Statistical analysis of meta-omics data

Sandra Plancade
INRA (French Institute of Research in Agriculture)

24 février 2016

Microbial ecosystems

- Microbial ecosystem = population of bacteria that interact in a given environment

  $\hookrightarrow$ Exple : soil, sea water, **gut**

- A varying proportion of bacteria are not genotyped neither cultivable.
- Before metagenomics : analysis of bacteria culture.
- Metagenomics = analysis of bacterial genes in a given biological sample. ($\neq$ genomics = analysis of the genome of a given organism)
- Metagenomics made possible by technological advances.
  $\hookrightarrow$ NGS (next generation sequencing)

Meta-omics data

Meta-omics data= omics data measured on a population of bacteria in a given environment.

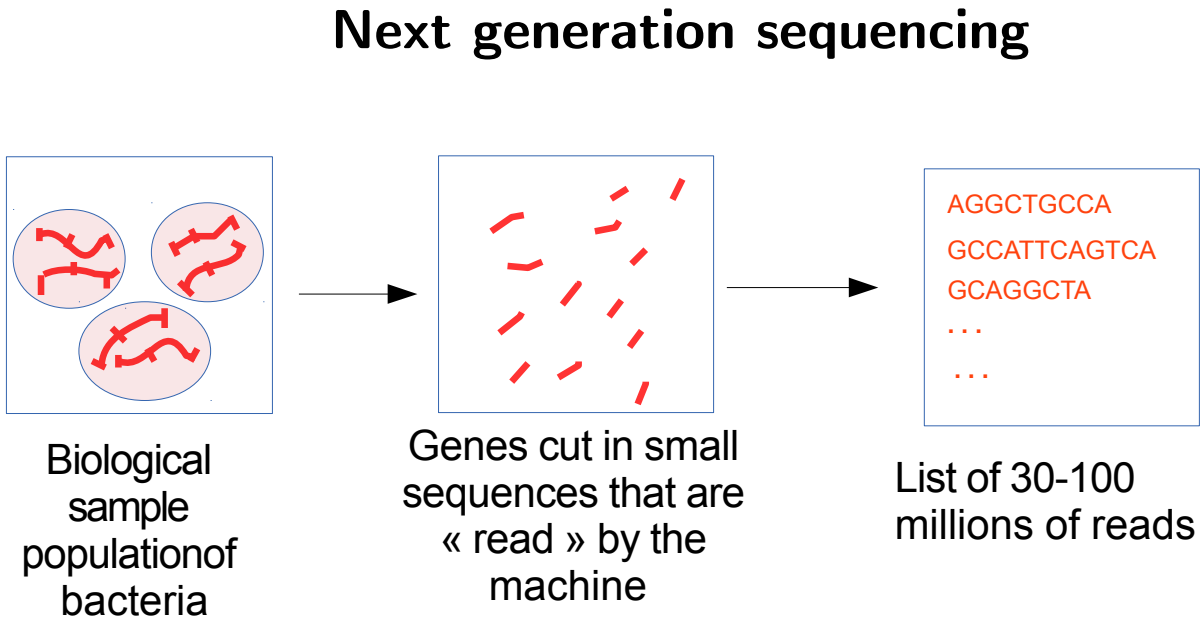
- Metagenomics data = DNA of bacteria. Two types of measures :
  - only 16S gene, characteristic of the species
  - all genes (Whole Genome Sequencing)

  $\hookrightarrow$ widely studied
- Meta-transcriptomics data = RNA of bacteria
- Meta-proteomics data = proteins of bacteria

