Nonnegative Matrix Factorization for analysis of metabolic pathways associated to fiber digestion from metagenomics data.

Sandra Plancade
Sébastien Raguideau
Béatrice Laroche
Marion Leclerc

17 Novembre 2015

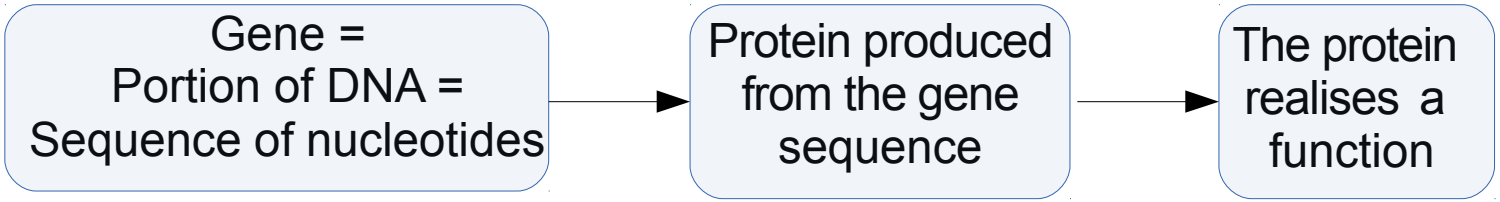
Metagenomics

Metagenomics

- **Microbial ecosystem** : population of bacteria in a given environment.
  ↪ Ex : gut, skin, soil, sea water...

Metagenomics

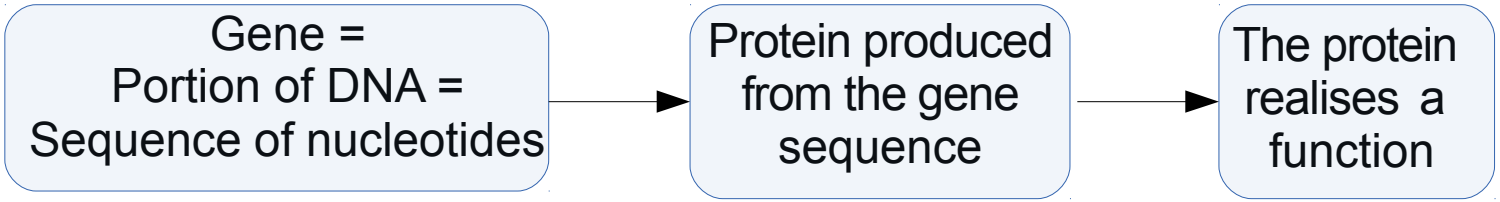
- **Microbial ecosystem** : population of bacteria in a given environment.
  ↪ Ex : gut, skin, soil, sea water...
- These bacteria ensure **functions** programmed by their genes (DNA).



  ↪ Ex : degradation of chemical compounds
  ↪ By extension, we talk about the **function of a gene**

Metagenomics

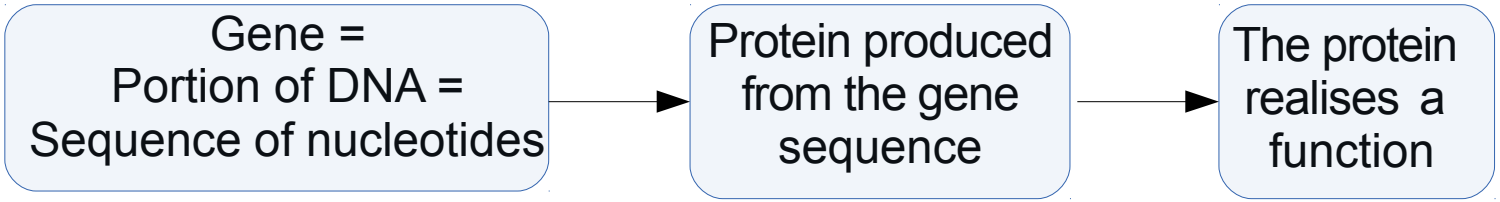
- **Microbial ecosystem** : population of bacteria in a given environment.
  ↪ Ex : gut, skin, soil, sea water...
- These bacteria ensure **functions** programmed by their genes (DNA).



  ↪ Ex : degradation of chemical compounds
  ↪ By extension, we talk about the **function of a gene**
- **Metagenomics** : analysis of the genetic material of bacteria from environment samples.
  ↪ A large proportion of these bacteria are unknown

Metagenomics

- **Microbial ecosystem** : population of bacteria in a given environment.
  ↪ Ex : gut, skin, soil, sea water...
- These bacteria ensure **functions** programmed by their genes (DNA).



  ↪ Ex : degradation of chemical compounds
  ↪ By extension, we talk about the **function of a gene**
- **Metagenomics** : analysis of the genetic material of bacteria from environment samples.
  ↪ A large proportion of these bacteria are unknown
- **Genomics and metagenomics**
  Genomics : analysis of the genome of a given organism (animal or plant)
  Metagenomics : analysis of genes in a given environment sample
  ↪ genes are not gathered by organisms.

Measurement of metagenomic abundances

Measurement of metagenomic abundances

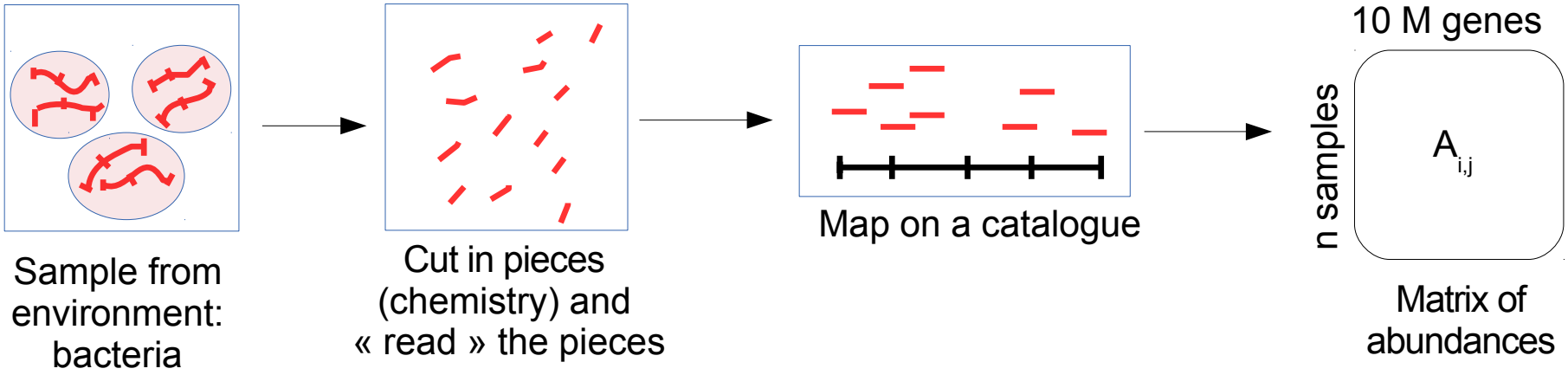
- Genes **from all bacteria** are measured together.

Measurement of metagenomic abundances

- Genes **from all bacteria** are measured together.
- A **catalogue** is built by assembling pieces of DNA from a large number of samples $\leftrightsquigarrow$ "metagenome" including genes from all bacteria.
  $\hookrightarrow$ catalogue for gut microbiote : 10^7 genes.

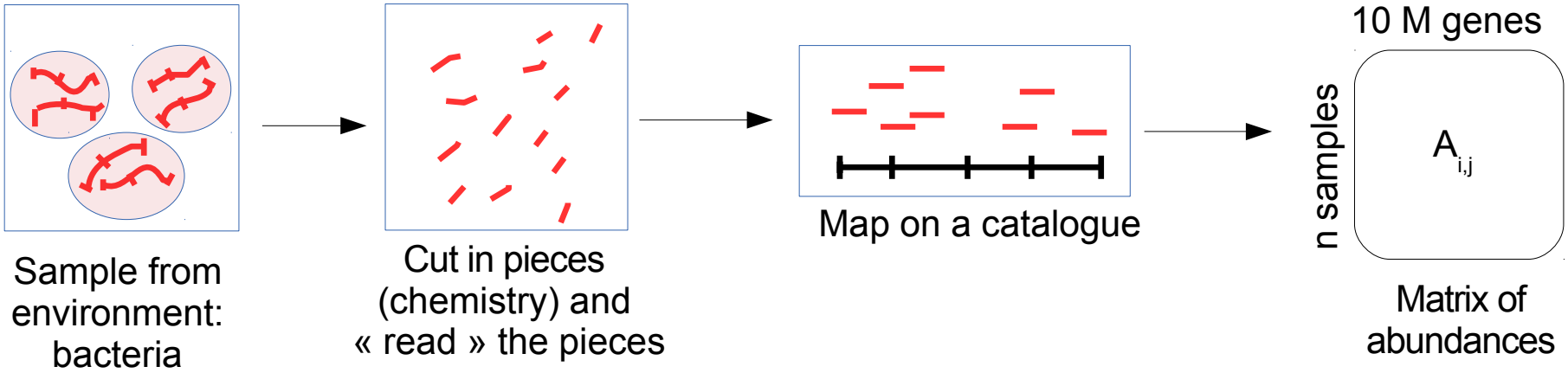
Measurement of metagenomic abundances

- Genes **from all bacteria** are measured together.
- A **catalogue** is built by assembling pieces of DNA from a large number of samples $\leftrightsquigarrow$ "metagenome" including genes from all bacteria.
  $\hookrightarrow$ catalogue for gut microbiote : 10^7 genes.
- Genes are cut in pieces, whose sequences are read and mapped on the catalogue.



Measurement of metagenomic abundances

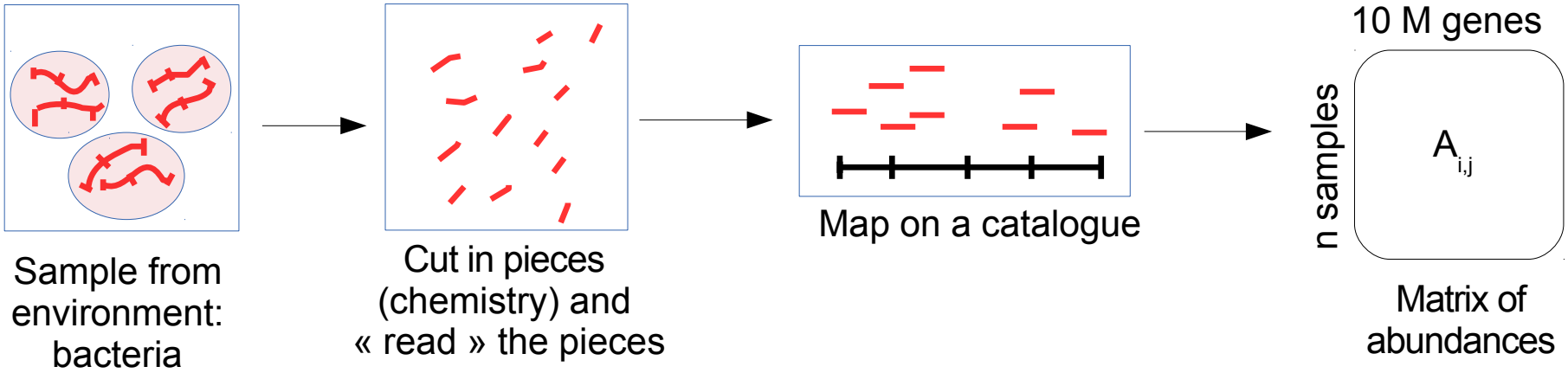
- Genes **from all bacteria** are measured together.
- A **catalogue** is built by assembling pieces of DNA from a large number of samples $\leftrightsquigarrow$ "metagenome" including genes from all bacteria.
  $\hookrightarrow$ catalogue for gut microbiote : 10^7 genes.
- Genes are cut in pieces, whose sequences are read and mapped on the catalogue.



- **Abundance of gene** j = proportion of gene j among all genes present in the sample.

