



HAL
open science

Analyse de données de métagénomique fonctionnelle par NMF pour la modélisation de la dégradation des fibres par le microbiote intestinal humain.

Sébastien Raguideau

► **To cite this version:**

Sébastien Raguideau. Analyse de données de métagénomique fonctionnelle par NMF pour la modélisation de la dégradation des fibres par le microbiote intestinal humain.. Médecine humaine et pathologie. Université Paris Saclay (COMUE), 2016. Français. NNT : 2016SACLA027 . tel-02371280

HAL Id: tel-02371280

<https://theses.hal.science/tel-02371280>

Submitted on 19 Nov 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

NNT: 2016SACLA027

THÈSE

PRÉSENTÉE POUR OBTENIR LE GRADE DE DOCTEUR DÉLIVRÉ PAR
L'UNIVERSITÉ PARIS-SACLAY

PRÉPARÉE À

L'INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
(AGROPARISTECH)

ECOLE DOCTORALE N°581

ABIÈS : AGRICULTURE, ALIMENTATION, BIOLOGIE, ENVIRONNEMENT ET SANTÉ
Spécialité : Mathématiques appliquées

Par

SÉBASTIEN RAGUIDEAU

ANALYSE DE DONNÉES DE MÉTAGÉNOMIQUE FONCTIONNELLE PAR NMF POUR LA MODÉLISATION DE LA DÉGRADATION DES FIBRES PAR LE MICROBIOTE INTESTINAL HUMAIN.

Soutenue publiquement le 6 Décembre 2016 à Jouy-en-josas devant le jury composé de :

Marie-Laure Martin-Magnette	Directeur de Recherche INRA	Présidente du Jury
Hidde de Jong	Directeur de recherche INRIA	Rapporteur
Guy Perriere	Directeur de recherche CNRS	Rapporteur
Théodore Bouchez	Ingénieur IRSTEA	Examineur
Béatrice Laroche	Directeur de Recherche INRA	Directrice de thèse
Marion Leclerc	Chargé de Recherche INRA	Co-directrice de thèse
Sandra Plancade	Chargé de recherche, Inra	Co-directrice de thèse

Table des matières

1	Microbiote intestinal humain et métagénomique	10
1.1	Caractérisation du microbiote intestinal humain	11
1.1.1	Du système digestif humain au microbiote	11
1.1.2	Le microbiote intestinal humain	14
1.1.3	Dynamique du microbiote intestinal humain	19
1.2	Quels types de données	22
1.2.1	Descriptions taxonomiques du microbiote intestinal humain	22
1.2.2	Descriptions génétiques du microbiote intestinal humain	22
1.2.3	Projets de métagénomique sur le microbiote humain	24
1.3	Quelle utilisation des données métagénomiques	25
1.3.1	Clusterisation taxonomique des individus	25
1.3.2	Reconstruction du génome d'espèces non cultivées	26
1.3.3	Mise en évidence des spécificités fonctionnelles des microbiotes humains .	28
1.4	Présentation des données métagénomiques utilisées dans ce travail de thèse . . .	29
1.4.1	Integrated Gene Catalog (IGC)	29
1.4.2	Abondances des gènes de l'IGC pour 1423 échantillons métagénomiques .	30
2	Modélisation biologique de la dégradation des fibres	32
2.1	Comment décrire la dégradation des fibres par l'éco-système du microbiote intestinal humain	33
2.1.1	Quelques notions d'écologie	33
2.1.2	Une description fonctionnelle plutôt que taxonomique	37
2.2	Sélection de traits fonctionnels de la dégradation anaérobie des fibres par le microbiote intestinal humain	38
2.2.1	Hydrolyse des polysaccharides	39
2.2.2	Sélection des marqueurs de la métabolisation des oses	42

2.2.2.1	Le catabolisme des monosaccharides en condition anaérobie . . .	42
2.2.2.2	Principe de sélection des marqueurs fonctionnels	49
2.2.2.3	Voies du catabolisme des sucres simples provenant de la dégra- dation des fibres	50
2.2.2.4	Glycolyse anaérobie	61
2.2.2.5	Du pyruvate aux accepteurs finaux d'électrons	65
2.2.2.6	Réutilisation des produits de fermentation	72
2.2.2.7	Graphe topologique des marqueurs fonctionnels sélectionnés . .	77
2.3	Constitution de la matrice des abondances des TFA	84
3	Modélisation Mathématique	86
3.1	Un modèle de Combinaisons de Trait Fonctionnels Agrégés (CTFA)	87
3.2	Prendre en compte l'information biologique pour la construction des CTFA . . .	90
3.2.1	Transformer une connaissance sur la structure des génomes en une contrainte algébrique sur les CTFA	90
3.2.2	Transformer une connaissance métabolique en contrainte algébrique . . .	91
3.3	Construction pratique de la contrainte de structure liée au métabolisme	94
3.3.1	Exploitation du graphe topologiques associé aux 61 KO	94
3.3.2	Étude de 190 génomes d'espèces prévalentes dans le microbiote intestinal humain afin de calibrer les contraintes	96
4	Factorisation en Matrices non-Négatives	101
4.1	Divergence matricielle	103
4.1.1	Divergences usuelles en NMF	104
4.1.1.1	Divergence associée à la norme de Frobenius	104
4.1.1.2	La divergence de Kullback-Leibler généralisée	105
4.1.2	Avantages et inconvénients de ces deux divergences	105
4.1.3	Choix d'implémentation	106
4.2	Terme de pénalisation	107
4.2.1	Pénalisations usuelles en NMF	107
4.2.2	Choix d'implémentation	108
4.2.3	Étude des constantes de pénalisation	108
4.3	Propriétés du problème d'optimisation de NMF	111
4.3.1	Le problème général	111
4.3.1.1	Point stationnaire et minimum local	112
4.3.1.2	Propriétés de la fonction de coût	114

4.3.1.3	Minimisation des problèmes bi-convexes	115
4.3.2	Particularité de la NMF sous contrainte	116
4.4	Résolution du problème de minimisation	117
4.4.1	Etat de l'art	117
4.4.1.1	Quelques algorithmes à résolution partielle	118
4.4.1.2	Quelques algorithmes de minimisation alternée	120
4.4.1.3	Gradient proximal accéléré	124
4.4.2	Implémentation	124
4.5	Détermination des paramètres du problème de NMF	127
4.5.1	Méthodes de sélection de la dimension k dans la littérature	127
4.5.1.1	Stabilité numérique	127
4.5.1.2	Validation Croisée adaptée à la NMF : Bi-Validation-Croisée	129
4.5.1.3	Autres méthodes	130
4.5.2	Sélections de α et k	130
4.5.2.1	Critères choisis	130
4.5.2.2	Démarche globale	131
4.5.2.3	Paramètres sélectionnés	132
5	Résultats de l'inférence : analyse des matrices W et H	137
5.1	La NMF, une approche de réduction de dimension	138
5.1.1	Interprétation géométrique des résultats	138
5.1.2	Conservation de l'information de A	143
5.2	Analyse des CTFA	145
5.2.1	Présentation générale	145
5.2.2	Comparaison des CTFA	153
5.3	Analyse des abondances des CTFA	159
5.3.1	Abondances des CTFA selon les méta-variables	160
5.3.2	Validation de la discrimination de la maladie de Crohn	165
	Perspectives	170
	Annexes	174
A	Valorisation scientifiques	175
B	Abondance des CTFA pour chaque catégorie de métavariable	176
C	Liste des 190 génomes choisis	181
D	Article soumis à Plos Computational Biology	187

Glossaire	231
Notations	231
Acronymes	231
Bibliographie	232

Introduction

Ce travail de thèse s'inscrit dans une approche globale et ambitieuse de représentation du microbiote intestinal humain, de ses fonctions et de ses capacités métaboliques. Mon travail de thèse est restreint à la dégradation des fibres alimentaires qui représentent la source d'énergie principale du microbiote intestinal humain et conditionnent de ce fait son organisation. Il s'agit d'un travail multidisciplinaire mettant en relation des sujets pointus de bioinformatique, d'optimisation et de microbiologie.

Des modèles de la dégradation des fibres par l'écosystème intestinal ont été proposés (voir Widder et al. (2016) pour un inventaire), comme le modèle de Muñoz-Tamayo (2012). Ces modèles représentent la dégradation des fibres alimentaires, les concentrations des principaux substrats (acides gras à chaînes courtes, sucres, gaz) et l'abondance de populations microbiennes par le biais d'un modèle dynamique à compartiments. Ils sont construits autour d'informations bibliographiques, d'expertise biologique et de données provenant de culture *in vitro* de microorganismes isolés. Ils représentent un nombre limité de micro-organismes typiques ou de populations de micro-organismes seulement définis par leurs fonctions métaboliques.

Mon travail de thèse se situe dans une approche différente dont l'ambition est d'utiliser des données métagénomiques pour rechercher et décrire des assemblages de micro-organismes qui varient conjointement dans différents échantillons métagénomiques. Il s'agit d'inférer à partir de données d'abondances de gènes, la structure de populations fonctionnelles composant le microbiote ; ce qui complète les approches précédentes.

Le microbiote intestinal humain, longtemps "oublié", est désormais considéré comme un organe de par son impact systémique et les nombreuses fonctions qu'il assure. Il est présenté dans le Chapitre 1 ainsi que les enjeux et les travaux associés.

Le Chapitre 2 présente les fondements biologiques et biochimiques de mon approche. J'y introduis des notions nécessaires à la compréhension de l'écologie du microbiote et décris la biochimie de la métabolisation des fibres alimentaires en milieu anaérobie. Le processus de sélection de marqueurs fonctionnels de la métabolisation des fibres est exposé ainsi que la

constitution du jeu de données sur lequel repose toute la suite de mon travail.

Le Chapitre 3 présente le modèle mathématique que j'ai construit pour représenter l'écosystème microbien à différentes échelles. J'y introduis une vision écologique de la dégradation des fibres par le microbiote et propose une contribution originale qui est l'introduction de contraintes de structure à l'échelle de l'écosystème provenant de connaissances métaboliques à l'échelle des génomes des microorganismes qui le composent.

Le Chapitre 4 est consacré à la formulation du problème d'inférence du modèle présenté au Chapitre 3 et aux outils et techniques mobilisés pour sa résolution. Je présente plus particulièrement les techniques de factorisation en matrices non-négatives sur lesquelles je me suis appuyé, ainsi que leur adaptation à mon problème.

Dans le Chapitre 5, j'expose et décrit les résultats obtenus en appliquant l'approche développée aux Chapitres 3 et 4. La pertinence des résultats est validée sur un nouveau jeu de donnée.

Enfin, j'esquisse des perspectives de recherche dans une partie de conclusion.

Chapitre 1

Microbiote intestinal humain et metagénomique

1.1 Caractérisation du microbiote intestinal humain

1.1.1 Du système digestif humain au microbiote

Digestion. La digestion chez l'homme se déroule principalement au niveau de l'estomac et de l'intestin grêle. Les aliments sont retenus dans l'estomac ce qui permet leur digestion et leur conditionnement. Le débit de sortie de l'estomac est contrôlé en rétro-action des calories du contenu, de sorte à limiter les phénomènes de pics de glycémie (McHugh and Moran (1979)). Le brassage de l'estomac ainsi que l'action filtrante du pyllore, organe régulant le débit de sortie de l'estomac, permettent d'obtenir une substance homogène et de granularité relativement fine (0.5 mm) appelée le chyme. Dès sa sortie de l'estomac le pH du chyme est neutralisé par l'action commune des sucs pancréatiques et de la bile dans le duodénum. Les sécrétions pancréatiques sont majoritairement responsables de la digestion en permettant de métaboliser les acides gras, les protéines et les polysaccharides. Les sécrétions biliaires émulsifient les graisses, les rendant accessibles aux enzymes. Le transit au niveau de la lumière intestinale, c'est-à-dire la partie intérieure et creuse, se fait par contraction périodique des muscles entourant l'intestin aussi nommé péristaltisme.

Résidus de digestion. Durant ce transit, eau et nutriments sont absorbés par l'épithélium. Cette absorption n'est pas complète car l'homme ne possède pas d'enzyme permettant de dégrader toutes les graisses et tous les polysaccharides. En effet, les fibres dites alimentaires sont des polysaccharides non-digestibles originaire principalement des parois ou du cytoplasme de cellules végétales. En d'autres termes ce sont des polymères de sucres simples autrement appelés oses, tels que le glucose, le mannose, le fucose, le xylose et d'autres, reliés selon des liaisons chimiques nécessitant un outillage enzymatique spécifique. Les amidons résistants sont aussi des fibres alimentaires, ce sont des polymères de glucose qui peuvent arriver presque intacts en fin d'intestin (ileum). Leur digestion est rendue difficile voire inexistante par leur structures particulières ou leur densité. D'autres nutriments digestibles peuvent arriver dans le côlon sans être absorbés, si l'absorption par la paroi intestinale n'est pas finie ou si elle est rendue impossible par la présence d'une matrice non digestible rendant les nutriments inaccessibles (exemple : graines). Tous les jours, le chyme parvenant jusqu'au cæcum est composé de : entre 12 et 18 grammes de protéines (Davila et al. (2013)), entre 21 et 95 grammes de polysides confondus (Egert et al. (2006)) et d'environ 7% des triglycérides ingérées (Gabert et al. (2011)).

Gradients biotiques et abiotiques. Ces résidus constituent les sources d'énergie consommées par le microbiote intestinal humain. Il est cependant trompeur de considérer le côlon comme seul lieu de colonisation et d'activité microbienne. Par exemple, de part son pH très faible, l'estomac a été longtemps considéré comme stérile, mais nous savons désormais que ce n'est pas le cas (Bik et al. (2006)). Bien que la colonisation existe dans tout le système digestif, la concentration ou biomasse bactérienne augmente drastiquement entre l'iléon et le côlon où elle atteint 10^{11} bactéries par gramme de contenu. Cette variation quantitative est aussi qualitative. Le pH et le potentiel d'oxydoréduction sont des facteurs abiotiques clés dans l'établissement et la persistance des communautés microbiennes. On observe ces gradients de l'estomac jusqu'au côlon (1.1). La variation d'oxygène s'explique par sa consommation par des bactéries anaérobies facultatives dans la partie proximale du système digestif. En effet l'oxygène permet l'utilisation de voies métaboliques énergétiquement très efficaces ce qui explique son épuisement rapide. D'un autre côté, l'absence d'oxygène dans la partie distale du système digestif permet le développement de bactéries anaérobies obligatoires qui ne peuvent survivre en présence d'oxygène. Le gradient de pH joue un rôle sélectif similaire puisque des espèces résistantes aux pH faibles sont sélectionnées vers l'estomac, tandis que les espèces microbiennes du côlon ne peuvent survivre dans de telles conditions. En conclusion, les espèces microbiennes sont distinctes entre intestin grêle et côlon.

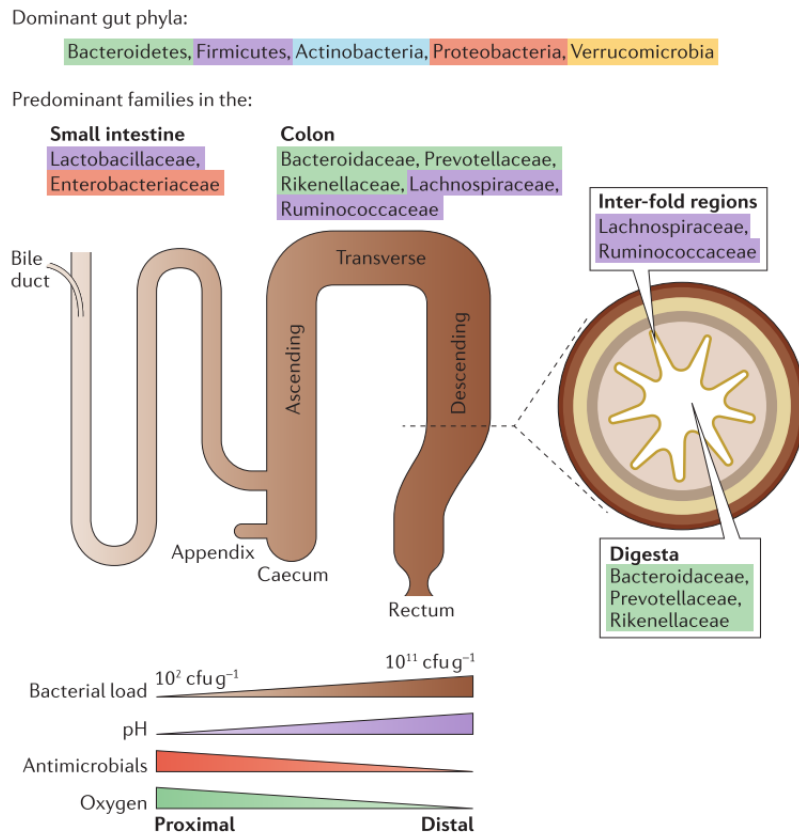


FIGURE 1.1: Évolution schématique du pH, de l'oxygène et répartition des phyla le long de l'intestin humain. "Antimicrobials" fait référence principalement au caractère anti-microbien de la bile dont l'action s'atténue à mesure que l'on s'éloigne du jejunum. L'unité cfu est l'abréviation de "colony-forming unit" qui est une mesure de quantité d'organismes microbiens viables. Tiré de Donaldson et al. (2015)

Colonisation épithéliale. On retrouve ainsi une colonisation bactérienne tout au long du système digestif, que ce soit par la colonisation du chyme ou du mucus l'épithélial. La colonisation épithéliale se fait au niveau du mucus recouvrant les parois du système digestif de l'œsophage au côlon. Il est principalement composé de glycoprotéines appelées mucines qui possèdent un squelette principal protéique ramifié par des branches latérales de polysaccharides. Les mucines constituent indépendamment du chyme une ressource énergétique qui est exploitée par de nombreux micro-organismes, une activité de dégradation des mucines a par exemple été mise en évidence pour *Bifidobacterium* (Ruas-Madiedo et al. (2008)). De manière similaire au système digestif, la structuration de la paroi intestinale est différenciée. En effet le mucus est présent au niveau de l'épithélium sous la forme de bi-couche (Johansson et al. (2011)). La pre-

mière et la plus proche de la paroi intestinale est une couche dense et protectrice, sa colonisation est difficile et souvent plutôt observée dans le cas de pathogène. La seconde couche recouvre la première et est en contact direct avec la lumière intestinale. Moins dense, plus hydratée et lâche, cette couche de mucus permet de lubrifier le passage du chyme et tient lieu de support de la colonisation intestinale. De même, les gradients abiotiques d'oxygène ne sont pas seulement longitudinaux mais aussi radiaux (voir Figure 1.2). L'oxygène sanguin diffuse à travers la paroi intestinale et en conséquence module localement les espèces microbiennes. On trouve ainsi des organismes anaérobies facultatifs colonisant le mucus épithélial (Albenberg et al. (2014)).

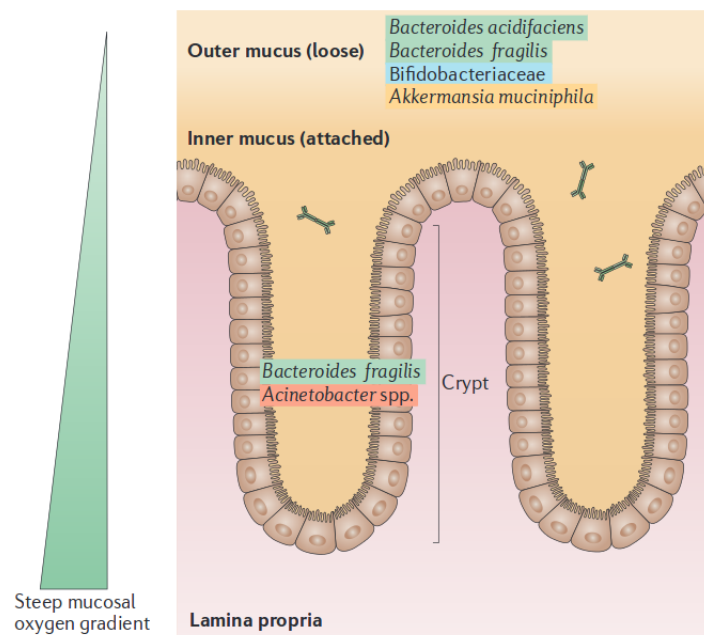


FIGURE 1.2: Gradient d'oxygène radial le long de la paroi du côlon humain et colonisation du mucus. Tiré de Donaldson et al. (2015)

1.1.2 Le microbiote intestinal humain

Répartition taxonomique. Le microbiote intestinal humain est constitué principalement de bactéries, on y trouve cependant aussi des archaea, des phages et des fungi. Environ 90% des bactéries sont représentés par deux phyla, les Bacteroidetes et les Firmicutes (Qin et al. (2010)). La diversité taxonomique inter-individuelle est élevée. Plusieurs études ont caractérisé la distribution taxonomique du microbiote intestinal humain retrouvant une structuration d'abondance similaire qui comporte à la fois un nombre limité de genres bactériens majoritaires

ainsi qu'une très grande diversité de genres faiblement abondants (Tap et al., 2009). Cette observation se retrouve à différents niveaux de description taxonomique (phylum, genre, OTU). Les observations de l'étude d'Arumugam et al. (2011) sont présentées en Figure 1.3.

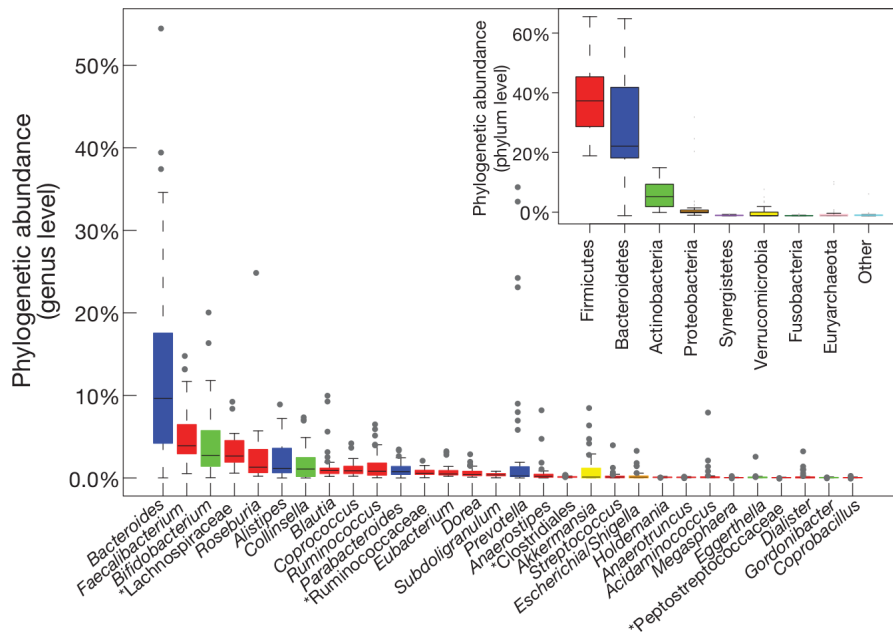


FIGURE 1.3: Genres prépondérants du microbiote intestinal humain. L'encart présente l'abondance par phylum. Tiré d'Arumugam et al. (2011)

La majorité des micro-organismes est présente en faible abondance cependant les abondances évoluent en fonction des repas ou autres perturbations extérieures. On pourrait supposer que l'abondance des micro-organismes est un bon indicateur de leur participation à l'activité métabolique totale du microbiote, mais des contre-exemples ont été observés comme l'archaea *Methanobrevibacter smithii* capable d'exploiter H_2 et CO_2 en générant du méthane, ce qui impacte effectivement le potentiel d'oxydoréduction.

Un organe oublié. La force principale du microbiote intestinal est la dégradation des polysaccharides non-digestibles qu'il transforme en Acides Gras à Chaine Courtes (AGCC), et qui sont ensuite réabsorbés par l'épithélium. Énergétiquement la métabolisation des AGCC par l'hôte lui apporte environ 50 % de l'énergie fournie par la digestion totale du polysaccharide. De plus, l'interaction des AGCC avec l'épithélium intestinal encourage sa croissance et sa différenciation (Cummings and Macfarlane (1997)). Il a été démontré que le Butanoate (AGCC sécrété par le microbiote) prévient l'apparition des cancers colo-rectaux Scharlau et al. (2009).

Les vitamines K et B sont produites par cet écosystème et réabsorbées par l'hôte. Enfin, le microbiote participe à la maturation du système immunitaire humain (Mazmanian et al. (2005)). Quelques fonctions supplémentaires de protection et de structure sont représentées en Figure 1.4. De plus, certains membres du microbiote intestinal humain (*Bacteroides dorei*) synthétisent des acides biliaires et le coprostanol dont le cholestérol est un précurseur, réduit de façon effective le cholestérol absorbé.

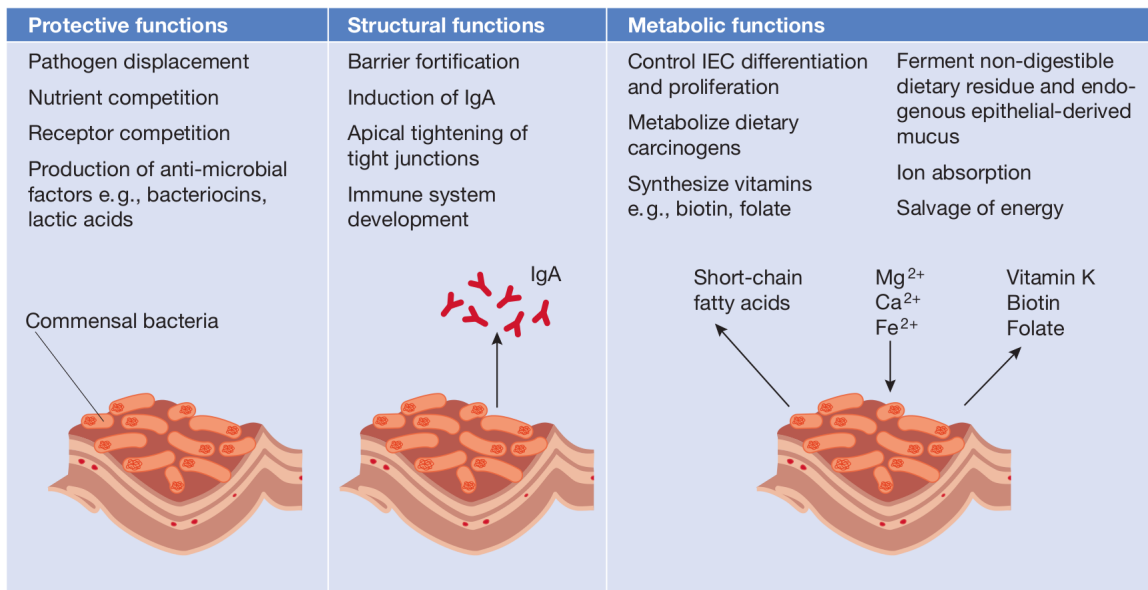


FIGURE 1.4: Différentes fonctions assurées par le microbiote intestinal humain. IEC est une abréviation pour Intestinal Epithelial Cell. IgA fait référence à l'immunoglobuline A. Tiré d'O'Hara and Shanahan (2006)

Le microbiote intestinal humain peut être considéré comme un organe à part entière de part les nombreux services qu'il rend et de part son impact systémique.

Pathologies en lien avec le microbiote intestinal. Les perturbations de cet écosystème ont des conséquences pour l'hôte. L'étude et la compréhension du microbiote prend alors une importance grandissante à mesure que l'observation des corrélations entre sa structuration et les pathologies humaines augmente.

Les Maladies Inflammatoires Chroniques Intestinales (MICI) sont en lien direct avec le microbiote intestinal : la maladie de Crohn , la colite ulcéreuse ou le syndrome de l'intestin irritable en font partie.

La maladie de Crohn est une maladie auto-immune et chronique caractérisée par des lésions localisées principalement au niveau du système digestif et alternant des phases de poussées et de rémission. Aucun traitement ne permet la rémission définitive, elle se traite par changement de régime alimentaire, immunothérapie et ablation chirurgicale des lésions. Concernant le microbiote intestinal, on constate des profils taxonomiques atypiques et notamment une perte de diversité importante en particulier pour le phylum Firmicutes (Lepage et al. (2013)). De même on constate une déficience de l'espèce *Faecalibacterium prausnitzii* identifiée comme anti-inflammatoire notamment pour sa production de Butanoate (Sokol et al. (2008)). La colite ulcéreuse est très similaire à la maladie de Crohn dans ses symptômes mais se limite uniquement au côlon. Dans une étude comprenant des jumeaux discordants (l'un est sain, l'autre est malade), des différences importantes ont été trouvées au niveau du microbiote : la présence de bactéries aérobies, la perte de diversité ainsi qu'une perte d'interaction entre microbiote et épithélium (Lepage et al. (2011)).

Qin et al. (2010) ont travaillé conjointement sur la maladie de Crohn et la colite ulcéreuse et ont montré par analyse en composante principale que les profils taxonomiques des individus atteints forment des clusters significativement distincts des individus sains comme on peut le voir en Figure 1.5.

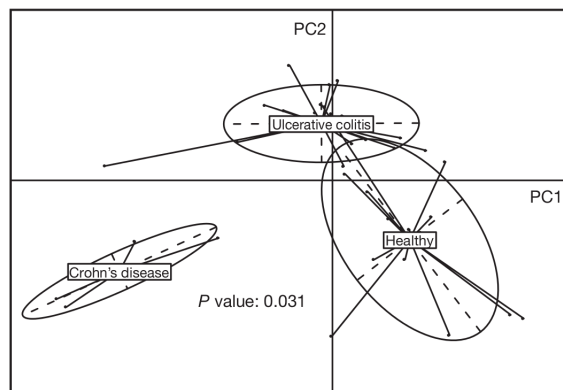


FIGURE 1.5: Clusters taxonomiques d'individus sains, atteints par la maladie de Crohn et par la colite ulcéreuse. Tiré de Qin et al. (2010)

Les MICI sont caractérisées par une réponse aberrante du système immunitaire au microbiote intestinal humain et sont d'origine multifactorielle. Il existe des facteurs de risque génétique chez l'homme, des différences au niveau du microbiote intestinal et un impact significatif du mode de vie (stress, fumeur, régime alimentaire).

De part l'impact systémique du microbiote intestinal humain, on observe des différences au niveau du microbiote associé à des pathologies humaines qui ne semblent pourtant pas en lien direct avec le microbiote intestinal. C'est le cas de l'autisme mais aussi des troubles anxieux et des états dépressifs (Foster and McVey Neufeld (2013)). Les molécules produites par le microbiote intestinal humain peuvent se retrouver dans le système circulatoire sanguin, parmi lesquelles on trouve des neurotransmetteurs comme par exemple certains acides gras à chaîne courte (Cryan and Dinan (2012)). De même, l'impact du microbiote intestinal sur la formation d'athérosclérose a été clairement mis en évidence chez la souris (Stepankova et al. (2010)).

Un faisceau d'observation met en relation la perte de diversité taxonomique avec diverses pathologies humaines en lien plus ou moins direct avec le digestif. Un exemple particulier est celui de l'obésité étudié par Le Chatelier et al. (2013) du consortium MetaHIT. Ils ont produit 207 nouveaux échantillons où sont comparés des individus obèses et non-obèses. La quantité de gènes des individus étudiés permet de les classer en deux groupes nommés respectivement LGC (Low Gene Count) et HGC (High Gene Count). Ils ont étudié les variations de gènes entre HGC et LGC et identifié 120000 gènes discriminant les deux populations. Ils identifient des covariations de groupes de gènes qui peuvent être parfois identifiés à des génomes microbiens de référence. Ils trouvent en particulier 9 espèces connues pour être inflammatoires ou anti-inflammatoires respectivement associées aux LGC et HGC. De plus, bien que les groupes LGC et HGC ne catégorisent pas strictement les individus obèses et non-obèses, une très forte prévalence d'individus obèses est observée dans le groupe LGC accompagné d'une tendance marquée à la prise de poids sur une période de 9 ans.

La Figure 1.6 tirée de Mosca et al. (2016) synthétise et source quelques exemples de pathologies associées à une perte de diversité taxonomique.

Disease	LOMD	References	Increased prevalence in western countries	References
IMMUNE DISEASES				
Crohn's Disease	+	Manichanh et al., 2006; Dey et al., 2013; Sha et al., 2013; Matsuoka and Kanai, 2015	+	Lehtinen et al., 2011
Ulcerative Colitis	+	Michail et al., 2012; Sha et al., 2013	+	Lehtinen et al., 2011
Type 1 diabetes mellitus	+	Giongo et al., 2010; De Goffau et al., 2013; Kostic et al., 2015	+	Harjutsalo et al., 2008; Andersson et al., 2014
Multiple Sclerosis	ND	Bhargava and Mowry, 2014	+	Mayr et al., 2003
Celiac Disease	+	Schippa et al., 2010	+	Lohi et al., 2007
Allergy	+	Wang et al., 2008; Ismail et al., 2012; Abrahamsson et al., 2014	+	Latvala et al., 2005; Eder et al., 2006
METABOLIC DISEASES				
Obesity	+	Turnbaugh et al., 2009	+	WHO Obesity and overweight
Type 2 diabetes mellitus	-		+	WHO Diabetes
CANCER				
Colorectal cancer	+	Ahn et al., 2013	+	European Cancer Observatory http://eco.iarc.fr/
OTHERS				
Irritable Bowel Syndrome	+	Carroll et al., 2012; Durbán et al., 2012	-	Lovell and Ford, 2012
Recurrent <i>Clostridium difficile</i> Infection	+	Chang et al., 2008	ND	
Autism	+	Kang et al., 2013	+	Atladóttir et al., 2015
Necrotising Enterocolitis	+	Stewart et al., 2013	ND	
Graft Versus Host Disease	+	Jenq et al., 2012	ND	

WHO | Diabetes WHO. Available at: <http://www.who.int/mediacentre/factsheets/fs312/en/> [Accessed September 29, 2015].
WHO | Obesity and overweight WHO. Available at: <http://www.who.int/mediacentre/factsheets/fs311/en/> [Accessed September 29, 2015].
ND, No Data.

FIGURE 1.6: Tableau regroupant les observations de faible diversité microbienne en relation avec diverses pathologies. LOMD désigne Loss Of Microbiota Diversity. Tiré de Mosca et al. (2016)

Il existerait donc un lien entre la biodiversité de l'écosystème microbien, les fonctions du microbiote, les molécules produites qui en résultent et le fonctionnement des organes humains.

1.1.3 Dynamique du microbiote intestinal humain

Le microbiote intestinal humain est un assemblage vivant et changeant, son évolution temporelle et sa résilience à diverses perturbations sont des questions importantes car elles pourraient conditionner le maintien de l'homéostasie digestive ou participer à l'apparition de pathologies.

Installation et évolution du microbiote intestinal humain. Des micro-organismes anaérobies facultatifs originaires du vagin suite à l'accouchement ou du lait maternel colonisent le système digestif initialement axénique (stérile). Ces micro-organismes consomment l'oxygène présent et établissent peu à peu un gradient d'oxygène permettant ensuite la colonisation par des bactéries anaérobies strictes. Le partage des lieux de vie, les contacts cutanés, le régime

alimentaire seraient autant de facteurs encourageant l'établissement d'un microbiote intestinal individuellement unique et diversifié mais toutefois similaire entre individus proches. Yatsunenko et al. (2012) ont étudié l'évolution de la diversité taxonomique en fonction de l'âge et des cultures, leur résultat est visible en Figure 1.7, ils constatent une diversité croissante avec l'âge, d'abord faible en bas âge puis plus importante à l'âge adulte. Dans cette étude, ils mettent aussi en évidence de très fortes différences entre les microbiotes d'individus originaires du Malawi et du Venezuela avec ceux des États-Unis. Cette observation pourrait provenir d'un mode de vie et d'un régime alimentaire complètement différents.

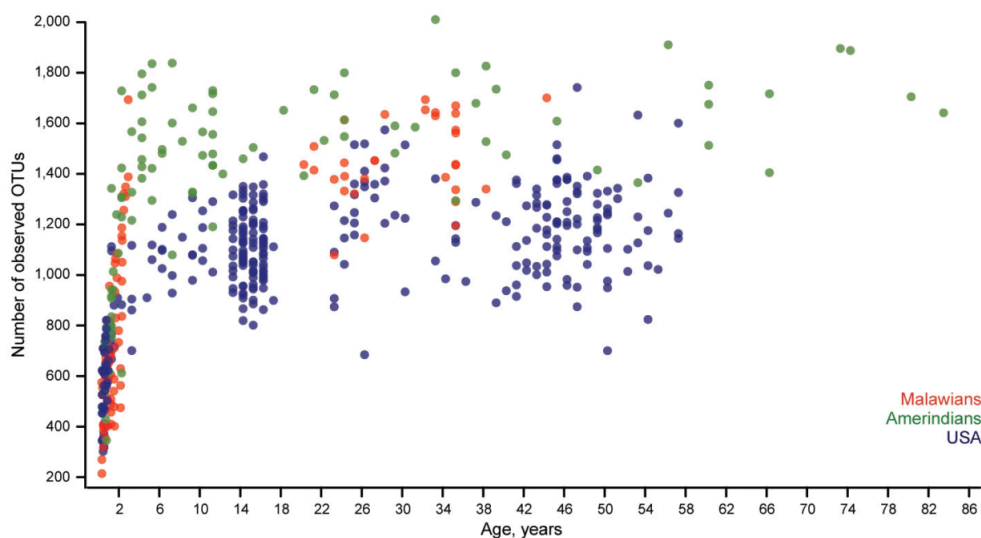


FIGURE 1.7: Distribution du nombre d'OTU en fonction de l'âge et de l'origine de l'hôte. Tiré de Yatsunenko et al. (2012)

Dans un autre registre, la spécificité du microbiote intestinal humain laisse entrevoir un nouveau mode d'identification similaire aux empreintes digitales (Eckburg et al. (2005), Franzosa et al. (2015)). Cependant la question se pose de savoir si cet écosystème est suffisamment stable temporellement. Franzosa et al. (2015) ont sélectionnés des marqueurs taxonomiques spécifiques du microbiote intestinal de 100 individus et les ont ensuite identifiés 10 mois plus tard avec un taux de réussite de 80%. Ils n'ont observé que très peu de faux positifs, et expliquent que les individus non identifiés ont perdu le marqueur spécifique pendant cet intervalle.

Impact du régime alimentaire sur le microbiote intestinal humain. Wu et al. (2011) ont exploré les variations taxonomiques du microbiote intestinal humain en corrélation avec

l'alimentation à long et court terme chez 98 volontaires sains. Ils trouvent des genres microbiens corrélés positivement ou négativement avec des catégories de nutriments. Par exemple, les Bacteroidetes et Actinobacteria sont corrélés positivement aux acides gras et négativement aux fibres alimentaires, tandis que les phyla Firmicutes et Proteobacteria se comportent de manière inverse. Dans une deuxième partie de leur étude Wu et al. (2011) on étudié l'évolution des profils taxonomiques à court terme en suivant dix individus soumis à deux régimes alimentaires spécifiques (riche en fibre ou riche en protéines et acides gras). Ils constatent un changement rapide, significatif et répétable des profils taxonomiques sous 24 heures. Ces changements significatifs mettent en évidence une plasticité du microbiote intestinal humain. Cependant, la plasticité reste limitée puisque même si les espèces microbiennes varient, les genres prédominants restent les mêmes, même lorsque le régime alimentaire leur était négativement corrélé. De même, une période de dix jours n'a pas suffi pour que les profils taxonomiques convergent vers des distributions communes aux individus partageant le même régime alimentaire. Cette étude met en évidence deux types de cinétiques d'évolution, une évolution rapide sous 24 heures en réponse à une perturbation (repas) et une évolution lente corrélée aux habitudes alimentaire individuelles sur le long terme.

D'autres études s'intéressent à l'impact du régime alimentaire sur le microbiote intestinal. En particulier, Cotillard et al. (2013) observent que les 49 individus obèses ou en surpoids de leur étude peuvent être répartis en deux groupes d'individus LGC (Low Gene Count) et HGC (High Gene Count). Au temps initial, des différences significatives d'indicateurs médicaux sont observées entre individus du groupe LGC et HGC, notamment en terme de résistance à l'insuline et niveaux de triglycérides à jeun, tous deux en défaveur du groupe LGC. Cette étude questionne l'impact d'un régime alimentaire sur le microbiote intestinal humain et sa diversité génétique. Les sujets de l'étude sont soumis à un régime alimentaire enrichi en protéines et faiblement énergétique pendant 6 semaines suivi d'un régime alimentaire de stabilisation pondérale pendant 6 semaines. Le premier résultat important est l'augmentation significative du nombre de gènes dans le groupe LGC après 6 semaines, cette observation est accompagnée de l'amélioration des indicateurs cliniques associés à l'obésité. Ils identifient 18 groupes de gènes discriminant les individus LGC et HGC qu'ils utilisent pour classifier presque parfaitement les individus HGC et LGC dans cette étude et dans celle de Le Chatelier et al. (2013).

1.2 Quels types de données

1.2.1 Descriptions taxonomiques du microbiote intestinal humain

La taxonomie est la classification hiérarchique et systématique des être vivants en groupes et catégories nommées rang taxonomiques. Les microbiens ont historiquement été établis selon leurs phénotypes, la caractérisation des parois cellulaires ou des activités métaboliques. Dans des écosystèmes complexes cette approche échoue par le nombre, la diversité et la variabilité des organismes. Plus particulièrement pour le microbiote intestinal, la difficulté première est l'isolement et la culture des micro-organismes rendant d'autant plus délicate l'identification taxonomique. Suau et al. (1999) considéraient, en estimation basse, que 60% des micro-organismes du microbiote intestinal humain sont non-cultivables. Rondon et al. (1999) s'intéressent aux mêmes questions dans le cadre des micro-organismes du sol : de la même façon seul 1% des cellules observées sont cultivables. La différence entre nombre d'organismes observés et cultivés pourrait s'expliquer par un verrou technologique concernant la culture d'espèces encore inconnues ou vivant en symbiose.

Depuis les années 90 un changement de paradigme s'est opéré dans la description du vivant par le biais des outils dit moléculaires (étude des séquences génétiques). En effet, Woese et al. (1990) proposent pour la première fois de reconstruire la taxonomie en s'inspirant des séquences génétiques des organismes et en particulier en exploitant les séquences de l'ARN ribosomique 16S (ARNr 16S). Il s'agit d'une petite sous-unité des ribosomes cellulaires, présente dans chaque organisme, qui possède des régions identiques d'un organisme à l'autre puisque nécessaires au fonctionnement du ribosome. En étudiant les différences de séquences d'ARNr 16S, il est possible d'identifier des événements de divergence génétique et de reconstruire la chronologie des différenciations des organismes vivants. L'ARNr 16S est utilisé à des fins de caractérisation taxonomique fine des micro-organismes de manière si extensive que de nombreuses espèces non cultivées ou actuellement non cultivables ne sont identifiées qu'en tant qu'OTU. Les OTU de l'anglais Operational Taxonomic Unit sont des organismes pragmatiquement et taxonomiquement définis par leur séquence d'ARNr 16S. Suau et al. (1999) ont les premiers utilisé l'ARNr 16S dans le but de caractériser taxonomiquement le microbiote intestinal humain et par ce biais identifié 76% d'organismes inconnus dans leurs échantillons.

1.2.2 Descriptions génétiques du microbiote intestinal humain

métagénomique. La génomique désigne l'étude du génome d'organismes individuels, de la même façon on définit la métagénomique comme l'étude du génome de populations d'orga-

nismes. Dans le cadre des populations bactériennes cette distinction est importante puisque l'incapacité d'isoler les micro-organismes se traduit par la nécessité d'étudier le matériel génétique de la population microbienne en entier. Rondon et al. (1999) ont proposé d'utiliser une approche métagénomique afin d'étudier les fonctions présentes dans l'écosystème cible sans passer par l'isolement et la culture d'organismes vivants.

Séquençage et assemblage. Les échantillons métagénomiques sont séquencés par Whole Genome Shotgun (WGS), cette méthode passe par une étape de fragmentation des fragments d'ADN en séquences de taille suffisamment petite pour être lue rapidement et précisément (de l'ordre de 100 bases azotées). Différentes technologies de haut-débit sont disponibles pour lire les séquences et il existe un compromis entre vitesse de séquençage, précision de lecture de la taille des séquences et prix. La Figure 1.8 présente différentes plateformes de séquençage.

Société	Roche			Illumina				Life Technologies			
Plateforme	GS Junior	454		MiSeq	HiSeq 1000	HiSeq 2000	Genome Analyzer Ix	Ion Torrent PGM	SOLID 4	SOLID 5500	SOLID 5500xl
Technologie	Titanium	FLX Titanium FLX+						Chip 314 Chip 316 Chip 318			
	Acides nucléiques (matrice)										
	Ligation adaptateurs										
Méthode d'amplification	PCR en émulsion			« Bridge PCR »				PCR en émulsion			
Méthode de séquençage	Synthèse (Pyroséquençage)			Synthèse				Ligation			
Durée de séquençage/run	10h	10h	20h	26h	8jrs	8jrs	14jrs	2h	12jrs	8jrs	8jrs
Capacité (Mb) séquençage/run	50	500	900	1500	100000	200000	95000	>10 >100 >1000	70000	80000	150000
Taille moyenne des reads	400	400	700	150+150	100+100	100+100	150+150	100 >100 >100	50+35	75+35	75+35
Coût (\$) /run	1100	6200		750	10000	20000	11500	500 750 950	8150	6100	10500
Coût machine + annexes ((K\$))	110+25	500+30		125	560	690	250	50+20	480+55	350+55	600+55
Exactitude de séquençage (%)	99	99		99,9	99,9	99,9	99,9	99	99,95	99,95	99,99

FIGURE 1.8: Plateforme de séquençage seconde génération. Synthèse datant de 2011, tiré de <http://www.biorigami.com/?p=1133>.

Une fois lues, les séquences sont appelées des reads. L'assemblage des reads utilise les recouvrements de séquences pour inférer leurs successions respectives. Les reads sont ainsi assemblés

en séquences plus longues nommées contigs. De manière générale, plus l'on dispose de reads pour un génome donné et plus l'assemblage permettra de construire des contigs de taille importante, le cas extrême étant l'obtention d'un unique contig correspondant au génome cible. Lorsque cet assemblage se fait sans l'aide de bases de données, on parle d'assemblage *de novo*. Dans le cas de la métagénomique, le séquençage donne accès à des fragments de matériel génétique originaires de multiples micro-organismes en proportions diverses. Ne pas savoir si les séquences appartiennent au même organisme rend l'assemblage *de novo* beaucoup plus complexe, et le limite à la reconstruction de contigs et non pas de génomes. On comprend que les contigs des espèces faiblement abondantes seront de qualité moindre (plus petits). Il est donc nécessaire de séquencer plus de reads dans un échantillon métagénomique si l'on veut bien reconstruire le matériel génétique des espèces faiblement abondantes.

1.2.3 Projets de métagénomique sur le microbiote humain

Les corrélations observées entre pathologies et microbiote alliées aux progrès techniques de séquençage ont rendu l'étude du microbiote humain très porteuse. Deux projets de grande envergure se détachent.

MetaHit. MetaHit (Metagenomic of the Humain Intestinal Tract) est un projet européen (budget 22 millions €) coordonné par l'INRA et regroupant 13 partenaires académiques et industriels de 2008 à 2012. L'objet d'étude a été le microbiote intestinal humain, ils ont produit 760 échantillons métagénomiques provenant d'individus Français, Espagnols, Danois et Néerlandais, pour divers états de santé (Maladie de Crohn, colite ulcéreuse, obésité...). La contribution la plus mémorable du consortium MetaHIT a été la constitution de catalogues de gènes assemblés *de novo* et annotés fonctionnellement de 3 millions de gènes (Qin et al. (2010)) puis de 9.9 millions (Li et al. (2014)). Ils ont aussi développé d'autres concepts tels que les entérotypes et les espèces métagénomiques que nous détaillons dans la section suivante.

HMP. HMP (Human Microbiome Project) est un projet (115 millions \$) du National Institutes of Health (NIH) aux États-Unis. Ils ont caractérisé le microbiote d'individus américains sains à différents sites du corps humain (bouche, peau, vagin, intestin, poumons). Ils ont caractérisés 139 échantillons intestinaux prélevés à différents temps en utilisant leurs outils d'annotation fonctionnelle HUMAnN (Huttenhower et al. (2012)). En parallèle, ils ont pour objectif ambitieux de séquencer 3000 génomes d'espèces bactériennes de référence du microbiote humain. En 2010, 178 génomes étaient séquencés (Nelson et al. (2010)), en 2012 800

génomomes avaient déjà été produits (Methé et al. (2012)) et enfin, en 2014, 1320 génomes de références étaient mis à disposition par HMP dont 457 concernant le microbiote intestinal humain (<http://www.hmpdacc.org/HMRGD/>).

1.3 Quelle utilisation des données métagénomiques

1.3.1 Clusterisation taxonomique des individus

La notion d'entérotype a été dans un premier temps proposée par le consortium MetaHit puis ré-examinée par le consortium HMP.

Des clusters de profils taxonomiques. Arumugam et al. (2011) de MetaHIT ont caractérisé le microbiote intestinal de 39 individus de différentes nationalités (Français, Danois, Espagnols, Américains et japonais). Ils trouvent que tous les métagénomomes considérés peuvent se classer taxonomiquement et fonctionnellement en trois groupes qu'ils appellent entérotypes. Ces entérotypes peuvent être caractérisés par le genre microbien prédominant qui les compose (voir Figure 1.9), en l'occurrence *Bacteroides*, *Prevotella* ou encore dans le cas du troisième entérotype, dont la composition varie en fonction des métagénomomes utilisés, *Ruminococcus*, *Blautia* ou *Lachnospiraceae*. Aucune corrélation n'a été trouvée entre entérotype et caractéristiques des individus considérés (âge, IMC, état de santé, etc...). Finalement, les auteurs expliquent les entérotypes par l'émergence répétable de communautés microbiennes construites autour de stratégies trophiques différentes et en réponse à des types d'intestin humain différents. Ils proposent notamment les trois voies métaboliques d'utilisation du dihydrogène comme origine vraisemblable des entérotypes et espèrent pouvoir utiliser l'appartenance à un entérotype pour prévoir des réponses caractéristiques à des traitements et régimes alimentaires.

Méthodologie des entérotypes. Koren et al. (2013) circonscrit l'observation d'entérotypes. Ils étudient l'impact de chaque choix méthodologique aboutissant aux résultats obtenus par Arumugam et al. (2011), notamment, le choix de la métrique de distance, le critère de choix du nombre de clusters, le niveau de description taxonomique, le choix de la région amplifiée du ARNr 16S, l'assignation des OTU, ou encore le type de données (ARNr 16S ou WGS). D'une manière générale tous les critères considérés ont un effet sur le nombre de clusters obtenus. Les auteurs ont identifié les choix méthodologiques importants dans l'inférence robuste d'entérotypes, et dans le cas du microbiote intestinal humain, préfèrent parler de distributions continues multimodales plutôt que de groupes discrets séparant les individus comme

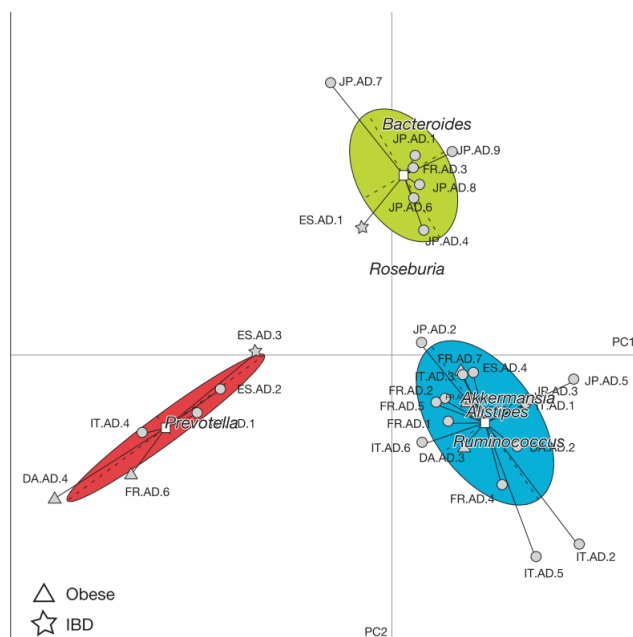


FIGURE 1.9: Analyse en composante principale des profils phylogénétiques des 39 individus de l'étude. Tiré de Arumugam et al. (2011)

des groupes sanguins, la Figure 1.10 permet d'illustrer ce point.

1.3.2 Reconstruction du génome d'espèces non cultivées

Nielsen et al. (2014) du consortium MetaHIT ont proposé une méthodologie permettant de construire des clusters de gènes s'apparentant à des génomes microbiens à partir de données WGS et sans utilisation de base de données. En exploitant un catalogue de 3.9 millions de gènes assemblés *de novo*, ils identifient les gènes corrélés entre eux et forment ainsi des CAG (Co-Abundance Gene groups). Ils trouvent 7381 CAG comprenant 68% des 3.9 millions de gènes. La distribution en taille des CAG est bi-modale (voir Figure 1.11), ils expliquent les CAG de petite taille comme provenant de phage et nomment les plus gros MGS (MetaGenomic Species) qu'ils identifient à des génomes microbiens. Dix-neuf individus échantillonnés avaient au préalable consommé des produits laitiers contenant *Bifidobacterium animalis subsp. lactis* CNCM I-2494. Le résultat phare de cette étude est la reconstruction d'une MGS contenant 95% des gènes de cette bactérie avec une similarité de 99.9%. Loin d'être une exception 44 MGS sont très proches de génomes de référence précédemment séquencés avec en moyenne 78% de couverture pour 98.4% d'identité. De plus 238 parmi les MGS sont de qualité suffisante pour être considérés comme ébauche de génome par le consortium HMP. Ils ont aussi étudié

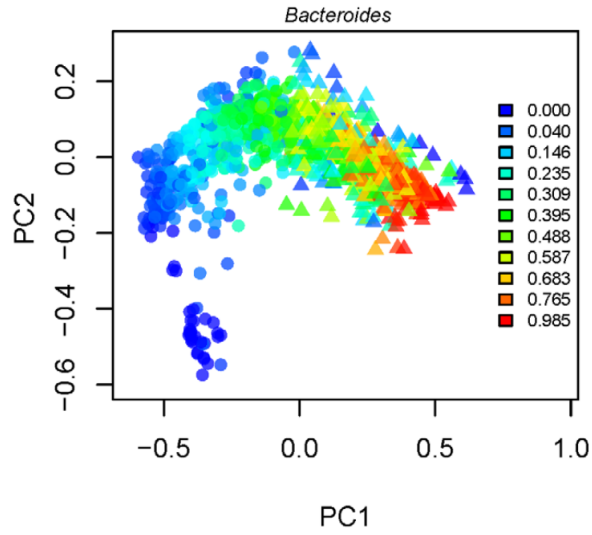


FIGURE 1.10: Analyse en composantes principales des profils des échantillons du consortium HMP. Tiré de Koren et al. (2013)

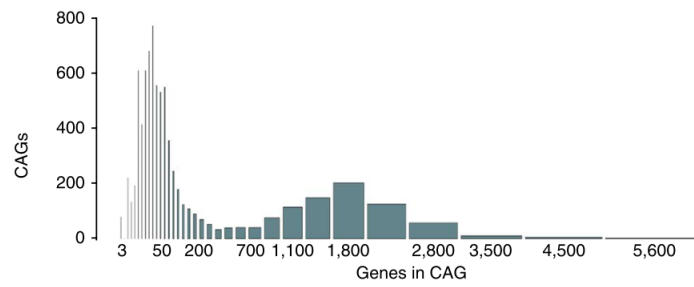


FIGURE 1.11: Distribution du nombre de gènes dans les CAG. Tiré de Nielsen et al. (2014)

l'association entre CAG et trouvent de nombreuses occurrences de MGS associées à de petits CAG qu'ils interprètent comme l'association de phages à des espèces microbiennes.

1.3.3 Mise en évidence des spécificités fonctionnelles des microbiotes humains

Le consortium HMP a proposé un processus de traitement des échantillons métagénomiques par le biais du logiciel HUMAnN (Abubucker et al. (2012)). Leur approche repose sur une quantification d'orthologies de la base de données KEGG (Kyoto Encyclopedia of Genes and Genomes, Kanehisa et al. (2009)) notée KO et sur l'identification de modules fonctionnels de KEGG. Ils comparent l'abondance de ces modules dans les différents échantillons métagénomiques du consortium HMP et identifient les particularités fonctionnelles des différents microbiotes du corps humain échantillonnés. La Figure 1.12 montre comment les échantillons fécaux se détachent nettement de par le niveau global d'abondance des 5 modules fonctionnels choisis et par leur distribution similaire inter-échantillons.

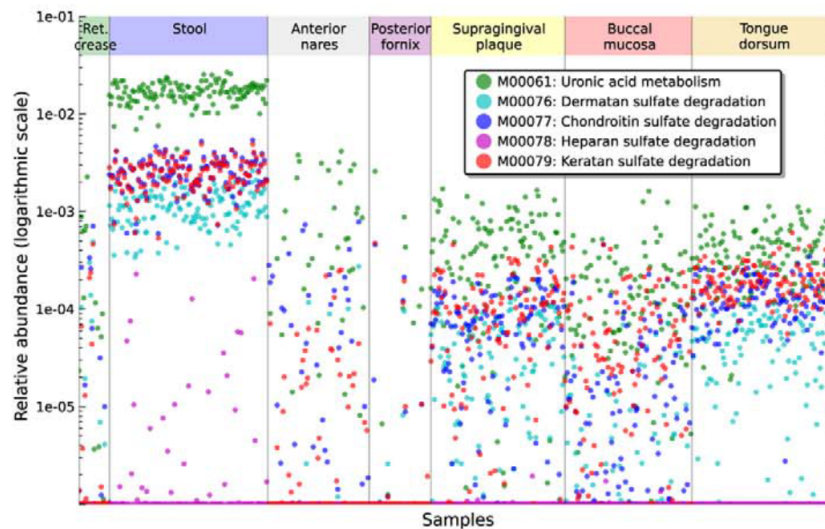


FIGURE 1.12: Distribution de 5 modules fonctionnels de KEGG dans les échantillons métagénomiques de 7 sites du corps humain. Tiré de Abubucker et al. (2012)

1.4 Présentation des données métagénomiques utilisées dans ce travail de thèse

1.4.1 Integrated Gene Catalog (IGC)

Origine du catalogue. Nous exploitons le catalogue produit par Li et al. (2014). Il a été construit à partir de 1263 échantillons métagénomiques :

- 245 nouveaux échantillons proviennent de cette étude, aucune métadonnées n’ont été fournies.
- 124 échantillons proviennent de Qin et al. (2010), parmi les individus échantillonnés, 21 sont atteints de colite ulcéreuse et 4 de la maladie de Crohn.
- 207 échantillons proviennent de Le Chatelier et al. (2013), les individus échantillonnés sont obèses ou non-obèses.
- 180 échantillons proviennent de Nielsen et al. (2014), 106 échantillons proviennent d’individus atteints de colite ulcéreuse et 17 de la maladie de Crohn.
- 139 échantillons proviennent de Methé et al. (2012).
- 368 échantillons proviennent de Qin et al. (2012), parmi lesquels 183 proviennent d’individus atteints de diabète de type 2.

L’IGC est un catalogue d’exactly 9.879896 gènes non redondants construits à partir de 511 génomes microbiens de référence du microbiote intestinal et par assemblage *de novo* de 1263 échantillons métagénomiques décrits ci-dessus. Il s’agit du troisième catalogue du consortium MetaHIT, le premier n’exploitant que 124 échantillons pour 3.3 millions de gènes (Qin et al. (2010)) et le deuxième 396 échantillons pour 3.9 millions de gènes (Nielsen et al. (2014)). Comparativement aux premiers catalogues et aux assemblages du consortium HMP (Méthé et al. (2012)), l’IGC semble être de meilleure qualité puisque les gènes obtenus sont en moyenne plus grands et puisque 10% de reads supplémentaires peuvent être associées aux gènes de l’IGC (comparativement à Qin et al. (2010)).

Construction du catalogue. La constitution du catalogue a été faite par le biais de MOCAT (Kultima et al. (2012)), un logiciel développé par MetaHIT qui associe différents outils spécialisés afin de proposer un processus de traitement standardisé prenant en entrée des échantillons métagénomiques et produisant en sortie un catalogue de gènes assemblés *de novo*. La première étape est la sélection des reads métagénomiques sur des critères de longueur et qualité minimale. La deuxième étape consiste à aligner les reads avec des bases de données, soit pour identifier du matériel génétique non pertinent (homme/animal) soit pour obtenir l’abondance

de certains gènes connus dans les échantillons métagénomiques. Cette sélection est suivie d'une étape d'assemblage, les séquences résultants de l'assemblage sont appelées contigs. Les contigs suivent alors une étape de révision où l'alignement des reads initiaux sur les contigs permet de corriger des erreurs de bases azotées, de suppression ou de délétions. Enfin, la dernière étape consiste à prédire la présence de gènes à partir des contigs.