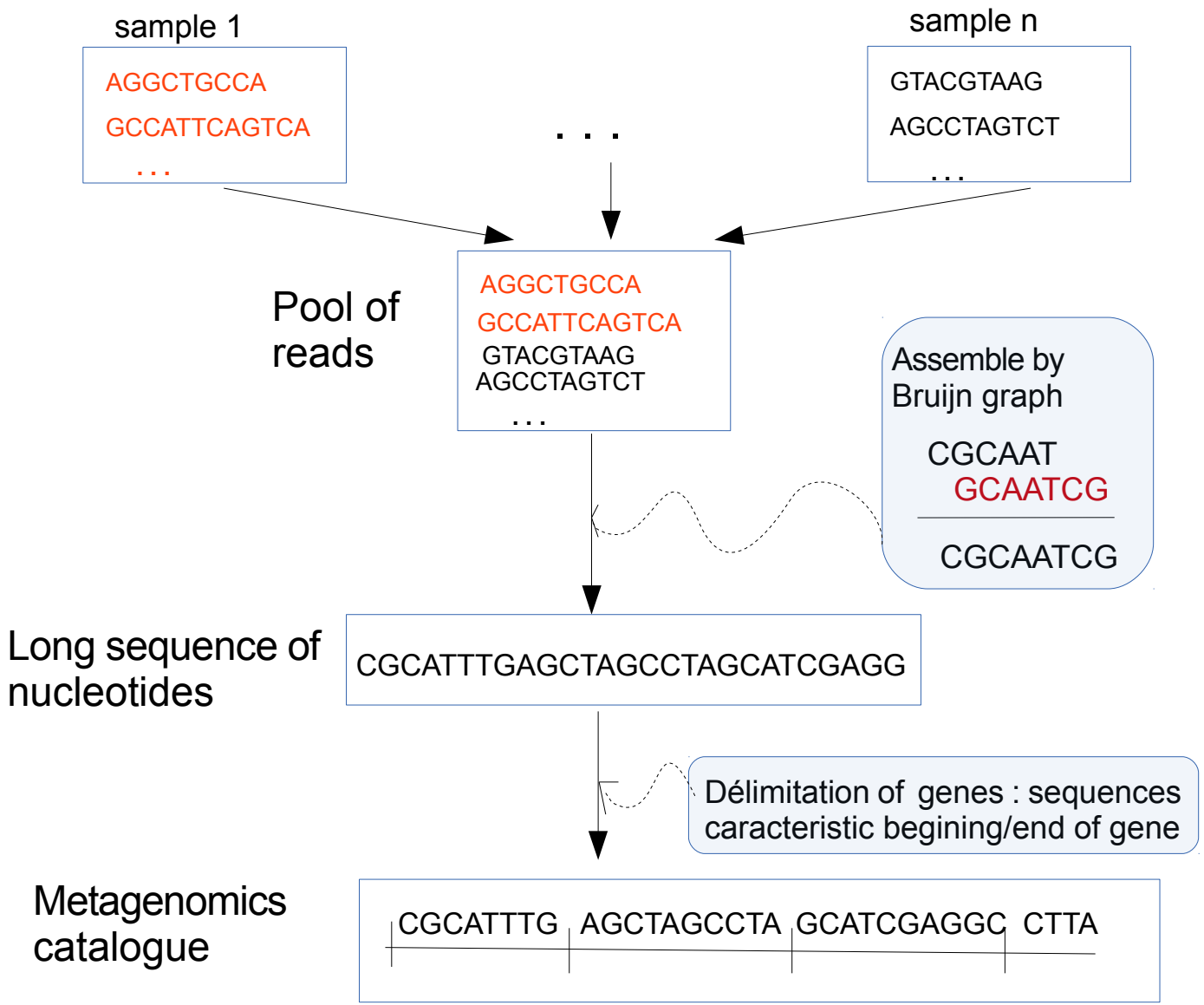
  $\hookrightarrow$ New

DNA	$\rightarrow$	RNA	$\rightarrow$	proteins	$\rightsquigarrow$	function
genomics		transcriptomics		proteomics		metabolomics