Measurement of metagenomic abundances

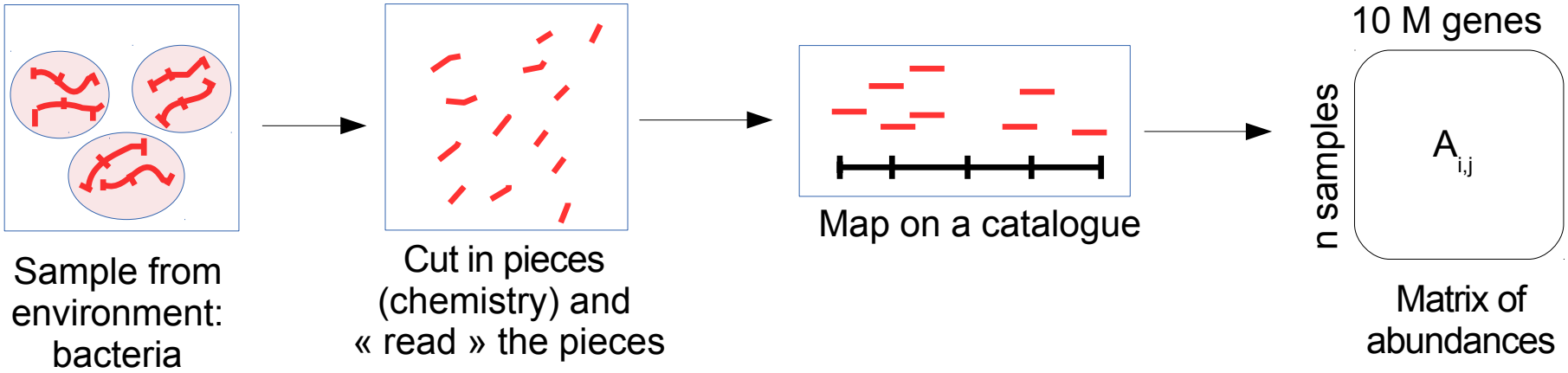
- Genes **from all bacteria** are measured together.
- A **catalogue** is built by assembling pieces of DNA from a large number of samples $\leftrightsquigarrow$ "metagenome" including genes from all bacteria.
  $\hookrightarrow$ catalogue for gut microbiote : 10^7 genes.
- Genes are cut in pieces, whose sequences are read and mapped on the catalogue.



- **Abundance of gene** j = proportion of gene j among all genes present in the sample.
- **Very large dimension** ("$\log p \not\ll n$") : dimension reduction is necessary

Measurement of metagenomic abundances

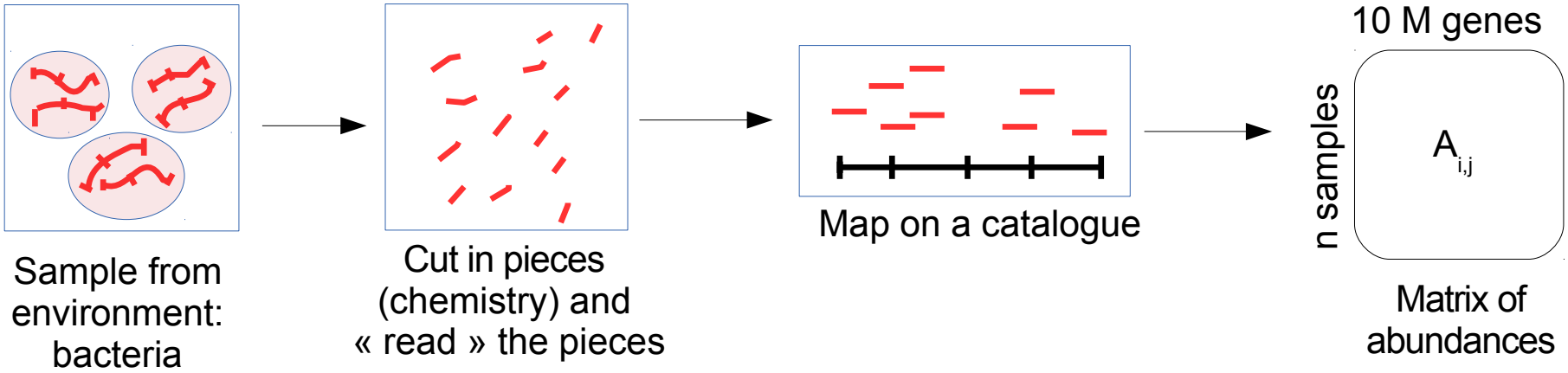
- Genes **from all bacteria** are measured together.
- A **catalogue** is built by assembling pieces of DNA from a large number of samples $\leftrightsquigarrow$ "metagenome" including genes from all bacteria.
  $\hookrightarrow$ catalogue for gut microbiote : 10^7 genes.
- Genes are cut in pieces, whose sequences are read and mapped on the catalogue.



- **Abundance of gene** j = proportion of gene j among all genes present in the sample.
- **Very large dimension** ("$\log p \not\ll n$") : dimension reduction is necessary
  $\hookrightarrow$ Phylogenetic grouping.

Measurement of metagenomic abundances

- Genes **from all bacteria** are measured together.
- A **catalogue** is built by assembling pieces of DNA from a large number of samples $\leftrightsquigarrow$ "metagenome" including genes from all bacteria.
  $\hookrightarrow$ catalogue for gut microbiote : 10^7 genes.
- Genes are cut in pieces, whose sequences are read and mapped on the catalogue.



- **Abundance of gene** j = proportion of gene j among all genes present in the sample.
- **Very large dimension** ("$\log p \not\ll n$") : dimension reduction is necessary
  - $\hookrightarrow$ Phylogenetic grouping.
  - $\hookrightarrow$ Groups of gene with similar functions from biological knowledge / sequence of translated proteins

Metabolism

- **Metabolism** : set of bio-chemical reactions occuring within a living organism.

Metabolism

- **Metabolism** : set of bio-chemical reactions occuring within a living organism.
- **Elementary bio-chemical reaction** : transformation of a chemical A into one or several chemical(s) B (C,D,...)
  - ↪ Each elementary reaction is realised by a protein (enzyme)

Metabolism

- **Metabolism** : set of bio-chemical reactions occuring within a living organism.
- **Elementary bio-chemical reaction** : transformation of a chemical A into one or several chemical(s) B (C,D,...)
  - ↪ Each elementary reaction is realised by a protein (enzyme)
- Metabolism of **fiber digestion** in gut is well known.
  - ↪ Fiber digestion is exclusively realised by bacteria (human genome does not include those genes)

Metabolism

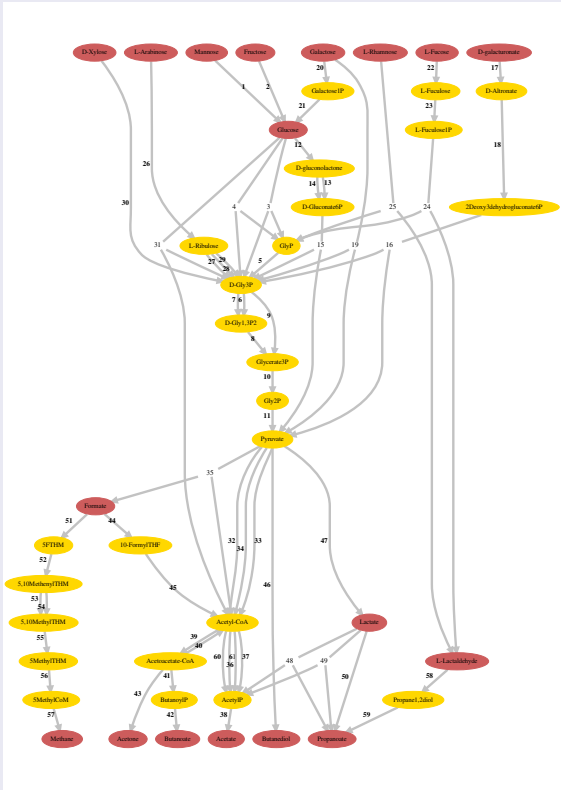
Metabolism

- **Metabolism** : set of bio-chemical reactions occuring within a living organism.
- **Elementary bio-chemical reaction** : transformation of a chemical A into one or several chemical(s) B (C,D,...)
  - ↪ Each elementary reaction is realised by a protein (enzyme)
- Metabolism of **fiber digestion** in gut is well known.
  - ↪ Fiber digestion is exclusively realised by bacteria (human genome does not include those genes)

Fiber digestion metabolism

- 86 elementary reactions extracted from literature
- nodes = chemical

  arrows= reactions



Metabolism and metagenomics

Usually, metabolism is observed by measuring chemicals called "metabolites".
↪ **How do we measure metabolism from metagenomics data ?**

Metabolism and metagenomics

Usually, metabolism is observed by measuring chemicals called "metabolites".
↪ **How do we measure metabolism from metagenomics data ?**

- **KEGG database** (Kyoto Encyclopedia of Genes and Genomes) gathers biological knowledge about functions of bacterial genes.
  ↪ For some elementary reactions, KEGG provides a list of bacterial genes which can realise this reaction.

Metabolism and metagenomics

Usually, metabolism is observed by measuring chemicals called "metabolites".
↪ **How do we measure metabolism from metagenomics data ?**

- **KEGG database** (Kyoto Encyclopedia of Genes and Genomes) gathers biological knowledge about functions of bacterial genes.
  ↪ For some elementary reactions, KEGG provides a list of bacterial genes which can realise this reaction.

  ↪ list of bacterial genes $\Leftrightarrow$ list of columns in the abundance matrix

Metabolism and metagenomics

Usually, metabolism is observed by measuring chemicals called "metabolites".
↪ **How do we measure metabolism from metagenomics data ?**

- **KEGG database** (Kyoto Encyclopedia of Genes and Genomes) gathers biological knowledge about functions of bacterial genes.
  ↪ For some elementary reactions, KEGG provides a list of bacterial genes which can realise this reaction.

  ↪ list of bacterial genes $\Leftrightarrow$ list of columns in the abundance matrix
- Therefore, from a metagene abundance matrix $A = n \times 10^7$, we can compute the **abundance of reaction r in sample i** as :

$$\sum_{\text{genes } g \text{ associated to reaction } r} A_{i,g}$$

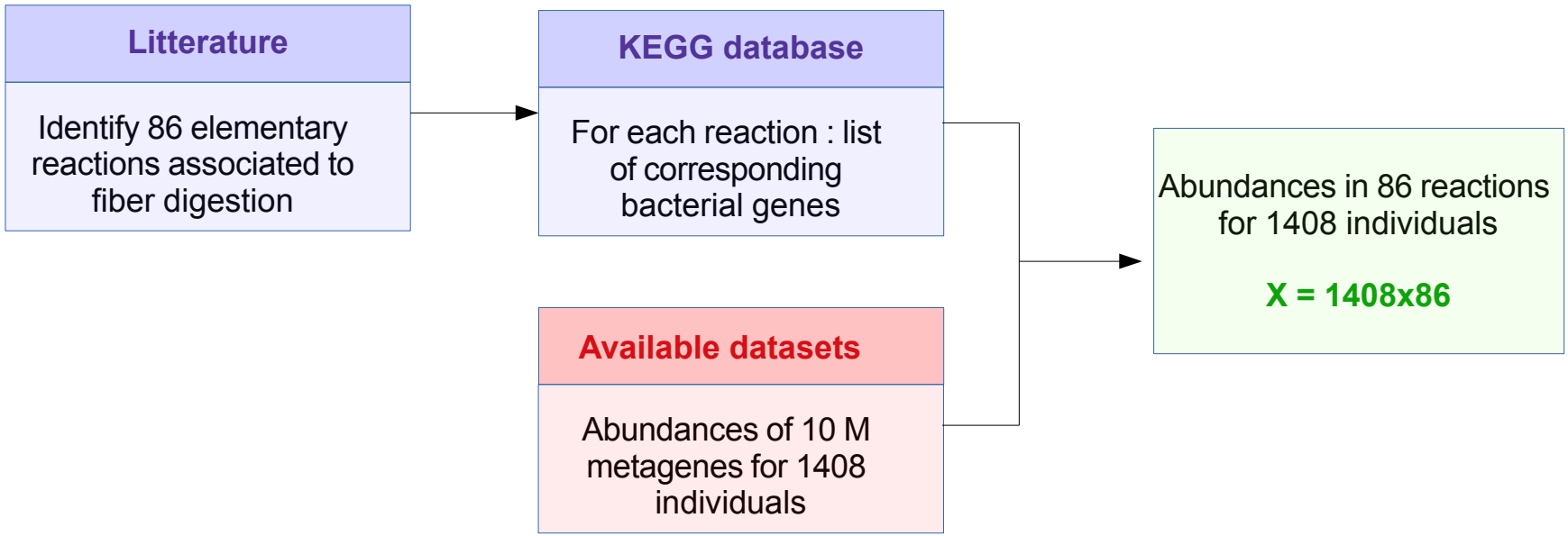
Construction of a data matrix for fiber digestion analysis

