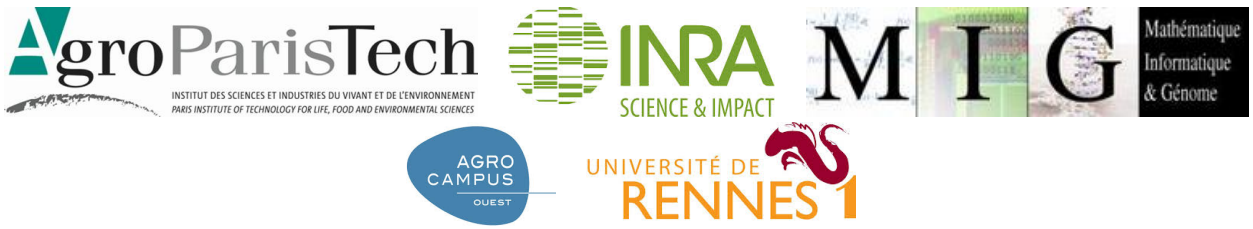
Metagenomics data analysis using a latent block model: application to plant-microbial communities interactions in the rhizosphere .

Julie Aubert, Sophie Schbath, Béatrice Laroche

Metagenomic data analysis using a LBM

Application to plant-microbial communities interactions in the rhizosphere

J. Aubert, C. Mougel, S. Schbath, S. Robin
julie.aubert@agroparistech.fr



Outline

Metagenomic data

Latent Block Model
- Model and inference
- Variational inference
- Poisson and Zero-Inflated Poisson Latent Block Model

Application to plant-microbial communities interactions in the rhizosphere

Conclusions and perspectives

Motivating example

Metagenomics : Study of genetic material recovered directly from environmental samples

Some applications

- Sequencing of the genomes of all the life on earth
- Monitoring the impact of pollutants on ecosystems
- Understanding the correlation between human health and changes in the human microbiome

One possible aim

Study the relationships between bacteria communities and environmental samples
Idea : **Simultaneous clustering** of the bacteria and samples

Metagenomic data

	S1	S2	S3	...	Sj	...	Sd
Bact. 1	0	0	0	...	y_{1j}	...	3
Bact. 2	1	15	0	...	y_{2j}	...	0
Bact. 3	59	17	43	...	y_{3j}	...	3
...	...	...	...	...	...	...	...
Bact. i	y_{i1}	y_{i2}	y_{i3}	...	y_{ij}	...	y_{id}
...	...	...	...	...	...	...	...
Bact. n	90	120	123	...	y_{nj}	...	95
Seq. depth	4738	5157	6010	...	$\sum_{i=1}^{n} y_{ij}$	...	5916

y_{ij} = number of sequences from sample j assigned to bacteria i.

Data characteristics

Count data with **excess of zeros** dependent from the sequencing effort

Latent Block Model (LBM) (Govaert et Nadif. 2003)

Assumptions

Identical distribution resp. for all rows and columns :
$(Z_i) \sim \mathcal{M}(1, \pi = (\pi_1, \ldots, \pi_g))$ $(W_j) \sim \mathcal{M}(1, \rho = (\rho_1, \ldots, \rho_m))$
Latent variables (Z_i) and (W_j) are independent
Latent class assumption : the random variables Y_{ij} are independent conditionally on labels

Model

$(Y_{ij}), i = 1, \ldots, n$ et $j = 1, \ldots, d$ matrix of observations.

$$Y_{ij}|(Z_{ik} = 1, W_{j\ell} = 1) \sim f(.; \alpha_{k\ell})$$

$\alpha = (\alpha_{k\ell})$ parameters for the distribution f.
$\theta = (\alpha, \pi, \rho)$ vector of parameters.

Inference - ML approach

Based on the observed data $Y = (Y_{ij})$, we want to infer :

- the parameters $\theta = (\alpha, \pi, \rho)$
- the hidden status $Z = (Z_i)$ and $W = (W_j)$

Direct approach

log-Likelihood : $logP(Y;\theta) = log \sum_{Z,W} P(Y, Z, W; \theta)$

Can not be computed : $g^n * m^d$ *terms*

Ex : $g = 2, m = 2, n = 100, m = 10 \rightarrow 1.29e^{33}$ *terms to calculate.*

E-M algorithm (Dempster et al. 1977)