Metagenomics WGS (Whole Genome Sequencing) or *shotgun*

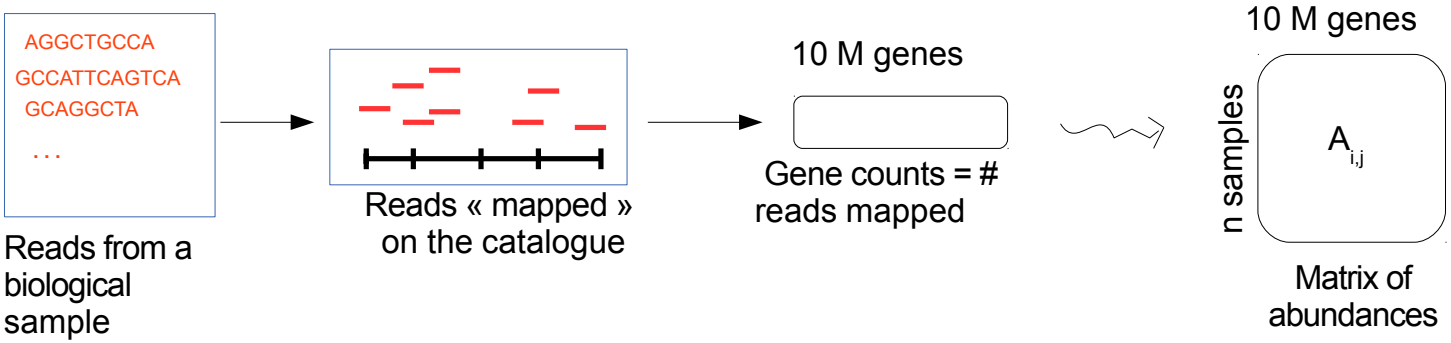


Construction of a catalogue from a large number of sample



↪ In gut, Metahit catalogue = 10 millions of genes.

- **Compute metagenomic abundances** in a biological sample :



$$\text{Abundance of gene g} = \frac{\text{counts of gene g}}{(\text{length of gene g}) \times (\#\text{reads mapped})}$$

- **Characteristics of the data**
  - ◇ High technical variability
  - ◇ Very large dimension : log(p)>n
  - ◇ In gut, 200-500,000 genes present in each sample : high sparsity
- **Dimension reduction**
  - ◇ Grouping of genes based on sequence (similarity between proteins translated in sillico) : COG (Cluster of Orthologous Genes)

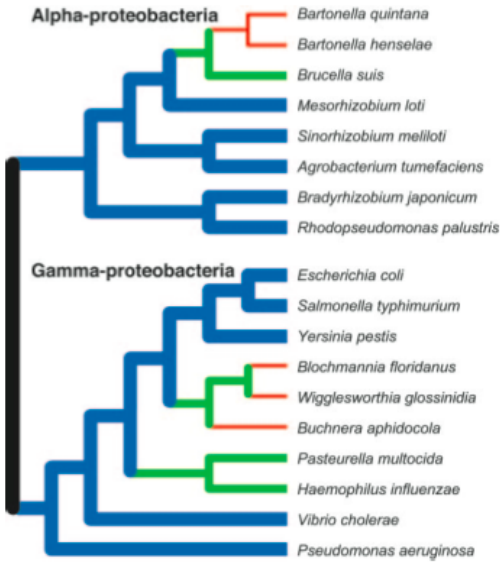
    ↪ Functional grouping.
  - ◇ MGS (MetaGenomics Species) : grouping by covariance of abundances.
  - ◇ Gene annotation (KEGG) : bank of genes whose function has been identified.

    ↪ Limited to known bacterial genes.

16s metagenomics data

- 16s : gene characteristic of species
- Data : matrix of abundances of bacterial species (100/1000 variables)
- Phylogenetic tree : tree that represents evolutionnary relashionships between species.

  ↪ built from distances between the nucleotide sequences of 16s genes.



↪ Structure in variables.

Comparison 16s/WGS

- **16S**
  - Less expensive
  - More widely used ($\Rightarrow$ more specific statistical methods)
  - Less technical variability.
  - Ecology issues : present/absent species in given conditions, co-presence...
- **WGS**
  - Large number of variables
  - High technical variability
  - Functional analysis.

Controverse : phylogenetic grouping correspond approximately to functional grouping

To sum up, metagenomics data are :

- of large/very large dimension
- (very) noisy
- highly correlated
- sparse
- potentially structured

Other meta-omics data

- **Meta-transcriptomics** : similar to metagenomics
- **Meta-proteomics and metabolomics** : Technologies similar to omics (GC-MS, MS-MS)
  - Fractionning of molecules (metabolites/proteins) in fragments (ions/peptides)
  - Identifications of fragments by their M/Z spectra compared to a bank of peptides/ions
  - Recovering of molecules abundances.

  Difficulty : identification requires alignement, more difficult for molecules present in few biological samples.

