column BarcodeSequence \
  --o-per-sample-sequences demux.qza
```

- Naming convention for options:
  - Inputs: `-i-<what-ever>`
  - Metadata: `-m-<what-ever>`
  - Parameter: `-p-<what-ever>`
  - Output: `-o-<what-ever>`

Metadata file

sample metadata

#SampleID	BarcodeSequence	LinkerPrimerSequence	BodySite
#q2:types	categorical	categorical	categoric
L1S8	AGCTGACTAGTC	GTGCCAGCMGCCGCGGTAA	gut
L1S57	ACACACTATGGC	GTGCCAGCMGCCGCGGTAA	gut
L1S76	ACTACGTGTGGT	GTGCCAGCMGCCGCGGTAA	gut
L1S105	AGTGCGATGCGT	GTGCCAGCMGCCGCGGTAA	gut
L2S155	ACGATGCGACCA	GTGCCAGCMGCCGCGGTAA	palm
L2S175	AGCTATCCACGA	GTGCCAGCMGCCGCGGTAA	palm
L2S204	ATGCAGCTCAGT	GTGCCAGCMGCCGCGGTAA	palm
L2S222	CACGTGACATGT	GTGCCAGCMGCCGCGGTAA	palm

Create a visualization for the demultiplexed object

```
qiime demux summarize \
  --i-data demux.qza \
  --o-visualization demux.qzv

# under local, or view *.qzv in https://view.qiime2.org
qiime tools view demux.qzv
```

3. Quality control

```
qiime quality-filter q-score \
  --i-demux demux.qza \
  --o-filtered-sequences demux-filtered.qza \
  --o-filter-stats demux-filter-stats.qza

qiime metadata tabulate \
  --m-input-file demux-filter-stats.qza \
  --o-visualization demux-filter-stats.qzv
```

4. Create feature table

```
qiime deblur denoise-16S \
  --i-demultiplexed-seqs demux-filtered.qza \
  --p-trim-length 120 \
  --o-representative-sequences rep-seqs.qza \
  --o-table table.qza \
  --p-sample-stats \
  --o-stats deblur-stats.qza

qiime deblur visualize-stats \
  --i-deblur-stats deblur-stats.qza \
  --o-visualization deblur-stats.qzv

qiime feature-table summarize \
  --i-table table.qza \
  --o-visualization table.qzv \
  --m-sample-metadata-file sample-metadata.tsv

qiime feature-table tabulate-seqs \
  --i-data reps-seqs.qza \
  --o-visualization rep-seqs.qzv
```

5. Phylogenetic analysis

```
qiime alignment mafft \
  --i-sequences rep-seqs.qza \
  --o-alignment aligned-rep-seqs.qza

qiime alignment mask \
  --i-alignment aligned-rep-seqs.qza \
  --o-masked-alignment masked-aligned-rep-seqs.qza

qiime phylogeny fasttree \
  --i-alignment masked-aligned-rep-seq.qza \
  --o-tree unrooted-tree.qza

qiime phylogeny midpoint-root \
  --i-tree unrooted-tree.qza \
  --o-rooted-tree rooted-tree.qza
```

6. Compute core diversity metrics