Aims at maximizing the log-likelihood

$$logP(Y;\theta)$$

through the alternation of two steps

EM-trick

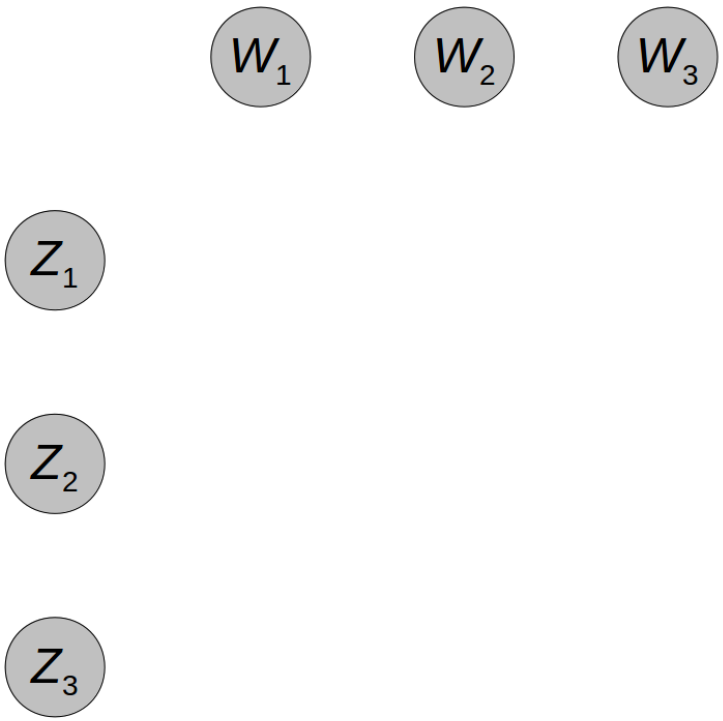
$$logP(Y;\theta) = logP(Y,H;\theta) - logP(H|Y;\theta)$$

$$logP(Y;\theta) = \mathbb{E}[logP(Y,H;\theta)|Y] + \mathcal{H}[P(H|Y;\theta)]$$

where $\mathcal{H}$ stands for the entropy and $H = (Z, W)$

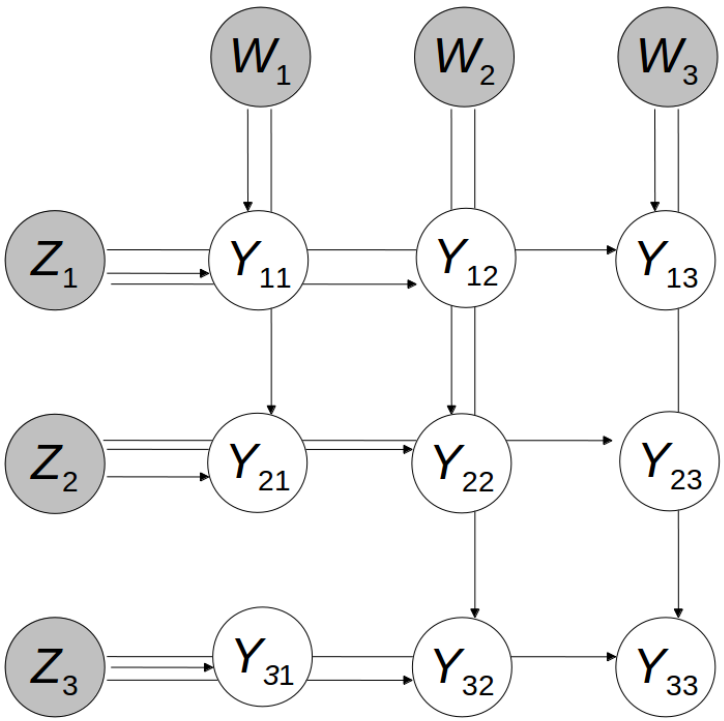
1. Expectation-step :
   calculate $P(H|Y) \rightarrow$ sometimes impossible
2. Maximization-step :
   maximize $\mathbb{E}[logP(Y,H;\theta)|Y]$ with respect to θ
   $\rightarrow$ generally similar to standard MLE.