General biological issues

- **Ecology** : description of species present in the environment.
  - ◇ Difference between conditions (ex :comparison of soil samples from different geographics area)
  - ◇ Co-presence of species.
- **Functionality** : how does microbiote works ?
  - ◇ Interactions between bacteria
  - ◇ Link between microbiote and phenotypes/omics data

↪ Related statistical questions may be unprecised.

Usual statistical approaches

- **Multiple testing** (differential analysis)
  - ◇ zero-inflated parametric models.
  - ◇ permutation tests [White *et al*, PLoS Comput. Bio. 2009]

- **Mixed models** (multiple time-points) [Le Cao *et al* 2015]

$$X_i^j(t) = \underbrace{f_j(t)}_{\text{time effect : splines}} + \underbrace{\alpha_i^j + \beta_i^j t}_{\text{random individual effect}} + \varepsilon_{i,j}(t)$$

- **Adaptation of multivariate analysis methods**
  - ◇ Centered Log-Ratio transformation + methods based on correlation (PLS...)
  - ◇ Variance decomposition (multi-sites measurements)
  - ◇ Methodes based on distance matrices
  - ◇ Penalisation contraining structure based on phylogenic trees [Chen 2012]

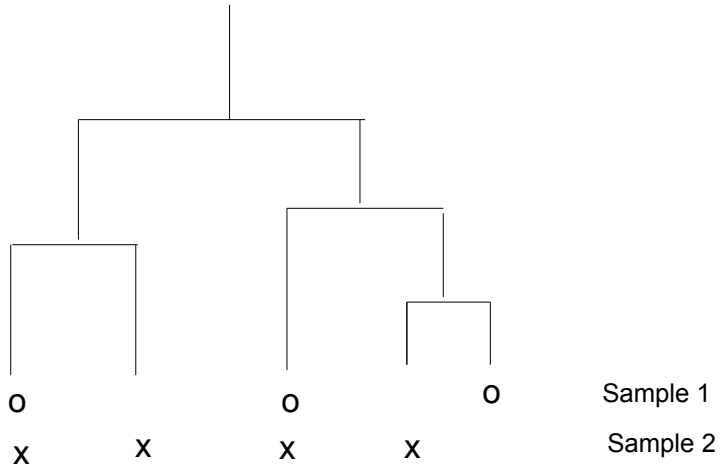
- **Variables selection** by sparse multivariate methods
- **Bi-clustering** : Non-negative Matrix Factorization
- **Network inference** : GGM

Example of anaysis based on distance matrices

- **Goal** : test the effect of race on rumen microbiote for cow.
- **Data** :
  - $(X_{u,k})$, $u = 1, \ldots, N$, $k = 1, \ldots, p$: 16S measurement of abundances in p bacterial species for N cows
  - $Y_u \in \{1, \ldots, a\}$: races
  - "ANOVA" notations : $X_{i,j,k}$: $i = 1, \ldots, a$: category (race)
    $j = 1, \ldots, n$: repetition (cow)
    $k = 1, \ldots, p$: variable (species)

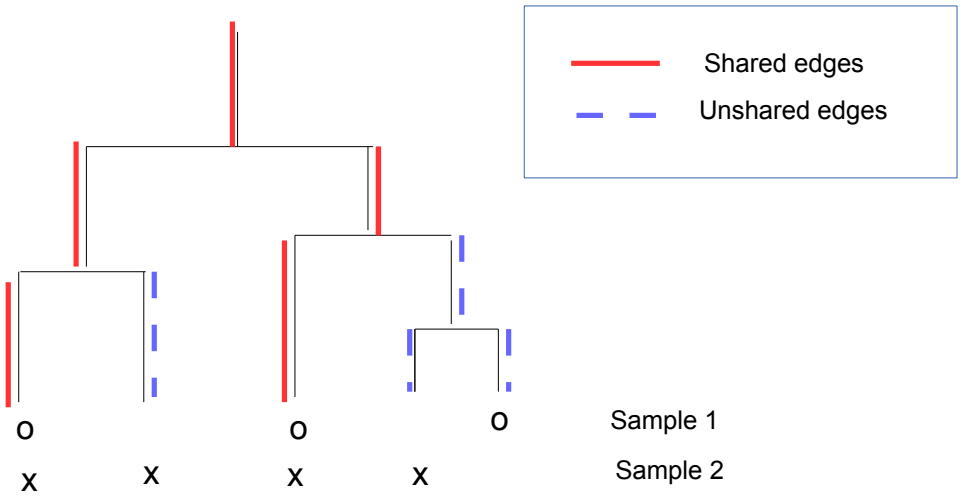
Example of anaysis based on distance matrices