```
qiime diversity core-metrics-phylogenetic \
  --i-phylogeny rooted-tree.qza \
  --i-table table.qza \
  --p-sampling-depth 100000 \
  --m-metadata-file sample-metadata.tsv \
  --output-dir metrics
```

Rarefaction

- Rarefaction is randomly subsampling the same number of sequences from each sample. (samples without that number of sequences are discarded.)
- Concern:
  - Too low: ignore a lot of sample's information
  - Too high: ignore a lot of samples
  - A good choice for normalization
- The sample-depth is the related issue.
- But how to choose? `qiime tools view table.qzv`
  - Work with your partners
  - Why did you choose this value?
  - How many samples will be excluded from your analysis?
  - How many total sequences will be analyzed in the core-metrics command?

Exploratory Analysis

- "Diversity" - alpha, beta(, gamma)
- Beta diversity in practice: Ecological Distances
- Unsupervised learning: Clustering, etc
  - Ordination: e.g., PCA, UniFrac/PCoA, DPCoA
- Testing: Permutational Multivariate ANOVA (`adonis` in vegan)

Core metrics

- Alpha diversity
  - Shannon's diversity index - a quantitative mesure of community richness
  - Observed OTUs - a quantitative measure of community richness
  - Faith's Phylogenetic Diversity - a quantitative measure of community richness that incorporates phylogenetic relationships between features
  - Evenness - a measure of community evenness

Species richness

- The community contains up to k "species".
- The relative abundance of the species are $P = \{p_1, \ldots, p_k\}$
- Richness is computed as

$$R = I(p_1) + I(p_2) + \ldots + I(p_k)$$

where

$$I(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

- Higher R indicates greater diversity.
- Very dependent upon sampling depth and sensitive to presence of rare species.

Shannon index

- Borrowed from Shannon entropy of information theory.
- Computed as:

$$S_k = -\sum_{i=1}^{k} p_i \log_2 p_i$$

- We define $p_i \log_2 p_i = 0$ when $p_i = 0$.
- Higher S_k means higher diversity.

Evenness index

- Shannon index for a community with k species has a maximum value at $\log_2 k$
- We can make different communities more comparable if we normalize it by the maximum.
- Evenness index is computed as:

$$E_k = S_k / \log_2 k$$

- $E_k = 1$ means complete evenness.

Simpson index

- Simpson index is the probability of resampling the same species on two consecutive draws with replacement.
- Suppose on the first draw we picked speces i, this event has probability p_i, hence the probability of drawing that species twice is p_i^2.
- Simpson index is usually computed as

$$D = 1 - \sum_{i=1}^{k} p_i^2$$

- The index represents the probability that two individuals randomly selected from a sample that will belong to *different* species.
- $D = 0$ means no diversity (1 species is completely dominant)
- $D = 1$ means complete diversity.

Alpha diversity: Equivalent units

- Alpha diversity can be defined in terms of equivalent units:

  How many equally abundant taxa will it take to get the same diversity as we see in a given community?

- For richness there is no difference in statistic.
- For Shannon, remember that $\log_2 k$ is the maximum which is attached when all species have equal abundance. Hence the diversity in equivalent units is 2^{S_k}.
- For Simpson the equivalent units measure of diversity is $1/(1-D)$, sometimes called "Inverse Simpson Index".

Core metrics

- Beta diversity
  - Jaccard distance - a quantitative measure of community dissimilarity
  - Bray-Curtis distance - a quantitative measure of community dissimilarity
  - Unweighted UniFrac distance - a quantitative measure of community dissimilarity that incorporates phylogenetic relationships
  - Weighted UniFrac distance - a quantitative measure of community dissimilarity that incorporates phylogenetic relationships.

Beta diversity in practice

1. Compute UniFrac/Bray-Curtis distances between samples.
2. Conduct MDS (a.k.a PCoA)
3. Plot first two or three axes
4. Admire clusters
5. Write paper
6. Choose new microbiomes
7. Return to Step 1, Repeat

This is in fact dimensionality reduction methods, that come from "unsupervised learning" in "exploratory data analysis".

Alpha diversity group significance

```
qiime diversity alpha-group-significance \
  --i-alpha-diversity metrics/faith_pd_vector.qza \
  --m-metadata-file sample-metadata.tsv \
  --o-visualization metrics/faith-pd-group-significance.
     qzv
```

7. Taxonomic assignment

```
qiime feature-classifier classify-sklearn \
  --i-classifier gg-13-8-99-515-806-nb-classifier.qza \
  --i-reads rep-seqs.qza \
  --o-classification taxonomy.qza

qiime metadata tabulate \
  --m-inpute-file taxonomy.qza \
  --o-visualization taxonomy.qzv