Une liste de gènes potentiels est alors obtenue dont on peut évaluer l'abondance dans chaque échantillon.

Annotation fonctionnelle. Il est possible d'annoter une partie de ces gènes en utilisant des bases de données afin de connaître leur fonction ou leur organisme d'origine. Cette information est cependant très dépendante de la qualité des bases de données. Sur ce catalogue, deux bases de données d'orthologie ont été utilisées, d'une part 6980 orthologies de KEGG ont été trouvées et d'autre part 36489 orthologies d'eggNOG ont été observées. Ces deux bases de données diffèrent dans la méthode de construction de leurs groupes d'orthologie. Tandis que KEGG construit manuellement des groupes d'orthologie assurant ainsi leur pertinence (Kanehisa et al. (2009)), eggNOG utilise des algorithmes automatiques permettant d'inférer des groupes orthologies (Powell et al. (2012)). Ces nombres d'orthologie décrivent la réunion des échantillons métagénomiques considérés. La diversité fonctionnelle individuelle est plus faible, de l'ordre de 13000 orthologies putatives d'après Qin et al. (2010). Tous les gènes du catalogue ne sont pas annotés et même les 36489 orthologies d'eggNOG ne représentent que 69.3% des 9.9 millions de gènes. Cette annotation nous permet tout de même de se faire une idée générale des fonctions prépondérantes présentes dans le microbiote intestinal comme le montre la Figure 1.13.

1.4.2 Abondances des gènes de l'IGC pour 1423 échantillons métagénomiques

Nous utilisons les 1263 échantillons métagénomiques utilisés pour construire l'IGC de pair avec 156 échantillons produits par Cotillard et al. (2013). Les reads de ces 1423 échantillons métagénomiques sont alignés sur les gènes de l'IGC au moyen de SOAP2 (Li et al. (2009)). Cette première étape permet d'obtenir des comptages de gènes calculés selon le nombre et la taille des reads alignés aux gènes cibles avec un critère de 95% d'identité. L'abondance d'un gène dans un échantillon métagénomique est calculé en prenant le comptage dudit gène dans l'échantillon normalisé par la taille du gène en nombre de nucléotides.

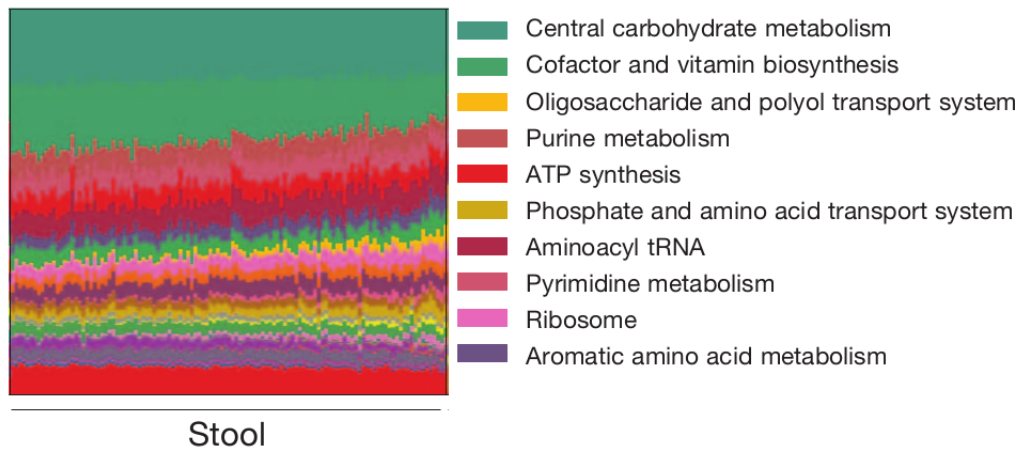


FIGURE 1.13: Visualisation des voies métaboliques les plus abondantes, sans surprise la métabolisation des fibres (carbohydrates) occupe une place importante. Tiré de Huttenhower et al. (2012)

Nous exploitons ces 9.9 millions d'abondances pour 1423 échantillons afin d'étudier la métabolisation des fibres alimentaires par le microbiote intestinal humain.

Chapitre 2

Modélisation biologique de la dégradation des fibres

Nous disposons de données métagénomiques qui doivent permettre la description de la dégradation des fibres dans l'intestin humain. Nous abordons dans un premier temps un formalisme écologique adapté à ce type de données (Section 2.1). Nous exposons dans un second temps le processus de sélection des marqueurs fonctionnels de la métabolisation anaérobie des fibres par le microbiote intestinal humain (Section 2.2). Enfin, nous présentons la constitution d'une matrice d'abondance des marqueurs fonctionnels associés aux 1423 échantillons métagénomiques considérés (Section 2.3).

2.1 Comment décrire la dégradation des fibres par l'écosystème du microbiote intestinal humain

Dans cette section, nous présentons quelques notions d'écologie (Section 2.1.1) avant de nous intéresser plus particulièrement à la description des écosystèmes complexes et en particulier les écosystèmes microbiens (Section 2.1.2).

2.1.1 Quelques notions d'écologie

Écologie des communautés. L'écologie des communautés est une branche de l'écologie s'intéressant particulièrement à l'évolution conjointe des abondances des individus d'un écosystème. On y définit les assemblages d'organismes vivants comme des ensembles phylogénétiquement bien caractérisés d'organismes qui co-varient temporellement et spatialement. Un assemblage peut désigner tout ou partie des individus d'un écosystème. Un assemblage entre plusieurs organismes peut avoir différentes origines. S'il s'agit d'une co-dépendance à une ressource, cet assemblage est appelé une *guilde* ; si les organismes sont potentiellement en interaction l'assemblage est appelé *communauté* (Nemergut et al., 2013).

Son application dans le cadre des écosystèmes microbiens se limite souvent à une étude taxonomique de la co-occurrence n'élucidant pas réellement les mécanismes fonctionnels sous-jacents. Un exemple de ce type d'approche sont les modèles de réseaux de co-occurrence où l'on estime une corrélation entre espèces dans plusieurs échantillons ; elle peut se visualiser de manière intuitive sous la forme d'un réseau où les espèces sont les nœuds et leurs corrélations sont symbolisées par des arêtes (Kurtz et al., 2015).

Mécanismes d'assemblage et de colonisation des écosystèmes. L'émergence d'assemblages est souvent observée mais leurs mécanismes sont plus difficilement expliqués. La littérature propose deux types de mécanismes complémentaires. Les mécanismes neutres modélisent

des organismes de fonctions équivalentes et pour lesquels des processus stochastiques de type naissance, mort, immigration ou spéciation sont les facteurs principaux expliquant les assemblages, ce qui donne un impact important à l'historique de l'écosystème. Le deuxième type de mécanisme est construit autour des particularités fonctionnelles des organismes d'un écosystème ; on peut encore les fractionner en deux branches, les mécanismes de type abiotique s'intéressent à l'adéquation d'un organisme à son environnement (Nemergut et al. (2010)) et les mécanismes biotiques aux interactions entre organismes.

Assemblages dans le microbiote intestinal humain. Différents points du premier chapitre de cette thèse illustrent la présence d'assemblages dans le microbiote intestinal humain. Par exemple, bien que taxonomiquement divers on retrouve toujours les mêmes genres prédominants entre individus (*Bacteroides* ou *Prevotella*) ; il existe de fortes différences entre les microbiotes d'individus originaires du Malawi et du Venezuela avec ceux des États-Unis ce qui témoigne de l'existence d'assemblages spécifiques des deux populations ; on observe également des corrélations entre groupes de genres bactériens et régimes alimentaires originaires d'une évolution conjointe et donc de la présence d'assemblages.

Les mécanismes en jeu sont à la fois neutres, biotiques et abiotiques :

- **Processus neutres** : les processus de type dispersifs ont un rôle important dans les premières années du système digestif. De même, la similarité des microbiotes des individus partageant un lieu de vie indique l'existence de processus neutres dans la construction de l'écosystème. Des études montrent une inertie temporelle du microbiote intestinal et mettent en exergue l'importance de l'historique de la colonisation.
- **Processus abiotiques** : le système digestif humain est contraignant ; l'absence d'oxygène, la structure de l'intestin, la présence de mucus ainsi que la fréquence des repas et le temps de rétention moyen du chyme sont autant de facteurs qui sélectionnent les individus du microbiote intestinal. En conséquence, le résultat de cette sélection sont des micro-organismes capables de fermentation anaérobie des fibres alimentaires non digestibles et de coloniser et consommer le mucus recouvrant l'épithélium intestinal.
- **Processus biotiques** : certaines espèces vivent en syntrophie c'est-à-dire qu'une espèce survie en détoxifiant l'environnement par consommation du "déchet" d'une autre espèce générant ainsi une co-dépendance (Glenn R Gibson, George T Macfarlane (1993)). Plus précisément, on peut prendre l'exemple des micro-organismes méthanogènes ou sulfato-réducteurs qui survivent en exploitant le dihydrogène de l'environnement dont l'accumu-

lation bloque certaines étapes clés de la glycolyse, leur association avec d'autres espèces du microbiote intestinal est syntrophique. De même une explication vraisemblable du très faible taux d'isolement des micro-organismes est l'existence d'interactions biotiques non élucidées. A l'exact opposé, des processus d'exclusions réciproques d'espèces bactériennes ont été mis en évidence (Koenig et al. (2011), Flint et al. (2007)).

Bien que participant aux assemblages du microbiote intestinal humain, les mécanismes neutres ne sont pas prédominants et étant par nature non fonctionnels, ne décrivent pas du tout la spécialisation métabolique préminente rencontrée dans cet écosystème. De même Jeraldo et al. (2012) ont développé une approche suggérant que les structures de diversités taxonomiques dans différents microbiotes intestinaux sont vraisemblablement les conséquences de mécanismes de sélection fonctionnelle et non pas de mécanismes stochastiques neutres. De ce fait, l'angle d'approche fonctionnelle semble être plus approprié pour caractériser les assemblages de micro-organismes dans le microbiote intestinal humain.

Traits et traits fonctionnels. Historiquement l'évolution des individus est décrite par l'adéquation de leur trait avec l'écosystème, dans ce contexte un trait désigne la performance d'un organisme relativement à un écosystème. De manière générale ce terme n'est pas bien défini; selon le domaine ou le type d'échelle considéré, un trait permet de caractériser tous les intermédiaires entre organismes et écosystèmes comme le montre la Figure 2.1.

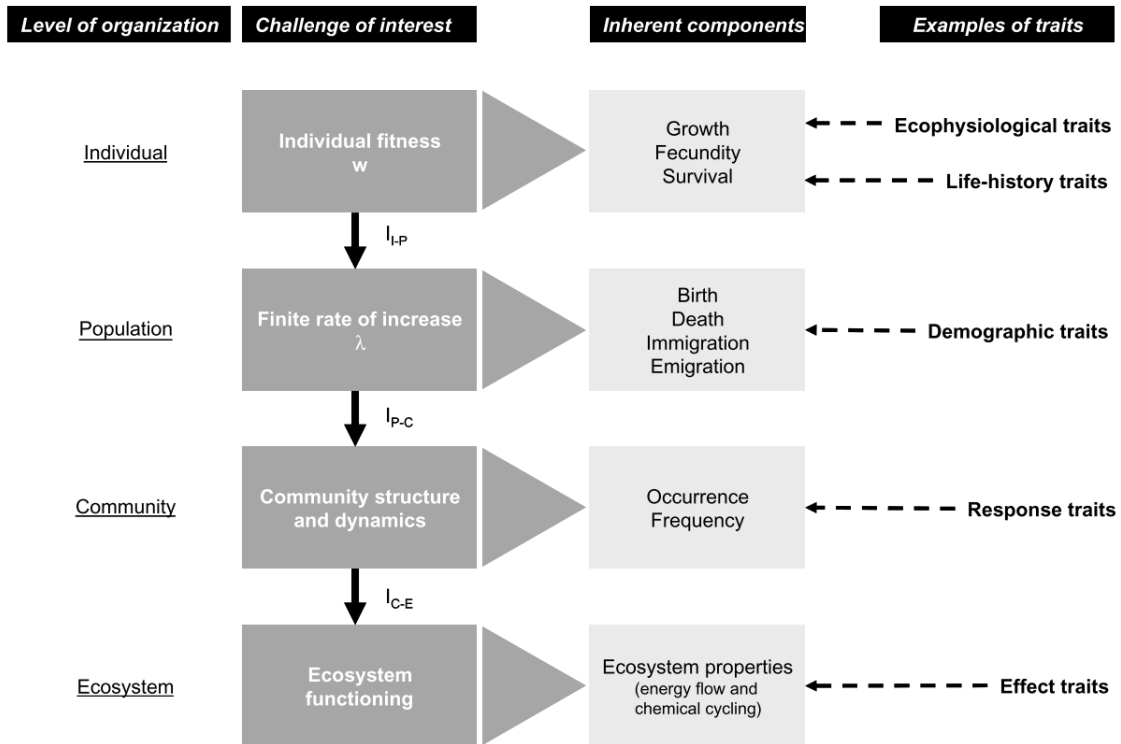


FIGURE 2.1: Utilisation de la notion de trait pour identifier des caractéristiques relevant de l'individu, la population d'individus, la communauté d'espèces jusqu'à l'écosystème. Tiré de Violle et al. (2007)

Violle et al. (2007) proposent de n'utiliser la notion de trait que pour les caractéristiques de l'individu, d'utiliser "paramètre démographique" pour une caractéristique de populations, "propriété de communauté" ou "d'écosystème" pour une propriété se rapportant à la communauté ou à l'écosystème.

Un trait fonctionnel permet de caractériser une fonction, il s'agit également d'un terme vague. Par exemple concernant les plantes, cette terminologie peut être utilisée pour désigner aussi bien une taille de feuille que la photosynthèse. Violle et al. (2007) proposent la définition suivante d'un trait fonctionnel.

Définition Un trait fonctionnel quantifie une caractéristique morphologique, physiologique ou phénologique ayant un impact direct ou indirect sur la performance individuelle (fitness).

2.1.2 Une description fonctionnelle plutôt que taxonomique

Si l'on s'intéresse aux fonctions exprimées dans un écosystème alors l'approche taxonomique peut ne pas être pertinente ; dans des écosystèmes complexes tels que les écosystèmes microbiens ou végétaux, la description taxonomique des organismes n'est pas toujours envisageable ou simplement possible :

- La quantité d'organismes différents est un obstacle pratique important.
- Dans le cas des écosystèmes microbiens la notion d'espèce en elle-même est mal définie puisqu'elle repose sur une notion de proximité de génomes ou de marqueurs taxonomiques.
- Certaines fonctions métaboliques sont clairement identifiables à certains taxons, par exemple seules les *archaea* ont développé la méthanogénèse, cependant connaître un rang taxonomique comme l'espèce ou le genre ne permet pas de prédire les fonctions métaboliques du micro-organisme ; des micro-organismes proches taxonomiquement peuvent posséder des fonctions différentes (présence ou absence).
- Le mécanisme de transfert de gène horizontal brouille l'association des fonctions aux espèces microbiennes.

Dans les écosystèmes microbiens, une description taxonomique donne une information fonctionnelle faible.

Pour répondre à ce genre de problématique, une alternative logique existe dans la description purement fonctionnelle et particulièrement dans l'utilisation de traits fonctionnels agrégés. D'abord introduits en écologie des plantes, les Traits Fonctionnels Agrégés (TFA) sont la valeur moyenne d'un trait fonctionnel sur l'ensemble des individus d'une communauté ou d'un écosystème. Bien que simplifiant drastiquement une information complexe, ils sont de bons indicateurs des propriétés d'une communauté ou d'un écosystème.

Par exemple, dans le cas de l'étude d'une communauté végétale il est possible de mesurer simplement une taille de feuille moyenne dans un écosystème (télédétection de spectres de canopées) alors qu'identifier toutes les variétés d'espèces est plus difficile ; de plus il est alors possible d'étudier une réponse fonctionnelle à des perturbations au niveau de l'écosystème et d'en fournir une interprétation mécaniste. Moreno García et al. (2014) ont ainsi étudié l'évolution de certains traits fonctionnels agrégés à l'échelle des communautés et ont montré une réponse fonctionnelle des communautés végétales en conséquence de leur mise en pâturage plus ou moins poussée et de leur proximité à une étendue d'eau. Ils trouvent par exemple que les végétaux mis en pâturage intensif sont plus résilients à l'écrasement et à la défoliation grâce à des mécanismes de repousse rapide.

Traits fonctionnels agrégés et abondances de marqueurs fonctionnels. La notion de TFA se révèle être particulièrement bien adaptée aux données méta-génomiques puisque toute mesure d'abondance de marqueurs fonctionnels revient à construire un TFA. En effet, un marqueur fonctionnel dont l'abondance est mesurée au sein d'un génome peut très clairement s'interpréter comme un trait fonctionnel : il s'agit d'une caractéristique physiologique ayant un impact direct ou indirect sur la performance de l'individu (fitness). De même un marqueur fonctionnel mesuré dans un échantillon métagénomique peut s'interpréter comme la résultante de la sommation des valeurs du trait fonctionnel de chacun des organismes de l'échantillon. Raes et al. (2011) se sont intéressés aux variations d'abondance de marqueurs fonctionnels de 25 échantillons métagénomiques de l'océan qu'ils ont mis en relation avec des métadonnées caractérisant l'environnement (localisation, données météorologiques, ... etc). Ils ont montré de fortes corrélations entre abondance de marqueurs fonctionnels et ensoleillement, température et concentration d'O₂ et de CO₂, expliquant et identifiant ainsi les mécanismes à l'origine des différences de profils taxonomiques.

2.2 Sélection de traits fonctionnels de la dégradation anaérobie des fibres par le microbiote intestinal humain

Nous choisissons d'adopter une approche strictement fonctionnelle et en particulier, nous souhaitons utiliser une approche par traits fonctionnels agrégés qui semble particulièrement bien adaptée aux données métagénomiques.

Nous travaillons sur les séquences de 9.9 millions de gènes qui nécessitent donc une étape d'annotation. Les Orthologies de KEGG ou les familles de Glycosides Hydrolases (GH) et Polysaccharides Lyases (PL) présentées dans cette section sont de bon exemples de marqueurs fonctionnels qui permettent d'associer une fonction à un regroupement de séquences de gènes similaires. Ce sont des traits fonctionnels.

Dans cette section notre objectif est alors de déterminer et choisir un ensemble de traits fonctionnels (marqueurs) pertinents pour décrire les différentes étapes de la métabolisation des fibres en condition anaérobie. Nous exposons dans un premier temps l'hydrolyse des fibres en sucres simples (Section 2.2.1) et sélectionnons 25 marqueurs avant de considérer la métabolisation des sucres simples (2.2.2) et la sélection de 61 traits fonctionnels associés.

2.2.1 Hydrolyse des polysaccharides

Les fibres alimentaires proviennent des légumes, céréales et fruits, elles ne sont pas digestibles et parviennent jusqu'au colon où le microbiote les dégrade. Elles consistent principalement en des polysaccharides de parois cellulaires végétales, des polysaccharide de stockage où de l'amidon résistant. Le Tableau 2.1 présentent les fibres alimentaires principales ainsi que leur composition en monosaccharide. La taille et la structure des fibres alimentaires ne leur permet de n'être dégradées que dans la lumière intestinale par des enzymes extra-cellulaires excrétées ou présentes à la surface du micro-organisme dans la lumière intestinale. Certaines fibres peuvent être de structure complexe, de par leur densité et ramification ou par le nombre de sucres et de type de liaisons nécessitant l'action conjointe de différentes enzymes.

Fibres alimentaires	Glucose	Fructose	Mannose	Rhamnose	Fucose	Galactose	Acide Galacturonique	Arabinose	Xylose
Pectines									
arabinan (Betterave à sucre)				~		~	~	X	
arabinogalactan						X		~	
pectic galactan (pomme de terre)				~		X	~	~	
homogalacturonan (citrus)							X		
rhamnogalacturonan I (pomme de terre)				X		~	X	~	~
rhamnogalacturonan II (Vin rouge)				~		~	X	~	~
Hemicelluloses and b-glucans									
arabinoxylan (blé)								X	X
Xylan (avoine)									X
xyloglucan (Tamarin)	X				~	~			X
glucomannan (konjac)	X		X						
galactomannan (caroube)			X			X			
B-glucan (orge)	X								
laminarin (algue)	X								
Starch, fructan, and a-glucans :									
amylopectin (pomme de terre, maïs)	X								
pullulan	X								
dextran	X								
inulin (chicorée)	~	X							
levan		X							

TABLE 2.1: Fibres alimentaires et leur composition en monosaccharide, certaines fibres contiennent marginalement certains sucres, nous le symbolisons par ~, tiré de Martens et al. (2011).

Dans le cas du microbiote intestinal humain les enzymes impliquées sont majoritairement les Glycosides Hydrolases (GH) et les Polysaccharides Lyases (PL). L'hydrolyse des fibres mobilise différents types d'enzymes qui ne libèrent pas directement de sucres simples. Dans le cas des GH, la distinction est faite entre enzymes dites *endo* et enzymes *exo*. Les enzymes *exo* sont uniquement capables de cliver les liaisons en bout de chaîne du polymère tandis que les enzymes

endo permettent de créer de nouvelles "extrémités" en attaquant des liaisons à l'intérieur de la chaîne, la Figure 2.2 illustre ce point. Les familles de GH peuvent être *endo* ou *exo* selon les familles d'enzymes considérées.

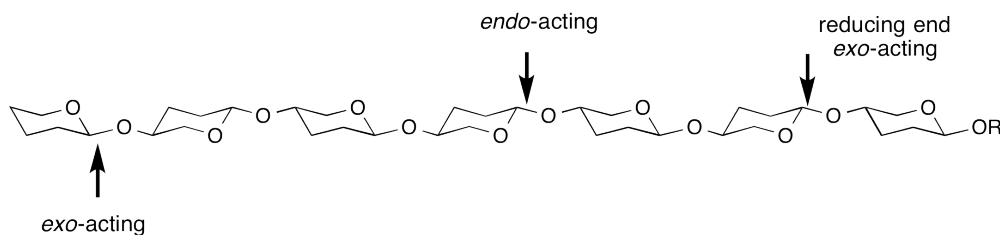


FIGURE 2.2: Mode d'action des enzymes *endo* et *exo* lors de l'hydrolyse d'un polymère linéaire. Tiré de http://www.cazylopedia.org/index.php/Glycoside_hydrolases.

Les PL sont des enzymes *endo* associées aux liaisons faisant intervenir des acides uroniques c'est-à-dire les sucres simples dont le dernier carbone a été oxydé, la Figure 2.3 en montre un exemple dans le cas du glucose. De manière abusive nous parlerons toujours de sucres ou de polysaccharides pour désigner de tels acides et leur polymères.

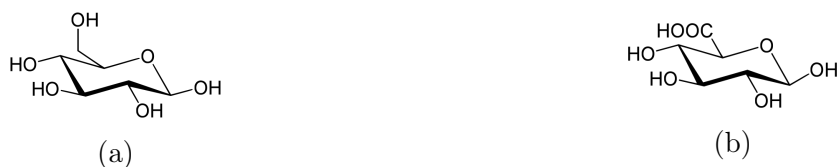


FIGURE 2.3: L'oxydation du carbone 6 du glucose (a) produit de l'acide glucuronique (b)

La méthode de classification actuelle est dérivée de la méthodologie proposée par Henrissat et al. (1991), les auteurs ont construit et maintenu la base de données de référence nommée CAZY (<http://www.cazy.org/>) comprenant 135 familles de GH et 24 de PL. Les familles de GH et PL y sont construites autour de domaines de séquences très conservées mais dans des gènes de longueurs très variables et de séquences peu homologues. De ce fait les stratégies habituelles d'annotation telles que l'utilisation de l'algorithme BLAST sur des séquences de référence offrent des résultats d'annotation médiocres. Yin et al. (2012) apportent une réponse au problème d'annotation à haut débit en modélisant les domaines de signature des enzymes de la base de donnée CAZY et en proposant une approche mixte utilisant à la fois BLAST et des modèles de GH et PL sous forme de chaînes de Markov cachées.

Les enzymes classifiées selon ce système partagent dans la grande majorité des cas un mécanisme moléculaire ainsi qu'une activité hydrolytique commune. Le principal atout de cette classification est l'interprétation mécaniste des réactions catalysées par les enzymes d'une même famille. Cependant l'appartenance à l'une de ces famille n'offre pas toujours d'information de spécificité des substrats. A quelques exceptions près, plusieurs activités hydrolytiques ont été mises en évidence pour chaque famille d'enzymes, mais cela ne signifie pas que toute ces activités soient assurées par chacune des enzymes de chaque famille. A titre d'exemple, le GH 13 est associé à une capacité enzymatique d'hydrolyse des fibres contenant du glucose ou du fructose, cependant l'appartenance à cette famille n'offre aucune garantie que l'enzyme catalyse la production de glucose seul, de fructose seul ou de fructose et de glucose ensemble. Cet état de fait est en lien direct avec l'aspect modulaire de ces enzymes, c'est-à-dire que l'enzyme se compose de plusieurs parties fonctionnellement indépendantes dont le GH ou le PL est l'un des composants. La fonction d'hydrolyse assurée n'est pas uniquement dépendante du GH ou PL mais peut nécessiter la présence conjointe de modules supplémentaires appelés Carbohydre-Binding Modules (CBM). Ces modules permettent l'absorption de l'enzyme sur le polysaccharide ce qui résulte en une proximité entre l'enzyme et l'objet de la catalyse. La connaissance de ces enzymes ainsi que les techniques d'annotation actuelles ne sont pas suffisantes pour fournir une information fonctionnelle spécifique et fiable. L'appartenance à une famille de GH ou PL offre donc une information faible concernant la spécificité de substrat que les enzymes catalysent.

Nous exploitons le travail conjoint de El Kaoutari et al. (2013), Lozupone et al. (2008), ainsi que des génomes annotés afin de choisir des GH et PL caractéristiques de l'hydrolyse des fibres alimentaires par le microbiote intestinal humain. Dans la mesure où notre travail porte sur les polysaccharides, nous ignorons les activités d'hydrolyses associées aux fibres contenant des protéines telles que les glycanes animales et les peptidoglycanes. Le Tableau 2.2 présente les 25 GH et PL sélectionnés ainsi que les sucres potentiellement produits suite à leur activité hydrolytique, ces informations sont tirées de la base de donnée CAZY (www.cazy.org). Ces marqueurs fonctionnels sont des traits fonctionnels de l'hydrolyse des fibres.

Marqueur de l'hydrolyse/Sucres	Glucose	Fructose	Mannose	Rhamnose	Fucose	Galactose	Acide Galacturonique	Arabinose	Xylose
GH 3	X					X		X	X
GH 5	X		X			X			X
GH 8	X								X
GH 9	X								X
GH 10									X
GH 13	X	X							
GH 16	X				X	X			X
GH 26			X						X
GH 28				X			X		
GH 30	X				X	X	X		X
GH 32	X	X							
GH 39									X
GH 43						X		X	X
GH 44	X								X
GH 48	X								
GH 51	X							X	X
GH 74	X				X	X			X
GH 91		X							
GH 94	X								
GH 115									X
GH 120									X
GH 127								X	
PL1							X		
PL9							X		
PL11							X		

TABLE 2.2: Sucres potentiellement produits par les enzymes de chaque famille de GH et PL.

2.2.2 Sélection des marqueurs de la métabolisation des oses

2.2.2.1 Le catabolisme des monosaccharides en condition anaérobie

On définit la capacité métabolique d'un organisme vivant comme la capacité interne à transformer un composé chimique en un autre par réactions successives. Dans ce cadre, les composés chimiques sont appelés métabolites et l'on peut les séparer en deux catégories. Les métabolites sont intra-cellulaires lorsque ils sont produits et consommés à l'intérieur de la cellule, ils sont extra-cellulaires lorsque ils sont transportés à l'intérieur ou excrétés à l'extérieur du micro-organisme.

Contrairement à l'hydrolyse des fibres, la métabolisation des sucres simple est un processus se déroulant à l'intérieur des micro-organismes. Il s'agit d'un catabolisme c'est-à-dire une catégorie de métabolismes où le processus réactif conduit à la dégradation des métabolites initiaux. La Figure 2.4 synthétise schématiquement ce catabolisme qui prend en entrée des sucres simples tel que le glucose et produit en sortie des Acides Gras à Chaînes Courtes (AGCC) comme le Lactate ou l'Acétate accompagnés de dégagement gazeux (H_2 et CO_2). De manière parallèle à la production d'AGCC, et de manière à illustrer le concept de syntrophie, les AGCC et les gaz sont réutilisés par des membres du microbiote responsables par exemple de la méthanogénèse et de l'acétogénèse.

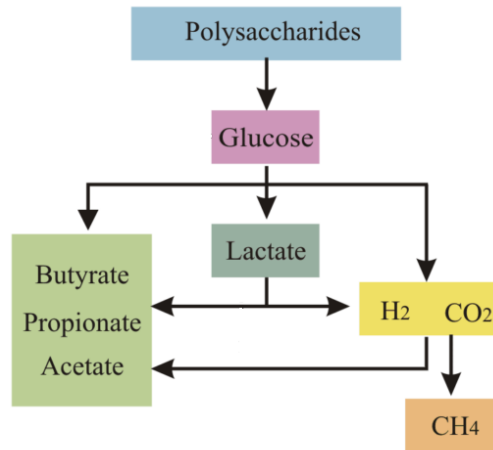


FIGURE 2.4: Représentation schématique de la métabolisation des sucres simples en AGCC et de leur réutilisation.

L'avantage principal associé à la capacité de fermentation est la possibilité d'extraire des sucres simples de l'énergie sous forme d'ATP. L'ATP ou Adénosine TriPhosphate intervient dans les réactions endo-énergétiques et assure de nombreuses fonctions vitales au sein de la cellule, telles que le transport trans-membranaire, la motilité ou la division cellulaire. De manière très générale, la production d'ATP se fait par réactions d'oxydo-réduction où un donneur d'électron X est oxydé par un accepteur d'électron A comme le montre la représentation schématique de la Figure 2.5.

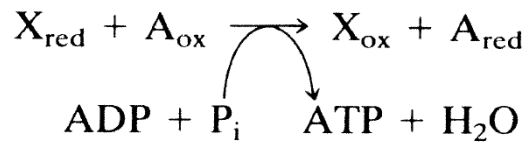


FIGURE 2.5: Représentation schématique de la production d'ATP, X est un donneur d'électron et A un accepteur. Tiré de Gottschalk (1986)

Selon les organismes et les milieux de vie, les composés A et X sont très variables. Chez l'homme la production d'ATP peut se faire en prenant le glucose comme donneur d'électron et le O₂ comme accepteur d'électron. La Figure 2.6 donne plusieurs exemples d'accepteurs et de donneurs d'électrons parmi plusieurs micro-organismes. Les organismes utilisant des substrats dits organiques (composés de carbone) en tant que donneurs d'électrons sont appelés chimio-organotrophes, et les organismes exploitant des substances minérales (O₂, CO₂,...) sont appelés chimio-lithotrophes. Les deux types d'organisme sont présents dans le microbiote intestinal humain.

type	electron donor	electron acceptor	carbon source	examples
chemoorganotrophy	organic substrate	O ₂	organic substrate	pseudomonads, bacilli
	organic substrate	NO ₃ ⁻	organic substrate	<i>Bacillus licheniformis</i>
	organic substrate	SO ₄ ²⁻	organic substrate	sulfate reducers
	organic substrate	organic substrate	organic substrate	clostridia, lactic acid bacteria
chemolithotrophy	H ₂	O ₂	CO ₂	hydrogen-oxidizing bacteria
	H ₂ S	O ₂	CO ₂	thiobacilli
	H ₂ S	NO ₃ ⁻	CO ₂	<i>Th. denitrificans</i>
	Fe ²⁺	O ₂	CO ₂	<i>Th. ferrooxidans</i>
	NH ₃	O ₂	CO ₂	<i>Nitrosomonas</i>
	NO ₂ ⁻	O ₂	CO ₂	<i>Nitrobacter</i>
	H ₂	CO ₂	CO ₂	methanogenic bacteria
	H ₂	CO ₂	CO ₂	acetogenic bacteria

FIGURE 2.6: Exemples de donneurs et d'accepteurs d'électrons chez différents organismes. Tiré de Gottschalk (1986)

Le processus central associé au catabolisme des sucres simples est la glycolyse ou encore la voie d'Embden-Meyerhof-Parnas illustrée en Figure 2.10. Il s'agit du processus métabolique permettant la transformation d'un mono-saccharide en deux molécules de pyruvate et produisant deux molécules d'ATP et de NADH (Nicotinamide adénine dinucléotide). De la même façon que l'ATP peut être décrit comme une molécule porteuse d'énergie, le NADH est le marqueur de la réduction. Le couple NAD^+/NADH joue localement (dans la cellule) le rôle du couple $A_{\text{ox}}/A_{\text{red}}$ de la Figure 2.5 et doit pouvoir réduire à son tour un composé chimique qui sera ensuite excrété par l'organisme. Dans le cas des organismes aérobies le O_2 est l'accepteur final d'électron qui permet de régénérer NADH en NAD^+ . Ce processus exploitant le O_2 est appelé chaîne respiratoire et permet de créer jusqu'à deux molécules d'ATP par NADH en parallèle de la glycolyse. Dans le cadre du catabolisme des sucres simples en milieu *anaérobie*, l'absence d'accepteur final librement disponible implique que les voies de fermentation anaérobie doivent être équilibrées autour de la production et régénération de NADH. Le pyruvate et ses dérivés sont souvent utilisés comme accepteur final d'électrons mais peuvent aussi permettre la production d'ATP. La Figure 2.7 présente un exemple où les deux molécules de pyruvate résultant de la glycolyse sont transformées respectivement en éthanol et en acétate. La fabrication d'acétate génère de l'ATP supplémentaire et la production d'éthanol régénère les deux NADH produits lors de la glycolyse. Ici, c'est l'acetyl-CoA qui fait office d'accepteur d'électron final. Ce type de voie métabolique est observé dans le cas de la fermentation d'acides mixtes d'*E.coli*. Il est à noter que si la réduction répétée des acides carboxyliques en aldéhyde puis en alcool semble être une stratégie viable de régénération de NADH, la production de solvant tels que l'Acétone, l'éthanol ou le butane-di-ol est toxique pour les micro-organismes et reste faible dans l'intestin humain.

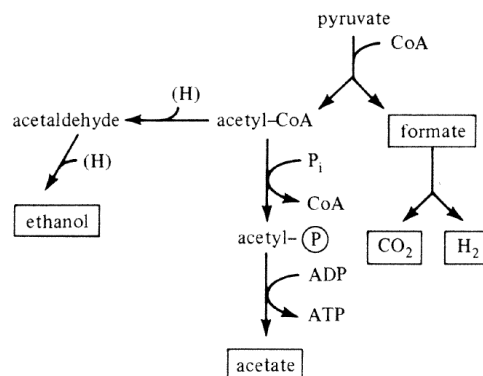


FIGURE 2.7: Production d'ATP possible après glycolyse, les (H) représentent les marqueurs de la réduction comme le NADH, le passage de l'acetyl-CoA à l'éthanol consomme deux NADH. Tiré de Gottschalk (1986)

D'autres molécules peuvent prendre le rôle d'accepteur d'électrons, c'est le cas des Ferrédoxines ou des FAD (Flavine adénine dinucléotide). Les Ferrédoxines peuvent être utilisées comme co-facteurs de la conversion du pyruvate en acetyl-CoA, mais leur particularité est la possibilité de régénérer NADH de manière presque unilatérale (très peu réversible en présence de NADH) et de réduire H^+ en H_2 en présence de catalyseurs microbiens, comme illustré par la Figure 2.8

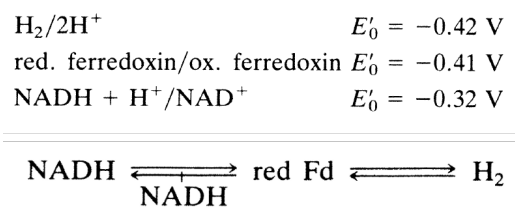


FIGURE 2.8: Le couple d'oxydoréduction associé à la Ferrédoxine noté Fd, est de potentiel redox plus faible que celui du NADH et très proche de celui du couple H_2/H^+ , les réactions entre NADH et Ferrédoxines sont faites principalement dans le sens de la réduction des Ferrédoxines. Tiré de Gottschalk (1986)

Des pressions partielles importantes en H_2 dans le milieu inhibe la régénération de NADH par le biais de la Ferrédoxine. L'association avec des micro-organismes consommateurs de H_2 permet une production supplémentaire d'ATP chez les micro-organismes exploitant les Ferrédoxines.

La Figure 2.9 illustre les comportements différents d'un même organisme lorsque la pression partielle en H_2 est forte et lorsqu'elle est faible.

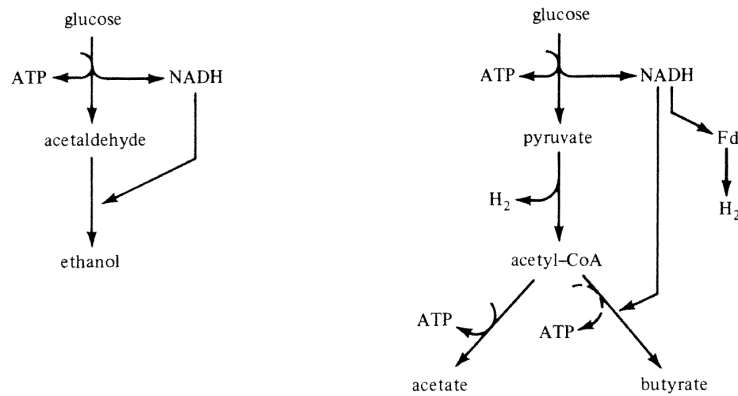


FIGURE 2.9: Illustration de stratégies différentes de régénération du NADH selon que la pression partielle de H_2 est importante (schéma de gauche) ou faible (schéma de droite). Tiré de Gottschalk (1986)

Dans la mesure où les organismes consommant les produits de la fermentation ont une influence sur le catabolisme des monosaccharides, nous choisissons de décrire les voies métaboliques hydrogénotrophes telles que l'acetogénèse et la méthanogénèse. De même, nous considérons aussi la réutilisation du lactate et du lactaldehyde.

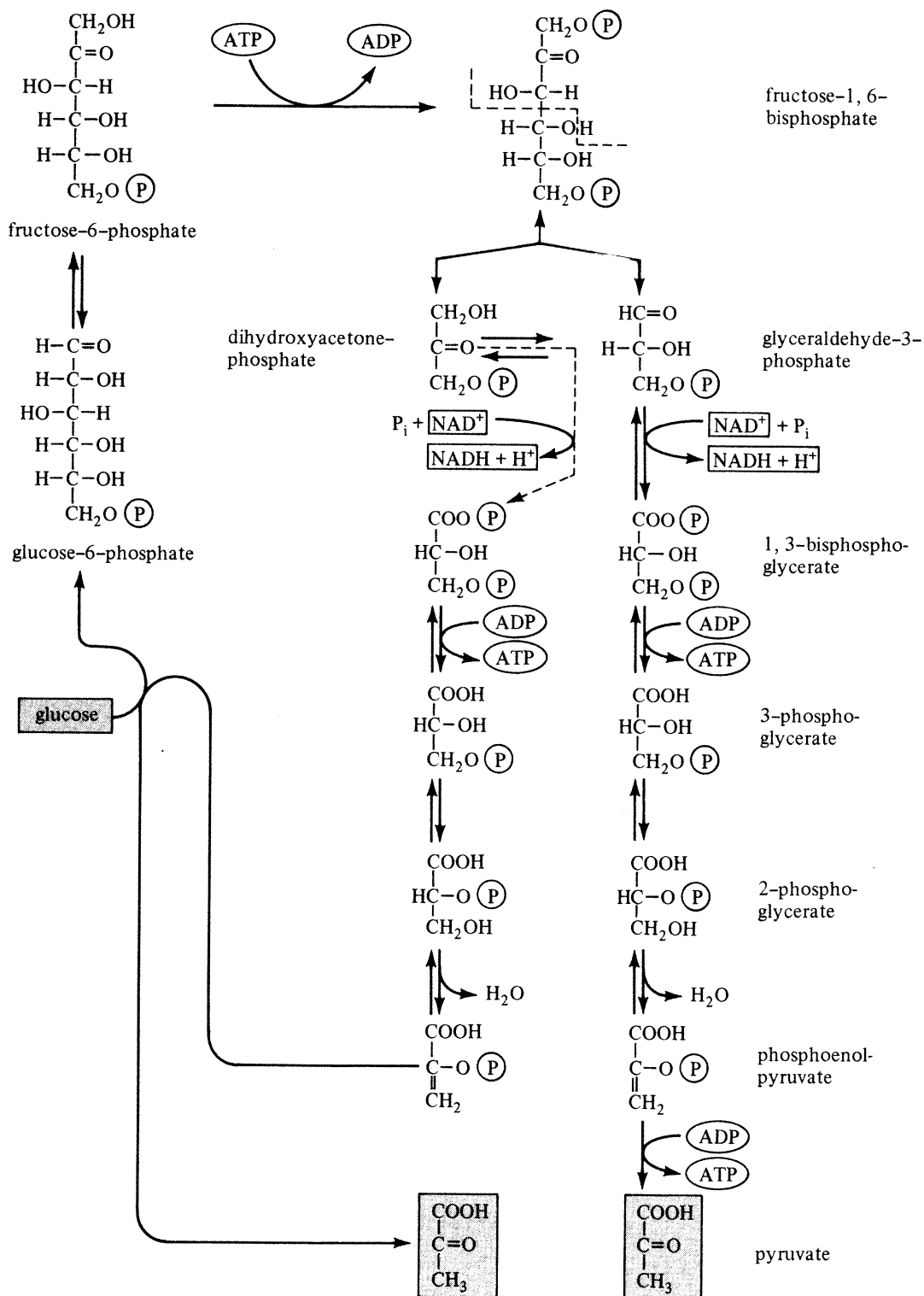


FIGURE 2.10: Représentation détaillée de la voie d'Embden-Meyerhof-Parnas. Tiré de Gottschalk (1986)

2.2.2.2 Principe de sélection des marqueurs fonctionnels

Nous choisissons les orthologies de KEGG (Kanehisa et al., 2009) en tant que marqueurs fonctionnels. Les orthologies sont des regroupements de gènes de séquences homologues qui possèdent une origine ancestrale commune. À l'intérieur d'une orthologie, la diversité de séquence provient d'événements de spéciation plutôt que de duplications (paralogies). Les gènes orthologues partagent la même fonction, cette classification est donc un marqueur fonctionnel pertinent. Les orthologies de KEGG sont construites manuellement ce qui garantit en général un bon niveau de qualité d'information fonctionnelle associée aux KO.

Il est important de rappeler que nous travaillons avec des abondances de gènes annotés ce qui donne une fonction probable. Indépendamment de la qualité de l'annotation, l'abondance d'un gène fonctionnellement annoté n'est l'indication que d'un *potentiel* de catalyse.

Pour chaque étape du catabolisme des monosaccharides il est presque toujours possible de trouver un KO associé. Cependant si l'on souhaite pouvoir interpréter l'abondance d'un KO comme la quantification du potentiel de réalisation d'une réaction, il faut posséder la garantie de la spécificité fonctionnelle de ce KO. En effet, certaines enzymes sont génériques et peuvent participer à plusieurs processus métaboliques n'ayant aucun rapport entre eux. L'abondance de ces enzymes ou des gènes qui leur sont associés est une information complexe qui ne permet de supposer la réalisation d'aucune réaction en particulier. De manière similaire, s'il existe plus d'une enzyme capable de catalyser une réaction, alors il n'est possible de rendre compte du potentiel de réalisation de la réaction que si toutes les enzymes sont spécifiques de la réaction.

Nous sélectionnons les KO de sorte à obtenir une liste de marqueurs fonctionnels spécifiques. Ce processus se déroule en plusieurs étapes et commence par le dénombrement des voies majeures du catabolisme des monosaccharides et de la réutilisation des AGCC. Notre sujet d'étude est la digestion anaérobie des fibres par le microbiote intestinal humain, nous écartons donc les voies improbables dans ce contexte telles que les voies aérobies ou anaboliques. Pour ce faire, nous exploitons un ouvrage de référence (Gottschalk, 1986) et nous explorons la base de donnée de voies métabolique MetaCyc (Caspi et al., 2014). Cette étape permet de constituer une liste putative de réactions.

Nous explorons ensuite la base de donnée KEGG afin de retrouver les réactions correspondant à la liste établie. Dans KEGG les réactions sont repérées par un code constitué de la lettre R suivie de 5 chiffres. Ces réactions sont reliées aux KO correspondants repérés par un code constitué de la lettre K suivie de 5 chiffres. Certaines réactions ne sont pas associées à des KO, dans ce cas elles sont éliminées de la liste.

Nous dressons alors une liste de réactions et de KO associés à partir de laquelle nous

procédons à une deuxième étape de sélection, de la façon suivante : si aucun gène d'un KO associé à une réaction n'est présent dans les 1423 échantillons métagénomiques, nous écartons ce KO. De même, si le KO est associé à la catalyse de plus d'une réaction, alors ce KO n'est pas spécifique et il est ignoré ainsi que la réaction et tous les autres KO qui y sont associés.

Nous détaillons ce processus de sélection dans les 4 sections qui suivent, nous abordons successivement les mécanismes reliant les monosaccharides à la glycolyse, le métabolisme central, l'utilisation du pyruvate et les voies syntrophiques. Dans la suite, les réactions et KO sélectionnés sont colorés en vert tandis que les réactions et KO rejetés sont rouges. Il sont désignés par leurs codes KEGG.

2.2.2.3 Voies du catabolisme des sucres simples provenant de la dégradation des fibres

Nous présentons les voies permettant de relier chaque sucre provenant de l'hydrolyse des fibres aux voies classiques de la glycolyse anaérobie. Les sucres considérés sont précisément le D-mannose, le L-fucose, le L-rhamnose, le D-xylose, le D-galactose, le L-arabinose, le D-fructose ainsi que l'acide galacturonique. Certaines voies métaboliques proposées par MetaCyc sont écartées, le Tableau 2.3 référence ces entrées ainsi que l'origine de la décision.

D-mannose. La dégradation du Mannose se fait selon :

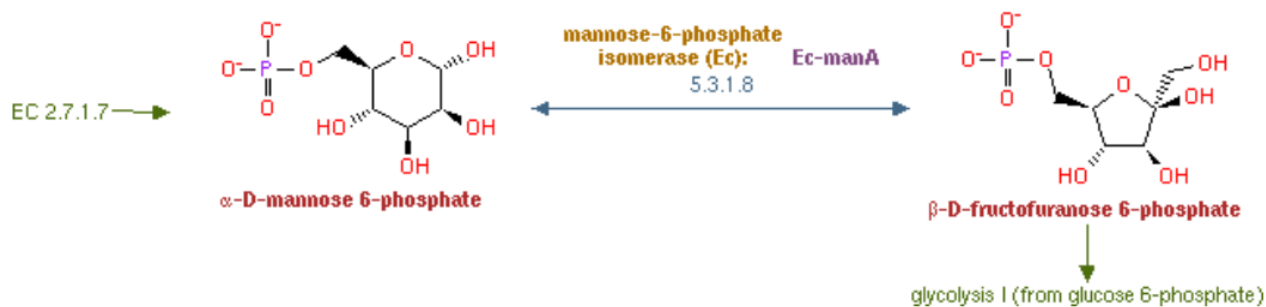


FIGURE 2.11: Voie de dégradation du Mannose. Tiré de <http://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=MANNCAT-PWY&detail-level=4>

Trois réactions sont associées à cette voie :

- **R02630 : protein-N(pi)-phosphohistidine :D-mannose 6-phosphotransferase.** Les KO associés sont les K02793, K02795, K02794, K02796. Ces KO sont spécifiques mais celui de la réaction R01326, réaction identique mais impliquant des co-facteurs différents, ne l'est pas.
- **R01326 : ATP :D-mannose 6-phosphotransferase** Le KO associé est le K00844 (hexokinase). Réaction dont le KO est non spécifique, transport du Mannose/ Glucose/ Fructose/ D-Glucosamine/ et différentes versions stéréospécifiques du glucose et fructose.
- **R01819 : D-mannose-6-phosphate aldose-ketose-isomerase** Le KO associé est le K01809. Ce KO est spécifique.

Bilan : Une seule réaction est gardée, responsable de l'interconversion du α -D-mannose 6-phosphate en β -D-fructofuranose 6-phosphate aussi appelé fructose 6-phosphate. Ce dernier peut être directement utilisé en glycolyse.

L-fucose. La dégradation du Fucose se fait selon :

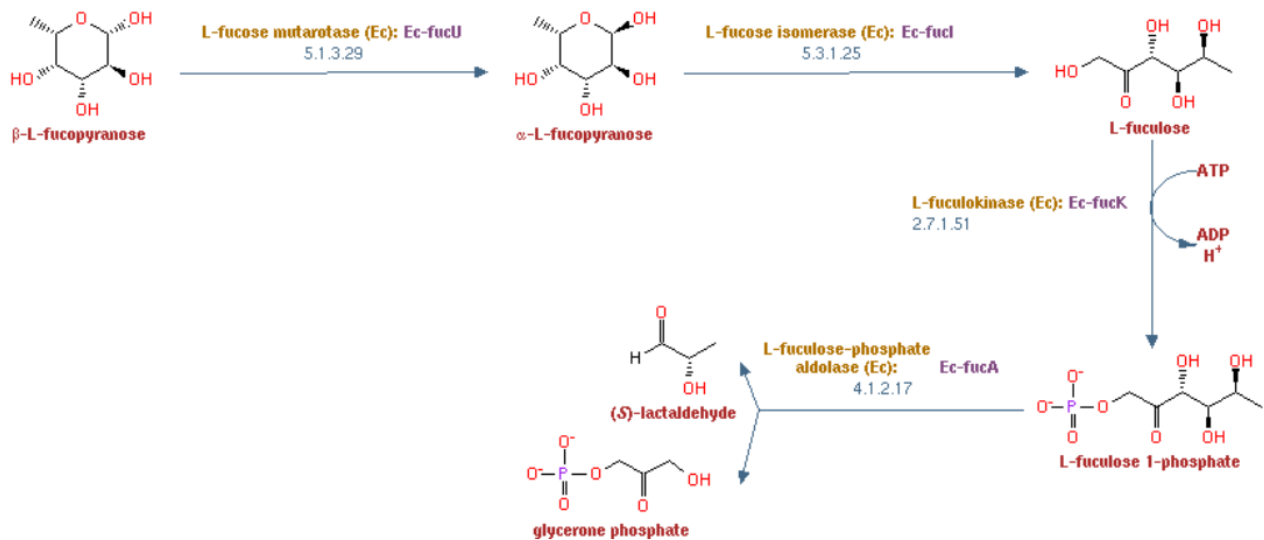


FIGURE 2.12: Voie de dégradation du Fucose. Tiré de <http://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=FUCCAT-PWY&detail-level=4>

Trois réactions sont associées à cette voie :

- **R03163** : **L-fucose aldose-ketose-isomerase** Le KO associé est le K01818.
Il est spécifique.
- **R03241** : **ATP :L-fucose 1-phosphotransferase** Le KO associé est le K00879.
Il est spécifique.
- **R02262** : **L-Fucose-1-phosphate lactaldehyde-lyase** Le KO associé est le K01628.
Il est spécifique.

Bilan : La première réaction de changement de conformation n'est pas décrite dans KEGG. Toutes les autres réactions sont associées à des KO spécifiques. Le Glycérone phosphate peut être directement ré-utilisé en glycolyse.

L-rhamnose. MétaCyc propose 3 voies de dégradation du Rhamnose. Après examen, deux parmi les trois sont associées à des organismes dont la présence est peu probable dans l'intestin humain et sont constituées de réactions ne permettant pas de dégager d'ATP. Nous ne considérons donc que la voie restante.

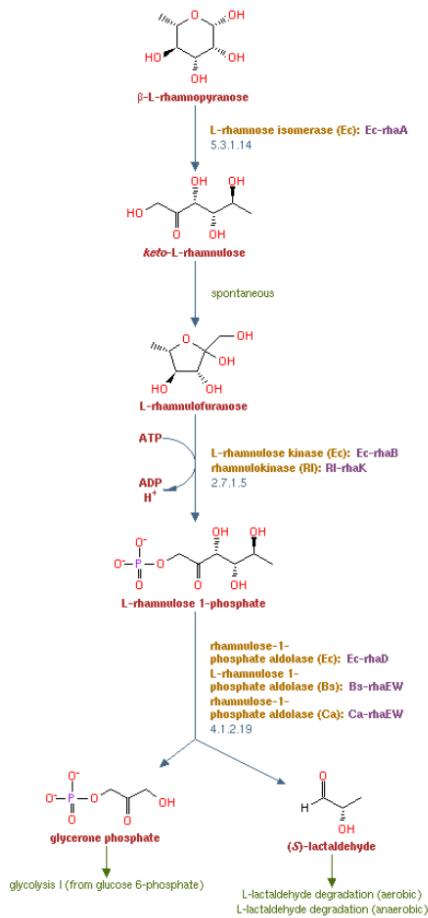


FIGURE 2.13: Voie de dégradation du rhamnose, tiré de <http://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=RHAMCAT-PWY&detail-level=4>.

Trois réactions sont associées à cette voie :

- **R02437** : **L-rhamnose ketol-isomerase**. Le KO associé est le K01813. Il est spécifique.
- **R03014** : **ATP :L-rhamnulose 1-phosphotransferase**. Le KO associé est le K00848. Il est non spécifique et peut aussi réagir sur du L-xylulose.
- **R02263** : **L-Rhamnulose-1-phosphate lactaldehyde-lyase**. Le KO associé est le K01629. Il est non spécifique et peut aussi réagir sur du L-Xylulose-P.

Bilan : Deux réactions sur les trois ont dues être supprimées. La réaction restante est associée à la production spécifique de L-rhamnulofuranose.

D-xylose. Metacyc dénombre 4 voies métaboliques associées à la dégradation du Xylose. Les KO associés à la voie numéro deux sont tous non spécifiques. Les voies numéro 3 et 4 sont ne produisent pas d'ATP en métabolisant ce sucre, de même ces voies semblent être associées à des *Archaea* extrémophiles peu vraisemblables dans le colon humain, ces voies sont donc ignorées.

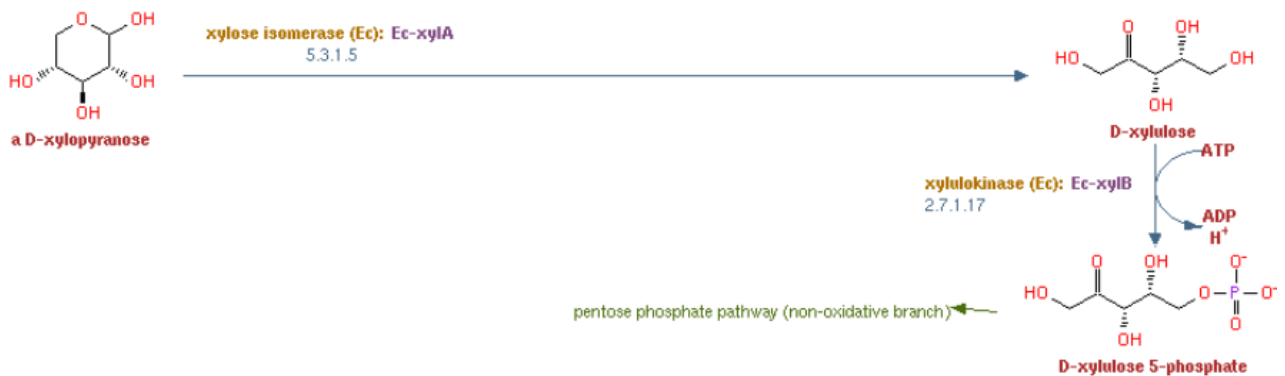


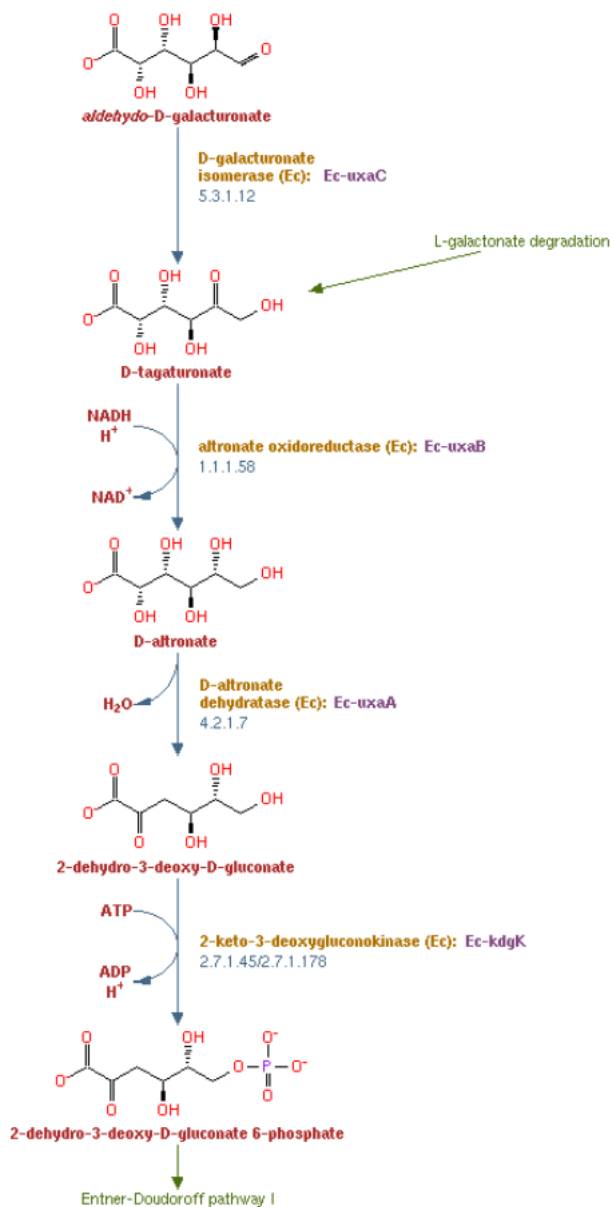
FIGURE 2.14: Voie de dégradation du xylose, tiré de <http://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=XYLCAT-PWY&detail-level=4>.

Deux réactions sont associées à cette voie :

- **R01432** : **D-xylose aldose-ketose-isomerase**. Le KO associé est le K01805. Il est non spécifique et permet aussi l'inter-conversion du D-glucose en D-fructose.
- **R01639** : **ATP :D-xylose 5-phosphotransferase**, Le KO associé est le K00854. Il est spécifique.

Bilan : Une seule réaction est conservée, elle permet la production de D-Xylose 5-phosphate à partir de D-xylose par le biais d'ATP. Le métabolite final peut être utilisé en suivant la voie du pentose phosphate ou du Bifidobacterium shunt, décrits plus loin dans ce document.

Acide Galacturonique. Metacyc dénombre trois voies associées à la métabolisation du galacturonate. Seules les réactions de la première voie sont associées à des KO. Nous ignorons les autres voies.



Cinq réactions sont associées à cette voie :

- **R01983** : **D-Galacturonate aldose-ketose-isomerase**. Le KO associé est le K01812. Il est non spécifique et pilote aussi l'isomérisation du Glucuronate en fructonate.
- **R02555** : **D-Altrunate :NAD⁺ 3-oxidoreductase**. Le KO associé est le K00041. Il est spécifique.
- **R01540** : **D-altrunate hydro-lyase (2-dehydro-3-deoxy-D-gluconate-forming)**. Le KO associé est le K00041. Il est spécifique.
- **R01541** : **ATP :2-dehydro-3-deoxy-D-gluconate 6-phosphotransferase**. Le KO associé est le K00874. Il est spécifique.

Bilan : Trois réactions sont conservées décrivant le passage du tagaturonate au 2-dehydro-3-deoxy-D-gluconate 6-phosphate faisant partie de la voie d'Entner-Doudoroff .

FIGURE 2.15: Voie de dégradation du galacturonate, tirée de <http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=GALACTUROCAT-PWY&detail-level=4>.

D-galactose. Metacyc dénombre 5 voies possibles. La Voie numéro 3 est principalement décrite chez les Embryophyta (clade végétale), ce qui n'est pas pertinent pour le microbiote intestinal. La quatrième voie de MetaCyc est principalement trouvée chez des *fungi* dont la présence dans le microbiote intestinal humain est peu vraisemblable. Enfin, la voie numéro 5 est une version stéréo-spécifique de la première voie, cette distinction se retrouve naturellement lors de l'énumération des KO dans KEGG.