- **Goal** : test the effect of race on rumen microbiote for cow.
- **Data** :
  - $(X_{u,k})$, $u = 1, \ldots, N$, $k = 1, \ldots, p$: 16S measurement of abundances in p bacterial species for N cows
  - $Y_u \in \{1, \ldots, a\}$: races
  - "ANOVA" notations : $X_{i,j,k}$: $i = 1, \ldots, a$: category (race)
    $j = 1, \ldots, n$: repetition (cow)
    $k = 1, \ldots, p$: variable (species)
- **Unifrac distance** based on phylogeny between 2 16S samples.



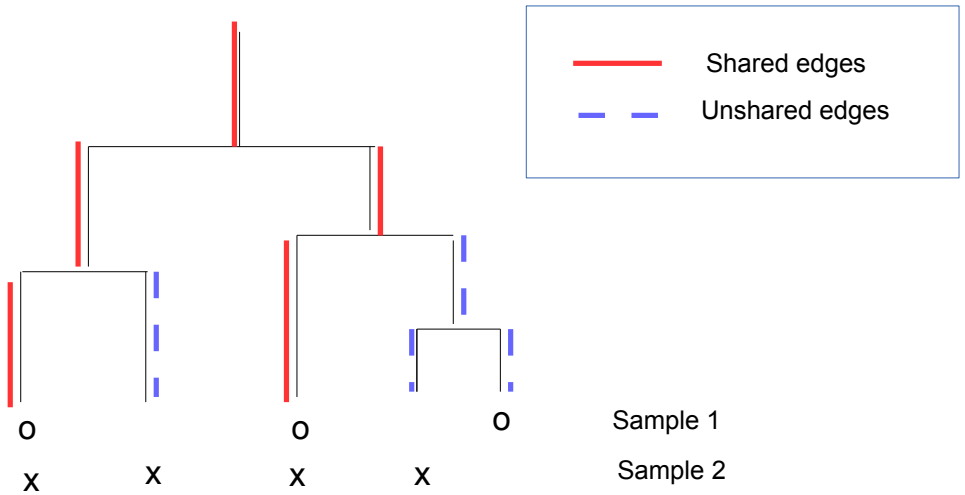
Example of anaysis based on distance matrices

- **Goal** : test the effect of race on rumen microbiote for cow.
- **Data** :
  - $(X_{u,k})$, $u = 1, \ldots, N$, $k = 1, \ldots, p$: 16S measurement of abundances in p bacterial species for N cows
  - $Y_u \in \{1, \ldots, a\}$: races
  - "ANOVA" notations : $X_{i,j,k}$: $i = 1, \ldots, a$: category (race)
    $j = 1, \ldots, n$: repetition (cow)
    $k = 1, \ldots, p$: variable (species)
- **Unifrac distance** based on phylogeny between 2 16S samples.



Example of anaysis based on distance matrices

- **Goal** : test the effect of race on rumen microbiote for cow.
- **Data** :
  - $(X_{u,k})$, $u = 1, \ldots, N$, $k = 1, \ldots, p$: 16S measurement of abundances in p bacterial species for N cows
  - $Y_u \in \{1, \ldots, a\}$: races
  - "ANOVA" notations : $X_{i,j,k}$: $i = 1, \ldots, a$: category (race)
    $j = 1, \ldots, n$: repetition (cow)
    $k = 1, \ldots, p$: variable (species)
- **Unifrac distance** based on phylogeny between 2 16S samples.



$$\text{dist(samp 1, samp 2)} = \frac{\text{sum length unshared edges}}{\text{sum length all edges}}$$