Graphical representation



- Z_i and W_j unobserved labels of row i and column j

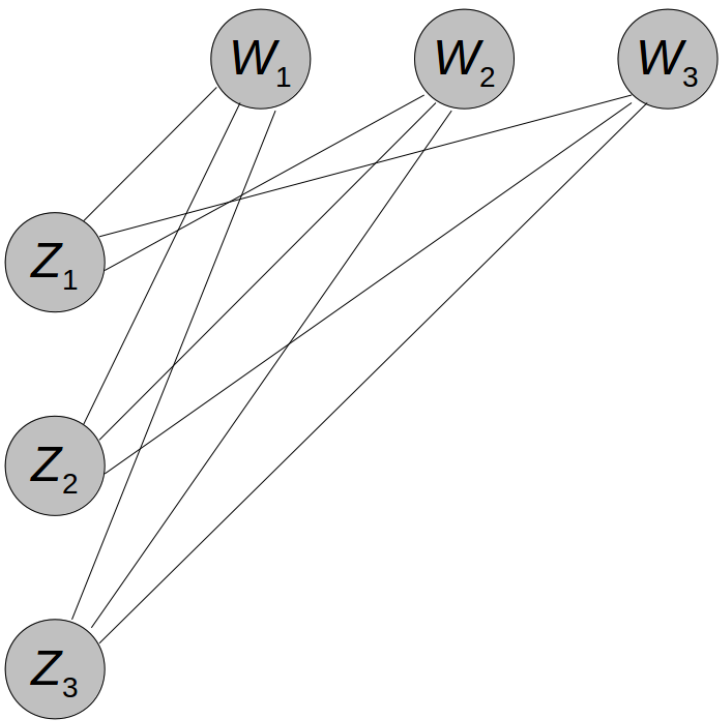
Graphical representation



- Z_i and W_j unobserved labels of row i and column j
- Y_{ij} depends on the labels.

Graphical representation

Conditional distribution of $H = (Z, W)$



- Z_i and W_j unobserved labels of row i and column j
- Y_{ij} depends on the labels.
- Z_i and W_j are not independent conditionally on Y_{ij}
  $\rightarrow P(H|Y)$ intractable
  Regular EM algorithm cannot be used

Variational approximation - VEM (Wainwright and Jordan 2008)

Notations : $H = (Z, W)$

Lower bound of the log-likelihood

For any distribution Q(H) :

$$logP(Y) \geq logP(Y) - KL[Q(H), P(H|Y)]$$

$$logP(Y) \geq \mathbb{E}_Q[logP(Y, H)] + \mathcal{H}(Q(H))$$

Link with EM :

$$logP(Y) = \mathbb{E}[logP(Y, H)] + \mathcal{H}(P(H|Y))$$

replacing $P(H|Y)$ with $Q(H)$

Variational EM algorithm

Variational E-step : replace the calculation of $P(H|Y)$ with the search of

$$Q^*(H) = \underset{Q \in \mathcal{Q}}{argmin} KL[Q(H), P(H|Y)]$$

M-step : compute

$$\hat{\theta} = \underset{\theta}{argmax} \mathbb{E}_{Q^*}[logP(Y, H; \theta)]$$

Mean-field approximation

$\mathcal{Q}$ = set of factorisable distr. with independence of the labels :

$$\mathcal{Q} = Q : Q(H) = \prod_{i=1}^{n} Q_i(Z_i) \prod_{j=1}^{d} Q_j(W_j),$$

with $\prod_i Q_i(Z_i) = \prod_i \prod_k s_{ik}^{Z_{ik}}, \prod_j Q_j(W_j) = \prod_j \prod_l t_{j\ell}^{W_{j\ell}}$,
where $s_{ik} = \mathbb{E}_Q(Z_{ik})$, $t_{j\ell} = \mathbb{E}_Q(W_{j\ell})$, $s_{ik}\, t_{j\ell} = \mathbb{E}_Q(Z_{ik} W_{j\ell})$

Generalized Variational EM (Govaert et Nadif 2010)

Plug-in of s_{ik}, $t_{j\ell}$ in place of Z_{ik}, $W_{j\ell}$, $s_{ik}t_{j\ell}$ in place of $Z_{ik}W_{j\ell}$.

$$F(s,t,\theta) = \mathbb{E}_Q log P(Y,Z,W;\theta) + \mathcal{H}(s) + \mathcal{H}(t)$$

$$(*) = \sum_{i,k} s_{ik} log \pi_k + \sum_{j,\ell} t_{j\ell} log \rho_\ell + \sum_{i,j,k,\ell} s_{ik}t_{j\ell} log(f(Y_{ij};\alpha_{k\ell})$$

1. Initialization of the unknown s, t, π, ρ, α

Generalized Variational EM (Govaert et Nadif 2010)

Plug-in of s_{ik}, $t_{j\ell}$ in place of Z_{ik}, $W_{j\ell}$, $s_{ik}t_{j\ell}$ in place of $Z_{ik}W_{j\ell}$.

$$F(s,t,\theta) = \mathbb{E}_Q log P(Y,Z,W;\theta) + \mathcal{H}(s) + \mathcal{H}(t)$$

$$(*) = \sum_{i,k} s_{ik} log \pi_k + \sum_{j,\ell} t_{j\ell} log \rho_\ell + \sum_{i,j,k,\ell} s_{ik}t_{j\ell} log(f(Y_{ij};\alpha_{k\ell})$$

1. Initialization of the unknown s, t, π, ρ, α
2. VEM on rows : Max $F(s,t,\theta)$ with respect to s, π and α with fixed ρ and t.