Construction of a data matrix for fiber digestion analysis

- **Original data** : abundances of 10^7 metagenes in gut for 1408 individuals (gather several studies)

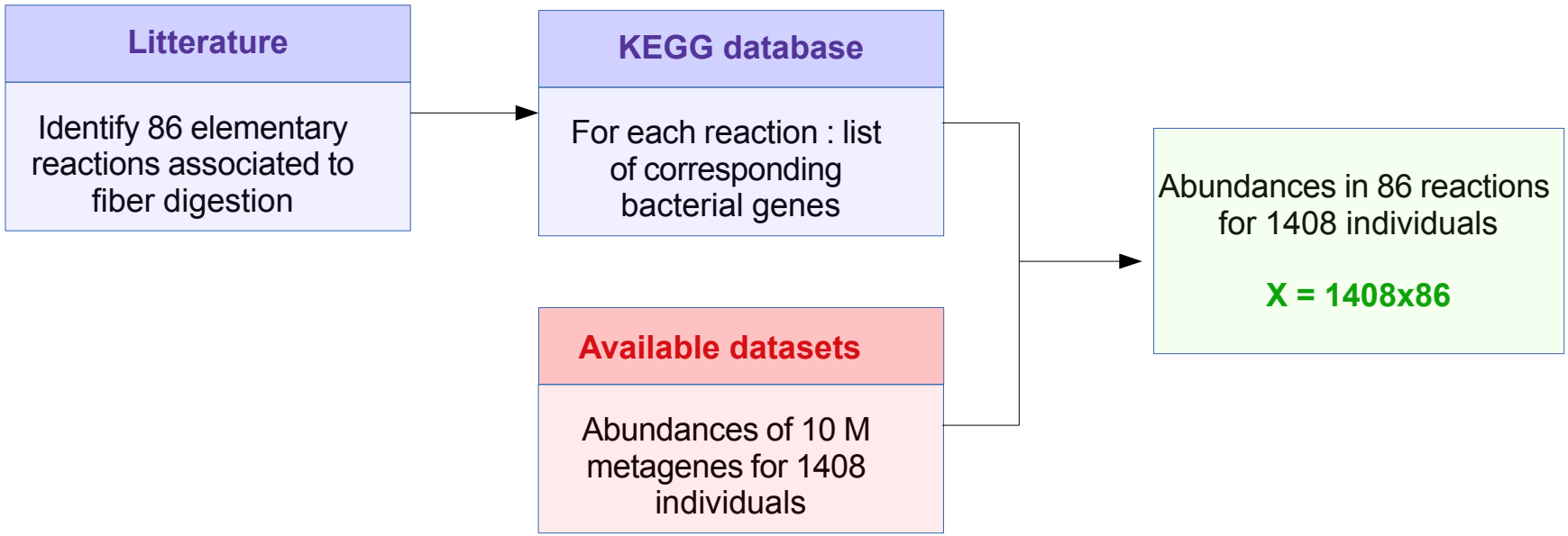
Construction of a data matrix for fiber digestion analysis

- **Original data** : abundances of 10^7 metagenes in gut for 1408 individuals (gather several studies)
- Construction of a matrix with abundances in reactions associated to fiber digestion, using prior biological knowledge



Construction of a data matrix for fiber digestion analysis

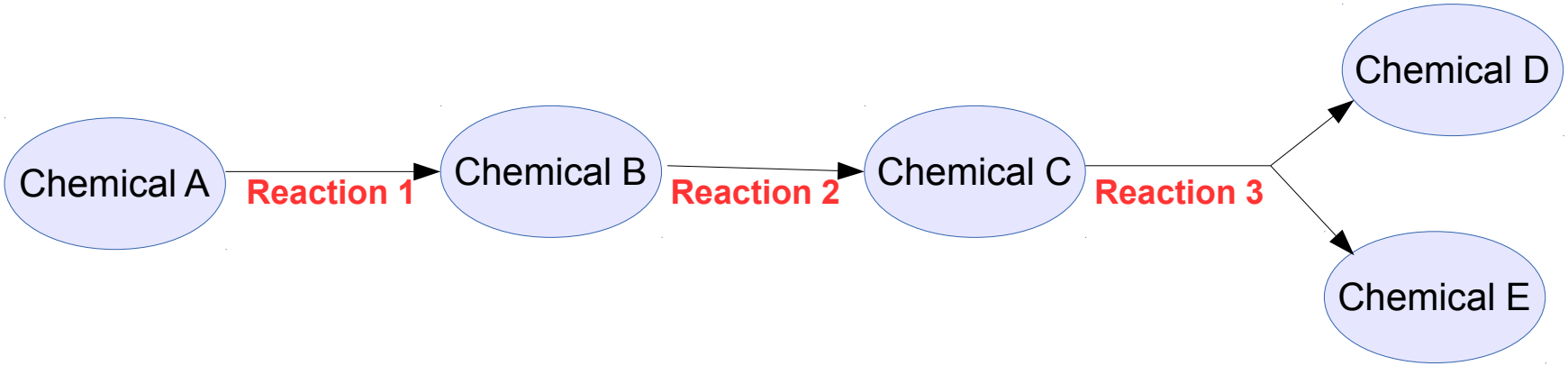
- **Original data** : abundances of 10^7 metagenes in gut for 1408 individuals (gather several studies)
- Construction of a matrix with abundances in reactions associated to fiber digestion, using prior biological knowledge



$\hookrightarrow$ $(X_{i,j})_{i=1,\dots,n,j=1,\dots,p}$: matrix with **abundances of p=86 elementary reactions** associated to fiber digestion, for **n=1408 individuals**.

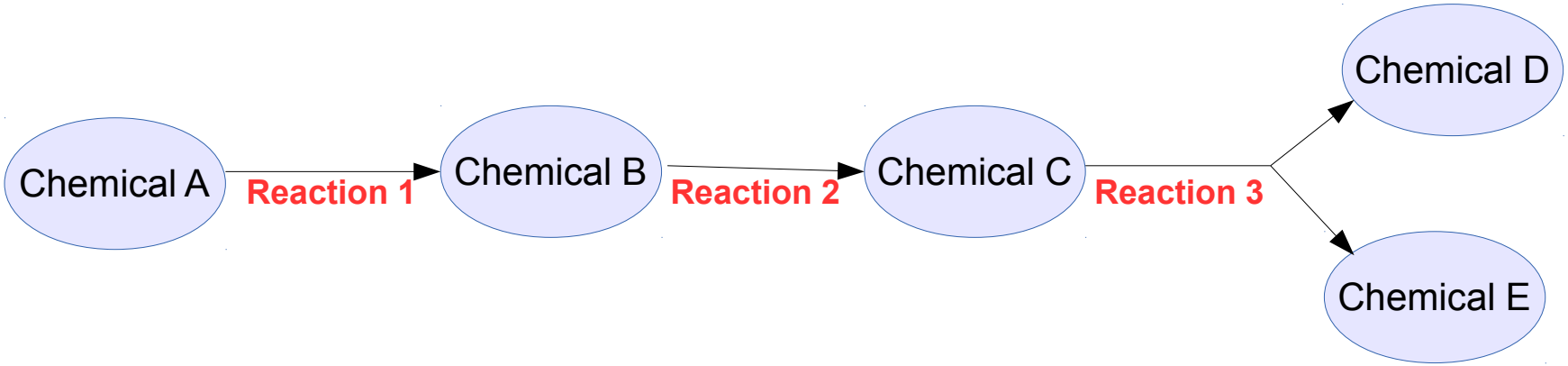
Metabolic pathways

- **Metabolic pathways** are series of bio-chemical reactions occuring together within a cell.



Metabolic pathways

- **Metabolic pathways** are series of bio-chemical reactions occuring together within a cell.

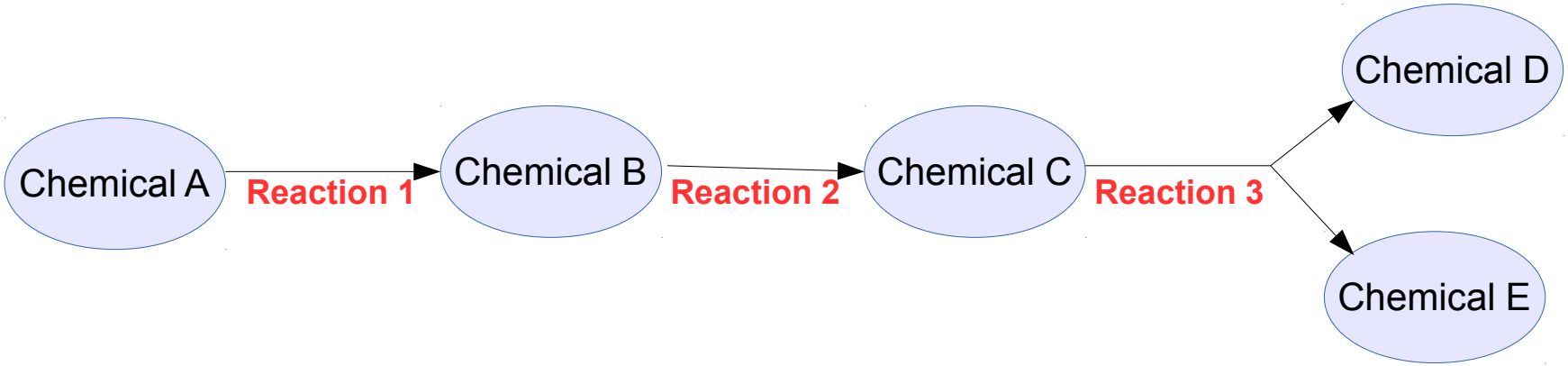


- **Modelling** : a metabolic pathway is characterised by its proportion in elementary reactions

  $\hookrightarrow$ A metabolic pathway associated to fiber digestion is defined by a vector $h \in (\mathbb{R}^+)^{86}$

Metabolic pathways

- **Metabolic pathways** are series of bio-chemical reactions occuring together within a cell.



- **Modelling** : a metabolic pathway is characterised by its proportion in elementary reactions

  $\hookrightarrow$ A metabolic pathway associated to fiber digestion is defined by a vector $h \in (\mathbb{R}^+)^{86}$

- **Our goal** : extract the main metabolic pathways associated to fiber digestion from data matrix X.

