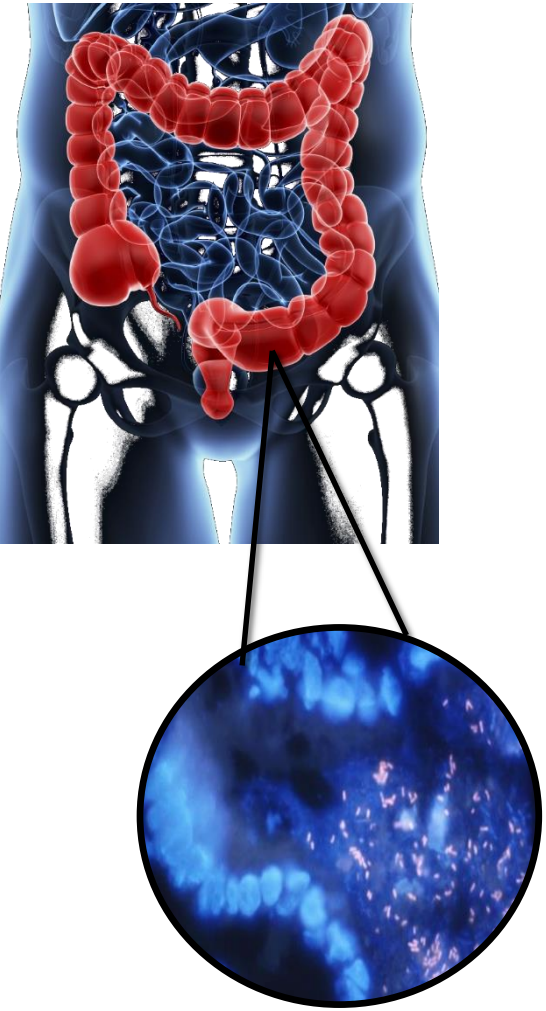
Modélisation du Microbiote intestinal

Béatrice Laroche
INRA, MaIAGE

Travaux conjoints avec
Sébastien Raguideau, Sandra Plancade (MaIAGE), Marion Leclerc (MICALIS)
Simon Labarthe (MaIAGE), Magali Ribot (U. Orléans)

22/09/2017 Journée chaire MMD

The human intestinal microbiota



All microorganisms in the human intestine : bacteria, fungi, archaea.

Germ free at birth, progressive colonization, increasing diversity until adulthood, driven by

- Environment (inc. health status)
- Genetics
- Diet

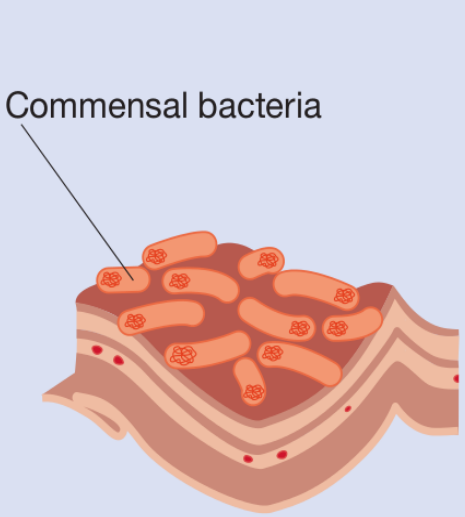
Very large number of individuals

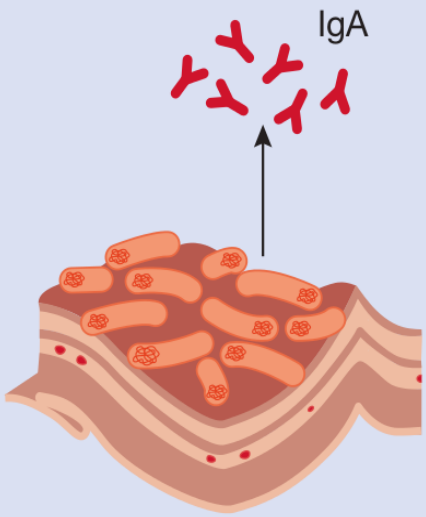
- 10^{14} microorganisms
- 1000 different species

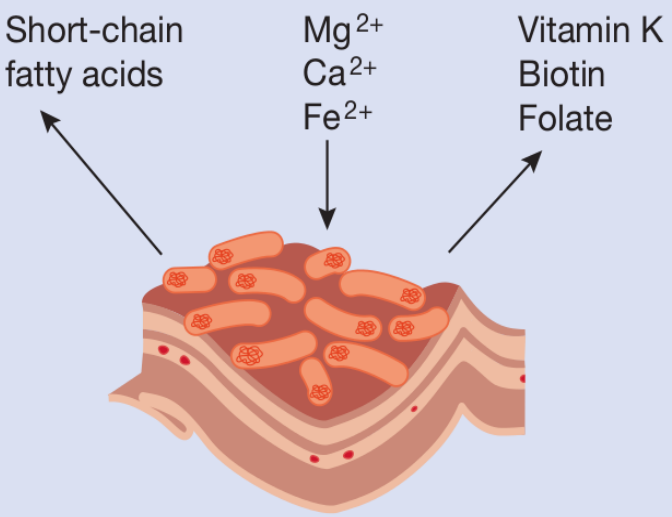
Symbiotic relationship between the microorganisms and their host.

Main Ecosystemic Functions

Protective functions	Structural functions	Metabolic functions
Ecological barrier	**Immune System Education**	**Fermentation of non digestible dietary fibers**





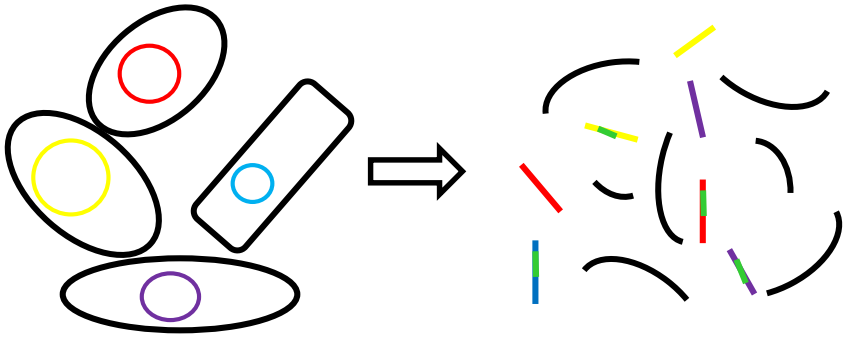


O'Hara and Shanahan (2006), Mosca et al. (2016)

Metagenomic Data

Metagenomics : allow culture free analysis of microorganisms genetic material in samples

2 approaches:



