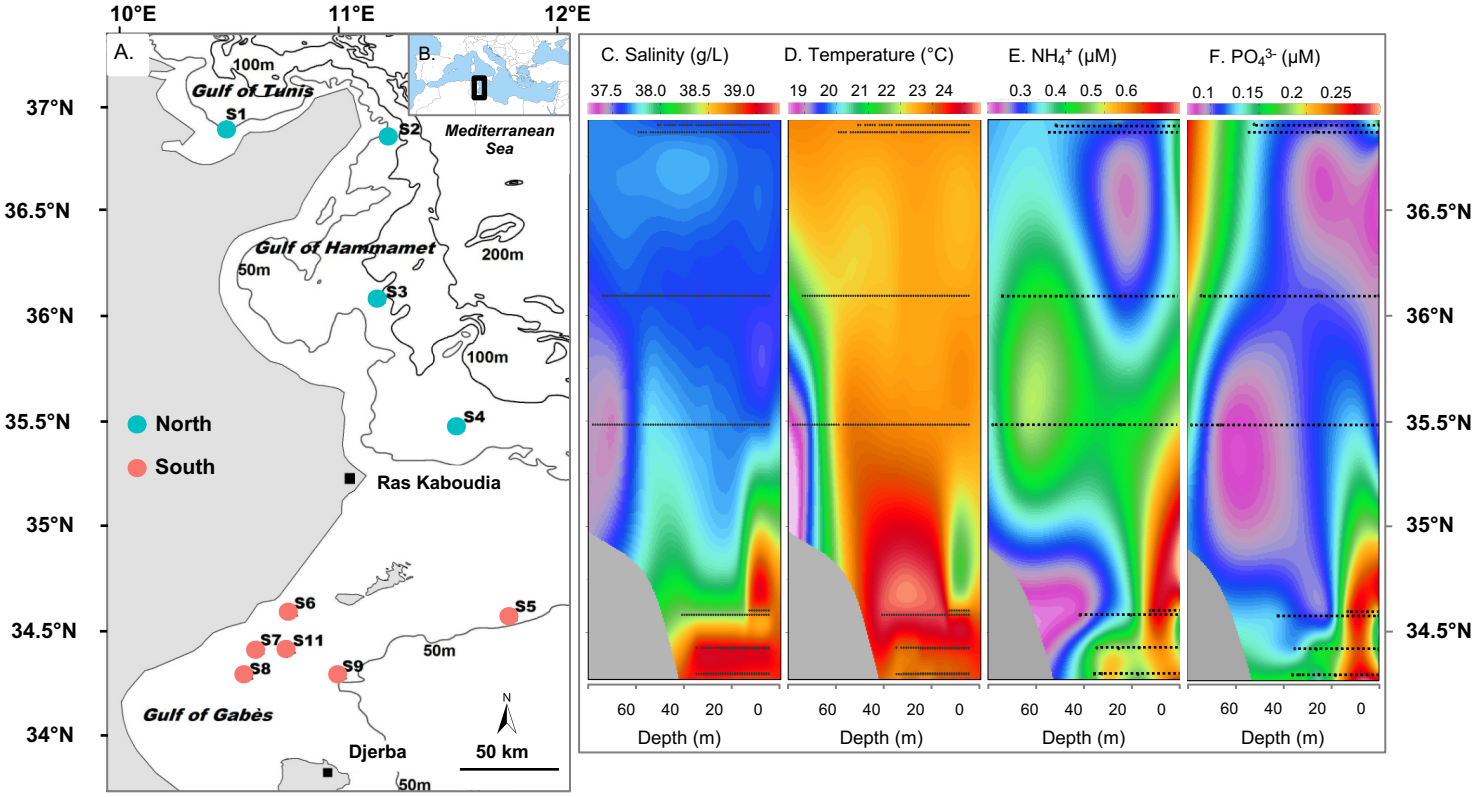
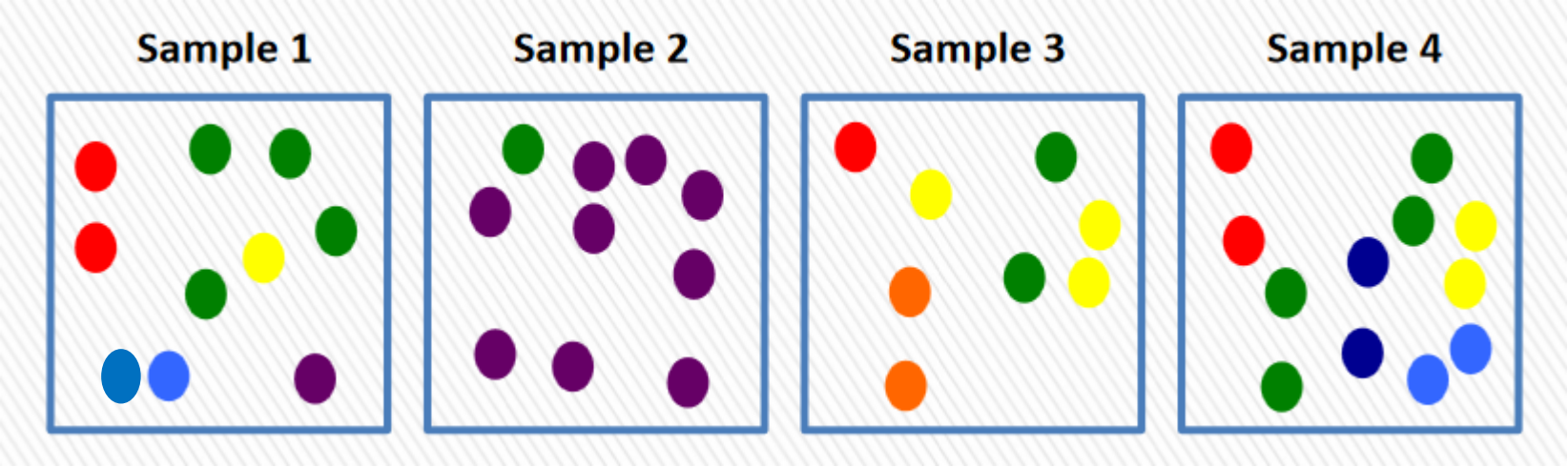
# MetaData


10°E
11°E
12°E
37°N
36.5°N
36°N
35.5°N
35°N
34.5°N
34°N
A.
B.
100m
Gulf of Tunis
S1
S2
Mediterranean Sea
Gulf of Hammamet
50m
200m
S3
100m
S4
North
South
Ras Kaboudia
S6
S5
S7
S11
S8
S9
50m
Gulf of Gabès
Djerba
50m
N
50 km
C. Salinity (g/L)
37.5 38.0 38.5 39.0
D. Temperature (°C)
19 20 21 22 23 24
E. NH4+ (µM)
0.3 0.4 0.5 0.6
F. PO43- (µM)
0.1 0.15 0.2 0.25
60 40 20 0
Depth (m)

# Alpha Diversity

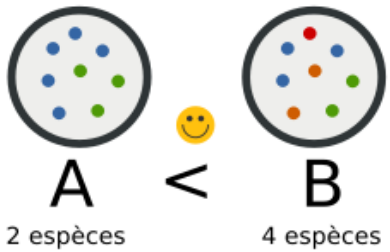
**intra-sample Diversity= Diversity observed in a uniq sample**




**Can you rank the samples according to their diversity (from the lowest to greatest )??**

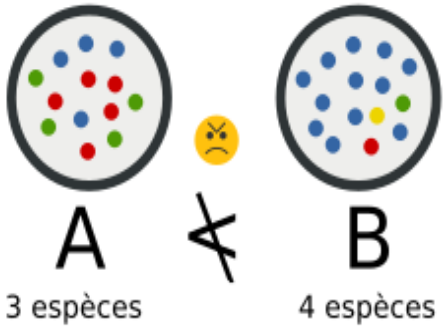
# Global Richness Vs. Specific Richness

**Global Richness (R) indice**: **The number of species observed** in a given sample




B has more species than A

→ It Is not to the best way to assess sample diversity
→ Because …




B has more species than A
But seems less diversified than A

Consequenly, Use indices of alpha diversity such as **Shannon, Simpson which reflect the species number and their relative abundance (Distribution):**

**It's the specific Richness**

# Shannon diversity Index

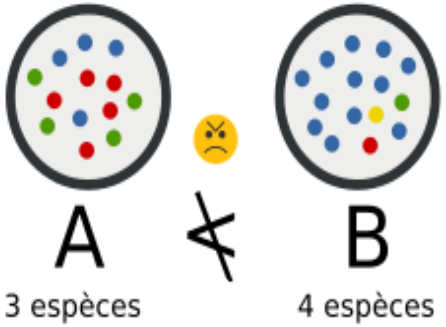
For each species : Sum of the frequency multiplicates by log Frequency

$$H(X) = H_2(X) = -\sum_{i=1}^{n} P_i \log_2 P_i.$$

- A consist of 3 species, of which 4 green, 5 red & 4 blue

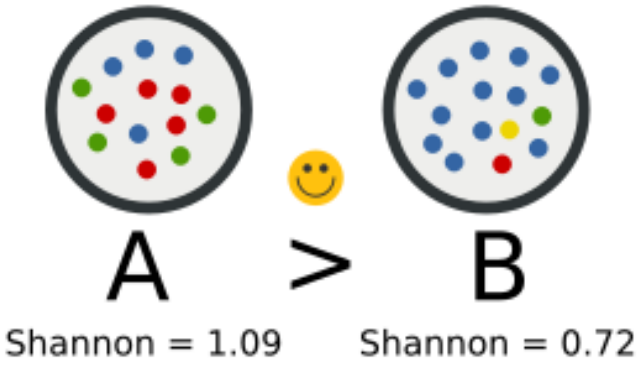
The Shannon indice will be :

$$-(\frac{4}{13}log(\frac{4}{13}) + \frac{5}{13}log(\frac{5}{13}) + \frac{4}{13}log(\frac{4}{13})) = 1.09$$




- B consist of 4 species, of which 1 reen, 1 red, 1 yellow & 11 blue

**Finally, after Shannon for B sample**




Shannon take into account species abundance

→ Influence by low abundant taxa

# Simpson diversity indice → affected by hi abundant texa

Idea : Probabily that 2 individuals selected randomly belong to the same species!