- Geometric MANOVA :

$$SS_W = \sum_{i=1}^{a}\sum_{j=1}^{n}\sum_{k=1}^{p}(X_{i,j,k} - X_{i,\cdot,k})^2 = \frac{1}{n}\sum_{\text{pairs}(u,v)} d_{u,v}^2\delta_{u,v}$$

with $d_{u,v}$ the euclidean distance betwen X_u et X_v and

$$\delta_{u,v} = \begin{cases} 1 & \text{if } (u,v) \text{ in same category} \\ 0 & \text{otherwise} \end{cases}$$

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}\sum_{k=1}^{p}(X_{i,j,k} - X_{\cdot,\cdot,k})^2 = \frac{1}{N}\sum_{\text{pairs}(u,v)} d_{u,v}^2$$

- PERMANOVA :
  - $d_{u,v}$ replaced by $D_{u,v}$
  - Test statistic : SS_W/SS_T
  - Distribution under H_0 : permutations

Nonnegative Matrix Factorization (NMF)

The NMF model : an interpretable dimension reduction

- $X^{n,p}$ matrix of abundances in p metagenomic groups in n samples.
- Hypothesis :
  - Abundances organised in $k << \min(n,p)$ pathways $h_1, \dots h_k$ characterised by their proportion in metagenomic groups

$$h_\ell = (H_{\ell,1}, \dots, H_{\ell,p})$$

  - Samples $i = 1, \dots, n$ carcterised by their abundances in pathways :

$$w_i = (W_{i,1}, \dots, W_{i,k})$$

- Therefore

$$X \approx WH$$

with $W, H \geqslant 0$.

Estimation of NMF

$$\arg \min_{W,H \geqslant 0} D(X, WH) + pen(W) + pen(H)$$

- Matrix distance $D \leftrightarrow$ log-likelihood of a parametric model
  - $X_{i,j} \sim \mathcal{N}\left((WH)_{i,j}, \sigma^2\right) \Leftrightarrow$ LL=$D_{Frob}(X, WH) + cte$
  - $X_{i,j} \sim \mathcal{P}\left((WH)_{i,j}\right) \Leftrightarrow$ LL=$D_{KL}(X, WH) + cte$

  $\hookrightarrow$ In practice : choice of distance depends on the field (signal theory : KL, genomics : Frobenius)
- Selection of dimension k of the reduced space : several empirical criteria
- Choice of penalisation (ex : favour sparse pathways)
- Algorithm : alternated minimisation/decreasing of the criterion (bi-convex)

Comment : Under constraints that individual profiles w_i have one non-zero term, the minimsation problem is equivalent to k-means

NMF in literature

- In omics, NMF often used for bi-clustering
- Methodological research : mainly algorithmic
- To my best knowlegde, no theoretical analysis with a statistical point of view.
- **PhD :** *Inferring agregated functional traits from metagenomics data : application to fiber digestion in gut microbiota* [Sebastien Raguideau, 2016]
  - Select groups of genes that catalyse elementary reactions associated to fiber digestion (KEGG)
  - Build a graph of constraints based on metabolites degradated and produced by elementary reactions
  - Build agregated functional traits by NMF under constraints of connectivity on the graph.

A statistical point of view on NMF

- Definition of a statistical model
- Analysis of criteria of selection of k
- Issue 1 : non-unicity of decomposition (W, H) ("ill-posed" problem)
  $\hookrightarrow$ Sufficient criterion for unicity : rows of H orthogonal.
- Issue 2 : general approach ?
  - Assume a predefined number k of pathways ? (parametric point of view)
  - Compromise bias/variance, reconstruction/stability, where optimal k depends on n ? (nonparametric point of view)

Meta-proteomics data : use of technical replicates

- Proteocardis project : 150 biological samples/4 pathologies, 8 samples with 6 technical replicates.
  $\hookrightarrow$ first large scale project (shotgun - 200 biological samples)
- Goal of the project : discriminant analysis /variable selection.
- Secondary goal : characterise technical variability in meta-proteomics data
  - ◇ Exple : thresholding of low counts : $X^r_{i,j}$ (sample $i = 1, \ldots, n$, variable j, replicate r), estimate

$$p_a = P[X^r_{i,j} = 0 | X^{r'}_{i,j} = a, r \neq r']$$

Question : Use of technical replicates in variable selection

- General idea : variations between replicates provide a "level" for the significance of biological difference.
- Mixed models ?
- Multivariate analysis ?