Selective Amplification

Barcoding: amplification of a target sequence

Taxonomic analyses through 16S rDNA amplification

Fragmentation & direct sequencing

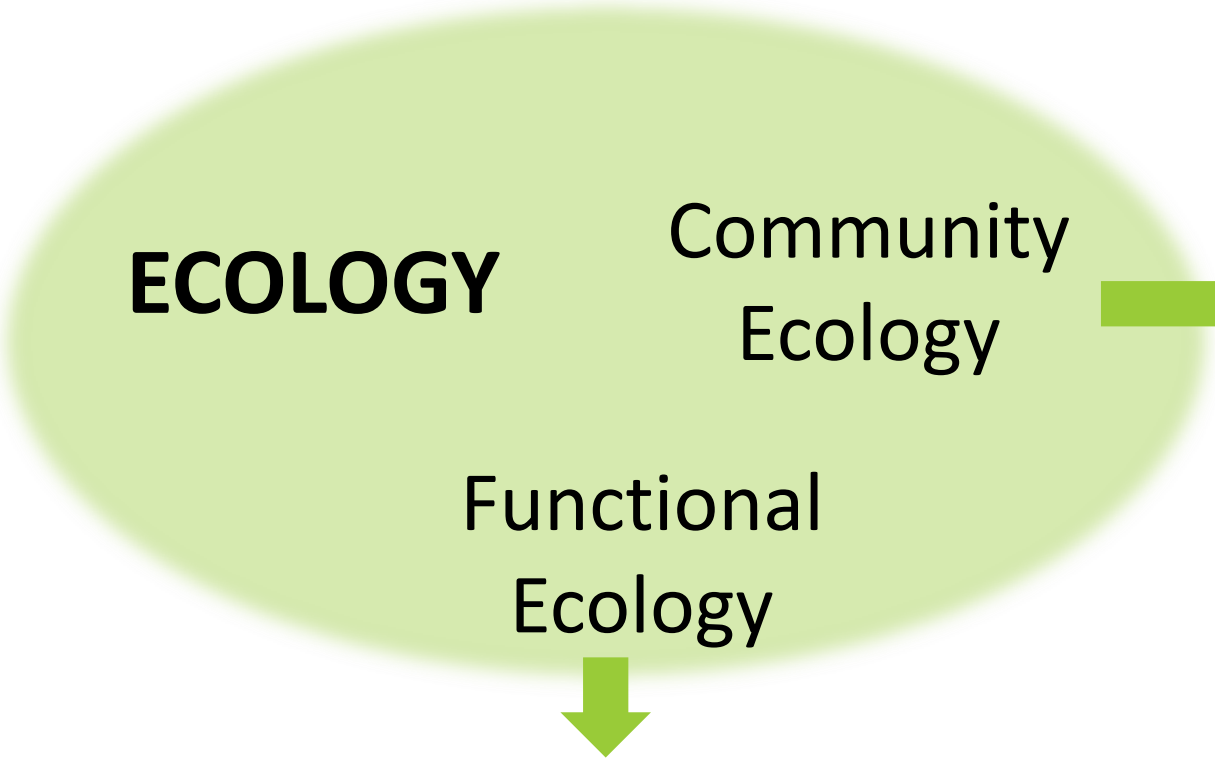
Whole Genome Shotgun (WGS)
total DNA fragmentation and production of small reads

Read mapping to annotated gene catalogues, Counting

Sample 1	1	2	0	13	4
Sample 2	2	0	30	1	14
⋮					
Sample n	10	5	3	14	

Access to all gene abundances, functional approaches

Ecological Framework



focuses on species diversity and Interactions

16S Data

Barrier effect against pathogens

Screening of probiotic strains

focuses on the functions of individuals in their ecosystem and environment

organisms described in terms of *functional traits* rather than taxonomy

WGS data

Fiber metabolism, host/microbe interaction

understanding structure & mechanisms

Outline

- Modelling Aggregated Functional Traits (AFT) for fiber degradation from WGS data

  Sébastien Raguideau, Sandra Plancade (MaIAGE), Marion Leclerc (MICALIS)

- Mechanistic model of fiber degradation in the human gut at the organ scale

- Dynamic interaction models & parameter inference from 16S data

Modelling AFT for fiber degradation

- How can we define AFT of fiber degradation
- Model & mathematical formulation
- Results

Functional Traits

Functional traits are *quantifiable* morphological or physiological characteristics of individuals that directly or indirectly impact their fitness or performance.

Trait : leaf surface
value : x cm^2

Very large number of individuals
$\Rightarrow$ *Aggregated Functional Traits* (AFT)

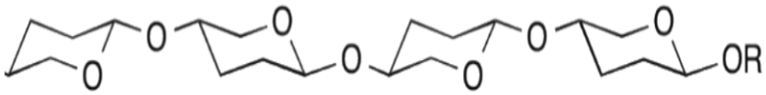
Leaf surface
value = total or mean surface of all the leaves (spectral remote sensing)

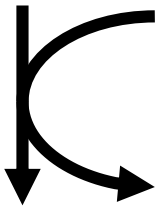
Violle et al. (2007), Homolová et al. (2013), Moreno García et al. (2014)

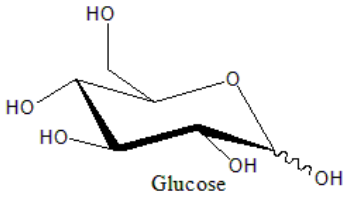
Defining AFT for fiber degradation

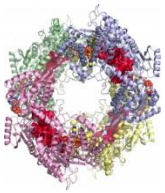
AFT = global catalytic potential for reaction steps in fiber degradation

Biochemical reaction



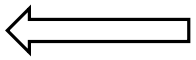






Enzyme catalyzing the reaction

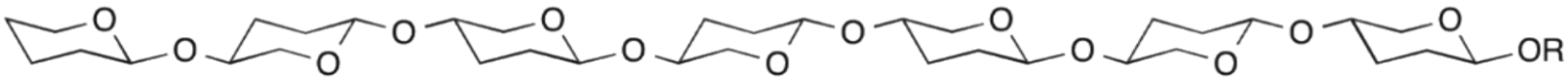
Transcription and synthesis



Gene family associated to a reaction

AFT value : Total abundance of all the genes that potentially code for the synthesis of this enzyme

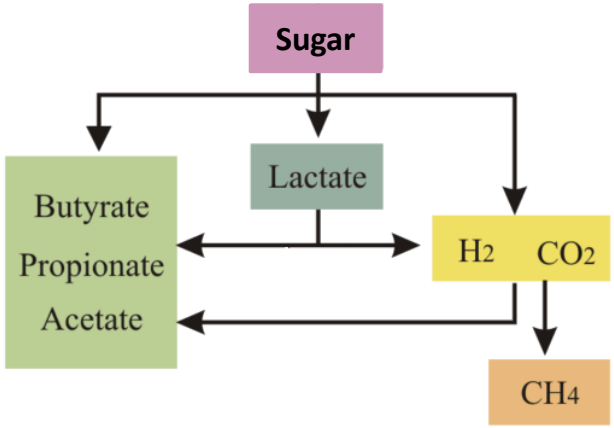
Functional Traits for Fiber Degradation



Hydrolysis

breakdown of fibers into simple sugar molecules

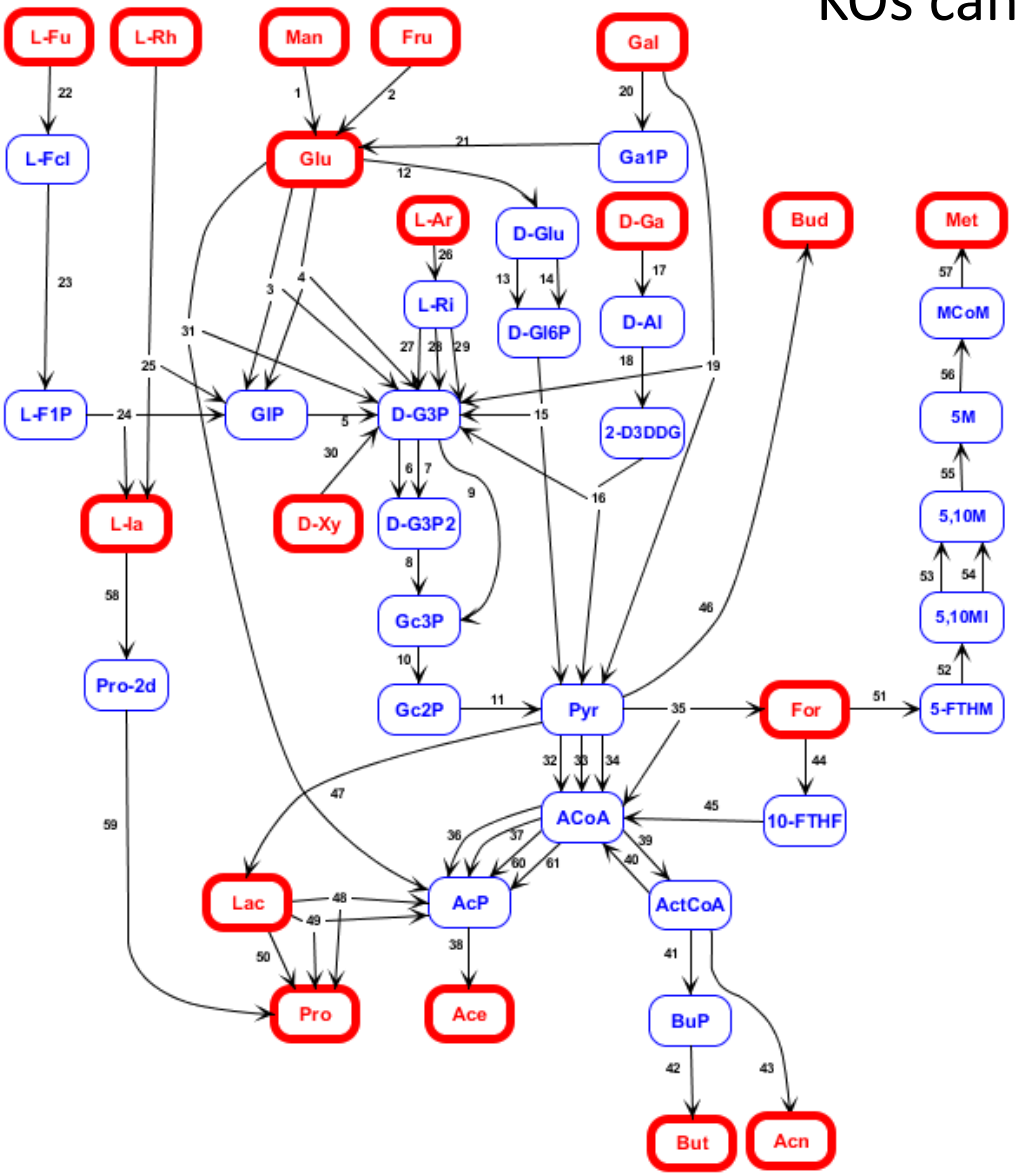
- Glycosides hydrolases (GH)
- Polysaccharides Lyases (PL)