Neuf réactions sont associées à la première voie :

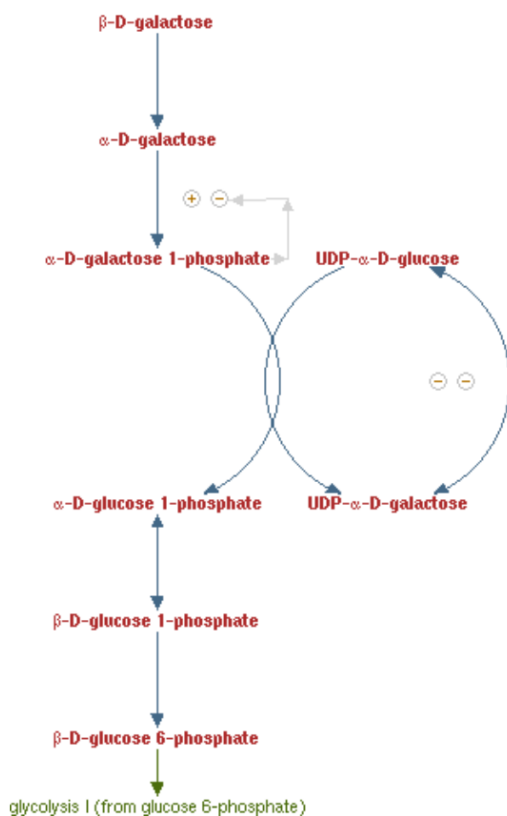


FIGURE 2.16: Voie de dégradation du galactose, voie de Leloir, tiré de <http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-6317&detail-level=1>

- **R10619** : **D-galactose 1-epimerase**. Le KO associé est le K01785. Il est non spécifique et permet l'épimérisation du glucose.
- **R01092** : **ATP :alpha-D-galactose 1-phosphotransférase**. Le KO associé est le K00849, il est spécifique.
- **R00955** : **UDP-glucose :alpha-D-galactose-1-phosphate uridylyltransférase**. Le KO associé est le K00965, il est spécifique.
- **R00502** : **UTP :alpha-D-galactose-1-phosphate uridylyltransférase**. Le KO associé est le K12447, il est non spécifique.
- **R00289** : **UTP :alpha-D-glucose-1-phosphate uridylyltransférase**. Les KO associés sont les K00963 et K12447. Un seul des deux est spécifique.
- **R00291** : **UDP-glucose 4-epimerase**. Le KO associé est le K01784. Il n'est pas spécifique.
- **R08639** : **alpha-D-glucose 1,6-phosphomutase**. Le KO associé est le K01835. Il n'est pas spécifique.

- **R02728** : **beta-D-Glucose 1-phosphate 1,6-phosphomutase**. Le KO associé est le K01838. Il est spécifique, cependant la réaction précédente, identique au détail de la stéréo-spécificité, ne l'est pas.
- **R00287** : **UDP-glucose glucophosphohydrolase**. Le KO associé est le K01513, il n'est

pas spécifique.

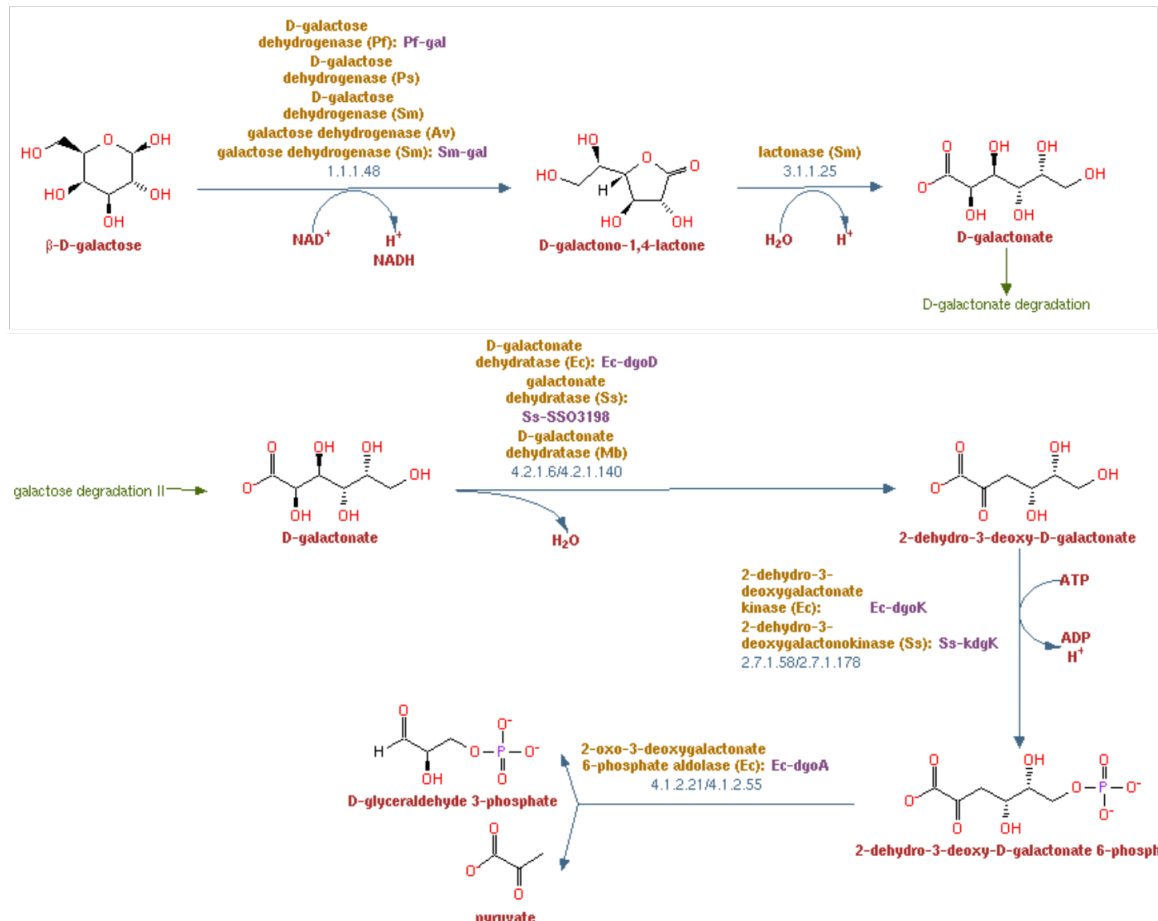


FIGURE 2.17: Voie de dégradation du galactose, voie de Leloir, tiré de <http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=GALDEG-PWY&detail-level=4> et <http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=GALACTCAT-PWY&detail-level=4>

Six réactions sont associées à cette deuxième voie :

- **R01094/R01097** : **D-galactose : $\text{NAD}^+/\text{NADP}^+$ 1-oxidoreductase** Le KO associé est le K00035, il est spécifique.
- **R03033** : **D-galactonate hydro-lyase** Les KO associés sont les K01684 et K05308, seuls le premier est spécifique.
- **R01064** : **2-dehydro-3-deoxy-6-phospho-D-galactonate D-glyceraldehyde-3-phospho-**

- lyase.** Les KO associés sont les K01631 et K11395. Seul le premier est spécifique.
- **R03387** : **ATP :2-dehydro-3-deoxy-D-galactonate 6-phosphotransferase.** Le KO associé est le K00883, il est spécifique.
 - **R10616** : **2-dehydro-3-deoxy-D-galactonate D-glyceraldehyde-lyase.** Le KO associé est le K11395, il n'est pas spécifique.

Bilan : Parmi toutes les réactions considérées, seules quatre satisfont nos critères de spécificité. La première voie peut être directement suivie par la voie EMP. La seconde crée du pyruvate et du D-glyceraldehyde 3 phosphate qui est un composé intermédiaire de la voie EMP.

L-arabinose. L'exploration de MetaCyc révèle 4 voies métaboliques. La voie numéro 2 est réalisée par des *fungi* peu vraisemblables. Les voies numéros 3 et 4 oxydent les co-facteurs sans produire d'ATP, elles sont de même associées respectivement à des bactéries du sol et à des *Archaea* vraisemblablement absents du microbiote intestinal humain.

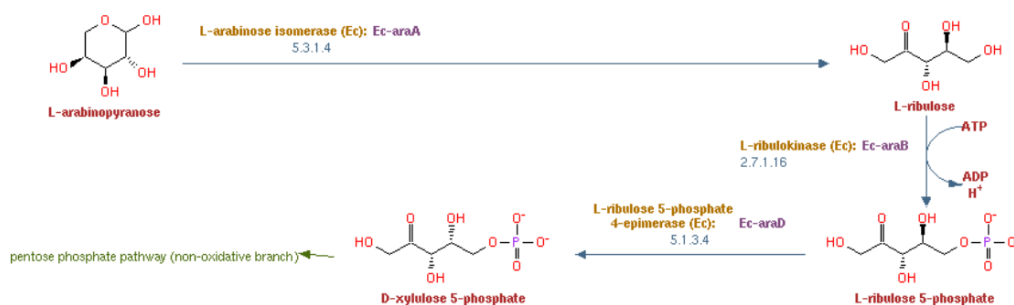


FIGURE 2.18: Voie de dégradation du L-arabinose, tiré de <http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=ARABCAT-PWY&detail-level=4>

Trois réactions sont associées à cette voie :

- **R01761** : **L-Arabinose aldose-ketose-isomerase.** Le KO associé est le K01804, il est spécifique.
- **R02439** : **ATP :L-ribulose 5-phosphotransferase.** Le KO associé est le K00853, il n'est pas spécifique et peut aussi catalyser du D-ribulose.
- **R05850** : **L-ribulose-5-phosphate 4-epimerase.** Les KO associés sont les K01786, K03077 et K03080. Ils sont spécifiques.

Bilan : Deux réactions sont conservées, elles mènent à la voie métabolique du pentose phosphate.

Fructose. MetaCyC dénombre une unique voie associée au fructose. L'exploration conjointe de KEGG met en avant plusieurs réactions alternatives permettant la transformation du D-Fructose en β -D-Fructose 6-phosphate, cependant les KO associés sont non spécifiques.

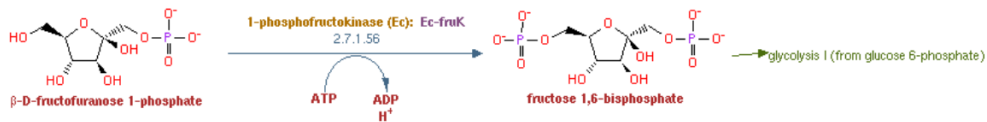


FIGURE 2.19: Voie de dégradation du Fructose, tiré de <http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY0-1314&detail-level=4>

Deux réactions sont associées à cette voie :

- **R00866** :ATP :D-fructose 1-phosphotransferase. Un seul KO non spécifique : K00846.
- **R02071** :ATP :D-fructose-1-phosphate 6-phosphotransferase. Un seul KO, spécifique : K00882.

Bilan : Une seule réaction est conservée. Le fructose 1,6-biphosphate fait partie de la voie EMP.

Entrée MetaCyc	URL de l'entrée	Critère de décision
Pathway : L-rhamnose dégradation II	http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-6713&detail-level=4	<i>Debaryomyces hansenii</i> et <i>Azotobacter vinelandii</i>
Pathway : L-rhamnose dégradation III	http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-6714&detail-level=4	<i>Sphingomonas</i>
Pathway : xylose dégradation III	http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-6760	<i>Haloferax volcanii</i>
Pathway : xylose dégradation IV	http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-7294	<i>Sulfolobus solfataricus</i>
Pathway : D-galacturonate dégradation II	http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-6486	Absence de KO
Pathway : D-galacturonate dégradation III	http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-6491	Absence de KO
Pathway : D-galactose dégradation III	http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-3821	<i>Embryophyta</i>
Pathway : D-galactose dégradation IV	http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-6693	<i>Aspergillus niger</i> , <i>Trichoderma reesei</i> et <i>Aspergillus nidulans</i>
Pathway : D-galactose dégradation V (Leloir pathway)	http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY66-422	stéro-spécificité
Pathway : L-arabinose dégradation II	http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-5515	<i>Aspergillus niger</i> et <i>Trichoderma reesei</i>
Pathway : L-arabinose dégradation III	http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-5517	<i>Azospirillum brasiliense</i> , <i>Pelomonas saccharophila</i> , <i>Herbaspirillum seropedicae</i> et <i>Bradyrhizobium japonicum</i>
L-arabinose dégradation IV	http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-7295	<i>Sulfolobus solfataricus</i>

TABLE 2.3: Tableau de référencement des voies métaboliques de MetaCyc qui ont été écartées, les organismes où les rangs taxonomiques référencés sont peu vraisemblables dans le microbiote intestinal humain.

2.2.2.4 Glycolyse anaérobie

Nous décrivons ici la deuxième partie du processus de glycolyse producteur d'ATP et de NADH. Ces processus prennent généralement en entrée du fructose 6-phosphate ou du glucose 6-phosphate et produisent du pyruvate. Nous dénombrons 3 processus principaux, et nous décrivons un 4^{ème} moins fréquent mais intéressant par son efficacité énergétique.

Embden-Meyerhof-Parnas. Cette voie métabolique est décrite précisément par la Figure 2.10. Les premières réactions associées au fructose sont non spécifiques, c'est aussi le cas de la dernière réaction de la voie métabolique produisant du pyruvate. Nous présentons les KO et réactions sélectionnées ci-dessous.

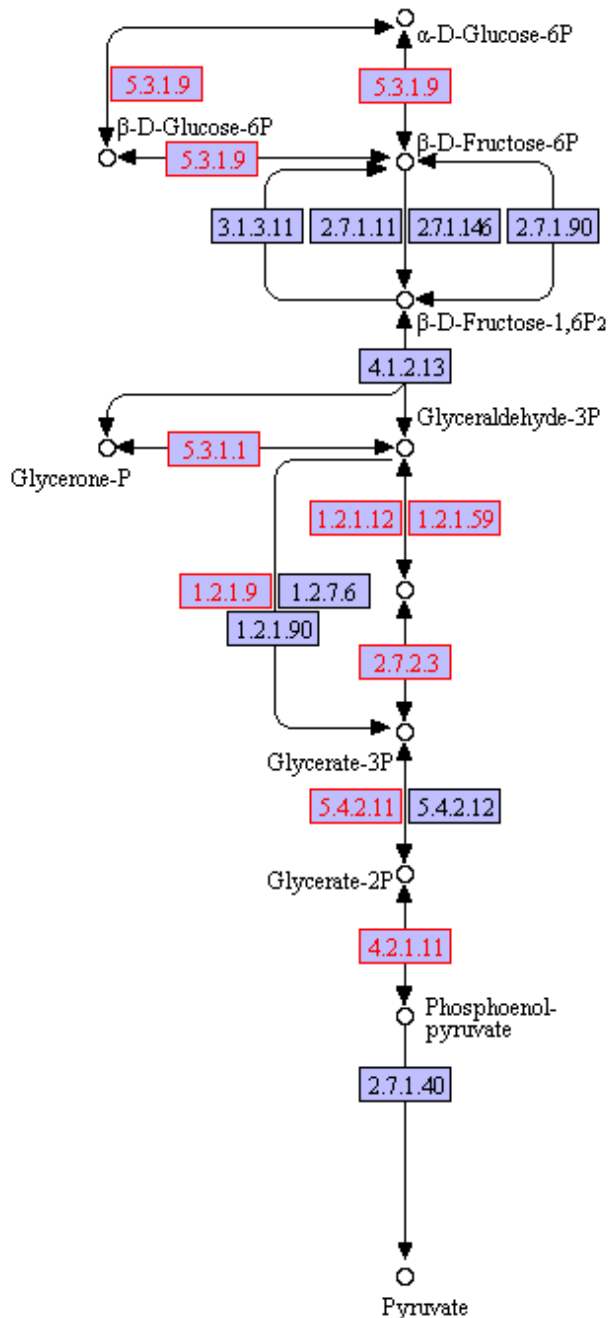


FIGURE 2.20: Voie d'Embden-Meyerhof-Parnas, les réactions sélectionnées sont colorées en rouge, chaque flèche est associée à un numéro EC. Tiré de http://www.genome.jp/kegg-bin/show_pathway?ko00010+K01803+K00150+K00134+K00927+K00131+K01834+K01689+K06859+K01810 62

Les 7 réactions sélectionnées sont les suivantes :

- R02739, R03321 et R02740 : **alpha-D-Glucose 6-phosphate ketol-isomerase**. Ces trois réactions sont spécifiques, elle permettent l'inter-conversion du glucose 6-phosphate en fructose 6-phosphate de manière stéréo-spécifique et sont toutes trois associées au même KO. Les enzymes associées sont capable de catalyser de façon spécifique l'ensemble de ces trois réactions. Les KO associés sont les K06859 et K01810. Le numéro EC est le 5.3.1.9.
- R01519 : **D-glyceraldehyde-3-phosphate aldose-ketose-isomerase**. Le KO associé est le K01803, le numéro EC est le 5.3.1.1.
- R01061 et R01063 : **D-glyceraldehyde-3-phosphate :NAD⁺/NADP⁺ oxidoreductase (phosphorylating)**. Les KO associés sont les K00150 et K00134, les nombres EC sont les 1.2.1.12 et 1.2.1.59.
- R01512 : **ATP :3-phospho-D-glycerate 1-phosphotransferase**. Le KO associé est le K00927, le nombre EC est le 2.7.2.3.
- R01058 : **D-glyceraldehyde 3-phosphate :NADP⁺ oxidoreductase**. Le KO associé est le K00131 et le nombre EC est le 1.2.1.9.
- R01518 : **D-phosphoglycerate 2,3-phosphomutase**. Le KO associé est le K01834 et le nombre EC est le 5.4.2.11.
- R00658 : **2-phospho-D-glycerate hydrolyase**. Le KO associé est le K01689 et le nombre EC est le 4.2.1.11.

Entner-Doudoroff. La voie d'Entner-Doudoroff possède un bilan similaire à celui de la glycolyse, avec la production de deux molécules de pyruvate pour la consommation d'une molécule de glucose, cependant seul un ATP est produit. En effet, l'un des deux pyruvates est obtenu directement par clivage d'une forme oxydée du glucose 6-phosphate. Bien que moins productrice d'énergie, la forme réduite d'un cofacteur de cette voie (NADPH) est utilisée ultérieurement dans des processus anaboliques.

Plusieurs voies alternatives existent et nous ne considérons que la voie entièrement phosphorylative, c'est à dire consommant le glucose 6-phosphate. Les voies d'Entner-Doudoroff et du Pentose-phosphate partagent un début similaire puisque nécessitant la production de D-gluconate 6-Phosphate à partir de D-Glucose 6-Phosphate.

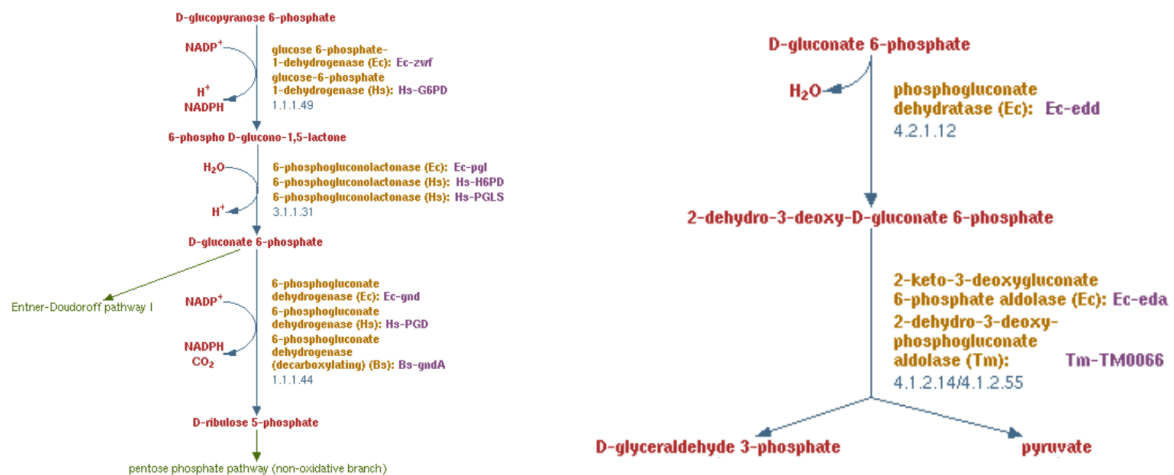


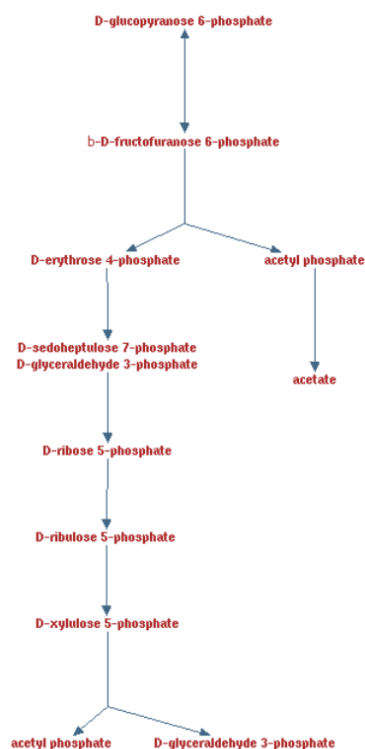
FIGURE 2.21: Voie d'Entner-Doudoroff, le schéma métabolique de gauche est commun à la voie d'Entner Doudoroff et du pentose phosphate, celui de droite lui est spécifique de la voie d'Entner-Doudoroff. Tirés respectivement de <http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=OXIDATIVEPENT-PWY> et de <http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=ENTNER-DOUDOROFF-PWY>

Les réactions sélectionnées sont les suivantes :

- R00835, R02736 et R10907 : **D-glucose-6-phosphate :NADP+/NAD+ 1-oxidoreductase.**
Le KO associé est le K00036.
- R02035 : **6-Phospho-D-glucono-1,5-lactone lactonohydrolase.** Les KO associés sont les K01057 et K07404.
- R02036 : **6-Phospho-D-gluconate hydro-lyase.** Le KO associé est le K01690.

Pentose-phosphate. La voie du Pentose-phosphate est au même titre que l'EMP et la voie d'Entner-Doudoroff, une voie majeure de l'oxydation des monosaccharides en sucres simples. Elle permet l'interconversion du glucose 6-phosphate en différents sucres de tailles différentes tels que le ribose à 5 carbones, l'érythrose à 4 carbones ou le sedoheptulose à 7 carbones. Bien que permettant la production d'ATP, cette voie métabolique est aussi exploitée à des fins anaboliques, en particulier en ce qui concerne le Ribose (ARN). Les réactions associées à cette voie sont peu spécifiques et nous n'en avons sélectionnée aucune.

Bifidobacterium shunt. La voie du Bifidobacterium shunt est caractérisée par la production d'actyl phosphate à partir de sucres phosphatés non oxydés là où les voies classiques passent par la production de NADH et de CO₂. La transformation de l'acetyl-phosphate en Acétate est accompagnée par la production d'ATP ce qui permet au total une production de 3 ATP par molécule de glucose, un rendement supérieur à EMP.



Deux réactions sont spécifiques de cette voie :

- R00761 : **D-fructose-6-phosphate D-erythrose-4-phosphate-lyase**. Le KO associé est le K01621.
- R01621 : **D-xylulose 5-phosphate D-glyceraldehyde-3-phosphate-lyase**. Le KO associé est le même que précédemment, le K01621.

Ces réactions sont spécifiquement catalysées par une unique enzyme ce qui fait du KO associé un bon marqueur de la voie.

FIGURE 2.22: Voie du Bifidobacterium shunt. Tiré de <http://biocyc.org/META/NEW-xIMAGE?type=PATHWAY&object=P124-PWY>

2.2.2.5 Du pyruvate aux accepteurs finaux d'électrons

Nous décrivons ici la sélection des marqueurs associés aux dernières étapes de la fermentation. Ce sont souvent des mécanismes de régénération des co-facteurs mais parfois aussi de production d'ATP.

Lactate. La fermentation du pyruvate en lactate permet la régénération de NADH en NAD⁺. Le lactate est excrété par la cellule.

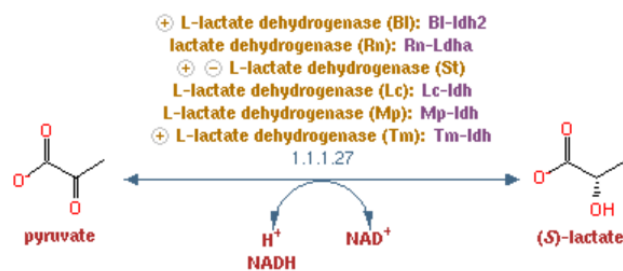


FIGURE 2.23: Régénération du NADH par réduction du pyruvate en lactate. Tiré de <http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-5481&detail-level=4>

La réaction associée est la R00703 : (S)-Lactate :NAD⁺ oxidoreductase. Le KO correspondant est le K00016.

Acétate. L'acétate est un acide gras à chaîne courte excrété par la cellule. La voie de l'acétate est associée à la production d'ATP. L'oxydation du pyruvate permet la production jointe d'acetyl-CoA et de CO₂. Cette réaction nécessite donc de consommer des co-facteurs oxydés plutôt que de régénérer les co-facteurs réduits. La production d'Acétate et donc d'ATP est généralement liée à la réalisation de réaction d'oxydation des co-facteurs. L'oxydation du pyruvate peut être réalisée par le NAD⁺ ; elle peut se faire par le biais des ferrédoxines que nous avons présentées plus haut. Les ferrédoxines permettent de transférer la réduction aux molécules de H₂ à condition que la pression partielle de ce gaz dans le milieu ne soit pas trop importante. Enfin le dernier mécanisme est lié à la production de formate que nous présentons dans paragraphe suivant.

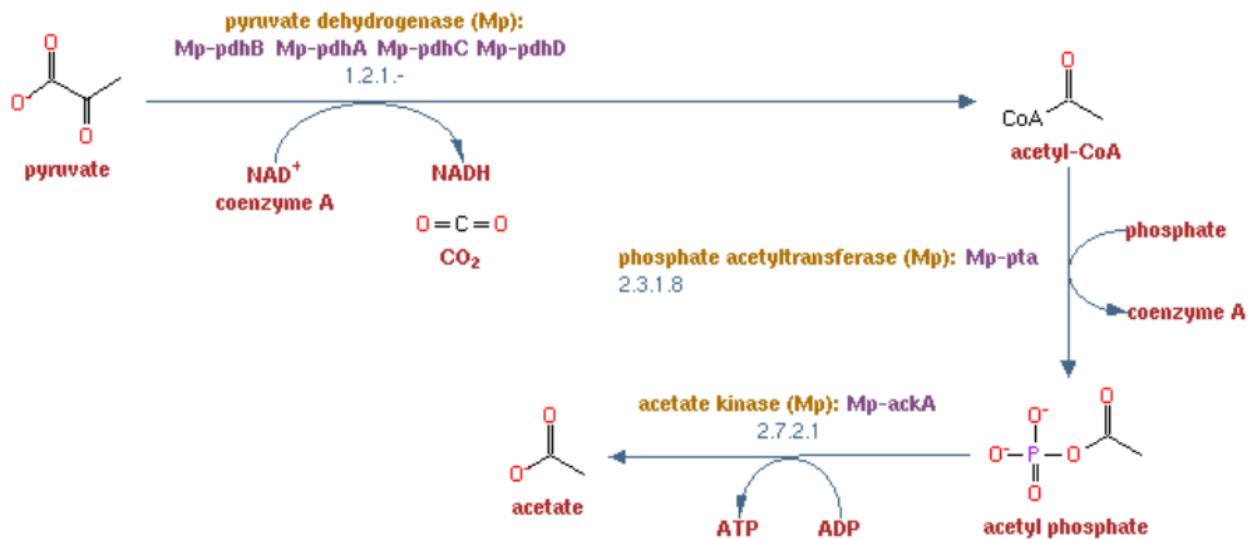


FIGURE 2.24: Illustration de la production d'acétate à partir de pyruvate. Ici, le pyruvate est oxydé par le biais du NAD⁺. Tiré de <http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-5482&detail-level=4>

Les réactions associées à cette voie sont les :

- R02569 : **acetyl-CoA :enzyme N6-(dihydrolipoyl)lysine S-acetyltransferase** le KO associé est le K00627. Cette réaction permet le passage du pyruvate à l'acetyl-CoA par réduction de NAD⁺.
- R01196 : **pyruvate :flavodoxin/ferredoxin 2-oxidoreductase** Les KO associé sont les K03737, K00169, K00169, K00170, K00171 et K00172. Cette réaction permet le passage du pyruvate à l'acetyl-CoA par réduction de férédoxines.
- R00230 : **acetyl-CoA :phosphate acetyltransferase**. Les KO associés sont les K04020, K00625, K15024 et K13788.
- R00315 : **ATP :Acétate phosphotransferase**. Le KO associé est le K00925.

Formate. La production de formate est conjointe à la production d'acetyl-CoA, cette réaction ne s'accompagne pas de dégagement gazeux ni de réduction de cofacteur : le formate est la forme réduite du CO₂ et est excrété par la cellule.

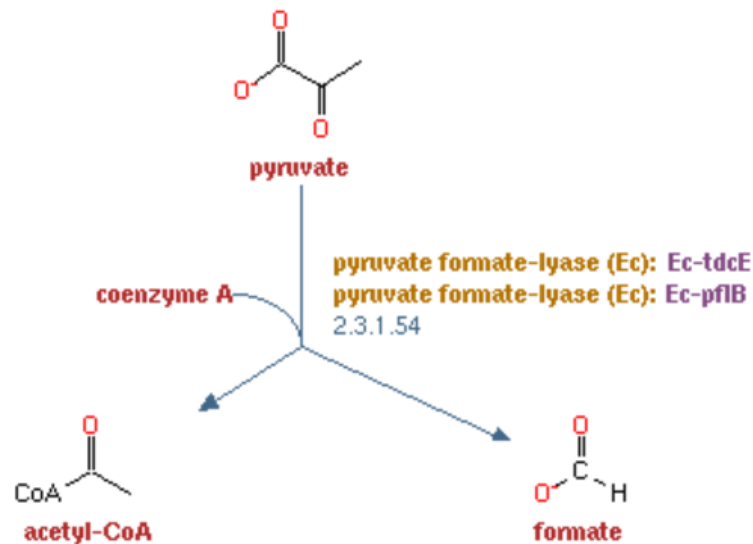


FIGURE 2.25: Production de formate lors de l'oxydation du pyruvate en acetyl-CoA. Tiré de <http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-5485&detail-level=4>

La réaction associée est la R00212 : **Acetyl-CoA :formate C-acetyltransferase**. Le KO correspondant est le K00656.

Butanoate. Le Butanoate est un acide gras à chaîne courte excrété par la cellule. Sa production fait intervenir des réactions associées à l'élongation des acide gras par ajout d'acetyl-CoA et les réactions considérées sont alors peu spécifiques. Sa production nécessite la présence de deux molécules d'acetyl-CoA qui produisent un acetoacetyl-CoA dont la réduction répétée permet la production de Butanoyl-CoA et la régénération des co-facteurs nécessaire à l'obtention des acetyl-CoA. Enfin, les dernières étapes de cette voie permettent la production d'ATP et de Butanoate.

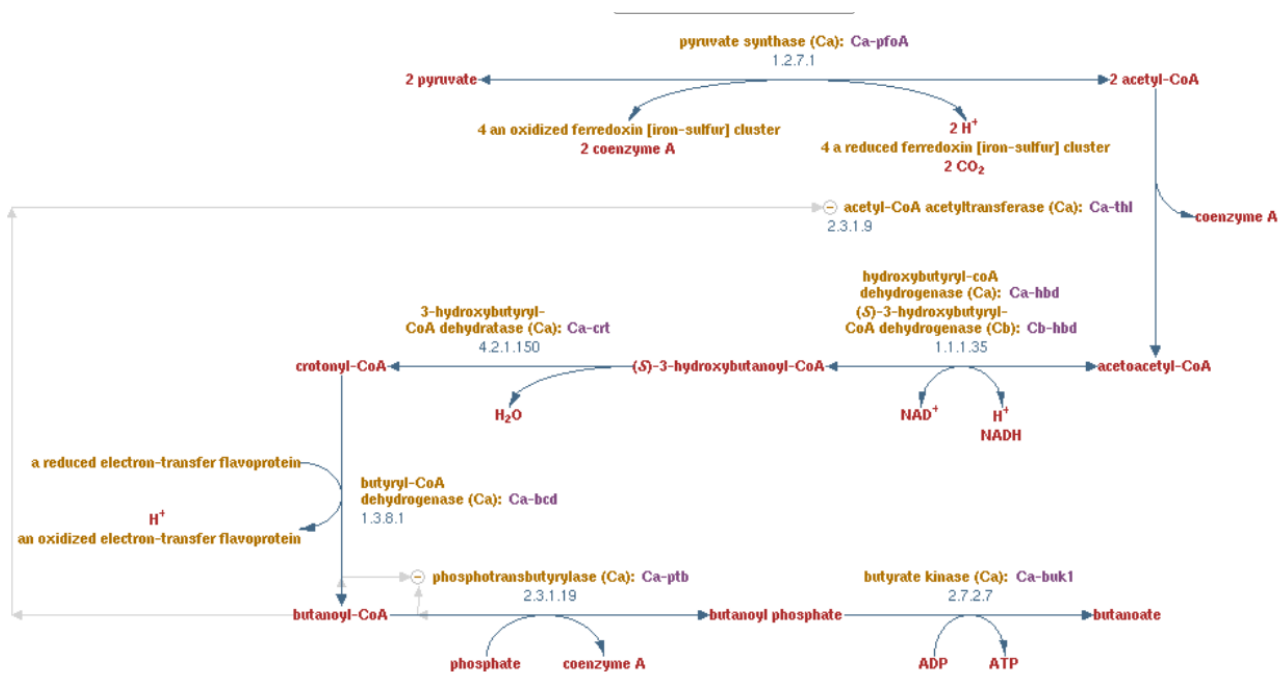


FIGURE 2.26: Voie du Butanoate. Tiré de <http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=CENTFERM-PWY>

Les réactions sélectionnées sont les suivantes.

- R00238 : **Acetyl-CoA :acetyl-CoA C-acetyltransferase**. Le KO associé est le K00626.
- R01174 : **butanoyl-CoA :phosphate butanoyltransferase**. Le KO associé est le K00634.
- R01688 : **ATP :Butanoate 1-phosphotransferase**. Le KO associé est le K00929.

CoA-transferase. Nous présentons ici un cas particulier associé à la production jointe d'Acétate et de Butanoate dans le microbiote intestinal. Dans cette voie, deux molécules de pyruvates sont utilisées pour produire du butanoyl-CoA. Une transferase intervient alors permettant de transférer le groupement CoA à une molécule d'Acétate. Dès lors, c'est l'acetyl-CoA qui est phosphorylé avant de produire de l'ATP.

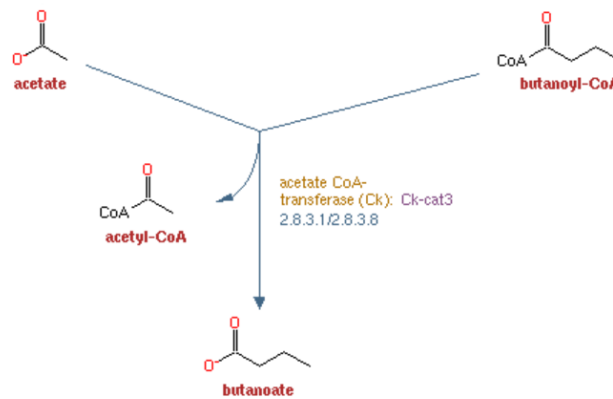


FIGURE 2.27: Production d'acétate par transférase. Tiré de <http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-5676>

La réaction associée est la R01179 : **butanoyl-CoA :Acétate CoA-transferase**. Le KO correspondant est le K01034.

Propane-1,2-diol. L'acetoacetyl-CoA peut produire du propane-1,2-diol après réductions successives. Les KO associés sont peu spécifiques.

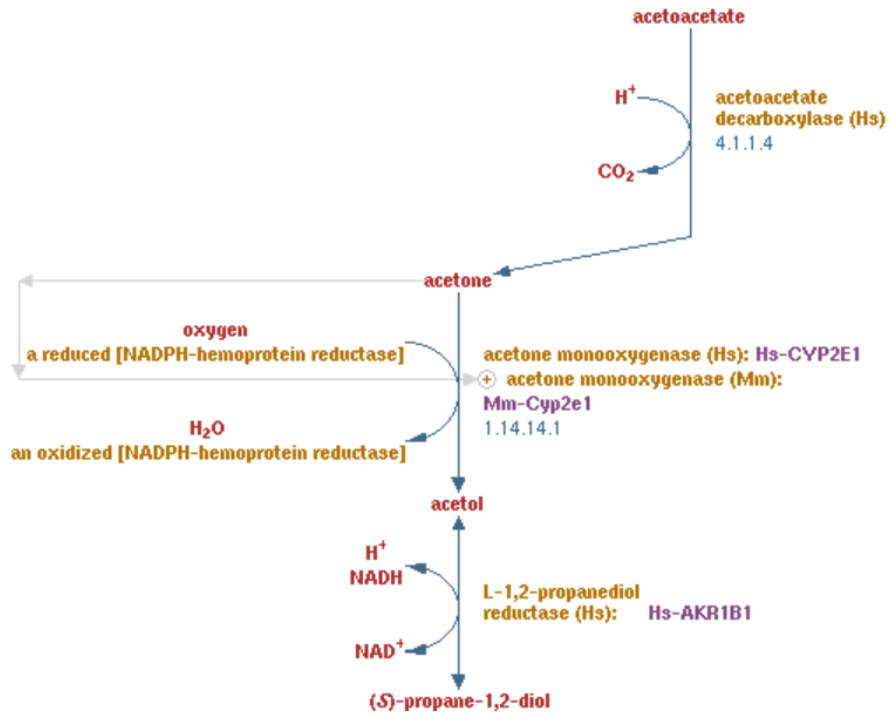


FIGURE 2.28: Production de Propane-1,2-diol. Tiré de <http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-7466>

La réaction associée est la R01366 : **Acetoacétate carboxy-lyase** pour laquelle le KO correspondant est le K01574.

Butane-diol. Le butane-diol est un accepteur final d'électron excrété par la cellule. Après condensation de deux molécules de pyruvate en acetolactate, ce métabolite est réduit entre 2 et 3 fois de façon plus ou moins stéréo-spécifique. Cette voie permet de régénérer les cofacteurs réduits de la glycolyse.

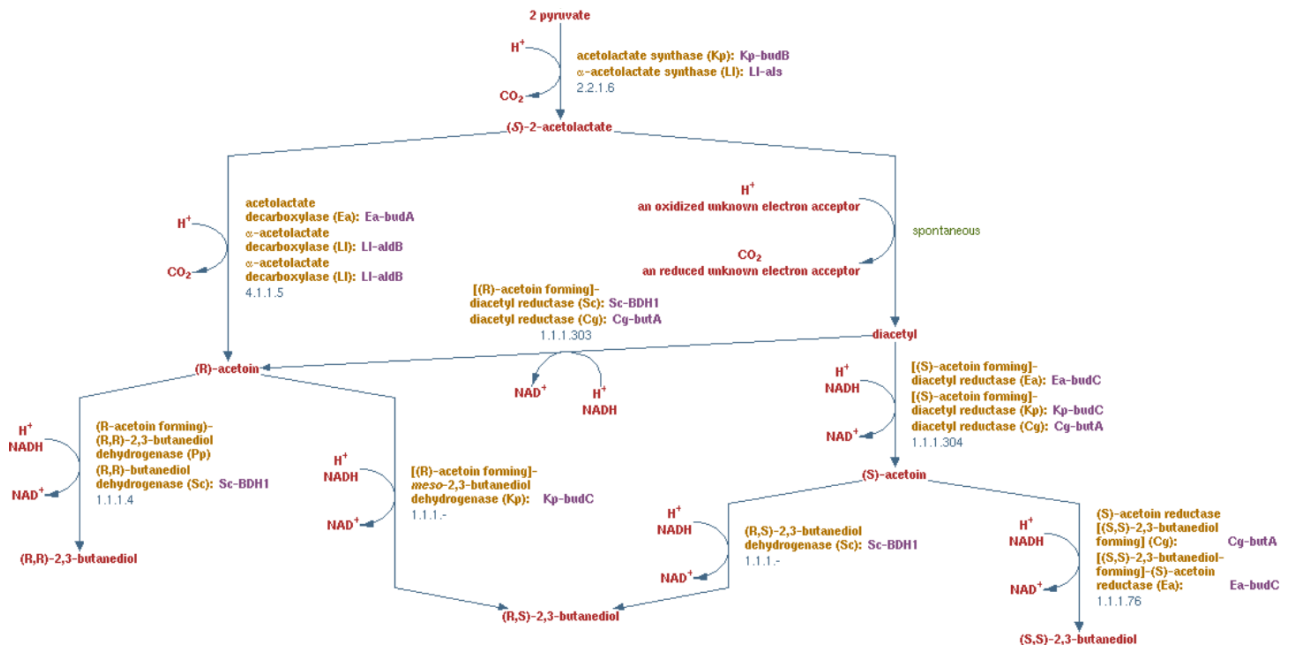


FIGURE 2.29: Réductions successives de l'acetolactate en différents butane-diols. Tiré de <http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-6396&detail-level=2>

Les réactions sélectionnées sont les suivantes.

- R02855 : **(R)-Acetoin :NAD⁺ oxidoreductase**. Les KO associés sont les K00004 et K03666.
- R02946 : **(R,R)-Butane-2,3-diol :NAD⁺ oxidoreductase**. Les KO associés sont les K00004 et K03666.
- R10504 : **meso-2,3-butanediol :NAD⁺ oxidoreductase**. Les KO associés sont les K00004 et K03666.

Éthanol. L'éthanol est un accepteur final d'électron excrété par la cellule. La production d'éthanol permet de régénérer les co-facteurs de la glycolyse puisque sa synthèse à partir de pyruvate consomme 1 à 2 NADH selon le mode de production d'acetyl-CoA. Les enzymes associées ne sont pas spécifiques. Nous ne pouvons pas représenter cette voie métabolique.

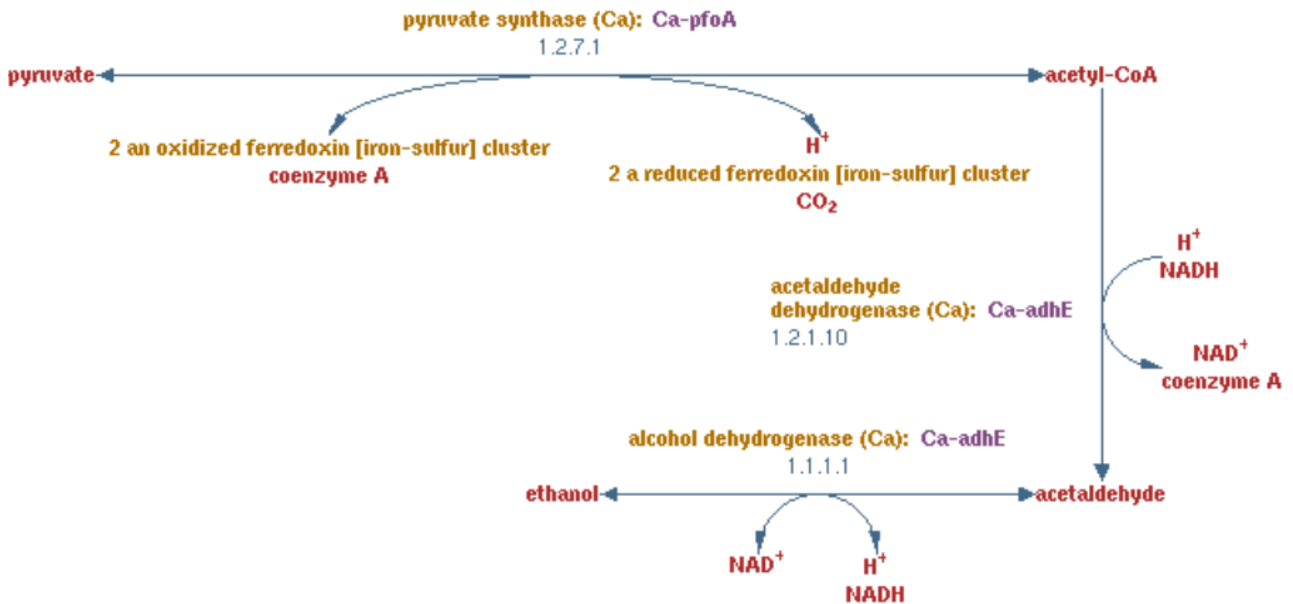


FIGURE 2.30: Voie de l'éthanol. Tiré de <http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-6396&detail-level=2>

2.2.2.6 Réutilisation des produits de fermentation

Nous présentons les voies syntrophiques associées à la réutilisation de composés résultant du catabolisme des sucres simples. Les plus intéressantes sont l'acétogénèse et la méthanogénèse qui consomment le H₂ permettant potentiellement aux ferrédoxines de régénérer le NADH.

Acétogénèse. L'acétogénèse réductrice consiste en l'utilisation de CO₂ et parfois de formate pour synthétiser de l'acetyl-CoA. Si l'acetyl-CoA est utilisé pour produire de l'ATP, cette voie ne produit pas d'ATP. Cependant, elle permet de réoxyder jusqu'à 4 cofacteurs réduits tels que le NADH ou le H₂ présent dans l'environnement. Dans le microbiote intestinal, cette voie est notamment utilisée par des organismes (*B. hydrogenotrophica*) capables de produire trois molécules d'acétate et 4 ATP par molécule de glucose, soit deux fois plus d'ATP que dans EMP.

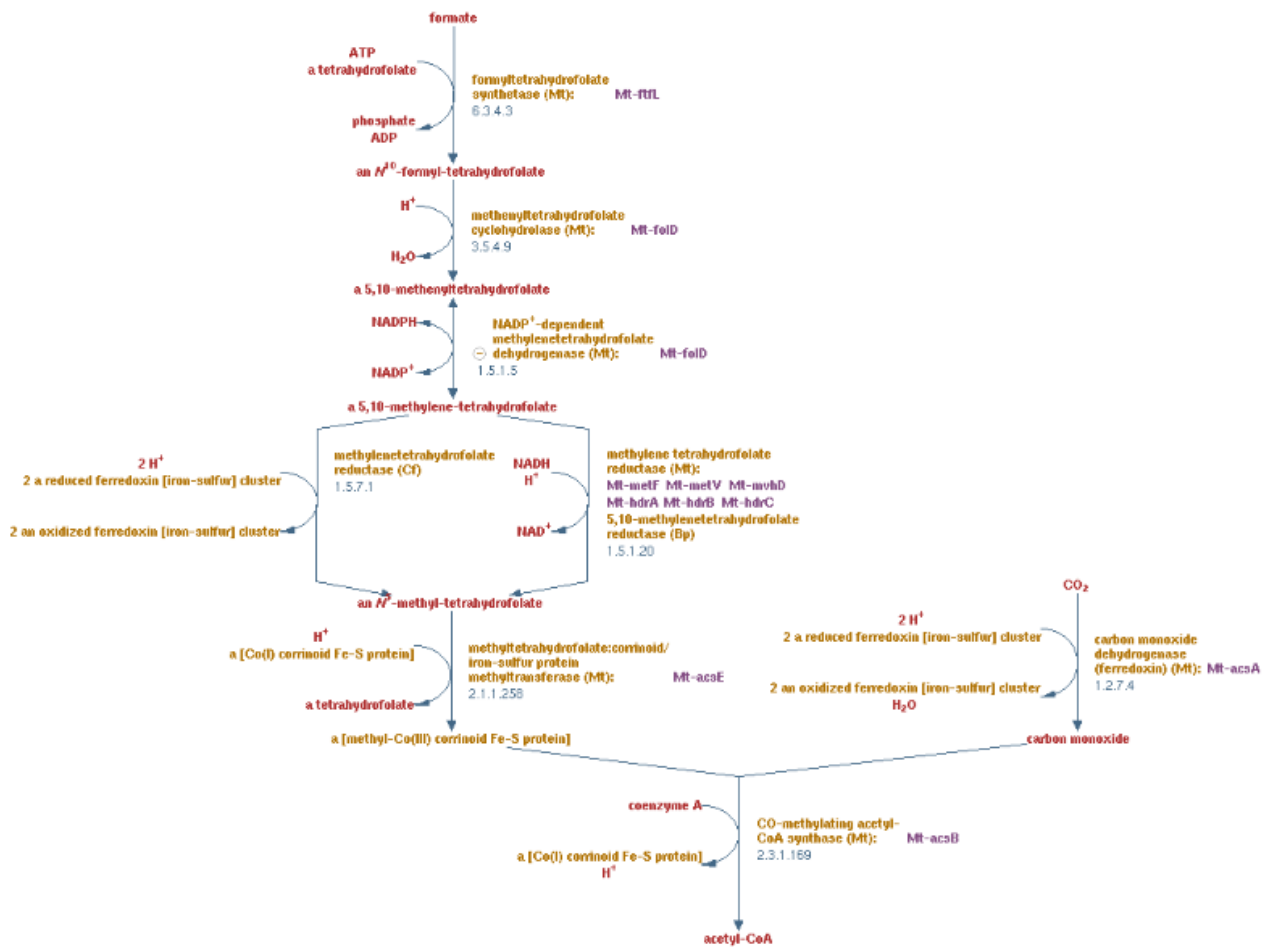


FIGURE 2.31: Voie de l'acétogénèse réductrice à partir de formate et de CO₂. Tiré de <http://metacyc.org/META/NEW-IMAGE?type=NIL&object=CODH-PWY>

Les réactions retenues sont les suivantes.

- R00943 :Formate :tetrahydrofolate ligase (ADP-forming), R01220 : 5,10-methylene tetrahydrofolate :NADP+ oxidoreductase et R01655 : 5,10-Methenyltetrahydrofolate 5-hydrolase (decyclizing). Les KO associés sont les K01938, K00288 et K01491.
- R01224 et R07168 : 5-methyltetrahydrofolate :NAD+/NADP+ oxidoreductase. Le KO associé est le K00297.

Méthanogénèse. La méthanogénèse est une voie spécifique des *Archaea*. C'est aussi une voie métabolique atypique pour laquelle chaque enzyme est très spécifique. Elle consiste en l'utilisation conjointe de CO_2 et de H_2 pour produire du méthane.

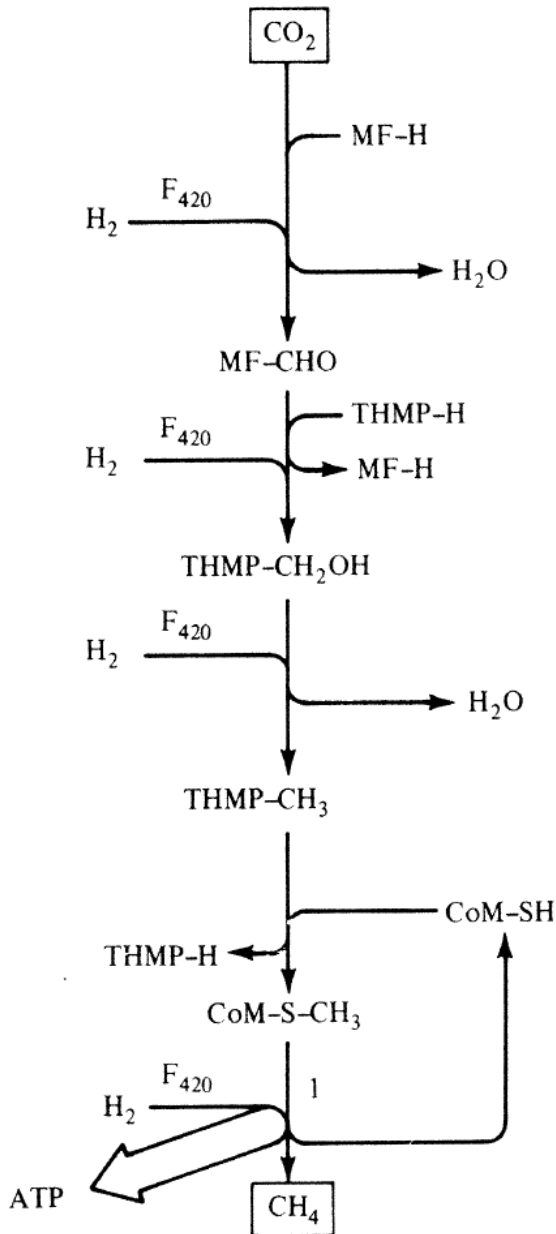


FIGURE 2.32: Voie de la Méthanogénèse. Gottschalk (1986)

Les réactions associées sont les suivantes.

- R03390 : **Formyl-methanofuran :5,6,7,8-tetrahydro-methanopterin 5-formyltransferase.** Le KO associé est le K00672.
- R03464 : **5,10-Methenyl-tetrahydro-methanopterin 10-hydrolase.** Le KO associé est le K01499.
- R04456 : **5,10-methylene-tetrahydro-methanopterin : coenzyme-F420 oxidoreductase.** Le KO associé est le K00319.
- R04455 : **hydrogen :N5,N10-methenyl tetraydro-methanopterin-oxidoreductase.** Le KO associé est le K13942.
- R04464 : **5,10-Methylene-tetrahydro-methanopterin : coenzyme-F420 oxidoreductase.** Le KO associé est le K00320.
- R04347 : **5-Methyl-5,6,7,8-tetrahydro-methanopterin : 2-mercapto ethane sulfonate 2-methyltransferas.** Les KO associés sont les K00577, K00578, K00579, K00580, K00581, K00582, K00583 et K00584.
- R04541 : **2-(methylthio)-ethanesulfonate : N-(7-thioheptanoyl)-3-O-phosphothreonine S-(2-sulfoethyl)-thiotransferase.** Les KO associés sont les K00399 K00401 K00402.

Production de propanoate. Le propanoate est un acide gras à chaîne courte, c'est un accepteur d'électron final excrété par la cellule. La production de propanoate peut se faire selon trois voies que nous présentons ci-après. Elles consistent à utiliser du lactate ou du L-lactaldéhyde afin de produire de l'ATP. Dans les deux premières voies, le lactate est d'une part réoxydé en pyruvate puis génère de l'ATP en synthétisant de l'acétate, et d'autre part le lactate est utilisé dans la synthèse de propanoate régénérant de ce fait les co-facteurs nécessaires à la production d'acétate et donc d'ATP. De la même façon, le lactaldéhyde est d'une part réduit en propanol et d'autre part utilisé pour synthétiser du propanoyl-CoA dont la phosphorylation permet ensuite la production d'ATP.

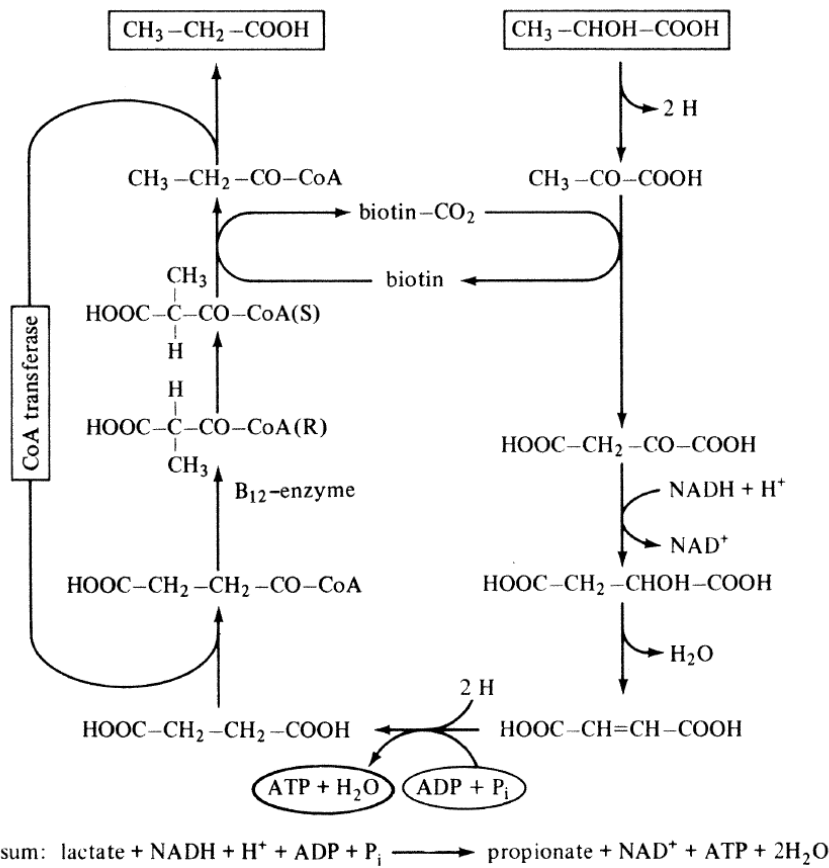


FIGURE 2.33: Production de propanoate par la voie du succinate. Tiré de Gottschalk (1986)

La réaction sélectionnée est la R00833 : **(R)-Methylmalonyl-CoA CoA-carboxylmutase**. Les KO associés sont les K01847, K01848 et K01849.

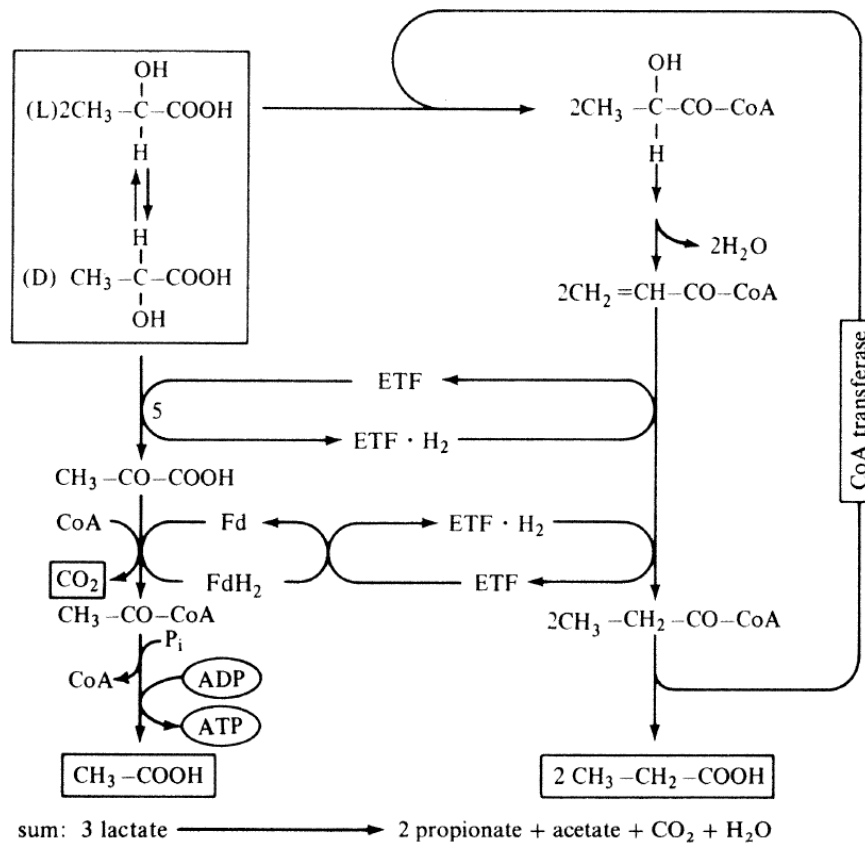


FIGURE 2.34: Production de propanoate par la voie de l'acrylate. Tiré de Gottschalk (1986)

La réaction sélectionnée est la R01449 : lactoyl-CoA :propanoate CoA-transferase, le KO correspondant est le K01026.

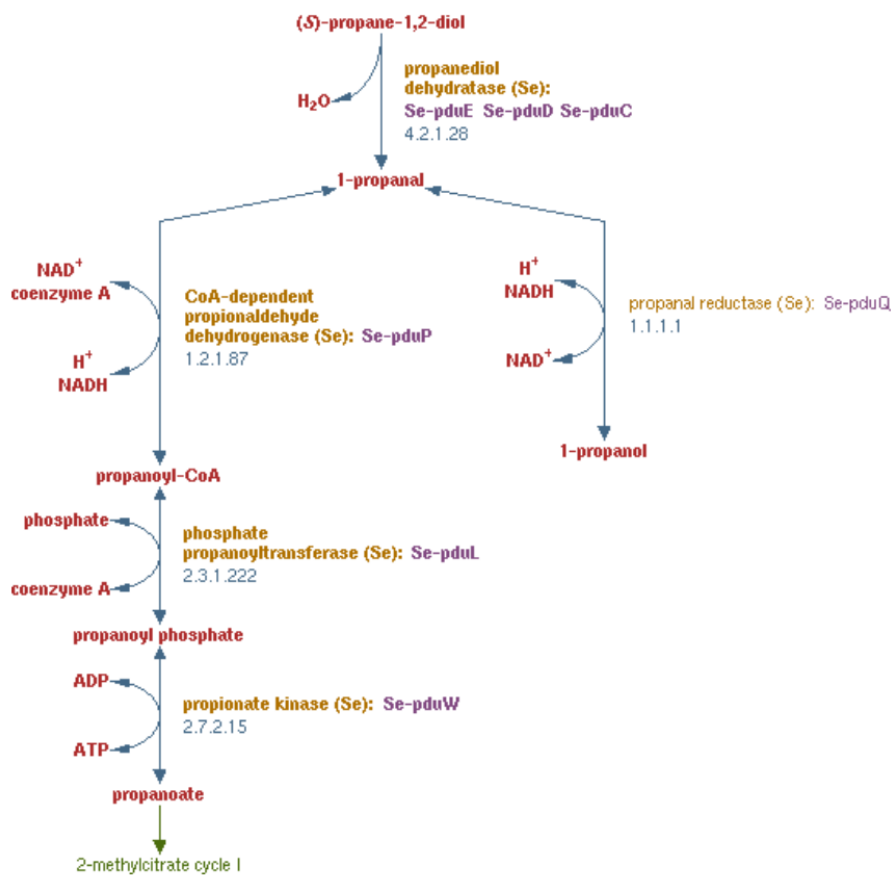


FIGURE 2.35: Voie du L-lactaldéhyde. Tiré de <http://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-7013>

Les réactions associées sont les suivantes :

- R02376 : **propane-1,2-diol hydro-lyase**. Les KO associés sont les K01699, K13919 et K13920.
- R09097 : **propanal :NAD+ oxidoreductase (CoA-propanoylating)**. Le KO associé est le K13922.