→ **Simpson indice (S)**

**A value of 0.8 means that 2 sequences randomly selected have 80% chance to belong to the same ASV/OTU!** Inversely proportional to the diversity

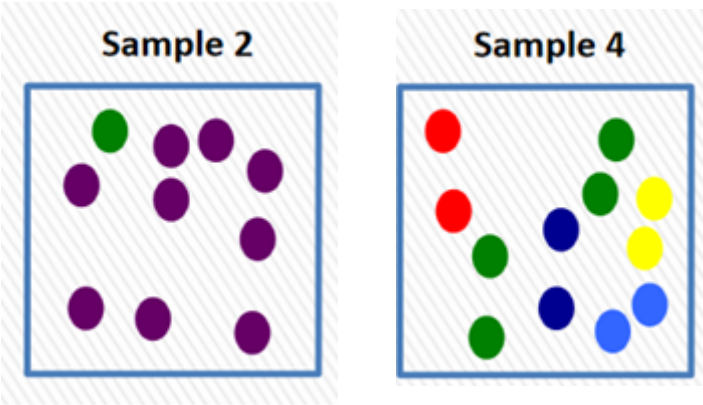
## Simpson Diversity indice = 1 - S

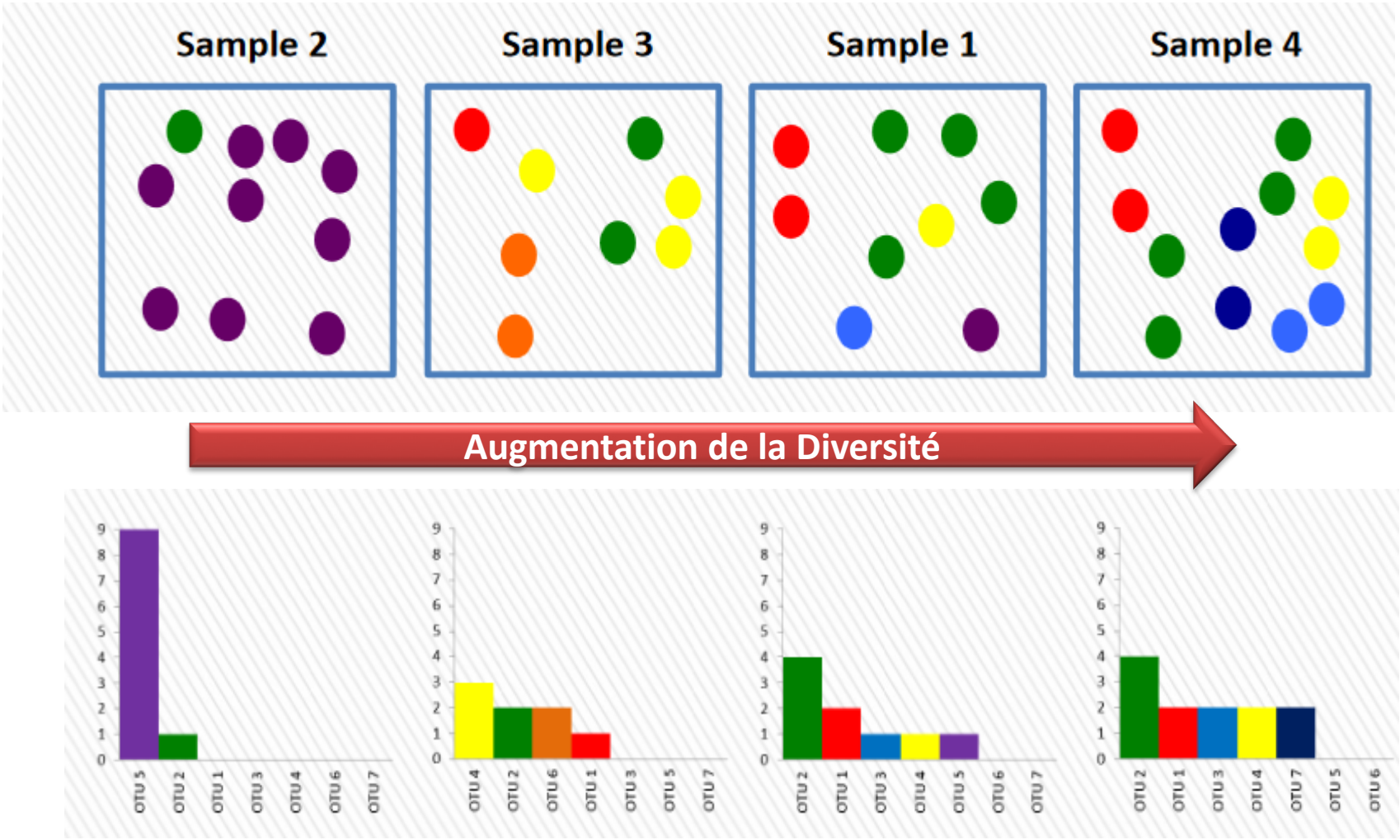
Influence by highly abundant Taxa

$$E = 1 - \sum_{s=1}^{S} p_s^2$$

S = species number

p= abundance contribution






**Diversity =Richness + evenness**

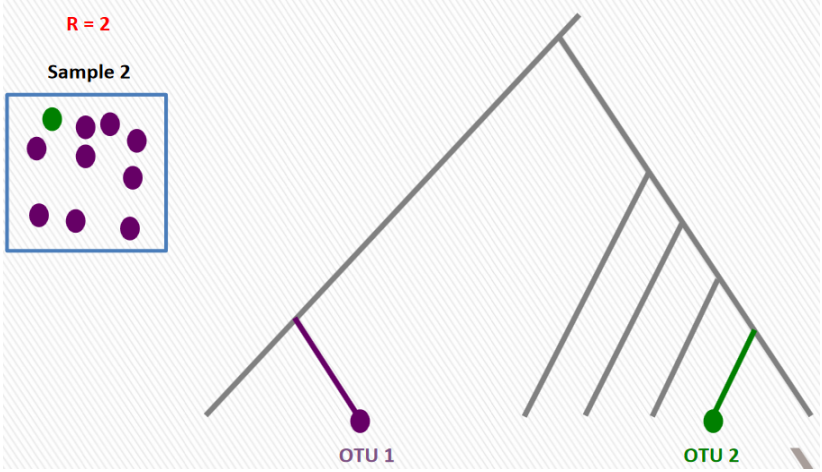
# Phylogenetic diversity

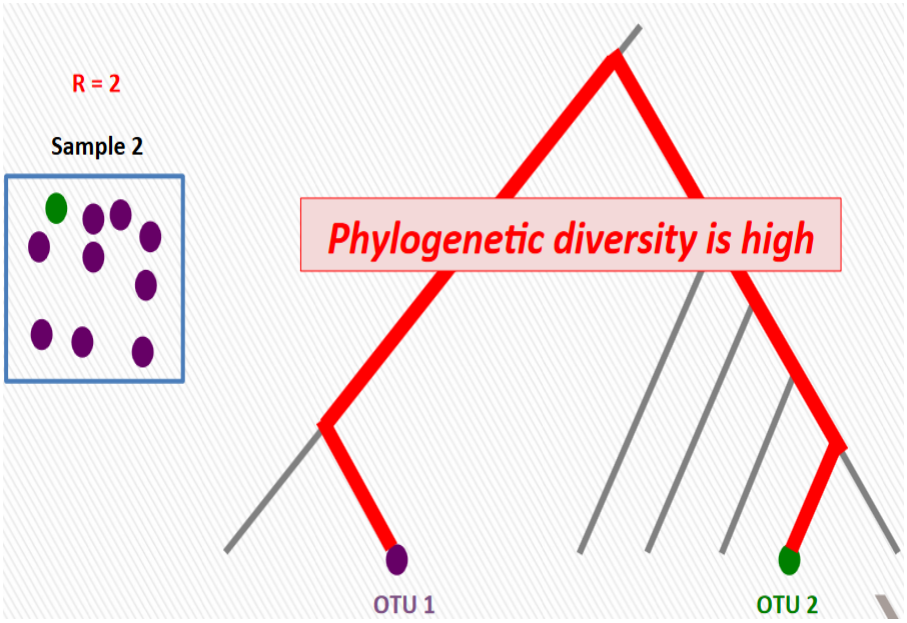
**Indice PD_whole_tree** = Faith's Phylogenetic Diversity,

- **Takes into account the phylogenetic information**

- The community diversity will be higher if the taxa belong to many different genera (rather than few)!

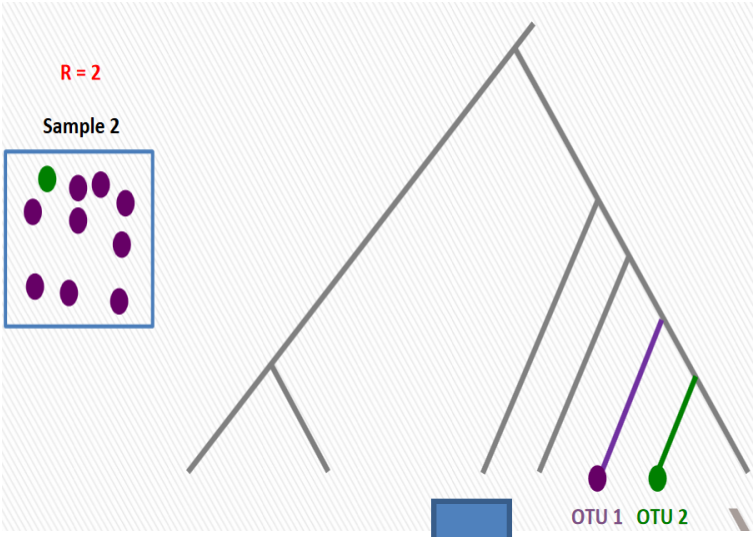
→ **Study the phylogenetic closeness of species/taxa**

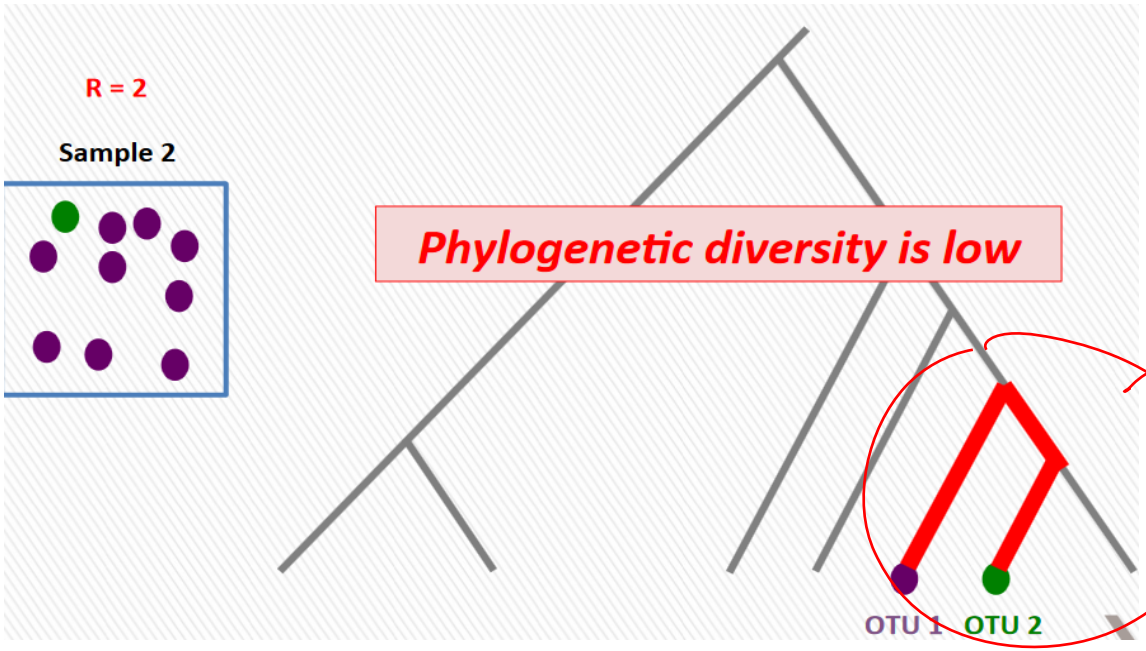



distance between OTU 1 and OTU 2 is high → it indicates a high diversity

→ calculate branch length




branch distance is small

# How it Works

→ Sum of branch lengths between taxa
→ The More ASVs/OTU in different phyla/genera you get, the higher diversity you have