Catabolic conversion of sugars into SCFA and syntrophic pathways

- KEGG Orthologies (KO)

Functional Traits for Fiber Degradation



KOs can be represented on a directed graph

Nodes : metabolic compounds

red : extracellular

blue : assumed intracellular

Edges : 61 AFT (=KOs)

simplified representation

of many reaction steps

ONLY A TOPOLOGICAL GRAPH

NO STOICHIOMETRY

NO FLUXES

Summary

- Selection of 86 AFT 25 GH PL + 61 KOs
- Topological graph whose edges are the 61 KOs
- AFT measured as the total abundance of genes for the GH or KO
- 1408 metagenomic samples from 3 projects (HMP, MetaHIT, MicrObes)
- AFT abundance matrix A size 1408 x 86

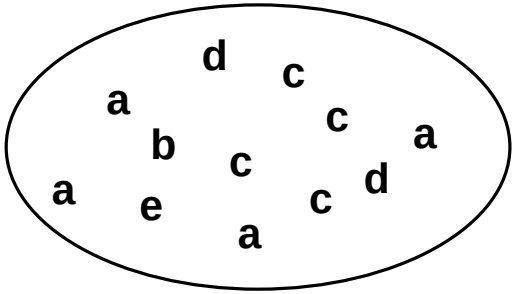
Outline

- How can we define AFT of fiber degradation
- Model & mathematical formulation
- Results

Ecological Model

Abundance of 5 AFT

	a	b	c	d	e
Sample	4	1	4	2	1



Annotated genes in a sample

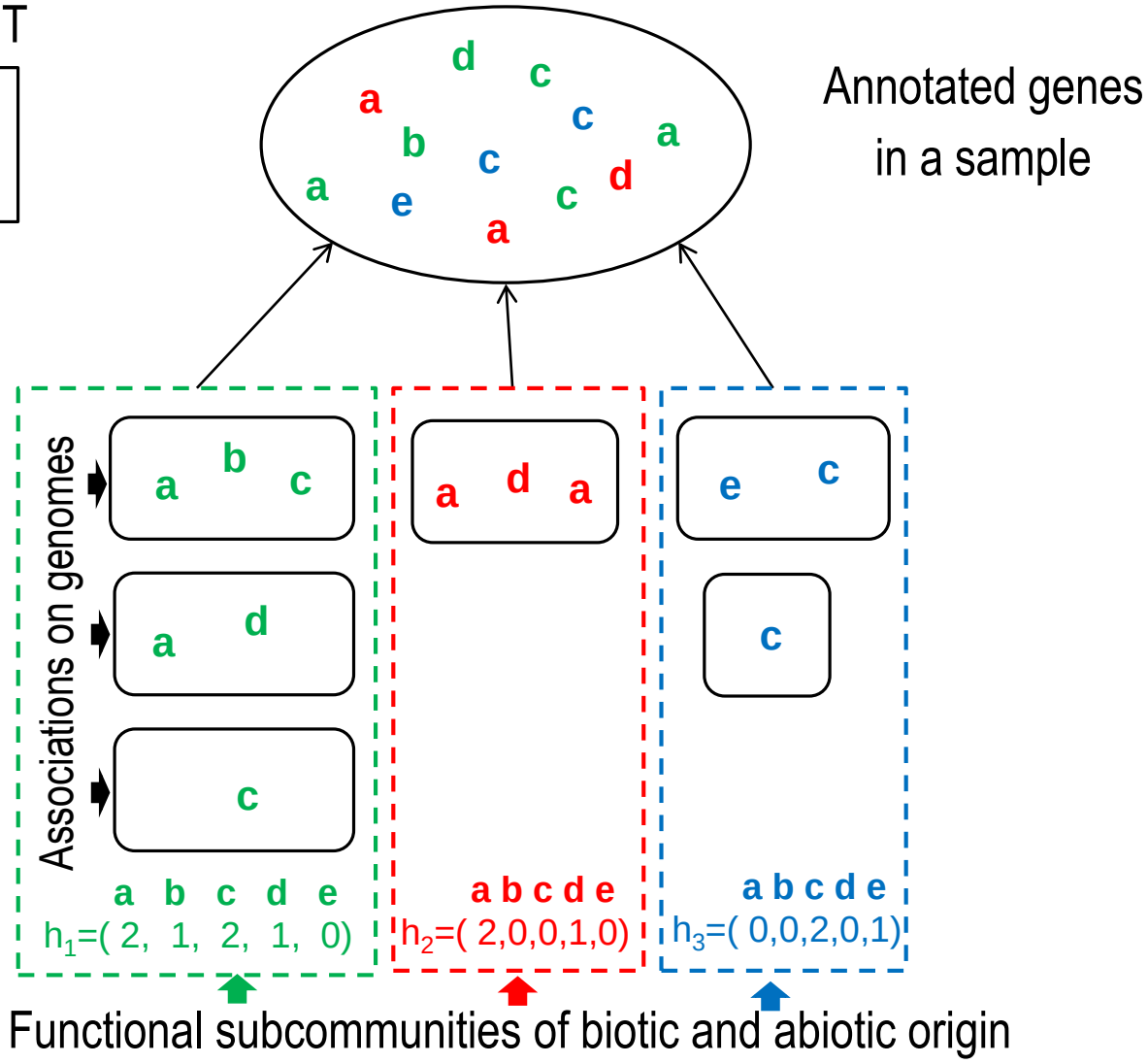
Ecological Model: latent structure

Abundance of 5 AFT

	a	b	c	d	e
Sample	4	1	4	2	1

Latent hierarchical structure :