Generalized Variational EM (Govaert et Nadif 2010)

Plug-in of s_{ik}, $t_{j\ell}$ in place of Z_{ik}, $W_{j\ell}$, $s_{ik}t_{j\ell}$ in place of $Z_{ik}W_{j\ell}$.

$$F(s,t,\theta) = \mathbb{E}_Q log P(Y,Z,W;\theta) + \mathcal{H}(s) + \mathcal{H}(t)$$

$$(*) = \sum_{i,k} s_{ik} log \pi_k + \sum_{j,\ell} t_{j\ell} log \rho_\ell + \sum_{i,j,k,\ell} s_{ik}t_{j\ell} log(f(Y_{ij};\alpha_{k\ell})$$

1. Initialization of the unknown s, t, π, ρ, α
2. VEM on rows : Max $F(s,t,\theta)$ with respect to s, π and α with fixed ρ and t.
3. VEM on columns : Max $F(s,t,\theta)$ with respect to t, ρ and α with fixed π and s.

Generalized Variational EM (Govaert et Nadif 2010)

Plug-in of s_{ik}, $t_{j\ell}$ in place of Z_{ik}, $W_{j\ell}$, $s_{ik}t_{j\ell}$ in place of $Z_{ik}W_{j\ell}$.

$$F(s,t,\theta) = \mathbb{E}_Q log P(Y,Z,W;\theta) + \mathcal{H}(s) + \mathcal{H}(t)$$

$$(*) = \sum_{i,k} s_{ik} log \pi_k + \sum_{j,\ell} t_{j\ell} log \rho_\ell + \sum_{i,j,k,\ell} s_{ik}t_{j\ell} log(f(Y_{ij};\alpha_{k\ell})$$

1. Initialization of the unknown s, t, π, ρ, α
2. VEM on rows : Max $F(s,t,\theta)$ with respect to s, π and α with fixed ρ and t.
3. VEM on columns : Max $F(s,t,\theta)$ with respect to t, ρ and α with fixed π and s.
4. Iterate steps 2 and 3 until convergence.

Generalized Variational EM (Govaert et Nadif 2010)

Plug-in of s_{ik}, $t_{j\ell}$ in place of Z_{ik}, $W_{j\ell}$, $s_{ik}t_{j\ell}$ in place of $Z_{ik}W_{j\ell}$.

$$F(s,t,\theta) = \mathbb{E}_Q log P(Y,Z,W;\theta) + \mathcal{H}(s) + \mathcal{H}(t)$$

$$(*) = \sum_{i,k} s_{ik} log \pi_k + \sum_{j,\ell} t_{j\ell} log \rho_\ell + \sum_{i,j,k,\ell} s_{ik}t_{j\ell} log(f(Y_{ij};\alpha_{k\ell})$$

1. Initialization of the unknown s, t, π, ρ, α
2. VEM on rows : Max $F(s,t,\theta)$ with respect to s, π and α with fixed ρ and t.
3. VEM on columns : Max $F(s,t,\theta)$ with respect to t, ρ and α with fixed π and s.
4. Iterate steps 2 and 3 until convergence.

Prop : the lower bound increases after each iteration.

Poisson and Zero-Inflated Poisson Latent Block Model

Poisson LBM $Y_{ij} \in \mathbb{N},\ Y_{ij}|k,\ell \sim \mathcal{P}(\mu_i \nu_j \alpha_{k\ell})$

- μ_i : mean level of presence of one particular bacteria
- ν_j : scale factor correcting for sequencing depth (around 1)
- $\alpha_{k\ell}$: interaction level between bacteria and environmental samples within the (k, ℓ) group

Poisson and Zero-Inflated Poisson Latent Block Model

Poisson LBM $Y_{ij} \in \mathbb{N},\ Y_{ij}|k,\ell \sim \mathcal{P}(\mu_i \nu_j \alpha_{k\ell})$

- μ_i : mean level of presence of one particular bacteria
- ν_j : scale factor correcting for sequencing depth (around 1)
- $\alpha_{k\ell}$: interaction level between bacteria and environmental samples within the (k, ℓ) group