qiime taxa barplot \
  --i-table table.qza \
  --i-taxonomy taxonomy.qza \
  --m-metadata-file sample-metadata.tsv \
  --o-visualization taxa-bar-plots.qzv
```

Next Topic ...

1. Introduction
2. Qiime2 Pipeline
3. Ordination
4. Unconstrained ordination analysis
   - Eigenanalysis-based approaches
     - PCA
     - CA
   - Distance-based approaches
     - PCoA
     - NMDS
5. Constrained ordination
   - RDA
   - CCA
   - Extension

Ordination methods

- Graphical tools for multivariate data:
  - Species abundance in sampling sites
  - OTUs of DNA sequence analysis in sampling units
  - Measurements of chemical and physical conditions in sampling stations
- Sometimes we can summarize multiple variables into one indicator (e.g., diversity), and then we can use univariate methods.
- Used to analyze the effects of multiple environmental factors on numerous species simultaneously.
- A method to arrange species along gradients:
  - indirect gradient analysis
  - direct gradient analysis
- Data Matrix:
  - Sample-species abundance matrix
  - Associated covariate data matrix (only for direct gradient analysis)

Ordination methods

Goal: represent sample and species relationships in a lower dimensional space.

Characteristics of community data

- *sparse*: lots of zeros, most of species are present in only a few samples.
- *low intrinsic dimensionality*: many factors may influence species composition but few are important.
- *noisy*: high variance between sample replicates.
- *redundancy*: many species have similar distribution.

Ordination methods

Pros

- focusing on a low dimensional space which will represent important and interpretable environmental gradients.
- noise reduction by focusing on a low dimensional space.
- statistical efficiency: one global analysis vs. many univariate analysis.

Cons

- Only exploratory analysis, no easy-to-use test
- each method has its own limitations
- good understanding of the mathematical framework behind each method is required in order
  - choose the appropriate method
  - to make accurate interpretations

Ordination methods

A set of popular multivariate methods applied to quantitative ecology.

- Unconstrained ordination of matrix **Y**
  - principal component analysis (PCA，主成分分析)
  - principal coordinate analysis (PCoA，主坐标分析)
  - correspondence analysis (CA，对应分析)
  - Detrended correspondence analysis (DCA，去趋势对应分析)
- Constrained ordination of **Y** under constraint of **X**
  - Redundancy analysis (RDA，冗余分析)
  - Canonical correspondence analysis (CCA，典范对应分析)
  - Detrended canonical correspondence analysis (DCCA，去趋势典范对应分析)
  - Canonical variate analysis (CVA，db-RDA，典型变量分析)

Ordination methods

Methods	based on	gradient link	data type
PCoA	dist	linear	
NMDS	dist	any	
PCA	eigen	linear	quantitative
CA	eigen	unimodal	contingency table or at least positive
DCA	eigen	unimodal	contingency table or at least positive

Linear versus unimodal response models

- **Linear models** are appropriate to community analysis when
  - species are abundant (few zeros)
  - small range of environmental variations
- **Unimodal model** can be obtained by
  - adding a quadratic term x^2 to the linear model but allow large negative values.
  - modeling the logarithm of species abundance by a quadratic form of the environmental variable.
- Gaussian response curve:

$$\log y = a - (x - u)^2 / 2t^2$$

where

  - u: optimum or mode
  - t: tolerance
  - $c = \exp(a)$: the maximum abundance value

Next Topic ...

1. Introduction
2. Qiime2 Pipeline
3. Ordination
4. Unconstrained ordination analysis
   - Eigenanalysis-based approaches
     - PCA
     - CA
   - Distance-based approaches
     - PCoA
     - NMDS
5. Constrained ordination
   - RDA
   - CCA
   - Extension

Eigenanalysis-based methods

- Eigenanalysis is a mathematical operation on a square, symmetric matrix.
- Eigenanalysis requires a numerical iterative approach (except if dim(matrix) ≤ 3
- Eigenanalysis provides a series of eigenvalues and eigenvectors
- The greatest eigenvalue is often called the "dominant" or "leading" eigenvalue
- The eigenvalue is a measure of the strength of an axis: the amount of variation along an axis and ideally the importance of an ecological gradient.
- The sum of the eigenvalues will equal the sum of the variance of all variables.
  - if performed on a correlation matrix, the sum of the eigenvalues will equal the number of variables/species.