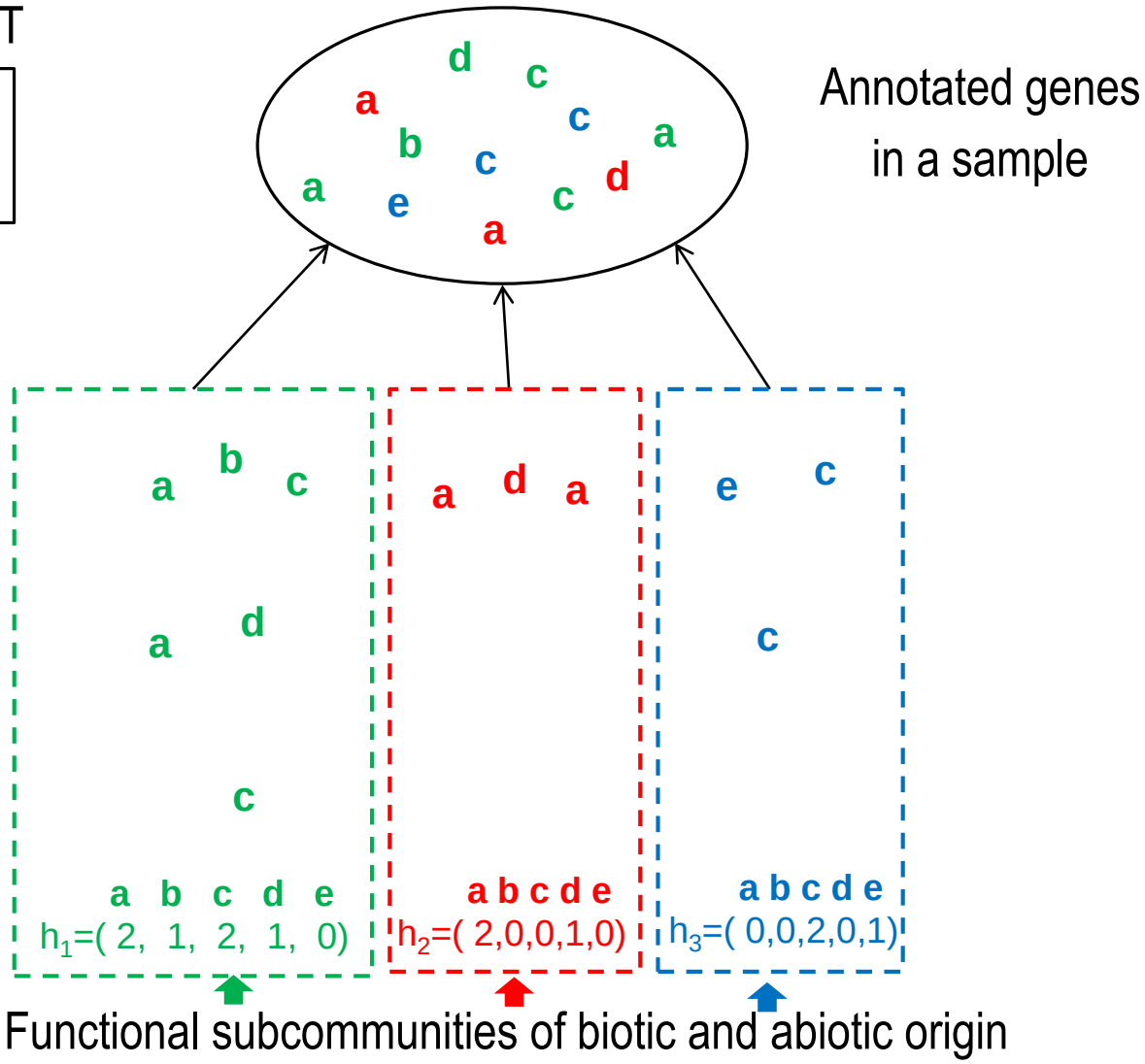
- genes associated on genomes,
- microorganisms associated within subcommunities.



Ecological Model: latent structure

Abundance of 5 AFT

	a	b	c	d	e
Sample	4	1	4	2	1



Latent hierarchical structure :

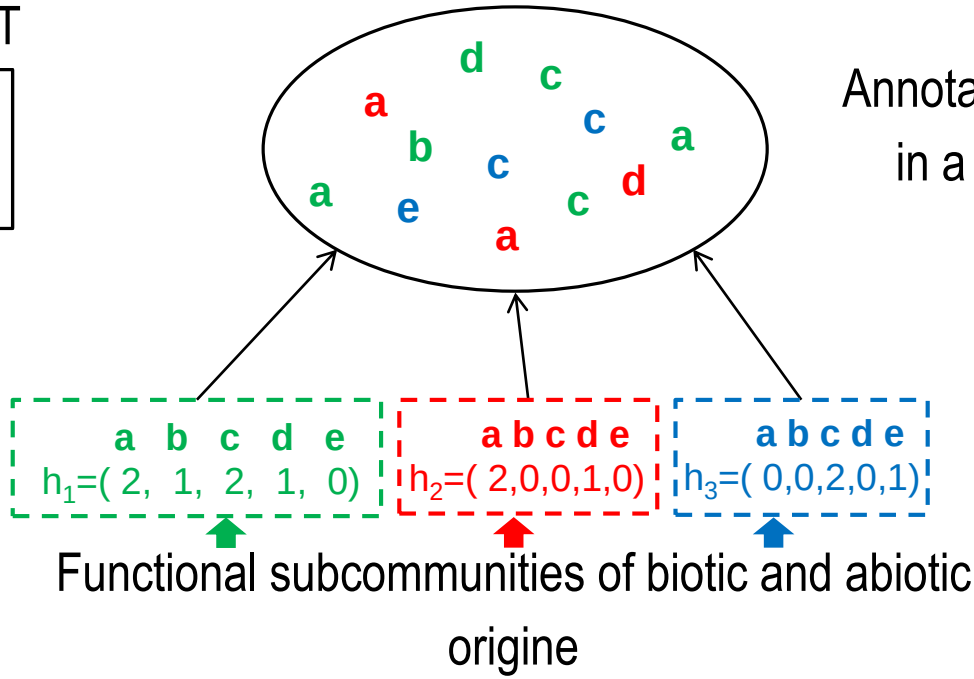
- genes associated on genomes,
- microorganisms associated within subcommunities.

- but we cannot access the ~~genomes~~

Ecological Model: Combined AFT

Abundance of 5 AFT

	a	b	c	d	e
Sample	4	1	4	2	1



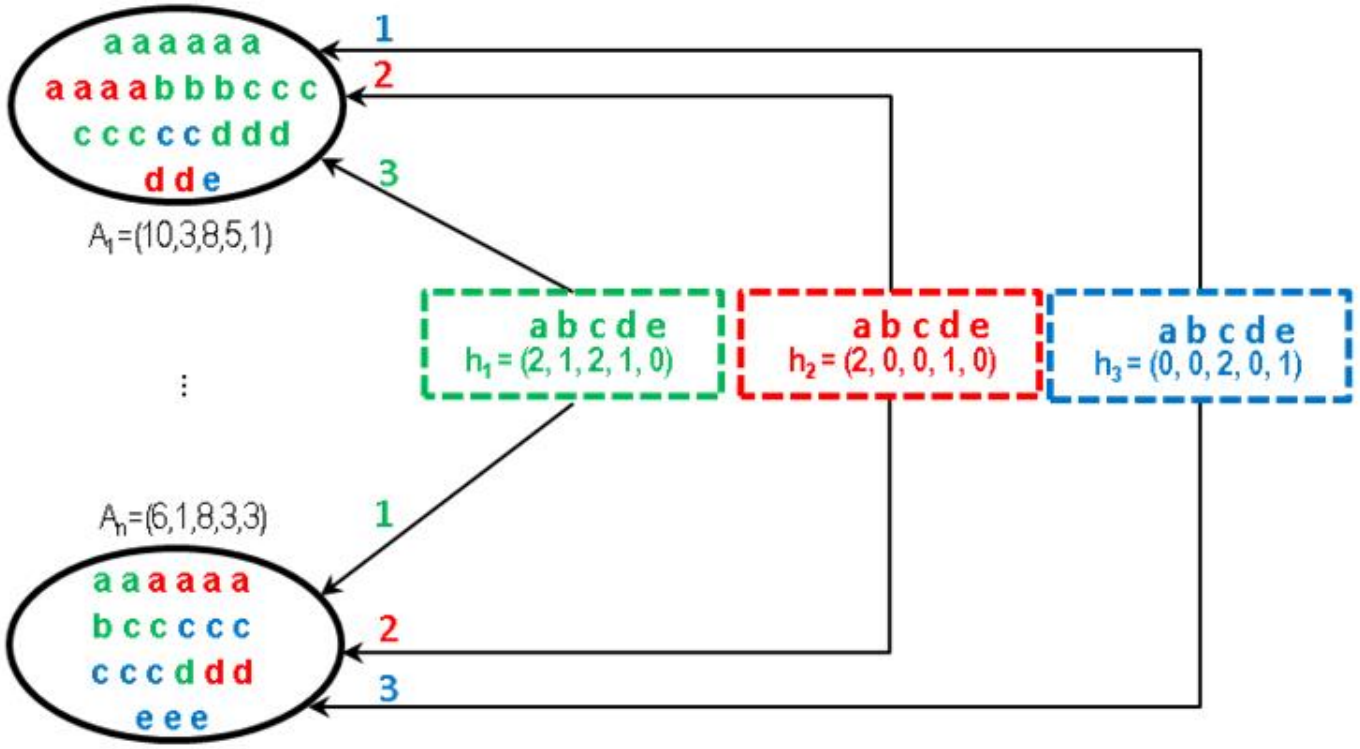
Latent hierarchical structure :

- genes associated within subcommunities.