Zero-inflated Poisson LBM $Y_{ij} \in \mathbb{N},\ Y_{ij}|k,\ell \sim ZIP(.; \alpha_{k\ell})$

$$Y_{ij}|(Z_{ik} = 1, W_{j\ell} = 1) \sim ZIP(.; \alpha_{k\ell})$$

$$\begin{cases} P(Y_{ij}|(Z_{ik} = 1, W_{j\ell} = 1) = 0) = \delta_{k\ell} + (1 - \delta_{k\ell}) e^{-\alpha_{k\ell}} \\ P(Y_{ij}|(Z_{ik} = 1, W_{j\ell} = 1) = y_{ij}) = (1 - \delta_{k\ell}) \frac{\alpha_{k\ell}^{y_{ij}} e^{-\alpha_{k\ell}}}{y_{ij}!} \text{ for } y_{ij} \geq 1 \end{cases}$$

Zero-Inflated Latent Block Model - Inference

Similar to Poisson Latent Block Model.
Part of the M-step :

$$\widehat{\delta_{k\ell}} = \frac{\bar{y_{k\ell}}}{\sum_{i,j} s_{ik} t_{j\ell} \widehat{\alpha_{k\ell}}}$$

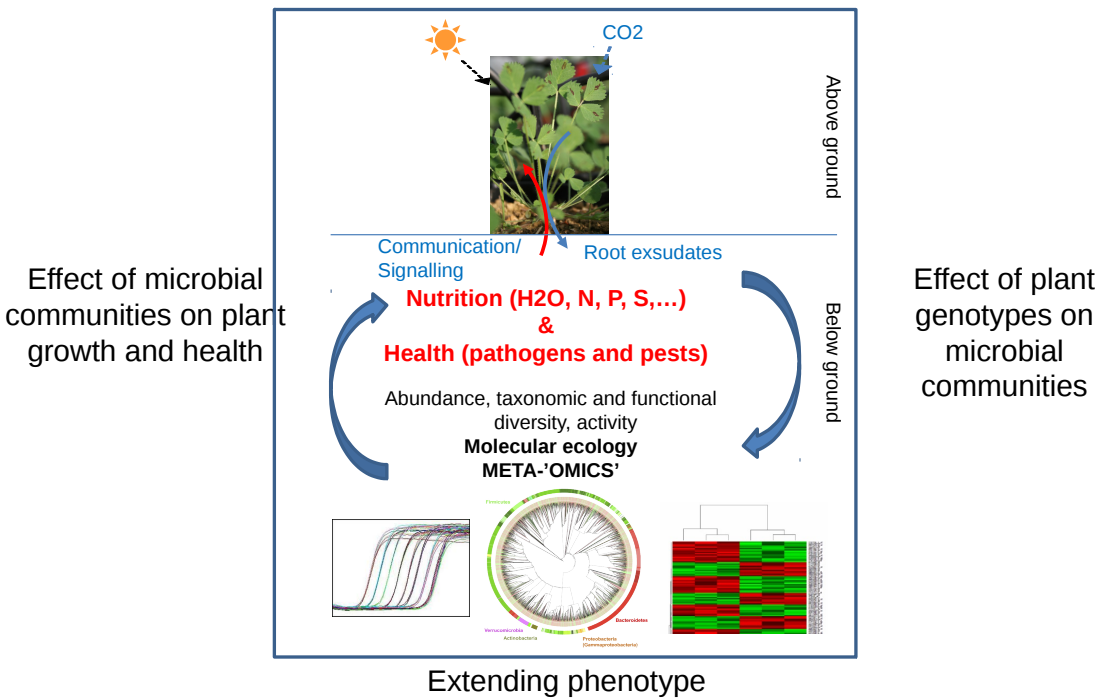
where $\bar{y_{k\ell}} = \frac{\sum_{i,j} s_{ik} t_{j\ell} y_{i,j}}{\sum_{i,j} s_{ik} t_{j\ell}}$

A fixed-point equation to solve :

$$\widehat{\alpha_{k\ell}} = \frac{\bar{y_{k\ell}}(1 - e^{1-\widehat{\alpha_{k\ell}}})}{1 - n_0^{k\ell}/n^{k\ell}} = G(\widehat{\alpha_{k\ell}})$$

Application - MétaRhizo project (C. Mougel)

Hyp : Plant genotype modifies the structure of the bacteria community in the rhizosphere (the region of soil directly influenced by root secretions and associated soil microorganisms).