2.2.2.7 Graphe topologique des marqueurs fonctionnels sélectionnés

Certaines réactions sont associées à plusieurs KO. Dans certains cas, c'est parce que l'enzyme est composée de multiples parties ou modules interdépendants correspondant à chaque KO. Dans cette situation, nous considérons que c'est l'abondance moyenne de ces KO qui peut être le marqueur d'une activité enzymatique *potentielle*. Dans d'autres situations, nous consta-

tons la présence de KO différents correspondant à des versions différentes de la même enzyme. Nous considérons alors que chacun des deux KO est un marqueur d'une activité catalytique *potentielle*. Suite au travail de sélection et suite à ces précisions, nous obtenons une liste de 61 marqueurs fonctionnels du catabolisme des monosaccharides et des voies syntrophiques associées. Ces marqueurs sont numérotés de 1 à 61 et légendés de manière exhaustive dans le Tableau 2.4.

Nous présentons l'ensemble des marqueurs sélectionnés sous la forme d'un graphe orienté en Figure 2.36. Il relie les marqueurs fonctionnels aux métabolites auxquels ils sont associés. Les 61 arêtes sont les marqueurs fonctionnels et les 43 nœuds sont les métabolites consommés ou produits par les réactions associées aux marqueurs. Les métabolites sont répartis en 25 métabolites intracellulaires colorés en bleu et 18 métabolites extra cellulaires colorés en rouge. Il est important de noter que ce graphe résume des relations de proximité entre les marqueurs fonctionnels, et que nous ne l'utilisons pas pour représenter des flux métaboliques comme cela peut se faire classiquement.

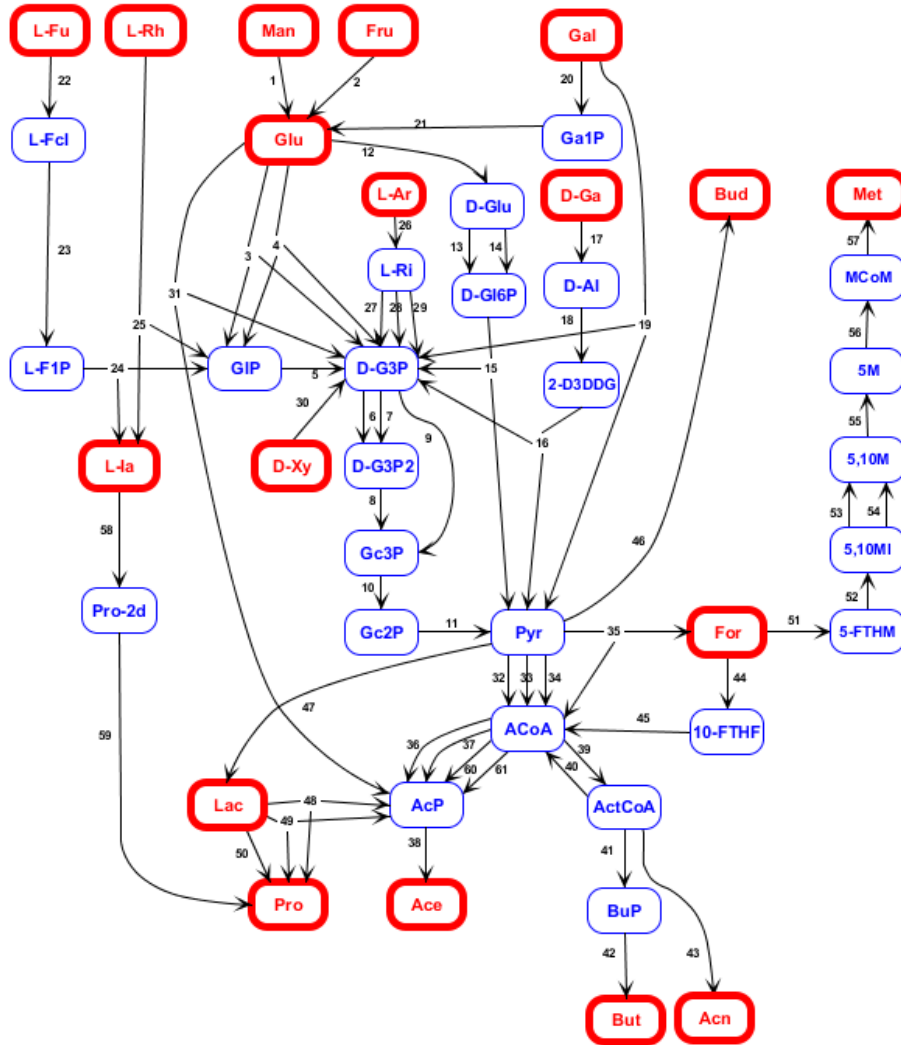


FIGURE 2.36: Représentation topologique des liens entre marqueurs fonctionnels au moyen des métabolites associés. Chaque flèche est associée à un des 61 marqueurs fonctionnels de la catalyse des monosaccharides. Chaque nœud est un métabolite, les métabolites extra-cellulaires sont colorés en rouge et les métabolites intra-cellulaires sont colorés en bleu.

De part le nombre important de réactions et KO écartés, la représentation de l'ensemble des voies métaboliques sous la forme d'un unique graphe connexe nécessite certaines approximations. A titre d'exemple, prenons la Figure 2.37 qui illustre la partie haute de la voie EMP décrite précédemment. Le schéma métabolique de gauche présente une version détaillée des métabolites et enzymes associés avec, en rouge, ceux que nous avons sélectionnés. Le schéma

métabolique de droite correspond à la réaction simplifiée. Cet exemple permet d'illustrer différents cas de figures rencontrés lors de la construction du graphe 2.36. Tout d'abord une unique enzyme permet l'interconversion du α -D-glucose 6P en β -D-fructose 6P et en β -D-Glucose 6P. Les deux KO associés à cette interconversion sont les marqueurs du début de la glycolyse. De même, de par la sélection stricte, les marqueurs associés au passage du β -D-fructose 6P au Glycerone-P et Glyceraldehyde 3P sont manquants.

- Nous utilisons alors les deux KO précédents comme marqueurs du passage du β -D-fructose 6P au Glycerone-P et Glyceraldehyde 3P.
- Nous amalgamons α -D-glucose 6P, β -D-fructose 6P et β -D-Glucose 6P en un unique métabolite abusivement nommé glucose.
- De plus, les deux KO associés à l'interconversion sont des versions différentes de la même enzyme et c'est pourquoi nous les représentons sous la forme de deux arêtes différentes.

De manière plus générale, certains métabolites sont assimilés en un unique métabolite permettant au graphe d'être connexe. De plus, un unique marqueur peut être utilisé comme marqueur d'un potentiel de réalisation de plus d'une réaction. La liste complète des simplifications est disponible dans le Tableau 2.4.

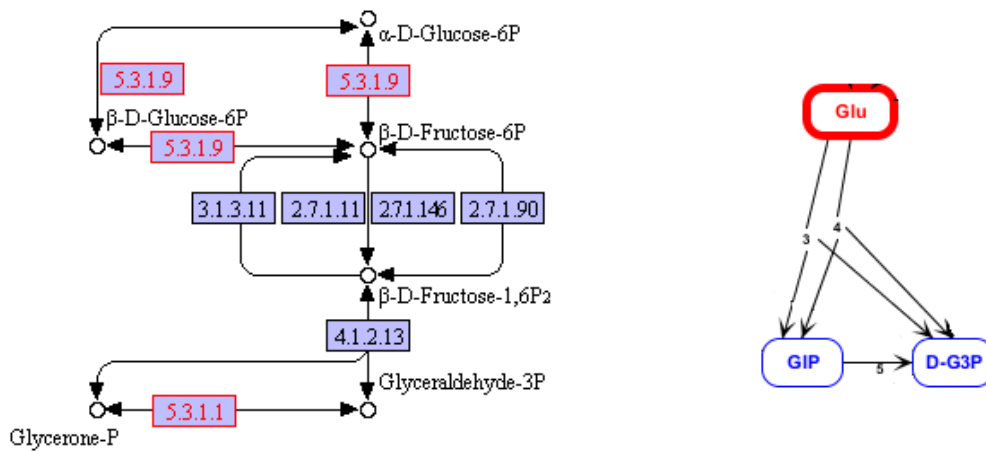


FIGURE 2.37: Illustration du processus de construction du graphe connexe 2.36 associé aux 61 KO sélectionnés. La figure de gauche correspond à la partie haute de la glycolyse présentée en 2.20 et la figure de droite en est sa représentation en n'utilisant que 3 KO différents.

TABLE 2.4: Tableau des marqueurs fonctionnels choisis. Ce tableau permet d'énumérer les simplifications permettant la construction du graphe 2.36.

Nb	Description	Convention de nommage	Réaction associé au KO sélectionnés	Ensemble des réaction représentées par le marqueur	Code Réaction	Code KO
1	Mannose→Glucose	Mannose est le D-Mannose6P, Glucose est utilisé comme nom générique pour plusieurs métabolites proches. Ici il remplace le β -D-Fructose6P	D-Mannose6P \Leftrightarrow β -D-Fructose6P	/	R01819	K01809
2	Fructose→Glucose	Le Fructose est le D-Fructose1P, Ici Glucose correspond au β -D-Fructose 1,6-biphosphate	ATP + D-Fructose1P \Leftrightarrow ADP + β -D-Fructose 1,6-biphosphate	/	R02071	K00882
3	Glucose→Glyceraldehyde3P + Dihydroxyacetone P	Ici Glucose remplace β -D-Fructose6P et β -D-Fructose 1,6-biphosphate	β -D-Glucose6P \Leftrightarrow β -D-Fructose6P \Leftrightarrow α -D-Glucose6P	β -D-Fructose6P \Leftrightarrow β -D-Fructose 1,6-biphosphate \Leftrightarrow Dihydroxyacetone P + D-Glyceraldehyde3P	R03321 R02740 R02739	K06859
4	Glucose→Glyceraldehyde3P + Dihydroxyacetone P				R03321 R02740 R02739	K01810
5	Dihydroxyacetone P→Glyceraldehyde3P		/	/	R01015	K01803
6	Glyceraldehyde3P→Glycerate1,3diP		/	/	R01063 R01061	K00150
7	Glyceraldehyde3P→Glycerate1,3diP		/	/	R01061	K00134
8	Glycerate1,3diP→Glycerate3P		/	/	R01512	K00927
9	Glyceraldehyde3P→Glycerate3P		/	/	R01058	K00131
10	Glycerate3P→Glycerate2P		/	/	R01518	K01834
11	Glycerate2P→Pyruvate		2-Phospho-D-glycerate \Leftrightarrow Phosphoenolpyruvate	2-Phospho-D-glycerate \Leftrightarrow Phosphoenolpyruvate \Leftrightarrow Pyruvate	R00658	K01689
12	Glucose→D-Glucono-1,5-lactone6P	Ici Glucose correspond au β -D-Glucose6P	β -D-Glucose6P + NADP+ \Leftrightarrow D-Glucono-1,5-lactone6P	/	R00835 R02736	K00036
13	D-Glucono-1,5-lactone6P→D-Gluconate6P		/	/	R02035	K01057
14	D-Glucono-1,5-lactone6P→D-Gluconate6P		/	/	R02035	K07404
15	D-Gluconate6P→Pyruvate+Glyceraldehyde3P		D-Gluconate6P \Leftrightarrow 2-Dehydro-3-deoxy-D-Gluconate6P	D-Gluconate6P \Leftrightarrow 2-Dehydro-3-deoxy-D-Gluconate6P \Leftrightarrow Pyruvate+ Glyceraldehyde3P	R02036	K01690
16	2-Dehydro-3-deoxy-D-gluconate→Pyruvate+Glyceraldehyde3P		2-Dehydro-3-deoxy-D-gluconate \Leftrightarrow 2-Dehydro-3-deoxy-6-phospho-D-gluconate	2-Dehydro-3-deoxy-D-gluconate \Leftrightarrow 2-Dehydro-3-deoxy-6-phospho-D-gluconate \Leftrightarrow Pyruvate+ Glyceraldehyde3P	R01541	K00874
17	D-galacturonate→D-Altronate		D-Tagaturonate \Leftrightarrow D-Altronate	D-galacturonate \Leftrightarrow D-Tagaturonate \Leftrightarrow D-Altronate	R02555	K00041
18	D-Altronate → 2-Dehydro-3-deoxy-6-phospho-D-gluconate		/	/	R01540	K01685
19	Galactose→Pyruvate+Glyceraldehyde3P		2-Dehydro-3-deoxy-D-galactonate \Leftrightarrow 2-Dehydro-3-deoxy-6-phospho-D-galactonate	D-galactose \Leftrightarrow D-Galactonono-1,4/5-lactone \Leftrightarrow D-Galactonate \Leftrightarrow 2-Dehydro-3-deoxy-D-galactonate \Leftrightarrow 2-Dehydro-3-deoxy-6-phospho-D-galactonate \Leftrightarrow Pyruvate + D-Glyceraldehyde3P	R03387	K00883
20	Galactose→Galactose1P	Galactose correspond au D-Galactose, Galactose1P correspond au α -D-Galactose1P	D-Galactose \Leftrightarrow α -D-Galactose	D-Galactose \Leftrightarrow α -D-Galactose \Leftrightarrow α -D-Galactose1P	R01092	K00849
21	Galactose1P→Glucose	Ici Glucose correspond au α -D-Glucose6P	UDP-glucose + α -D-Galactose1P \Leftrightarrow D-Glucose1P + UDP- α -D-galactose	α -D-Galactose1P \Leftrightarrow UDP- α -D-galactose \Leftrightarrow UDP-glucose \Leftrightarrow D-Glucose1P \Leftrightarrow α -D-Glucose6P	R00955	K00965

Nb	Description	Convention de nommage	Réaction associée au KO sélectionnés	Ensemble des réaction représentées par le marqueur	Code Réaction	Code KO
22	L-Fucose→L-Fucose		/	/	R03163	K01818
23	L-Fucose→L-Fucose1P		/	/	R03241	K00879
24	L-Fucose1P→Dihydroxyacetone Lactaldehyde	P+L-	/	/	R02262	K01628
25	L-Rhamnose→Dihydroxyacetone Lactaldehyde	P+L-	L-Rhamnose ⇌ L-Rhamnulose	L-Rhamnose ⇌ L-Rhamnulose⇌L-Rhamnulose ⇌ L-Rhamnulose1P ⇌ Dihydroxyacetone P + L-Lactaldehyde	R02437	K01813
26	L-Arabinose→L-Ribulose		/	/	R01761	K01804
27	L-Ribulose→Glyceraldehyde3P		L-Ribulose5P ⇌ D-Xylulose5P	L-Ribulose⇌L-Ribulose5P⇌ D-Ribulose5P⇌Xylulose5P⇌Glycer-aldehyde-3P	R05850	K01786
28					R05850	K03077
29					R05850	K03080
30	D-Xylose→Glyceraldehyde3P		D-Xylose ⇌ D-Xylulose	D-Xylose ⇌ D-Xylulose⇌D-Xylulose5P⇌Glyceraldehyde3P	R01432	K00854
31	Glucose→Acetyl P+Glyceraldehyde3P	Ici Glucose correspond au D-Fructose-6P	D-Fructose6P ⇌ Acetyl P + D-Erythrose4P D-Xylulose5P ⇌ Acetyl P + D-Glyceraldehyde3P	D-Erythrose4P ⇌ D-sedoheptulose7P⇌D-ribose 5 P⇌D-ribulose5P⇌D-Xylose5P⇌D-xylulose5P⇌ Acetyl P + D-Glyceraldehyde3P	R01621 R00761	K01621
32	Pyruvate→Acetyl-CoA		/	/	R01196	K00169 K00170 K00171 K00172
33	Pyruvate→Acetyl-CoA		S-Acetyldihydroliipoamide-E⇌ Acetyl-CoA	Pyruvate ⇌ 2-(alpha-Hydroxyethyl)thiamine diphosphate ⇌S-Acetyldihydroliipoamide-E⇌ Acetyl-CoA	R02569	K00627
34	Pyruvate→Acetyl-CoA		/	/	R01196	K03737
35	Pyruvate→Acetyl-CoA+Formate		/	/	R00212	K00656
36	Acetyl-CoA→Acetyl P		/	/	R00230	K04020
37	Acetyl-CoA→Acetyl P		/	/	R00230	K00625
38	Acetyl P→Acétate		/	/	R00315	K00925
39	Acetyl-CoA→Acetoacetyl-CoA		/	/	R00238	K00626
40	Acetoacetyl-CoA→Acetyl-CoA		Butanoyl-CoA + Acétate ⇌ Butanoate + Acetyl-CoA	Acetoacetyl-CoA ⇌(S)-3-Hydroxybutanoyl-CoA⇌Crotonoyl-CoA⇌ Butanoyl-CoA Butanoyl-CoA + Acétate ⇌ Butanoate + Acetyl-CoA	R01179	K01034 K01035
41	Acetoacetyl-CoA→Butanoyl P		Butanoyl-CoA ⇌ Butanoyl P	Acetoacetyl-CoA⇌(S)-3-Hydroxybutanoyl-CoA⇌Crotonoyl-CoA⇌ Butanoyl-CoA ⇌ Butanoyl P	R01174	K00634
42	Butanoyl P→Butanoate		/	/	R01688	K00929
43	Acetoacetyl-CoA→Acetone		/	/	R01366	K01574 K01938
44	Formate→10-Formyltetrahydrofolate		/	/	R00943	K00288 K01491
45	10-Formyltetrahydrofolate→Acetyl-CoA		5,10-Methylenetetrahydrofolate⇌5-Methyltetrahydrofolate	10-Formyltetrahydrofolate⇌5,10-Methylenetetrahydrofolate⇌ 5,10-Methylenetetrahydrofolate⇌5-Methyltetrahydrofolate⇌Acetyl-CoA	R01224 R07168	K00297
46	Pyruvate→Butanediol		(R)-Acetoin ⇌ Diacetyl (R,R)-Butane-2,3-diol ⇌ (R)-Acetoin meso-2,3-Butanediol ⇌ (S)-Acetoin	Pyruvate ⇌ (S)-2-Acetolactate ⇌ 2-Acetoin ⇌ Butane-diol	R02855 R02946 R10504	K00004 K03366
47	Pyruvate→Lactate	Lactate correspond au L-Lactate	/	/	R00703	K00016

Nb	Description	Convention de nommage	Réaction associé au KO sélectionnés	Ensemble des réaction représentées par le marqueur	Code Réaction	Code KO
48	Lactate → Acetyl P + Propanoate		(R)-Methylmalonyl-CoA ⇌ Succinyl-CoA	L-Lactate ⇌ Pyruvate ⇌ Oxaloacetate ⇌ (S)-malate ⇌ Fumarate ⇌ Succinate ⇌ Succinyl-CoA ⇌ (R)methylmalonyl-CoA ⇌ methylmalonyl-CoA ⇌ propanoyl-CoA Cette voie a besoin d'être associée avec Lactate ⇌ Pyruvate ⇌ acetyl-CoA ⇌ AcetylP ⇌ Acétate Par conséquent, nous relient la production de Propanoate avec la production d'Acétyl P	R00833	K01847
49	Lactate → Acetyl P + Propanoate				R00833	K01848 K01849
50	Lactate → Propanoate		Lactoyl-CoA + Propanoate ⇌ L-Lactate + Propanoyl-CoA	Cette réaction a lieu au début et à la fin de la voie : Lactate ⇌ Lactoyl-CoA ⇌ Acryloyl-CoA ⇌ Propanoyl-CoA	R01449	K01026
51	Formate → 5-Formyl-5,6,7,8-tetrahydromethanopterin		Formylmethanofuran ⇌ 5-Formyl-5,6,7,8-tetrahydromethanopterin	Lactate ⇌ Co2 ⇌ Formylmethanofuran ⇌ 5-Formyl-5,6,7,8-tetrahydromethanopterin	R03390	K00672
52	5-Formyl-5,6,7,8-tetrahydromethanopterin → 5,10-Methenyltetrahydromethanopterin		/	/	R03464	K01499
53	5,10-Methenyltetrahydromethanopterin → 5,10-Methylenetetrahydromethanopterin		/	/	R04456	K00319
54	5,10-Methenyltetrahydromethanopterin → 5,10-Methylenetetrahydromethanopterin		/	/	R04455	K13942
55	5,10-Methylenetetrahydromethanopterin → 5-Methyl-5,6,7,8-tetrahydromethanopterin		/	/	R04464	K00320 K00577 K00578 K00579 K00580 K00581 K00582 K00583 K00584 K00399
56	5-Methyl-5,6,7,8-tetrahydromethanopterin → Methyl-CoM	Methyl-CoM correspond au 2-(Methylthio)ethanesulfonate	/	/	R04347	K00401 K00402 K01699 K13919 K13920
57	Methyl-CoM → Methane		/	/	R04541	K13922
58	L-Lactaldehyde → Propane 1,2-diol		L-Lactaldehyde ⇌ Propane 1,2-diol	L-Lactaldehyde ⇌ Propane 1,2-diol	R02376	
59	Propane 1,2-diol → Propanoate		Propanal ⇌ Propanoyl Coa	Propane 1,2-diol ⇌ Propanal ⇌ Propanoyl Coa ⇌ Propanoyl P ⇌ Propanoate	R09097	
60	Acetyl-CoA → Acetyl P		/	/	R00230	K13788
61	Acetyl-CoA → Acetyl P		/	/	R00230	K15024

2.3 Constitution de la matrice des abondances des TFA

À l'issue de la section précédente nous obtenons une liste de 25 GH et PL et de 61 KO. Ce sont des marqueurs fonctionnels spécifiques du catabolisme des polysaccharides en milieu anaérobie. Nous présentons dans cette section la construction pratique des TFA et de leur valeurs. Nous abordons séquentiellement l'annotation puis l'obtention des abondances de GH, PL et KO.

Annotation. La particularité des GH et PL ne permet pas d'utiliser les outils d'alignement classique (BLAST). Nous utilisons alors la solution évoquée précédemment c'est-à-dire, l'utilisation des modèles de domaine des familles de GH et PL développées dans dbCan version3 Yin et al. (2012) au moyen du programme Hmmer tool (<http://hmmer.org/>). Nous prenons les paramètres par défaut proposés. Une première étape de vérification du procédé a été réalisée au moyen 145 séquences de protéines manuellement annotées dans Tasse et al. (2010). En considérant hypothétiquement les annotations de Tasse et al. (2010) comme vraies, il est possible d'utiliser des métriques de classification pour comparer les deux annotations. Les faux positifs trouvés sont principalement associés aux carbohydrate-binding modules que nous n'exploitons pas dans ce travail. Si l'on ignore ces modules, la précision est supérieure à 98%.

Nous considérons donc que l'annotation automatique que nous réalisons est de qualité comparable à l'annotation experte de Tasse et al. (2010). Nous utilisons alors cette approche pour annoter les 9.9 millions de gènes de l'IGC. Cette étape nous permet de sélectionner environ 180 000 gènes, dont 50 000 sont annotés par au moins l'un des 25 GH PL sélectionnés. Certaines annotations se chevauchent c'est-à-dire qu'une fraction de la séquence de protéine est annotée par deux marqueurs. Dans cette situation, (1) si la taille du chevauchement est inférieure à 10% du plus petit domaine détecté, le chevauchement est ignoré, (2) sinon, nous considérons les e-values qui mesurent la qualité de l'annotation et écartons l'annotation la plus mauvaise si son e-value est 100 fois plus forte (mauvaise) que l'autre, (3) dans tous les autres cas, les annotations sont considérées comme erronées et sont écartées.

L'IGC tel que produit par Li et al. (2014) est annoté en KO, nous exploitons cette annotation. Cependant, certains gènes sont annotés de façon multiple par plusieurs KO. Dans cette situation, nous choisissons de diviser l'abondance du gène par le nombre de KO et nous considérons que le gène est annoté par tous les KO trouvés.

Les gènes de l'IGC sont désormais annotés en KO, GH et PL et nous obtenons donc pour

chacun des 86 KO, GH et PL sélectionnés une liste de gènes associés ainsi que leur abondances.

Abondance de TFA. L'IGC donne l'abondance relative de 1423 échantillons concernant chacun des 9.9 millions de gènes qui le composent. Après alignement des 1423 échantillons sur le catalogue, pour chaque échantillon i et pour chaque gène g , des comptages de gènes C_{ig} sont obtenus. L'abondance A_{ig} du gène g dans l'échantillon i se calcule alors selon

$$A_{ig} = \frac{C_{ig}/L_g}{\sum_{l=1}^{l=9.9.10^6} C_{il}/L_l}, \quad (2.1)$$

où L_g désigne la taille en nombre de nucléotides du gène g , la division par L_g permet de corriger un biais de taille dans le nombre de reads associés à un gène (Li et al., 2014). Ces abondances donnent la fréquence de la présence de chaque gène.

Les 86 marqueurs fonctionnels sélectionnés sont des traits fonctionnels. Nous n'avons pas accès aux quantités de marqueurs fonctionnels associés à chaque micro-organisme de chaque échantillon. Nous pouvons toutefois observer leur abondance totale dans chaque échantillon et construire ainsi des TFA. L'abondance d'un TFA est obtenue en additionnant les abondance des gènes qui sont annotés par le marqueur fonctionnel correspondant.

Outliers. Nous étudions les 1423 échantillons par Analyse en Composante Principale et observons la présence de 15 échantillons aberrants que nous supprimons.

Après sélection des traits fonctionnels et construction des TFA et de leur abondances, nous avons produit une matrice A d'abondance relative de 86 marqueurs fonctionnels de la métabolisation des polysaccharides par le microbiote intestinal humain pour 1408 échantillons.

Chapitre 3

Modélisation Mathématique

Nous avons construit dans le chapitre précédent (Section 2.3) une matrice d'abondance A de 86 TFA de la dégradation des fibres par le microbiote intestinal humain pour chacun des 1408 échantillons métagénomiques considérés. De par la bonne connaissance de cet écosystème et grâce à la bibliographie extensive associée à ce métabolisme il est possible de suivre une approche fonctionnelle qui comme nous expliqué au Chapitre 3 est particulièrement pertinente.

3.1 Un modèle de Combinaisons de Trait Fonctionnels Agrégés (CTFA)

Nous construisons un modèle reposant sur la notion d'assemblage fonctionnel. Le terme assemblage fonctionnel désigne des assemblages de micro-organismes dont la taxonomie n'est pas définie mais dont les fonctions sont bien caractérisées. Notre modèle s'appuie sur les deux hypothèses suivantes.

Hypothèse 1 *Le microbiote intestinal de chaque individu est constitué d'un nombre limité d'assemblages fonctionnels.*

L'écosystème est par définition limité, que ce soit par les ressources nutritives ou la structuration spatiale de l'intestin humain ; nous interprétons les différents assemblages fonctionnels comme des stratégies spécialisées d'exploitation des ressources disponibles. Cette hypothèse est illustrée par la figure 3.1.

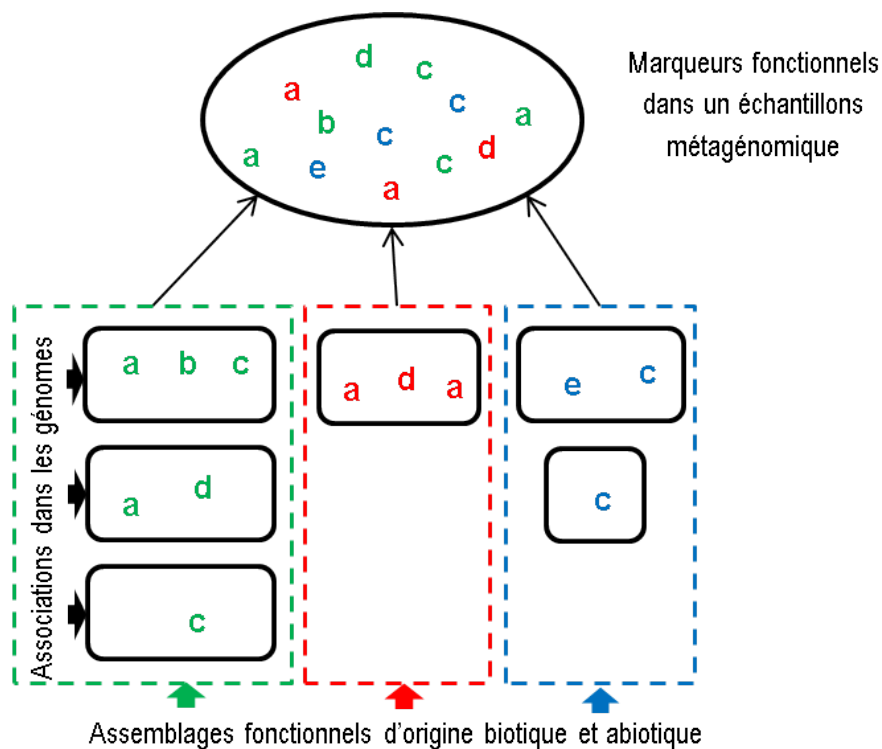


FIGURE 3.1: Illustration de la notion d'assemblages fonctionnels : dans un échantillon métagénomique, les abondances de marqueurs fonctionnels (a,b,c,d et e) peuvent s'expliquer comme provenant de micro-organismes (6 boîtes noires) regroupés en assemblages (encarts verts,rouge et bleu).

Les vecteurs d'abondances de ces TFA définissent des Combinaisons de Traits Fonctionnels Agrégés (CTFA), qui caractérisent chacune de ces populations.

Hypothèse 2 *Les processus d'assemblages fonctionnels du microbiote intestinal humain sont identiques d'un individu à l'autre.*

Par définition de la notion d'assemblage fonctionnel, cette hypothèse ne présuppose pas de stabilité taxonomique mais signifie que ces assemblages, uniquement caractérisés en terme de TFA, auront des structure fonctionnelles identiques dans tous les microbiotes intestinaux humains. Ceci revient à supposer un processus de sélection des fonctions microbiennes par l'intestin humain, reproductible d'un individu à l'autre. Dans le figure 3.2 des CTFA sous forme de vecteurs d'abondances de gènes sont présents dans chacun des échantillons en proportions différentes.

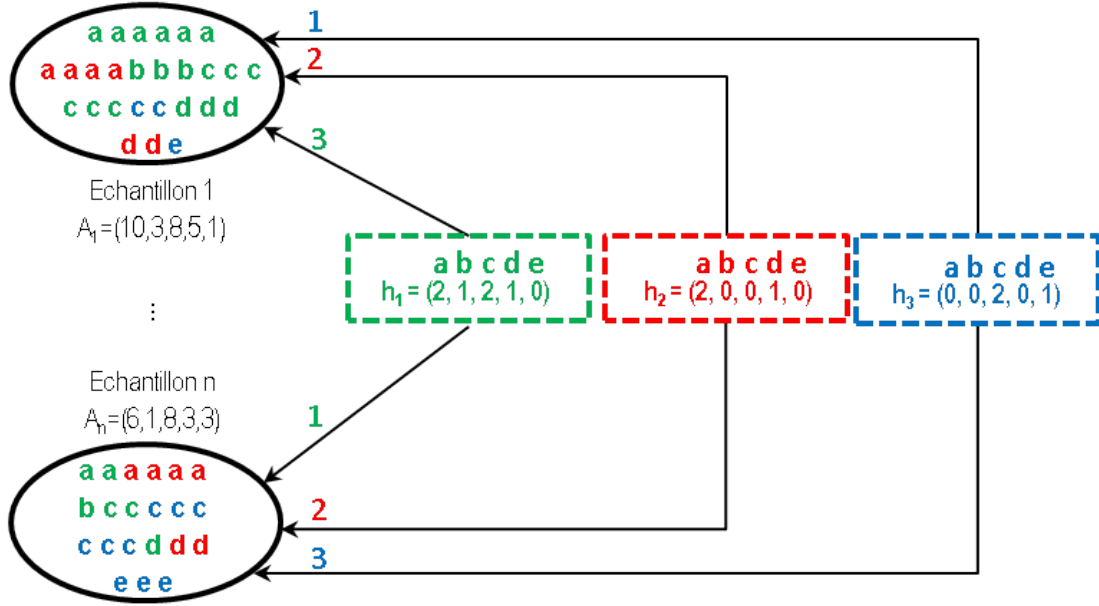


FIGURE 3.2: Illustration de la notion d'assemblages fonctionnels reproductibles. Les encarts en pointillés symbolisent les CTFA associés aux assemblages fonctionnels. Ils sont décrits par des vecteurs h_i d'abondances relatives des marqueurs a, b, c, d et e . Dans cet exemple simple les abondances (A_i) de marqueurs dans l'échantillon métagénomique i sont reconstruites exactement par les CTFA. Les chiffres colorés sur les flèches représentent le poids de chaque CTFA présent dans l'échantillon.

En accord avec les hypothèses précédentes, nous modélisons tous les métagénomés observés comme un mélange d'un nombre limité de CTFA dans des proportions variables. Notre modèle s'écrit alors : pour tout échantillon $i \in [1, \dots, n]$, et pour tout marqueur fonctionnel $j \in [1, \dots, r]$, la fréquence A_{ij} du marqueur fonctionnel j dans l'échantillon i est la résultante des contributions des k différents CTFA,

$$A_{ij} \simeq \sum_{l=1}^k W_{il} H_{lj} \quad (3.1)$$

où $H_l = (H_{l1}, \dots, H_{lr})$ est le CTFA numéro l , et W_{il} est la proportion du CTFA l dans l'échantillon i . La matrice $H \in \mathcal{M}_{k,r}(\mathbb{R}^+)$ de terme H_{lj} est appelée la matrice des CTFA et la matrice $W \in \mathcal{M}_{n,k}(\mathbb{R}^+)$ de terme W_{il} est appelée la matrice des proportions. De même la matrice $A \in \mathcal{M}_{n,r}(\mathbb{R}^+)$ de terme A_{ij} est appelée matrice des données ou d'abondances. Il est important

de noter que par définition, les matrices W et H sont à termes positifs.

$$W \geq 0, \quad H \geq 0 \quad (3.2)$$

Par convention, nous normalisons la matrice H par ligne de façon à interpréter les CTFA comme des fréquences de marqueurs dans un assemblage fonctionnel.

3.2 Prendre en compte l'information biologique pour la construction des CTFA

3.2.1 Transformer une connaissance sur la structure des génomes en une contrainte algébrique sur les CTFA

Par définition les CTFA caractérisent des assemblages de micro-organismes. Dans un échantillon i , l'abondance des marqueurs fonctionnels directement imputable au CTFA l s'écrit $W_{il}H_l$. Cette abondance est aussi égale à la somme des marqueurs fonctionnels présents dans les génomes de l'ensemble $T_l^{(i)}$ des micro-organismes qui composent l'assemblage fonctionnel caractérisée par ce CTFA dans l'échantillon i . Ceci se traduit par la relation

$$W_{il}H_{lj} = \sum_{x \in T_l^{(i)}} g_{xj} \quad (3.3)$$

où g_{xj} est l'abondance du marqueurs fonctionnel j dans le génome du micro-organisme x .

De ce fait si nous possédons une information sur la structure de chacun des génomes qui composent cet assemblage, cette information va pouvoir être traduite à l'échelle du CTFA.

A titre d'exemple, si deux marqueurs fonctionnels j et j' sont tels que pour tout génome x de l'écosystème Σ , il existe une constante $R_{x,j,j'} \geq 0$ vérifiant

$$g_{xj} = R_{x,j,j'} g_{xj'}, \quad (3.4)$$

alors en posant

$$\delta = \sup_{x \in \Sigma} R_{x,j,j'}, \quad (3.5)$$

tout CTFA l vérifie

$$H_{lj} \leq \delta H_{lj'}. \quad (3.6)$$

Il serait possible d'essayer de construire des contraintes pour toutes les paires ordonnées

de marqueurs fonctionnels. Cependant, tous les génomes ne vérifient pas nécessairement par la propriété (3.4), auquel cas la quantité (3.5) n'existe pas. Nous allons donc par la suite exploiter l'information métabolique que nous avons présentée dans le chapitre 2 afin de choisir un nombre limité de contraintes pertinentes.

3.2.2 Transformer une connaissance métabolique en contrainte algébrique

À l'échelle des micro-organismes, la réalisation de fonctions métaboliques nécessite la présence conjointe de gènes dans le génome. La connaissance a priori des voies métaboliques et l'annotation fonctionnelle doivent permettre d'anticiper ces co-occurrences. La prise en compte de cette structuration permet de compléter la figure 3.1 en la figure 3.3.

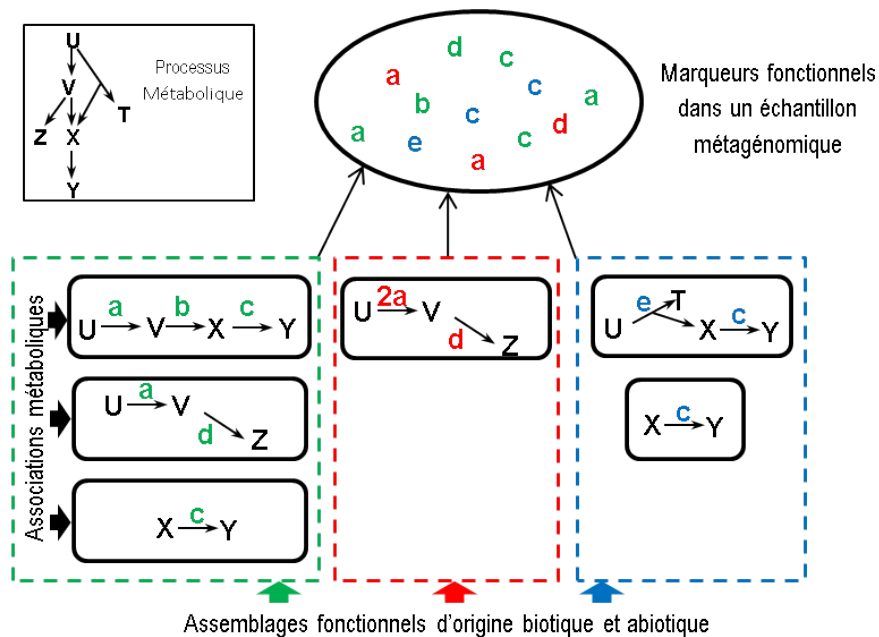


FIGURE 3.3: Illustration de l'organisation hiérarchique des gènes : les micro-organismes (6 boîtes noires) possèdent des gènes (a,b,c,d et e) dont la présence conjointe permet la réalisation de processus métaboliques. Ces organismes pour multiple raisons mais aussi de part leurs capacités métaboliques sont regroupés en assemblages fonctionnels (encarts verts,rouge et bleu). Les gènes une fois exprimés permettent la production d'enzymes qui catalysent des réaction du processus métabolique. C'est ce que signifie l'association d'un gène (a,b,c,d,e) avec une flèche du schéma métabolique dans les micro-organismes.

En étudiant le catabolisme des fibres nous nous intéressons à un ensemble de réactions dont la réalisation simultanée conduit à un processus orienté de dégradation de composés énergétiques (fibres/sucres) en composés plus simples (Acides Gras à Chaines Courtes/Gas). De ce fait nous nous faisons l'hypothèse suivante.

Hypothèse 3 *Dans un génome bactérien, la présence d'un marqueur fonctionnel spécifique du catabolisme, indique un potentiel de réalisation d'une réaction orientée. Cette orientation se fait selon sens du catabolisme. Il est donc possible d'associer à cette réaction des métabolites consommés et des métabolites produits.*

Cette hypothèse permet de définir pour chaque métabolite considéré des réactions qui le produisent et le consomment.

Pour construire un ensemble de contraintes liées au métabolisme, une première idée serait de supposer que dans un micro-organisme, un métabolite intracellulaire ne peut exister que si il existe au moins un gène associé à sa production et un second associé à sa consommation. Si tel est le cas, il est possible de formuler une règles simples de co-occurrences des marqueurs fonctionnels : dans un micro-organisme le nombre de marqueurs associés aux réactions qui produisent un métabolite intracellulaire et le nombre de marqueurs associés aux réactions qui consomment ce même métabolite sont soit tous les deux nuls soit tous les deux non nuls. Cette règle peut se formaliser algébriquement : soit un métabolite intracellulaire m , nous définissons deux ensembles : P_m l'ensemble d'indices des marqueurs associés à une réaction produisant m et C_m l'ensemble d'indices des marqueurs associés à une réaction consommant m . Soit g_{ij} , le nombre de copies du marqueur j dans le génome du micro-organisme i , alors la règle se formule selon :

Propriété Pour un micro-organisme x , (3.7) ou (3.8) est vraie.

$$\sum_{j \in P_m} g_{xj} = \sum_{j' \in C_m} g_{xj'} = 0 \quad (3.7)$$

$$\sum_{j \in P_m} g_{xj} = R_{xm} \sum_{j' \in C_m} g_{xj'} > 0 \quad (3.8)$$

R_{xm} est le rapport de la somme des gènes produisant m sur la somme des gènes consommant m dans le micro-organisme x .

Toutefois, cette première approche souffre de certains défauts. Tout d'abord certaines cellules sont capables d'accumuler de grandes quantités de métabolites pour des utilisations ultérieures

(graines, féculés, etc..), cependant ce n'est pas le cas pour les processus métaboliques du catabolisme des fibres.

Ensuite, nous ne décrivons qu'un processus métabolique particulier qui est le catabolisme des fibres. Or certains des métabolites que nous avons pris en compte se sont pas spécifiques et sont également utilisés par d'autres fonctions métaboliques. Il est donc possible que certaines abondances de marqueurs fonctionnels s'expliquent partiellement par la présence d'autres marqueurs non décrits réalisant d'autre voies métaboliques. Par exemple, les catabolismes des lipides et des protéines ont des parties communes avec le catabolisme des sucres simples. Enfin nous n'avons pas la garantie absolue de connaître parfaitement le catabolisme des fibres et certaines voies métaboliques peuvent être inconnues.

La propriété précédent est adaptée de façon analogue à la formulation présentée en (3.4).

Propriété Pour un micro-organisme x , (3.9) ou (3.10) est vraie.

$$\sum_{j' \in C_m} g_{xj'} > 0 \quad \text{et} \quad 0 \leq \sum_{j \in P_m} g_{xj} = R_{xm} \sum_{j' \in C_m} g_{xj'} \quad (3.9)$$

$$\sum_{j \in P_m} g_{xj} > 0 \quad \text{et} \quad 0 \leq R'_{xm} \sum_{j' \in C_m} g_{xj'} = \sum_{j \in P_m} g_{xj} \quad (3.10)$$

Cette deuxième version permet de prendre en compte pour un micro-organisme x le cas où l'on ne trouve aucun marqueur fonctionnel permettant la production du métabolite m alors que l'on trouve un nombre non nul de marqueurs associés à sa consommation (3.9) et le cas où l'on ne trouve aucun marqueur permettant la consommation du métabolite m alors que l'on trouve un nombre non nul de marqueurs associés à sa production (3.10). On remarque que si $\sum_{j \in P_m} g_{xj} \neq 0$ et $\sum_{j' \in C_m} g_{xj'} \neq 0$ on peut prendre $R_{xm} = R'_{xm}$ et on retrouve l'égalité 3.8.

La traduction des contraintes (3.9) et (3.10) à l'échelle des CTFA se fait de façon analogue à l'équation (3.5). Pour cela nous définissons M_S (respectivement M_I) l'ensemble des métabolites m tels que pour tous les micro-organismes x de l'écosystème Σ , l'inégalité (3.9) (resp. (3.10)) est vraie. Définissons

$$\forall m \in M_S \quad \delta_m^S = \sup_{x \in \Sigma} R_{xm} \quad (3.11)$$

$$\forall m \in M_I \quad \delta_m^I = \inf_{x \in \Sigma} R'_{xm}. \quad (3.12)$$

D'après (3.9) et (3.10), quelque soit le CTFA l on a

$$\begin{aligned} \sum_{x \in T_l^{(i)}} \sum_{j \in P_m} g_{xj} &\leq \delta_m^S \sum_{x \in T_l^{(i)}} \sum_{j' \in C_m} g_{xj'} \\ \sum_{x \in T_l^{(i)}} \sum_{j' \in C_m} g_{xj'} &\leq \delta_m^I \sum_{x \in T_l^{(i)}} \sum_{j \in P_m} g_{xj}. \end{aligned}$$

Il s'ensuit d'après (3.3) que quel que soit l'échantillon i et le CTFA l

$$\begin{aligned} \sum_{j \in P_m} W_{il} H_{lj} &\leq \delta_m^S \sum_{j' \in C_m} W_{il} H_{lj'} \\ \sum_{j' \in C_m} W_{il} H_{lj'} &\leq \delta_m^I \sum_{j \in P_m} W_{il} H_{lj}. \end{aligned}$$

et donc en supposant la présence du CTFA dans au moins un échantillon

$$\sum_{j \in P_m} H_{lj} \leq \delta_m^S \sum_{j' \in C_m} H_{lj'} \quad (3.13)$$

$$\sum_{j' \in C_m} H_{lj'} \leq \delta_m^I \sum_{j \in P_m} H_{lj}. \quad (3.14)$$

On remarque que pour les métabolites appartenant à $M_I \cap M_S$ la relation suivante est vérifiée

$$\delta_m^I \leq \frac{\sum_{j \in P_m} H_{lj}}{\sum_{j' \in C_m} H_{lj'}} \leq \delta_m^S. \quad (3.15)$$

3.3 Construction pratique de la contrainte de structure liée au métabolisme

3.3.1 Exploitation du graphe topologiques associé aux 61 KO

La contrainte métabolique explicitée en (3.13) et (3.14) requiert pour chaque métabolite la constitution des ensembles P_m et C_m . Pour cela nous exploitons le graphe construit dans la Section 2.2.2.7.

Nous avons sélectionné 86 marqueurs se rapportant à la métabolisation anaérobie des fibres

alimentaires. L'information fonctionnelle des marqueurs de la dégradation des fibres en sucres simples (GH/PL) ne peut pas être reliée de façon spécifique à des métabolites intracellulaires. De ce fait, nous ne contraignons la co-occurrence que des 61 marqueurs (KO) de la fermentation anaérobie des sucres simples provenant de la dégradation des fibres alimentaires.

En construisant un graphe topologique orienté reliant des métabolites (nœud) et des marqueurs fonctionnels spécifiques (arêtes) nous avons défini de manière implicite les ensembles P_m et C_m . Sur ce graphe, pour un nœud m l'ensemble des marqueurs (arêtes) aboutissant au composé m (nœud) constitue P_m et l'ensemble des marqueurs partant de m constitue C_m . Nous utilisons la matrice d'incidence de ce graphe pour formuler l'ensemble des contraintes sur la structure des CTFA. Le graphe est composé de 43 nœuds pour 61 arêtes, Q_0 la matrice d'incidence correspondante est de taille 43×61 et de terme q_{mj} . Par définition d'une matrice d'incidence, $q_{mj} = -1$ si l'arête j part du nœud m et $q_{mj} = 1$ si l'arête j pointe sur le nœud m , $q_{mj} = 0$ dans tous autres cas. Nous prenons la sous-matrice Q constituée des lignes associées aux 25 métabolites intracellulaires bien identifiés lors de la construction du graphe. Nous définissons alors deux sous-matrices Q^+ et Q^- composées respectivement des termes positifs de Q et de la valeur absolue des termes négatifs de Q de sorte que : $Q = Q^+ - Q^-$. On remarque que :

$$\sum_{j \in P_m} H_{lj} = q_m^+ H_l^T \quad \text{et} \quad \sum_{j' \in C_m} H_{lj'} = q_m^- H_l^T \quad (3.16)$$

en prenant q_m^+ et q_m^- respectivement les $m^{\text{ème}}$ lignes des matrices Q^+ et Q^- . Les contraintes (3.13) et (3.14) n'étant appliquées qu'aux métabolites appartenant respectivement aux ensembles M_S et M_I , deux sous-matrices Q_I^+ et Q_I^- de Q^+ et Q^- ne contenant que les lignes des métabolites de M_I sont définies. De manière symétrique Q_S^+ et Q_S^- sont construits à partir de Q^+ , Q^- et M_S . On écrit :

$$\begin{aligned} \forall m \in M_S, \quad \sum_{j \in P_m} H_{lj} \leq \delta_m^S \sum_{j' \in C_m} H_{lj'} \\ \forall m \in M_I, \quad \sum_{j' \in C_m} H_{lj'} \leq \delta_m^I \sum_{j \in P_m} H_{lj} \end{aligned} \quad \iff \quad F_{\Delta} H^T = \begin{pmatrix} Q_S^+ - \Delta_S Q_S^- \\ Q_I^- - \Delta_I Q_I^+ \end{pmatrix} H^T \leq 0 \quad (3.17)$$

où Δ_S et Δ_I sont les matrices diagonales de termes δ_m^S et δ_m^I respectivement définies pour les métabolites de M_S et M_I .

3.3.2 Étude de 190 génomes d'espèces prévalentes dans le microbiote intestinal humain afin de calibrer les contraintes

En théorie, les ensembles M_S et M_I ainsi que les grandeurs δ_m^S et δ_m^I introduits dans la Section 3.2.2 doivent être définis à partir de l'ensemble des génomes microbiens présents dans l'écosystème. En pratique ce n'est pas possible, puisque le nombre de génomes du microbiote intestinal humain séquencé est important mais non exhaustif. Il est donc nécessaire de définir ces objets à partir d'une description partielle des génomes présents dans l'écosystème.

Nous exploitons 190 génomes microbiens faisant partie d'un noyau de micro-organismes communs aux microbiotes intestinaux humains (Qin et al. (2010) et Tap et al. (2009)). La liste de ces organismes et leur origine est présentée en Annexe C. Nous les annotons en KO en utilisant BLAST avec un bitscore minimum de 60. L'annotation en GH se fait de manière similaires à l'annotation de l'IGC présentée en 2.3. La construction de l'abondance des 86 TFA dans chacun de ces génomes est similaire à celle réalisée pour les 1408 échantillons que nous avons présenté en 2.3. Notons que nous manipulons ici des nombres de copies de gènes provenant de génomes. Nous construisons de cette façon une matrice G donnant le nombre de gènes annotés pour chacun des 86 marqueurs dans les 190 génomes.

Construction d'une approximation de M_S et M_I . En nous appuyant sur les équation (3.9), (3.10) et (3.16), une approximation \tilde{M}_S et \tilde{M}_I des ensembles M_S et M_I est obtenue en cherchant les métabolites m pour lesquels les sommes respectives $q_m^+ g_x$ et $q_m^- g_x$ sont non nulles dans le génome x pour moins de (5%) des 190 génomes (soit 10). Nous faisons l'hypothèse de travail que ce faible nombre de génomes peut être négligé dans la mesure où les majorations (3.13) et (3.14) utilisent une définition très conservative des bornes (3.11) et (3.12). \tilde{M}_S et \tilde{M}_I sont construits selon le protocole suivant.

1. si $q_m^- g_x$ est nul pour plus de 10 génomes parmi les 190, cela signifie que nous connaissons le ou les marqueurs associés à la production de m sans en avoir identifié un seul associé à sa consommation. Nous vérifions que m est un métabolite spécifique de la dégradation anaérobie des sucres simples.
 - (a) Si m n'est pas spécifique, il n'est pas ajouté à l'ensemble \tilde{M}_S et la contrainte associée n'est pas implémentée pour les marqueurs associés à ce métabolite.
 - (b) Sinon, nous reconsidérons le nombre de génomes dans lequel $q_m^- g_x$ est nul avec un nouveau seuil de 15%

- i. Le nouveau seuil permet de considérer que la contrainte s'applique et m est ajouté à \tilde{M}_S
 - ii. Le nouveau seuil ne permet pas de considérer que la contrainte s'applique et m n'est pas ajouté à \tilde{M}_S
2. si $q_m^+g_x$ est nul pour plus de 10 génomes parmi les 190, cela signifie que nous connaissons le ou les marqueurs associés à la consommation de m sans en avoir identifié un seul associé à sa production. Nous vérifions que m est un métabolite spécifique de la dégradation anaérobie des sucres simples.
- (a) Si m n'est pas spécifique, il n'est pas ajouté à l'ensemble \tilde{M}_I et la contrainte associée n'est pas implémentée pour les marqueurs associés à ce métabolite.
 - (b) Sinon, nous reconsidérons le nombre de génomes dans lequel $q_m^+g_x$ est nul avec un nouveau seuil de 15%
 - i. Le nouveau seuil permet de considérer que la contrainte s'applique et m est ajouté à \tilde{M}_I
 - ii. Le nouveau seuil ne permet pas de considérer que la contrainte s'applique et m n'est pas ajouté à \tilde{M}_I
3. $q_m^+g_x$ et $q_m^-g_x$ sont tous les deux nuls dans plus de 5% des génomes, alors les marqueurs fonctionnels associés à m ne sont pas contraints. De même nous décidons de supprimer arbitrairement les contraintes associées à certains métabolites dont le métabolisme est mal connu et pour lesquels l'une des deux quantités $q_m^+g_x$ et $q_m^-g_x$ est nulle dans un grand nombre de génomes.

Ensembles \tilde{M}_S et \tilde{M}_I obtenus. Le Tableau 3.1 énumère les métabolites où pour plus de 5% des génomes, $\tilde{\delta}_m^S$ ou $\tilde{\delta}_m^I$ ne sont pas définis.

Suite à l'application des règles exposées précédemment, les fréquences des marqueurs associés à 4 métabolites intracellulaires sur 25 ne sont pas contraintes. Il s'agit du D-Gluconate6P, du D-gluconolactone, du L-Fuculose1P et du Propane1,2diol. Cela implique que pour une part non négligeable des 190 organismes étudiés, il existent d'autres mécanismes associés à la présence des marqueurs de ces 4 métabolites. Concernant le D-Gluconate 6P, nous ne le représentons qu'associé à la voie d'Entner-Doudoroff alors qu'il peut aussi être utilisé dans la voie du pentose-phosphate que nous ne représentons pas. Il est aussi possible que des problèmes d'annotation, de faux positifs ou de faux négatifs brouillent les abondances associés à ces marqueurs.

Les 21 métabolites restants sont répartis dans les ensembles \tilde{M}_S et \tilde{M}_I constitués de 19 métabolites chacun ce qui résulte donc en 38 contraintes, 19 de type (3.13) et 19 de type (3.14).

Métabolite	Nb de génomes x où $q_m^- g_x$ ou $q_m^+ g_x$ s'annule	δ	Action	Cas
Acetoacetyl-CoA	34	δ^S	-	cas 1.(a)
ButanoylP	28	δ^S	Ajouté à \tilde{M}_S	cas 1.(b).i
D-Altronate	10	δ^S	Ajouté à \tilde{M}_S	cas 1.(b).i
D-Gluconate6P	12	δ^S	-	cas 3
D-gluconolactone	12	δ^S	-	cas 3
Galactose1P	68	δ^S	-	cas 1.(a)
L-Fucose1P	14	δ^S	-	cas 3
2Dehydro3deoxy-Dgluconate	89	δ^I	-	cas 2.(a)
D-Gluconate6P	38	δ^I	-	cas 3
D-gluconolactone	24	δ^I	-	cas 3
L-Fucose	17	δ^I	Ajouté à \tilde{M}_I	cas 2.(b).i
L-Fucose1P	26	δ^I	-	cas 3
L-Ribulose	27	δ^I	-	cas 2.(a)
Propanal 1,2di-ol	92	δ^I	-	cas 3

TABLE 3.1: Construction de \tilde{M}_S et \tilde{M}_I dans les cas litigieux

Construction d'une approximation de δ_m^S et δ_m^I . Nous construisons une approximation $\tilde{\delta}_m^S$ et $\tilde{\delta}_m^I$ des bornes δ_m^S et δ_m^I de la façon suivante. Pour tout m de \tilde{M}_S et \tilde{M}_I ,

$$\tilde{\delta}_m^S > \sup_{x \in [1, \dots, 190]} \left(\frac{q_m^+ g_x}{q_m^- g_x} \right)$$

$$\tilde{\delta}_m^I > \sup_{x \in [1, \dots, 190]} \left(\frac{q_m^- g_x}{q_m^+ g_x} \right)$$

où g_x est le vecteur colonne des abondances de marqueurs dans le génome du micro-organismes numéros x . Nous parcourons les 190 génomes d'espèce microbiennes prépondérantes afin de calibrer les valeurs de $\tilde{\delta}_m^S$ et $\tilde{\delta}_m^I$. Nous prenons pour chaque m de \tilde{M}_S :

$$\delta_m^{S,0} = 1.5 \max_x \left(\frac{q_m^+ g_x}{q_m^- g_x} \right)$$

et pour chaque métabolite m de \tilde{M}_I :

$$\delta_m^{I,0} = 1.5 \max_x \left(\frac{q_m^- g_x}{q_m^+ g_x} \right)$$

Nous majorons une deuxième fois en arrondissant par valeur supérieure les valeurs de $\tilde{\delta}_m^S$ et $\tilde{\delta}_m^I$ à l'entier divisible par 5 le plus proche de $\delta_m^{S,0}$ et $\delta_m^{I,0}$:

$$\tilde{\delta}_m^S = 5 \lceil \frac{\delta_m^{S,0}}{5} \rceil \quad (3.18)$$

$$\tilde{\delta}_m^I = 5 \lceil \frac{\delta_m^{I,0}}{5} \rceil \quad (3.19)$$

$\lceil \cdot \rceil$ dénotant la partie entière par excès. Le Tableau 3.2 contient les valeurs obtenues pour $\tilde{\delta}_m^S$ et $\tilde{\delta}_m^I$. Le symbole "-" indique l'absence de contraintes.

Métabolites	δ_m^S	δ_m^I
10-FormylTHF	5	20
2Dehydro3deoxyDgluconate	5	-
5,10MethenylTHM	5	5
5,10MethylTHM	5	5
5FTHM	5	5
5MethylCoM	5	5
5MethylTHM	5	5
Acetoacetyl-CoA	-	5
Acetyl-CoA	20	5
AcetylP	10	10
ButanoylP	5	5
D-Altronate	10	10
D-Gluconate6P	-	-
D-Gly1,3P2	5	5
D-Gly3P	30	5
D-gluconolactone	-	-
Galactose1P	-	10
Gly2P	15	5
GlyP	10	10
Glycerate3P	10	15
L-Fuculose	5	5
L-Fuculose1P	-	-
L-Ribulose	5	-
Propane1,2diol	-	-
Pyruvate	5	10

TABLE 3.2: Valeurs de δ_m^S et δ_m^I pour chaque métabolites intra-cellulaire m

Chapitre 4

Factorisation en Matrices non-Négatives

Introduction

Dans le précédent chapitre, nous avons défini un modèle dont l’ambition est de représenter des populations fonctionnelles à partir de données méta-génomiques. Nous avons construit une matrice A d’abondance de 86 traits fonctionnels agrégés dans 1408 échantillons, que nous cherchons à approcher par une matrice de Combinaisons de Traits Fonctionnels Agrégés H et une matrice des proportions de traits W , soumises à un ensemble de contraintes. Dans ce chapitre, nous explicitons nos choix de formulation du problème d’inférence statistique.

Nous avons écrit précédemment :

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

ce qui peut se réécrire matriciellement

$$A \simeq WH$$

Par la suite, nous désignerons l’ensemble des couples de matrices (W, H) à termes positifs par

$$\mathcal{C}_0 = \mathcal{M}_{n,k}(\mathbb{R}^+) \times \mathcal{M}_{k,r}(\mathbb{R}^+). \quad (4.1)$$

L’estimation de deux matrices $(W, H) \in \mathcal{C}_0$ dont le produit matriciel approche une matrice de données $A \in \mathcal{M}_{n,r}(\mathbb{R}^+)$ se résout classiquement par NMF (Non-negative Matrix Factorisation ou Factorisation en Matrices Non negatives).