The functional subcommunities are characterized by **Combined Aggregated Functional Traits (CAFT)** defined by their composition in terms of AFT (h_1, h_2, h_3)

Ecological Model: Mixture of AFT

- The fiber degradation potential can be represented by a limited number of CAFTs
- All samples are mixtures of these CAFTs, in variable proportions



Ecological Model: Mathematical Formulation

AFT abundances are modelled as mixture of ***k*** CAFTs in variable proportions

$$A_{ij} \simeq \sum_{l=1}^{\mathbf{k}} W_{il} H_{lj}$$

A_{ij} abundance AFT j in sample i

W_{il} abundance of CAFT l in sample i

H_{lj} proportion of AFT j in CAFT l

- ***k*** is unknown,
- *A*, *W*, *H* nonnegative => NMF (Nonnegative Matrix Factorization)

Inference Problem: NMF formulation

$$(W^\star, H^\star) = \underset{W\geq 0, H\geq 0}{\arg\min} \quad \mathcal{D}(A|WH) + \mathrm{pen}(W) + \mathrm{pen}(H)$$

- Frobenius norm (Gaussian error model), other possible choices KL divergence (Poisson model)

$$\mathcal{D}(A|WH) = \|A - WH\|_F^2 = \sum_{ij} (A_{ij} - (WH)_{ij})^2$$

- Penalization to lift identifiability problem and impose sparsity

$$\|W\|_F^2 = \sum_{il} (W_{il})^2$$

Ridge

$$\|H\|_{1,2}^2 = \sum_{j=1}^{r} \left(\sum_{l=1}^{k} H_{lj} \right)^2 = \|\mathbb{1}^T H\|_2^2$$

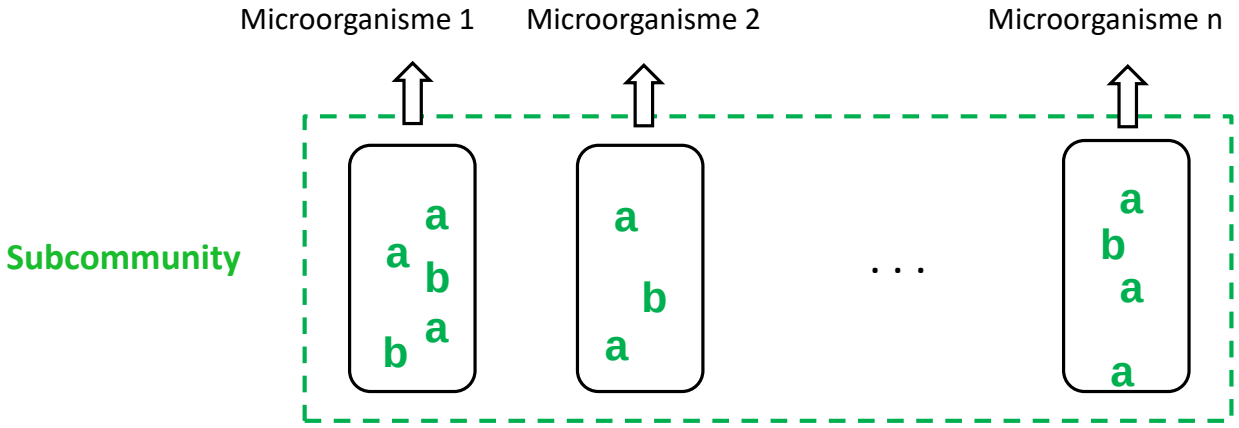
Exclusive LASSO

- Nonnegative Matrix Factorization (NMF) problem

$$(W^\star, H^\star) = \underset{W\geq 0, H\geq 0}{\arg\min} \quad \|A - WH\|_F^2 + \alpha(\|W\|_F^2 + \|H\|_{1,2}^2)$$

Building constraints from genomic information

- CAFT characterize AFT frequencies in subcommunities
- A subcommunity is a set of microorganisms



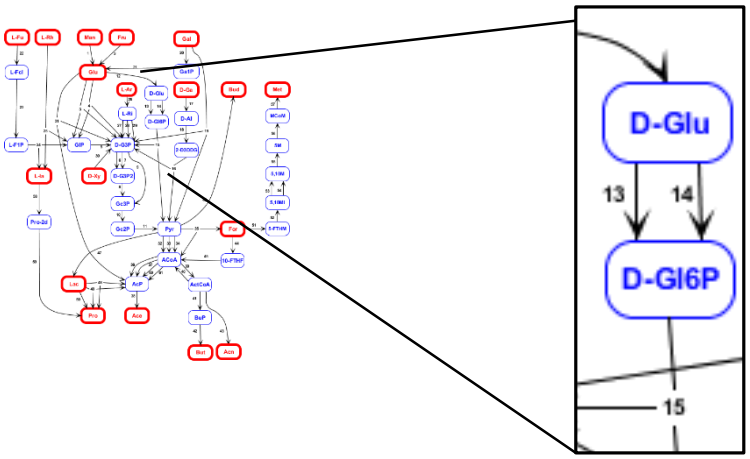
- Find upper bound on the FT ratios valid for all genomes (if possible!)

$$N_a \leq 3N_b \qquad N_b \leq \frac{2}{3}N_a$$

- Derive constraints for all CAFTs

$$H_{la} \leq 3H_{lb} \qquad H_{lb} \leq \frac{2}{3}H_{la}$$

Building constraints from genomic information



look for specific constraints associated to blue nodes (intracellular metabolites)

$$H_{l13} + H_{l14} \leq r_1 H_{l15}$$
$$H_{l15} \leq r_2 (H_{l13} + H_{l14})$$

Use 190 genomes of most prevalent microorganisms in the gut

- Check if bounds exist for blue nodes
- Calibrate the ratio bounds

$$\sum_{AFT\ x\ producing\ m} H_{lx} \leq r_{1m} \sum_{AFT\ y\ consuming\ m} H_{ly}$$
$$\sum_{AFT\ y\ consuming\ m} H_{ly} \leq r_{2m} \sum_{AFT\ x\ producing\ m} H_{lx}$$

25 blue nodes => 38 effective constraints out of 50 possible

$$FH^T \leq 0$$

Inference Problem: solution

Constrained NMF problem

$$(W^\star, H^\star) = \underset{W\geq 0, H\geq 0, FH^T\leq 0}{\arg\min} \quad \|A - WH\|_F^2 + \alpha(\|W\|_F^2 + \|H\|_{1,2}^2)$$

Bi-convex problem,
Alternate minimization (shown to converge to a stationnary point)

$$W^{(t+1)} = \underset{W\geq 0}{\arg\min} \quad \|A - WH^{(t)}\|_F^2 + \alpha(\|W\|_F^2 + \|H^{(t)}\|_{1,2}^2)$$

Nesterov accelerated projected gradient

$$H^{(t+1)} = \underset{H\geq 0, FH^T\leq 0}{\arg\min} \quad \|A - W^{(t+1)}H\|_F^2 + \alpha(\|W^{(t+1)}\|_F^2 + \|H\|_{1,2}^2)$$

Semi explicit solution of the Lagrangian min-max problem and Nesterov

Outline

- Aggregated Functional Traits (AFTs) of fiber degradation
- Model & mathematical formulation
- Results

Inference: Selection of k

- Solve for a range of values for k (use of SVD)
- Selection according to 3 criteria