# Diversity Estimators

predictors → what you observed vs. what you could have observed if you had sampled more

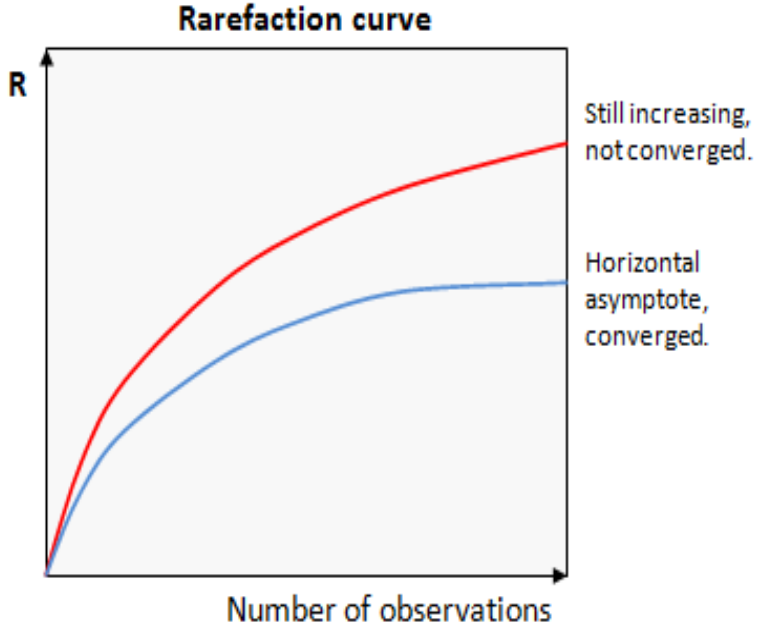
- **Chao1 & ACE are estimators of the species number that would be observed if sampled at infinity**

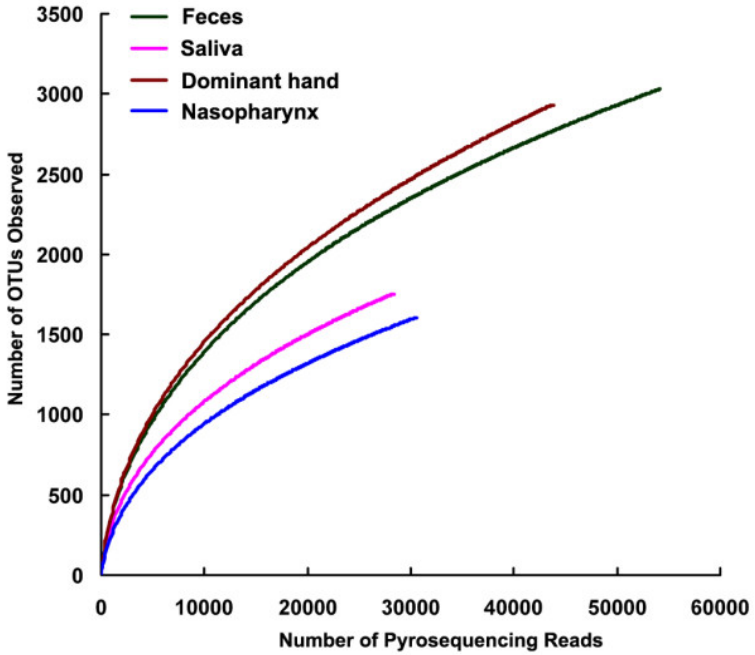
- **Good sampling gives you a total number of ASV/OTU observed not far from the Chao1 / ACE value (predicted for the sampled environment)**

**Chao1/ACE= $S_{obs}$ + Adjustment (linked to the rare)**

# Rarefaction Curves

**« Is the sequencing effort performed (sequencing depth) for a sample (s) sufficient for the number of species observed ? »**




**→Reach the asymptote ???**

**Asymptote means that sequencing more (depth), will not increase your number of OTU/ASVs observed**

# Beta Diversity

- **Inter-sample comparison of the community composition**
- **Measure of the similarities/dissimilarities between the samples** according to specific criteria of the MEASURE under consideration (Unifrac, Bray-curtis)

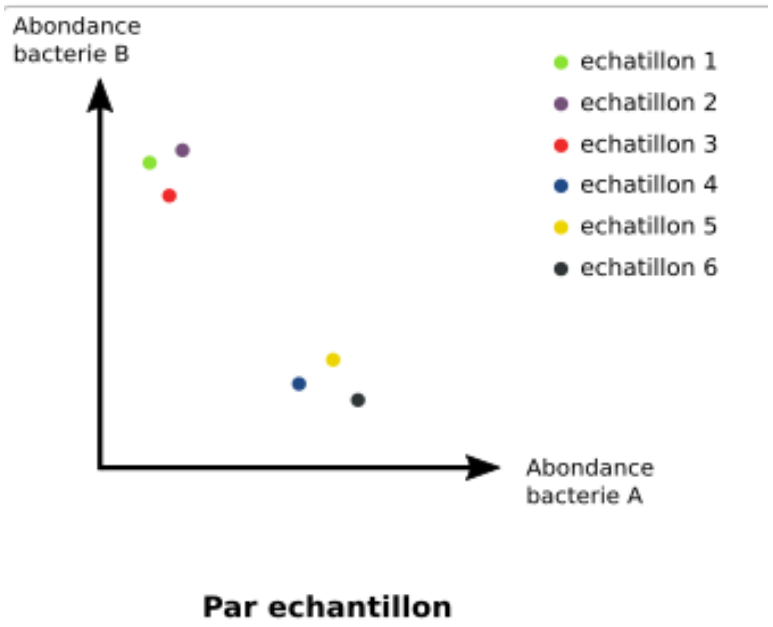
# Approach

- OTUs/ASV table composed with **2 taxa & 6 samples**
- Easy plot using two axes (1 by taxa)

- Coordinates (y1,y2) of each sample use the abundance values of Taxa/ASV1 and Taxa2/ASV2

**Mini OTU Table ☺**

| | Objets | Variable 1 OTU-A | Variable 2 OTU-B |
|---|---|---|---|
| Sample 1 | Objet 1 | Val.Abondance $y_{11}$ | $y_{12}$ |
| Sample 2 | Objet 2 | $y_{21}$ | $y_{22}$ |
| | . | | |
| Sample 6 | Objet *i* | $y_{i1}$ | $y_{i2}$ |
| | . | | |
| | Objet *n* | $y_{n1}$ | $y_{n2}$ |




**Par echantillon**

# What is your problem …

**Your initial matrix : Sample (object) and Variables (ASV/OTU)**

| | Descripteurs | | | | |
|---|---|---|---|---|---|
| | Variable 1 | Variable 2 | Variable $j$ | | Variable $p$ |
| Objets | | | | | |
| Objet 1 | $y_{11}$ | $y_{12}$ | $y_{1j}$ | … | $y_{1p}$ |
| Objet 2 | $y_{21}$ | $y_{22}$ | $y_{2j}$ | … | $y_{2p}$ |
| . | | | | | |
| Objet $i$ | $y_{i1}$ | $y_{i2}$ | $y_{ij}$ | … | $y_{ip}$ |
| . | | | | | |
| Objet $n$ | $y_{n1}$ | $y_{n2}$ | $y_{nj}$ | … | $y_{np}$ |

Val.Abondance

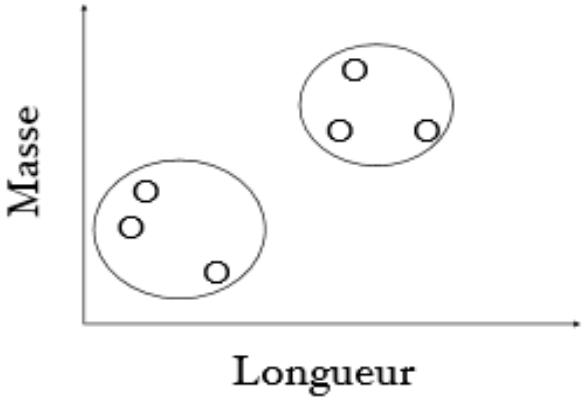
**Variables = Descriptors (Taxa/OTUs/ASV)**

**Objets = Observations (Site, Stations, Env)**

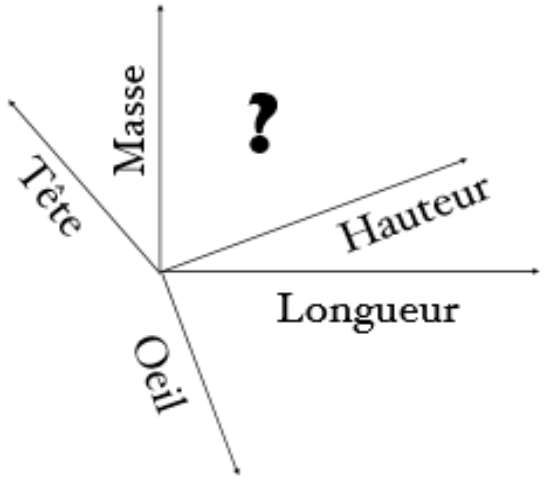
→ **Need a number of axes equal to the number of Descriptors!!!!! ☹**

# How to visualize data in more than 3 dimensions ??

- Problème : visualisation des données en plus de 3 dimensions




2 Variables




p Variables

**Impossible to graphically display all the axes**
**The ordination methods respond to this problem by projecting the variability of all these axes over 2 or even 3 axes that can be visualized**