La problématique de NMF est relativement récente puisqu’on retrouve sa trace pour la première fois en 1994 (Paatero and Tapper (1994)). La NMF constitue un modèle générique qui peut être utilisé à des fins de réduction de dimension ou de clustering, de manière similaire à la décomposition en valeurs singulières. La contrainte de positivité permet une représentation réaliste de nombreux problèmes, ce qui explique la popularité de la NMF dans des domaines variés tels que la séparation de sources aveugles en traitement du signal (Cichocki et al. (2006)), le traitement d’image (Hoyer (2004)), la fouille de texte (Lee and Seung (1999)) ou l’analyse de données de spectroscopie de masse (Zhang et al. (2011)). Plus récemment la NMF a été introduite en métagénomique, pour identifier des reads appartenant aux mêmes OTU (Baran and Halperin (2012)), ou dans le cadre de la comparaison d’écosystèmes. En particulier, Jiang et co-auteurs ont utilisés la NMF à des fins de clustering pour étudier et caractériser des écosystèmes par des abondances de voies métaboliques (Jiang et al. (2012b)), par des abondances

de familles de protéines pour des échantillons d'écosystèmes marins (Jiang et al. (2012a)) et par des abondances en 16S et en KO simultanément pour les différents sites du corps humain (Jiang et al. (2014)).

La NMF se formule classiquement comme la minimisation d'une fonction de coût de la forme :

$$\mathcal{D}(A|\mathbf{WH}) + pen(\mathbf{W}, \mathbf{H}) \quad (4.2)$$

où \mathcal{D} est une divergence permettant de quantifier la dissimilarité entre matrices et pen une fonction de pénalité. La définition précise de la fonction de coût (4.2) et sa minimisation requièrent de fixer quatre quantités :

- **La divergence \mathcal{D}** : le choix d'une divergence peut être associé au choix d'un modèle statistique sous-jacent en écrivant la divergence $\mathcal{D}(A|WH)$ comme l'opposé de la log-vraisemblance des observations A conditionnellement à WH .
- **La fonction de pénalité pen** : elle peut notamment être choisie pour encourager une structure dans les matrices W et H .
- **La méthode d'optimisation** : une littérature abondante documente la résolution algorithmique et fournit des garanties empiriques ou théoriques de convergence.
- **La procédure de choix de la dimension optimale** : elle peut dépendre de l'utilisation ultérieure des matrices W et H .

Dans ce chapitre, pour chacun des points ci-dessus, nous décrivons la démarche et les choix classiques dans la littérature, puis nous explicitons nos choix d'implémentation. Nous abordons également les propriétés générales du problème d'optimisation de NMF.

Dans le modèle exposé au chapitre précédent, nous avons décrit un domaine de contraintes convexe dans lequel nous recherchons W et H . Il s'écrit formellement :

$$\mathcal{C} = \{(W, H) \in \mathcal{C}_0 \mid F_{\Delta} H^T \leq 0\}. \quad (4.3)$$

L'ensemble \mathcal{C} est plus restrictif que \mathcal{C}_0 , ainsi nous développons une variante contrainte du problème de NMF.

4.1 Divergence matricielle

Une divergence est une fonction permettant de mesurer la dissimilarité entre distributions, vecteurs ou matrices. Comme les distances, les divergences prennent des valeurs positives et ne s'annulent que si leur deux arguments sont égaux. Par contre, elles ne vérifient pas nécessaire-

ment la symétrie en les arguments et l'inégalité triangulaire. Dans notre cas, \mathcal{D} est donc une fonction qui vérifie :

$$\begin{aligned} \forall X, Y \in \mathcal{M}_{n,r}(\mathbb{R}^+), \quad \mathcal{D}(X|Y) &\geq 0 \\ \mathcal{D}(X|Y) = 0 &\Leftrightarrow X = Y \end{aligned}$$

De plus, dans ce travail les divergences considérées sont convexes différentiables à dérivées continues en leur deuxième argument.

4.1.1 Divergences usuelles en NMF

Les deux divergences principalement considérées dans la littérature sont la norme de Frobenius et la divergence de Kullback-Leibler généralisée. Sous l'hypothèse d'indépendance des coefficients $\{A_{ij}\}$ conditionnellement au produit WH , on montre que ces deux mesures de dissimilarité correspondent à deux modélisations statistiques de l'erreur d'approximation de A par WH .

4.1.1.1 Divergence associée à la norme de Frobenius

La divergence induite par la norme de Frobenius est définie par :

$$\mathcal{D}_F(Y|X) = \sum_{i=1}^n \sum_{j=1}^r (Y_{i,j} - X_{i,j})^2. \quad (4.4)$$

Supposons que les $\{Y_{i,j}\}$ sont des variables aléatoires indépendantes conditionnellement à X , et que chaque $\{Y_{i,j}\}$ suit une loi gaussienne de moyenne $\{X_{i,j}\}$ et de même variance σ^2 :

$$Y = X + \varepsilon \quad (4.5)$$

avec $\{\varepsilon_{i,j}\}$ i.i.d $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$.

La quantité $\mathcal{D}_F(Y|X)$ est égale, à une constante additive près, à l'opposée de la log-vraisemblance des observations Y .

4.1.1.2 La divergence de Kullback-Leibler généralisée

La divergence KLG (Kullback-Leibler Généralisée) est définie pour toutes matrices Y et X telles que $X_{i,j} = 0 \Rightarrow Y_{i,j} = 0$:

$$\mathcal{D}_{KL}(Y|X) = \sum_{i=1}^n \sum_{j=1}^r \left(Y_{i,j} \log \frac{Y_{i,j}}{X_{i,j}} - Y_{i,j} + X_{i,j} \right) \quad (4.6)$$

avec la convention

$$Y_{i,j} \log \frac{Y_{i,j}}{X_{i,j}} = 0 \quad \text{si} \quad Y_{i,j} = 0.$$

Cette convention est justifiée car $\lim_{x \rightarrow 0^+} x \log x = 0$.

Utiliser la divergence KLG revient à modéliser les observations selon un modèle Poissonien. En effet, en considérant chaque terme $Y_{i,j}$ de la matrice Y comme une réalisation d'une loi de Poisson de paramètre $X_{i,j}$, la divergence KLG entre Y et X est égale à l'opposé de la log-vraisemblance des observations Y conditionnellement aux paramètres X , à une constante additive près. La log-densité d'une loi de Poisson de paramètre λ est égale à $\log p(x|\lambda) = x \log \lambda - \lambda - \log(x!)$, ainsi la log-vraisemblance de Y s'écrit :

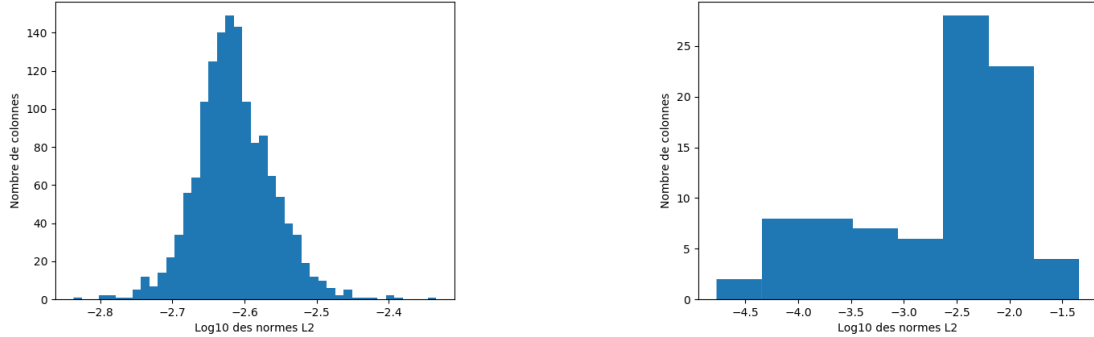
$$\begin{aligned} \sum_{i,j} (Y_{i,j} \log(X_{i,j}) - X_{i,j} - \log(Y_{i,j}!)) &= -D_{KL}(Y|X) \\ &+ \sum_{i,j} (-\log(Y_{i,j}!) + Y_{i,j} \log Y_{i,j} - Y_{i,j}). \end{aligned}$$

4.1.2 Avantages et inconvénients de ces deux divergences

La question d'un modèle statistique sous-jacent associé au choix d'une mesure de dissimilarité est rarement abordée dans la littérature relative à la NMF, dans ses aspects appliqués comme méthodologiques. Le choix semble plutôt lié au domaine d'application. La divergence KLG est majoritairement utilisée dans les applications en traitement du signal et la norme de Frobenius est principalement considérée dans les applications en génomique.

Le modèle statistique poissonnien associé à la divergence KLG offre une meilleure interprétation pour des données positives, alors que la modélisation gaussienne associée à la norme de Frobenius ne constitue une approximation raisonnable pour des données positives que si la moyenne est positive et la variance faible devant la moyenne, c'est à dire si $(WH)_{i,j}$ est strictement positif et grand devant la variance des erreurs.

La résolution de la NMF en utilisant la norme de Frobenius est à la fois bien documentée, rapide et plus facilement implémentée.



(a) Distribution du logarithme en base 10 des normes L2 des 1408 lignes de A

(b) Distribution du logarithme en base 10 des normes L2 des 86 colonnes de A

FIGURE 4.1: Distribution des normes des colonnes et lignes de A

4.1.3 Choix d'implémentation

Nous avons choisi d'utiliser la divergence induite par la norme de Frobenius comme mesure de dissimilarité entre données et produit matriciel.

Le modèle (4.5) suppose une variance identique de l'erreur de reconstruction de tous les termes de la matrice de donnée. Si les ordres de grandeur des valeurs des colonnes (respectivement lignes) de la matrice sont très différents, les erreurs de reconstruction relatives seront donc très différentes entre colonnes. Dans notre étude, les abondances totales par échantillons (lignes) sont relativement similaires, tandis que les abondances totales par marqueurs fonctionnels (colonnes) sont très dispersées (voir Figure 4.1).

De ce fait, les marqueurs dont l'ordre de grandeur est faible (colonnes de la matrice de données) seront mal reconstruits alors que nous souhaitons tout de même leur assurer une reconstruction relative de même niveau que celle des autres marqueurs. C'est pourquoi nous normalisons la matrice d'abondance A en norme L_2 par colonnes selon

$$\tilde{A}_{ij} = \frac{A_{ij}}{\sqrt{\sum_{p=1}^n (A_{pj}^2)}}. \quad (4.7)$$

En introduisant la matrice diagonale N de termes diagonaux $N_j = \sqrt{\sum_{p=1}^n (A_{pj}^2)}$, cette normalisation s'écrit $\tilde{A} = AN^{-1}$.

Le problème à résoudre s'écrit

$$(W^*, \tilde{H}^*) = \underset{(\mathbf{W}, \tilde{\mathbf{H}}) \in \tilde{\mathcal{C}}}{\operatorname{argmin}} \quad \|\tilde{A} - \mathbf{W}\tilde{\mathbf{H}}\|_F^2 + \operatorname{pen}(\mathbf{W}, \tilde{\mathbf{H}}) \quad (4.8)$$

où $\tilde{\mathcal{C}}$ est l'ensemble

$$\tilde{\mathcal{C}} = \{(W, \tilde{H}) \in \mathcal{C}_0 \mid W \geq 0, \tilde{H} \geq 0, \tilde{F}_\Delta \tilde{H}^T \leq 0\} \quad (4.9)$$

et $\tilde{F}_\Delta = F_\Delta N^T$.

$\tilde{\mathcal{C}}$ est tel que $(W, H) \in \mathcal{C} \Leftrightarrow (W, \tilde{H}) \in \tilde{\mathcal{C}}$, la solution du problème d'inférence de CTFA est alors définie par $(W^*, H^*) = (W^*, \tilde{H}^* N^{-1})$.

Comme développé dans la Section 4.1.1.1, le problème 4.8 revient à une estimation du maximum de vraisemblance sous le modèle $\tilde{A} \sim \mathcal{N}(W\tilde{H}, \operatorname{Id})$ qui est équivalent à $A \sim \mathcal{N}(WH, \Sigma)$ avec $\Sigma = (NN^T)$, où encore

$$A_{ij} \sim \mathcal{N}((WH)_{ij}, \sum_i A_{ij}^2). \quad (4.10)$$

4.2 Terme de pénalisation

4.2.1 Pénalisations usuelles en NMF

La pénalisation est très utilisée dans différentes approches telles que la régression ou l'optimisation. Son utilisation dans le cadre de la NMF ne fait pas l'objet d'adaptations particulières dans la littérature. La pénalisation peut permettre de favoriser une structure donnée dans les matrices W et H , comme par exemple la parcimonie, et éventuellement d'inclure des connaissances a priori (groupes de variables). Elle permet également d'éliminer certains problèmes d'identifiabilité et ce faisant d'améliorer significativement les temps de calculs algorithmiques. Classiquement dans le cadre de la NMF, la pénalité s'écrit comme la somme d'une pénalité en W et d'une pénalité en H :

$$\operatorname{pen}(W, H) = \operatorname{pen}_1(W) + \operatorname{pen}_2(H)$$

Les pénalités usuelles comportent des constantes à calibrer. Les mêmes critères que ceux utilisés pour sélectionner la dimension k peuvent être utilisés.

4.2.2 Choix d'implémentation

La fonction de coût est pénalisée en W en utilisant une norme de Frobenius au carré, également appelée régularisation de Tikhonov ou pénalité ridge dans le contexte de la régression.

La pénalisation en la matrice H que nous avons choisie est moins classique (Kim and Park (2008)), elle a été généralisée par Yang et al. (2010) sous le nom de Lasso exclusif et permet d'encourager la parcimonie parmi un groupement de variables préalablement défini. Ici il s'agit des termes de la même colonne :

$$\|\tilde{H}\|_{1,2}^2 = \sum_{j=1}^r \left(\sum_{l=1}^k \tilde{H}_{lj} \right)^2 = \|\mathbf{1}^T \tilde{H}\|_2^2 \quad (4.11)$$

avec $\mathbf{1} \in \mathbb{R}^k$ le vecteur colonne dont tous les termes sont égaux à 1. Ainsi cette pénalité favorise la parcimonie dans la répartition des abondances de marqueurs fonctionnels au sein des CTFA.

Nous pouvons désormais compléter la fonction de coût (4.2) et écrire le problème d'optimisation suivant :

$$T(W, \tilde{H}, \alpha, \beta) = \|\tilde{A} - W\tilde{H}\|_F^2 + \alpha\|W\|_F^2 + \beta\|\mathbf{1}^T \tilde{H}\|_2^2 \quad (4.12)$$

et

$$(W^*, \tilde{H}^*) = \underset{(W, \tilde{H}) \in \tilde{\mathcal{C}}}{\operatorname{argmin}} T(W, \tilde{H}, \alpha, \beta), \quad (4.13)$$

où (W^*, \tilde{H}^*) est le couple solution.

4.2.3 Étude des constantes de pénalisation

Le problème de NMF sans terme de pénalisation est "mal posé" puisque l'on peut trouver une infinité de matrices inversibles $Q \in \mathcal{M}_{k,k}(\mathbb{R})$ telles que

$$WH = WQQ^{-1}H \text{ et } (WQ, Q^{-1}H) \in \tilde{\mathcal{C}}.$$

En particulier, toute matrice Q de la forme $\lambda \operatorname{Id}_k$ où $\lambda > 0$ et Id_k est la matrice identité de dimension k et toute matrice Q de permutation de taille k ont cette propriété.

Nous allons montrer dans la proposition 4.2.1 que la pénalisation permet en particulier de lever le problème d'identifiabilité lié à l'invariance par multiplication par $\lambda \operatorname{Id}_k$. Nous en déduisons ensuite qu'il est possible de formuler le problème d'inférence de (W^*, \tilde{H}^*) en n'utilisant qu'une

seule constante de pénalisation.

Proposition 4.2.1 *Si (W^*, H^*) est un optimum local de T , alors soient $\lambda, \alpha, \beta \in \mathbb{R}^{+*}$*

$$T(\lambda W^*, \tilde{H}^*/\lambda, \alpha, \beta) = T(W^*, \tilde{H}^*, \alpha, \beta) \quad \Rightarrow \quad \lambda = 1 \quad (4.14)$$

Preuve Nous travaillons dans l'espace de dimension finie $\mathcal{M}_{n,k}(\mathbb{R}) \times \mathcal{M}_{k,r}(\mathbb{R})$, toute les normes sont équivalentes donc nous choisissons de le munir de la norme infinie associée au produit cartésien des espaces métriques $(\mathcal{M}_{n,k}(\mathbb{R}), \|\cdot\|_F)$ et $(\mathcal{M}_{k,r}(\mathbb{R}), \|\cdot\|_F)$, c'est à dire $\|(W, H)\|_\infty = \max(\|W\|_F, \|H\|_F)$.

Soit (W^*, \tilde{H}^*) un minimum local de T dans $\tilde{\mathcal{C}}$. Il existe $\rho \in \mathbb{R}^{+*}$ tel que

$$\begin{aligned} \forall (W, \tilde{H}) \in \mathcal{B}((W^*, \tilde{H}^*), \rho) \cap \tilde{\mathcal{C}}, \\ T(W, \tilde{H}, \alpha, \beta) \geq T(W^*, \tilde{H}^*, \alpha, \beta). \end{aligned}$$

et $\mathcal{B}((W^*, \tilde{H}^*), \rho)$ la boule de $\mathcal{M}_{n,k}(\mathbb{R}) \times \mathcal{M}_{k,r}(\mathbb{R})$ définie par

$$\mathcal{B}((W^*, \tilde{H}^*), \rho) = \{(W, H) \in \mathcal{M}_{n,k}(\mathbb{R}) \times \mathcal{M}_{k,r}(\mathbb{R}) \mid \max(\|W - W^*\|_F, \|\tilde{H} - \tilde{H}^*\|_F) \leq \rho\}. \quad (4.15)$$

Soient $\lambda, \alpha, \beta \in \mathbb{R}^{+*}$, $(\lambda W^*, \tilde{H}^*/\lambda) \in \tilde{\mathcal{C}}$ et

$$\begin{aligned} T(\lambda W^*, \tilde{H}^*/\lambda, \alpha, \beta) &= T(W^*, \tilde{H}^*, \alpha, \beta) \\ &+ (\lambda^2 - 1)\alpha\|W^*\|_F^2 + \left(\frac{1}{\lambda^2} - 1\right)\beta\|\mathbf{1}^T \tilde{H}^*\|_2^2 \end{aligned}$$

Considérons la fonction $g(\lambda) = (\lambda^2 - 1)\alpha\|W^*\|_F^2 + (1/\lambda^2 - 1)\beta\|\mathbf{1}^T \tilde{H}^*\|_2^2$. Ils suffit de prouver que si $g(\lambda)$ s'annule alors $\lambda = 1$.

La dérivée de g : $g'(\lambda) = 2(\lambda\alpha\|W^*\|_F^2 - \beta\|\mathbf{1}^T \tilde{H}^*\|_2^2/\lambda^3)$ s'annule en

$$\lambda_0 = \left(\frac{\beta\|\mathbf{1}^T \tilde{H}^*\|_2^2}{\alpha\|W^*\|_F^2}\right)^{1/4}. \quad (4.16)$$

La fonction g est strictement décroissante sur $]0, \lambda_0[$, atteint son minimum en λ_0 et est strictement croissante sur $]\lambda_0, +\infty[$. Supposons $\lambda_0 \neq 1$, la fonction g est strictement croissante ou décroissante au voisinage de 1, il existe donc λ' suffisamment proche de 1 tel que $(\lambda'W^*, \tilde{H}^*/\lambda') \in \mathcal{B}((W^*, \tilde{H}^*), \rho) \cap \tilde{\mathcal{C}}$ et $g(\lambda') < g(1) = 0$. Ceci contredit le fait que (W^*, \tilde{H}^*) est un minimum

local de T et prouve l'égalité.

Ainsi $\lambda_0 = 1$ et pour tout $\lambda \in \mathbb{R}^{+*}$ différent de 1, $g(\lambda) > 0$. Ceci prouve la proposition. \square

En conséquence de la démonstration précédente et par définition de λ_0 , nous avons montré le corollaire suivant

Corollaire 4.2.1 *Soit (W^*, \tilde{H}^*) un minimum local de T , alors*

$$\alpha \|\mathbf{1}^t \tilde{H}^*\|_2^2 = \beta \|W^*\|_F^2.$$

Ce résultat nous permet qu'on peut n'utiliser qu'une seule constante de pénalisation et que ceci équivaut à fixer le rapport $\|\mathbf{1}^t \tilde{H}^*\|_2^2 / \|W^*\|_F^2$ à 1. Pour cela, définissons

$$E(W, \tilde{H}, \alpha) = T(W, \tilde{H}, \alpha, \alpha). \quad (4.17)$$

Nous allons établir une équivalence entre les minima locaux de (4.12) et ceux de (4.17) dans la proposition suivante.

Proposition 4.2.2 *Pour tout $\alpha, \beta > 0$, soit $\lambda = (\beta/\alpha)^{1/4}$.*

(W^, \tilde{H}^*) est un minimum local de $T(W, \tilde{H}, \alpha, \beta)$ si et seulement si $(\lambda W^*, \tilde{H}^*/\lambda)$ est un minimum local de $E(W, \tilde{H}, \sqrt{\alpha\beta})$.*

Preuve Par définition du minimum local, il existe $\rho \in \mathbb{R}^{+*}$ tel que pour tout $(W, \tilde{H}) \in \mathcal{B}((W^*, \tilde{H}^*), \rho) \cap \tilde{\mathcal{C}}$

$$T(W, \tilde{H}, \alpha, \beta) \geq T(W^*, \tilde{H}^*, \alpha, \beta), \quad (4.18)$$

où $\mathcal{B}((W^*, \tilde{H}^*), \rho)$ est la boule fermée de centre (W^*, \tilde{H}^*) et de rayon ρ telle que définie en (4.15). Soit $\rho' = \min(\rho\lambda, \rho/\lambda)$. Alors pour tout $(W, \tilde{H}) \in \mathcal{B}((\lambda W^*, \tilde{H}^*/\lambda), \rho') \cap \tilde{\mathcal{C}}$

$$\begin{aligned} \|W/\lambda - \lambda W^*/\lambda\|_F &\leq \rho'/\lambda \leq \rho \\ \|\lambda \tilde{H} - \lambda \tilde{H}^*/\lambda\|_F &\leq \lambda \rho' \leq \rho \end{aligned}$$

donc $(W/\lambda, \lambda \tilde{H}) \in \mathcal{B}(W^*, \tilde{H}^*, \rho) \cap \tilde{\mathcal{C}}$. On en déduit que $T(W/\lambda, \lambda \tilde{H}, \alpha, \beta) \geq T(W^*, \tilde{H}^*, \alpha, \beta)$.

Soient $\alpha, \beta > 0$, définissons $\lambda = (\alpha/\beta)^{1/4}$, de telle sorte que $\beta\lambda^2 = \sqrt{\alpha\beta}$ et $\alpha/\lambda^2 = \sqrt{\alpha\beta}$. Alors

$$\begin{aligned} T(W/\lambda, \lambda\tilde{H}/, \alpha, \beta) &= \|\tilde{A} - W/\lambda\tilde{H}\lambda\|_F^2 + \alpha/\lambda^2\|W\|_F^2 + \beta\lambda^2\|\mathbf{1}^t\tilde{H}\|_2^2 \\ &= \|\tilde{A} - W\tilde{H}\|_F^2 + \alpha\|W\|_F^2 + \beta\|\mathbf{1}^t\tilde{H}\|_2^2 \\ &= T(W, \tilde{H}, \sqrt{\alpha\beta}, \sqrt{\alpha\beta}) \\ &= E(W, \tilde{H}, \sqrt{\alpha\beta}) \end{aligned}$$

De même, on montre que $T(W^*, \tilde{H}^*, \alpha, \beta) = E(\lambda W^*, \tilde{H}^*/\lambda, \sqrt{\alpha\beta})$. Nous avons donc montré que pour tout $(W, \tilde{H}) \in \mathcal{B}((\lambda W^*, \tilde{H}^*/\lambda), \rho') \cap \tilde{\mathcal{C}}$,

$$E(W, \tilde{H}, \sqrt{\alpha\beta}) \geq E(\lambda W^*, \tilde{H}^*/\lambda, \sqrt{\alpha\beta})$$

Ceci prouve que $(\lambda W^*, \tilde{H}^*/\lambda)$ est un minimum local de $E(W, \tilde{H}, \sqrt{\alpha\beta})$. La réciproque se montre de manière symétrique. \square

Cette proposition démontre que la minimisation de E permet d'obtenir un minimum de T et donc qu'il suffit de résoudre le problème de NMF pénalisé à une constante pour obtenir directement une solution du problème pénalisé à deux constantes.

4.3 Propriétés du problème d'optimisation de NMF

Nous présentons dans un premier temps les propriétés du problème de NMF classique. Dans un deuxième temps nous revenons sur l'impact de l'ajout de la contrainte de $\tilde{\mathcal{C}}$ sur ces propriétés.

4.3.1 Le problème général

Nous considérons le problème de minimisation de la fonction de coût suivante :

$$f(W, H) = \mathcal{D}(A|WH) + pen_1(W) + pen_2(H) \quad (4.19)$$

pour lequel nous faisons les hypothèses suivantes

- pen_1 et pen_2 sont convexes, dérivables et de dérivées continues sur \mathcal{C}_0
- il existe $\kappa_1, \kappa_2 \in \mathbb{R}^{*+}$ et $\gamma_1, \gamma_2 \in \{1, 2\}$ tels que pour tout $(W, H) \in \mathcal{C}_0$

$$\begin{aligned} pen_1(W) &\geq \kappa_1\|W\|_F^{\gamma_1} \\ pen_2(H) &\geq \kappa_2\|H\|_F^{\gamma_2} \end{aligned} \quad (4.20)$$

On peut appliquer à la NMF les outils classiques d'optimisation sous contrainte, ce sera l'objet de la Section 4.3.1.1. Cependant, il s'agit d'un problème d'optimisation bi-convexe. On peut donc exploiter cette structure particulière, dont les propriétés sont rappelées dans la Section 4.3.1.2, pour proposer des approches de minimisation adaptées et garantir des résultats théoriques de convergence (section 4.3.1.3).

4.3.1.1 Point stationnaire et minimum local

Dans le cas d'un problème d'optimisation soumis à des contraintes *linéaires* d'égalité ou d'inégalité, des résultats en théorie de l'optimisation (Bertsekas, 1999) garantissent que les conditions de Karush-Kuhn-Tucker (KKT) sont des conditions *suffisantes* pour obtenir un point stationnaire. Le problème de minimisation de NMF peut s'écrire sous cette forme

$$\begin{aligned} & \min_{\mathbf{W} \in \mathcal{M}_{n,k}(\mathbb{R}), \mathbf{H} \in \mathcal{M}_{k,r}(\mathbb{R})} f(\mathbf{W}, \mathbf{H}) \\ \text{tel que} & \quad (-\mathbf{W}) \leq 0 \\ & \quad (-\mathbf{H}) \leq 0. \end{aligned}$$

L'écriture des conditions de KKT fait appel à une fonction intermédiaire, le Lagrangien, dont l'expression est ici donnée par

$$\mathcal{L}(W, H, U_W, U_H) = f(W, H) - \text{tr}(U_W^T W) - \text{tr}(U_H^T H), \quad (4.21)$$

où $U_W \in \mathcal{M}_{n,k}(\mathbb{R}^+)$ et $U_H \in \mathcal{M}_{k,r}(\mathbb{R}^+)$, sont les matrices de multiplicateurs de Lagrange et $\text{tr}(\cdot)$ l'opérateur trace.

Dans ces conditions (W°, H°) vérifie les conditions de KKT pour le problème de NMF si il existe $U_W \in \mathcal{M}_{n,k}(\mathbb{R}^+)$ et $U_H \in \mathcal{M}_{k,r}(\mathbb{R}^+)$ tels que :

$$\nabla_{[W,H]} \mathcal{L}(W^\circ, H^\circ, U_W, U_H) = \begin{bmatrix} \nabla_W f(W^\circ, H^\circ) - \nabla_W \text{tr}(U_W^T W^\circ) \\ \nabla_H f(W^\circ, H^\circ) - \nabla_H \text{tr}(U_H^T H^\circ) \end{bmatrix} = \begin{bmatrix} 0_W \\ 0_H \end{bmatrix}$$

et

$$\begin{cases} (U_W)_{i,l} W_{i,l}^\circ = 0, & \text{pour tout } i \in \{1, \dots, n\} \text{ et } l \in \{1, \dots, k\}, \\ (U_H)_{l,j} H_{l,j}^\circ = 0, & \text{pour tout } l \in \{1, \dots, k\} \text{ et } j \in \{1, \dots, r\}, \end{cases} \quad (4.22)$$

où $0_W \in \mathcal{M}_{n,k}(\mathbb{R}^+)$ et $0_H \in \mathcal{M}_{k,r}(\mathbb{R}^+)$ sont des matrices nulles. On dit alors que (W°, H°) est un point stationnaire de f sur \mathcal{C}_0 .

Dans le cas d'une fonction convexe, un point est stationnaire si et seulement si le minimum

de cette fonction est atteint en ce point.

Dans le cas général, tout minimum local de $f(W, H)$ sur \mathcal{C}_0 est un point stationnaire. Cependant la réciproque est fautive : tout point stationnaire n'est pas un minimum local, comme le montre l'exemple ci-dessous.

Un point stationnaire n'est pas toujours un minimum local. Considérons

$$f(W, H) = \|A - WH\|_F^2 + \alpha(\|W\|_F^2 + \|H\|_F^2), \quad (4.23)$$

où A est une matrice non nulle à termes positifs et α sont tel qu'il existe (i, j) pour lesquels $A_{ij} > \alpha > 0$. Considérons le point nul $(0_W, 0_H)$, et $(U_W, U_H) = (0_W, 0_H)$. Alors

$$\nabla_{[W, H]} \mathcal{L}(0_W, 0_H, 0_W, 0_H) = \begin{bmatrix} 0_W \\ 0_H \end{bmatrix}.$$

Les conditions supplémentaires (4.22) sur les termes de 0_W et 0_H sont également vérifiées, donc $(0_W, 0_H)$ est un point stationnaire.

Par ailleurs, soit $\mathcal{B}((0_W, 0_H), \rho)$ la boule de $\mathcal{M}_{n,k}(\mathbb{R}) \times \mathcal{M}_{k,r}(\mathbb{R})$ de centre $(0_W, 0_H)$ et de rayon ρ . Si $(0_W, 0_H)$ est un minimum local, il doit exister $\rho \in \mathbb{R}^{+*}$ tel que pour tout $(W, H) \in \mathcal{B}((0_W, 0_H), \rho)$, $f(W, H) \geq f(0_W, 0_H)$.

Or quel que soit $\rho \in \mathbb{R}^{+*}$, on peut construire $(\hat{W}, \hat{H}) \in \mathcal{B}((0_W, 0_H), \rho)$ qui contredit l'inégalité précédente. Soit (i, j) tel que $A_{ij} - \alpha > 0$, considérons les matrices \hat{W} et \hat{H} suivantes :

$$\hat{W}_{lm} = \begin{cases} V^{1/2} & \text{si } (l, m) = (i, 1) \\ 0 & \text{sinon} \end{cases} \quad \hat{H}_{lm} = \begin{cases} V^{1/2} & \text{si } (l, m) = (1, j) \\ 0 & \text{sinon} \end{cases}$$

où V est une constante vérifiant $V < \min(\rho^2, 2(A_{ij} - \alpha))$. Alors par construction $(\hat{W}, \hat{H}) \in \mathcal{B}((0_W, 0_H), \rho)$. D'autre part,

$$\begin{aligned} f(\hat{W}, \hat{H}) &= \|A - \hat{W}\hat{H}\|_F^2 + \alpha(\|\hat{W}\|_F^2 + \|\hat{H}\|_F^2) \\ &= \|A\|_F^2 - 2VA_{ij} + V^2 + 2\alpha V \\ &= \|A\|_F^2 + (2\alpha + V - 2A_{ij})V \\ &< \|A\|_F^2 = f(0_W, 0_H), \end{aligned}$$

ce qui prouve que $(0_W, 0_H)$ est un point stationnaire sans être un minimum local.

4.3.1.2 Propriétés de la fonction de coût

Soit g une fonction définie sur un ensemble $X = X_1 \times X_2$ tel que X_1 et X_2 soient des ensembles convexes, on dit que g est bi-convexe sur X si et seulement si, pour tout $x_1 \in X_1$ la fonction $x_2 \mapsto g(x_1, x_2)$ est convexe sur X_2 et pour tout $x_2 \in X_2$, la fonction $x_1 \mapsto g(x_1, x_2)$ est convexe sur X_1 .

Dans le cas de la NMF, la fonction de coût f définie en (4.19) est bi-convexe sur \mathcal{C}_0 car :

$$f_H(\mathbf{W}) = \mathcal{D}(A|\mathbf{W}H) + pen_1(\mathbf{W}) \quad \text{est convexe en } \mathbf{W} \text{ sur } \mathcal{M}_{n,k}(\mathbb{R}^+) \quad (4.24)$$

et

$$f_W(\mathbf{H}) = \mathcal{D}(A|\mathbf{W}H) + pen_2(\mathbf{H}) \quad \text{est convexe en } \mathbf{H} \text{ sur } \mathcal{M}_{k,r}(\mathbb{R}^+) \quad (4.25)$$

Ces propriétés de convexité impliquent que f_H et f_W possèdent chacune un unique minimum global et aucun autre minimum local. En revanche les minimiseurs associés ne sont pas nécessairement uniques.

Par contre une fonction bi-convexe n'est pas nécessairement convexe, f peut donc avoir plusieurs minima locaux. Toutefois, elle possède un minimum global ainsi qu'un minimiseur associé, c'est l'objet de la proposition qui suit.

Proposition 4.3.1 *La borne inférieure de f sur \mathcal{C}_0 , notée $\mu = \inf_{(W,H) \in \mathcal{C}_0} f(W, H)$, est finie et il existe $(W^*, H^*) \in \mathcal{C}_0$ tel que $\mu = f(W^*, H^*)$*

Preuve Comme f est positive, μ est finie et positive.

Soit $0_W \in \mathcal{M}_{n,k}(\mathbb{R}^+)$ et $0_H \in \mathcal{M}_{k,r}(\mathbb{R}^+)$ les matrices nulles et $K = f(0_W, 0_H)$, par définition $\mu \leq K$.

Définissons

$$\mathcal{L}_K = \{(W, H) \in \mathcal{C}_0 \mid f(W, H) \leq K\}, \quad (4.26)$$

on a donc par construction $\mu = \inf_{(W,H) \in \mathcal{C}_0} f(W, H) = \inf_{\mathcal{L}_K} f(W, H)$.

Or, comme f est continue, \mathcal{L}_K est un ensemble fermé. De plus, pour tout $(W, H) \in \mathcal{L}_K$,

grâce à l'hypothèse (4.20),

$$\begin{aligned}\|W\|_F &\leq \left(\frac{1}{\kappa_1} \text{pen}_1(W)\right)^{1/\gamma_1} \leq \left(\frac{1}{\kappa_1} f(W, H)\right)^{1/\gamma_1} \leq \left(\frac{K}{\kappa_1}\right)^{1/\gamma_1} \\ \|H\|_F &\leq \left(\frac{1}{\kappa_2} \text{pen}_2(H)\right)^{1/\gamma_2} \leq \left(\frac{1}{\kappa_2} f(W, H)\right)^{1/\gamma_2} \leq \left(\frac{K}{\kappa_2}\right)^{1/\gamma_2}\end{aligned}$$

On en déduit que \mathcal{L}_K est fermé borné donc compact. Le minimum de f est donc atteint sur \mathcal{L}_K et il existe $(W^*, H^*) \in \mathcal{L}_K$ tel que $\mu = f(W^*, H^*)$. \square

4.3.1.3 Minimisation des problèmes bi-convexes

Au vu des propriétés de f , une idée intuitive pour améliorer la résolution du problème de minimisation de f consiste à exploiter la bi-convexité en minimisant (4.25) et (4.24) de façon alternée. En effet, comme souligné précédemment, même en présence de contraintes de positivité, la minimisation des fonctions convexes est très bien documentée dans la littérature. L'algorithme général de minimisation alternée est présenté ci-dessous.

Algorithme 1 Minimisation alternée dans le cas de la NMF

Initialisation aléatoire de $W^{(0)}$ et $\tilde{H}^{(0)}$

while Convergence \neq True **do**

$$W^{(t+1)} = \underset{\mathbf{W} \in \mathcal{M}_{\mathbf{n}, \mathbf{k}}(\mathbb{R}^+)}{\text{argmin}} \quad f(\mathbf{W}, H^{(t)}) \quad (4.27)$$

$$H^{(t+1)} = \underset{\mathbf{H} \in \mathcal{M}_{\mathbf{k}, \mathbf{r}}(\mathbb{R}^+)}{\text{argmin}} \quad f(W^{(t+1)}, \mathbf{H}) \quad (4.28)$$

end while

Cet algorithme est une méthode de type Gauss-Seidel, encore appelée méthode de minimisation par blocs dans la littérature (Grippo and Sciandrone (2000)). En appliquant l'Algorithme 1, on génère une suite de valeurs de la fonction de coût qui converge puisqu'elle est minorée par 0, monotone et décroissante. En effet, pour tout $t \in \mathbb{N}^*$,

$$f(W^{(t+1)}, H^{(t+1)}) \leq f(W^{(t+1)}, H^{(t)}) \leq f(W^{(t)}, H^{(t)}).$$

Cependant, ceci ne donne aucune information sur la convergence de la suite $\{(W^{(t)}, H^{(t)})\}_{t \in \mathbb{N}}$. Or il à été supposé ou démontré dans les sections précédentes que :

- f est continuellement différentiable sur \mathcal{C}_0
- $\mathcal{C}_0 = \mathcal{M}_{n,k}(\mathbb{R}^+) \times \mathcal{M}_{k,r}(\mathbb{R}^+)$ est le produit cartésien des deux sous-espaces sur lesquels (4.25) et (4.24) sont minimisées.
- les sous-espaces $\mathcal{M}_{n,k}(\mathbb{R}^+)$ et $\mathcal{M}_{k,r}(\mathbb{R}^+)$ sont fermés, non-vides et convexes.
- enfin, nous avons montré dans la Section 4.3.1.2 que pour tout $(W^0, H^0) \in \mathcal{C}_0$ l'ensemble $\mathcal{L}_{f(W^0, H^0)}$ défini en (4.26) est un compact.

D'après Grippo and Sciandrone (2000), ces hypothèses sont suffisantes pour garantir la convergence d'une suite extraite de $\{(W^{(t)}, H^{(t)})\}_{t \in \mathbb{N}}$ dont la limite est un point stationnaire du problème de NMF.

4.3.2 Particularité de la NMF sous contrainte

Dans notre modèle, la fonction de coût $f(W, H)$ s'écrit

$$E(W, \tilde{H}, \alpha) = \|\tilde{A} - W\tilde{H}\|_F^2 + \alpha(\|W\|_F^2 + \|\mathbf{1}^T \tilde{H}\|_2^2). \quad (4.29)$$

les pénalités considérées vérifient bien les hypothèses (4.20), car quel que soit $H \in \mathcal{M}_{k,r}(\mathbb{R}^+)$

$$\|\mathbf{1}^T \tilde{H}\|_2^2 \leq \|\tilde{H}\|_F^2. \quad (4.30)$$

Nous cherchons à minimiser la fonction de coût $E(W, \tilde{H}, \alpha)$ sur l'ensemble $\tilde{\mathcal{C}}$ défini en (4.9). $\tilde{\mathcal{C}}$ s'écrit :

$$\begin{aligned} \tilde{\mathcal{C}} &= \mathcal{M}_{n,k}(\mathbb{R}^+) \times \{\tilde{H} \in \mathcal{M}_{k,r}(\mathbb{R}^+) \mid \tilde{F}_\Delta \tilde{H}^T \leq 0\} \\ &= \mathcal{C}_W \times \tilde{\mathcal{C}}_H \end{aligned}$$

Cet ensemble est encore un ensemble convexe, produit de deux ensembles convexes.

Le problème d'optimisation sous contraintes linéaires s'écrit

$$\begin{aligned} &\min_{\mathbf{W} \in \mathcal{M}_{n,k}(\mathbb{R}), \mathbf{H} \in \mathcal{M}_{k,r}(\mathbb{R})} E(\mathbf{W}, \tilde{\mathbf{H}}, \alpha) \\ &\quad (-\mathbf{W}) \leq 0 \\ \text{tel que} \quad &(-\mathbf{H}) \leq 0 \\ &(\tilde{F}_\Delta \tilde{\mathbf{H}}^T) \leq 0. \end{aligned}$$

Le lagrangien correspondant est le suivant :

$$\mathcal{L}(W, \tilde{H}, U_W, U_{\tilde{H}}, U_{\tilde{F}}) = E(W, \tilde{H}, \alpha) - \text{tr}(U_W^T W) - \text{tr}(U_{\tilde{H}}^T \tilde{H}) + \text{tr}(U_{\tilde{F}}^T \tilde{F}_\Delta \tilde{H}^T), \quad (4.31)$$

où $U_W \in \mathcal{M}_{n,k}(\mathbb{R}^+)$, $U_{\tilde{H}} \in \mathcal{M}_{k,r}(\mathbb{R}^+)$ et $U_{\tilde{F}} \in \mathcal{M}_{c,r}(\mathbb{R}^+)$ sont les matrices de multiplicateurs de Lagrange. Les conditions de KKT introduites en Section 4.3.1.1 s'écrivent :

$$\nabla_{[W,\tilde{H}]} \mathcal{L}(W^\circ, \tilde{H}^\circ, U_W, U_{\tilde{H}}) = \begin{bmatrix} \nabla_W E(W^\circ, \tilde{H}^\circ, \alpha) & -\nabla_W \text{tr}(U_W^T W^\circ) \\ \nabla_{\tilde{H}} E(W^\circ, \tilde{H}^\circ, \alpha) & -\nabla_{\tilde{H}} \text{tr}(U_{\tilde{H}}^T \tilde{H}^\circ) + \nabla_{\tilde{H}} \text{tr}(U_{\tilde{F}}^T \tilde{F}_\Delta \tilde{H}^{\circ T}) \end{bmatrix} = \begin{bmatrix} 0_W \\ 0_{\tilde{H}} \end{bmatrix}$$

$$\text{et } \begin{cases} (U_W)_{i,l} W_{i,l}^\circ = 0, & \text{pour tout } i \in \{1, \dots, n\} \text{ et } l \in \{1, \dots, k\}, \\ (U_{\tilde{H}})_{l,j} \tilde{H}_{l,j}^\circ = 0, & \text{pour tout } j \in \{1, \dots, r\} \text{ et } l \in \{1, \dots, k\}, \\ (U_{\tilde{F}})_{m,l} (\tilde{F}_\Delta \tilde{H}^{\circ T})_{m,l} = 0, & \text{pour tout } m \in \{1, \dots, c\} \text{ et } l \in \{1, \dots, k\}, \end{cases}$$

E est bi-convexe donc les propriétés de convergence définies dans le cas général de la NMF non contrainte restent valables. La minimisation peut se faire selon l'Algorithme 2.

Algorithme 2 Minimisation alternée dans le cas de la NMF contrainte

Initialisation aléatoire de $W^{(0)}$ et $\tilde{H}^{(0)}$

while Convergence \neq True **do**

$$W^{(t+1)} = \underset{\mathbf{W} \in \mathcal{C}_W}{\text{argmin}} E(\mathbf{W}, \tilde{H}^{(t)}, \alpha) \quad (4.32)$$

$$\tilde{H}^{(t+1)} = \underset{\tilde{\mathbf{H}} \in \tilde{\mathcal{C}}_H}{\text{argmin}} E(W^{(t+1)}, \tilde{\mathbf{H}}, \alpha) \quad (4.33)$$

end while

En suivant cet algorithme nous avons la garantie qu'il existe une sous-suite extraite de la suite des $\{W^{(t)}, \tilde{H}^{(t)}\}_{t \in \mathbb{N}}$ qui converge vers un point stationnaire.

4.4 Résolution du problème de minimisation

4.4.1 Etat de l'art

La résolution des problèmes de NMF se fait généralement en exploitant la structure bi-convexe du problème à l'aide de mises à jour alternées des matrices W et H . On peut distinguer deux classes d'approches. La première regroupe des méthodes où l'on effectue des résolutions partielles des sous-problèmes d'optimisation au cours desquelles on fait simplement décroître la fonction de coût sans calculer de minimum. La deuxième regroupe les méthodes où l'on effectue des résolutions complètes des sous-problèmes d'optimisation, ce sont les méthodes de minimisation alternée comme présenté dans l'Algorithme 1.

Dans une première partie, deux algorithmes appartenant à la première classe sont présentés, la Multiplicative Update Rule (MUR) (Lee and Seung (1999)) et la projection alternée des moindres carrés proposée par Berry et al. (2007).

Depuis, de nombreuses approches ont été proposées pour aboutir en 2006 à un consensus sur la minimisation alternée de l'Algorithme 1. Dès lors la recherche s'est focalisée sur la résolution efficace des sous-problèmes (4.27) et (4.28). Plusieurs algorithmes classiques d'optimisation ont été proposés. Parmi eux, Kim and Park (2008) proposent une méthode d'Active Set. Lin (2007) et Han et al. (2009) utilisent des variantes de gradient projeté. Enfin, Guan et al. (2012) utilisent un algorithme d'accélération de gradient de Nesterov. Quelques unes de ces méthodes sont présentées dans une deuxième partie.

4.4.1.1 Quelques algorithmes à résolution partielle

Multiplicative Update Rule. La multiplicative update rule se fait selon l'algorithme suivant dans le cas de la norme de Frobenius et l'absence de pénalisation.

Algorithme 3 Multiplicative Update Rule dans le cas de la norme de Frobenius

Initialisation aléatoire de $W^{(0)}$ et $H^{(0)}$

while Convergence \neq True **do**

$$\forall i, j, \quad W_{i,j}^{(t+1)} = \mathbf{W}_{ij}^{(t)} \frac{(AH^{(t)T})_{ij}}{(\mathbf{W}^{(t)}H^{(t)}H^{(t)T})_{ij}}$$

$$\forall i, j, \quad H_{i,j}^{(t+1)} = \mathbf{H}_{ij}^{(t)} \frac{(W^{(t)T}A)_{ij}}{(W^{(t)T}W^{(t)}\mathbf{H}^{(t)})_{ij}}$$

end while

Cette méthode repose sur l'idée suivante (Lee and Seung, 2001). Lorsque W est fixé, $\|A - WH\|_F^2$ s'écrit

$$\|A - WH\|_F^2 = \sum_{j=1}^r \|A_{(:,j)} - WH_{(:,j)}\|_2^2, \quad (4.34)$$

où la notation $M_{(:,j)}$ (respectivement $M_{(i,:)}$) désigne la colonne numéro j (respectivement la ligne numéro i) de la matrice M . Définissons sur \mathbb{R}^k la fonction F par

$$F(h) = \frac{1}{2} \|A_{(:,j)} - Wh\|_2^2 \quad (4.35)$$

L'étape de calcul de H dans le problème de NMF peut donc se traiter en réduisant successive-

ment la valeur de F pour chaque colonne de la matrice A .

Définissons sur $\mathbb{R}^k \times (\mathbb{R}^*)^k$ la fonction G par

$$G(h, h') = F(h') + (h - h')^T \nabla F(h') + \frac{1}{2}(h - h')^T K(h')(h - h'), \quad (4.36)$$

où $K(h')$ est la matrice diagonale de dimension k de terme diagonal $(K(h'))_{aa} = (W^T W h')_a / h'_a$.

On montre que pour tout $(h, h') \in (\mathbb{R}^*)^k \times (\mathbb{R}^*)^k$, on a

$$G(h, h') \geq F(h), \quad \text{et } G(h, h) = F(h). \quad (4.37)$$

On dit alors que G est une fonction auxiliaire pour F . On montre alors que la suite définie par

$$h^{(t+1)} = \underset{h}{\operatorname{argmin}} G(h, h^{(t)})$$

fournit une suite décroissante $\{F(h^{(t)})\}_{n \in \mathbb{N}}$ de valeurs de la fonction F comme le montre la Figure 4.2. L'Algorithme 3 correspond à l'implémentation de cette procédure.

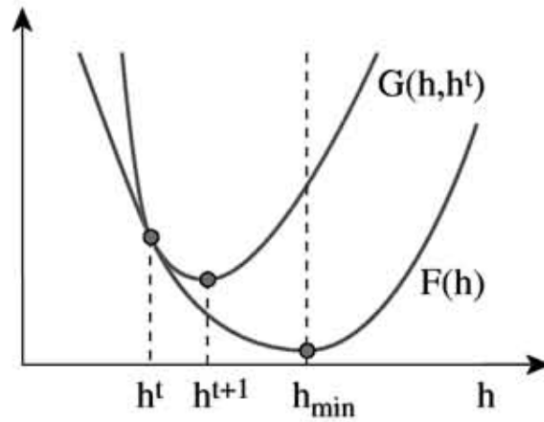


FIGURE 4.2: La minimisation de la fonction $G(h, h^{(t)})$ garantit que $F^{(t+1)} \leq F^{(t)}$. Figure tirée de Lee and Seung (2001).

Des formulations adaptées ont été proposées pour les divergences de Frobenius et KLG, avec et sans pénalisation (Lee and Seung (1999), Hoyer (2004)) et dans un cadre plus général concernant d'autres mesures de divergence (Dhillon and Sra, 2005). Comme nous l'avons rappelé ci-dessus, cette approche garantit la positivité de la suite de matrices $\{(W^{(t)}, H^{(t)})\}_{t \in \mathbb{N}}$ générée ainsi qu'une diminution monotone de la fonction de coût. Donc la valeur de celle-ci converge.

Cependant cette méthode présente plusieurs inconvénients. En pratique, lorsqu'un terme est nul à l'initialisation, il le restera. De même, aucun terme ne peut formellement s'annuler pendant l'optimisation. De plus cet algorithme n'effectue pas de minimisation alternée donc les résultats de convergence de la Section 4.3.1.3 ne s'appliquent pas. On ne peut pas garantir la convergence de la suite $\{(W^{(t)}, H^{(t)})\}_{t \in \mathbb{N}}$ et encore moins la convergence d'une sous-suite vers un point stationnaire (Gonzalez and Zhang (2005)). Enfin tous les autres algorithmes présentés dans ce chapitre offrent de meilleures performances en ce qui concerne les temps de calculs.

Moindre carrés projetés. Cette méthode n'est utilisable que pour une norme de Frobenius. Cet algorithme met à jour les matrices W et H en calculant la solution du problème des moindres carrés non contraints avant de faire la projection euclidienne sur le cône positif, de sorte que tous les termes négatifs de la solution deviennent nuls. Les termes de l'Algorithme 4 s'écrivent par annulation simple du gradient de $\|A - WH\|_F^2$ sans pénalisation.

Algorithme 4 Moindres carrés projetés

```

Initialisation aléatoire de  $W^{(0)}$  et  $H^{(0)}$ 
while Convergence  $\neq$  True do
  Résolution de :  $H^{(t)} H^{(t)T} \mathbf{W}_{\text{temp}} = H^{(t)} A^T$ 
   $W^{(t+1)} = [W_{\text{temp}}]_+$ 
  Résolution de :  $W^{(t+1)T} W^{(t+1)} \mathbf{H}_{\text{temp}} = W^{(t+1)T} A$ 
   $H^{(t+1)} = [H_{\text{temp}}]_+$ 
end while                                % Avec  $[\cdot]_+$  la projection euclidienne sur le cône positif

```

Cet algorithme souffre des mêmes défauts de convergence que la MUR, car le problème de minimisation exact n'est pas résolu à chaque étape.

Ces deux algorithmes présentent l'avantage de la simplicité d'implémentation et de la rapidité ce qui les rend attractifs malgré le manque de garanties de convergence vers un point stationnaire.

4.4.1.2 Quelques algorithmes de minimisation alternée

Active Set. L'implémentation efficace de cet algorithme dans le cadre de la NMF à été proposée par Kim and Park (2008) qui l'appliquent à la détermination successive des colonnes notées x de W et H du problème de NMF.

Cette approche nécessite la formulation du problème de minimisation comme un problème de

moindres carrés sous contrainte de positivité et repose sur la notion de contraintes dites actives et passives. Par convention, les différentes contraintes sont formulées comme des égalités ou inégalités relatives à 0. On dit que la contrainte de positivité associée à la coordonnée x_i d'un point x est active si $x_i = 0$ et passive si $x_i > 0$. On note \mathcal{E} l'ensemble des indices des contraintes actives et \mathcal{S} celui des indices des contraintes passives.

Pour toute matrice M , on introduit la notation $M^{\mathcal{S}}$ pour désigner la matrice de même taille que M , dont les colonnes d'indices appartenant à \mathcal{S} sont égales à celles de M et les autres sont nulles.

L'algorithme d'Active Set résout le problème de minimisation en identifiant ces deux ensembles. Lorsque ces deux ensembles sont identifiés, on construit x^* la solution du problème de la façon suivante. Pour tout indice i de l'ensemble passif \mathcal{S} , la valeur de x_i^* est donnée par résolution des moindres carrés sans contrainte et pour tout indice j de l'ensemble actif \mathcal{E} la valeur de x_j^* est mise à 0. La détermination de \mathcal{S} et \mathcal{E} se fait selon l'algorithme suivant.

Algorithme 5 Active Set : pour $\min_{\mathbf{x} \geq 0} \|Bx - y\|_2$, $B \in \mathcal{M}_{m,n}(\mathbb{R})$, $y \in \mathbb{R}^m$

Initialisation :

$$x = 0$$

$$\mathcal{E} = \{1, \dots, n\}$$

% Active set

$$\mathcal{S} = \emptyset$$

% Passive set

$$\nabla^x = B^T(y - Bx)$$

% opposé du Gradient en x de $\|Bx - y\|_2$

while ($\mathcal{E} \neq \emptyset$ et $\exists j \in \mathcal{E} \mid (\nabla^x)_j > 0$) **do**

$$j = \underset{t}{\operatorname{arg\,max}} (\nabla^x)_t$$

j passe de l'ensemble \mathcal{E} à l'ensemble \mathcal{S}

$$z = \underset{\mathbf{w}}{\operatorname{arg\,min}} \|B^{\mathcal{S}}\mathbf{w} - y\|_2$$

$$z_j = 0, \text{ si } j \in \mathcal{E}$$

% Les éléments de \mathcal{E} saturent la contrainte et sont annulés

while ($\exists j \in \mathcal{S} \mid z_j \leq 0$) **do**

$$\mathcal{J} = \{j \in \mathcal{S} \mid z_j \leq 0 \text{ et } z_j = \min_j x_j / (x_j - z_j)\}$$

tous les éléments de \mathcal{J} sont retirés de \mathcal{S} et passent dans \mathcal{E}

$$z = \underset{\{z_j \mid j \in \mathcal{S}\}}{\operatorname{arg\,min}} \|B^{\mathcal{S}}x - y\|_2$$

% Seul les z_j pour $j \in \mathcal{S}$

end while

$$x = z$$

$$\nabla^x = B^T(y - Bx)$$

end while

$$x^* = x$$

Sa vitesse d'exécution et le nombre d'itérations le rend comparable aux gradients projetés accélérés présentés dans les paragraphes suivants (Guan et al. (2012)). Le point faible de cette approche est la nécessité d'hypothèses de forte convexité pour garantir la convergence.

Gradient proximal et gradient projeté. Les algorithmes de gradient proximal permettent de résoudre des problèmes de minimisation de la forme :

$$\min_{\mathbf{X} \in \mathbb{R}^n} f(\mathbf{X}) + P(\mathbf{X}) \quad (4.38)$$

où

- $f(x)$ et $P(x)$ sont convexes, propres et semi continus inférieurement
- $P(x)$ est défini sur un convexe fermé de \mathbb{R}^n noté $domP$,
- $f(x)$ est défini sur un ensemble ouvert de \mathbb{R}^n contenant $domP$,
- f est C^1 sur $domP$.

L'algorithme s'écrit :

Algorithme 6 Gradient proximal

$X^{(0)} = X_{ini}$

while Convergence \neq True **do**

$$X^{(k+1)} = \text{Prox}_{\lambda^{(k)}P} [X^{(k)} - \lambda^{(k)} \nabla f(X^{(k)})] \quad (4.39)$$

end while

$x^* = X^{k+1}$

où $\text{Prox}_{\lambda^{(k)}P}$ désigne l'opérateur proximal de $\lambda^{(k)}P$ défini par :

$$\text{Prox}_{\lambda^{(k)}P}[v] = \underset{u}{\text{argmin}} \quad \lambda^{(k)}P(u) + \frac{1}{2}\|u - v\|_F^2, \quad (4.40)$$

et $\{\lambda^{(k)}\}_{k \in \mathbb{N}}$ est la suite des pas du gradient de valeur strictement positive. On montre que si ∇f est Lipschitzien de constante L la méthode converge linéairement pour un choix de pas fixe $\lambda^{(k)} = \lambda \in]0, 1/L]$.

Un des intérêts de cette méthode est qu'on ne fait pas beaucoup d'hypothèses de régularité sur P .

- P peut être non-différentiable, ce qui permet par exemple de traiter des problèmes de minimisation avec des pénalisations de type Lasso.
- P peut être à valeur dans $\bar{\mathbb{R}}^+$, en particulier elle peut être la fonction caractéristique d'un ensemble de contraintes. Dans ce cas l'opérateur $\text{Prox}_{\lambda^{(k)}P}$ est indépendant de $\lambda^{(k)}$ et est réduit à la projection euclidienne sur l'ensemble des contraintes; on parle alors d'algorithme de gradient projeté.

Cependant, elle nécessite le calcul de l'opérateur proximal ce qui restreint son intérêt aux problèmes où ce calcul n'est pas trop coûteux.

Dans le cas de la NMF, le gradient proximal s'applique aux sous-problèmes de minimisation alternée (4.27) et (4.28). Par exemple pour résoudre (4.27), on prend $f(\mathbf{W}) = E(\mathbf{W}, H, \alpha)$ et $P(\mathbf{W})$ est la fonction caractéristique des matrices à termes positifs, ce qui s'écrit :

$$P(\mathbf{W}) = \begin{cases} 0 & \text{if } \mathbf{W} \geq 0 \\ +\infty & \text{sinon.} \end{cases} \quad (4.41)$$

Dans ce cas, Prox_P se calcule très simplement puisqu'il s'agit de la projection euclidienne sur le cône positif :

$$\text{Prox}_P[W] = [W]_+, \text{ où } ([W]_+)_{i,j} = \begin{cases} W_{i,j} & \text{si } W_{i,j} \geq 0 \\ 0 & \text{sinon} \end{cases} \quad (4.42)$$

L'algorithme devient :

Algorithme 7 Gradient Projeté

```

 $W_{local}^{(0)} = W^{(t)}$                                      % Même notation que pour algorithme 1
while Convergence  $\neq$  True do                               %  $\nabla_W^E$ , le gradient en  $W$  de  $E$ 

 $W_{local}^{(\tau+1)} = \left[ W_{local}^{(\tau)} - \lambda^{(\tau)} \nabla_W^E(W_{local}^{(\tau)}) \right]_+$ 

end while
 $W^{(t+1)} = W_{local}^\infty$                                      %  $W_{local}^\infty$  est la valeur de  $W_{local}^{(\tau)}$  après convergence.

```

4.4.1.3 Gradient proximal accéléré

Des méthodes d'accélération de convergence des algorithmes à gradient proximal ont été proposées par Yurii Nesterov (voir Nesterov (1998) et Tseng (2008)). Sous les hypothèses du paragraphe précédent et en supposant que le gradient de la fonction f définie en (4.38) est Lipschitzien, il est possible de garantir une vitesse de convergence quadratique au lieu de linéaire. En gardant les mêmes notations que le paragraphe précédent, l'algorithme de gradient projeté accéléré s'écrit :

Algorithme 8 Gradient projeté accéléré

$$W_{local}^{(0)} = W^{(t)}$$

$$Z^{(0)} = W^{(t)}$$

$$\theta^{(0)} = 1$$

while Convergence \neq True **do**

$$Y^{(\tau)} = (1 - \theta^{(\tau)})W_{local}^{(\tau)} + \theta^{(\tau)}Z^{(\tau)}$$

$$Z^{(\tau+1)} = \left[Z^{(\tau)} - \frac{1}{L\theta^{(k)}} \nabla_W^E(Y^{(\tau)}) \right]_+$$

$$W_{local}^{(\tau+1)} = (1 - \theta^{(\tau)})W_{local}^{(\tau)} + \theta^{(\tau)}Z^{(\tau+1)}$$

$$\theta^{(\tau+1)} = \frac{1}{2}\theta^{(\tau)}(\sqrt{(\theta^{(\tau)})^2 + 4} - \theta^{(\tau)})$$

end while

$$W^{(t+1)} = W_{local}^{(\infty)}$$

Notons que dans le cas de la divergence KLG, le gradient n'est pas Lipschitzien ce qui fait perdre les garanties théoriques de vitesse de convergence et amène à utiliser d'autres méthodes.

4.4.2 Implémentation

Le processus de résolution du problème (4.13) est décrit dans l'Algorithme 9. Nous commençons par normaliser la matrice de données A , normalisation ensuite répercutée sur la matrice de contraintes. N'ayant aucune garantie d'atteindre le minimum global, nous prenons le meilleur couple solution $(W^\circ, \tilde{H}^\circ)$ parmi les 5 résultats de 5 NMF initialisées aléatoirement. Les matrices d'initialisations aléatoires sont telles que la norme de leur produit est égale à celle de \tilde{A} .