Relative reconstruction Error

$$Err(W^\star, H^\star, k) = \frac{\|A - W^\star H^\star\|_F}{\|A\|_F}$$

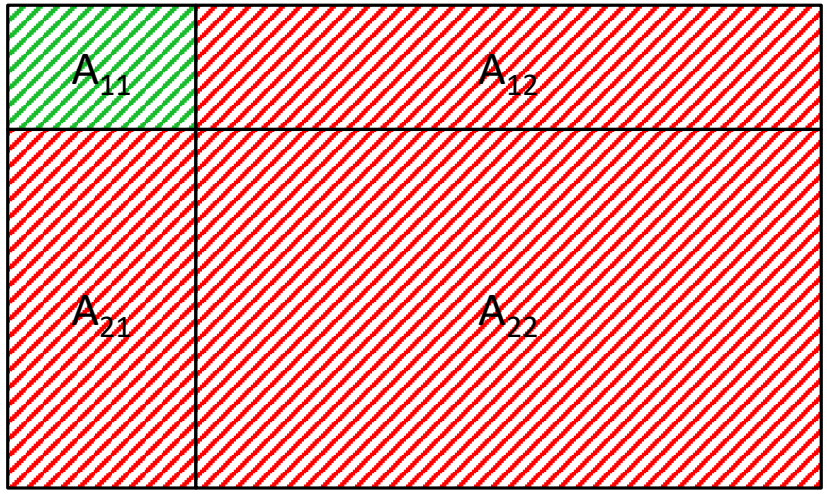
➢ Decreasing with k

➢ Optimal k = slope change

Inference: Selection of k

Bi Crossvalidation

Assess the predictive capacity



Random 10-fold split on lines and columns

Training set (red)
Validation set (green)

- Optimal k = minimum BiCV error
- AFT (columns) are not independant => no theoretical guarantee
- Not possible to use with constraints

Owen and Perry (2009)

Inference: Selection of k

Concordance index on H

Assess the robustesse of CAFT inference

Separate NMF on two-fold splits

$$Conc(k) = 1 - \frac{1}{86}\|\bar{H}_1^T \bar{H}_1 - \bar{H}_2^T \bar{H}_2\|_F$$

$A_1 \approx W_1 H_1$

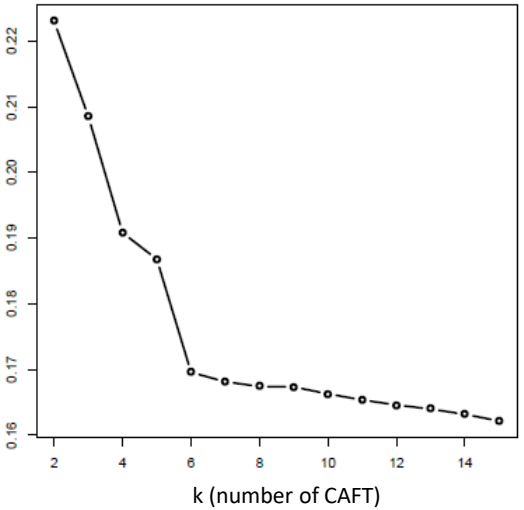
$A_2 \approx W_2 H_2$

similarity between H_1 et H_2 up to an orthogonal transformation

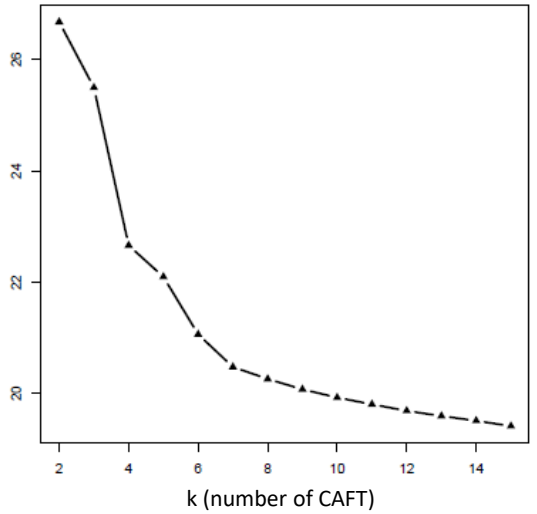
Mean over several random splits

Jiang et al. (2012)

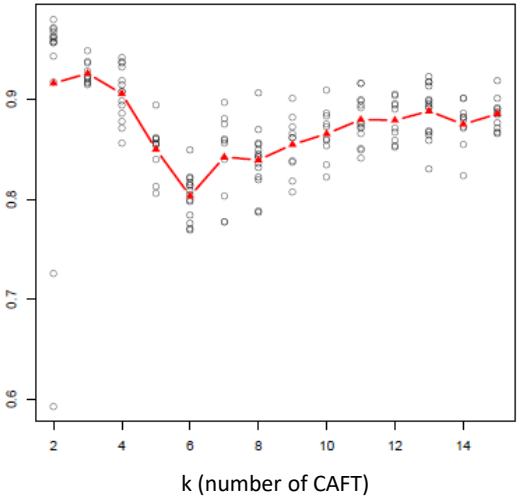
Inference: Selection of k



Relative reconstruction error



Bi-crossvalidation error

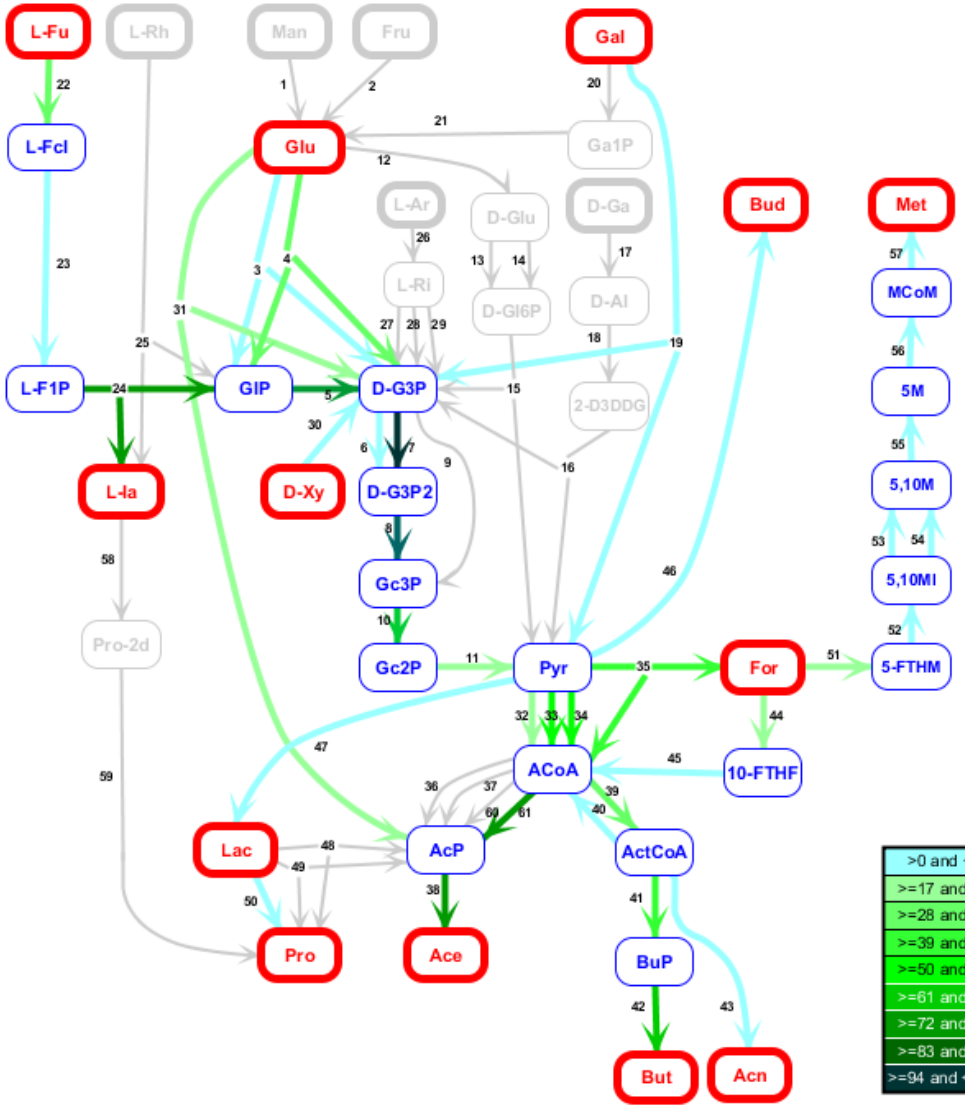


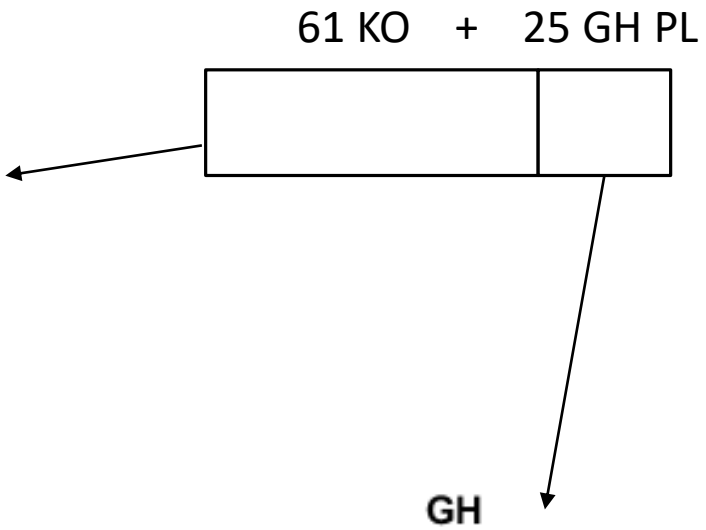
Concordance

Relative reconstruction error : k=4 or k=6
BiCV error : k=4 or k=7
Concordance : k=4 (or k=6?)