# β diversity & Multifactorial approaches

**1- Measuring the similarity between objects in a data table (i.e UNIFRAC = phylogenetic distance)**

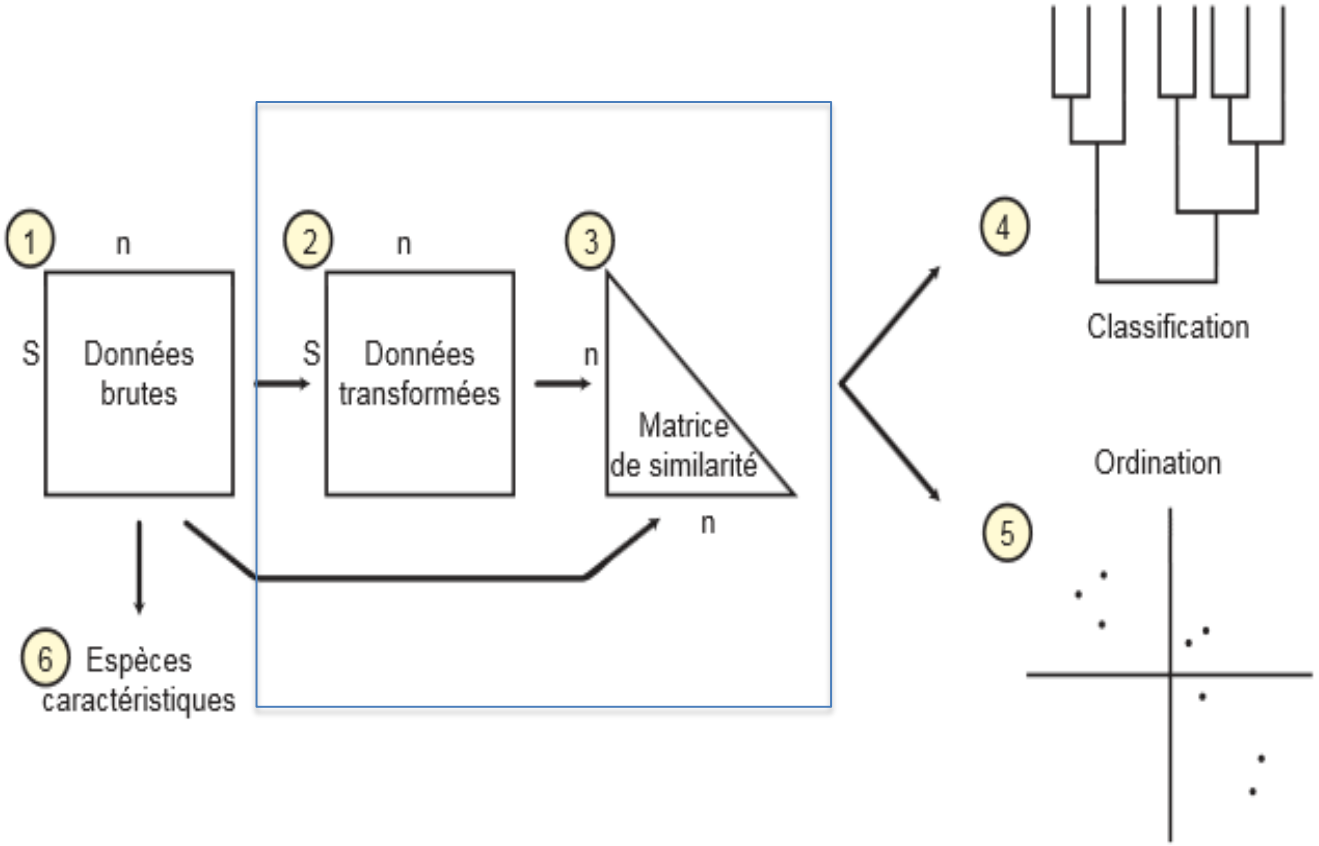
**2- Grouping objects or variables according to these similarities (i.e Hierarchical Clustering)**

**2- AND/OR Ordination of objects and/or variables in a small space to highlight their main structures (i.e PCA, PCoA, AFC, biplots)**

Obtain plots that provide the best possible summary of the information contained in a large data table

→ Minimize the loss of information !! because there will be!

# Overview of this approach



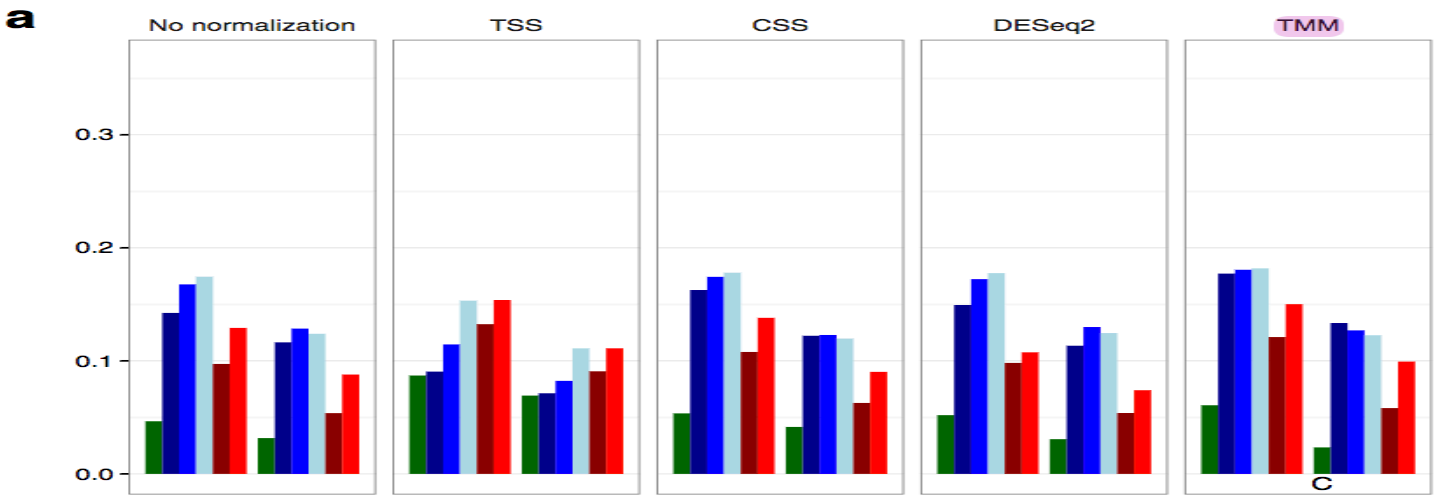
- Normalization
- Transformation
- Distance

# Data Normalization

**Normalization of the sequencing library?**
**→ Depth sequencing effect …**

- **Subsampling**
- **Deseq2**
- **TMM (EdgeR)**
- **CSS (MetagenomeSeq)**

**Negligible effect for the "separation" of the samples !!!**




*Thorsen et al. Microbiome. 2016*

# Transformation des données

**Cas données hétérogènes (unités)**
**Standardisation : Centrage/reduction= Changement d'Unité!**

→ **Transformation centrée-réduite est la solution pour palier aux problèmes d'hétérogénéité des Unités (exple: données physico-chimiques)**

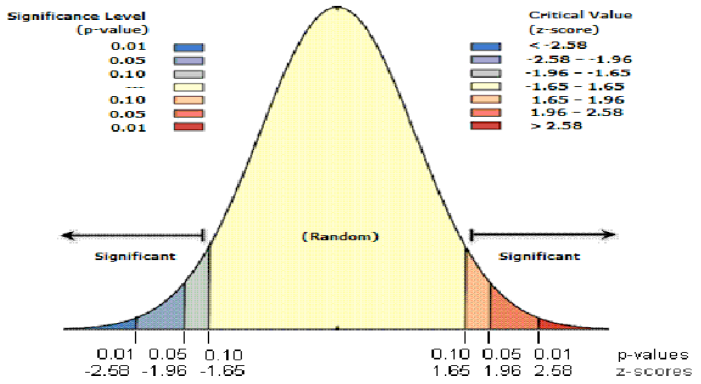
→ **Fonction *scales* dans package Vegan.**

not for ONV table, but for environ table when we have variable that have different unit

**Centrage = retrancher la moyenne à chaque valeur d'une variable.**
Correspond simplement à un changement d'origine qui place la moyenne de la distribution au point 0.

**Réduire une variable = diviser toutes ces valeurs par l'écart type.**

Les données deviennent indépendantes
de l'unité ou de l'échelle choisie
On parle en Unité d'Ecart-type → Z-Scores




- Déformation des données!

# Cas données d'Unités Homogènes (table des abondances)

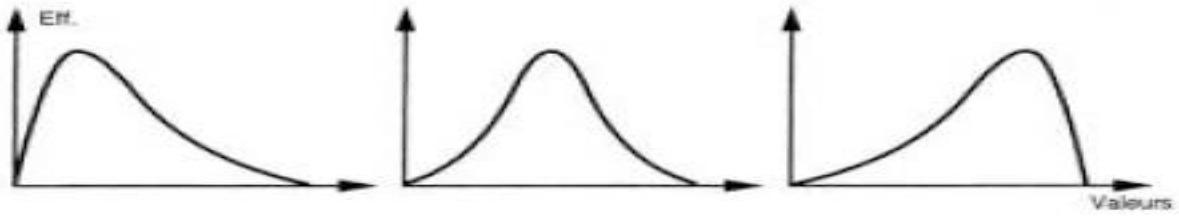
Transformation simples:

- Log x+1 (*log1p()*)
- Racine carrée (*sqrt()*)
- double racine carrée (*sqrt(sqrt()*

→ Réduire les gammes de variations (ne pas donner trop de poids aux valeurs extrêmes)
→ Le zéro représente l'absence d'une espèce
→ Déformation ne sont pas équivalentes…

Idéal c'est réduire les gammes de variations sans trop les écraser!

Réduction des gammes de variations Log>double sqrt >sqrt

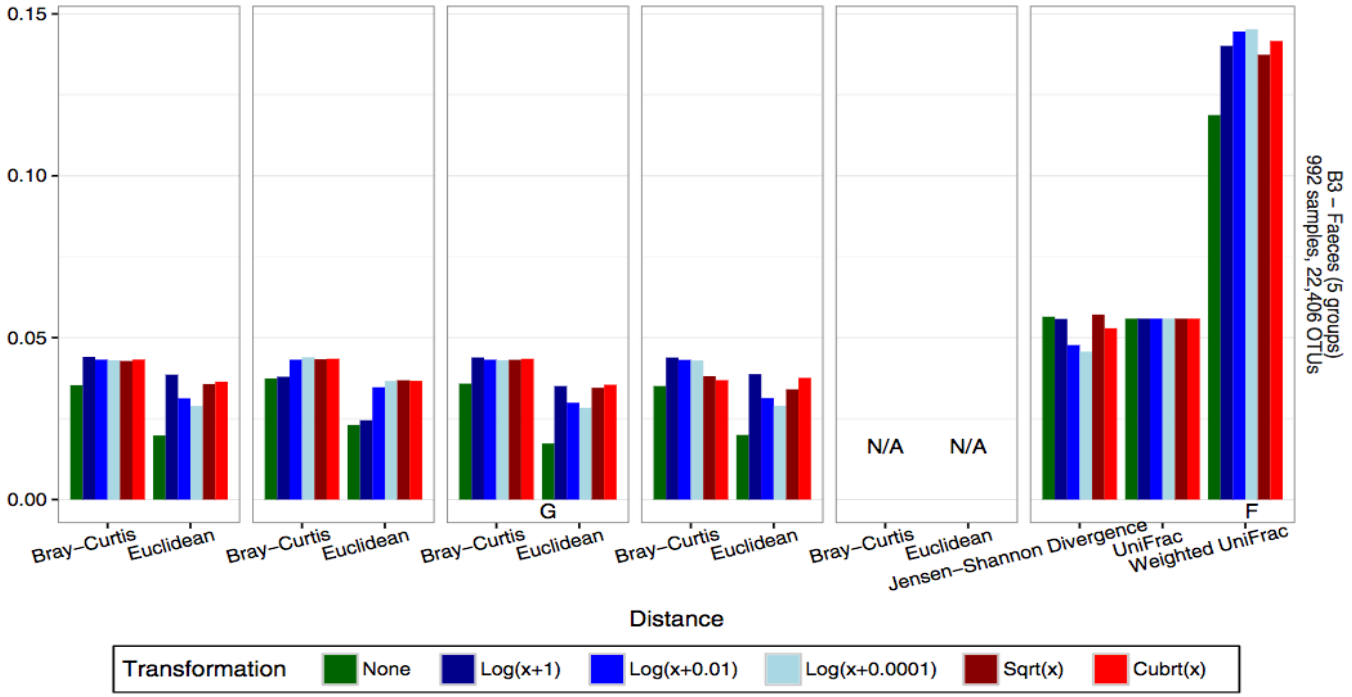



Skewness coef >0

Skewness coef <0

**skewness function R**

# Transformation & distance Impact




Log transformation

→ Reduced the weight of highly abundant ASV/OTUs

→ Increase the weight of low abundant ASV/OTUs

**Distance : Drives the separation of samples**

# « Double Zeros »…. Co-absence

| | Species A | Species B | Species C |
|---|---|---|---|
| Site 1 | 0 | 44 | **0** |
| Site 2 | 11 | 50 | **0** |

is this true??
is this related to depth of sequencing?