Algorithme 9 NMF

$$\tilde{A} = AN^{-1}$$

$$\tilde{F}_\Delta = F_\Delta N^T$$

for Iter= 1 to 5 **do**

$$(W_{0,\text{Iter}}, \tilde{H}_{0,\text{Iter}}) \sim \mathcal{U}(0, 1)$$

$$W_{\text{ini},\text{Iter}} = W_{0,\text{Iter}} \sqrt{\frac{\|\tilde{A}\|_F}{\|W_{0,\text{Iter}} \tilde{H}_{0,\text{Iter}}\|_F}} \quad \% W_{\text{ini},\text{Iter}} \text{ est } W^{(0)} \text{ dans l'Algorithme 2}$$

$$\tilde{H}_{\text{ini},\text{Iter}} = H_{0,\text{Iter}} \sqrt{\frac{\|\tilde{A}\|_F}{\|W_{0,\text{Iter}} \tilde{H}_{0,\text{Iter}}\|_F}} \quad \% \tilde{H}_{\text{ini},\text{Iter}} \text{ est } \tilde{H}^{(0)} \text{ dans l'Algorithme 2}$$

$$(W_{\text{Iter}}^\circ, \tilde{H}_{\text{Iter}}^\circ) = \underset{(\mathbf{W}, \tilde{\mathbf{H}}) \in \tilde{\mathcal{C}}}{\operatorname{argmin}} \|\tilde{A} - \mathbf{W} \tilde{\mathbf{H}}\|_F^2 + \alpha(\|\mathbf{W}\|_F^2 + \|\mathbf{1}^T \tilde{\mathbf{H}}\|_2^2)$$

end for

$$(W^\circ, \tilde{H}^\circ) = \underset{\text{Iter} \in [1,5]}{\operatorname{argmin}} \|\tilde{A} - W_{\text{Iter}}^\circ \tilde{H}_{\text{Iter}}^\circ\|_F^2$$

$$H^\circ = \tilde{H}^\circ N$$

À chaque itération nous utilisons la minimisation alternée de l'Algorithme 2 afin de résoudre le problème de NMF contraint. Il nécessite l'optimisation des sous-problèmes (4.32) et (4.33) que nous présentons ci-après. Dans les deux cas, nous avons choisi d'exploiter le gradient proximal accéléré proposé par Nesterov. La présence d'une contrainte supplémentaire nécessite toutefois une adaptation dans le cas de la minimisation de H . Nous rappelons que l'ensemble de minimisation $\tilde{\mathcal{C}}$ s'écrit

$$\begin{aligned} \tilde{\mathcal{C}} &= \mathcal{M}_{n,k}(\mathbb{R}^+) \times \{\tilde{H} \in \mathcal{M}_{k,r}(\mathbb{R}^+) \mid \tilde{F}_\Delta \tilde{H}^T \leq 0\} \\ &= \mathcal{C}_W \times \tilde{\mathcal{C}}_H \end{aligned}$$

Optimisation en W : Le problème de minimisation (4.32) est le suivant :

$$W^{(t+1)} = \underset{\mathbf{W} \in \mathcal{C}_W}{\operatorname{argmin}} \|\tilde{A} - \mathbf{W} \tilde{H}^{(t)}\|_F^2 + \alpha(\|\mathbf{W}\|_F^2 + \|\mathbf{1}^T \tilde{H}\|_2^2)$$

La résolution se fait par l'algorithme de gradient proximal accéléré 8 en définissant la fonction à minimiser comme la somme de la fonction caractéristique du cône positif et de la fonction de coût $E(W, \tilde{H}, \alpha)$ pour \tilde{H} fixé. L'opérateur proximal associé à la fonction caractéristique du cône positif est la projection euclidienne sur le cône positif. L'algorithme nécessite le calcul

du gradient $E(W, H, \alpha)$ selon W de :

$$\nabla_W^E(W) = 2W(\tilde{H}\tilde{H}^T + \alpha\text{Id}_k) - 2\tilde{A}\tilde{H}^T, \quad (4.43)$$

où $\text{Id}_k \in \mathcal{M}_{k,k}(\mathbb{R}^+)$ est la matrice identité. On montre facilement que L , la constante de Lipschitz de $\nabla_W^E(W)$, est égale à la plus grande valeur propre de la matrice $2(\tilde{H}\tilde{H}^T + \alpha\text{Id}_k)$.

Optimisation en \tilde{H} : Le problème de minimisation (4.32) est le suivant :

$$\tilde{H}^{(t+1)} = \underset{\tilde{\mathbf{H}} \in \tilde{\mathcal{C}}_H}{\text{argmin}} \quad \|\tilde{A} - W^{(t+1)}\tilde{\mathbf{H}}\|_F^2 + \alpha(\|W\|_F^2 + \|\mathbf{1}^T \tilde{\mathbf{H}}\|_2^2)$$

Si l'on souhaite utiliser l'approche du gradient proximal, il faut pouvoir calculer facilement son expression or il n'existe pas de formule analytique permettant de calculer la projection sur l'ensemble $\tilde{\mathcal{C}}$, nous avons donc choisi de procéder différemment. Nous utilisons une approche classique de Lagrangien augmenté (Arrow et al. (1958)), la méthode d'Uzawa, que nous combinons avec une étape de gradient proximal accéléré.

Tout d'abord, nous écrivons le lagrangien associé au problème de minimisation dans le cône positif :

$$\mathcal{L}(\tilde{H}, U_{\tilde{F}}) = E(W^{(t+1)}, \tilde{\mathbf{H}}, \alpha) + \text{tr}(\mathbf{U}_{\tilde{\mathbf{F}}}^T \tilde{F}_\Delta \tilde{\mathbf{H}}^T),$$

avec $\text{tr}(\cdot)$ l'opérateur trace et $U_{\tilde{F}} \in \mathcal{M}_{c,k}(\mathbb{R}^+)$ la matrice de multiplicateurs de Lagrange. Alors, l'Algorithme 10 permet d'obtenir la solution du problème.

Algorithme 10 Itération d'Uzawa

$$\tilde{H}_{local}^{(0)} = \tilde{H}^{(t)}$$

while Convergence \neq True **do**

$$\tilde{H}_{local}^{(\tau+1)} = \underset{\tilde{\mathbf{H}} \geq 0}{\text{argmin}} \quad \mathcal{L}(\tilde{\mathbf{H}}, U_{\tilde{F}}^{(\tau)}) \quad (4.44)$$

$$U_{\tilde{F}}^{(\tau+1)} = \left[U_{\tilde{F}}^{(\tau)} + \frac{2a}{C^2} \tilde{F}_\Delta \tilde{H}_{local}^{(\tau+1)T} \right]_+.$$

end while

$$H^{(t+1)} = H_{local}^{(\infty)}$$

Le paramètre a est la constante de forte convexité du Lagrangien en \tilde{H} . Il s'obtient en calculant la plus petite valeur propre de la matrice $W^T W + \alpha \mathbf{1} \mathbf{1}^T$. Le paramètre C est la plus grande valeur propre de la matrice de contraintes \tilde{F}_Δ .

Le sous-problème de minimisation (4.44) se résout en utilisant le gradient proximal accéléré. Si l'on transpose l'Algorithme 8 au cas de (4.44), le calcul du gradient de $\mathcal{L}(\tilde{H}, U_{\tilde{F}})$ est nécessaire, il s'écrit :

$$\nabla_{\tilde{H}}^{\mathcal{L}}(H) = 2(W^T W + \alpha \mathbf{1}\mathbf{1}^T)H - 2W^T \tilde{A} + U_{\tilde{F}}^T \tilde{F}_{\Delta} \quad (4.45)$$

et la constante de Lipschitz associée s'obtient en prenant la plus grande valeur propre de $(2W^T W + \alpha \mathbf{1}\mathbf{1}^T)$.

4.5 Détermination des paramètres du problème de NMF

Le problème de NMF (4.29) dépend de deux paramètres : k le nombre de CTFA et α la constante de pénalisation. Le choix de k est une problématique bien décrite dans la littérature, ce qui n'est pas le cas du choix des constantes de pénalisation. Il est cependant possible d'appliquer les approches de détermination de k au choix de α . Nous présentons dans la Section 4.5.1 les approches développées dans la littérature avant de présenter celles que nous avons choisies dans la Section 4.5.2.

4.5.1 Méthodes de sélection de la dimension k dans la littérature

4.5.1.1 Stabilité numérique

Les algorithmes de NMF présents dans la littérature offrent des garanties de décroissance de la fonction de coût à minimiser ou de convergence vers un point stationnaire. En l'absence de garanties de convergence de l'Algorithme 9 vers un de minimum local ou global, des répétitions utilisant des initialisations aléatoires différentes sont implémentées. Cette procédure est également utilisée pour choisir la dimension k optimale : la valeur de k conduisant à la plus grande stabilité vis-à-vis de l'initialisation aléatoire est sélectionnée. Ce critère est le plus couramment utilisé dans les applications aux données omiques.

Une approche alternative basée sur un consensus entre les matrices W obtenues à partir d'initialisations différentes est également considérée. Deux types de critères existent, le premier porte sur le clustering des échantillons induit par la matrice W , le second quantifie le consensus en un sens plus général.

Consensus basé sur le clustering induit par W . Dans de nombreuses applications en biologie, la NMF est considérée comme un outil de bi-clustering (des échantillons biologiques

et des variables) avec une attention particulière portée au clustering des échantillons. Pour chaque valeur de k , chaque échantillon i est affecté à son profil majoritaire défini comme $\arg \max_{j=1,\dots,k} W_{i,j}$. Un indice de consensus est construit, qui mesure la similarité entre les clusters d'échantillons provenant d'initialisations différentes. La matrice des indices de consensus $C^{(k,s)} \in \mathcal{M}_{n,n}(\mathbb{R}^+)$ est définie par $C_{ij}^{(k,s)} = 1$ si les échantillons i et j sont classifiés dans le même cluster lors de la répétition s pour un nombre k de clusters et $C_{ij}^{(k,s)} = 0$ sinon. La moyenne $\bar{C}^{(k)}$ est calculée sur les répétitions et le critère final s'écrit :

$$\rho_k = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n 4(\bar{C}^{(k)}(i,j) - \frac{1}{2})^2$$

Lorsque $\rho_k = 1$, les assignations des échantillons aux clusters sont parfaitement cohérentes entre répétitions. Cette méthode est présentée notamment dans Brunet et al. (2004) et Kim and Park (2007).

Consensus mesuré par concordance. Jiang et al. (2012b) proposent une mesure plus générale du consensus qui repose sur la composition en profils de chacun des échantillons. Celle-ci est donnée par les lignes de la matrice W . Cette approche ne suppose pas une structure de clusters sous-jacente dans les données. Plus précisément, pour chaque répétition s , on calcule la matrice symétrique

$$S_{jj'}^{(s)} = \frac{\sum_{l=1}^m W_{lj}^{(s)} W_{lj'}^{(s)}}{\left(\sum_{l=1}^m (W_{lj}^{(s)})^2\right)^{1/2} \left(\sum_{l=1}^m (W_{lj'}^{(s)})^2\right)^{1/2}} \quad (4.46)$$

dont le terme général est compris entre 0 et 1. L'indice de concordance est alors défini par

$$Conc(k) = 1 - \frac{1}{\binom{s_{max}}{2}} \frac{1}{m(m-1)} \sum_{(s \neq s') \in \{1, \dots, s_{max}\}^2} \sum_{i \neq j} \left(S_{i,j}^{(s)} - S_{i,j}^{(s')}\right)^2 \quad (4.47)$$

Il est par construction compris entre 0 et 1 et vaut 1 si les répétitions sont identiques. On le calcule pour chaque valeur de k , et la dimension conduisant à la meilleure concordance est sélectionnée.

4.5.1.2 Validation Croisée adaptée à la NMF : Bi-Validation-Croisée

Owen and Perry (2009) ont proposé une procédure de validation croisée adaptée à la NMF, la Bi-Validation-Croisée (BiCV), dans laquelle les lignes et les colonnes de la matrice de données sont découpées en ensembles d'apprentissage et de validation. Les indices des échantillons biologiques ainsi que les indices des descripteurs sont partitionnés en d sous-groupes de taille égale notés respectivement $\{I_1, \dots, I_d\}$ et $\{J_1, \dots, J_d\}$. On définit pour chaque couple $(l, l') \in \{1, \dots, d\}^2$, 4 sous-matrices de \tilde{A} : $\tilde{A}_{11}^{(l, l')} \in \mathcal{M}_{n^l, r^{l'}}(\mathbb{R}^+)$, $\tilde{A}_{21}^{(l, l')} \in \mathcal{M}_{n-n^l, r^{l'}}(\mathbb{R}^+)$, $\tilde{A}_{12}^{(l, l')} \in \mathcal{M}_{n^l, r-r^{l'}}(\mathbb{R}^+)$ et $\tilde{A}_{22}^{(l, l')} \in \mathcal{M}_{n-n^l, r-r^{l'}}(\mathbb{R}^+)$ où n^l et $r^{l'}$ sont les cardinaux respectifs des sous-groupes I_l et $J_{l'}$.

$$\left[\begin{array}{c|c} \tilde{A}_{11}^{(l, l')} & \tilde{A}_{12}^{(l, l')} \\ \hline \tilde{A}_{21}^{(l, l')} & \tilde{A}_{22}^{(l, l')} \end{array} \right]$$

De même à partir de la matrice de normalisation N définie en (4.7), on construit $N^{l'}$ la sous-matrice de N contenant toutes les colonnes dont les indices appartiennent à $J_{l'}$.

L'erreur de BiCV $E_{BiCV}^{(l, l')}$ est calculée pour tous les couples (l, l') selon l'Algorithme 11.

Algorithme 11 Bi-Validation-Croisée

```

for  $(l, l') \in [1, \dots, d]^2$  do
   $(W_2^{(l, l')}, \tilde{H}_2^{(l, l')}) = \operatorname{argmin}_{(\mathbf{W}, \mathbf{H}) \in \mathcal{C}_0} \|\tilde{A}_{22}^{(l, l')} \mathbf{W} \tilde{\mathbf{H}}\|_F^2 + \alpha (\|\mathbf{W}\|_F^2 + \|\mathbf{1}^T \tilde{\mathbf{H}}\|_2^2)$ 
   $(W_1^{(l, l')}) = \operatorname{argmin}_{\mathbf{W} \in \mathcal{M}_{n, k}(\mathbb{R}^+)} \|\tilde{A}_{12}^{(l, l')} \mathbf{W} \tilde{H}_2^{(l, l')}\|_F^2 + \alpha \|\mathbf{W}\|_F^2$ 
   $(\tilde{H}_1^{(l, l')}) = \operatorname{argmin}_{\mathbf{H} \in \mathcal{M}_{k, r}(\mathbb{R}^+)} \|\tilde{A}_{21}^{(l, l')} W_2^{(l, l')} \tilde{\mathbf{H}}\|_F^2 + \alpha \|\mathbf{1}^T \tilde{\mathbf{H}}\|_2^2$ 
   $A_{11}^{(l, l')} = \tilde{A}_{11}^{(l, l')} N^{l'}$ 
   $H_1^{(l, l')} = \tilde{H}_1^{(l, l')} N^{l'}$ 
   $E_{BiCV}^{(l, l')} = \|A_{11}^{(l, l')} - W_1^{(l, l')} H_1^{(l, l')}\|_F$ 

```

end for

L'erreur de BiCV finale E_{BiCV} s'obtient en prenant la médiane des $E_{BiCV}^{(l, l')}$:

$$E_{BiCV} = \operatorname{mediane}_{(l, l') \in \{1, \dots, d\}^2} E_{BiCV}^{(l, l')}.$$

Cette procédure ne possède pas à notre connaissance de justification théorique.

4.5.1.3 Autres méthodes

L'erreur de reconstruction $\mathcal{D}(A|WH)$ décroît automatiquement quand le nombre de profils k augmente et la vitesse de décroissance est plus forte lorsque l'ajout d'une composante est significatif. Ainsi, Kim and Tidor (2003) ont proposés une procédure dans laquelle la vitesse de décroissance est comparée à celle obtenue sur une matrice générée aléatoirement selon une distribution gaussienne de mêmes moyennes et variances.

D'autres approches basées sur le nombre de valeurs propres significatives de la matrice \tilde{A} ont également été proposées dans la littérature (Fogel et al. (2007)).

4.5.2 Sélections de α et k

Nous recherchons des valeurs de k et α qui offrent un consensus entre la qualité de reconstruction des abondances de la matrice A et la reproductibilité biologique.

4.5.2.1 Critères choisis

(1) Erreur de reconstruction.

Pour un couple (α, k) , l'erreur de reconstruction s'écrit

$$E^{rec}(\alpha, k) = \|A - W^\circ H^\circ\|_F / \|A\|_F$$

où (W°, H°) est obtenu par la procédure de NMF décrite dans l'Algorithme 9.

L'erreur de reconstruction décroît automatiquement lorsque k augmente et peut atteindre 0 lorsque $k = r$, puisque dans ce cas $W = A$ et $H = \text{Id}_r$ est une solution triviale. De la même façon l'erreur de reconstruction est minimale pour $\alpha = 0$. Cet indicateur permet d'évaluer la qualité de la reconstruction des données. Les valeurs de paramètres optimales ne sont donc pas celles qui minimisent l'erreur de reconstruction. Nous étudions la diminution de l'erreur de reconstruction en fonction de k et α et choisissons les valeurs à partir desquelles le critère ne diminuera plus significativement. Cela se visualise en étudiant les cassures de pentes.

(2) Bi validation-croisée.

La BiCV est un critère qui doit permettre d'évaluer la capacité de prédiction de données non observées de la décomposition par NMF évaluée sur des données d'entraînement, pour des

paramètres (α, k) fixés. Ce critère s’applique à la NMF non contrainte. En effet, la contrainte $\tilde{F}_\Delta \tilde{H}^T \leq 0$ s’applique à l’ensemble des marqueurs fonctionnels, il n’est pas possible de formuler des contraintes de ce type pour des sous-ensembles de marqueurs.

Nous utilisons donc la procédure décrite plus haut dans l’Algorithme 11 mais résolvons le problème de NMF non contraint. L’ensemble des échantillons biologiques ainsi que l’ensemble des marqueurs fonctionnels sont découpés en 10 parties.

(3) Concordance des CTFA.

A la différence de Jiang et al. (2012b) qui utilisent le critère de concordance (4.47) afin de quantifier la robustesse numérique de la NMF vis-à-vis des initialisations, nous proposons d’utiliser la concordance pour évaluer la reproductibilité biologique des CTFA. Pour cela nous évaluons la concordance entre les matrices de CTFA obtenues en partitionnant aléatoirement la matrice A selon les lignes en deux sous-matrices de même taille.

Plus précisément, pour chaque répétition $s \in \{1, \dots, 40\}$, la matrice d’abondance A est découpée aléatoirement en deux sous matrices $A_{1,s}$ et $A_{2,s}$ contenant chacune la moitié des 1408 échantillons. Nous appliquons l’Algorithme 9 et obtenons deux couples solutions $(W_{1,s}^\circ, H_{1,s}^\circ)$ et $(W_{2,s}^\circ, H_{2,s}^\circ)$. La concordance est évaluée à partir de ces 40 répétitions appliquée aux couples $(H_{1,s}^\circ, H_{2,s}^\circ)$ selon

$$\text{Conc}(\alpha, k) = 1 - \frac{1}{40 \times 86} \sum_{s=1}^{40} \|S_1^{(s)}(\alpha, k) - S_2^{(s)}(\alpha, k)\|_F.$$

4.5.2.2 Démarche globale

Nous évaluons les 3 critères pour des couples paramètres (α, k) appartenant à l’ensemble $\mathcal{G}_\alpha \times \mathcal{G}_k$, où $\mathcal{G}_\alpha = (10^{-3}, 10^{-2}, 10^{-1.5}, 10^{-1})$ et $\mathcal{G}_k = (2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15)$. La valeur maximale de α est choisie par calculs numériques préalables de sorte que le rang de W° et \tilde{H}° soit toujours égal k ; bien que dépendante de la valeur de k , cette valeur maximale est toujours proche de 10^{-1} . La valeur minimale de α notée α_{min} est prise telle que $|E^{rec}(\alpha_{min}, k) - E^{rec}(0, k)| > 10^{-4}$. Concernant la valeur maximale de k , nous considérons le ratio du nombre de données sur le nombre de paramètres $nr/(nk + kr)$ et choisissons d’explorer les valeurs de k pour lesquelles nous avons au minimum 5 fois plus de données que de paramètres.

Pour des raisons de temps de calculs, nous ne testons pas tous les couples (α, k) mais commençons par déterminer une valeur de α pour $k \in \{2, 4, 6, 8, 11, 13\}$. Dans le cas de α , l’évaluation

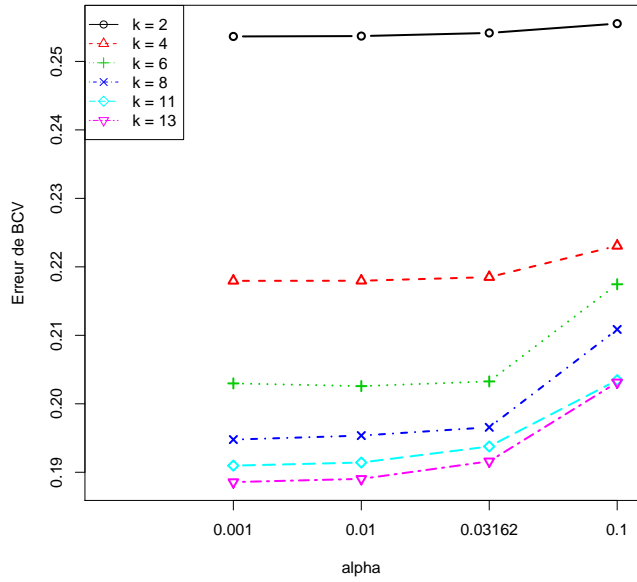
des 3 critères se fait en utilisant la même initialisation aléatoire pour chaque valeur de α . Une fois α fixé, nous évaluons les critères sur les valeurs de k non testées.

4.5.2.3 Paramètres sélectionnés

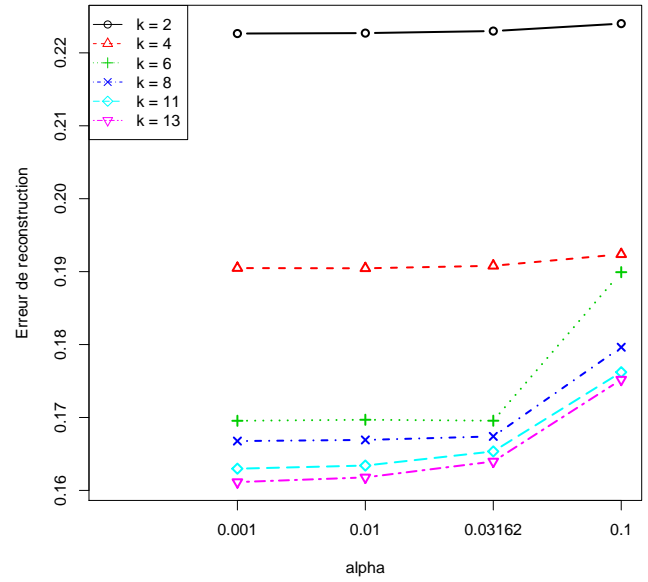
La maximisation de la concordance et la minimisation des erreurs de reconstruction et de BiCV sont chacune obtenues pour des valeurs de α et k non envisageables telles que $k = 86$ où $\alpha = 0$. De ce fait, la détermination de ces paramètres se fait par l'étude des pentes des critères : pour chaque critère, on cherche les valeurs de α et k à partir desquelles l'amélioration du critère devient négligeable.

Dans un deuxième temps et puisque les valeurs optimales de paramètres pour chaque critère ne sont forcément identiques, nous retenons un couple (α, k) issu d'un compromis entre les trois critères.

Sélection de α . Les Figures 4.3 et 4.4 représentent l'évolution des trois critères considérés en fonction de α pour différentes valeurs de k . On observe que les trois critères sont globalement constants pour $\alpha \leq 10^{-1.5} = 0.0316$ et se détériorent de façon significative pour $\alpha > 10^{-1.5}$ pour tout k , à l'exception de la concordance pour $k = 6$ où le critère diminue de façon constante lorsque α augmente. Cependant la distribution de la concordance pour chaque partitionnement aléatoire indique que les variations de la moyenne ne sont pas fortement significatives. De ce fait nous avons choisis $\alpha = 10^{-1.5}$.



(a) Erreur de BiCV en fonction de α pour différentes valeurs de k



(b) Erreur de reconstruction de A en fonction de α pour différentes valeurs de k

FIGURE 4.3: Erreur de reconstruction et de BiCV en fonction de α pour différentes valeurs de k

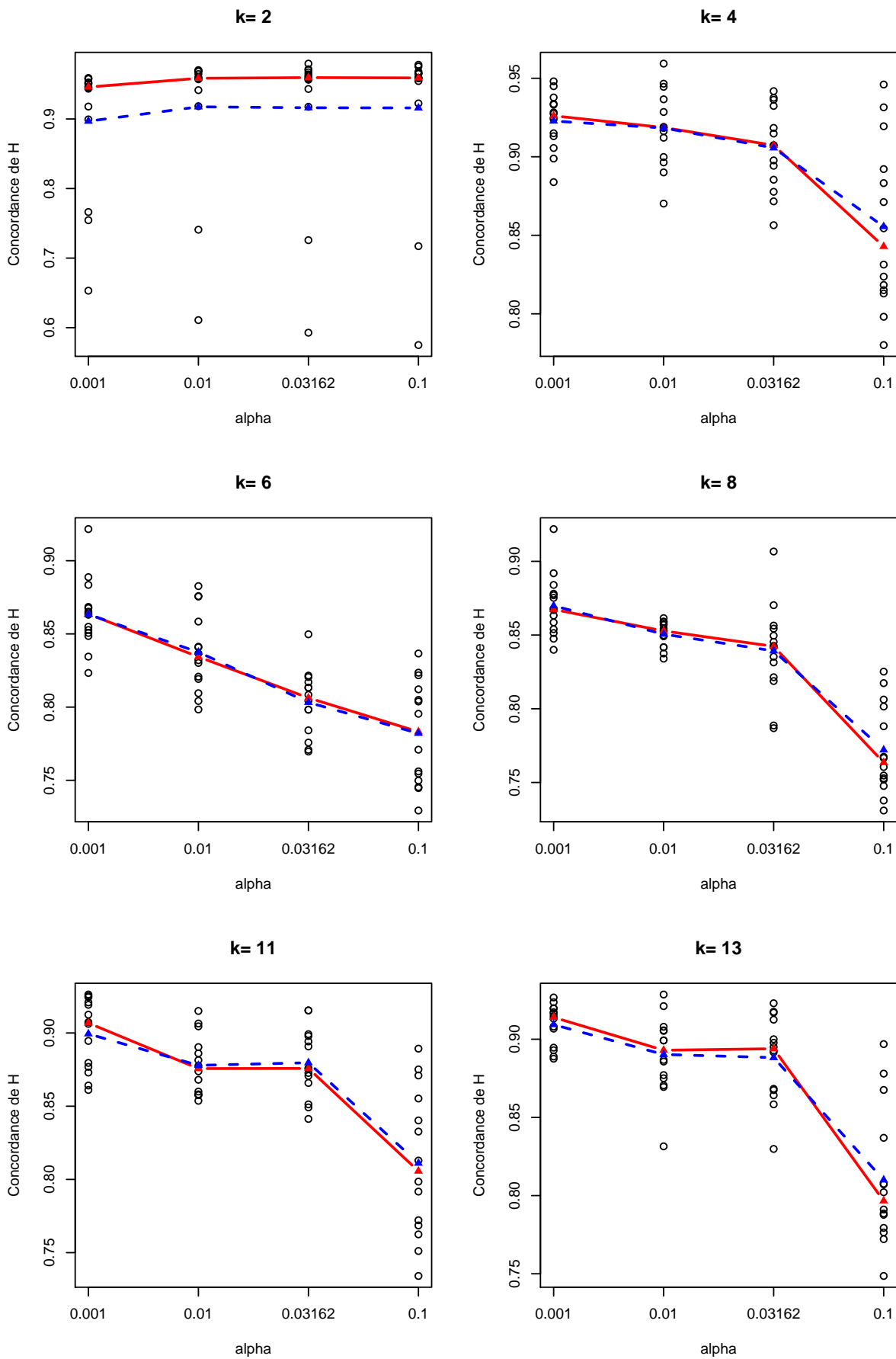
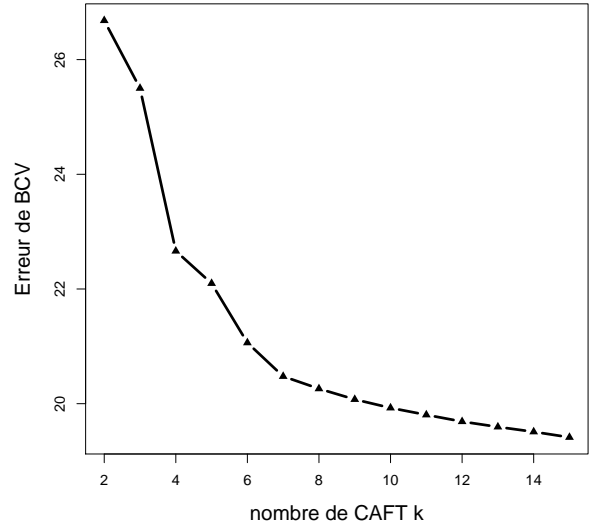
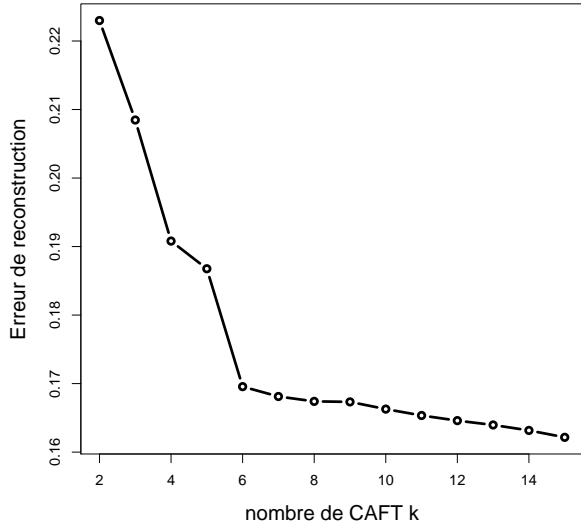


FIGURE 4.4: Concordance des CTFA pour chaque partitionnement aléatoire (cercles noirs), médiane (lignes bleues) et moyenne (ligne rouge) en fonction de α , pour différentes valeurs de k

Sélection de k . Les courbes présentées en Figure 4.5.2.3 représentent l'évolution des trois critères considérés en fonction de k . Nous observons qu'ils présentent des tendances différentes : les erreurs de reconstruction et de BiCV diminuent quand k augmente tandis que la concordance est maximale pour k faible.

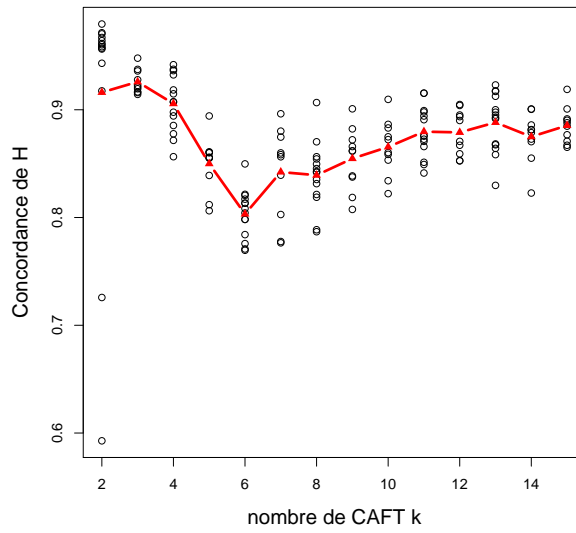
Les courbes d'erreur de reconstruction et d'erreur de BiCV possède deux ruptures de pentes respectivement pour $k = 4$ et 6 , et pour $k = 4$ et 7 . Enfin, la concordance assure une bonne reproductibilité des CTFA pour $k \leq 4$.

Dans le contexte de notre modèle l'objectif principal est la détermination de CTFA qui caractérisent des assemblages fonctionnels présent dans tous les échantillons. Dans ce contexte, nous avons décidé de favoriser le critère de concordance qui quantifie stabilité des CTFA entre les échantillons, d'autre part la dégradation des erreurs de reconstruction et de BiCV pour $k = 4$ reste modérée c'est pourquoi nous avons choisi la valeur $k = 4$.



(a) Erreur de reconstruction en fonction de k pour $\alpha = 10^{-1.5}$

(b) Erreur de BiCV en fonction de k pour $\alpha = 10^{-1.5}$



(c) Concordance pour chaque partitionnement aléatoire (cercles noirs) et moyenne (ligne rouge) en fonction de k pour $\alpha = 10^{-1.5}$

Chapitre 5

Résultats de l'inférence : analyse des matrices W et H

La Factorisation en Matrices Non-négatives permet d’approcher une matrice d’abondance A par le produit de deux matrices positives de rang plus faible. Nous étudions dans cette partie les propriétés du couple de matrices (W^*, H^*) obtenues dans le chapitre précédent en prenant $k = 4$ et $\alpha = 10^{-1.5}$; nous le noterons (W, H) dans ce chapitre. Nous examinons dans un premier temps quelques propriétés de cette réduction de dimension en faisant le parallèle avec une approche classique de réduction de données, la Décomposition en Valeurs Singulières (Singular Values Decomposition, SVD), avant de nous intéresser aux abondances des CTFA dans les 1408 échantillons en liens avec leurs métadonnées et avant de conclure sur les particularités et propriétés des CTFA.

5.1 La NMF, une approche de réduction de dimension

Cette section propose l’interprétation géométrique des résultats ainsi que le calcul d’indicateurs de qualité de la réduction des données en établissant lorsque c’est nécessaire un parallèle avec la SVD.

5.1.1 Interprétation géométrique des résultats

En traitement des données, la réduction de dimension est un ensemble de techniques permettant d’approcher des données décrites dans un espace de grande dimension par une approximation appartenant à un espace de dimension plus faible. Ces traitements peuvent être utilisés à des fins de modélisation, d’exploration (recherche de structure) ou de simplification permettant d’agréger une information complexe.

L’approche de NMF peut être décrite comme la recherche d’une région d’un sous-espace de dimension k contenant des approximations de l’ensemble des données. Plus précisément, comme illustré par la Figure 5.1, l’approximation $A \sim WH$ où H possède k lignes indépendantes peut s’interpréter comme la recherche d’un cône convexe de dimension k , engendré par les vecteurs lignes de H contenus dans l’orthant positif \mathbb{R}_+^r . Par définition, tout point de ce cône convexe s’exprime de façon unique comme une combinaison linéaire à termes positifs des vecteurs lignes de H . Les vecteurs lignes de W correspondent donc aux coordonnées des points du cône convexe approchant les vecteurs lignes de A .

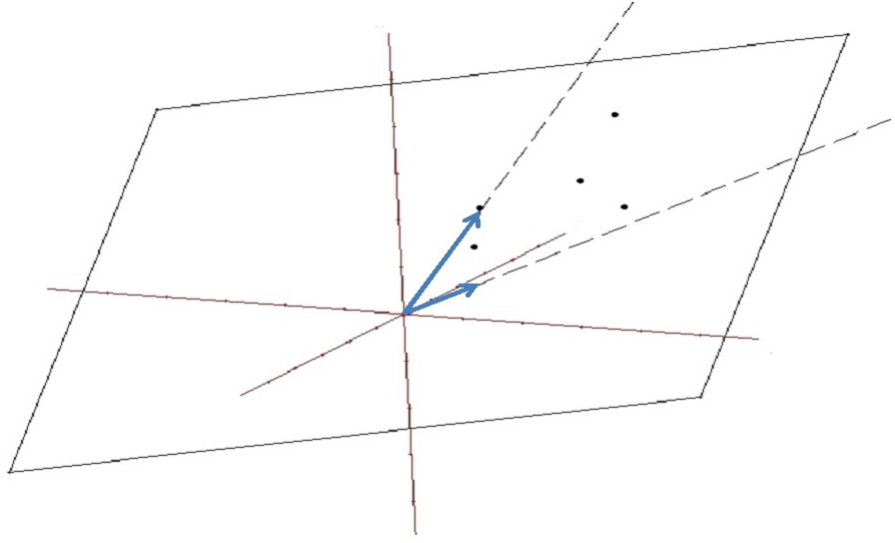


FIGURE 5.1: Interprétation géométrique des matrices W et H , dans cet exemple les vecteurs lignes contenus dans H (trait pleins bleus) génèrent un sous-espace de dimensions 2, dans lequel ils engendrent un cône convexe (traits pointillés). Les points approchant au mieux (selon une mesure de dissimilarité) les données initiales se trouvent à l'intérieur de ce cône (points noirs). Les vecteurs lignes de W sont les coordonnées de ces points dans le cône.

Dans ce travail de thèse, nous imposons une contrainte supplémentaire sur la recherche des vecteurs lignes de H qui doivent appartenir au cône $\mathbb{R}_+^r \cap \{v \in \mathbb{R}_+^r \mid F_{\Delta} v^T \leq 0\}$.

Normalisation conjointe. Nous normalisons les vecteurs ligne de H (CTFA) en norme L_1 . Les poids W sont donc modifiés en conséquence, de façon à conserver le produit WH .

$$\begin{aligned} \bar{H}_{lj} &= \frac{H_{lj}}{\sum_{p=1}^r H_{lp}} \\ \bar{W}_{il} &= W_{il} \times \sum_{p=1}^r H_{lp} \end{aligned} \quad (5.1)$$

Cette normalisation permet d'interpréter la composition des CTFA comme des fréquences de marqueurs fonctionnels dans des sous-populations. D'autre part, elle permet la comparaison poids associés aux différents CTFA.

Quantification de l'information. Il s'agit ici de déterminer une mesure pertinente de la quantité d'information reconstruite par chacun des vecteurs ligne de H . Nous allons tout d'abord

rappeler comment cette mesure se fait dans le cadre de la SVD, qui est une des méthodes les plus couramment utilisées pour l'analyse de données (Antoulas, 2005). La SVD repose sur la détermination d'une matrice $U \in \mathcal{M}_{n,n}(\mathbb{R})$ dont les colonnes U_i forment une base orthonormée de vecteurs propres de AA^T , et une matrice $V \in \mathcal{M}_{r,r}(\mathbb{R})$ dont les colonnes V_j forment une base orthonormée de vecteur propres de $A^T A$. On introduit également la matrice $\Sigma \in \mathcal{M}_{n,r}(\mathbb{R})$ de terme $\Sigma_{ii} = \sigma_i$ pour $i \in [0, \min(n,r)]$, où σ_i est la $i^{\text{ème}}$ parmi les valeurs singulières de A (valeurs propres de $A^T A$) ordonnées par ordre décroissant. La SVD de la matrice A s'écrit alors

$$A = U\Sigma V^T = \sum_{i=1}^{\min(n,r)} U_i \sigma_i V_i^T. \quad (5.2)$$

En posant $P = U\Sigma$ et $Q = V^T$, on retrouve une factorisation matricielle $A = PQ$ de même forme que la NMF, mais sans contrainte de positivité. On a alors

$$\|A\|_F = \left(\sum_{i=1}^{\min(n,r)} \sigma_i^2 \right)^{1/2}. \quad (5.3)$$

La contribution du vecteur ligne $Q_i = V_i^T$ dans la reconstruction de A est alors égale à

$$\frac{\|U_i \sigma_i V_i^T\|_F}{\|A\|_F} = \frac{\sigma_i}{\left(\sum_{i=1}^{\min(n,r)} \sigma_i^2 \right)^{1/2}} \quad (5.4)$$

Il serait tentant d'utiliser un formalisme similaire dans le cas de la NMF, puisqu'on a encore une égalité de type (5.2) sous la forme de

$$A \sim WH = \sum_i W_{(:,i)} H_{(i,:)}, \quad (5.5)$$

en introduisant des contributions des CTFA de la forme $\|W_{(:,i)} H_{(i,:)}\|_F / \|WH\|_F$.

Cependant et par construction, les CTFA c'est à dire les lignes de la matrice H ne sont pas orthogonaux entre eux comme indiqué par la Figure 5.2. De ce fait, le carré de la norme de Frobenius de WH n'est pas égal à la somme des carrés des normes de Frobenius des $W_{(:,i)} H_{(i,:)}$.

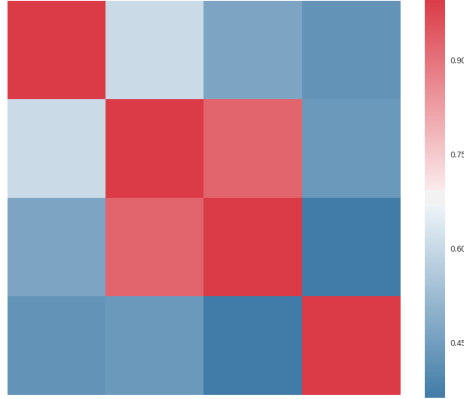


FIGURE 5.2: Cosinus des angles entre les CTFA de la matrice H_4 , le cosinus se calcule comme le produit scalaire des CTFA normés en norme 2 et permet de quantifier l'orthogonalité : si les vecteurs lignes de H sont orthogonaux entre eux, la matrice doit être la matrice identité.

En effet, sous contrainte de positivité, l'orthogonalité impliquerait que chaque CTFA de H ne partage aucun marqueur fonctionnel avec aucun autre. Or un aspect intéressant de l'approche de NMF est précisément la possibilité de pouvoir expliquer l'abondance d'un marqueur fonctionnel comme provenant des contributions de différents CTFA. A titre d'exemple, le métabolisme central des sucres simples nécessite la présence de certains marqueurs fonctionnels dans chaque CTFA comme par exemple les marqueurs associés au passage du Glyceraldehyde-3P au pyruvate.

Nous proposons une alternative en considérant les quantités

$$\frac{\|W_{(:,i)}H_{(i,:)}\|_1}{\|WH\|_1} = \frac{\sum_{i=1}^n \sum_{j=1}^r W_{i,l}H_{l,j}}{\sum_{i=1}^n \sum_{j=1}^r \sum_{l=1}^k W_{i,l}H_{l,j}} = \frac{\sum_{i=1}^n \bar{W}_{i,l}}{\sum_{i=1}^n \sum_{l=1}^k \bar{W}_{i,l}}. \quad (5.6)$$

Ceci revient à quantifier la contribution de chaque CTFA en calculant le pourcentage de l'abondance totale reconstruite par chaque CTFA.

Comparaison des cônes approximant pour différentes valeurs de k . Si nous considérons encore le cadre classique de la SVD, le problème de la réduction de dimension qui consiste à approcher la matrice de données A par une matrice $A_{(k)}$ de rang k inférieur à celui de A est résolu par le théorème de Schmidt-Eckart-Young-Mirsky (voir remarque 3.2.2 de Antoulas

(2005)). Il montre qu'il existe une approximation de rang k de A , optimale au sens de la norme de Frobenius donnée par la troncature à l'ordre k de la SVD donnée par

$$A_{(k)} = \sum_{p=1}^k U_p \sigma_p V_p^T = P_{(k)} Q_{(k)}. \quad (5.7)$$

Cette approximation est unique si $\sigma_k < \sigma_{k+1}$. On voit donc que dans ce cadre, lorsque la valeur de k augmente, Q_{k+1} est construit en ajoutant une ligne à Q_k ce qui implique que les espaces engendrés par les Q_k sont emboîtés. On montre également que l'erreur de reconstruction de A est nécessairement décroissante.

La NMF ne garantit aucune relation entre le cône convexe engendré par H^k et celui engendré par H^{k+1} . Par analogie avec la SVD, on pourrait souhaiter que H^k soit contenu dans le cône convexe engendré par H^{k+1} , mais ceci ne peut être garanti.

Nous illustrons ceci en comparant les matrices H^4 et H^{15} résultant de la NMF pour $k = 4$ et $k = 15$ respectivement. Pour évaluer le recouvrement entre les deux cônes associés, nous recherchons par moindres carrés positifs (Non-Negative Least Square, NNLS) une matrice $\tau \in \mathcal{M}_{4,15}(\mathbb{R}^+)$ solution du problème suivant

$$\tau = \underset{\tau \geq 0}{\operatorname{argmin}} \quad \|H^4 - \tau \hat{H}^{15}\|_F^2 \quad (5.8)$$

où \hat{H}^{15} est la matrice H^{15} après normalisation L_2 de ses lignes. La Figure 5.3 retranscrit sous forme d'un gradient de couleurs les différentes valeurs de la matrice τ .

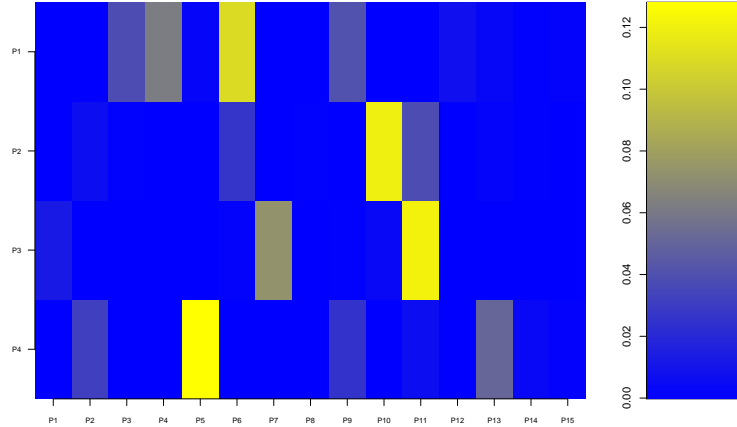


FIGURE 5.3: Visualisation de la matrice τ résultat de la NNLS (5.8). Les coefficients sont représentés par un gradient de couleur du bleu (0) au jaune (1).

la proximité de chaque CTFA i de H_4 avec le cône convexe engendré par H^{15} est évalué grâce à l'erreur de reconstruction relative définie par

$$E_i = \frac{\|H_{(i,.)}^4 - (\tau H^{15})_{(i,.)}\|_F}{\|H_{(i,.)}^4\|_F}. \quad (5.9)$$

Les erreurs de reconstruction des 4 CTFA sont respectivement de 0.456, 0.045, 0.032 et 0.109. On constate que les erreurs de reconstruction sont non nulles et donc que les deux cônes ne sont pas emboîtés. Cependant, mis à part le CTFA 1, les vecteurs lignes de H^4 sont bien reconstruits par ceux de H^{15} .

5.1.2 Conservation de l'information de A

Nous nous intéressons par ailleurs aux propriétés de A conservées par \bar{W} .

Conservation du pourcentage de marqueurs de la métabolisation des fibres. La matrice A est composée des abondances relatives de 86 marqueurs fonctionnels pour chacun des 1408 échantillons. De ce fait, la somme sur les marqueurs j des A_{ij} correspond au pourcentage de marqueurs fonctionnels de métabolisation des fibres de l'échantillon i . Cette grandeur varie de 0,008 à 0,018 et évolue selon les études comme illustré par la Figure 5.4-(a) et en particulier pour les individus de MicroObes qui possèdent une proportion plus importante de marqueurs dédiés à la métabolisation des fibres. Quelque soit i un échantillon métagénomique, la grandeur

$\sum_j A_{ij}$ est approximée par $\sum_j \sum_l W(il)H_{lj}$. Après la normalisation précédemment décrite cette approximation peut se réécrire

$$\sum_j A_{ij} \sim \sum_l \bar{W}_{il}.$$

La pertinence de cette approximation est examinée en calculant la corrélation entre $\sum_j A_{ij}$ et $\sum_l \bar{W}_{il}$ illustrée par la Figure 5.4-(b). La forte corrélation (corrélation de Pearson de 0.91) implique que la variabilité individuelle des proportions de marqueurs de la dégradation des fibres est très bien conservée par NMF.

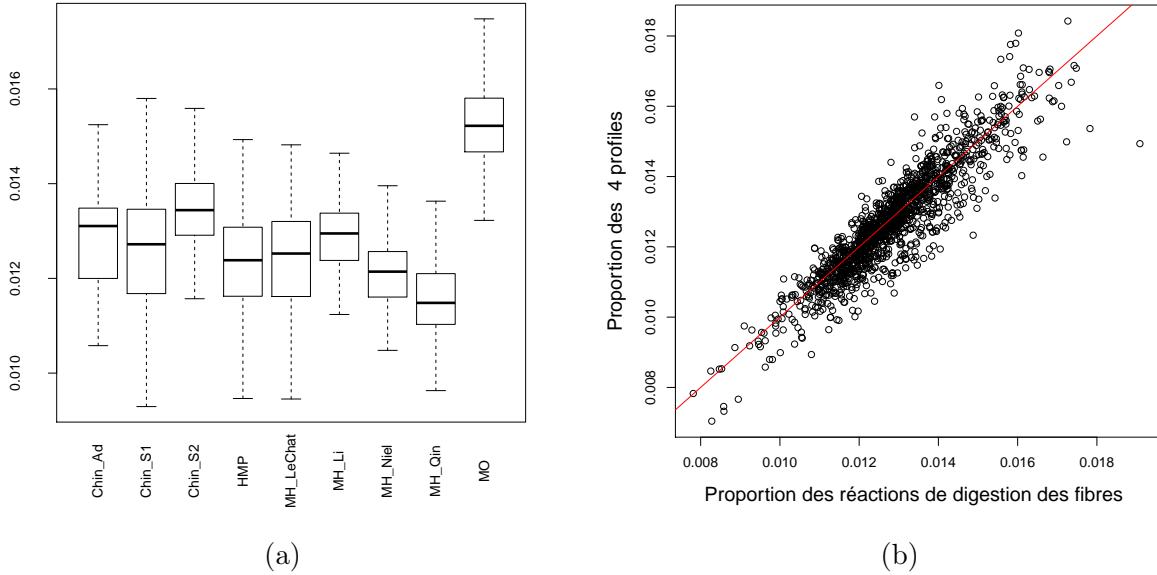


FIGURE 5.4: (a) Diagramme en boîte de l'abondance relative totale des marqueurs fonctionnels de métabolisation des fibres par études. Les échantillons chinois provenant de Qin et al. (2012) ont été scindés en trois groupes nommés Chin_Ad, Chin_S1 et Chin_S2 relatifs à différents lots d'analyse dans l'étude. Les autres abréviations correspondent aux études suivantes : HMP Methé et al. (2012), MH_LeChat Le Chatelier et al. (2013), MH_Li Li et al. (2014), MH_Niel Nielsen et al. (2014), MH_Qin Qin et al. (2010) et MO Cotillard et al. (2013); (b) proportions de marqueurs fonctionnels imputable aux CTFA (somme par ligne des W) en fonction des proportions de marqueurs fonctionnels de métabolisation des fibres (somme des lignes de A)

Conservation de la structure géométrique de A . Soit D^A la matrice associée à A dont le terme D_{ij}^A est la distance L_2 entre l'échantillon i et l'échantillon j . Soit de même $D_{ij}^{\bar{W}}$ la matrice des distances entre les lignes de la matrice \bar{W} . Les corrélations entre ces deux matrices sont calculées selon $C_i = \text{corr}((D_{ii'}^{\bar{W}})_{i \neq i'}, (D_{ii'}^A)_{i \neq i'})$ et représentées sur l'histogramme de la Figure 5.5.

Toutes les corrélations sont supérieures à 0.7 et parmi elles, 99% sont supérieures à 0.85. Ceci indique que la similarité entre échantillons biologiques est bien conservée dans \bar{W} .

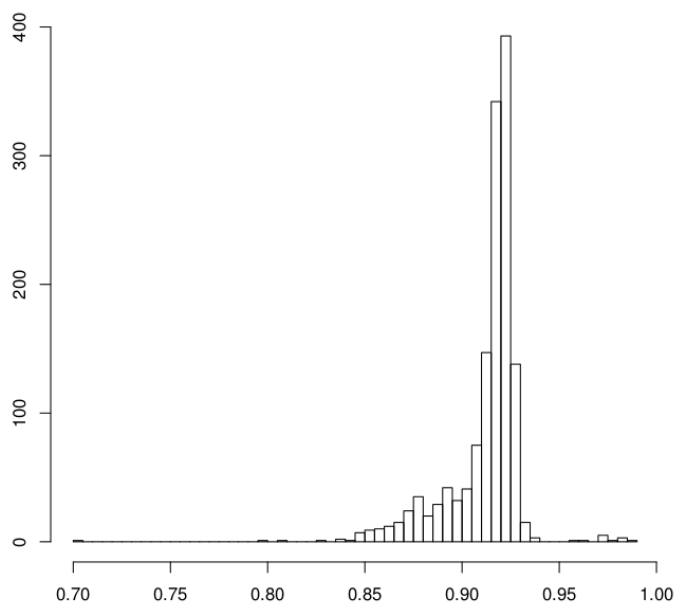


FIGURE 5.5: Corrélation entre géométries de A et de W

5.2 Analyse des CTFA

Après une présentation rapide des CTFA, nous examinons l'impact de la contrainte et décrivons plus en détails les particularités fonctionnelles de chacun d'eux.

5.2.1 Présentation générale

Visualisation. Les 4 CTFA sont présentées en Figures 5.8, 5.9, 5.10 et 5.11. Nous utilisons le graphe construit dans le Chapitre 2 comme support pour cette visualisation. Chacun des 61 marqueurs fonctionnels de la fermentation des sucres simples est associé à une flèche du graphe. La couleur correspond à l'abondance du marqueur dans le CTFA exprimé en pourcentage du marqueur le plus présent dans le CTFA. Les parties grisées sont absentes ou d'abondance plus

faible que 0.5% de l'abondance de ce marqueur. Les marqueurs de l'hydrolyse des fibres en sucre simple sont représentés dans une matrice dont les cases sont colorées selon la même convention.

Il est important de bien distinguer ces représentation graphique, qui n'est qu'une façon commode et facilement interprétable de visualiser les profils, des graphes de réactions utilisés dans le contexte des modèles métabolique de type Flux Balance Analysis. En effet comme nous l'avons vu dans les chapitres précédents, nous modélisons des abondances de gènes et en aucun cas des flux de métabolites.

Distribution. Les CTFA 1 et 4 sont peu présents, en effet leurs abondances sont nulles dans respectivement 653 et 444 échantillons contre seulement 38 et 9 pour les CTFA 2 et 3.

La Figure 5.6 représente la distribution des CTFA pour lesquelles les valeurs nulles ont été tronquées ; les distributions d'abondances des CTFA 1 et 4 se distinguent par une grande quantité d'échantillons d'abondances très faibles accompagnées par une queue de distribution all

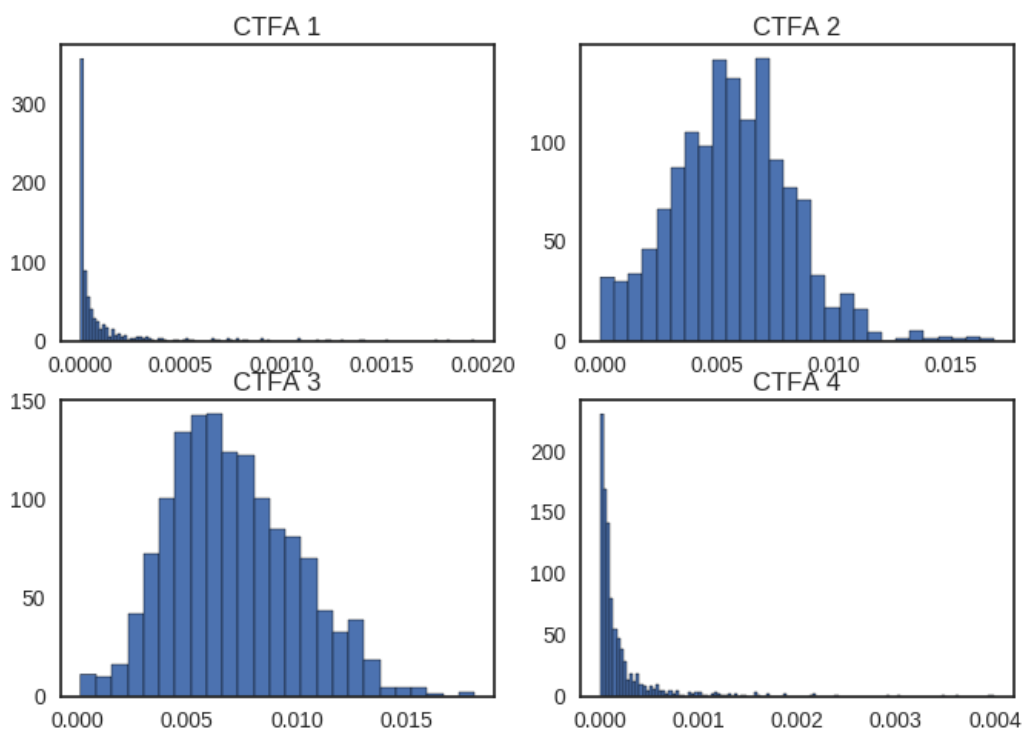


FIGURE 5.6: Distributions des abondance des 4 CTFA, les valeurs nulles ont été tronquées afin de ne pas écraser les distributions.

La Figure 5.7 illustre les différences importantes entre les apports de chacun des CTFA ; les CTFA 2 et 3 contribuent majoritairement à la reconstruction de A tandis que les CTFA 1 et 4 n'y contribuent que très marginalement. Dans l'ordre les contributions sont respectivement de 0.4%, 43.2%, 55.3% et 1.1%.

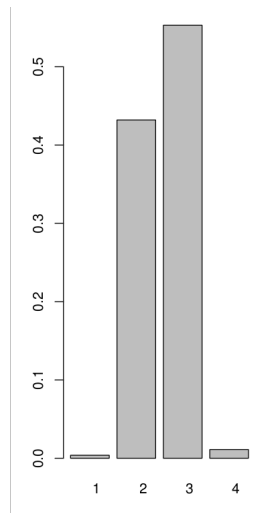


FIGURE 5.7: Contribution des 4 CTFA en pourcentage de la reconstruction de la matrice d'abondance A

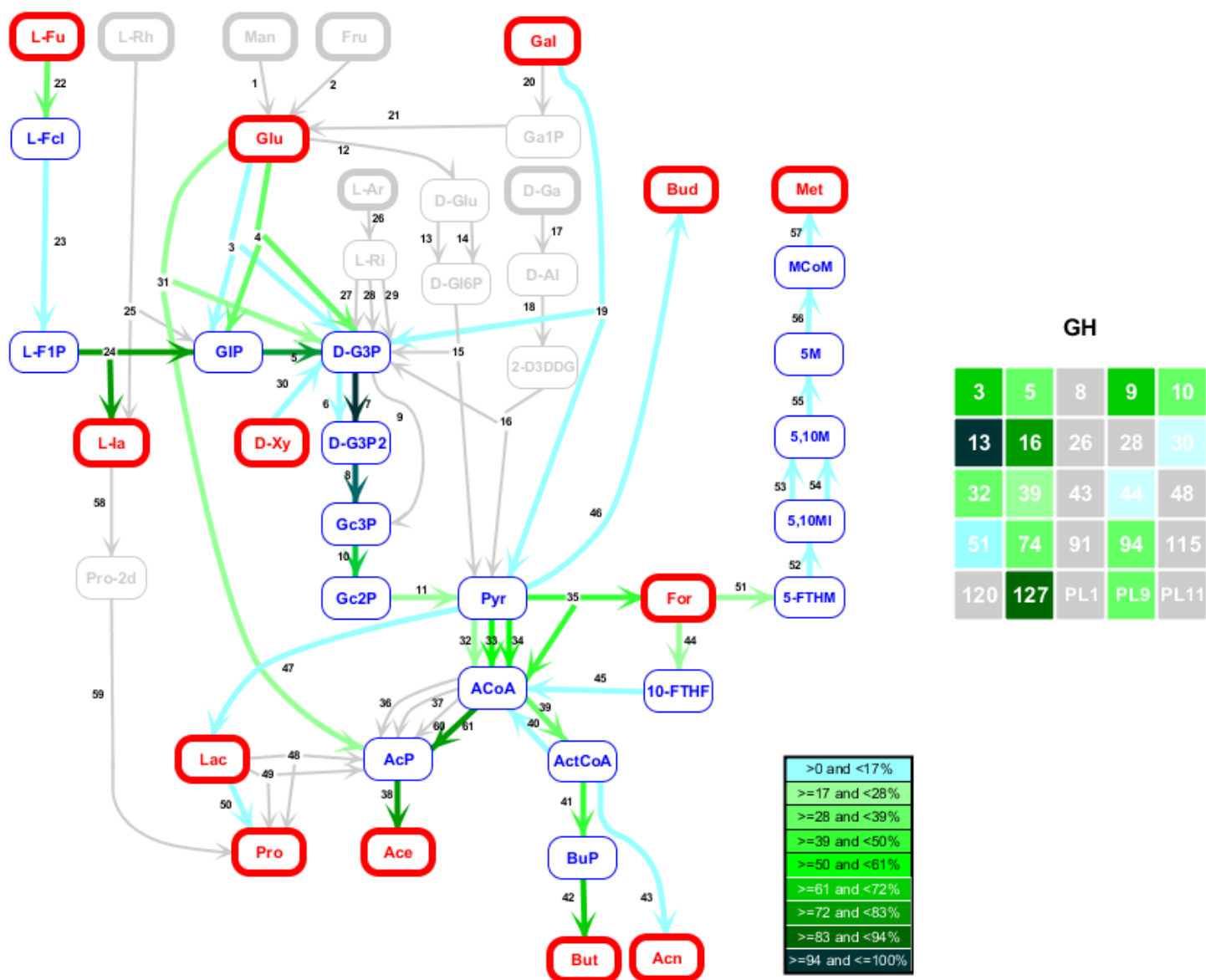


FIGURE 5.8: Représentation graphique du CTFA 1. Chaque flèche correspond à un des 61 marqueurs fonctionnels associés à la fermentation des sucres simples. La couleur des flèches symbolise l'abondance du marqueur dans le CTFA 1 en pourcentage du marqueur le plus abondant (ici marqueur 7). Les 25 marqueurs associés à l'hydrolyse sont représentés sous forme d'un tableau et de même, leur couleur représente un pourcentage de l'abondance du marqueur le plus abondant (ici GH13). Si l'abondance du marqueur est plus faible que 0.5% du maximum, l'arête où la case associée est grisée.