k=4

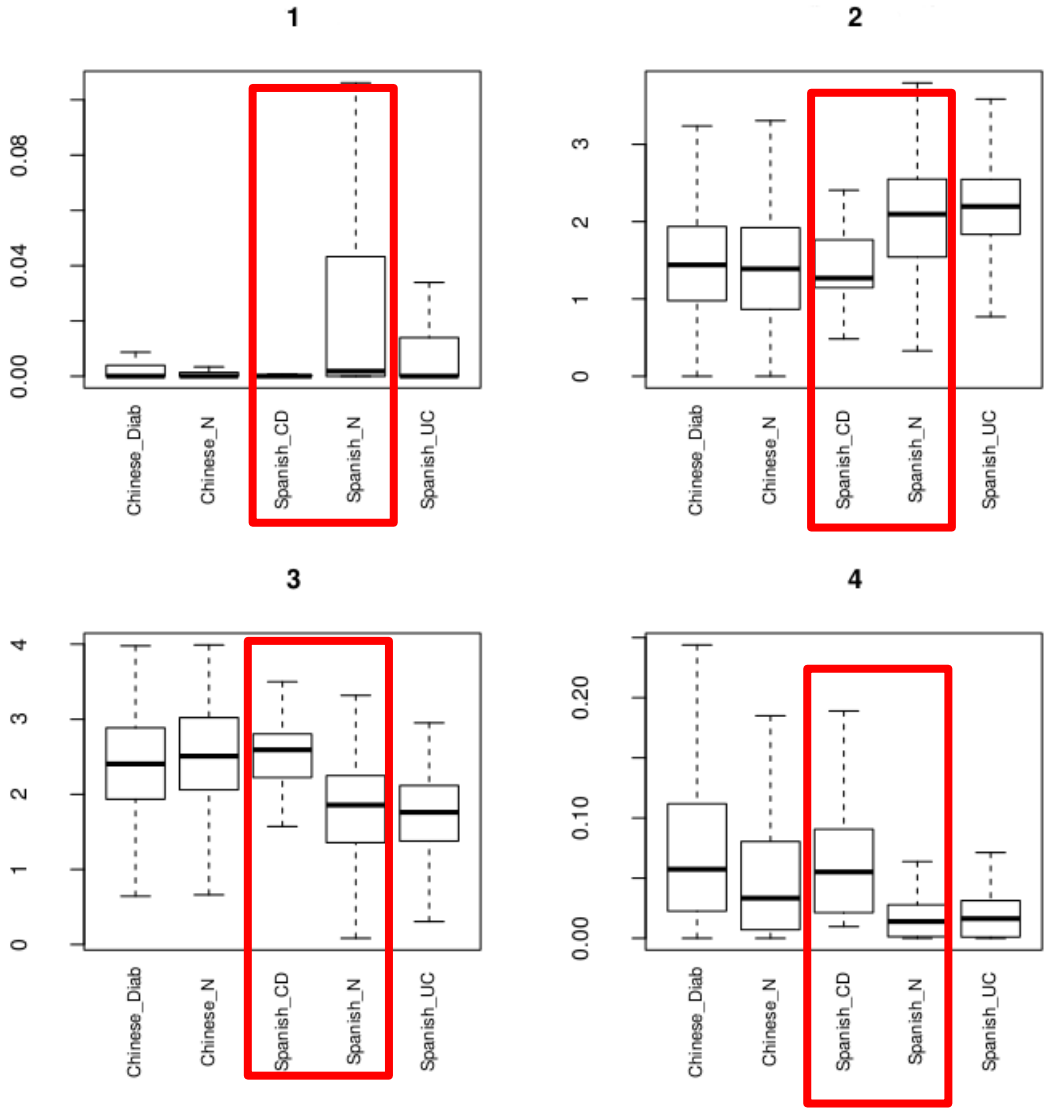
CAFT Visualization





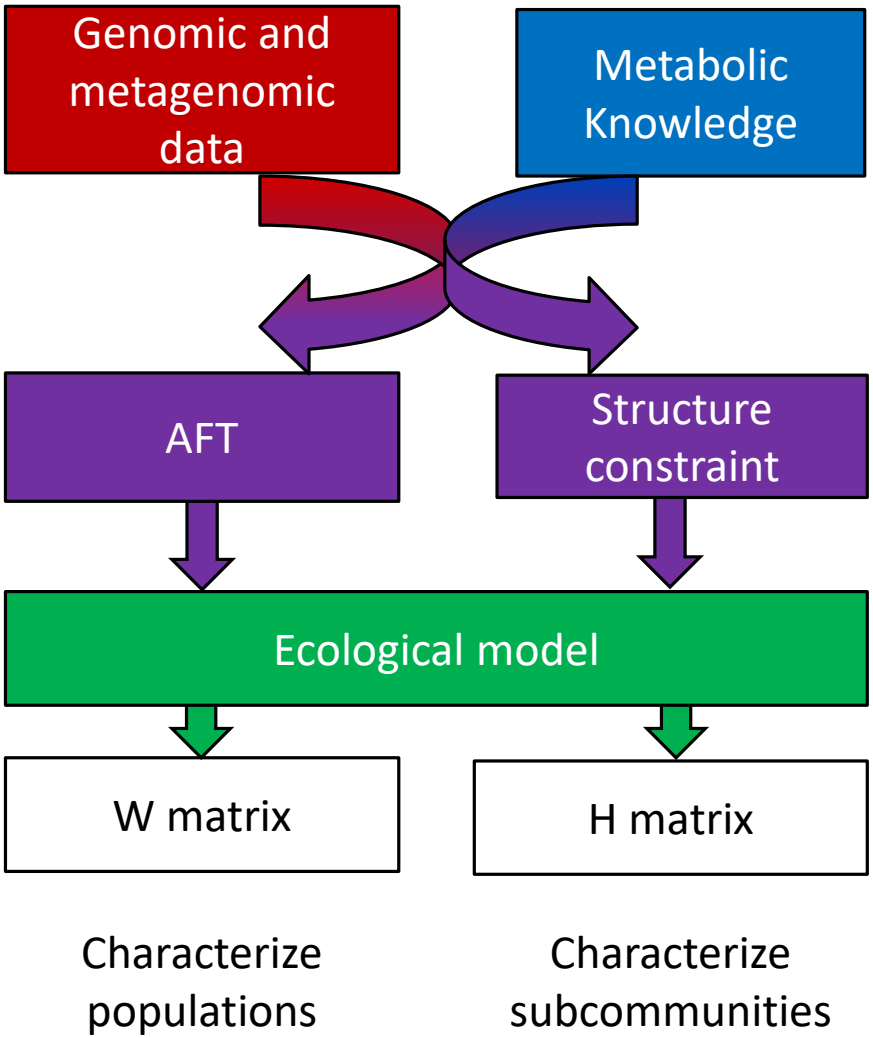
3	5	8	9	10
13	16	26	28	30
32	39	43	44	48
51	74	91	94	115
120	127	PL1	PL9	PL11

Analysis of Weights W



- Métavariables : clinical study x health status
- Abundances of 4 CAFT in each group
  - Chinese x Diabetes
  - Chinese x Healthy
  - Spanish x Crohn
  - Spanish x Healthy
  - Spanish x Ulcerative Colitis
- Difference between Spanish Crohn/Healthy
- Confirmed on external data (ongoing work)

Wrap up



Raguideau et al. Inferring Aggregated Functional Traits from Metagenomic Data Using Constrained Non-Negative Matrix Factorization: Application to Fiber Degradation in the Human Gut Microbiota.