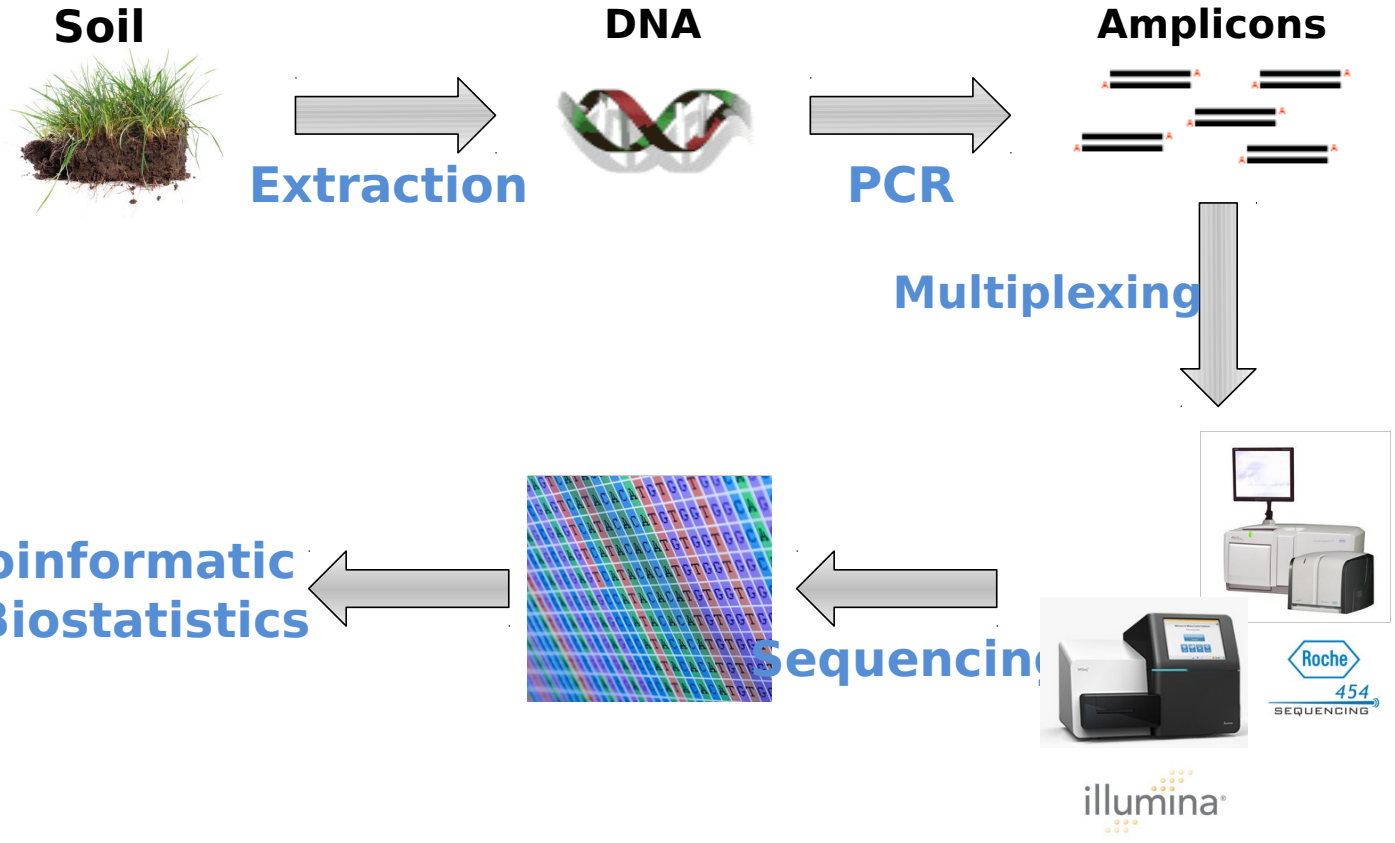


Medicago truncatula

Small annual legume : model organism in genomic research.
Forms symbioses with nitrogen-fixing rhizobia and arbuscular mycorrhizal fungi.
$\rightarrow$ **Reduction of nitrogen inputs** responsible for various pollution (substainable agriculture)