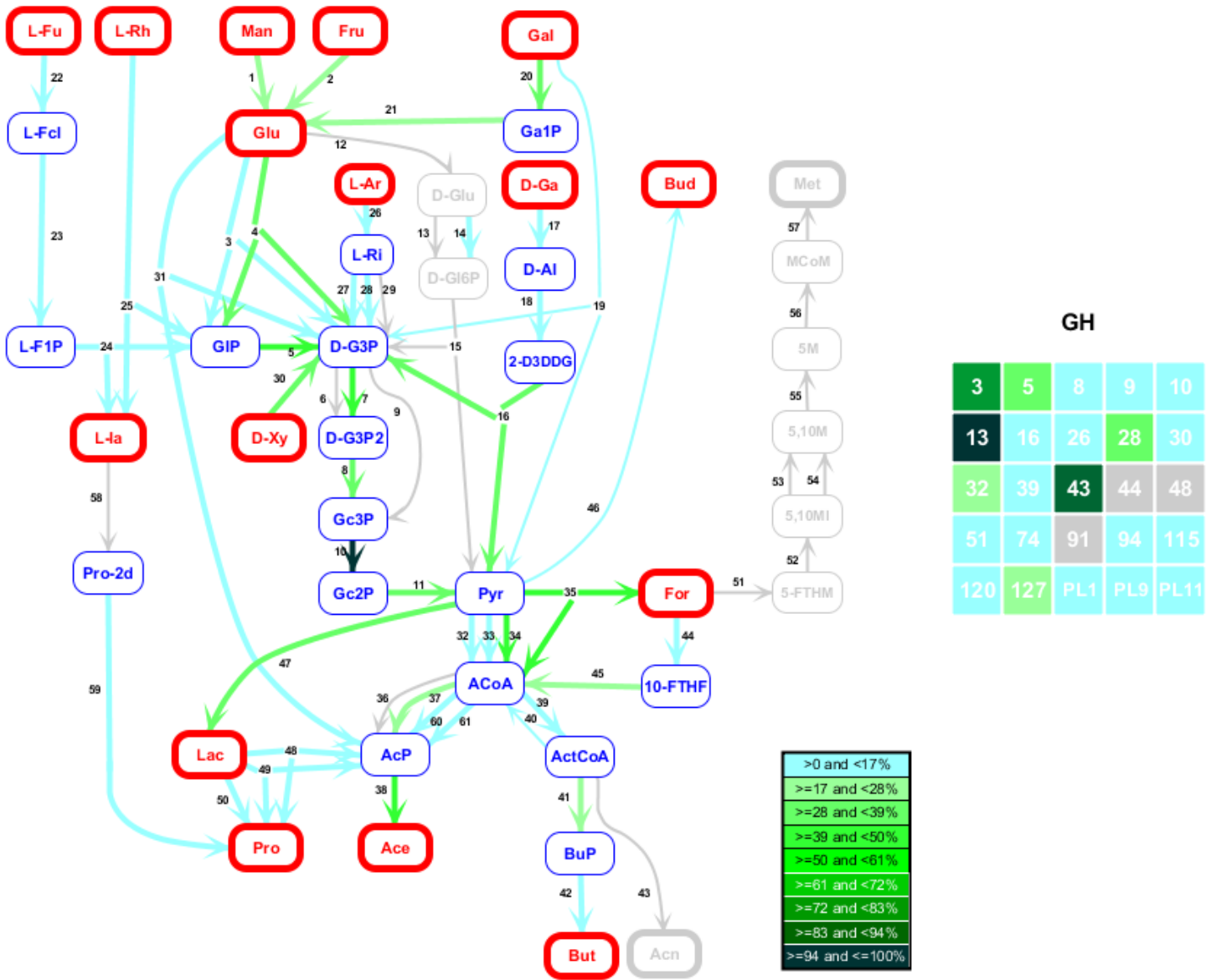


FIGURE 5.9: Représentation graphique du CTFA 2. Les conventions sont les mêmes que pour le CTFA 1.

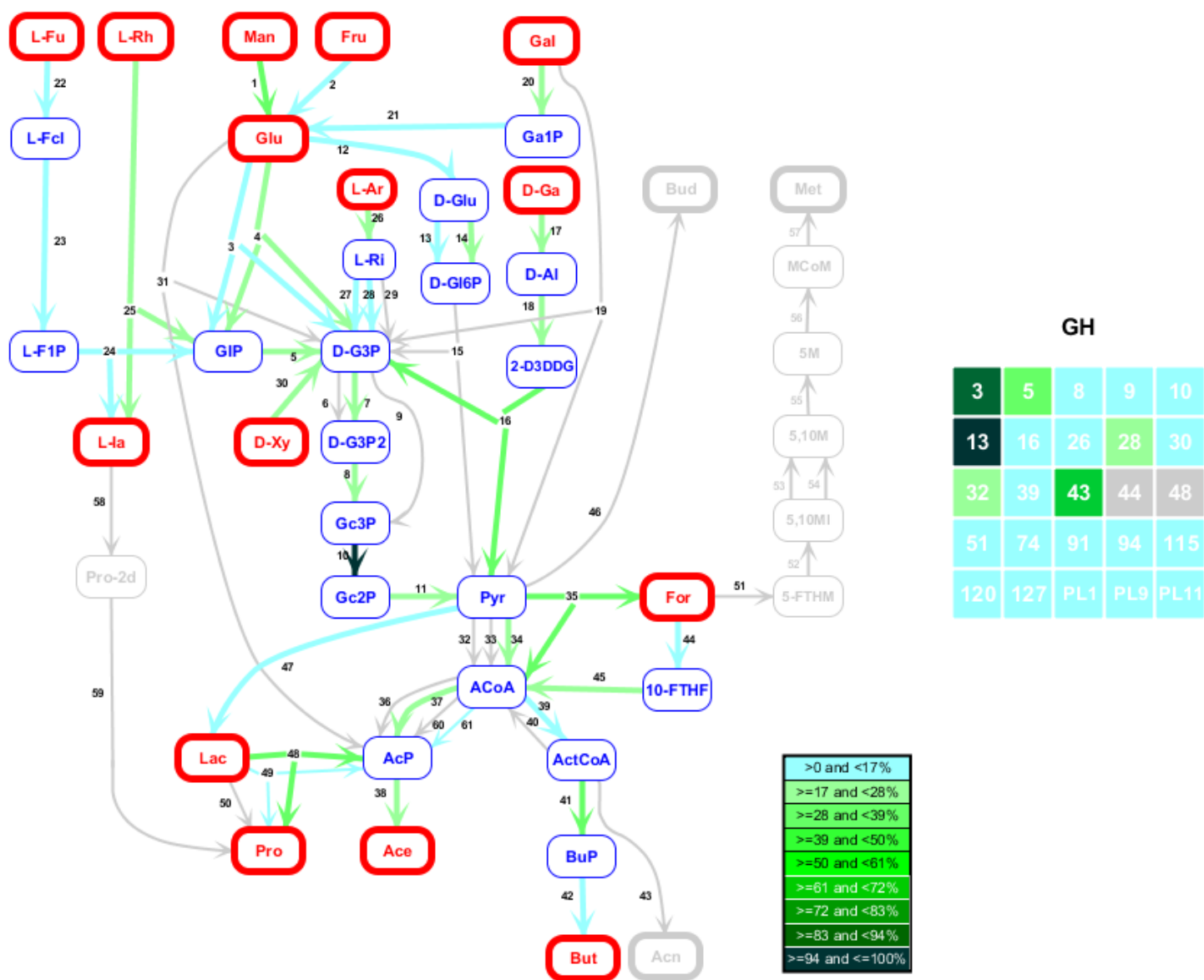


FIGURE 5.10: Représentation graphique du CTFA 3. Les conventions sont les mêmes que pour le CTFA 1.

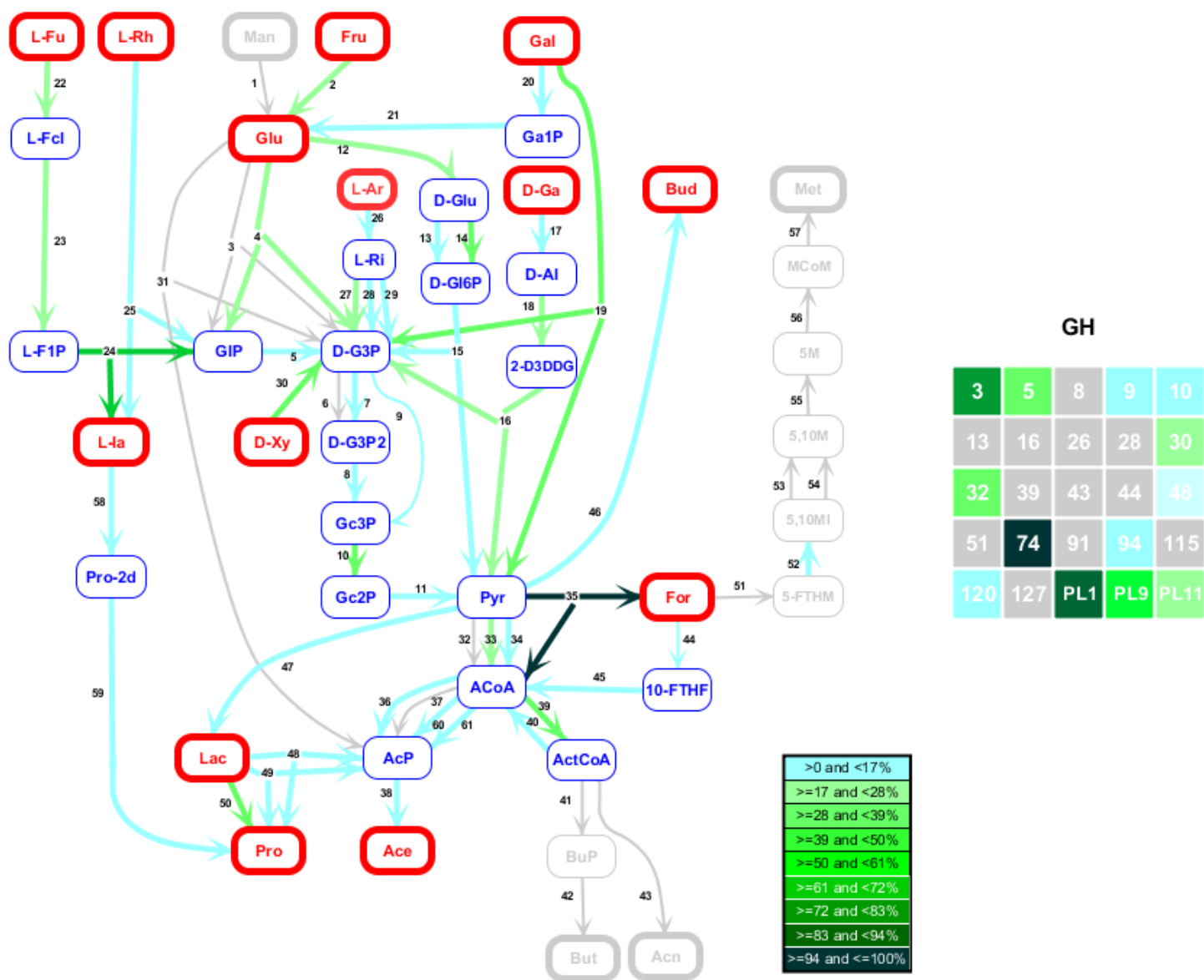


FIGURE 5.11: Représentation graphique du CTFA 4. Les conventions sont les mêmes que pour le CTFA 1.

Conséquence de la contrainte. Sur les 38 contraintes implémentées, on observe entre 3 et 7 contraintes actives par CTFA, principalement concernant les marqueurs fonctionnels 52 à 57 associés à la méthanogénèse. Bien que ce nombre soit faible, cela ne signifie pas que la plupart des autres contraintes n'aient pas un rôle important dans la détermination des CTFA. En effet, le

problème non contraint possède plusieurs minima locaux, la présence des contraintes permet de délimiter une région spécifique dans laquelle le minimum sera recherché, même si ce minimum n'est pas situé sur le bord du domaine délimité par les contraintes.

Les contraintes que nous imposons ont un effet local c'est à dire relatif à un nombre réduit de marqueurs fonctionnels associé à un métabolite. Elles induisent cependant un effet à l'échelle du CTFA en permettant l'obtention d'une succession cohérente de marqueurs fonctionnels qui peuvent s'interpréter, comme un potentiel de réalisation de voies métaboliques. En effet il est possible de retrouver des chemins ininterrompus entre sucres simples et acides gras à chaînes courtes.

Nous observons la présence non-désirable du marqueur fonctionnel 52 dans la Figure 5.11. Il s'agit ici d'un problème de précision numérique. Lors de l'exécution d'un algorithme d'optimisation, sa convergence est toujours évaluée relativement à un seuil. L'abondance non nulle du marqueur 52 est due à un non respect des contraintes qui rentre dans la marge d'erreur numérique définie par le seuil.

Nous avons choisi en Section 3.3.2 de ne pas contraindre les abondances de marqueurs fonctionnels associées aux métabolites suivant, D-gluconate 6 Phosphate, D-Glucono-1,5-lactone 6-phosphate, Propane1,2diol et L-Fuculose 1-phosphate. Nous avons fait ce choix après avoir observé la présence, dans un groupe de génomes de référence (voir Section 3.3.2), de marqueurs fonctionnels dont il est impossible de justifier l'occurrence par la présence conjointe des 60 autres marqueurs. Nous retrouvons dans les CTFA des caractéristiques similaires, ce qui est cohérent avec la réalité biologique décrite ci-avant. De manière exhaustive, le marqueur 58 est manquant et le marqueur 14 est présent en Figure 5.9, le marqueur 15 est manquant en Figure 5.10.

La présence conjointe de marqueurs fonctionnels non contraints dans un même CTFA indique que les abondances de ce CTFA au sein des 1408 échantillon permettent de bien reconstruire ces marqueurs non contraints dans les échantillons méta-génomiques. Dans la mesure où les abondances des CTFA sont entièrement dépendantes de leur composition en marqueurs fonctionnels, l'implémentation d'une contrainte concernant les abondances d'une partie des marqueurs à un effet direct sur la présence de tous les marqueurs dans les CTFA.

Nous observons plusieurs occurrences d'associations cohérentes de marqueurs fonctionnels non contraints dans les CTFA qui pourraient être expliquées par ce phénomène. En premier lieu, les marqueurs fonctionnels associés au L-Fuculose 1-phosphate (23 et 24) sont toujours présents

de façon conjointes contrairement à ce qui à été observé sur les génomes de références, il est donc possible pour chaque CTFA d'interpréter la présence non nul de ces marqueurs comme un potentiel de réalisation de la fermentation du Fucose. De même nous observons un phénomène similaire limité au CTFA 4, concernant les marqueurs associés aux métabolites D-gluconate 6 Phosphate et D-Glucono-1,5-lactone 6-phosphate (12 à 15). Cette dernière observation est intéressante puisque cela signifie qu'une partie de l'abondance de ces marqueurs peut effectivement s'interpréter, pour le CTFA 4, comme le potentiel de réalisation de la voie d'Entner Doudoroff. De même, le GH spécifique de la production du Rhamnose n'est présent que dans les CTFA 2 et 3, seuls CTFA possédants des marqueurs associé à sa consommation.

5.2.2 Comparaison des CTFA

Afin d'étudier les CTFA, nous considérons deux informations complémentaires, la composition en terme de pourcentage du marqueur le plus abondant présenté dans les Figures 5.8 à 5.11, et le pourcentage de marqueurs reconstruits par chaque profil. Cette deuxième information prend en compte les abondances des CTFA et permet de visualiser quel CTFA est le plus impliqué dans la reconstruction de chaque marqueurs fonctionnels. Elle peut se calculer comme une normalisation particulière de H noté \check{H} et de terme

$$\check{H}_{lj} = \frac{\sum_i W_{il} H_{lj}}{\|WH\|_1}. \quad (5.10)$$

Les Figures 5.12, 5.13, 5.14 et 5.15 illustrent sous forme d'histogrammes l'importance de chaque CTFA dans la reconstruction de chaque marqueur fonctionnel. Cette deuxième information permet d'observer que des marqueurs très faiblement présent (moins de 0.5%) dans un CTFA, peuvent tout de même avoir un rôle majeur dans la reconstruction de l'abondance d'un marqueurs fonctionnel. C'est par exemple le cas du marqueur 43 associé à la création d'acétone qui représenter une part minoritaire des marqueurs du CTFA 2 alors que ce CTFA reconstruit 89% de l'abondance totale en marqueur 43.

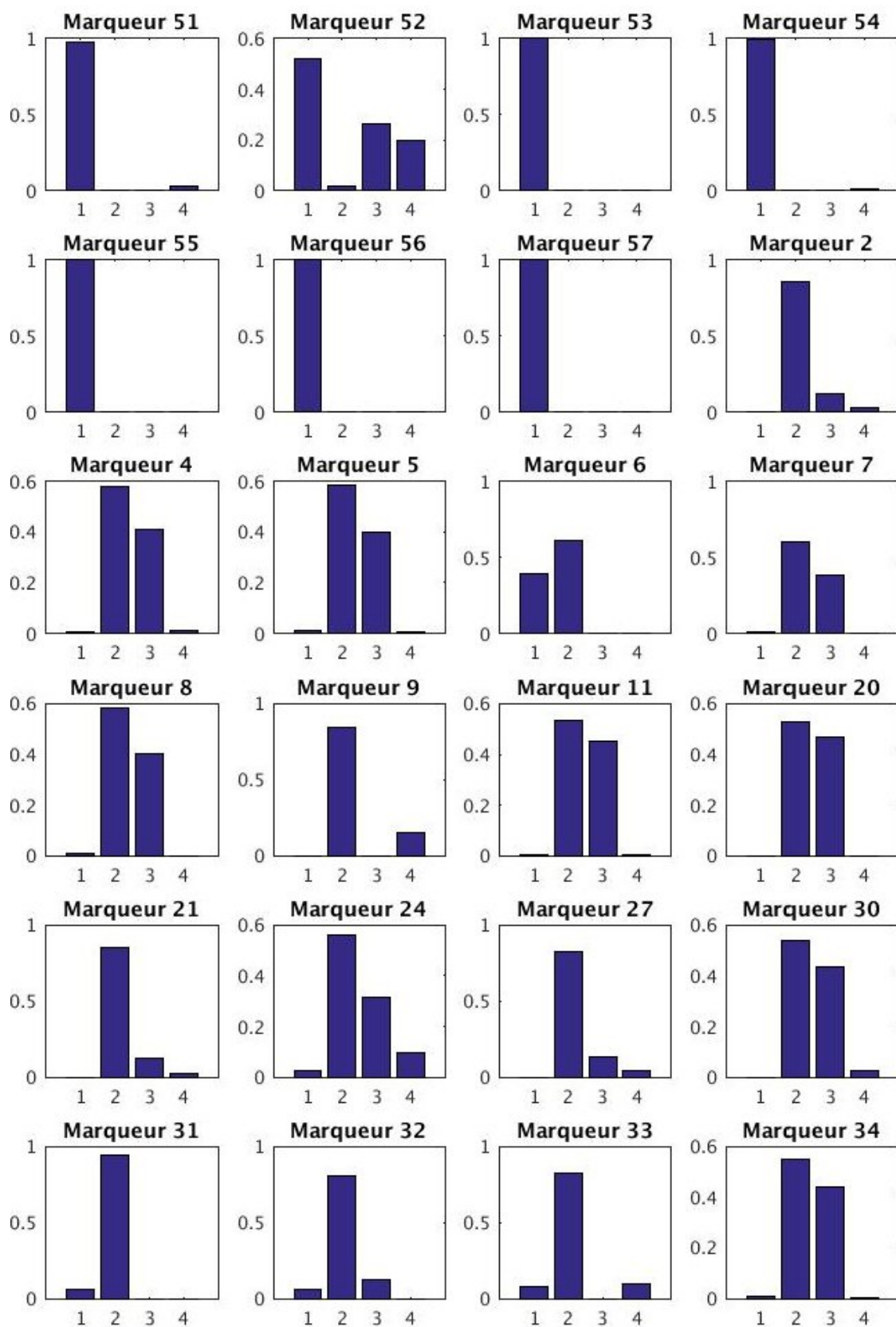


FIGURE 5.12: Histogrammes représentant la contribution de chaque CTFA dans la reconstruction de chaque marqueurs fonctionnels. Les 7 premières figures montrent les marqueurs principalement reconstruits par le CTFA 1. Les suivantes correspondent aux marqueurs principalement reconstruits par le CTFA 2.

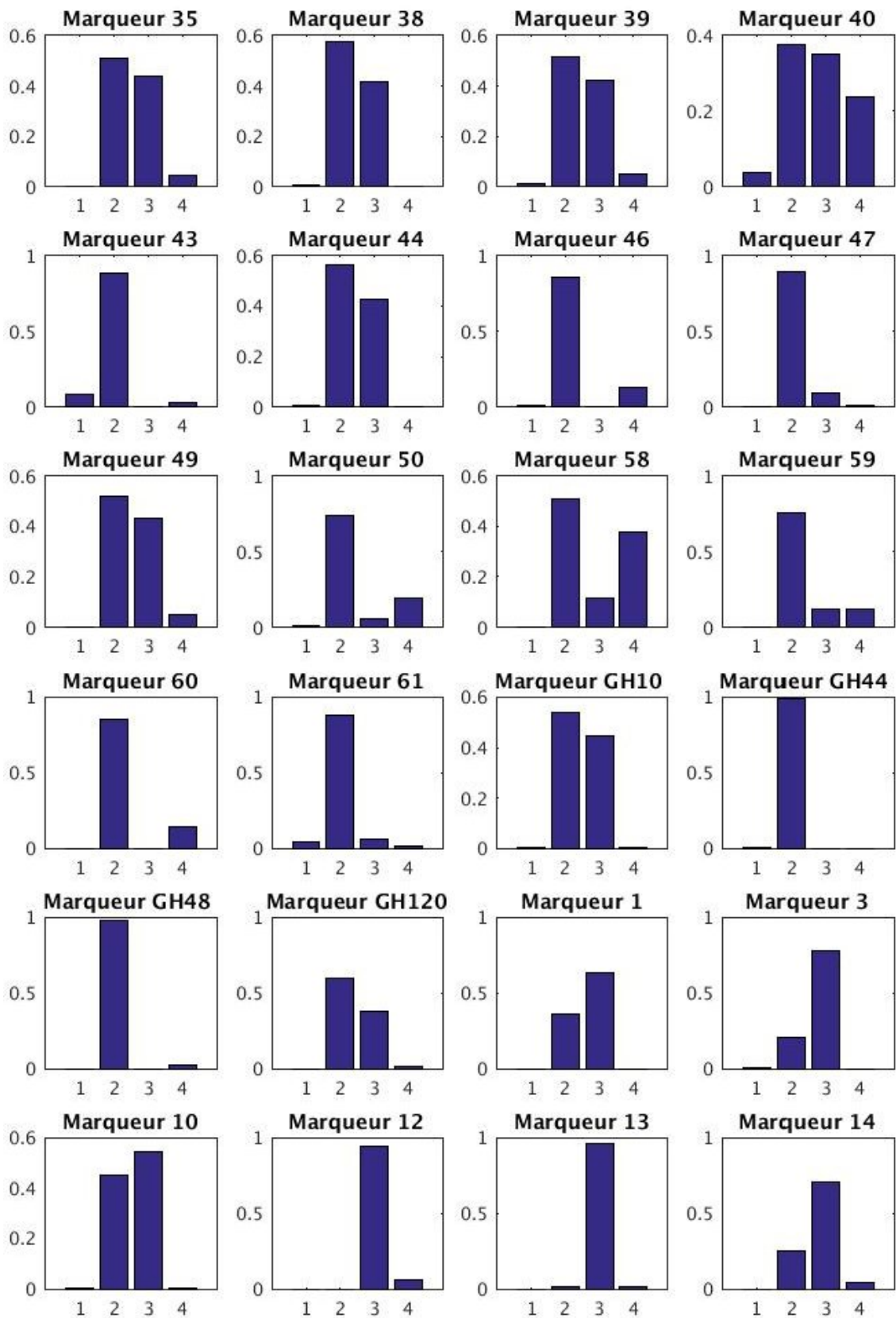


FIGURE 5.13: Histogrammes représentant la contribution de chaque CTFA dans la reconstruction de chaque marqueurs fonctionnels. Les 6 dernières figures montrent les marqueurs principalement reconstruits par le CTFA 3. Les précédentes correspondent aux marqueurs principalement reconstruits par le CTFA 2.

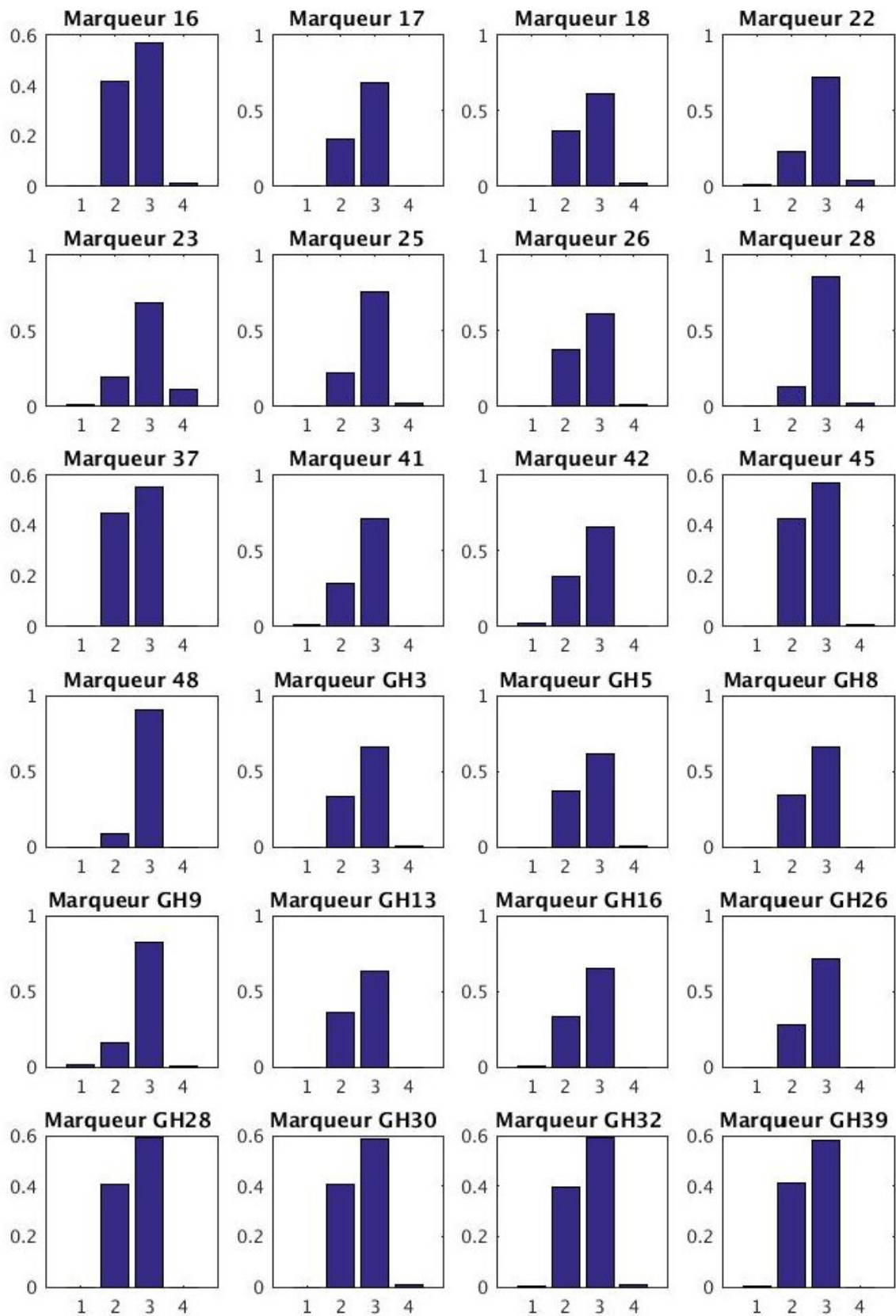


FIGURE 5.14: Histogrammes représentant la contribution de chaque CTFA dans la reconstruction de chaque marqueurs fonctionnels. Ici ces figures montrent les marqueurs principalement reconstruits par le CTFA 3.

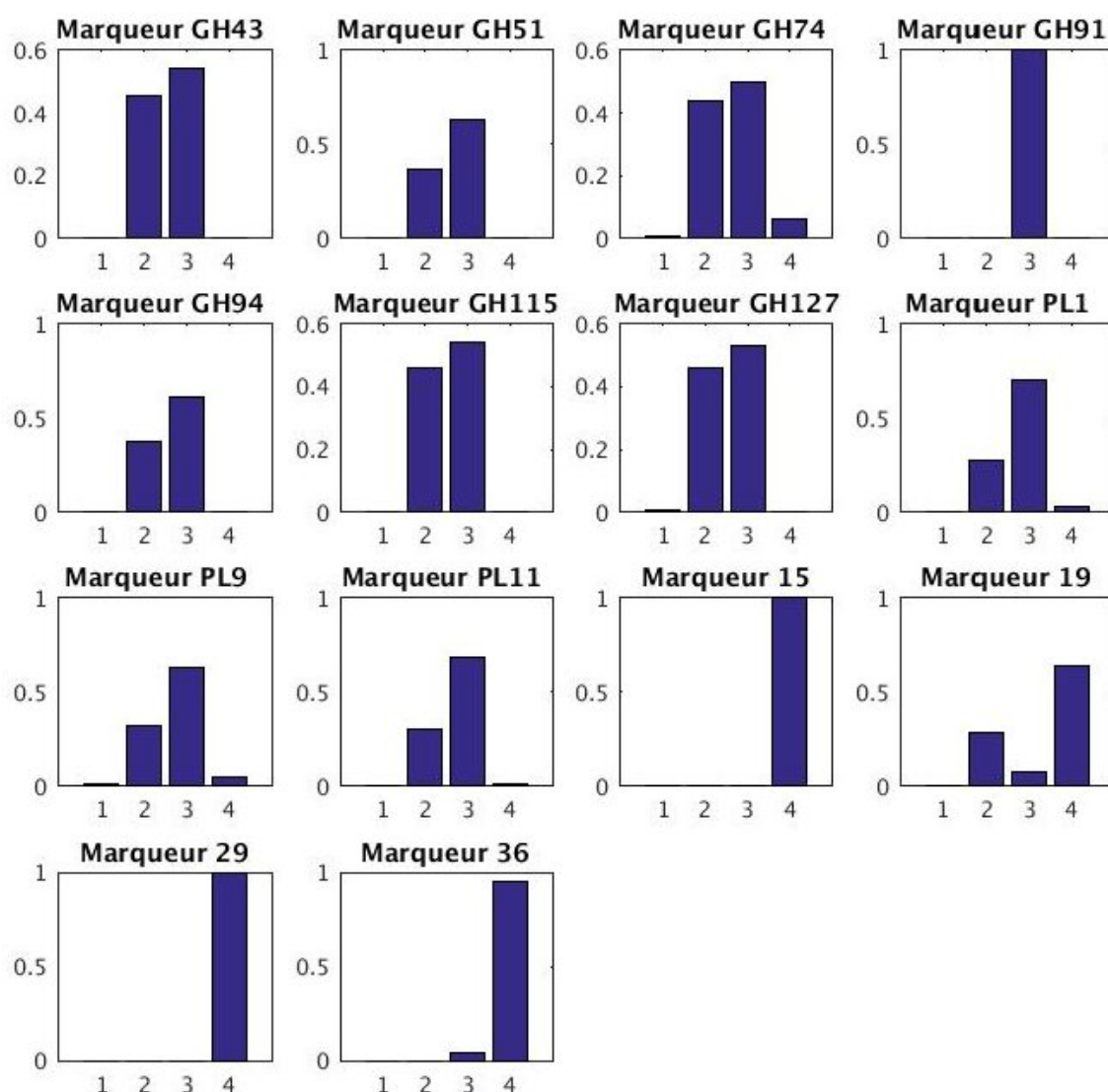


FIGURE 5.15: Histogrammes représentant la contribution de chaque CTFA dans la reconstruction de chaque marqueurs fonctionnels. Les 4 dernières figures montrent les marqueurs principalement reconstruits par le CTFA 4. Les précédentes correspondent aux marqueurs principalement reconstruits par le CTFA 3.

CTFA 1. Le premier CTFA est d'abondance la plus faible. Il est aussi le plus parcimonieux avec 34 marqueurs strictement nuls. Il est caractérisé par un nombre réduit de voies métaboliques potentiellement réalisées. La plus remarquable est la méthanogénèse dont il à l'exclusivité comme présenté en Figure 5.8. Cette capacité métabolique est assurée par les Archae métha-

nogènes hydrogenotrophiques présents dans environ 50% des microbiotes intestinaux humains en faible quantités. Nous retrouvons ici cette information puisque ce CTFA est peu prévalent et absent dans 653 échantillons sur 1408. Aucun autres marqueurs ne sont uniquement ou majoritairement reconstruit par ce CTFA.

Les voies métaboliques potentiellement réalisées par l'assemblage fonctionnel décrit par ce CTFA sont la fermentation du fucose, glucose, galactose, xylose en formate, Lactose, Acétate, Acétone, butanoate, butane-diol et Propanoate par le biais de la voies métabolique Emben-Meyerhoff-Parnas. La production de Propanoate est réalisée par le voie de l'acrylate. La présence du marqueur 31 dénote d'un potentiel de réalisation du Bifidobacterium shunt.

CTFA 2. Ce CTFA est le second plus abondant et aussi le moins parcimonieux. Toutes les voies métaboliques considérées dans ce travail sont potentiellement réalisées par l'assemblage fonctionnel décrit par ce CTFA à l'exception de la méthanogénèse et de la voie d'Entner-Doudoroff. Seuls le GH 91 spécifique du fructose est strictement absent de ce CTFA, la catalyses des fibres du fructose est par ailleurs assurée par les GH 13 et 32.

De nombreux marqueurs fonctionnels sont majoritairement présent dans ce CTFA. La fermentation du fructose, la production d'acétone de butanediol et de lactate sont essentiellement expliqués par ce CTFA. C'est aussi le cas de la production du propanoate par la voie de l'acrylate et du propanediol, et de la voie du bifidobacterium shunt. Enfin, les deux voies de fermentation du galactose sont présente et l'une des deux est essentiellement reconstruite par ce CTFA.

CTFA 3. Ce CTFA est le plus abondant et comme le CTFA2, très peu de voies métaboliques ne sont pas présente, nommément la méthanogénèse, la production d'acétone, de butanediol et le bifidobacterium shunt.

De nombreux marqueurs sont majoritairement expliqués par ce CTFA, c'est d'abord le cas de 21 des 25 GH et PL à l'exception des GH 44 et 48 absents et des GH 120 et 10 associés de façon plus proéminente au CTFA2. De même la fermentation de 5 des 9 sucres simples représentés sont essentiellement reconstruit par le CTFA3 (mannose, le D-galacturonate, le fucose le rhamnose et l'arabinose) ce qui est cohérent avec la forte reconstruction des *GH*. C'est aussi le cas des marqueurs de la production de propanote par la voie du succinate.

CTFA 4. Le CTFA 4 est similaire au premier puisque d'abondance faible et possédant un nombre élevé de marqueurs fonctionnels absents. De ce fait, des voies sont absentes comme la méthanogénèse, le bifidobacterium shunt, la fermentation de mannose et la production de butanoate et d'acétone.

Un nombre limité de marqueurs sont exclusivement présent dans ce CTFA, c'est le cas des marqueurs 15 et 29. Le marqueur 29 est associé à la fermentation du L-arabinose tandis que le marqueur 15 de par sa présence conjointe avec les marqueurs 12 à 14 est indique un potentiel de réalisation de la voie d'Entner Doudoroff. C'est le seul CTFA possédant cette voie complète. Quelques autres marqueurs fonctionnels sont notablement reconstruits tels que les marqueurs 50,58 et 59 associé à la production de propanoate par les voies de l'acrylate et du propanediol.

5.3 Analyse des abondances des CTFA

Dans cette section et les suivantes, nous normalisons en normes L_1 les lignes de \bar{W} de façon à obtenir des proportions de CTFA pour chaque échantillon, la matrice obtenue est notée \tilde{W} . Cette normalisation permet d'éliminer les variations d'abondances totales entre échantillons illustrées par la Figure 5.4-(a).

Méta-variables. Dans cette section, la matrice \tilde{W} est analysée à la lumière des métadonnées. Nous rappelons que les 1408 échantillons étudiés proviennent de populations différentes et sont caractérisés par : la nationalité, le sexe, l'état de santé, l'âge, l'indice de masse corporelle ainsi que l'étude d'origine. L'ensemble de ces informations n'est pas disponible pour tous les échantillons, en particulier pour ceux de Methé et al. (2012) et certains de Li et al. (2014), pour lesquels aucune information n'est disponible. Il est cependant important de noter que les métadonnées sont par construction très corrélées. En particulier, les études ont été réalisées par maladies, il peut donc y avoir une confusion entre effet d'étude et effet de maladie. D'autre part, certaines études ont été réalisées pour des états de santé spécifiques dans des pays spécifiques il peut donc y avoir une confusion entre les effets d'état de santé et de nationalité. Il est donc nécessaire de rester prudent dans les interprétations causales. C'est pourquoi nous construisons une nouvelle méta-variable par concaténation de l'état de santé et de la nationalité ; ce faisant il est possible de séparer l'effet de l'état de santé de l'effet de la nationalité.

5.3.1 Abondances des CTFA selon les méta-variables

Nous analysons l'abondance des CTFA en fonction des méta-variables. En ce qui concerne les métavariabiles catégorielles telles que l'étude, la nationalité, le sexe ou l'état de santé, les abondances des CTFA dans les échantillons appartenant aux différentes catégories sont présentés sous la formes de diagramme en boîte. Pour ce qui est des métavariabiles quantitatives telles que l'âge ou l'Indice de Masse Corporelle (IMC), nous représentons les abondances de CTFA en fonction de la méta-variable accompagné d'une courbe de régression non paramétrique. Les abondances tracées correspondent aux lignes de la matrice \bar{W} définie en (5.1). Nous présentons ici une sélection de résultats, l'intégralité de ces figures est fournie en Annexe B.

Indice de masse corporelle (IMC). Des expériences de transfert de microbiote en modèle animal ont démontré une relation entre microbiote intestinal et prise de poids. De même plusieurs publications présentées au chapitre 1 suggèrent chez l'homme des disparités de composition taxonomique entre individus obèses et d'IMC plus faible. La Figure 5.16 illustre l'abondance des CTFA en fonction de l'IMC. Les IMC disponibles dans nos échantillons varient de 15 à 45. Les CTFA ne semblent pas permettre de discriminer des groupes d'individus d'IMC différents.

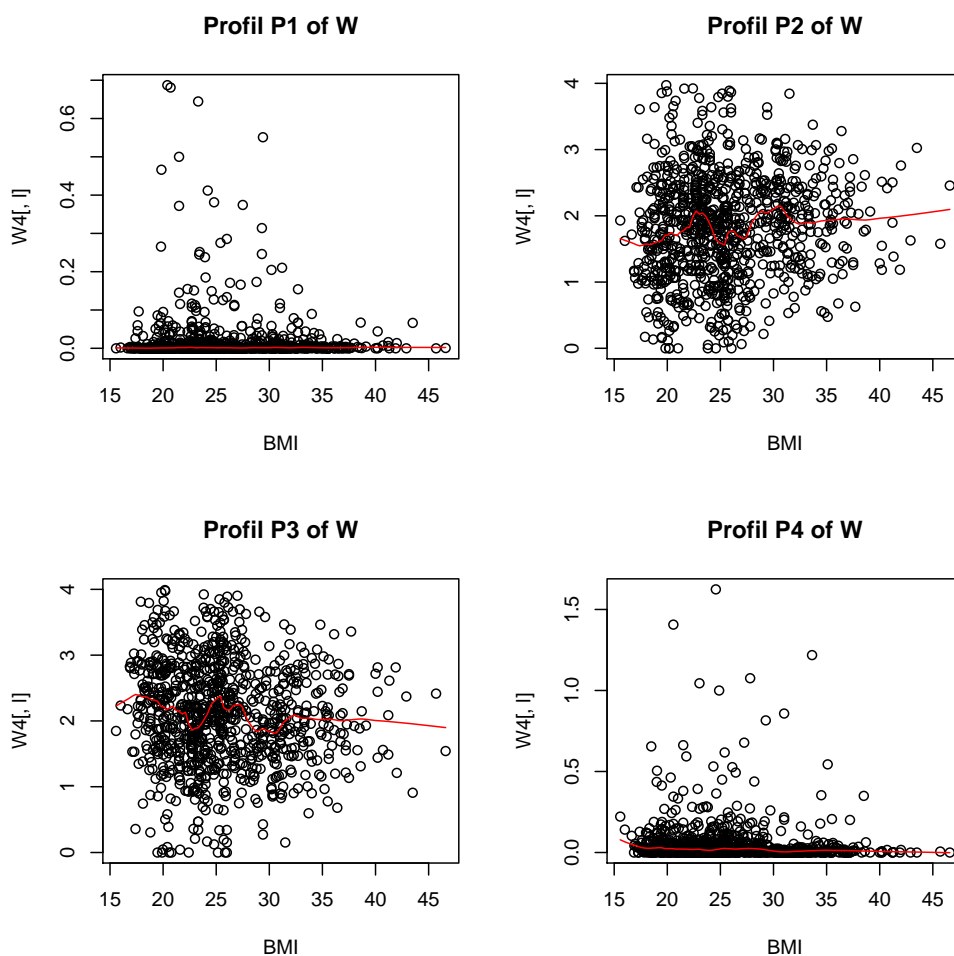


FIGURE 5.16: Distributions d'abondance des CTFA en fonction de l'IMC des individus, la courbe rouge est une régression non paramétrique du nuage de point (Librairie R LOESS).

État de santé. L'état de santé, la nationalité et l'étude d'origine sont très liés, nos données nous permettent cependant de comparer l'état de santé au sein d'une nationalité. Trois états de santé sont considérés, les individus atteints de diabète, colite ulcéreuse et de maladie de Crohn.

Les individus diabétiques proviennent tous de la même étude (Qin et al., 2012) et ont tous la même nationalité. Cette étude portant aussi sur des individus sains, il est possible de quantifier l'impact du diabète sur les abondances de CTFA en s'affranchissant d'un effet étude et nationalité. A la lumière de la Figure 5.17 qui montre les abondances en CTFA par méta-variable, les abondances sont très similaires. Il ne semble pas y avoir d'effet visible lié au diabète sur

l'abondance des CTFA.

Tout les individus espagnols, proviennent de 3 études (Qin et al. (2010),Nielsen et al. (2014) et Li et al. (2014)), celles-ci ont été réalisées par le consortium MetaHit suivant le même protocole. Parmi eux se trouvent tous les individus atteints de colite ulcéreuse et de la maladie de Crohn ainsi qu'un groupe important d'individus sains. Il est donc possible d'isoler l'effet de l'état de santé sur l'abondance des CTFA pour les individus espagnols étudié par le consortium MetaHit. La Figure 5.17 met en évidence des répartition d'abondances en CTFA très similaires entre espagnols sains et atteints de colite ulcéreuse. En revanche, nous constatons que les malades de Crohn présentent des abondances de CTFA nettement différentes de celles des individus sains. Il semblerait donc qu'il existe un effet la maladie de Crohn sur l'abondance des CTFA.

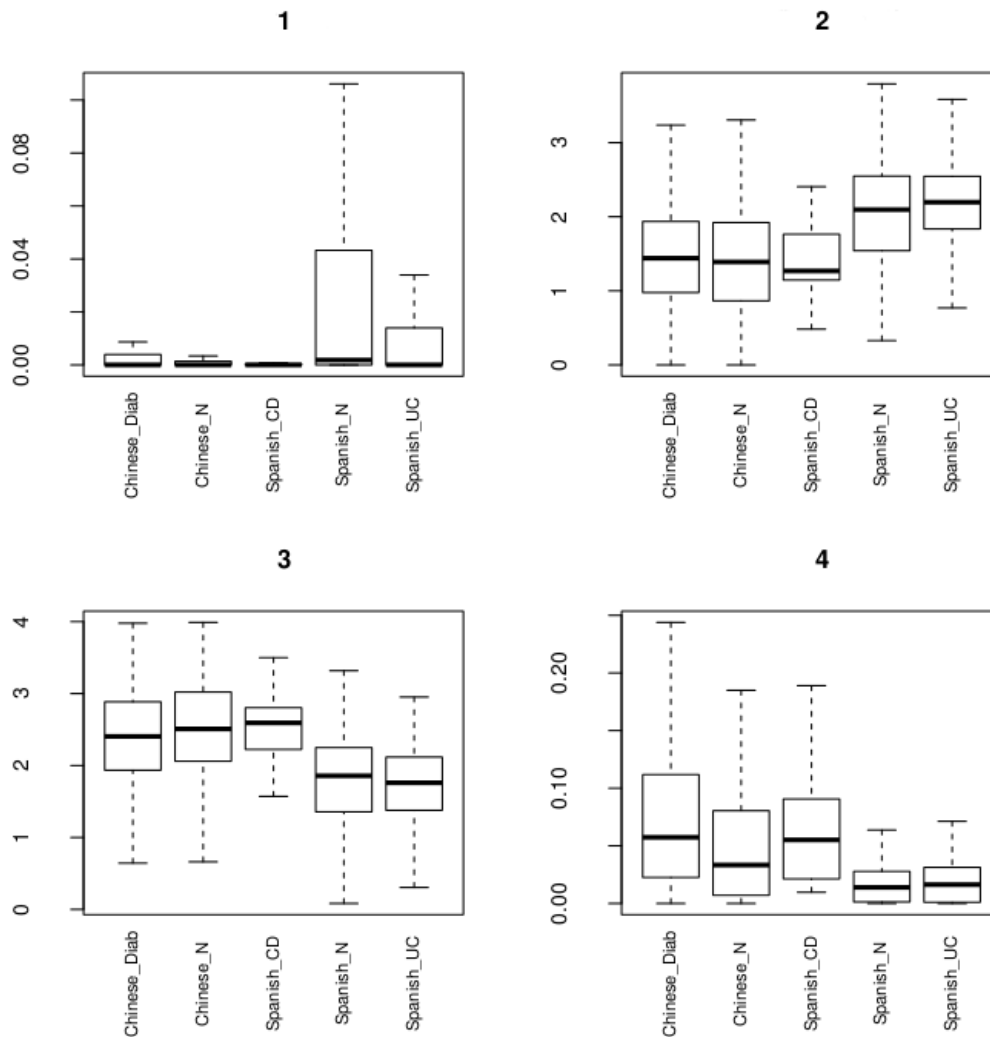


FIGURE 5.17: Diagramme en boîte des abondances des CTFA 1 à 4 au sein des groupes d'échantillons possédant les mêmes couples (nationalité, état de santé). Diab signifie Diabétique, N correspond à Non malade, CD est l'abréviation de Crohn Disease et UC de Colite Ulcéreuse.

Individus HMP. Les échantillons provenant de l'étude du consortium HMP (Methé et al., 2012) sont tous sains et de nationalité américaine. La Figure 5.18 montre que ces échantillons se distinguent des autres. Il est difficile de savoir si la discrimination de ces échantillons métagénomiques par les CTFA du métabolisme des fibres alimentaires est due à des effets techniques d'étude ou si elle relève d'un régime alimentaire ou d'un mode de vie différents.

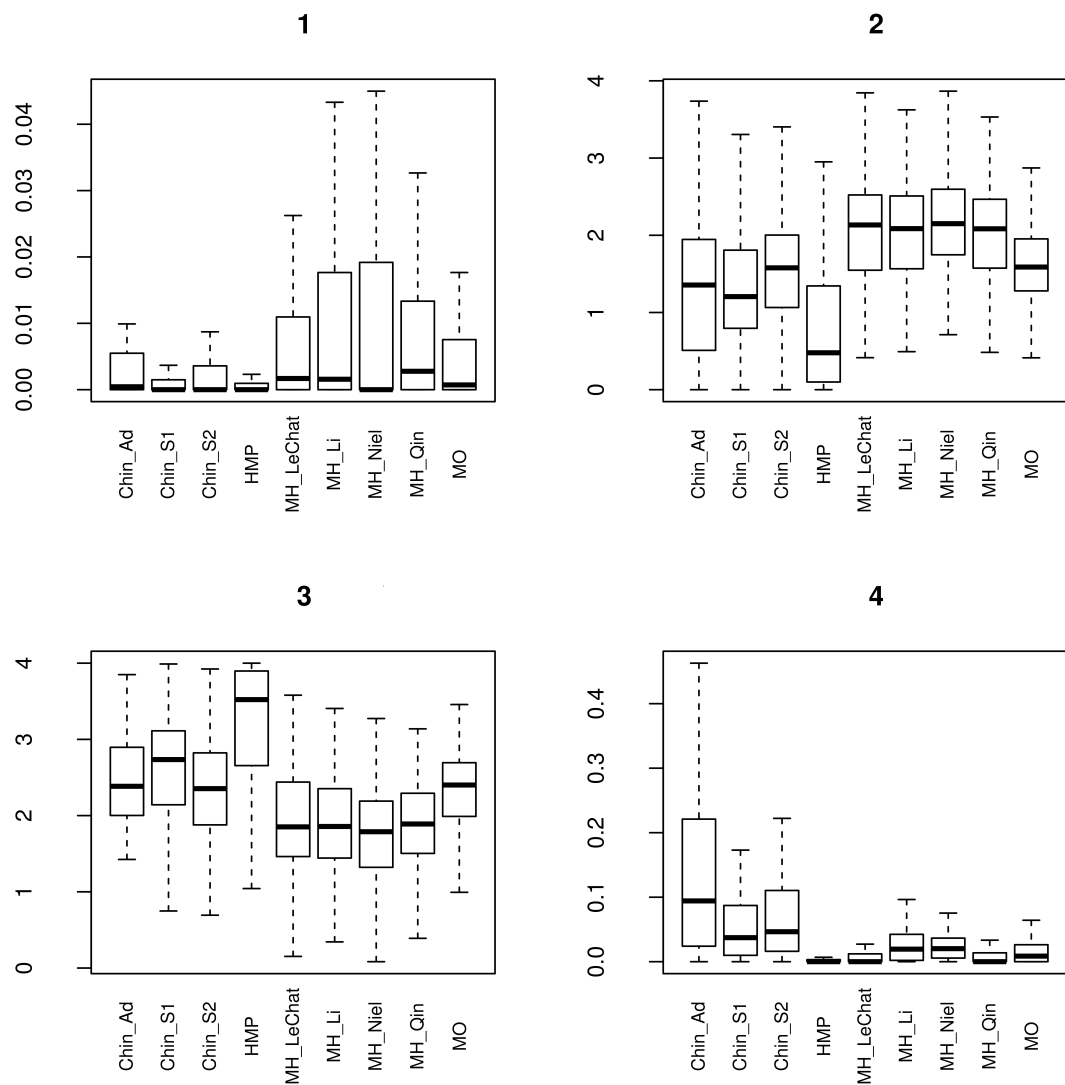


FIGURE 5.18: Diagramme en boîte des abondances des CTFA 1 à 4 au sein des groupes d'échantillons provenant des mêmes études. Les échantillons chinois provenant de Qin et al. (2012) ont été scindés en trois groupes nommés Chin_Ad, Chin_S1 et Chin_S2 relatifs à différents lots d'analyse dans l'étude. Les autres abréviations correspondent aux études suivantes : HMP Me-thé et al. (2012), MH_LeChat Le Chatelier et al. (2013), MH_Li Li et al. (2014), MH_Niel Nielsen et al. (2014), MH_Qin Qin et al. (2010) et MO Cotillard et al. (2013).

5.3.2 Validation de la discrimination de la maladie de Crohn

Afin de valider les différences de composition en CTFA induites par la maladie de Crohn, nous exploitons des données supplémentaires provenant d'une collaboration avec Christopher Quince, chez qui j'ai effectué un séjour de 2 mois à l'université de Warwick. Les échantillons proviennent d'une étude publiée dans Quince et al. (2015). Nous estimons pour ces nouvelles données l'abondance des CTFA construits précédemment et nous vérifions si les tendances observées auparavant pour les échantillons de malade de Crohn et leur contrôles du consortium MetaHit sont reproductibles.

Étude métagénomique temporelle de la maladie de Crohn. Quince et al. (2015) étudient l'impact sur le microbiote intestinal du traitement de la maladie de Crohn par l'alimentation entérique exclusive, traitement dont l'efficacité a été démontrée en pédiatrie. Une cohorte de 23 enfants dont l'âge varie entre 7 et 15 ans est suivie pendant 8 semaines. Un premier échantillon fécal est prélevé avant traitement puis 3 autres de façon régulière jusqu'à la semaine 8, où les individus reprennent un régime alimentaire normal ; un dernier échantillon est prélevé 2 mois après traitement. En parallèle, les auteurs prélèvent 21 enfants sains de la même gamme d'âge. Les échantillons fécaux sont exploités pour obtenir des données d'abondances de d'ARNr 16S ainsi que des données de métagénomique WGS. De manière identique aux études précédentes (Lepage et al., 2013), des profils taxonomiques et fonctionnels très différents entre individus malades et sains sont observés. Les malades de Crohn possèdent une diversité taxonomique plus faible avec toutefois une diversité fonctionnelle plus forte. Le traitement par alimentation entérique exclusive agit de façon paradoxale puisque, bien que permettant d'alléger les symptômes, il ne permet pas de réduire les disparités entre microbiote sains et dysbiotique et en augmente au contraire l'amplitude.

Bien que ces données se rapportent effectivement à des individus malades de Crohn, il est important de ne pas oublier leurs particularités. Les individus concernés sont des enfants. Or il a été observé que chez l'enfant, la diversité du microbiote intestinal est plus faible que chez l'adulte. De même, les patients sont sous traitement par alimentation entérique exclusive, ce qui perturbe significativement leur microbiote. Des disparités avec les échantillons d'apprentissage sont donc vraisemblables.

Validation. Christopher Quince nous a fourni la matrices des abondances des 9.9 millions de gènes du catalogue (Li et al., 2014) pour 69 échantillons métagénomiques. Parmi ces 69 échantillons, 13 proviennent de 12 individus sains et 56 de 12 individus malades à différentes

étapes du traitement. Ces données sont comparables à celles utilisées lors de la construction de la matrice A . Nous suivons le protocole de construction de la matrice d'abondances des marqueurs fonctionnels décrit en Section 2.3 afin d'obtenir une matrice A^V des abondances de 86 marqueurs fonctionnels de la dégradation anaérobie des fibres dans 69 échantillons de validation. Nous nous limitons à une comparaison entre individus sains et malades en omettant l'aspect longitudinal des données.

Nous avons observé précédemment des différences d'abondances des CTFA entre malades de Crohn et individus sains. Nous souhaitons donc vérifier la reproductibilité biologique de ces observations sur de nouvelles données. Pour ce faire nous recherchons une matrices W^V des coordonnées des échantillons de validation A^V dans le cône convexe engendré par \bar{H} . Nous procédons de manière similaire à précédemment en normalisant par colonne la matrice A^V , normalisation reportée sur \bar{H} selon

$$\begin{aligned}\tilde{A}_{ij}^V &= \frac{A_{ij}^V}{(\sum_p (A_{pj}^V)^2)^{1/2}} \\ \tilde{H}_{ij} &= \frac{\bar{H}_{ij}}{(\sum_p (A_{pj}^V)^2)^{1/2}}.\end{aligned}$$

La matrice W^V est obtenue pas moindres carrés positifs

$$W^V = \operatorname{argmin}_{W^V \geq 0} \|\tilde{A}^V - W^V \tilde{H}\|_F^2. \quad (5.11)$$

Nous calculons l'erreur de reconstruction relative des données de validation notée E^V

$$E^V = \frac{\|A^V - W^V H\|_F}{\|A^V\|_F}. \quad (5.12)$$

Cette erreur est égale à 62% ce qui est considérablement plus important que l'erreur de reconstruction de 19% obtenue sur les 1408 échantillons initiaux. Plus en détails, les erreurs relatives des échantillons d'individus sains sont en moyenne de 58% contre 64% pour les échantillons de malades de Crohn. Comme exposé précédemment, les échantillons proviennent d'enfant dont le microbiote est connue pour posséder une diversité taxonomique moins importante que celle de l'adulte. De ce fait l'erreur de reconstruction élevée pourrait provenir d'un manque de robustesse de notre approche où d'une dissimilarité des objets étudiés.

Les Figure 5.19 et 5.20 juxtaposent les diagrammes en boites comparant individus sains

et malades de Crohn pour les données initiales et de validation. Nous observons que les diagramme en boîte sont similaires entre les deux jeux de données. De plus, l'indépendance des échantillons de validation permet de tester la significativité des différences de distribution. De ce fait nous réalisons un test de Wilcoxon unilatéral qui donne les p-value suivantes pour les CTFA respectifs 1.32×10^{-2} , 1.13×10^{-2} , 6.73×10^{-2} et 4.98×10^{-4} . Nous constatons que si les abondances de CTFA 3 ne sont pas significativement distinctes, il n'en est rien pour les autres CTFA et en particulier pour le CTFA 4 qui semble être particulièrement discriminant.

Malgré une reconstruction médiocre des données de validation, cette procédure nous permet de valider la capacité de séparation de 3 CTFA entre malades de Crohn et individus sains.

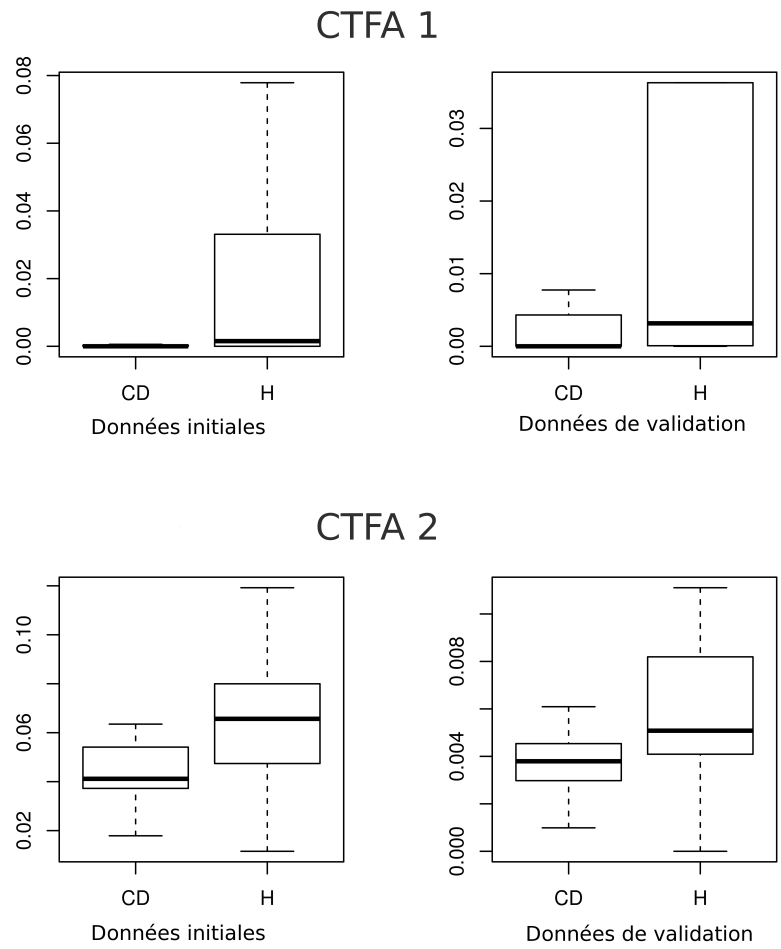
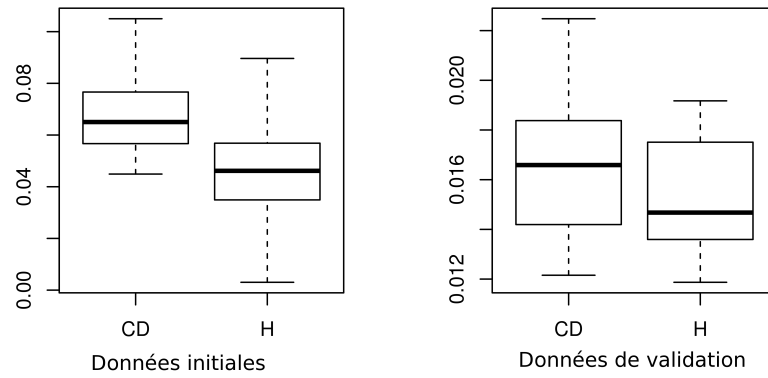


FIGURE 5.19: Diagramme en boîte représentant les distributions d'abondances des CTFA 1 et 2 dans chacun des échantillons d'apprentissage et de validation, regroupés selon l'état de santé, H signifie Healthy et CD Crohn Disease.