  - if performed on a covariance matrix, the sum of the eigenvalues will equal the sum of the variance of all species.

Principal component analysis (PCA)

Pearson 1901

Objectives

- To describe a matrix of data consisting of objects by reducing its dimensions.
- To find uncorrelated linear combinations of the original variables with maximal variance
- To suggest new combined variables for further study

Description/characteristics

- Rotation of the data matrix: not change the positions of points relative to each other; just change the coordinate system
- Axes are created such that the perpendicular distance from each object to the ordination axes is minimized

PCA: Limitations

- PCA is intended mainlly for continuous data; it is inefficient for data not well summarized by variances and covariances.
- The procedure considers only linear combinations of the raw variables, so it will not discover nonlinear combinations.
- The procedure has a serious problem for vegetation data: the horseshoe effect caused by the curvilinearity of species distributions along gradients, since species response curves are typically unimodal (very very strongly curvilinear), horseshoe effects are common.

Correspondence analysis (CA)

Reciprocal Averaging (RA), Hirschfeld 1935

Objective

- To describe data consisting of counts by reducing the number of dimensions to graph and new combined variables in further analysis.
- To simultaneously ordinate species and samples by maximizing the correlation between species scores and sample scores.

Principle

- For each sample i, calculate the weighted average of all of the species scores:

$$\text{score}_{\text{sample}_i} = \frac{\sum_j w_{ij} \times \text{score}_{\text{species}_j}}{\sum_j w_{ij}}$$

  where the weights w_{ij} is the abundance of species j in sample i.
- For each species j calculate the

$$\text{score}_{\text{species}_j} = \frac{\sum_i w_{ij} \times \text{score}_{\text{sample}_i}}{\sum_i w_{ij}}$$

CA: Characteristics

- The end results is that Speces scores and sample scores will be maximally correlated with each other.
- The eigenvalue is a measure of how well the species scores correspond with the sample scores
- The eigenvalues of of an axis will equal the correlation coeffcient between species scores and sample scores.
- An eigenvalue of 1.0 implies that one sample (or group of samples) shares no species with other samples.
- The first axis usually turns out to be related to important environmental gradients.
- One can put new samples in a correspondence analysis without affecting the rest of the ordination.

CA: Limitations

- The procedure is inefficient for data that are not counts because they will not be described by χ^2-distances
- The procedure is not suitable for nonlinear data; it will not discover nonlinear relationship - Arch effect
- The arch is not serious as the horseshoe effect of PCA, because the ends of the gradient are not convoluted.
- The ends of the gradient are compressed.

Distance-based approaches

Distance-based methods rely on a square, symmetric distance or similarity matrix:

$$D = \begin{bmatrix} d_{11} & d_{12} & \dots & d_{1j} & \dots & d_{1n} \\ d_{21} & d_{22} & \dots & d_{2j} & \dots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{i1} & d_{i2} & \dots & d_{ij} & \dots & d_{in} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & d_{nj} & \dots & d_{nn} \end{bmatrix}$$

where d_{ij} is the distance between sample i and j. The analysis may change according to the *distance* function used.

- Principal coordinate analysis (PCoA)
- Nonmetric multidimensional scaling (NMDS)

Principal coordinate analysis (PCoA): Multidimensional scaling (MDS)

Objective

To describe the data by reducing the dimensions of a distance matrix among objects to a graph (2D or 3D).

Characteristics

- Maximizes the linear correlation between distance measures and (Euclidean) distance in the ordination.
- The underlying model assumes a fixed number of gradients. In contrast, PCA, RA, and DCA assume potentially many gradients, but declining importance.
- Generalization of PCA in which non-Euclidean distance may be used (Euclidean distance $\Rightarrow$ PCoA = PCA)
- Generalization of CA: χ^2-distance $\Rightarrow$ PCoA = CA

PCoA

Principle (MDS)