PLoS Computational Biology, 2016.

Set of scripts and code developped in python and C, available soon.

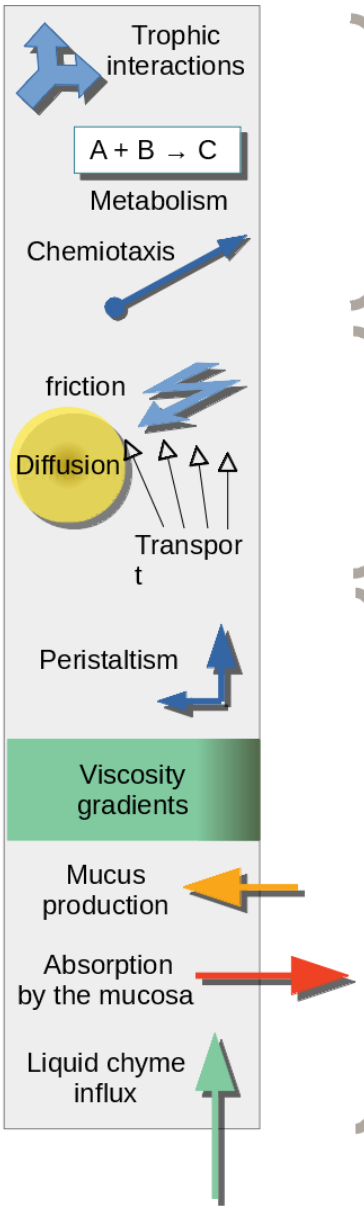
Outline

- Modelling Aggregated Functional Traits (AFT) for fiber degradation from WGS data
- Mechanistic model of fiber degradation in the human gut at the organ scale

  integrate all the knowledge we gain in a deterministic mechanistic model, started in 2009 (ODE)

  Simon Labarthe (MaIAGE), Magali Ribot (U. Orléans)
- Dynamic interaction models & parameter inference from 16S data

Mechanistic model: now PDEs



$$f = (f_i)_{i\in I_C},\ c = (c_j)_{j\in I_S},\ u,\ (\vartheta_i)_{i\in I_C},\ (\Phi_k)_{k\in I_C\cup I_S} \text{ and } p$$

$x \in \Omega$ and $t \in (0,T)$: Cylindric geometry

$$\sum_{i\in I_C} f_i = 1$$

$$\partial_t f_i - \operatorname{div}(\sigma \nabla f_i) + \operatorname{div}(u_i f_i) = F_i(t,x,f,c)$$

$$\partial_t c_j - \operatorname{div}(\theta_j \nabla c_j) + \tilde{u}\cdot\nabla c_j = G_j(t,x,f,c)$$

$$u_i = u + \vartheta_i, \text{ with } \vartheta_i = \vartheta_{i,chem} + \vartheta_{i,fr} \quad \text{and } \tilde{u} = \sum_{i\in I_C} f_i u_i$$

$$\vartheta_{i,chem} = \sum_{j\in I_S\cup I_C} \lambda_{i,j}\nabla\Phi_j \quad \text{where } -\Delta\Phi_j = c_j - \frac{1}{|\omega|}\int_\omega c_j(x,z)dx$$

$$\vartheta_{i,fr} = -\delta_{i,fr}(1-f_l)u$$

$$\nabla p - \operatorname{div}(\mu(f)D(u)) = \operatorname{div}\Big(\sum_{i\in I_C}\mu_i D(\vartheta_i)\Big)$$

$$\operatorname{div}(u) = -\operatorname{div}\Big(\sum_{i\in I_C} f_i\vartheta_i\Big),$$

+ Boundary conditions

Mechanistic model

Mixture model,
Transport, reaction, diffusion, friction
+ Stokes (fluid mechanics)
+ Keller Segel (swimming)

Solved using MAC grid and adapted numerical schemes (time splitting)

Simplified model (homogeneization) => highly efficient simulation

Next step (PhD starting soon)

- improve bacterial population structuration using previous results and build dFBA like models
- Improve boundary terms= host/microbe crosstalk, mucus/microbe interaction…

Outline

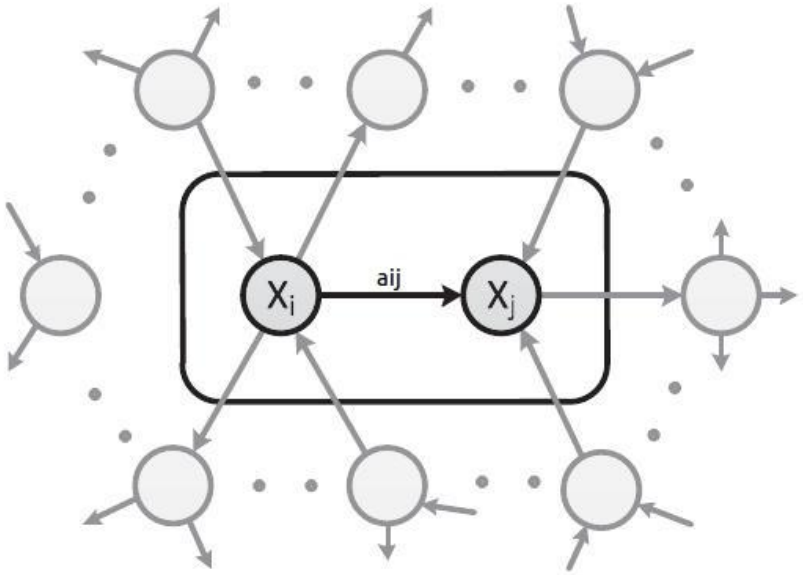
- Modelling Aggregated Functional Traits (AFT) for fiber degradation from WGS data
- Mechanistic model of fiber degradation in the human gut at the organ scale
- Dynamic interaction models & parameter inference from 16S data

  recently started, Simon Labarthe, Nicolas Brunel (U Evry) for deterministic, Generalized Lotka Volterra models

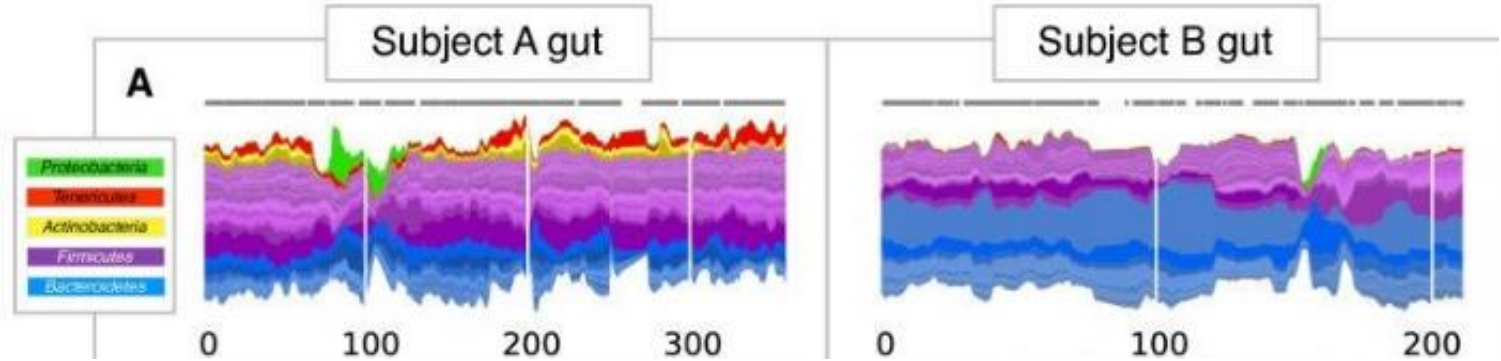
  S. Widder (U Vienna) for stochastic IBM+collab. INRA

Stochastic IBM

Individual interaction rules
and rates (interaction, death,
Immigration)
Marvov jump process



Objective= inference of parameters
& model selection
High dimension (data/model reduction)
Sparse interactions
Noisy and poorly sampled data



Merci de votre attention