Modelling metabolic pathways

We assume that fiber digestion mainly occures through **a small number k of metabolic pathways**

Modelling metabolic pathways

We assume that fiber digestion mainly occures through **a small number k of metabolic pathways**

- **Each metabolic pathway** ℓ is characterised by a given proportion of elementary reactions :

$$H_\ell = (H_{\ell,1}, \ldots, H_{\ell,p}) \in (\mathbb{R}^+)^p$$

Modelling metabolic pathways

We assume that fiber digestion mainly occures through **a small number k of metabolic pathways**

- **Each metabolic pathway** ℓ is characterised by a given proportion of elementary reactions :
$$H_\ell = (H_{\ell,1}, \ldots, H_{\ell,p}) \in (\mathbb{R}^+)^p$$
- **Each individual sample** i includes abundances of pathways $1, \ldots, k$:
$$W_i = (W_{i,1}, \ldots, W_{i,k}) \in (\mathbb{R}^+)^k$$

Modelling metabolic pathways

We assume that fiber digestion mainly occures through **a small number k of metabolic pathways**

- **Each metabolic pathway** ℓ is characterised by a given proportion of elementary reactions :
$$H_\ell = (H_{\ell,1}, \ldots, H_{\ell,p}) \in (\mathbb{R}^+)^p$$
- **Each individual sample** i includes abundances of pathways $1, \ldots, k$:
$$W_i = (W_{i,1}, \ldots, W_{i,k}) \in (\mathbb{R}^+)^k$$
- Then the abundance of reaction j in the sample i is
$$X_{i,j} \approx \sum_{\ell=1}^{k} W_{i,\ell} H_{\ell,j}, \quad \forall i = 1, \ldots, n, \ j = 1, \ldots, p.$$

Modelling metabolic pathways

We assume that fiber digestion mainly occures through **a small number k of metabolic pathways**

- **Each metabolic pathway** ℓ is characterised by a given proportion of elementary reactions :
$$H_\ell = (H_{\ell,1}, \ldots, H_{\ell,p}) \in (\mathbb{R}^+)^p$$
- **Each individual sample** i includes abundances of pathways $1, \ldots, k$:
$$W_i = (W_{i,1}, \ldots, W_{i,k}) \in (\mathbb{R}^+)^k$$
- Then the abundance of reaction j in the sample i is
$$X_{i,j} \approx \sum_{\ell=1}^{k} W_{i,\ell} H_{\ell,j}, \quad \forall i = 1, \ldots, n, \ j = 1, \ldots, p.$$

$\Leftrightarrow$ $\mathbf{X \approx WH}$ with $\begin{cases} W \in (\mathbb{R}^+)^{n*k} & \text{individual profiles in the } k \text{ metabolic pathways} \\ H \in (\mathbb{R}^+)^{k*p} & \text{composition of the metabolic pathways.} \end{cases}$

Modelling metabolic pathways

We assume that fiber digestion mainly occures through **a small number k of metabolic pathways**

- **Each metabolic pathway** ℓ is characterised by a given proportion of elementary reactions :
$$H_\ell = (H_{\ell,1}, \ldots, H_{\ell,p}) \in (\mathbb{R}^+)^p$$
- **Each individual sample** i includes abundances of pathways $1, \ldots, k$:
$$W_i = (W_{i,1}, \ldots, W_{i,k}) \in (\mathbb{R}^+)^k$$
- Then the abundance of reaction j in the sample i is
$$X_{i,j} \approx \sum_{\ell=1}^{k} W_{i,\ell} H_{\ell,j}, \quad \forall i = 1, \ldots, n, \ j = 1, \ldots, p.$$

$\Leftrightarrow$ $\mathbf{X \approx WH}$ with $\begin{cases} W \in (\mathbb{R}^+)^{n*k} & \text{individual profiles in the } k \text{ metabolic pathways} \\ H \in (\mathbb{R}^+)^{k*p} & \text{composition of the metabolic pathways.} \end{cases}$

$\hookrightarrow$ **Nonnegative Matrix Factorization (NMF)**

- **Formulation of the NMF problem.** Let $X \in (\mathbb{R}^+)^{n \times p}$ be a matrix with non-negative coefficients. The NMF decomposition is the solution of the following minimisation problem :

$$(\hat{W}, \hat{H}) = \arg\min_{W,H} \mathsf{D}(X, WH) + pen(W, H) \quad \text{for} \quad \begin{cases} W \in (\mathbb{R}^+)^{n \times k} \\ H \in (\mathbb{R}^+)^{k \times p} \end{cases}$$

with $k \leqslant \min(n, p)$, D a distance on $(\mathbb{R}^+)^{n \times p}$ and pen a penalty function.

- **Formulation of the NMF problem.** Let $X \in (\mathbb{R}^+)^{n \times p}$ be a matrix with non-negative coefficients. The NMF decomposition is the solution of the following minimisation problem :

$$(\hat{W}, \hat{H}) = \arg\min_{W,H} \mathsf{D}(X, WH) + pen(W, H) \quad \text{for} \quad \begin{cases} W \in (\mathbb{R}^+)^{n \times k} \\ H \in (\mathbb{R}^+)^{k \times p} \end{cases}$$

  with $k \leqslant \min(n, p)$, D a distance on $(\mathbb{R}^+)^{n \times p}$ and pen a penalty function.

- **Questions to be addressed :**
  - Distance
  - Number of profiles k : various criteria
  - Choice of a penalty function
  - Numerical computing.

- **Formulation of the NMF problem.** Let $X \in (\mathbb{R}^+)^{n \times p}$ be a matrix with non-negative coefficients. The NMF decomposition is the solution of the following minimisation problem :

$$(\hat{W}, \hat{H}) = \arg\min_{W,H} \mathsf{D}(X, WH) + pen(W, H) \quad \text{for} \quad \begin{cases} W \in (\mathbb{R}^+)^{n \times k} \\ H \in (\mathbb{R}^+)^{k \times p} \end{cases}$$

  with $k \leqslant \min(n, p)$, D a distance on $(\mathbb{R}^+)^{n \times p}$ and pen a penalty function.

- **Questions to be addressed :**
  - Distance
  - Number of profiles k : various criteria
  - Choice of a penalty function
  - Numerical computing.
- **NMF in literature**
  - Originally developed in signal theory for source separation.
  - Then applications to genomics.
  - Methodological developments : mainly focused on algorithms

Choice of a distance : underlying statistical model

In literature, two distances mainly used : Frobenius and Kullback-Leiber.

Choice of a distance : underlying statistical model

In literature, two distances mainly used : Frobenius and Kullback-Leiber.

- **Frobenius distance** : $D_F^2(X, WH) = \sum_{i,j} \left(X_{i,j} - (WH)_{i,j}\right)^2$

  Assume that the $(X_{i,j})_{i=1:n,j=1:p}$ are independent given W, H and

  $$(X_{i,j})|(WH)_{i,j} \sim \mathcal{N}((WH)_{i,j}, \sigma^2)$$

  then $-D_F^2(X, WH)$ is equal to the log-likelihood of the observations X.

Choice of a distance : underlying statistical model

In literature, two distances mainly used : Frobenius and Kullback-Leiber.

- **Frobenius distance** : $D_F^2(X, WH) = \sum_{i,j} \left(X_{i,j} - (WH)_{i,j}\right)^2$

  Assume that the $(X_{i,j})_{i=1:n,j=1:p}$ are independent given W, H and

  $$(X_{i,j})|(WH)_{i,j} \sim \mathcal{N}((WH)_{i,j}, \sigma^2)$$

  then $-D_F^2(X, WH)$ is equal to the log-likelihood of the observations X.

- **Generalized Kullback-Leiber distance**

  $$D_{KL}(X|WH) = \sum_{i,j} \left(X_{i,j} \log \frac{X_{i,j}}{(WH)_{i,j}} - X_{i,j} + (WH)_{i,j} \right)$$

  Assume that the $(X_{i,j})_{i=1:n,j=1:p}$ are independent given W, H and

  $$X_{i,j}|(WH)_{i,j} \sim \mathcal{P}\left((WH)_{i,j}\right)$$

  then $-D_{KL}(X, WH)$ is equal to the log-likelihood of the observations X.

Choice of a distance (2)

Choice of a distance (2)

- No methodology proposed to choose the distance (analysis of residuals, etc)

$\hookrightarrow$ Choice based on a priori modelling and implementation practicality

  - Frobenius : easy computings, classic in regression and modelling, but homoscedastic measurement error : not very realistic.
  - KL : problem when $(WH)_{i,j} = 0$ (problem in iterative computing) but more realistic variance modelling

Choice of a distance (2)

- No methodology proposed to choose the distance (analysis of residuals, etc)

↪ Choice based on a priori modelling and implementation practicality

  - Frobenius : easy computings, classic in regression and modelling, but homoscedastic measurement error : not very realistic.
  - KL : problem when $(WH)_{i,j} = 0$ (problem in iterative computing) but more realistic variance modelling

- We have chosen Frobenius.

  ↪ Project : develop algorithms for other distances (zero-inflated KL, negative binomial)

Choice of a distance (2)

- No methodology proposed to choose the distance (analysis of residuals, etc)

↪ Choice based on a priori modelling and implementation practicality

  - Frobenius : easy computings, classic in regression and modelling, but homoscedastic measurement error : not very realistic.
  - KL : problem when $(WH)_{i,j} = 0$ (problem in iterative computing) but more realistic variance modelling

- We have chosen Frobenius.

  ↪ Project : develop algorithms for other distances (zero-inflated KL, negative binomial)

- Rq : KL and Frobenius present computing facilities that can not be generalised.

Numerical computing

- If D is convex, then $(W, H) \to D(X, WH)$ is biconvex, but not convex : **alternate minimisation**

Numerical computing

- If D is convex, then $(W, H) \to D(X, WH)$ is biconvex, but not convex : **alternate minimisation**
  - Initialisation : (W_0, H_0)
  - Update
$$\begin{cases} W_j \leftarrow \arg\min_{W \geqslant 0} \mathsf{D}(X, WH_{j-1}) + pen(W, H_{j-1}) \\ H_j \leftarrow \arg\min_{H \geqslant 0} \mathsf{D}(X, W_j H) + pen(W_j, H) \end{cases}$$
  - Stopping criterion : $W_j H_j \approx W_{j-1} H_{j-1}$.

Numerical computing

- If D is convex, then $(W, H) \to D(X, WH)$ is biconvex, but not convex : **alternate minimisation**
  - Initialisation : (W_0, H_0)
  - Update
$$\begin{cases} W_j \leftarrow \arg\min_{W \geqslant 0} \mathsf{D}(X, WH_{j-1}) + pen(W, H_{j-1}) \\ H_j \leftarrow \arg\min_{H \geqslant 0} \mathsf{D}(X, W_j H) + pen(W_j, H) \end{cases}$$
  - Stopping criterion : $W_j H_j \approx W_{j-1} H_{j-1}$.
- In practise, minimisation is replaced by **decreasing of the criterion** (first step of a convex optimisation algo).

$$\begin{cases} D(X, W_j H_{j-1}) + pen(W_j, H_{j-1}) \leqslant \mathsf{D}(X, W_{j-1} H_{j-1}) + pen(W_{j-1}, H_{j-1}) \\ D(X, W_j H_j) + pen(W_j, H_j) \leqslant \mathsf{D}(X, W_j H_{j-1}) + pen(W_j, H_{j-1}) \end{cases}$$

Numerical computing

- If D is convex, then $(W, H) \to D(X, WH)$ is biconvex, but not convex : **alternate minimisation**
  - Initialisation : (W_0, H_0)
  - Update
$$\begin{cases} W_j \leftarrow \arg\min_{W \geqslant 0} \mathsf{D}(X, WH_{j-1}) + pen(W, H_{j-1}) \\ H_j \leftarrow \arg\min_{H \geqslant 0} \mathsf{D}(X, W_j H) + pen(W_j, H) \end{cases}$$
  - Stopping criterion : $W_j H_j \approx W_{j-1} H_{j-1}$.
- In practise, minimisation is replaced by **decreasing of the criterion** (first step of a convex optimisation algo).

$$\begin{cases} D(X, W_j H_{j-1}) + pen(W_j, H_{j-1}) \leqslant \mathsf{D}(X, W_{j-1} H_{j-1}) + pen(W_{j-1}, H_{j-1}) \\ D(X, W_j H_j) + pen(W_j, H_j) \leqslant \mathsf{D}(X, W_j H_{j-1}) + pen(W_j, H_{j-1}) \end{cases}$$

$\hookrightarrow$ For Frobenius and KL distances, we obtain additive and multiplicative update rules.

  - No theroretical guarantee of convergence.
  - No generalisation to other distance or additional constraints.