1. Start from a distance matrix $D = [D_{ij}]$;
2. Use choose a dimension n for the representation space.
3. Assign randomly point coordinates in the n-D space.
4. Repeat the following steps until reaching the minimal *stress*:
   (1) Compute $d = [d_{ij}]$ euclidean distance matrix in the n-d space;
   (2) Regress d_{ij} on D_{ij} with a linear model $\hat{d}_{ij} = \alpha + \beta D_{ij}$;
   (3) Compute the *stress* as $\sqrt{\sum_{i,j}(d_{ij} - \hat{d}_{ij})^2 / \sum_{i,j} d_{ij}^2}$;
   (4) Move point coordinates to decrease *stress*.
5. The **optimal solution** is the eigenvectors of D.

Limitations

- Results depend on the distance measure chosen.
- Cannot indicate combinations of variables, because only the distance matrix among objects is provided.
- Not trivial to put new objects in a PCoA.

Study questions: NMDS

- What are the main differences between NMDS and PCA?
- Which three matrices are computed during NMDS?
- Outline the major elements of the algorithm used to compute the NMDS.
- Discuss the limitations of NMDS.

Nonmetric Multidimensional Scaling (NMDS)

Kruskal 1964

Objectives

- To describe data by reducing the number of dimensions to a graph.
- To descover nonlinear relationships.

Characteristics

- NMDS is very computer-intensive; becoming feasible for large data set recently.
- NMDS maximizes rank-order correlation between distance measures and distance in ordination space.
- Objects are moved to minimize *stress* (a measure of mismatch between the two kinds of distance)
- To increase the likelihood of finding the correct solution, a DCA is often performed first.
- A recent variant of NMDS is Local-NMDS, which only minimizes *stress* locally.

NMDS

Principle

1. Start from $D = [D_{ij}]$ distance matrix (not necessarily symmetric);
2. Use choose a dimension n for the representation space;
3. Assign randomly to each object coordinates in the n-d space;
4. Repeat the following steps until reach the minimum *stress*:
   (1) compute $d = [d_{ij}]$ between objects in the n-d space (with Euclidean metric, for example);
   (2) regress d_{ij} on D_{ij}: $\hat{d}_{ij} = \alpha + \beta D_{ij}$;
   (3) compute the *stress* as $\sqrt{\sum_{i,j}(d_{ij} - \hat{d}_{ij})^2 / \sum_{i,j} d_{ij}^2}$;
   (4) move object coordinates to decrease *stress*.

NMDS

Limitations

- The procedure uses rank order information only.
- The configuration will change as the number of axes (chosen by user) changes.
- It assumes that dissimilarity is monotonically related to ecological distance.
- There is not guarantee that the correct (lowest stress) solution will be found, though it is widely assumed that this is not a big issue.

Recommendations

- try a lot of starting point to find a "good" minimum.
- try different dimensions, the "optimal" is selected according to the slope heuristic.

Distance-based methods: Summary

- Unlike methods derived from eigenanalysis, distance-based methods do not provide species and sample scores simultaneously.
- CA, CCA are also distance-based methods, where the distance is based on the chi-squared distribution.
- The philosophy of the eigenanalysis-based methods is different: these methods attempt to faithfully place species along gradients (either inferred or directly related to measured variables), and not to faithfully relate difference to distance.

Next Topic ...

1. Introduction
2. Qiime2 Pipeline
3. Ordination
4. Unconstrained ordination analysis
   - Eigenanalysis-based approaches
     - PCA
     - CA
   - Distance-based approaches
     - PCoA
     - NMDS
5. Constrained ordination
   - RDA
   - CCA
   - Extension

Direct gradient analysis

Constrained ordination

- Species are directly related to measured environmental factors $(y_1, \ldots, y_n) \sim (x_1, \ldots, x_m)$.
- Results in axes are then constrained to be a function of measured factors (constrained ordination)
- Sample scores are constrained to be linear combinations of explanatory variables.
- Principle
  - $\hat{y}_i = b_0 + b_1 x_{i1} + \cdots + b_m x_{im}$
  - maximize $R^2 = cor(y, \hat{y})$
- The two most commonly used constrained ordination techniques:
  - Redundancy Analysis (RDA) = constrained PCA (linear models)
  - Canonical correspondence analysis (CCA) = constrained CA (unimodal model)

Study questions

The basics of RDA

- How many constrained axes has an RDA and how are they related to the descriptors?