- **Sur des tables d'abondance**:

- Co-présence : conditions similaires en terme de niches
- Présence/absence: opposer deux niches
- Co-absence ?

→ **L' absence d'une espèce dans 2 samples simultanément, co-absence** (double zéros), n'est, en règle générale, pas prise en compte pour estimer la ressemblance entre samples/sites!

→Car le doute subsiste sur : Espèces rares non collectées

→ Pas que … et cas très nombreux (trop de poids) vs. Présence

- **Sur des Données/mesures physico-chimiques**:

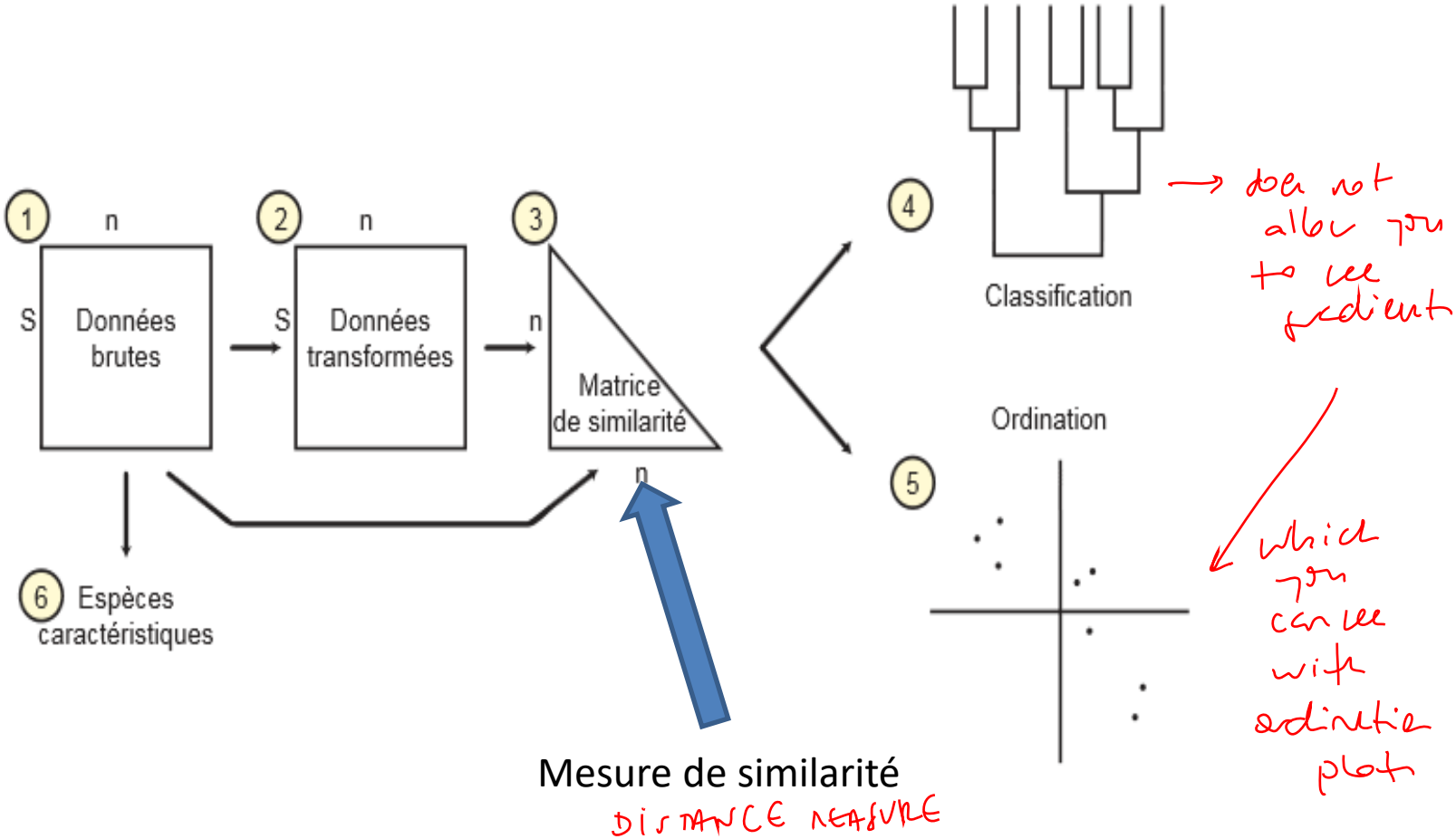
- Double Zéros à une vraie valeur (deux stations avec température de 0°C, etc) & doit être considérée

→ Takes into account of « double zeros » : Symmetric coefficient (Euclidean distance- > PCA)

→ Do not take into account « double zero » : Asymmetric coefficient (Bray-Curtis, Unifrac)

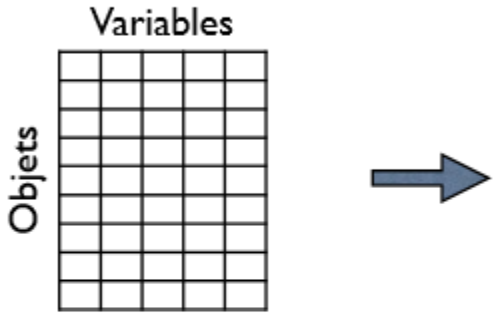
"Euclidean distance is known to be unsuited for ecological distance measurements due to what has been termed the "double zeros" problem,
the fact that it is not possible to distinguish if a species is absent from two samples due to undersampling"
Thorsen et. al, 2016

# DISTANCES

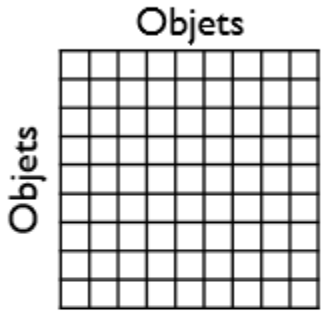

1
n
S
Données brutes
2
n
S
Données transformées
3
n
Matrice de similarité
n
4
Classification
Ordination
5
6
Espèces caractéristiques
Mesure de similarité
DISTANCE MEASURE
→ does not allow you to see gradient
which you can see with ordination plot

# Distance Measure

## Matrices d'association

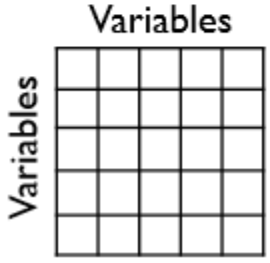





Bray - Curtis
Unifrac
→ ressemblance entre les sites

**Mode Q** : Objets X Objets
*Ressemblance*

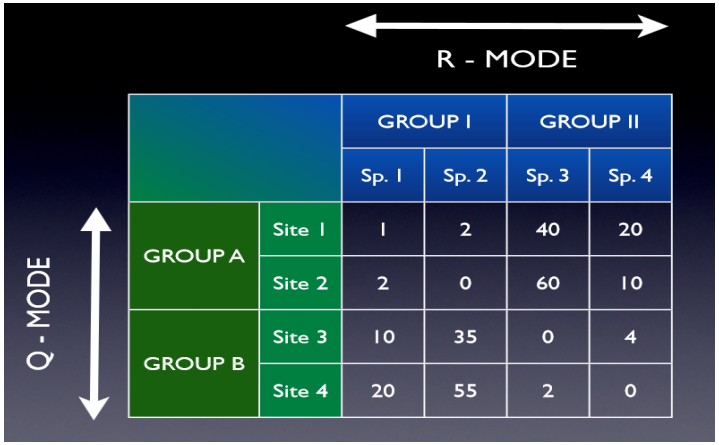




**Mode R** : Variables X Variables
*Dépendance*

# Mesure : Indices d'associations: ressemblance/dissemblance

## Comparaison de toutes les paires possibles d'Objets ou de Descripteurs

- **Entre les paires d'Objets (Env, Sites) : on parle de mode Q**
- **Entre les paires de Descripteurs (OTUs): on parle de mode R**




| | | GROUP I | | GROUP II | |
|---|---|---|---|---|---|
| | | Sp. 1 | Sp. 2 | Sp. 3 | Sp. 4 |
| GROUP A | Site 1 | 1 | 2 | 40 | 20 |
| | Site 2 | 2 | 0 | 60 | 10 |
| GROUP B | Site 3 | 10 | 35 | 0 | 4 |
| | Site 4 | 20 | 55 | 2 | 0 |

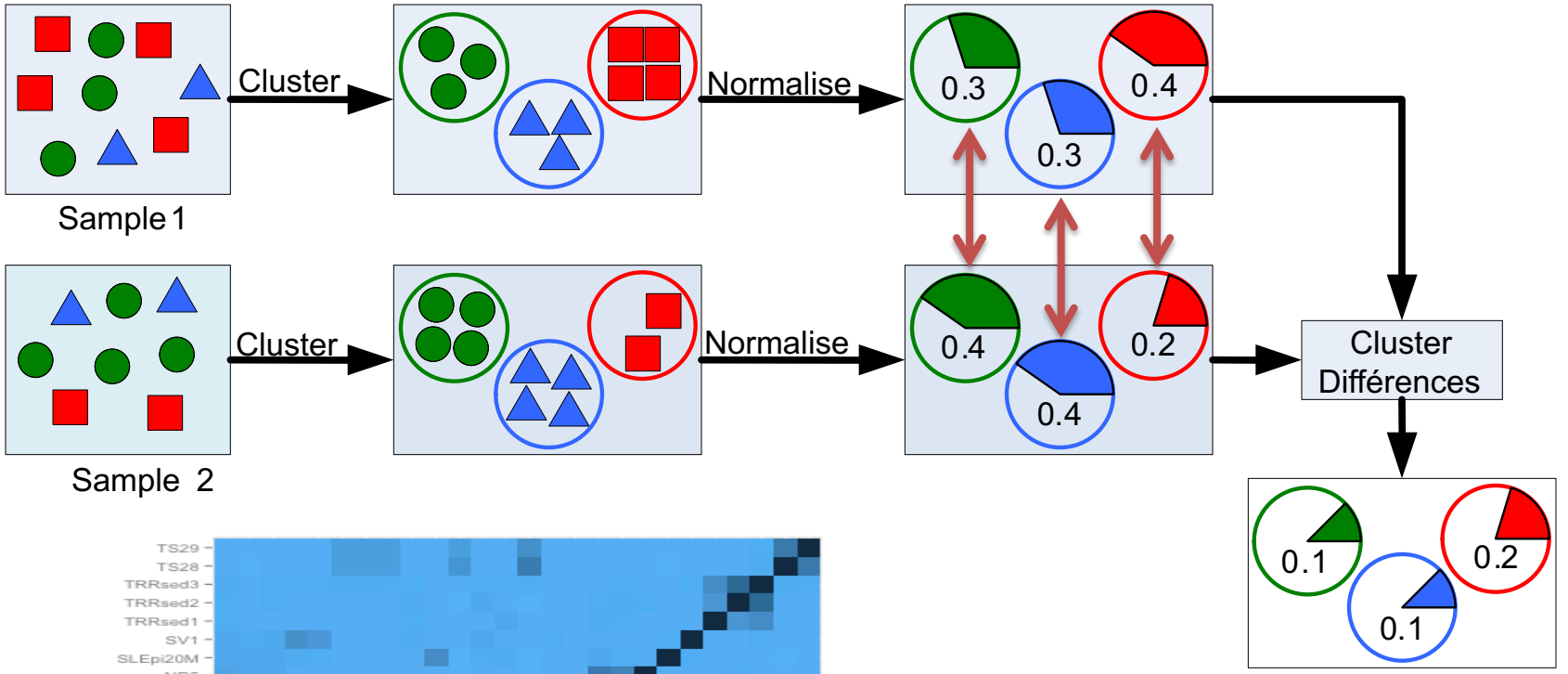
**Pas les mêmes indices d'associations selon le mode :**