Numerical computing

- If D is convex, then $(W, H) \to D(X, WH)$ is biconvex, but not convex : **alternate minimisation**
  - Initialisation : (W_0, H_0)
  - Update
$$\begin{cases} W_j \leftarrow \arg\min_{W \geqslant 0} \mathsf{D}(X, WH_{j-1}) + pen(W, H_{j-1}) \\ H_j \leftarrow \arg\min_{H \geqslant 0} \mathsf{D}(X, W_j H) + pen(W_j, H) \end{cases}$$
  - Stopping criterion : $W_j H_j \approx W_{j-1} H_{j-1}$.
- In practise, minimisation is replaced by **decreasing of the criterion** (first step of a convex optimisation algo).

$$\begin{cases} D(X, W_j H_{j-1}) + pen(W_j, H_{j-1}) \leqslant \mathsf{D}(X, W_{j-1} H_{j-1}) + pen(W_{j-1}, H_{j-1}) \\ D(X, W_j H_j) + pen(W_j, H_j) \leqslant \mathsf{D}(X, W_j H_{j-1}) + pen(W_j, H_{j-1}) \end{cases}$$

$\hookrightarrow$ For Frobenius and KL distances, we obtain additive and multiplicative update rules.

  - No theroretical guarantee of convergence.
  - No generalisation to other distance or additional constraints.
- We implemented alternate minimisation using general algorithm for constrainted convex optimisation.

Choice of the reduced dimension k

- k has to be chosen **from data**

Choice of the reduced dimension k

- k has to be chosen **from data**
- Contrary to PCA :
  - NMF pathways are not automatically ranked and weighted by importance.
  - The metabolic pathways for a given k are not included in pathways for $k + 1$.

Choice of the reduced dimension k

- k has to be chosen **from data**
- Contrary to PCA :
  - NMF pathways are not automatically ranked and weighted by importance.
  - The metabolic pathways for a given k are not included in pathways for $k + 1$.
- **Mostly used criteria** : stability of W over various initialisations.

Choice of the reduced dimension k

- k has to be chosen **from data**
- Contrary to PCA :
  - NMF pathways are not automatically ranked and weighted by importance.
  - The metabolic pathways for a given k are not included in pathways for $k+1$.
- **Mostly used criteria** : stability of W over various initialisations.
  - Consensus in clustering of samples : each sample i is assigned to the profile $j = \arg\max(W_{i,1}, \ldots, W_{i,p})$, and the number of pathways k for which clustering is best preserved along repetitions is selected.

Choice of the reduced dimension k

- k has to be chosen **from data**
- Contrary to PCA :
  - NMF pathways are not automatically ranked and weighted by importance.
  - The metabolic pathways for a given k are not included in pathways for $k+1$.
- **Mostly used criteria** : stability of W over various initialisations.
  - Consensus in clustering of samples : each sample i is assigned to the profile $j = \arg\max(W_{i,1}, \ldots, W_{i,p})$, and the number of pathways k for which clustering is best preserved along repetitions is selected.
  - Concordance of W : for a individual profile matrix W, let S be the two-by-two correlation matrix between individual profiles W_i and $W_{i'}$: S is called the **similarity** matrix associted to W.
    Then the concordance index between two initialisations r and r' is defined as

$$1 - \sqrt{\sum_{i \neq j} \left(S_{i,j}^{(r)} - S_{i,j}^{(r')} \right)^2} \quad (1)$$

    and k which maximises concordance between initialisations is selected.

Choice of the reduced dimension k (2)

In alternative, we consider more heuristic criteria.

Choice of the reduced dimension k (2)

In alternative, we consider more heuristic criteria.

- **Reconstruction error** $D(X, WH)$
  - Automatically decreases as k increases
  - Stop when adding a new profile do not improve significantly the reconstruction error.

Choice of the reduced dimension k (2)

In alternative, we consider more heuristic criteria.

- **Reconstruction error** $D(X, WH)$
  - Automatically decreases as k increases
  - Stop when adding a new profile do not improve significantly the reconstruction error.
- **Concordance of** H when randomly spliting samples $\{1, \ldots, n\} = I_1 \sqcup I_2$: for each $k = 2, \ldots k_0$
  - Compute NMF with $X[I_1,]$ and $X[I_2,]$: (W_1, H_1), (W_2, H_2).
  - Compute concordance index (1) between H_1 and H_2.
  - Stop when the concordance gets poor.

Choice of the reduced dimension k (2)

In alternative, we consider more heuristic criteria.

- **Reconstruction error** $D(X, WH)$
  - Automatically decreases as k increases
  - Stop when adding a new profile do not improve significantly the reconstruction error.
- **Concordance of** H when randomly spliting samples $\{1, \ldots, n\} = I_1 \sqcup I_2$: for each $k = 2, \ldots k_0$
  - Compute NMF with $X[I_1,]$ and $X[I_2,]$: (W_1, H_1), (W_2, H_2).
  - Compute concordance index (1) between H_1 and H_2.
  - Stop when the concordance gets poor.
- **Bi-cross-validation** : version of CV adapted to NMF.

Recall : cross-validation in regression.

Let $X \in \mathbb{R}^{n*p}$, $Y \in \mathbb{R}^n$, and $\{\hat{f}_\lambda, \lambda \in \Lambda\}$ a set of estimating procedures of the regression function $r(x) = \mathbb{E}[Y|X = x]$.

Let $I_1 \sqcup I_2 \sqcup \cdots \sqcup I_N = \{1, \ldots, n\}$. For $\ell = 1, \ldots, N$,

- Compute $\{\hat{f}_\lambda^{(\ell)}, \lambda \in \Lambda\}$ from $X[-I_\ell,]$ and $Y[-I_\ell]$.
- For $i \in I_\ell$, $\hat{Y}_i^\lambda = \hat{f}_\lambda^{(\ell)}(X_i)$.
- CV-error : $E^{\ell,\lambda} = \sum_{i=1,\ldots,n}(Y_i - Y_i^\lambda)^2$

Then, we choose the λ which minimises $\sum_{\ell=1}^N E^{\ell,\lambda}$.

Recall : cross-validation in regression.

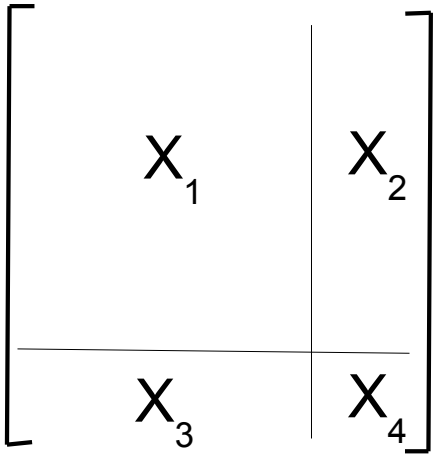
Let $X \in \mathbb{R}^{n*p}$, $Y \in \mathbb{R}^n$, and $\{\hat{f}_\lambda, \lambda \in \Lambda\}$ a set of estimating procedures of the regression function $r(x) = \mathbb{E}[Y|X = x]$.

Let $I_1 \sqcup I_2 \sqcup \cdots \sqcup I_N = \{1, \ldots, n\}$. For $\ell = 1, \ldots, N$,

- Compute $\{\hat{f}_\lambda^{(\ell)}, \lambda \in \Lambda\}$ from $X[-I_\ell,]$ and $Y[-I_\ell]$.
- For $i \in I_\ell$, $\hat{Y}_i^\lambda = \hat{f}_\lambda^{(\ell)}(X_i)$.
- CV-error : $E^{\ell,\lambda} = \sum_{i=1,\ldots,n}(Y_i - Y_i^\lambda)^2$

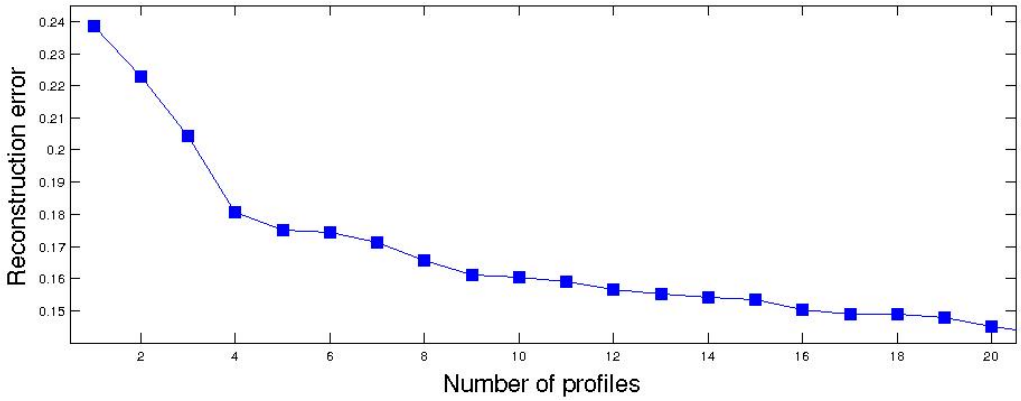
Then, we choose the λ which minimises $\sum_{\ell=1}^N E^{\ell,\lambda}$.

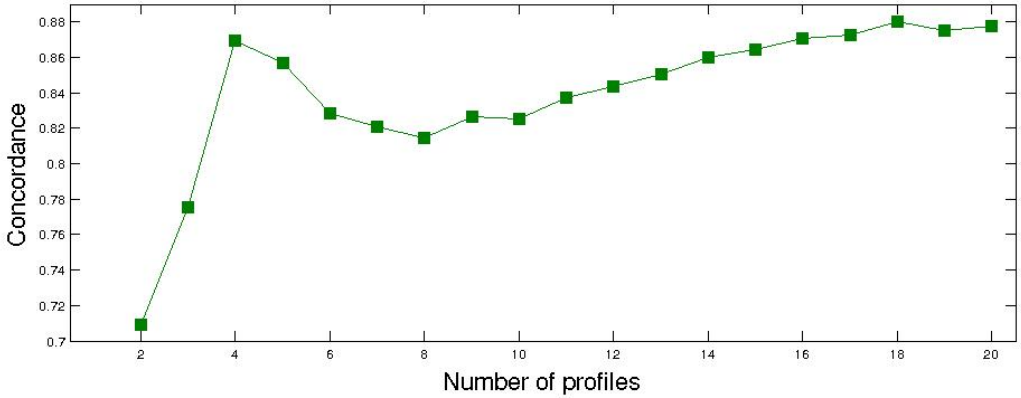
Bi-cross-validation in NMF

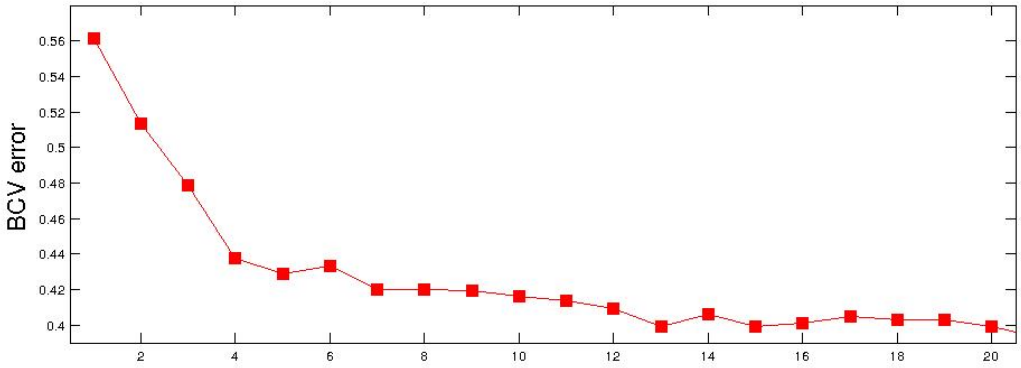


- $(W^{train}, H^{train}) = \arg\min D(X_1, WH) + pen(W, H)$
- $H^{val} = \arg\min D(X_2, W^{train}H) + pen(W^{train}, H)$
- $W^{val} = \arg\min D(X_3, WH^{train}) + pen(W, H^{train})$
- Bi-CV error : $D(X, W^{val}H^{val})$.

Results on our data







Choice of penalty

- For alternate estimation of W and H : $pen(W, H) = pen_W(W) + pen_H(H)$

Choice of penalty

- For alternate estimation of W and H : $pen(W, H) = pen_W(W) + pen_H(H)$
- Similarly to regression :
  - L^1-penalty favor sparsity (possible lost of significant variables)
  - L^2-penalty usually offer a better reconstruction but no sparsity (less easy to interprete)
  - Elasticnet : trade-off
  - Group-lasso, etc : favor biological model of pathway/individual profiles

A few word about unicity of NMF decomposition

- Let M be an invertible matrix of dimension $k \times k$ such that $M^{-1} \geqslant 0$ and $M \geqslant 0$. Then,

$$WH = (WM^{-1})(MH)$$

A few word about unicity of NMF decomposition

- Let M be an invertible matrix of dimension $k \times k$ such that $M^{-1} \geqslant 0$ and $M \geqslant 0$. Then,

$$WH = (WM^{-1})(MH)$$

- Trivial examples : M diagonal, M permutation matrix.