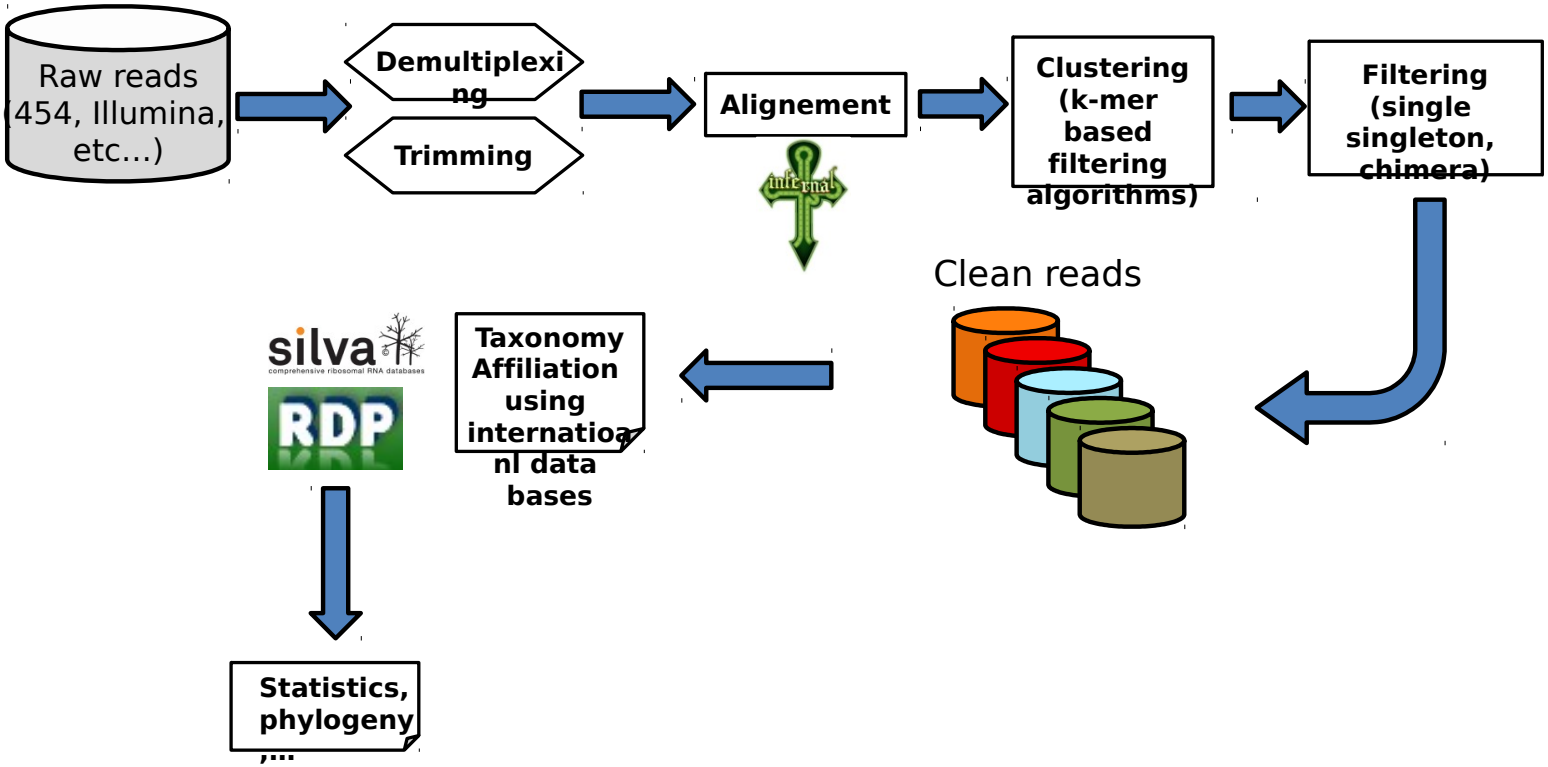
Sampling strategy of the M. truncatula core collection, Ronfort et al. (2006)

Standard operating procedures from soil collection to mathematical analysis



Agroécologie Dijon Unité de Recherche
GenoSol PLATEFORME
INRA SCIENCE & IMPACT

A dedicated bioinformatic workflow for microbial communities diversity analysis: the GnS-PIPE



Agroécologie Dijon Unité de Recherche GenoSol PLATEFORME INRA SCIENCE & IMPACT

First results

After filtering : 442 bacteria * 508 samples

Poisson LBM with g=5 and m=6

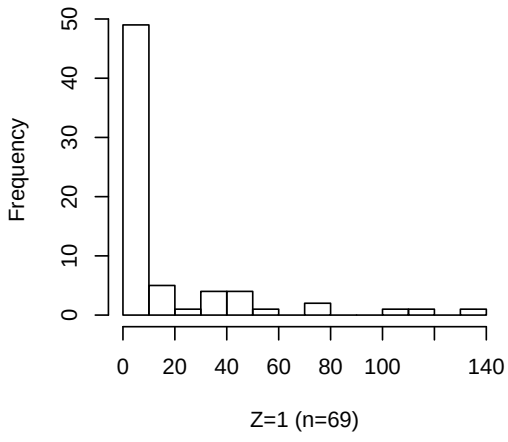
	1	2	3	4	5	6
1	-0.01	-0.00	-0.09	0.01	0.14	0.01
2	-0.06	-0.04	0.10	-0.14	1.67	-0.17
3	0.41	-0.13	0.16	-0.07	-0.38	0.06
4	-0.49	0.02	-0.54	0.19	-0.62	0.02
5	-0.25	0.15	0.26	-0.13	-0.51	-0.08