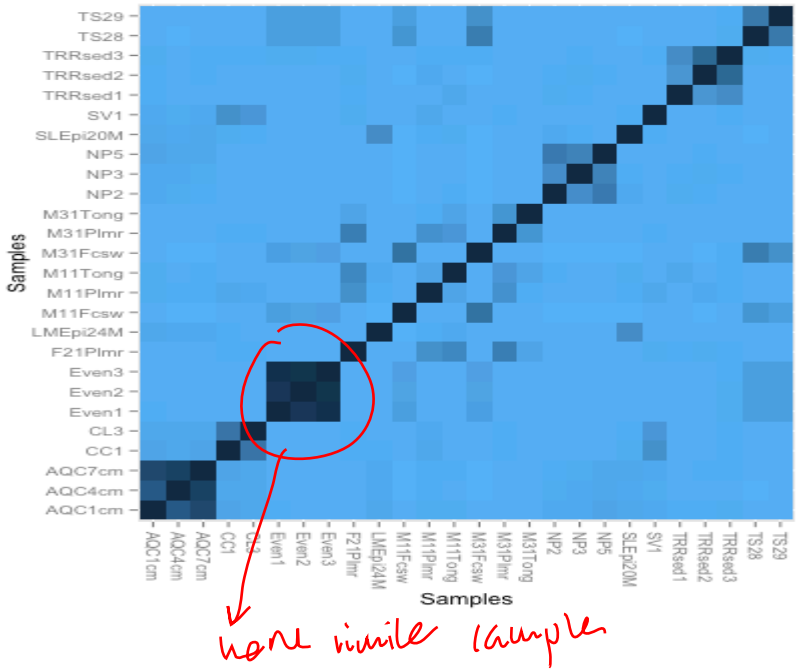
**Mode Q (sites/samples)**: indice mesure la **similarité** ou la **distance** (UNIFRAC)

**Mode R** : utilise coefficient de **dépendance entre les variables (OTU, genus, pH etc)** (indices de corrélation/covariance)

**NB: l'ensemble des comparaisons forment une matrice CARREE (n*n)**

# Similarity measure between Objets: Bray-curtis





Normalisée mesure est entre [0, 1]

**Somme des différences de toutes les abondances**

→OTU faible/OTU fort

# Dissimilarity measure between Objets : UNIFRAC Distance

**UNIFRAC:** Comparison of microbial communities using phylogenetic information

## Measure the difference between the composition of communities from diverse environments

→ **Estimate the proportion of Branch length unique to an environment**

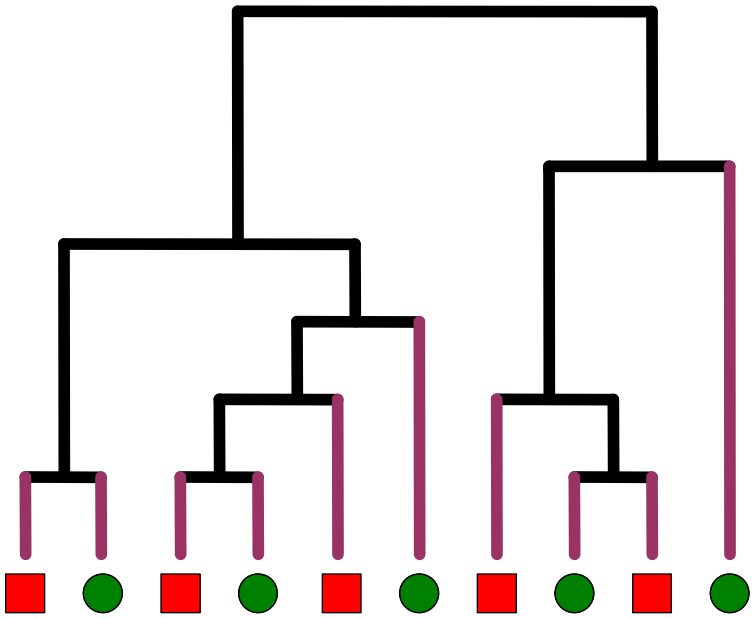
→ **Unique Vs. Shared**

# Concept Unifrac

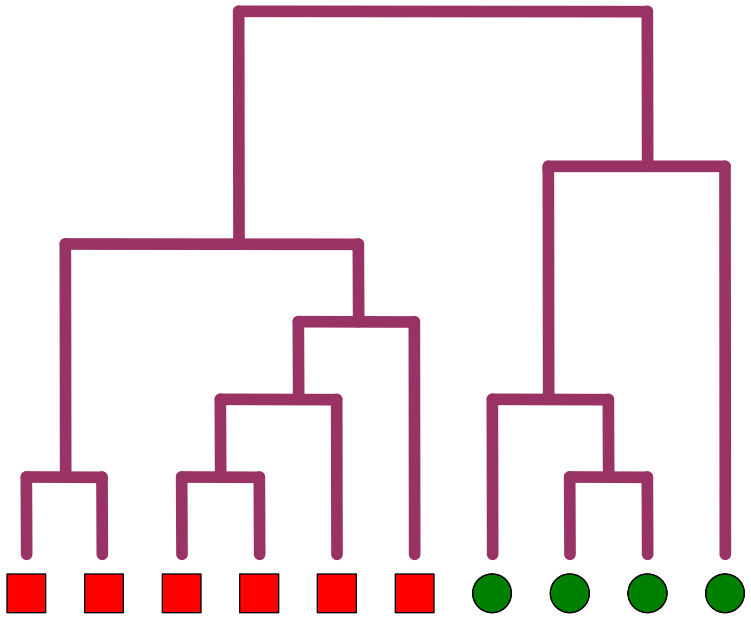
Espèces EnvA  Espèces EnvB

## Unweighted

Communautés Similaires



Différence Max entre Communautés



Distance Measure of UniFrac = (unique) / (partagé + unique)

1 = max dissimilarité (everything is different)