  $\hookrightarrow$ the NMF is not unique.

A few word about unicity of NMF decomposition

- Let M be an invertible matrix of dimension $k \times k$ such that $M^{-1} \geqslant 0$ and $M \geqslant 0$. Then,

$$WH = (WM^{-1})(MH)$$

- Trivial examples : M diagonal, M permutation matrix.

  $\hookrightarrow$ the NMF is not unique.

- H. Laurberg (PhD 2008) : geometric conditions on X such that the NMF solution is unique up to rescaling and permutations
  $\hookrightarrow$ not easily interpretable.

A few word about unicity of NMF decomposition

- Let M be an invertible matrix of dimension $k \times k$ such that $M^{-1} \geqslant 0$ and $M \geqslant 0$. Then,

$$WH = (WM^{-1})(MH)$$

- Trivial examples : M diagonal, M permutation matrix.

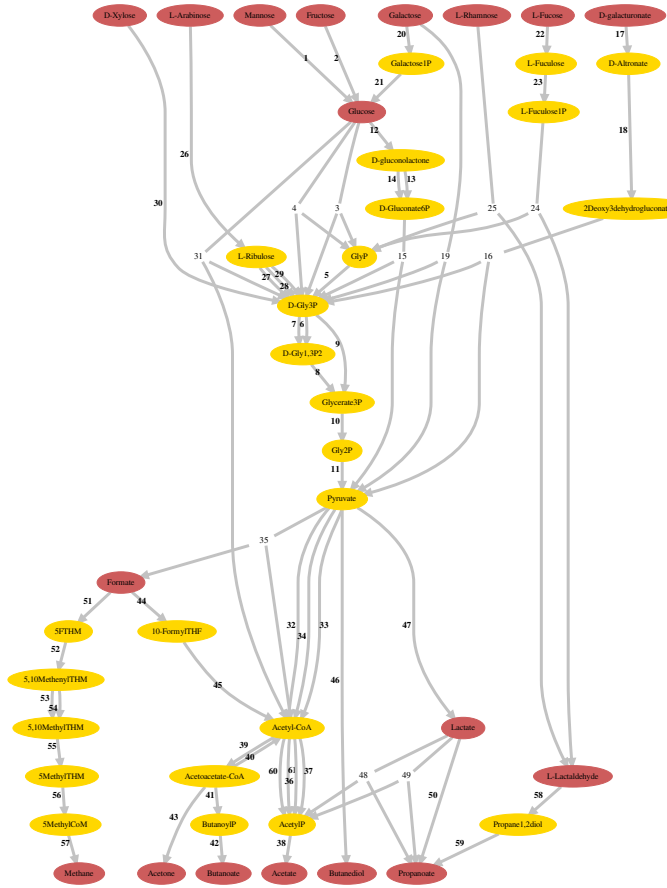
  $\hookrightarrow$ the NMF is not unique.

- H. Laurberg (PhD 2008) : geometric conditions on X such that the NMF solution is unique up to rescaling and permutations
  $\hookrightarrow$ not easily interpretable.

- In practise, we observe almost identical results up to scaling and permutations over repeated initialisations of the algorithm.

Inclusion of a priori knowledge

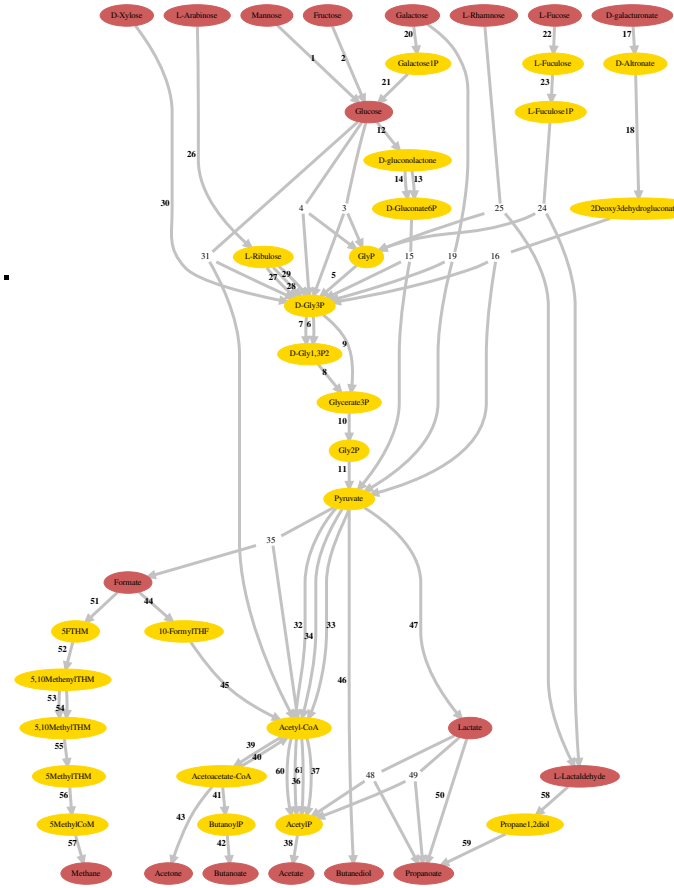
- **Graph of reactions involved in fiber digestion**
  - Arrow = elementary reaction
  - Node = chemical
  - Intra/extra cellular chemicals



Inclusion of a priori knowledge

- **Graph of reactions involved in fiber digestion**
  - Arrow = elementary reaction
  - Node = chemical
  - Intra/extra cellular chemicals

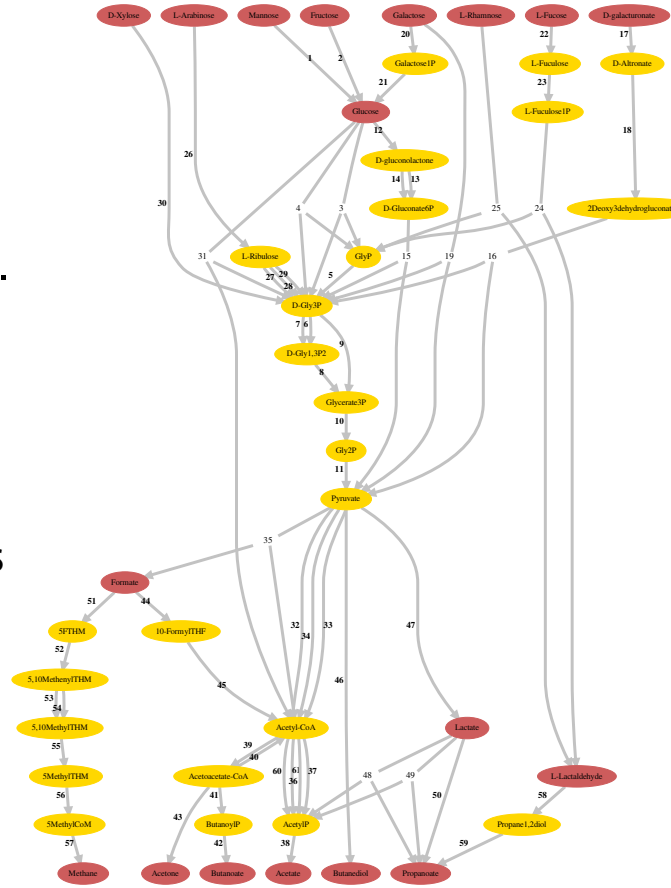
$\hookrightarrow$ Metabolic pathways in H are weighted subgraphs.



Inclusion of a priori knowledge

- **Graph of reactions involved in fiber digestion**
  - Arrow = elementary reaction
  - Node = chemical
  - Intra/extra cellular chemicals

  $\hookrightarrow$ Metabolic pathways in H are weighted subgraphs.
- We want to constraint metabolic pathways to be unions of subgraphs which are :
  - **Starting and ending** with extra-cellular chemicals.
  - **Connex in the graph** : if an intra-cellular chemical is created in a pathway, then it has to be degradated.



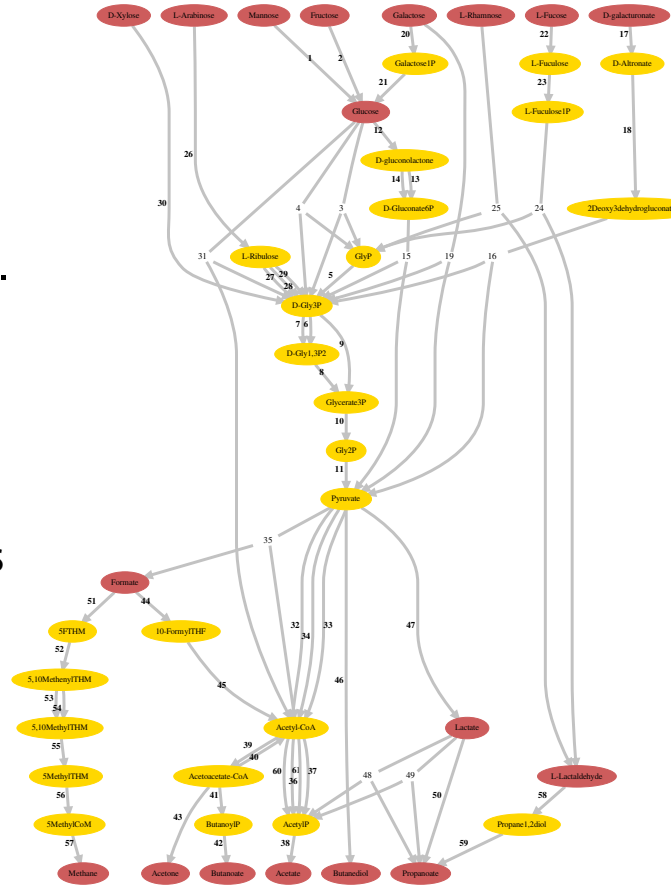
Inclusion of a priori knowledge

- **Graph of reactions involved in fiber digestion**
  - Arrow = elementary reaction
  - Node = chemical
  - Intra/extra cellular chemicals

  $\hookrightarrow$ Metabolic pathways in H are weighted subgraphs.
- We want to constraint metabolic pathways to be unions of subgraphs which are :
  - **Starting and ending** with extra-cellular chemicals.
  - **Connex in the graph** : if an intra-cellular chemical is created in a pathway, then it has to be degradated.
- We show that it's equivalent to a linear constraint :

$$\frac{1}{\delta}\mathbf{Q}^{-}\mathbf{H}^{\mathbf{t}} \leqslant \mathbf{Q}^{+}\mathbf{H}^{\mathbf{t}} \leqslant \delta\mathbf{Q}^{-}\mathbf{H}^{\mathbf{t}} \quad \text{with} \quad \delta > 1$$

with Q^+ and Q^- defined from the graph.



More precisely :

- Let Q be the adjacence matrix of the graph without extra-cellular chemical : for each reaction j and chemical c

$$Q_{c,j} = \begin{cases} 1 \text{ if reaction j reaches metabolite c} \\ -1 \text{ if reaction j originates on metabolite c} \\ 0 \text{ otherwise} \end{cases}$$

- Let $Q = Q^+ - Q^-$ the decomposition of Q in positive/negative part.
- For a metabolic pathway ℓ,
  $(Q^+H)_{\ell,c}$ = sum of proportion in reactions which reach metabolite c
  $(Q^-H)_{\ell,c}$ = sum of proportion in reactions which originate from metabolite c
- Thus, the proportion of reactions which create and degradate metabolite c have same order if :

$$\frac{1}{\delta}(Q^-H)_{\ell,c} \leqslant (Q^+H)_{\ell,c} \leqslant \delta(Q^-H)_{\ell,c} \quad \text{with} \quad \delta > 1$$

NMF procedure

- X is the matrix of abundances in 86 reactions associated to fiber digestion, for 1408 individuals.