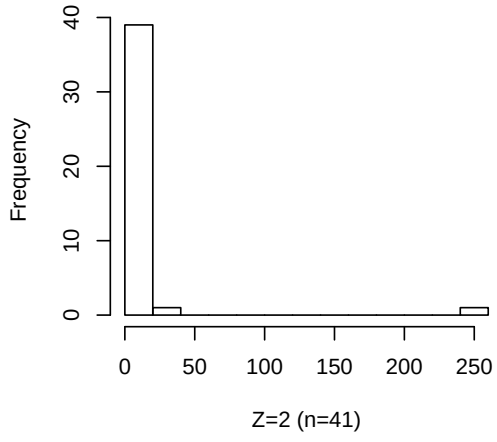
TABLE: $log(\widehat{\alpha_{k\ell}})$

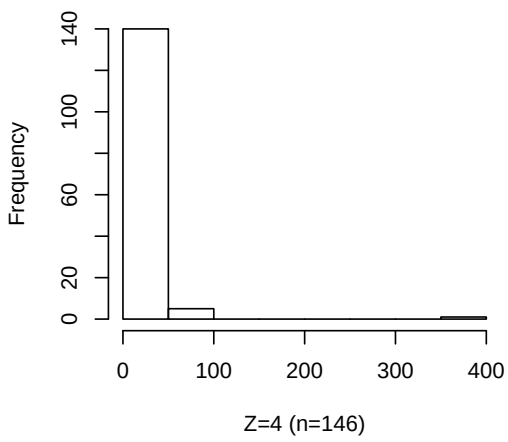
First results

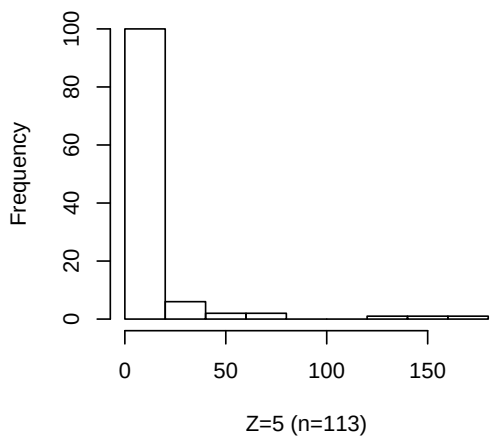
Group $(2, 5)$: 8 plants et 41 bacteria.

Groups are not correlated with the abundance level of bacteria.









First results

Plants

originate from mediterranean Middle-East + Tunisia and Libya,
belongs to 2 genomics group among 4,
but nothing obvious on the ecophysiology of plants.

Bacteria

With one of these following ecophysiological traits :

- acidophilus
- anaerobic
- metabolism of single carbon
- nitrogen metabolism

Conclusions and perspectives

- Parameters biologically interpretable
- First encouraging results
- LBM : parcimonious and complex model able to reduce data dimension

Short-term perspectives

- Application of $ZIP(\mu_i \nu_j \alpha_{k\ell})$ LBM on the data : what advantage ?
- Discussion with my favorite biologist in order to understand the group deciphered by LBM

Perspectives

- Choice of the number of groups
- Variance for $\hat{\alpha_{k\ell}}$

Model with covariates

$$Y_{ij}|(Z_{ik} = 1, W_{j\ell} = 1) \sim \mathcal{P}(\lambda_{k\ell} e^{\beta^\intercal x_{ij}})$$

with x_{ij} vector of covariates.

Acknowledgments

For experiments and biological expertise

C. Mougel (INRA, IGEPP)

Other collaborators and helpful discussion

T. Ha, T. Mary-Huard, C. Keribin and

S. Robin S. Schbath

References

[1] Govaert, G. et Nadif, M. (2010), **Latent Block Model for contingency table**, *Communications in Statistics - Theory and Methods*, 39(3), 416–425.

[2] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), **Maximum likelihood from incomplete data via the EM algorithm**, *Journal of the Royal Statistical Society, B* 39(1), 1-38.

[3] Ronfort, J. et al. (2006), **Microsatellite diversity and broad scale geographic structure in a model legume : building a set of nested core collection for studying naturally occurring variation in Medicago truncatula**, BMC Plant Biology, 6(1), 28.

[4] Wainwright, M. J. and Jordan, M. I. (2008), **Graphical models, exponential families, and variational inference. Found**, *Trends Mach. Learn.*,1, 1–305.