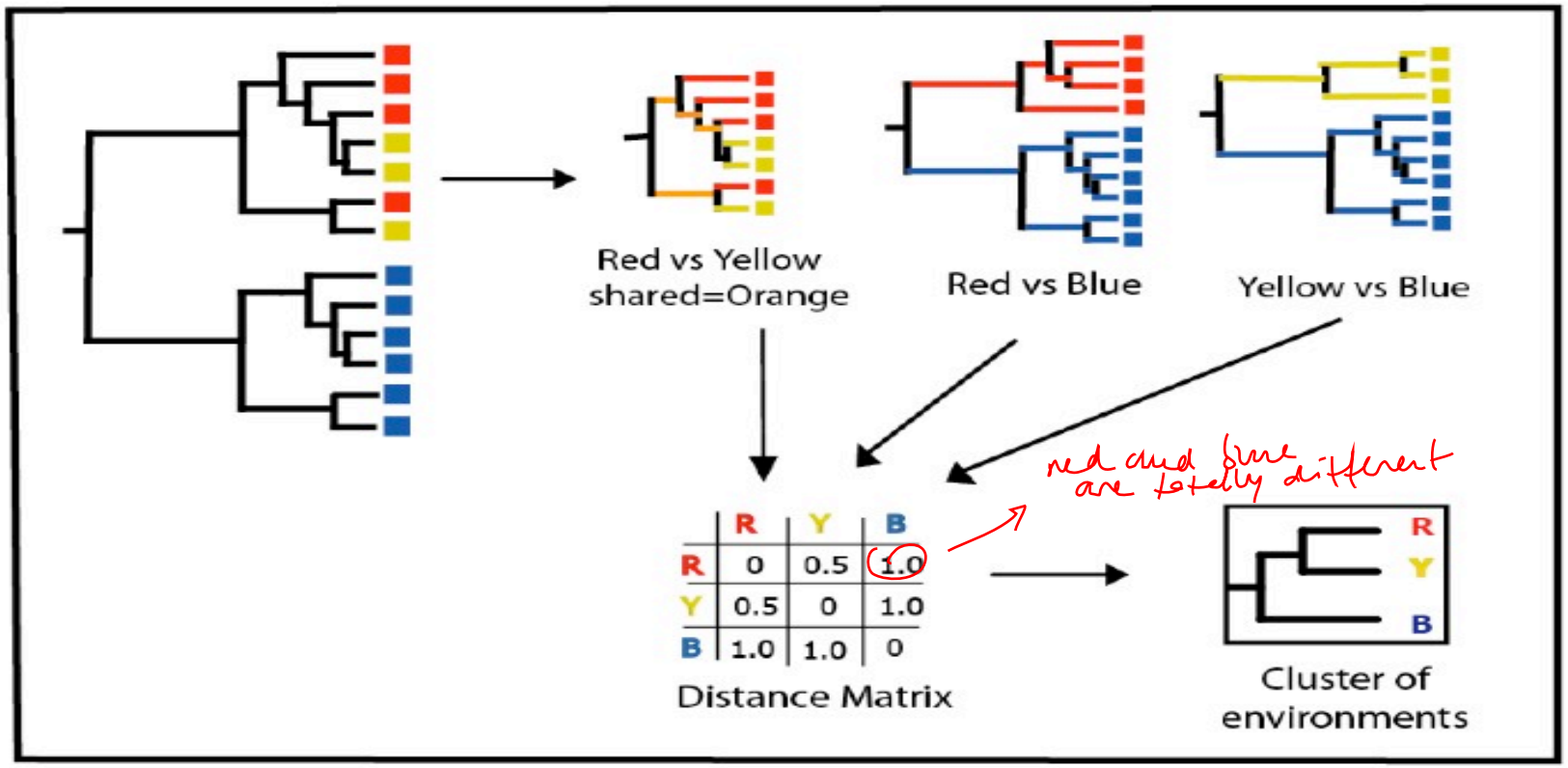
0 = min (everything is equal) = similar

Shared

Unique

D(U)=max=1->

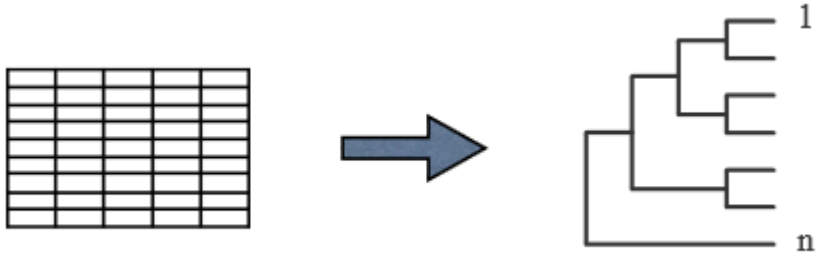
if Share =0



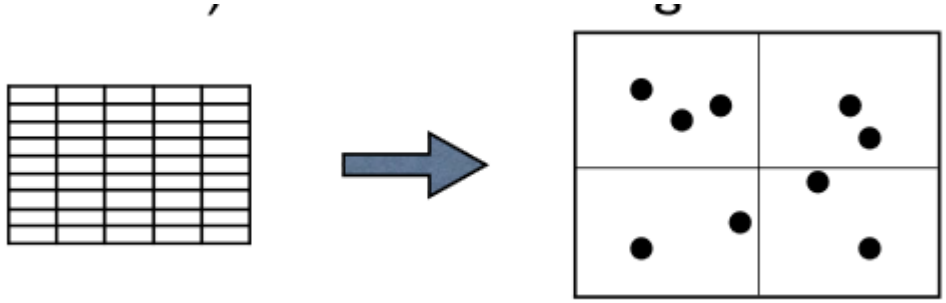

→ **Weighted Unifrac : Weight the length branch by the abundance of taxa**

# Plots from Distance Matrix

- **Clustering : Partition identification**



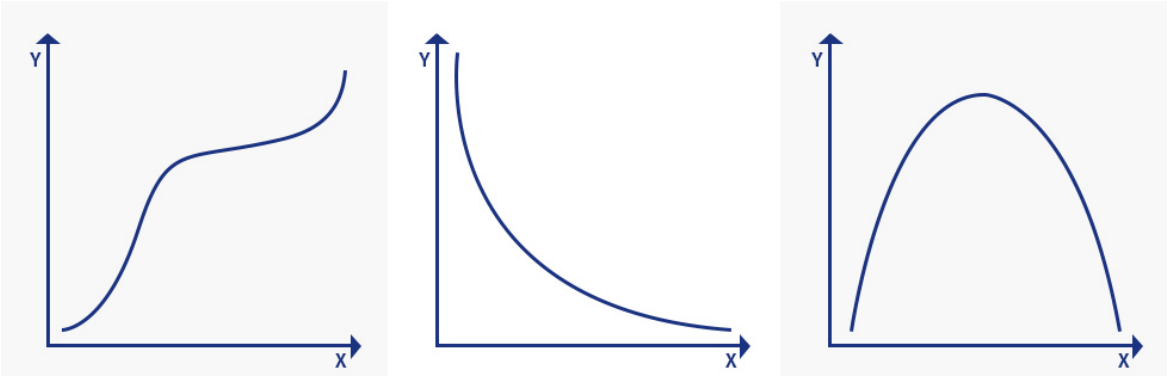

- **Ordination : Gradient identification**



# **Correlation coefficent:** strength & direction of association between two variables

- **Linear : Pearson** (parametric) = covariance → ACP
- **Monotonic non linear**: Spearman (non parametric)
- **Monotonic non linear :** Kendall (non parametric) : alternative to Spearman when small sample size,

NB: Non parametric → rank test, permutation tests



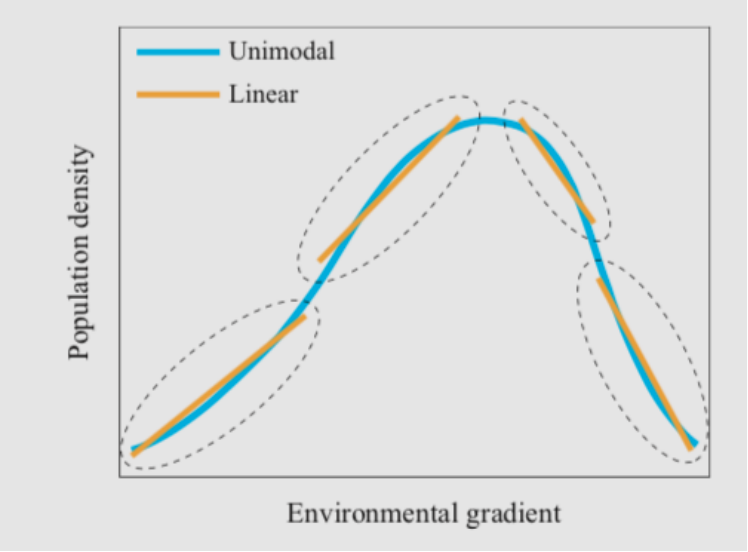
# Models of variable response to environmental gradients

- **The goal of ecological studies is to assess and contrast the relationships between biological entities (species) and their environment**

→ make a specific assumption of the type of the relationship, called « variable response model » -> Mathematics

- linear relationships (rarely in nature)
- Non monotonic relationship with the environment : unimodal shape

Gradient : Spatial, temporal, Ph, nutrients, pertubations etc

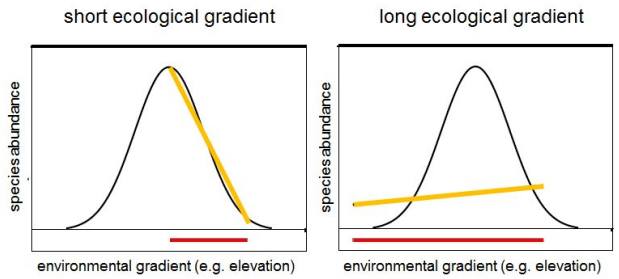


- Linear relationship : **PCA/RDA**
- Unimodal relationship: **CA/CCA**
- Not based on specific underlying model of variable–environment relationship : **PCoA, NMDS**

| Relation | Not constrained | Constrained |
|---|---|---|
| Linear | PCA | RDA |
| Unimodal | CA | CCA |
| **Transformation-based** | Tb-PCA | Tb-RDA |
| **distance-based** | NMDS, PCoA | db-RDA |

**Linear based-method can be used :**
**Transformation of data not sensitive to double-zeros, Euclidean distance**
→ Hellinger distance
→ Chord distance

**Remarque: Linearity → when gradient are low…**



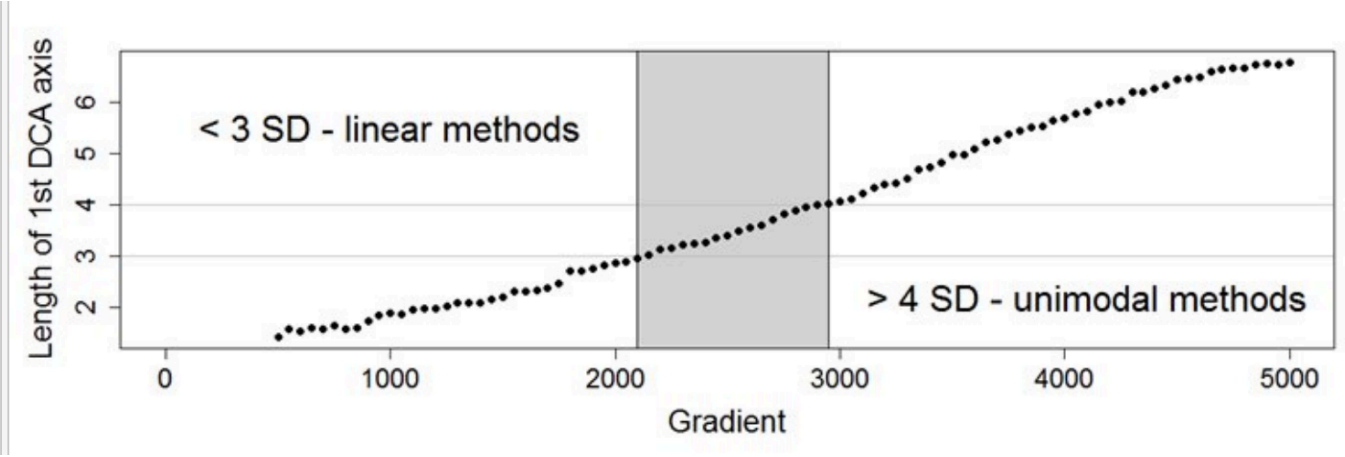
# DCA

- Apply **linear or unimodal ordination method** on your data?
- Use DCA R package
- check the length of the *first* DCA axis

The length of first DCA axis > 4 S.D.

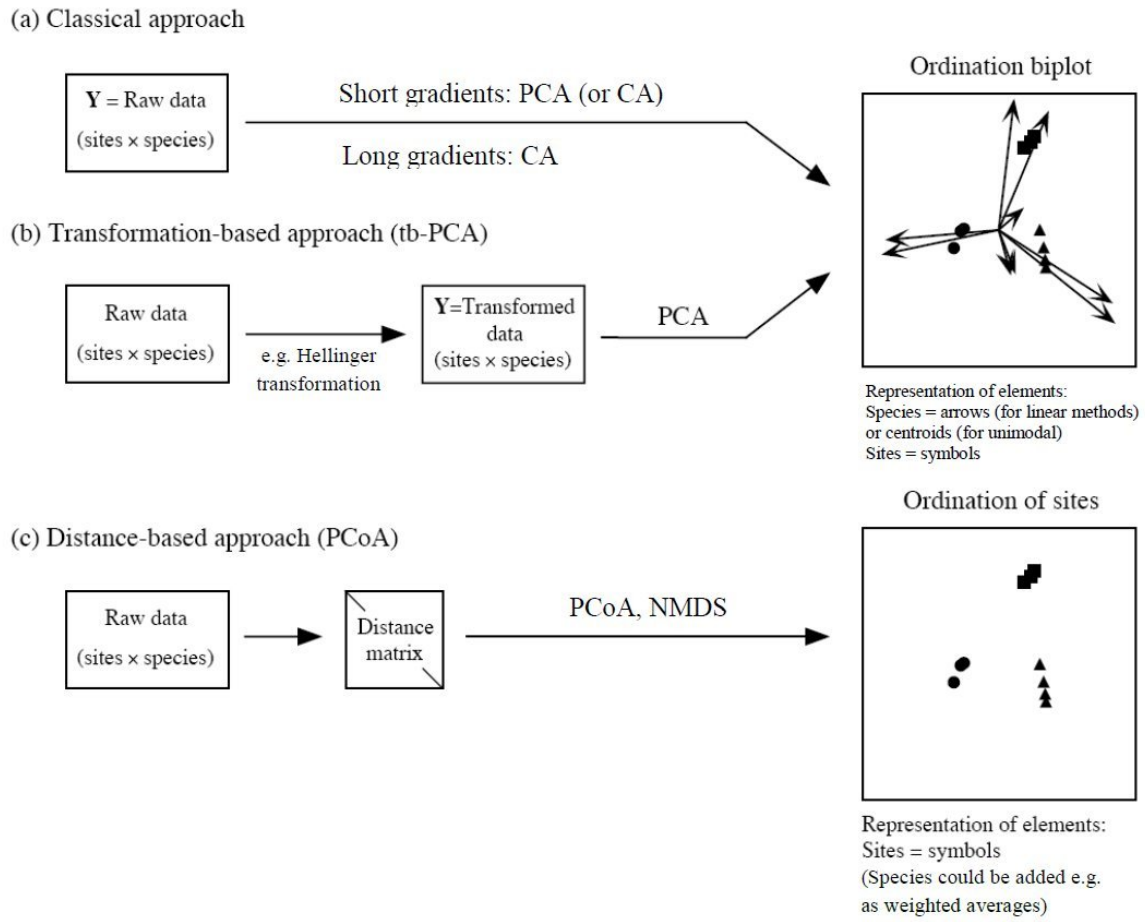
= heterogeneous dataset on which unimodal methods should be used

The length of first DCA axis < 3 S.D. indicates = homogeneous dataset for which linear methods are suitable



# Unconstrained Ordination

## (1) Unconstrained ordination analysis



# The rules

→ Normal distribution

Toutes **les Variables suivent une loi normale** : Multinormalité des données!! Rarement le cas!