NMF procedure

- X is the matrix of abundances in 86 reactions associated to fiber digestion, for 1408 individuals.
- Criterion minimization

$$(W, H) = \arg \min_{W \geqslant 0, H \geqslant 0} \{\|X - WH\|_F^2 + \frac{\alpha}{k}(\|W\|_F^2 + \|1^t H\|_2^2) \quad \text{such that}$$

$$\frac{1}{\delta} Q^- H^t \leqslant Q^+ H^t \leqslant \delta Q^- H^t \quad \text{with} \quad \delta = 5$$

NMF procedure

- X is the matrix of abundances in 86 reactions associated to fiber digestion, for 1408 individuals.
- Criterion minimization

$$(W, H) = \arg \min_{W \geqslant 0, H \geqslant 0} \{\|X - WH\|_F^2 + \frac{\alpha}{k}(\|W\|_F^2 + \|1^t H\|_2^2) \quad \text{such that}$$

$$\frac{1}{\delta} Q^- H^t \leqslant Q^+ H^t \leqslant \delta Q^- H^t \quad \text{with} \quad \delta = 5$$

- $k = 4$ is chosen based on the 3 criteria presented above (reconstruction error, concordance of H and bi-CV).

NMF procedure

- X is the matrix of abundances in 86 reactions associated to fiber digestion, for 1408 individuals.
- Criterion minimization

$$(W, H) = \arg \min_{W \geqslant 0, H \geqslant 0} \{\|X - WH\|_F^2 + \frac{\alpha}{k}(\|W\|_F^2 + \|1^t H\|_2^2) \quad \text{such that}$$

$$\frac{1}{\delta} Q^- H^t \leqslant Q^+ H^t \leqslant \delta Q^- H^t \quad \text{with} \quad \delta = 5$$

- $k = 4$ is chosen based on the 3 criteria presented above (reconstruction error, concordance of H and bi-CV).
- α is chosen by bi-cross-validation.

Analysis of the NMF results

- Our first goal is to prove that our whole method :
  - Constitution of the reaction abundance matrix using KEGG
  - Exploration of metabolic pathways through NMF decomposition

  may be of interest for biological interpretation.

Analysis of the NMF results

- Our first goal is to prove that our whole method :
  - Constitution of the reaction abundance matrix using KEGG
  - Exploration of metabolic pathways through NMF decomposition

  may be of interest for biological interpretation.
- Thus, we are interested in
  - what type of information can be recovered from this analysis
  - existing biological knowledge (not included as a priori) actually recovered.

Analysis of the NMF results

- Our first goal is to prove that our whole method :
  - Constitution of the reaction abundance matrix using KEGG
  - Exploration of metabolic pathways through NMF decomposition

  may be of interest for biological interpretation.
- Thus, we are interested in
  - what type of information can be recovered from this analysis
  - existing biological knowledge (not included as a priori) actually recovered.
- $X \approx WH$ can be seen as a dimension reduction method preserving interpretability,
  - The rows of W gives the individual profiles in the reduced space.
  - The columns of H gives the composition of the metabolic pathways.

Analysis of the NMF results

- Our first goal is to prove that our whole method :
  - Constitution of the reaction abundance matrix using KEGG
  - Exploration of metabolic pathways through NMF decomposition

  may be of interest for biological interpretation.
- Thus, we are interested in
  - what type of information can be recovered from this analysis
  - existing biological knowledge (not included as a priori) actually recovered.
- $X \approx WH$ can be seen as a dimension reduction method preserving interpretability,
  - The rows of W gives the individual profiles in the reduced space.
  - The columns of H gives the composition of the metabolic pathways.
- Our results are mainly descriptive.

Contribution of pathway to the total signal

Contribution of pathway to the total signal

- Contrary to PCA or ICA, the part of explained variance of the metabolic pathway is not relevant since pathways are not orthogonal/independent.

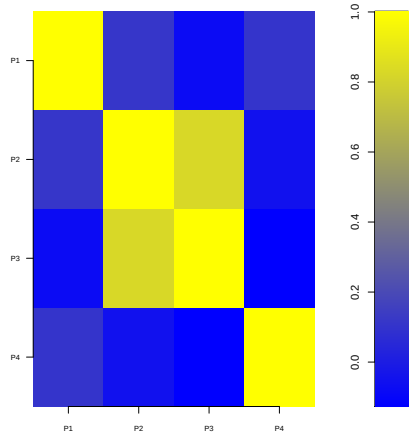
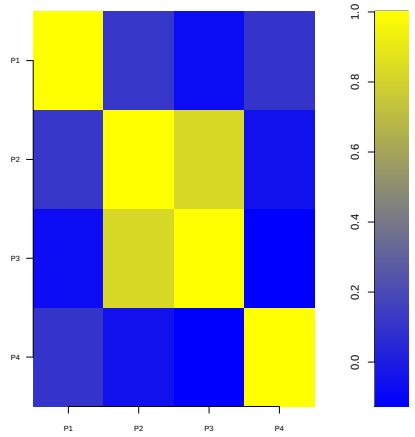


Fig : Pearson correlation between the 4 pathways $\{H[\ell,], \ell = 1, \ldots, 4\}$

Contribution of pathway to the total signal

- Contrary to PCA or ICA, the part of explained variance of the metabolic pathway is not relevant since pathways are not orthogonal/independent.

  Fig : Pearson correlation between the 4 pathways $\{H[\ell,], \ell = 1, \ldots, 4\}$



- Alternative : **contribution of each metabolic pathway to the total signal** :

$$\text{"Total signal"} = \sum_{i=1}^{n}\sum_{j=1}^{86} X_{i,j} \approx \sum_{i=1}^{n}\sum_{j=1}^{86}\sum_{l=1}^{k} W_{i,l}H_{l,j} = \sum_{l=1}^{k}\left(\sum_{i=1}^{n}\sum_{j=1}^{86} W_{i,l}H_{l,j}\right)$$

Contribution of pathway to the total signal

- Contrary to PCA or ICA, the part of explained variance of the metabolic pathway is not relevant since pathways are not orthogonal/independent.

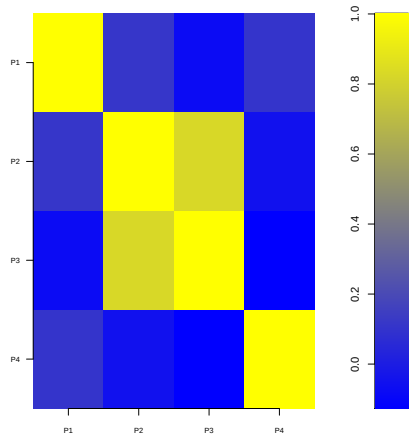


Fig : Pearson correlation between the 4 pathways $\{H[\ell,], \ell = 1, \ldots, 4\}$

- Alternative : **contribution of each metabolic pathway to the total signal** :

$$\text{"Total signal"} = \sum_{i=1}^{n}\sum_{j=1}^{86} X_{i,j} \approx \sum_{i=1}^{n}\sum_{j=1}^{86}\sum_{l=1}^{k} W_{i,l}H_{l,j} = \sum_{l=1}^{k}\left(\sum_{i=1}^{n}\sum_{j=1}^{86} W_{i,l}H_{l,j}\right)$$

We define the contribution of pathway ℓ as :

$$c_\ell = \frac{\sum_{i=1}^{n}\sum_{j=1}^{86} W_{i,l}H_{l,j}}{\sum_{i=1}^{n}\sum_{j=1}^{86}(WH)_{i,j}}$$

Contribution of pathway to the total signal

- Contrary to PCA or ICA, the part of explained variance of the metabolic pathway is not relevant since pathways are not orthogonal/independent.

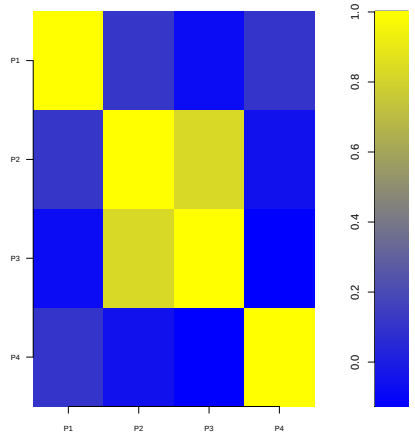


Fig : Pearson correlation between the 4 pathways $\{H[\ell,], \ell = 1, \ldots, 4\}$

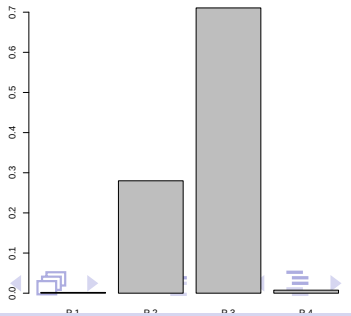
- Alternative : **contribution of each metabolic pathway to the total signal** :

$$\text{"Total signal"} = \sum_{i=1}^{n}\sum_{j=1}^{86} X_{i,j} \approx \sum_{i=1}^{n}\sum_{j=1}^{86}\sum_{l=1}^{k} W_{i,l}H_{l,j} = \sum_{l=1}^{k}\left(\sum_{i=1}^{n}\sum_{j=1}^{86} W_{i,l}H_{l,j}\right)$$

We define the contribution of pathway ℓ as :

$$c_\ell = \frac{\sum_{i=1}^{n}\sum_{j=1}^{86} W_{i,l}H_{l,j}}{\sum_{i=1}^{n}\sum_{j=1}^{86}(WH)_{i,j}}$$

Contribution of the 4 pathways

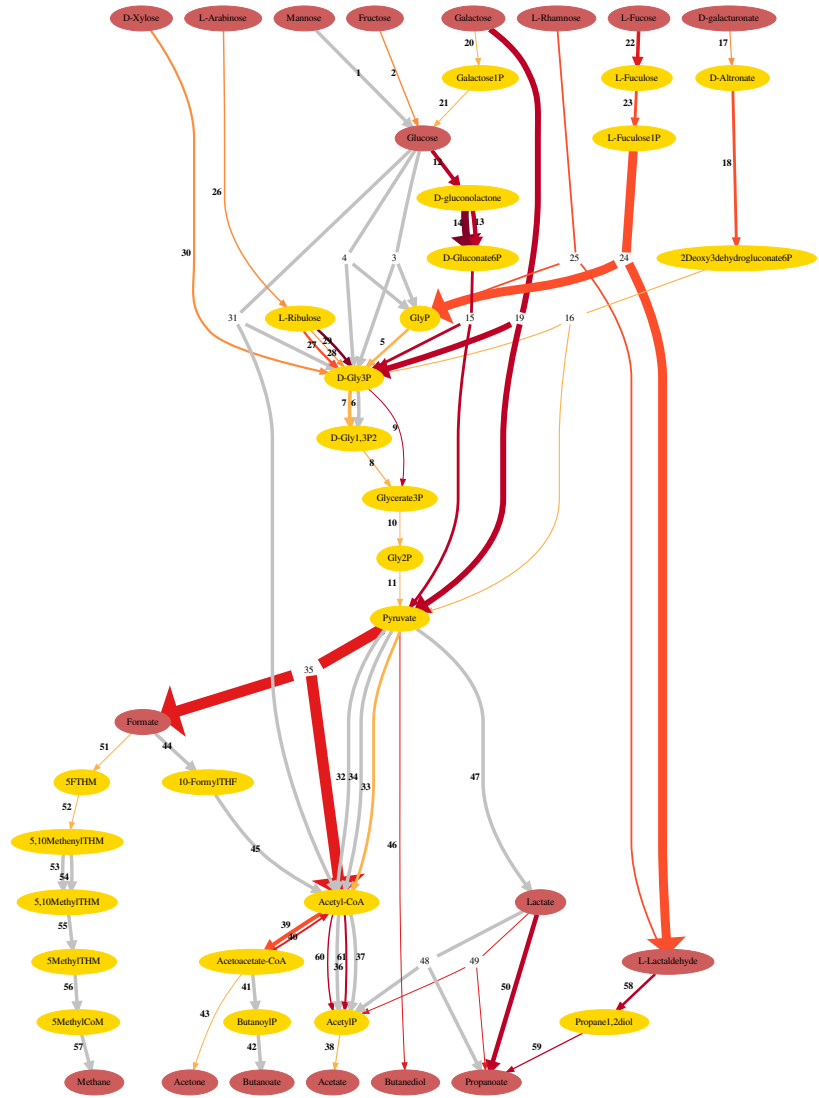


Analysis of H

Pathways in H are represented as weighted subgraphs

$\hookrightarrow$ Biological interpretation .