- How does scaling influence the interpretation of a triplot?

Commonly-used association measures

- Which association is measured with similarity measures?
- Outline the calculation of the Bray-Curtis and the Jaccard coefficient.

Redundancy Analysis (RDA)

- **Aim**: Display and explain variation in a set of response variables constrained by second set of predictor variables (e.g., links multivariate multiple regression and PCA)
- **Example**: Which variables do best explain the variation in fungal communities sampled along a gradient of fungicide toxicity?

RDA

- Input: centered **Y** and **X**
- Fit multivariate regression model:

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}(\mathbf{X}^{\mathbf{T}}\mathbf{Y})$$

- RDA uses variance-covariance matrix of $\hat{\mathbf{Y}} \Rightarrow \Sigma_{\mathbf{Y}^T\mathbf{Y}}$
- Generally this is unknown and the sample variance- covariance matrix (also called Dispersion matrix) will be estimated from the observations:

$$\mathbf{S}_{\hat{Y}^T\hat{Y}} = \frac{1}{n-1}\hat{Y}^T\hat{Y}$$

- Obtain the eigenvectors U through eigendecomposition of the matrix (PCA) S.
- The matrix YU is the ordination in the space of variable **Y**.
- The matrix $\hat{Y}U$ is the ordination in the space of variable **X**
- Compute the residual values: $Y_{res} = Y - \hat{Y}$
- Apply PCA to Y_{res}, obtain U_{res}
- The matrix $Y_{res}U_{res}$ is the ordination in the space of residuals.

RDA results

- Triplot with relationship between species, sites, and environmental variables.
- Eigenvalues and variance partitioning (constrained and unconstrained)
- Sites scores
- Species scores
- Biplot scores for variables

RDA axes and variable importance

- How many RDA axes are required?
- Hypothesis tests (permutation-based) recommended (Legendre et al., MEE 2011)
- How many environmental variables are needed and how important are they?
- Manual and automatic model-building with adjusted R^2 as goodness of fit criteria (as for multiple linear regression)
- Variance partitioning between different models to determined explained variance of individual variable

RDA: Assumptions and extensions

- Independence of observations (sites)
- Linear relationship between explanatory and response variable
- No multicollinearity between explanatory variables
- n (sites) $>>$ p (predictors) to realiably infer p importance.
- RDA can be employed for multivariate ANOVA (Borcard et al. 2011: 185 ff)

Interpret RDA Results

Rao 1964

- Species as well as environmental variables are represented by arrows.
- CCA focuses more on species composition, i.e., relative abundance. Thus if you have a gradient along which all species are positively correlated, RDA will detect such a gradient while CCA will not.
- It is possible to use "species" that are measured in different units. If so, the data should be centered and normalized.
- Useful when gradients are short, thus RDA is the method of choice in a short-term experiment study.
- Variance partitioning, and interpretation of eigenvalues, are more straightforward than for CCA.

The RDA is not more or less valuable than CCA. It is simply used for different purposes.

Canonical correspondence analysis (CCA)

ter Braak 1986

- CCA is the marriage between CA and multiple regression.
- Mathematical is the same as RDA except that $\hat{Y}$ is obtained with a weighted multiple regression.
- CCA maximizes the correlation between species scores and sample scores.
- The sample scores are constrained to be linear combinations of explanatory variables.
- Because of the constraint, eigenvalues in CCA will be lower than in CA.
- Some variants: Detrended CCA (DCCA), partial CCA (pCCA), ...

Further constrained ordination

- Canonical correspondence analysis (CCA)
  - Widely used.
  - Extension of unconstrained correspondence analysis
  - Simiar to RDA, but assumes unimodal distribution (χ^2-distance) of species along environmental gradient
  - R function: `vegan::cca()`
- Constrained additive ordination (CAO)
  - Comparatively new
  - derives response of each species to main environmental gradient from data - no linear or unimodal model assumed.
  - Mixture of generalized additive models (GAMs) and canonical Gaussian ordination.
  - Computationally demanding.
  - R function: `VGAM::cao()`

When to use what?

Legendre & Legendre 2012, 533