→S'abstenir de cette condition si on dispose de **10 fois plus d'objets que de variables** …

→Acceptée pour les jeux de données contenant **au moins plus d'objets que de variables**

# ACP…

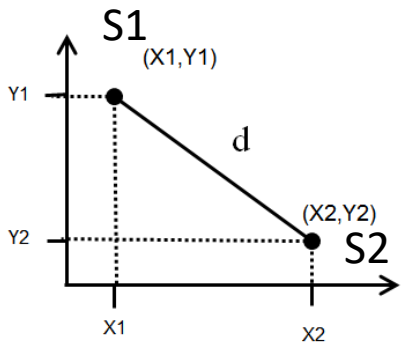
- → Basée sur **les distances euclidiennes**
- → **Double zéro est considéré -> provoque des aberrations (prouvé)**
- → Centrer-réduire (si variabilité des unités)
- → Mode direct sur données quantitatives/semi-quantitatives
- → Si relation existe entre les variables : alors de type linéaire
- → Loi normale but Accepte plus d'objets que de variables

- → Chaque axe : Combinaison linéaire des différentes variables : CP

- → Valeur propre : Représente la variance expliquée par un axe, quantité d'information portée par l'axe

- → Vecteur propre: Direction de l'axe dans l'espace des x paramètres (combinaison linéaire)

| | Objets | Variable 1 OTU-A | Variable 2 OTU-B |
|---|---|---|---|
| Sample 1 | Objet 1 | Val.Abondance | $y_{12}$ |
| Sample 2 | Objet 2 | $y_{21}$ | $y_{22}$ |
| | . | | |
| Sample 6 | Objet $i$ | $y_{i1}$ | $y_{i2}$ |
| | . | | |
| | Objet $n$ | $y_{n1}$ | $y_{n2}$ |




$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

# ACP

## Use Euclidean distance for measuring resemblance among sites?

Abundance table

| | species1 | species2 | species3 |
|---|---|---|---|
| Site1 | 0 | 4 | 8 |
| Site2 | 0 | 1 | 1 |
| Site3 | 1 | 0 | 0 |

Euclidean distance among site are calculated

| | Site1 | site2 | site3 |
|---|---|---|---|
| Site1 | 0 | 7.6 | 9 |
| Site2 | 7.6 | 0 | 1.7 |
| Site3 | 9 | 1.7 | 0 |

→ **Site2-Site3 smaller distance than that of site2-site1!!!??**
→ **Look the shared species, site2 & site3 share 0 species (Orloci paradox)!**
→ **Non sense in ecology**
→ **PCA, excessive impact of the rare species**

Legendre & legendre, Numerical Ecology

# Ordination: PCoA

- Utilise n'importe quelle **matrice de distance** ou de (dis)similarité entre chaque paire d'objets (mode Q)

→ Populaire par l'utilisation distance phylogénétique : Unifrac

## Quand utiliser la PCoA?

Elle permet de réaliser une analyse avec la distance ou dissmilarité de son choix :
Exple: **Bray-curtis, Unifrac**

- Nombre d'objets > Nombre de variables…
- Coefficient asymétrique (Bray Curtis, Unifrac)
- Non parametric

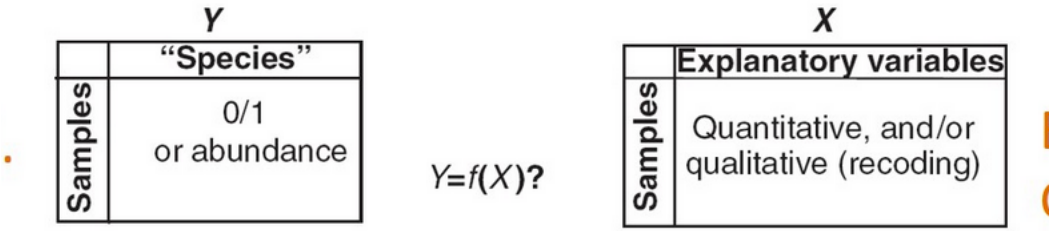
# NMSD

- Assumption: any relation (non parametric)
- A “ mapping technic” : Iteratively repositiones objects in the ordination space
- Goal : **Minimize the Stress function**
- Stress : Obtain the best fit between the object distances in the ordination space and the calculated dissimilarities among objects
- Stress <0.05 good
- Stress < 0.1/0.15 acceptable

# Constrained Analysis
# Canonical Correspondence Analysis

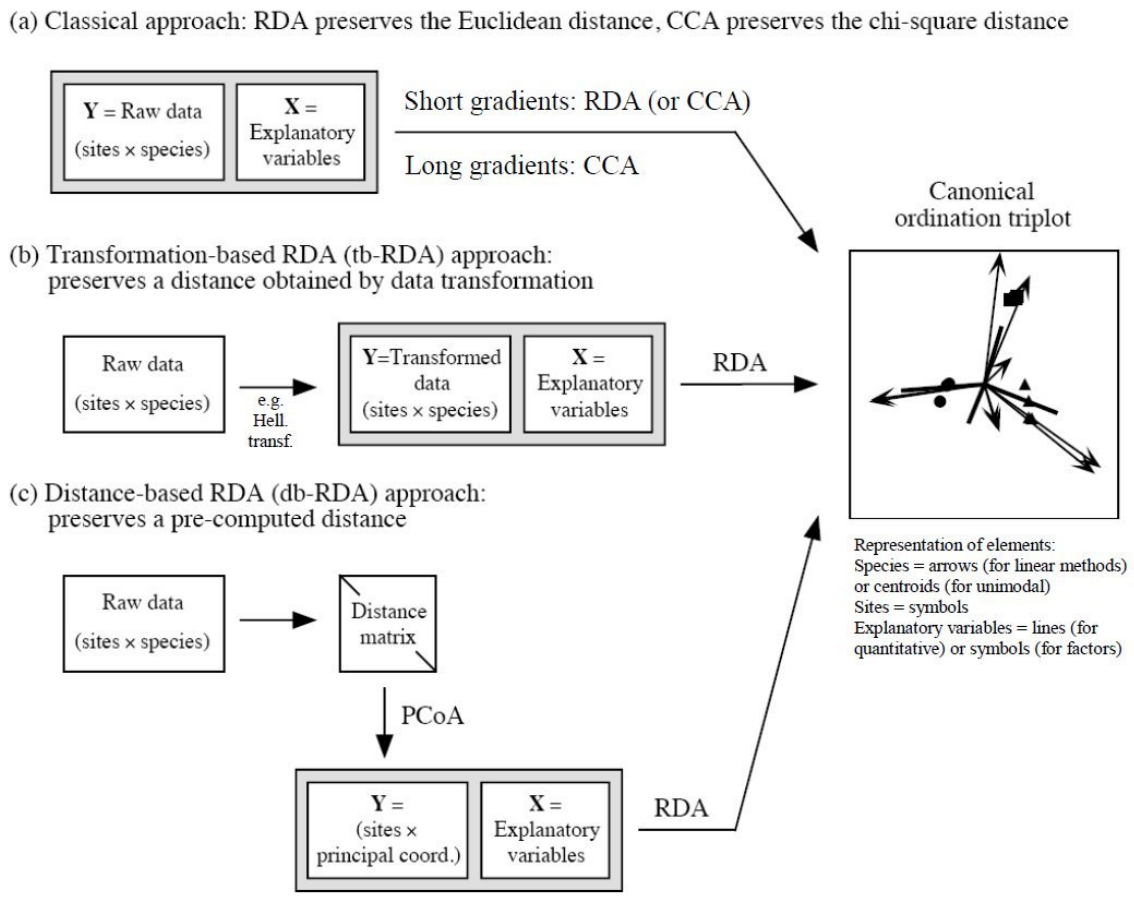
**Link Anbundance species (otu_table) to environmental variables (table Y)**

= Explain the variability of species distribution using environmental variables



# Constrained Ordination

## (2) Constrained ordination analysis




(slightly) modified from Legendre & Legendre (2012)

```
              Inertia Proportion Rank
Total          1.3313     1.0000
Constrained    0.8181     0.6145    8
Unconstrained  0.5132     0.3855    9
Inertia is scaled Chi-square
```

**61% de l'inertie totale expliquée avec les variables env.**
**→ Explique une grande partie de la variabilité dans tableau abondance des espèces**

```
Eigenvalues for constrained axes:
   CCA1    CCA2    CCA3    CCA4    CCA5    CCA6    CCA7    CCA8
0.4143 0.1148 0.0838 0.0761 0.0531 0.0341 0.0227 0.0192
```

→ 0.8181

```
Eigenvalues for unconstrained axes:
     CA1      CA2      CA3      CA4      CA5      CA6      CA7      CA8      CA9
0.12753 0.10441 0.07388 0.06510 0.04901 0.03955 0.02637 0.01551 0.01186
```

→ 0.5132

```
> anova.cca(cca_model)
Permutation test for cca under reduced model
Permutation: free
Number of permutations: 999

Model: cca(formula = abundantlog1 ~ SiOH4 + NO2 + NO3 + NH4 + PO4 + T + S + Sigma_t, data = env.z)
         Df ChiSquare      F Pr(>F)
Model     8   0.81810 1.7933  0.009 **
Residual  9   0.51323
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Les données espèces/site sont linéairement liées aux données variables env.**

# Degré d'importance des variables environnementales : Significativité des variables

```
Model: capscale(formula = otu_table(Final2_rar) ~ SiOH4 + NO2 + I
ray")
          Df SumOfSqs      F Pr(>F)
SiOH4      1  0.58488 3.9908  0.010 **
NO2        1  0.27065 1.8467  0.115
NO3        1  0.14910 1.0174  0.385
NH4        1  0.20366 1.3896  0.205
PO4        1  0.23690 1.6164  0.133
T          1  0.49456 3.3745  0.015 *
S          1  0.29949 2.0435  0.075 .
Sigma_t    1  0.23883 1.6296  0.126
Residual   9  1.31900
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

## Triplots Species, Sites & Environnmental variables

**Samples (sites):** distances between points approximate compositional dissimilarity among samples

**Species:** centroids for unimodal and distance-based methods; centroids indicate the position of the species optima in ordination diagram from which the abundance of given species decreases in all directions!

The distance between site and species position on the triplot is indicative of the abundance of that species at that site

Environmental variables : arrows indicate in which direction the value of environment increases

# Distance-based Redundancy Analysis (db-RDA)

- **Redundancy analysis** is a type of constrained ordination that assesses **how much of the variation** in **one set of variables (species ab.)** can be explained by the variation in **another set of variables (env)**

- Constrained ordination using **non-Euclidean distance measures**

- **Distance matrix** is calculated using the distance measure of choice (bray, unifrac)

# Double Projection : Biplot (ACP)

**Un diagramme  d'ordination qui représente à la fois les Objets et les Variables est un Biplot**

**En ordination, le cadrage (type 1 & type 2) sert à représenter les résultats d'une ordination**

**Il n'existe pas de moyen de représenter de manière optimale à la fois les objets et les descripteurs dans un diagramme d'ordination!!!!**

**La représentation optimale de l'un se fait au prix d'une déformation de l'autre!**

scaling = 1: interet dans l'ordination des sites
# scaling = 2 :interet dans la relation entre les especes