$\hookrightarrow$ Pathways with a strong total signal include more reactions.



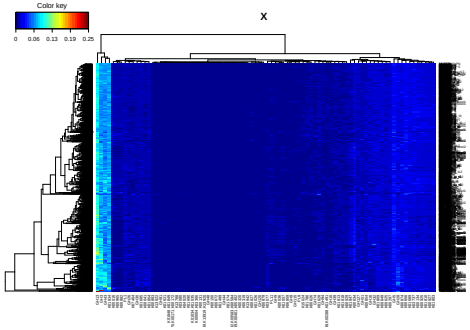
Analysis of the individual profiles in W (1)

Is there an underlying clustering structure in matrix W ?

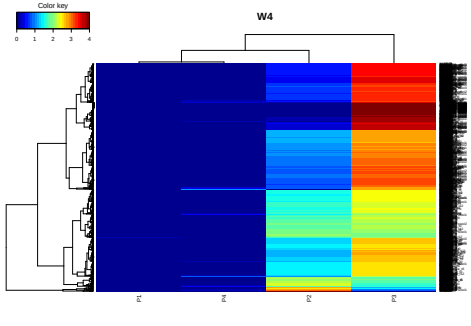
Analysis of the individual profiles in W (1)

Is there an underlying clustering structure in matrix W ?

- Visual examination : **hierarchical clustering** of rows and columns.



X

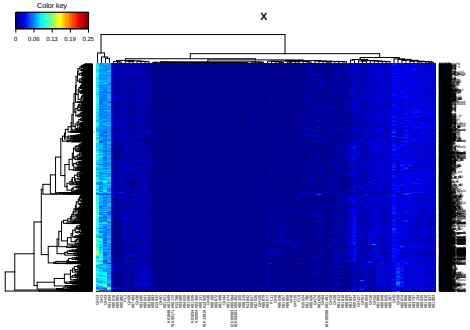


W

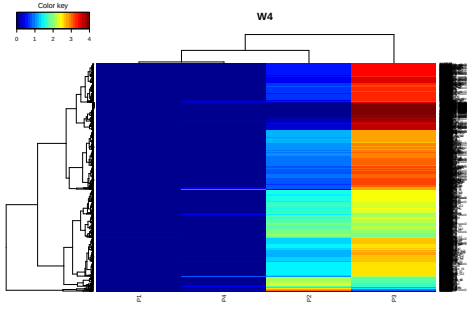
Analysis of the individual profiles in W (1)

Is there an underlying clustering structure in matrix W ?

- Visual examination : **hierarchical clustering** of rows and columns.



X



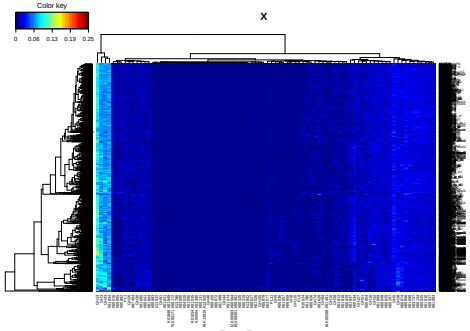
W

- Gap = difference between the **ratio "variance intra cluster / total variance"** between **true and bootstrapped** data
  - For a given clustering method, compute the gap for $k = 1, 2, ...$ clusters
  - Asa the gap decreases when a new cluster is added, this cluster is not significant.

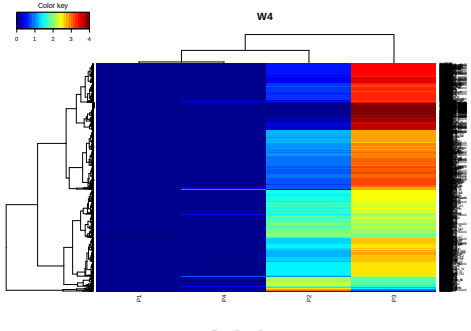
Analysis of the individual profiles in W (1)

Is there an underlying clustering structure in matrix W ?

- Visual examination : **hierarchical clustering** of rows and columns.

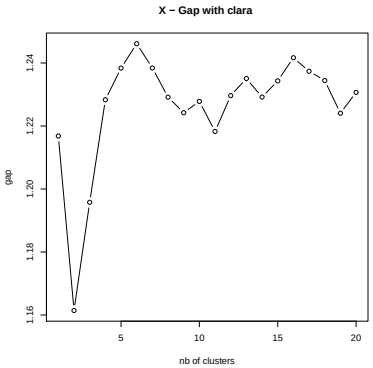


X

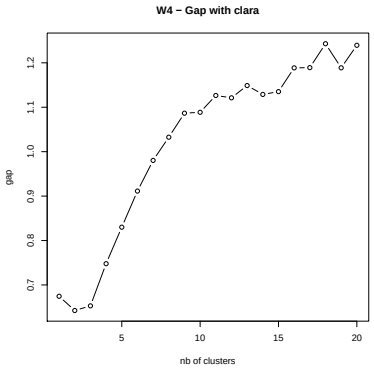


W

- Gap = difference between the **ratio "variance intra cluster / total variance"** between **true and bootstrapped** data
  - For a given clustering method, compute the gap for $k = 1, 2, ...$ clusters
  - Asa the gap decreases when a new cluster is added, this cluster is not significant.



X

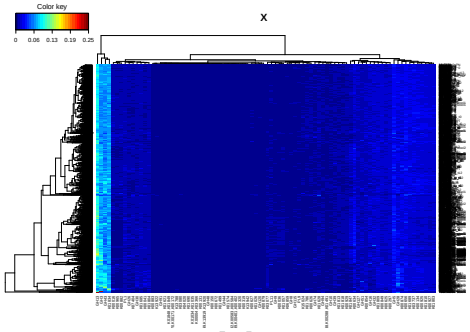


W

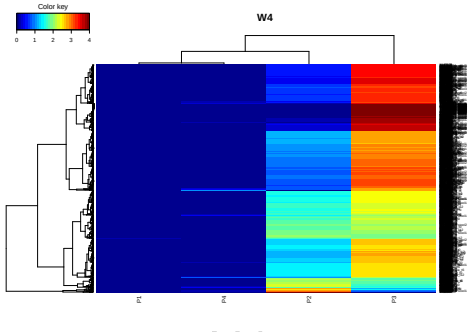
Analysis of the individual profiles in W (1)

Is there an underlying clustering structure in matrix W ?

- Visual examination : **hierarchical clustering** of rows and columns.

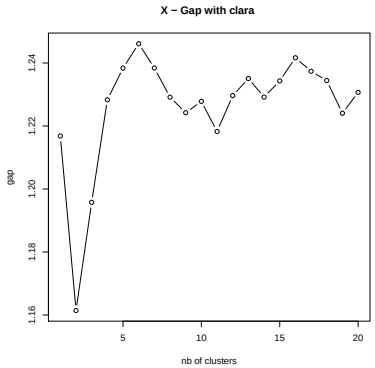


X

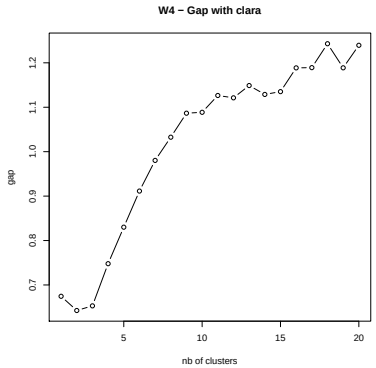


W

- Gap = difference between the **ratio "variance intra cluster / total variance"** between **true and bootstrapped** data
  - For a given clustering method, compute the gap for $k = 1, 2, ...$ clusters
  - Asa the gap decreases when a new cluster is added, this cluster is not significant.



X



W

Conclusion :
- no strong clustering.
- smaller significant clusters in W.

Analysis of the individual profiles in W (2)

- 8 **metavariables** (age, disease, nationality, BMI...) are available.

Analysis of the individual profiles in W (2)

- 8 **metavariables** (age, disease, nationality, BMI...) are available.
- Are metabolic pathways associated with some metavariables ?

Analysis of the individual profiles in W (2)

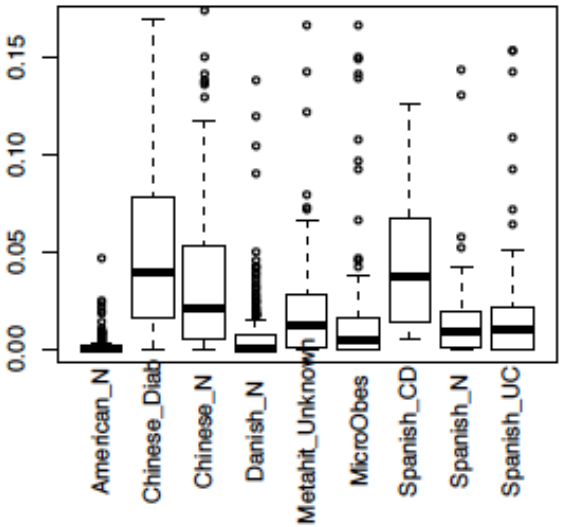
- 8 **metavariables** (age, disease, nationality, BMI...) are available.
- Are metabolic pathways associated with some metavariables ?
- The metavariables are highly correlated by construction (e.g. study focused on a specific disease and done in a given country).

  $\hookrightarrow$ Caution in interpretation

Analysis of the individual profiles in W (2)

- 8 **metavariables** (age, disease, nationality, BMI...) are available.
- Are metabolic pathways associated with some metavariables ?
- The metavariables are highly correlated by construction (e.g. study focused on a specific disease and done in a given country).

  $\hookrightarrow$ Caution in interpretation
- **Example of visual result** : boxplot of abundances of metabolic pathway 4 by study*disease :



- In Spanish study, pathway 4 is more present in Crohn disease patients.
- **Difficult to validate statistically**
  - Multiple testing problem.
  - Pathways have been built to discriminate individuals : p-values may be biased.

  $\hookrightarrow$ Perspective for validation on an independent data set of Crohn/healthy indivduals.

Summary of the biological analysis

- Goal : **explore potential of our methodology** to analyse metabolic pathways based on metagenomics data
  $\hookrightarrow$ Focus on fiber digestion in gut
- Procedure :
  - **Build a matrix** with abundances in reactions associated to fiber digestion, **combining metagenomic data and a priori knowlegde**.
  - **Decompose** this matrix in pathways by NMF
- Results
  - This procedure provides results easy to handle and interprete.
  - Some biological knowledge is recovered
  - NMF can enlight metabolic pathways which represent a low signal but may be significantly discriminative.
- Perspective : validation on independent data.

Summary on the NMF methodology

NMF procedure is defined by the following elements :

- Distance = statistical model
  $\hookrightarrow$ We have chosen Frobenius as a first approach.
- Criterion for the choice of the reduced dimension : linked to the global goal of NMF decomposition.
  $\hookrightarrow$ We have proposed more interpretable criteria than the usual numerical stability.
- Penalty : linked to general biological assumptions / posterior use of the NMF results
- Possibly : additional constraints
- Algorithm
  $\hookrightarrow$ We have developed an alternate minimisation algorithm that can be adapted to various distances and constraints

Methodological perspectives

- Compare results on our data with Frobenius and KL distances (adapt algorithm for KL distance)
- Develop a simulation framework for NMF
  $\hookrightarrow$ Further : pseudo-simulations which keep the correlation structure of experimental data and generate a controlled signal on them
- Propose an analysis of the criteria used for the choice of the reduced dimension k, at first using simulations.
  $\hookrightarrow$ In particular, the concordance of H could require a rescaling by a function of k
- Develop residuals analysis to choose between various distances.
- Theoretical analysis of NMF estimator in a statistical context : convergence, consistence...