CTFA 3



CTFA 4

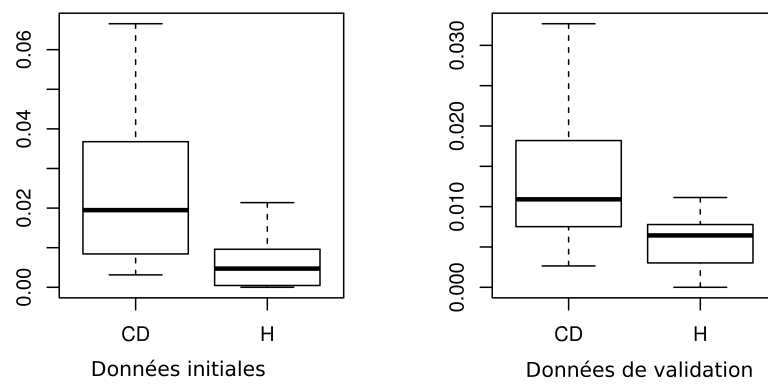


FIGURE 5.20: Diagramme en boîte représentant les distributions d'abondances des CTFA 1 et 2 dans chacun des échantillons d'apprentissage et de validation, regroupés selon l'état de santé, H signifie Healthy et CD Crohn Disease.

Perspectives

J'ai exploité des données métagénomiques conjointement à des informations métaboliques pour construire une représentation, réduite et vraisemblable métaboliquement, de l'organisation fonctionnelle d'assemblages microbiens, dans le cadre la dégradation des fibres alimentaires par le microbiote intestinal humain. Cette approche peut être étendue et j'expose mes perspectives au regard du travail accompli. Cet exposé est structuré en abordant tour à tour les parties mathématiques, modélisation et biologiques.

Factorisation en Matrices non-Négatives

J'ai adopté dans mon travail de thèse une approche d'optimisation, reposant sur la minimisation d'une fonction de coût $E(W, H, \alpha)$ définie en (4.29) sur l'ensemble convexe construit à l'aide de différentes contraintes \mathcal{C} (4.3). Toutefois, il est aussi possible de formuler la résolution de la NMF comme un problème d'inférence bayésienne. Le processus de minimisation peut s'écrire comme la recherche d'un maximum a posteriori où la mesure de dissimilarité correspond à une modélisation statistique des données et les termes de pénalisation correspondent à des lois a priori des matrices de W et H qui sont alors considérées comme des variables aléatoires. A titre d'exemple, dans le cas de la minimisation de E sur \mathcal{C} , on écrit alors

$$P(W, H|A, \alpha) \propto P(A|W, H)P(W|\alpha)P(H|\alpha)$$

où $P(A|W, H) \sim \mathcal{N}(WH, \Sigma)$ et

$$P(W|\alpha) = \prod_{i=1}^n \prod_{j=1}^k \begin{cases} 2 \frac{\sqrt{\alpha}}{\sqrt{\pi}} e^{-\alpha(W_{ij})^2} & \text{si } W_{ij} \geq 0 \\ 0 & \text{sinon} \end{cases}$$

$$P(H|\alpha) \propto \begin{cases} \prod_{i=1}^r e^{-\alpha(\sum_{j=1}^k H_{ij})^2} & \text{si } F_{\Delta}H^T \leq 0 \text{ et } H \geq 0 \\ 0 & \text{sinon.} \end{cases}$$

Le principe est ensuite d'échantillonner la probabilité a posteriori $P(W, H|A, \alpha)$ de manière à obtenir, soit une bonne approximation de la distribution, soit son maximum. L'équivalence entre les deux approches devient apparente en prenant le logarithme du produit de la vraisemblance multipliée par les a priori puisque l'on retrouve alors la fonction de coût 4.29. Différents types d'algorithmes permettent l'exploration de la distribution a posteriori, c'est le cas de l'échantillonneur de Gibbs, de l'algorithme d'Espérance-Maximisation ou l'algorithme de Bayésien variationnel (Cemgil, 2009).

J'envisage dans le cadre de l'approche Bayésienne d'utiliser des lois différentes en a priori et en vraisemblance. En effet, le modèle gaussien actuel est mal adapté aux données positives de variance importante devant l'espérance. D'autres lois permettent la modélisation de données de comptages plutôt que de fréquences, c'est le cas de la loi Multinomiale, la loi de Poisson où la Binomiale Négative. Ce sujet fait l'objet de publications dans le domaine de la génomique (Robinson and Smyth, 2008; Lu et al., 2005). Il est aussi intéressant de changer les a priori sur les matrices W et H . Une loi de Dirichlet adaptée à la modélisation de proportions permettrait d'interpréter les lignes de W et de H respectivement comme des compositions en CTFA et des compositions en marqueurs fonctionnels.

La formulation bayésienne de la NMF rend possible le calcul d'intervalles de crédibilité pour W et H , caractérisant la qualité de l'estimation. De même, elle permet l'utilisation de critères de sélection de modèles tels que le facteur de Bayes et le DIC (Deviance information criterion) (Gelman et al., 2004) que j'aimerais comparer à mon approche actuelle.

Dans mon travail de thèse, la dimension k des matrices W et H a été choisie grâce à un compromis entre trois critères. Il serait intéressant d'étudier de manière plus approfondie un critère de sélection de k qui reposerait sur la comparaison des CTFA estimés pour k et $k + 1$ de manière similaire à la comparaison entre H^4 et H^{15} dans le Chapitre 5.

Options de modélisation

Étendre l'approche par contrainte. J'ai développé au cours de mon travail, une contrainte associée aux abondances de marqueurs fonctionnels en m'inspirant des voies de la métabolisation des mono-saccharides. Le principe en est de sélectionner des marqueurs fonctionnels

spécifiques qui sont représentés comme les arêtes d'un graphe topologique dont les sommets sont des métabolites. Les arêtes du graphe, c'est à dire le rôle de ces marqueurs fonctionnels, sont orientées dans le sens du catabolisme. Ceci permet de définir pour chaque métabolite, une liste de marqueurs associés à la consommation et la production du métabolite. Les rapports entre l'abondance des marqueurs produisant et consommant les métabolites sont bornés au moyen de 190 génomes de référence ce qui permet de définir une contrainte.

La voie de la sulfato-réduction, qui consomme de façon syntrophique du H_2 dans l'écosystème intestinal humain, est en liens directs avec le catabolisme des polysaccharides et peut être ajoutée de façon assez immédiate. De même j'envisage d'étendre cette approche à d'autres voies métaboliques telle que le catabolisme des protéines. Ceci nécessite toutefois un travail de sélection de marqueurs très lourd.

Pendant les trois années qu'ont duré cette thèse, de nouveaux génomes d'espèces du microbiote intestinal humain ont été séquencés, en particulier par le projet HMP. La prise en compte de ces génomes dans la construction des contraintes pourrait améliorer notre approche.

Comme nous l'avons au Chapitre 3, il est possible de construire des contraintes sur les CTFA entre les abondances de couples de marqueurs à partir des 190 génomes considérés. Une perspective intéressante serait de ne plus se limiter à un métabolisme particulier mais d'exploiter tous les marqueurs fonctionnels de KEGG présents dans ces génomes pour construire de façon systématique des contraintes. J'envisage de ne contraindre que les abondances de marqueurs associés à un métabolite commun. Ce travail peut se faire de façon automatique puisque la base de donnée KEGG met en liens marqueurs fonctionnels, réactions et métabolites. Dans cette approche je propose de s'affranchir du travail de sélection de marqueurs spécifiques et de l'orientation des réactions. En contrepartie cette démarche donne une importance beaucoup plus grande à la structure des génomes considérés. L'apprentissage sur un ensemble de génomes représentatifs est alors d'autant plus nécessaire. Des efforts de séquençage importants ont été réalisés récemment, cette limitation est donc amenée à disparaître.

Mon approche peut être utilisée dans d'autres écosystèmes à condition de disposer de données suffisantes pour calibrer les contraintes.

Exploiter une information taxonomique. Dans ce travail, j'ai utilisé une information purement fonctionnelle. Sans ajout de données supplémentaires il est possible d'introduire une description taxonomique dans la présente approche. En effet, le catalogue IGC contient plusieurs annotations et notamment une information taxonomique pour certains gènes. À titre d'exemple admettons que la description au niveau du genre soit disponible, soit S le nombre de

genres différents considérés, l'abondance des 86 marqueurs fonctionnels pourrait être subdivisée en S origines. Ceci multiplierait alors par S le nombre de marqueurs fonctionnels et permettrait potentiellement d'associer des profils taxonomiques aux CTFA. Des pénalisations encourageant la parcimonie en genre dans un même CTFA sont envisageables. Ces approches sont potentiellement très intéressantes mais dépendent fortement de la qualité de l'annotation initiale de l'IGC.

La possibilité d'obtenir des données supplémentaires telles que des données d'ADNr 16S, rendrait envisageable de nouvelles approches. Pour rappel, les abondances d'ADNr 16S donnent une information sur la diversité et l'abondance des OTU dans un échantillon. Notons $A^{16S} \in \mathcal{M}_{1408,s}(\mathbb{R}^+)$ la matrice hypothétique de données d'ADNr 16S associée aux abondance des s OTU ou autre classe taxonomique (genre, espèce, etc ...) considérée. Dans les exemples suivant je prend s comme étant un nombre d'OTU.

Une première piste serait de s'inspirer de la démarche proposée par Jiang et al. (2014) et de considérer que l'information taxonomique est complémentaire de l'information métagénomique. Il s'agit d'utiliser l'approche complète de NMF sous contrainte décrite dans ce travail de thèse, appliquée aux matrices A et A^{16S} concaténées en une unique matrice $A^{Tot} \in \mathcal{M}_{1408,86+s}(\mathbb{R}^+)$.

Le résultat en serait une matrice H de dimension $k \times (86 + s)$ caractérisant les CTFA et leur composition taxonomique et une matrice W des abondances de commune des CTFA et OTU. Le processus de détermination de la dimension k serait à nouveau nécessaire. Mon approche se distingue de celle de Jiang et al. (2014) par l'utilisation d'information biologique incorporée dans les contraintes.

Une deuxième possibilité serait de rechercher des regroupements de marqueurs fonctionnels qui co-varient avec les abondances 16S. L'idée de départ serait de résoudre un problème du type

$$S^{16S} = \underset{\mathbf{X} \geq 0, F_{\Delta} \mathbf{X}^T \leq 0}{\operatorname{argmin}} \|A - A^{16S} \mathbf{X}\|_F^2, \quad (5.13)$$

où $S^{16S} \in \mathcal{M}_{S,86}(\mathbb{R}^+)$ est une matrice de 86 marqueurs fonctionnels associée aux s OTU. Cette piste est une tentative d'élucidation des marqueurs fonctionnels associés aux OTU. La contrainte que j'ai construite doit permettre d'apporter suffisamment d'information pour que les marqueurs fonctionnels correspondant à chaque OTU puissent être vraisemblables biologiquement. Une difficulté prévisible est le nombre limité d'échantillons et de marqueurs fonctionnels devant le nombre d'OTU. Si seuls 86 marqueurs sont considérés, au maximum le rang de la matrice S^{16S} sera de 86 et de ce fait il ne sera pas possible de décrire plus de 86 OTU différents. De plus, A^{16S}

sera selon toute vraisemblance parcimonieuse, il se peut que certains OTU ne soient présent que dans très peu d'échantillons parmi les 1408. Ce défaut réduit la capacité à évaluer les marqueurs associés aux OTU. Il est toutefois possible de chercher les marqueurs correspondant plutôt à des groupes d'OTU rassemblés par exemple par distance entre séquences d'ADNr 16S.

Problématique biologiques

Approche de NMF supervisée. Dans le travail présenté jusqu'ici, le processus de construction des CTFA à été mené de façon totalement non supervisée, sans tenir compte des métavariabiles disponibles sur les échantillons. Il serait possible d'utiliser l'approche développée dans ce travail de thèse dans une optique supervisée où l'on chercherait à construire des CTFA qui discriminent des groupes d'échantillons. Ce modèle peut être utilisé pour caractériser différentes populations en exploitant les métavariabiles dans le processus de construction des CTFA. Par exemple dans le cas de la maladie de Crohn, il serait possible de créer un CTFA dont on encourage l'association aux échantillons de cette pathologie par le biais d'une pénalisation structurée (Cho and Saul, 2011). De même nous pourrions essayer de construire des CTFA permettant de mettre en évidence les différences hommes/femmes, ou les différences d'IMC.

Validation du modèle. Une méthode simple de validation consisterait à utiliser la matrice de CTFA obtenue pour décrire de nouveaux échantillons métagénomiques, de qualité comparable, de manière similaire à la Section 5.3.2. L'erreur de reconstruction des nouvelles données serait un critère permettant de valider le modèle.

Il serait aussi intéressant d'adopter une approche de validation en laboratoire à caractère plus biologique. Je propose de construire des assemblages synthétiques correspondants aux CTFA inférés. Cette idée nécessite réflexion puisqu'elle pose le défi de pouvoir associer des espèces cultivables à chaque CTFA. En admettant que cette étape soit réalisée, il serait alors possible de construire des microbiotes artificiels comme mélanges d'espèces associées aux CTFA. Ils pourraient être testés chez des souris axéniques et l'existence des assemblages et leur permanence dans le temps pourraient être évalués. Des perturbations par changements de régime alimentaire sont aussi envisageables. La variation jointe des espèces associées aux CTFA serait alors une validation biologique des CTFA trouvés et du concept d'assemblage fonctionnel.

Annexes

A Valorisation scientifiques

European Conferance on Mathematical and Theorical Biology (ECMTB). Présentation orale de 20 minute : Metagenomic data analysis and integration in a functional population model of fiber degradation by the human intestinal microbiota.

Mathematical Models in Ecology and Evolution (MMEE). Présentation d'un poster : Modeling Fiber Degradation by Human Gut Microbiome.

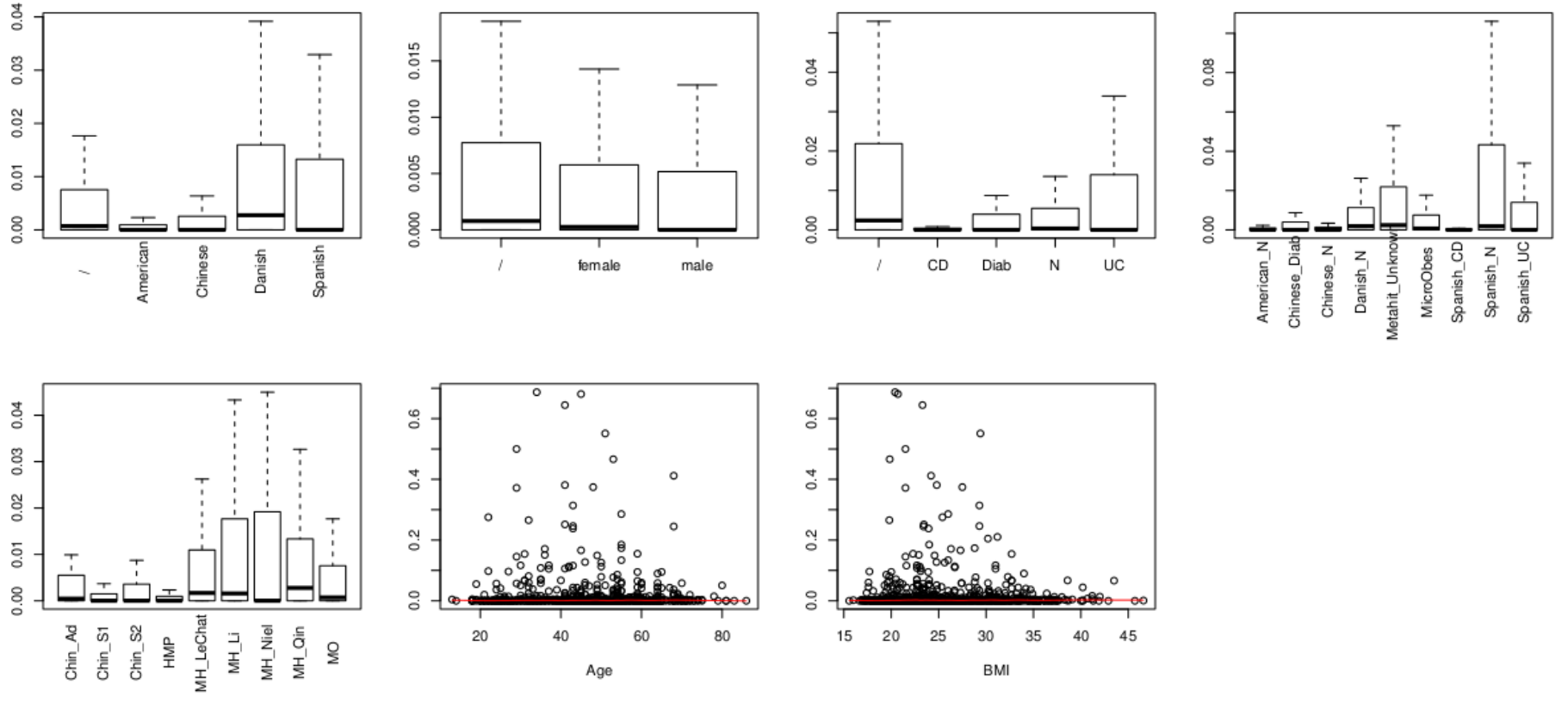
Foundations of Systems Biology in Engineering (FOSBE). Article de conférence accepté. Présentation d'un poster : Modeling Fiber Degradation by Human Gut Microbiome.

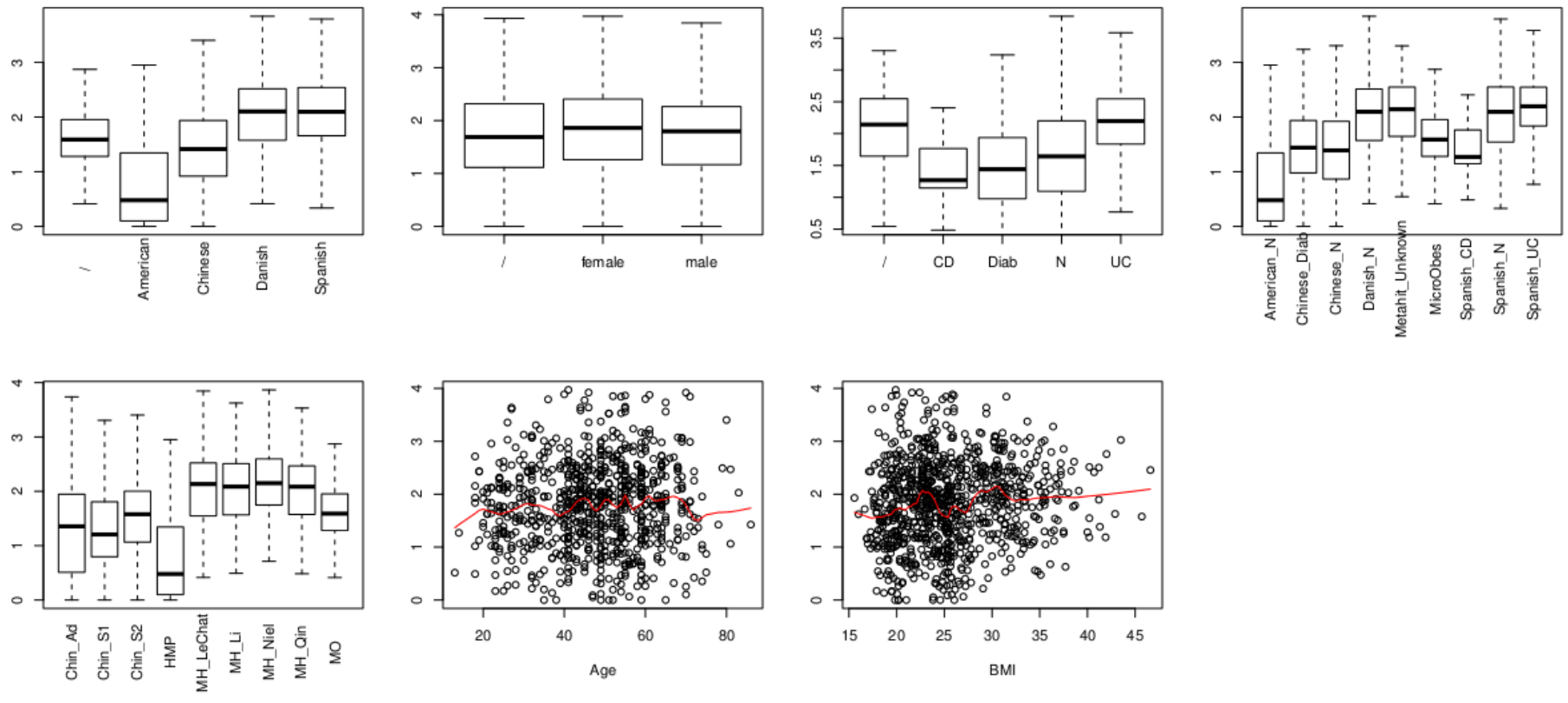
Plos Computational Biology. Un article soumis, actuellement en révision. L'intégralité de l'article est présenté en Annexe D.

B Abondance des CTFA pour chaque catégorie de méta-variable

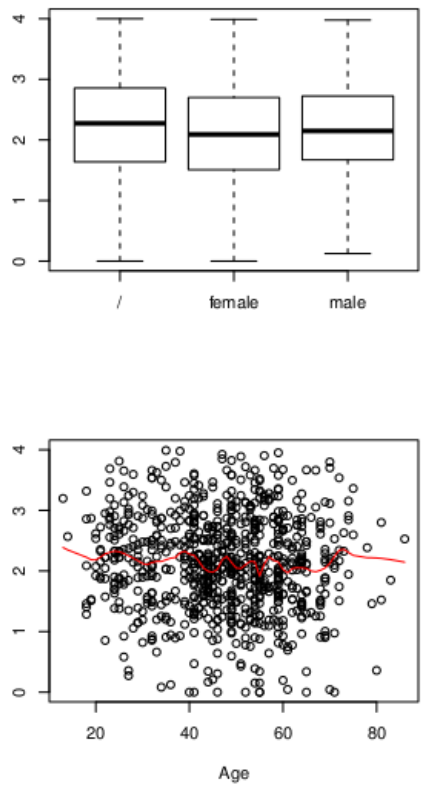
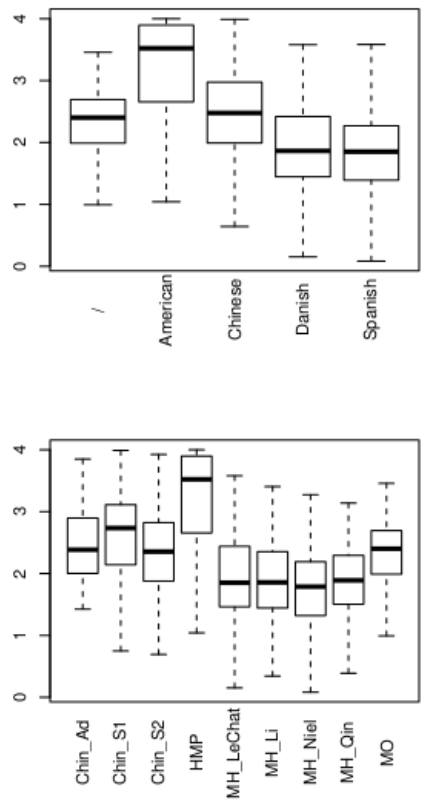
L'abondance des CTFA pour chaque catégorie disponible est représentée sous forme de diagrammes en boîtes. Nous rappelons que Diab signifie Diabétique, N correspond à Non malade, CD est l'abréviation de Crohn Disease et UC de Colite Ulcéreuse. De même les études associées à chaque abréviation sont les suivantes , HMP Methé et al. (2012), MH_LeChat Le Chatelier et al. (2013), MH_Li Li et al. (2014), MH_Niel Nielsen et al. (2014), MH_Qin Qin et al. (2010) et MO Cotillard et al. (2013). Enfin, les échantillons chinois provenant de Qin et al. (2012) ont été scindés en trois groupes nommés Chin_Ad, Chin_S1 et Chin_S2 relatifs à différents lots d'analyse dans l'étude.

CTFA 1

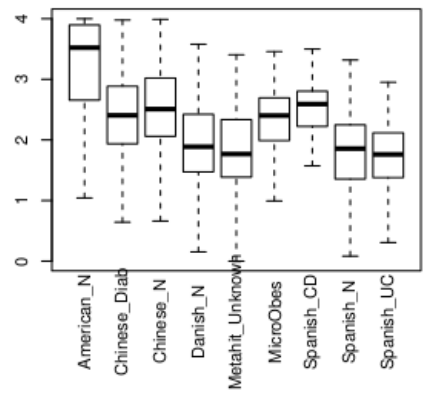
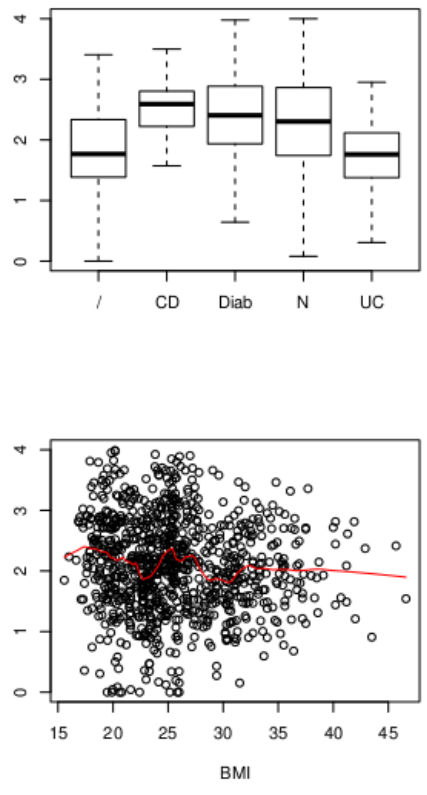


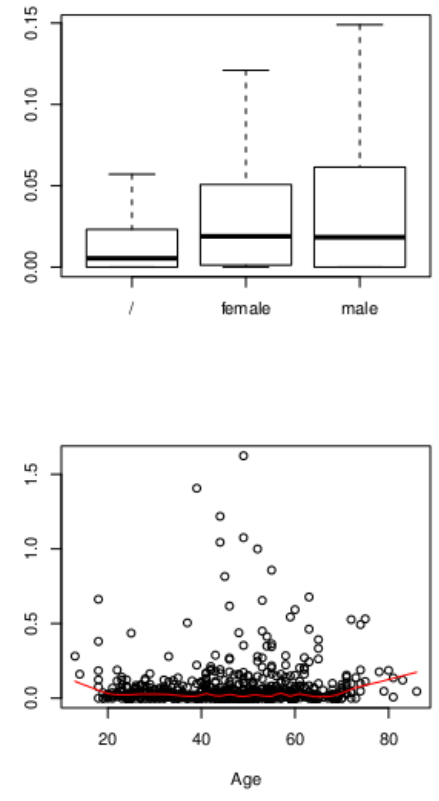
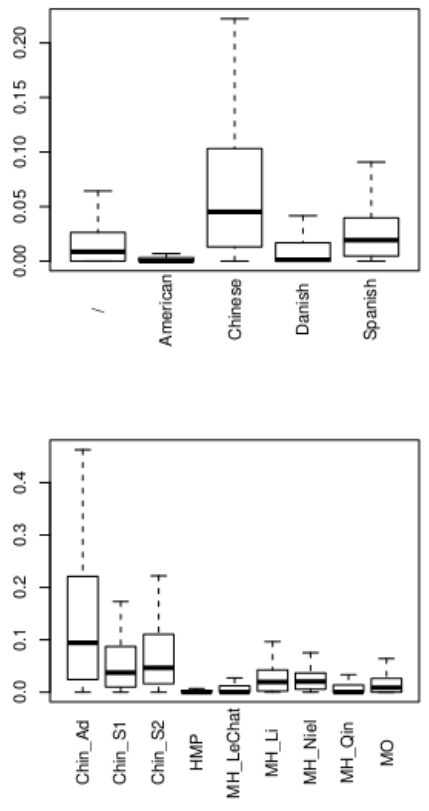


CTFA 2

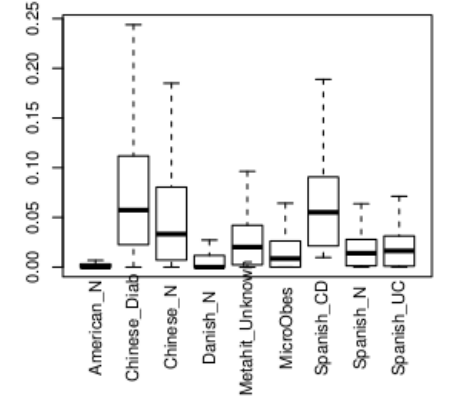
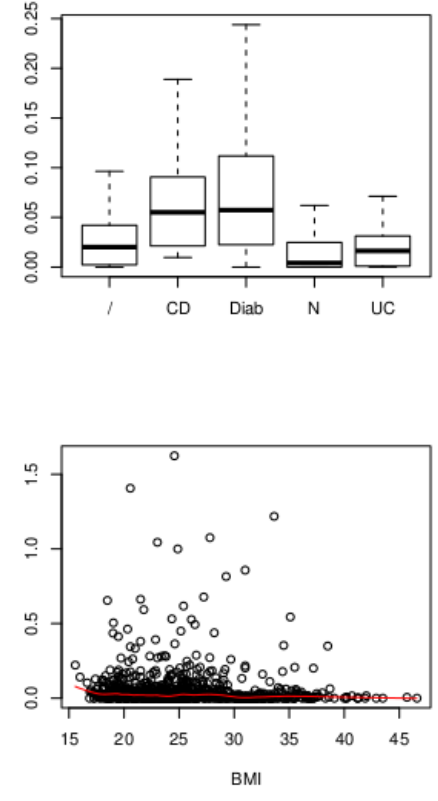


CTFA 3





CTFA 4



C Liste des 190 génomes choisis

Le choix des génomes se fait principalement en examinant les publications Tap et al. (2009) et Qin et al. (2010), nous symbolisons l'origine du génomes choisi par un 1 ou un 0.

Tap & al 2009	Qin & al 2010	id taxonomique	Nom
0	1	563191	Acidaminococcus sp. D21
0	1	1203555	Acidaminococcus sp. HPA0509
0	1	349741	Akkermansia muciniphila ATCC BAA-835
1	0	445970	Alistipes putredinis DSM 17216
1	0	717959	Alistipes shahii WAL 8301
1	0	445972	Anaerotruncus colihominis DSM 17241
1	0	411901	Bacteroides caccae ATCC 43185
1	0	997873	Bacteroides caccae CL03T12C61
0	1	997874	Bacteroides cellulosilyticus CL02T12C19
0	1	537012	Bacteroides cellulosilyticus DSM 14838
0	1	470145	Bacteroides coprocola DSM 17136
0	1	547042	Bacteroides coprophilus DSM 18228 = JCM 13818
1	0	556260	Bacteroides dorei 5_1_36/D4
1	0	997875	Bacteroides dorei CL02T00C15
1	0	997876	Bacteroides dorei CL02T12C06
1	0	997877	Bacteroides dorei CL03T12C01
1	0	483217	Bacteroides dorei DSM 17855
1	0	665953	Bacteroides eggerthii 1_2_48FAA
1	0	483216	Bacteroides eggerthii DSM 20697
1	0	997888	Bacteroides finegoldii CL09T03C10
1	0	483215	Bacteroides finegoldii DSM 17565
1	0	457424	Bacteroides fragilis 3_1_12
1	0	997880	Bacteroides fragilis CL05T00C42
1	0	997881	Bacteroides fragilis CL05T12C13
1	0	1073386	Bacteroides fragilis HMW 610
1	0	1073388	Bacteroides fragilis HMW 616
1	0	471870	Bacteroides intestinalis DSM 17393
1	0	1121098	Bacteroides massiliensis B84634 = Timone 84634 = DSM 17679 = JCM 13223

1	0	1235788	<i>Bacteroides massiliensis</i> dnLKV3
1	0	665954	<i>Bacteroides ovatus</i> 3_8_47FAA
1	0	411476	<i>Bacteroides ovatus</i> ATCC 8483
1	0	997885	<i>Bacteroides ovatus</i> CL02T12C04
1	0	997886	<i>Bacteroides ovatus</i> CL03T12C18
1	0	702443	<i>Bacteroides ovatus</i> SD CMC 3f
0	1	484018	<i>Bacteroides plebeius</i> DSM 17135
1	0	457388	<i>Bacteroides</i> sp. 2_1_7
1	0	469590	<i>Bacteroides</i> sp. 2_2_4
1	0	457394	<i>Bacteroides</i> sp. 4_3_47FAA
1	0	457395	<i>Bacteroides</i> sp. 9_1_42FAA
1	0	556258	<i>Bacteroides</i> sp. D1
1	0	449673	<i>Bacteroides stercoris</i> ATCC 43183
1	0	1073351	<i>Bacteroides stercoris</i> CC31F
1	0	226186	<i>Bacteroides thetaiotaomicron</i> VPI-5482
1	0	1235785	<i>Bacteroides thetaiotaomicron</i> dnLKV9
1	0	411479	<i>Bacteroides uniformis</i> ATCC 8492
1	0	997889	<i>Bacteroides uniformis</i> CL03T00C23
1	0	997890	<i>Bacteroides uniformis</i> CL03T12C37
1	0	1235787	<i>Bacteroides uniformis</i> dnLKV2
1	0	435590	<i>Bacteroides vulgatus</i> ATCC 8482
1	0	997891	<i>Bacteroides vulgatus</i> CL09T03C04
1	0	702446	<i>Bacteroides vulgatus</i> PC510
1	0	1235786	<i>Bacteroides vulgatus</i> dnLKV7
1	0	997892	<i>Bacteroides xylanisolvens</i> CL03T12C04
1	0	702447	<i>Bacteroides xylanisolvens</i> SD CC 1b
1	0	702444	<i>Bacteroides xylanisolvens</i> SD CC 2a
1	0	657309	<i>Bacteroides xylanisolvens</i> XB1A
0	1	367928	<i>Bifidobacterium adolescentis</i> ATCC 15703
0	1	411481	<i>Bifidobacterium adolescentis</i> L2-32
0	1	703613	<i>Bifidobacterium animalis</i> subsp. <i>animalis</i> ATCC 25527
0	1	911598	<i>Bifidobacterium animalis</i> subsp. <i>lactis</i> BS 01
0	1	486409	<i>Bifidobacterium animalis</i> subsp. <i>lactis</i> HN019
0	1	1207543	<i>Bifidobacterium bifidum</i> IPLA 20015

0	1	398513	Bifidobacterium bifidum NCIMB 41171
0	1	883062	Bifidobacterium bifidum S17
0	1	1036735	Bifidobacterium breve CECT 7263
0	1	1074482	Bifidobacterium breve DPC 6330
0	1	518634	Bifidobacterium breve DSM 20213 = JCM 1192
0	1	1261072	Bifidobacterium breve JCP7499
0	1	326426	Bifidobacterium breve UCC2003
0	1	566552	Bifidobacterium catenulatum DSM 16992 = JCM 1194
0	1	871562	Bifidobacterium dentium ATCC 27679
0	1	401473	Bifidobacterium dentium Bd1
0	1	553191	Bifidobacterium dentium JCVIHMP022
1	0	391904	Bifidobacterium longum subsp. infantis ATCC 15697 = JCM 1222
1	0	1161744	Bifidobacterium longum subsp. longum 1-6B
1	0	1161745	Bifidobacterium longum subsp. longum 2-2B
1	0	1161904	Bifidobacterium longum subsp. longum 35B
1	0	722911	Bifidobacterium longum subsp. longum F8
0	1	547043	Bifidobacterium pseudocatenulatum DSM 20438 = JCM 1200
1	0	537007	Blautia hansenii DSM 20583
1	0	511680	Butyrivibrio crossotus DSM 2876
0	1	451640	Catenibacterium mitsuokai DSM 15897
0	1	469595	Citrobacter sp. 30_2
0	1	1173691	Citrobacter sp. A1
1	0	518636	Clostridium asparagiforme DSM 15981
0	1	445973	Clostridium bartlettii DSM 16795
0	1	997893	Clostridium bolteae 90A5
0	1	997894	Clostridium bolteae 90A9
0	1	997895	Clostridium bolteae 90B3
0	1	997896	Clostridium bolteae 90B7
0	1	411902	Clostridium bolteae ATCC BAA-613
1	0	742735	Clostridium clostridioforme 2_1_49FAA
1	0	999403	Clostridium clostridioforme 90A1
1	0	999404	Clostridium clostridioforme 90A3

1	0	999407	<i>Clostridium clostridioforme</i> 90A7
1	0	999409	<i>Clostridium clostridioforme</i> 90B1
1	0	428125	<i>Clostridium leptum</i> DSM 753
0	1	537013	<i>Clostridium methylpentosum</i> DSM 5476
1	0	500632	<i>Clostridium nexile</i> DSM 1787
1	0	411468	<i>Clostridium scindens</i> ATCC 35704
1	0	457396	<i>Clostridium</i> sp. 7_2_43FAA
1	0	411489	<i>Clostridium</i> sp. L2-50
1	0	411486	<i>Clostridium</i> sp. M62/1
1	0	411484	<i>Clostridium</i> sp. SS2/1
1	0	1042156	<i>Clostridium</i> sp. SY8519
1	0	411903	<i>Collinsella aerofaciens</i> ATCC 25986
0	1	469596	<i>Coprobacillus</i> sp. 29_1
0	1	665941	<i>Coprobacillus</i> sp. 3_3_56FAA
0	1	469597	<i>Coprobacillus</i> sp. 8_2_54BFAA
0	1	556262	<i>Coprobacillus</i> sp. D6
0	1	556270	<i>Coprobacillus</i> sp. D7
1	0	470146	<i>Coprococcus comes</i> ATCC 27758
1	0	411474	<i>Coprococcus eutactus</i> ATCC 27759
0	1	411464	<i>Desulfovibrio piger</i> ATCC 29098
1	0	742765	<i>Dorea formicigenerans</i> 4_6_53AFAA
1	0	411461	<i>Dorea formicigenerans</i> ATCC 27755
1	0	411462	<i>Dorea longicatena</i> DSM 13814
0	1	888066	<i>Enterococcus casseliflavus</i> ATCC 12755
0	1	565655	<i>Enterococcus casseliflavus</i> EC20
0	1	565652	<i>Enterococcus casseliflavus</i> EC30
0	1	565650	<i>Enterococcus faecalis</i> D6
0	1	565643	<i>Enterococcus faecalis</i> DS5
1	0	1157352	<i>Enterococcus faecalis</i> EnGen0111
1	0	1157353	<i>Enterococcus faecalis</i> EnGen0119
1	0	1157382	<i>Enterococcus faecalis</i> EnGen0198
1	0	749501	<i>Enterococcus faecalis</i> TX0102
1	0	491074	<i>Enterococcus faecalis</i> TX0104
0	1	565656	<i>Enterococcus faecium</i> 1,230,933
0	1	565663	<i>Enterococcus faecium</i> Com15

0	1	585035	<i>Escherichia coli</i> S88
0	1	439855	<i>Escherichia coli</i> SMS-3-5
0	1	364106	<i>Escherichia coli</i> UTI89
0	1	469598	<i>Escherichia</i> sp. 3_2_53FAA
0	1	518637	<i>Eubacterium bifforme</i> DSM 3989
1	0	515620	<i>Eubacterium eligens</i> ATCC 27750
1	0	411469	<i>Eubacterium hallii</i> DSM 3353
1	0	1256908	<i>Eubacterium ramulus</i> ATCC 29099
1	0	515619	<i>Eubacterium rectale</i> ATCC 33656
1	0	657318	<i>Eubacterium rectale</i> DSM 17629
1	0	657317	<i>Eubacterium rectale</i> M104/1
1	0	657319	<i>Eubacterium siraeum</i> 70/3
1	0	428128	<i>Eubacterium siraeum</i> DSM 15702
1	0	717961	<i>Eubacterium siraeum</i> V10Sc8a
1	0	411463	<i>Eubacterium ventriosum</i> ATCC 27560
1	0	748224	<i>Faecalibacterium</i> cf. <i>prausnitzii</i> KLE1255
1	0	411483	<i>Faecalibacterium prausnitzii</i> A2-165
1	0	718252	<i>Faecalibacterium prausnitzii</i> L2-6
1	0	411485	<i>Faecalibacterium prausnitzii</i> M21/2
1	0	657322	<i>Faecalibacterium prausnitzii</i> SL3/3
0	1	657308	<i>Gordonibacter pamelaetae</i> 7-10-1-b
1	0	545696	<i>Holdemania filiformis</i> DSM 12042
0	1	507522	<i>Klebsiella pneumoniae</i> 342
0	1	478749	<i>Marvinbryantia formatexigens</i> DSM 14469
0	1	420247	<i>Methanobrevibacter smithii</i> ATCC 35061
0	1	521002	<i>Methanobrevibacter smithii</i> DSM 2374
0	1	483214	<i>Methanobrevibacter smithii</i> DSM 2375
0	1	500635	<i>Mitsuokella multacida</i> DSM 20544
1	0	435591	<i>Parabacteroides distasonis</i> ATCC 8503
1	0	999416	<i>Parabacteroides distasonis</i> CL03T12C09
1	0	999417	<i>Parabacteroides distasonis</i> CL09T03C24
1	0	999419	<i>Parabacteroides johnsonii</i> CL02T12C29
1	0	537006	<i>Parabacteroides johnsonii</i> DSM 18315
1	0	411477	<i>Parabacteroides merdae</i> ATCC 43184
1	0	999420	<i>Parabacteroides merdae</i> CL03T12C32

1	0	999421	Parabacteroides merdae CL09T00C40
0	1	537011	Prevotella copri DSM 18205
1	0	411467	Pseudoflavonifractor capillosus ATCC 29799
1	0	536231	Roseburia intestinalis L1-82
1	0	657315	Roseburia intestinalis M50/1
1	0	718255	Roseburia intestinalis XB6B4
1	0	622312	Roseburia inulinivorans DSM 16841
1	0	657321	Ruminococcus bromii L2-63
1	0	411470	Ruminococcus gnavus ATCC 29149
1	0	1073375	Ruminococcus gnavus CC55_001C
1	0	471875	Ruminococcus lactaris ATCC 29176
1	0	1073376	Ruminococcus lactaris CC59_002D
1	0	657314	Ruminococcus obeum A2-162
1	0	411459	Ruminococcus obeum ATCC 29174
1	0	457412	Ruminococcus sp. 5_1_39BFAA
1	0	1095771	Ruminococcus sp. JC304
1	0	657323	Ruminococcus sp. SR1/5
1	0	411460	Ruminococcus torques ATCC 27756
1	0	657313	Ruminococcus torques L2-14
1	0	1051074	Streptococcus thermophilus JIM 8232
1	0	322159	Streptococcus thermophilus LMD-9
1	0	264199	Streptococcus thermophilus LMG 18311
1	0	1187956	Streptococcus thermophilus MN-ZLW-002
1	0	767463	Streptococcus thermophilus ND03
1	0	411471	Subdoligranulum variabile DSM 15176
1	0	483218	[Bacteroides] pectinophilus ATCC 43243

D Article soumis à Plos Computational Biology



Inferring Aggregated Functional Traits from Metagenomic Data Using Constrained Non-negative Matrix Factorization: Application to Fiber Degradation in the Human Gut Microbiota - Submission to PLOS Journals

Raguideau Sébastien^{1,2}, Plancade Sandra², Pons Nicolas³, Leclerc Marion¹, Laroche Béatrice^{2,*}

1 PhylHom Team, Micalis Institute, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France

2 MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France

3 MGP, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France

* Beatrice.Laroche@jouy.inra.fr

Abstract

Whole Genome Shotgun (WGS) metagenomics are increasingly used to study the structure and functions of complex microbial ecosystems, both from the taxonomic and functional point of view. Gene inventories of otherwise uncultured microbial communities make the direct functional profiling of microbial communities possible. The concept of community aggregated trait has been adapted from environmental and plant functional ecology to the framework of microbial ecology. Community aggregated trait are quantified from WGS data by computing the abundance of relevant marker genes. They can be used to study key processes at the ecosystem level and correlate environmental factors and ecosystem functions. In this paper we propose a novel model based approach to infer combinations of aggregated traits characterizing specific ecosystemic metabolic processes. We formulate a model of these Combined Aggregated Functional Traits (CAFTs) accounting for a hierarchical structure of genes, which are associated on microbial genomes, further linked at the ecosystem level by complex co-occurrences or interactions. The model is completed with constraints specifically designed to exploit available genomic information, in order to favor biologically relevant CAFTs. The CAFTs structure, as well as their intensity in the ecosystem, is obtained by solving a constrained Non-negative Matrix Factorization (NMF) problem. We developed a multicriteria selection procedure for the number of CAFTs. We illustrated our method on the modelling of ecosystemic functional traits of fiber degradation by the human gut microbiota. We used 1408 samples of gene abundances from several high-throughput sequencing projects and found that four CAFTs only were needed to represent the fiber degradation potential. This data reduction highlighted biologically consistent functional patterns while providing a high quality preservation of the original data. Our method is generic and can be applied to other metabolic processes in the gut or in other ecosystems.

Author Summary

Microbial communities are highly complex ecosystems, harboring a high taxonomic diversity, rapidly varying in time. Focusing on the functional traits of microorganisms

1

2

instead of species offers an interesting and complementary insight since quite often, 3
owing to environmental selection pressure, traits are less variable in time and shared by 4
many species. By providing gene inventories of otherwise uncultured microbial 5
communities, Whole Genome Shotgun (WGS) metagenomics made the functional 6
profiling of microbial communities possible, since community aggregated traits can be 7
quantified by the abundance of relevant marker genes. Considering a global metabolic 8
process in an ecosystem, we propose a method that exploits gene abundance data to 9
determine combinations of functional traits characterizing this process. When applied to 10
fiber degradation in the human gut microbiota, the method shows that only four 11
complex functional traits are needed to characterize this process. The approach is 12
generic and could be applied to other processes or ecosystems. It allows summarizing 13
complex datasets by a limited number of biologically interpretable functional patterns. 14

Introduction

Whole Genome Shotgun (WGS) metagenomics are increasingly used to study the structure and functions of complex microbial ecosystems. Thanks to a huge research effort, these high-throughput approaches are constantly better mastered and have already provided valuable information.

Recent achievements [1–3] for analyzing metagenomic reads have focused on genome or species reconstruction. They constitute an alternative to amplicon sequencing approaches such as 16S rDNA and provide potentially more robust tools for community structure assessment.

As a complement to these taxonomy oriented approaches, WGS metagenomics allow the direct functional profiling of microbial communities through gene inventories of otherwise uncultured microbial communities. The functional annotation of these genes, as well as the evaluation of their abundances requires the mapping of metagenomic reads on one or several functional databases which may include annotated, de novo assembled gene catalogs. It is a challenging task for which bioinformatic tools are still being developed (see [4] for a very recent review).

As early as 2005, Tringe et al. [5] demonstrated that functional binning could discriminate environments and determine the functional potential of microbial communities. Since then, several other groups used functional data to shed light on the processes taking place in the ecosystem, even in the absence of additional expression data (see e.g. [6,7]), and developed frameworks to functionally characterize and compare ecosystems [4,8]. Investigating complex microbial communities from a purely functional point of view provides an interesting insight because their taxonomic diversity may be very high and varying in time while, owing to environmental selection pressure, their ecosystemic functions are often ubiquitous and much more stable in time. The functional profiling of microbial communities has motivated the adaptation of concepts borrowed from environmental and plant functional ecology to the framework of microbial ecology, such as community aggregated traits [9,10], which can be quantified through abundance computation of specific genes or predefined pathways from WGS data. These taxon-free approaches allow studying key processes at the ecosystem level and correlate environmental factors and ecosystem functions [11].

Choosing relevant descriptors for ecosystem functions raises several issues. Using arbitrarily predefined pathways helps simplifying the systems description by creating combined descriptors, however it may not reflect all the information in the data. We propose a novel model based approach to overcome these issues. We focus on specific ecosystemic metabolic processes, described by a possibly high number of genes or functional markers abundances. Our aim is to infer a limited number of combined functional traits characterizing these processes, as well as their intensity in the ecosystem.

Our approach relies on Non-negative Matrix Factorization (NMF), a popular machine learning technique in data and image analysis. It was previously used for genomic data mining in the context of microarray data analysis [12], [13]. More recently, it was introduced in the context of metagenomics for reads binning [14], or analyzing datasets from various ecosystems. In particular Jiang and co-authors used this approach as a "soft" clustering tool in different ecosystems. They studied pathway abundances in diverse environmental ecosystems [15], compared habitats in marine ecosystems using protein families profiles [16] and human body sites using phylogenetic and functional data [17].

The originality of our work lies in the design of a constrained NMF approach that allows selecting biologically relevant combined functional traits to describe these processes, and a careful multicriteria selection procedure.

As a proof of concept, we applied our approach to the modelling of ecosystemic functional traits of fiber degradation by the human gut microbiota based on 1408 samples of gene abundances.

Results

The Model

Combined Aggregated Functional Traits. Quantitative metagenomics allow to study metabolic processes at the ecosystem level, by producing metagenes abundances with functional annotation. Taking advantage of this information, we consider a metabolic process occurring in an ecosystem. We assume that this process is described

by a list of biochemical reactions, each associated to what we call a *functional marker*.
 A functional marker is a group of genes, or modular elements in genes, able to control
 the production of enzymes involved in a reaction. Typical functional markers are Kegg
 Orthologies (KO). The abundance of a functional marker is defined as the sum of the
 abundances of all the metagenes identified as belonging to the group.

Our first modelling hypothesis is the existence of patterns underlying functional
 markers abundances in the ecosystem. Indeed, within a microbial ecosystem, genes can
 be looked at in a hierarchical fashion. They are first associated on microbial genomes,
 which are further grouped at the ecosystem level into subpopulations of microorganisms.
 These subpopulations involve hundreds of different bacterial species linked by complex
 co-occurrences or interactions. Moreover they are influenced by environmental factors
 (e.g. nutrients or temperature). The set of functional marker abundances within each of
 these subpopulations form a characteristic pattern in the ecosystem.

This hypothesis is illustrated in the simplistic example of metabolic process in Fig. 1.
 Here, the observed functional marker abundances originate from three subpopulations,
 resulting in three patterns. The ecological interpretation of the first pattern (green box)
 could be a trophic chain in which metabolite X is produced by bacterial species of type
 1 and used by bacterial species of type 3 as well as an association between bacterial
 species of type 1 and 2 for other ecological reasons which cannot be elucidated with the
 observation of this particular metabolic process only. Therefore functional marker
 abundances in samples should reflect this structure.

Our second modelling hypothesis is motivated by the fact that even if in many
 microbial communities microbial species composition varies over time and samples,
 metabolic processes at the community level are ubiquitous and much more stable in
 time. Consequently, we assume that the patterns mentioned above are shared by all
 samples of the ecosystem, in variable proportions because they were selected by specific
 environmental constraints. For instance, for the gut ecosystem these include anaerobic
 conditions, temperature or digestion residue composition coming from the host diet. As
 illustrated on Fig. 2, when considering the abundances of functional markers in several
 samples, each of these patterns can be characterized by a line vector of functional
 marker frequencies, which we define to be a Combined Aggregated Functional Trait
 (CAFT), in the meaning that it represents a combination of quantifiable characteristics

(functional markers) of aggregated microbial populations. 107

Following our hypotheses, functional marker abundances in each sample are 108
 modelled as mixtures of underlying CAFTs. For each sample $i = 1, \dots, n$ and marker 109
 $j = 1, \dots, r$, the observed frequency a_{ij} of marker j in sample i is the sum of the 110
 contributions of all CAFTs. This leads to 111

$$a_{ij} \simeq \sum_{l=1}^k w_{il} h_{lj} \quad (1)$$

where $h_l = (h_{l1}, \dots, h_{lr})$ is the characterization in terms of r markers frequencies of the 112
 l^{th} CAFT and $w_i = (w_{i1}, \dots, w_{ik})$ is the vector of the intensities or weights of the k 113
 CAFTs in sample i . 114

Note that each h_l is defined up to a multiplication by a positive constant, provided 115
 that its weights in the samples are scaled accordingly. The ℓ^1 -scaled version $h_l / (\sum_j h_{lj})$ 116
 can be interpreted as the vector of relative frequencies of functional markers within the 117
 subpopulation characterized by trait l . 118

In the rest of the paper, we denote by A the $(n \times r)$ data matrix, H the $(k \times r)$ 119
 matrix whose lines are the h_l and W the $(n \times k)$ matrix whose lines are the w_i . H will 120
 be called the trait matrix and W the weight matrix. 121

Figure 1. Hierarchical structure underlying metagenomic data. We consider 122
 a metabolic process in a microbial ecosystem involving a substrate U, metabolites V, X, 123
 Y, Z, T and the set of reactions $U \rightarrow V$, $V \rightarrow X$, $X \rightarrow Y$, $V \rightarrow Z$ and $U \rightarrow T+X$, 124
 respectively catalyzed by proteins synthesized by genes in KEGG Orthology (KO) 125
 groups **a**, **b**, **c**, **d** and **e**. Gene counts stem from an underlying hierarchical organization. 126
 Genes are associated within bacterial genomes (solid black lines) and through ecosystem 127
 level association patterns (green, red and blue ticked lines). In this example, the green 128
 and blue boxes can be interpreted as trophic chains corresponding to two distinct 129
 pathways for substrate degradation. The red box can be interpreted as an alternative to 130
 the green one, involving different bacterial groups, depending on the host diet or life 131
 history. Note that the red box involves (possibly several) species harbouring two copies 132
 of gene **a** in their genomes.

Figure 2. Combined Aggregated Functional Traits. The gene association 133
 patterns in Fig. 1 correspond to stable structures, representing different ways to fulfill 134
 the metabolic function of interest at the ecosystem level. At the bottom of each box, 135
 vectors h_1 , h_2 and h_3 represent functional markers abundances. We call these vectors 136
 Combined Aggregated Functional Traits (CAFTs). They should be found in all samples, 137
 possibly in varying proportions. Sample 1 is decomposed as $A_1 = 3h_1 + 2h_2 + h_3$ and 138
 sample n as $A_n = h_1 + 2h_2 + 3h_3$. 139

Constraints By definition the entries of the weight and trait matrices W and H are assumed to be non-negative, we denote this condition by

$$W \geq 0, \quad H \geq 0. \tag{2}$$

We formulate additional hypotheses that allow to take advantage of prior knowledge on microbial metabolism and genomics in order to select biologically relevant CAFTs.

First, we assume that a subset of reactions in the metabolic process under consideration are well known, so that a list of metabolites that are consumed or produced by each reaction can be extracted.

Second, we assume that metabolites can be partitioned into intra- and extra-cellular, namely metabolites staying within microbial cells and metabolites that can be exported out of the cell.

Then there is a subset M_1 of intracellular metabolites such that for each m in M_1 , each CAFT h_l should satisfy

$$\sum_{j \text{ producing } m} h_{lj} \leq \delta_{1,m} \sum_{j' \text{ consuming } m} h_{lj'} \tag{3}$$

In the same way, there is a subset M_2 of extracellular metabolites such that for each m in M_2 , each CAFT h_l should satisfy

$$\sum_{j' \text{ consuming } m} h_{lj'} \leq \delta_{2,m} \sum_{j \text{ producing } m} h_{lj} \tag{4}$$

The intuitive idea behind such constraints is the following. Assuming that functional markers are relevant proxies of metabolic functions, it could be expected that if an intracellular metabolite is produced in the subpopulation characterized by a given CAFT it has to be used. This inspires constraint (3), since it imposes that the number of markers associated to reactions that produce a given intracellular metabolite is zero if the number of markers associated to reactions that consume the same intracellular metabolite is zero. Conversely, it could be expected that a metabolite cannot be used if it is not produced, which is enforced by constraint (4). This justifies the absence of constraints imposed on extracellular metabolites as they can be imported or exported

outside bacterial cells in the ecosystem.

However, the set of metabolites and reactions under consideration is defined in the context of a specific metabolic process, given the available knowledge. Thus it may be that in this process, some intracellular metabolites accumulate in the cells. It may also be that the selected reactions ignore metabolic routes, because they are unknown or correspond to other metabolic processes, that could produce or use the metabolites. Consequently, depending on metabolites, either none, one or both constraints (3) and (4) may be implemented. In the cases where both constraints apply, the frequencies of markers associated to production and consumption are both either zero or positive with consumption over production marker frequency ratio lying in $[1/\delta_{1,m}, \delta_{2,m}]$.

The sets M_1 , M_2 as well as the positive constants $\delta_{1,m}$ and $\delta_{2,m}$ can be derived from genomic information on the microorganisms present in the ecosystem as detailed in the Material and Method section. Finally, the set of constraints (3) and (4) can be expressed in matrix notation as

$$F_{\Delta} H^T \leq 0 \tag{5}$$

where Δ is the set of parameters $\Delta = \{\delta_{1,m}, m \in M_1\} \cup \{\delta_{2,m}, m \in M_2\}$ and F_{Δ} is a coefficient matrix depending on Δ .

Inference of CAFT and weights

Nonnegative Matrix Factorization formulation The determination of matrices W and H which satisfy (1) and (2) can be formulated as an approximate Nonnegative Matrix Factorization (NMF) problem (see e.g. [18]). The mathematical formulation of the problem requires first to choose a dissimilarity measure for the approximate factorization (1). In our algorithm, we considered the Frobenius norm. Moreover, as functional markers may display different order of magnitude, a scaling per functional marker is operated by considering the ℓ^2 -columnwise normalization of A , denoted \tilde{A} . Therefore, the dissimilarity considered is

$$\|\tilde{A} - W\tilde{H}\|_F^2 = \sum_{i=1}^n \sum_{j=1}^r \left(\tilde{a}_{ij} - \sum_{l=1}^k w_{il} \tilde{h}_{lj} \right)^2 \tag{6}$$

where \tilde{H} is the scaled trait matrix. Matrix H is then recovered by multiplying each column of \tilde{H} by the ℓ^2 -norm of the corresponding column in A .

The NMF algorithms which aim at minimizing this distance require the use of regularizing terms in order to lift identifiability problems and filter noisy data during the minimization process. We consider a regularizing term that promotes sparsity over the columns of \tilde{H} , which encourages markers to be present in a limited number of CAFTs (see Material and Methods).

Finally, we impose the additional constraint (5). Let \tilde{F}_Δ be the scaled matrix F_Δ such that the constraint (5) is equivalent to $\tilde{F}_\Delta \tilde{H}^T \leq 0$. Therefore, the inference of the CAFTs and weights is obtained by solving the following minimisation problem

$$\min_{W, \tilde{H}} \|\tilde{A} - W\tilde{H}\|_F^2 + \alpha(\|W\|_F^2 + \|\mathbf{1}^T \tilde{H}\|_2^2) \quad (7)$$

$$W \geq 0, H \geq 0, \tilde{F}_\Delta \tilde{H}^T \leq 0$$

where $\mathbf{1}$ denotes the vector of length k whose entries are equal to 1.

Selection of parameters. The minimisation problem (7) requires to choose a value for the number of CAFTs k and the tuning parameter α . Parameter selection procedures proposed in most applications of NMF to biological data are based upon stability of the weight matrix W over several initializations of the iterative algorithm used to solve the problem [12,13,15,16], in particular regarding the clustering of samples. Since our use of NMF focuses on extracting reproducible biological mechanisms characterized by the trait matrix H , we are more interested in biological stability than numerical one. Therefore, we propose a new criterion to evaluate the concordance of the CAFTs computed on independent data sets. The reproducibility of H on a new data set is mimicked by repeatedly splitting the set of biological samples, performing the NMF decomposition on each subset and evaluating the concordance between the two trait matrices. We use the concordance index introduced in [16], but our method differs since the authors assessed the reproducibility between various initialisations of the NMF algorithm implemented on the whole data set. To enhance interpretation, this criterion is associated with two more classical ones: change of slope of the reconstruction error and bi-cross-validation error [19]. The selected value is the one which optimizes the

concordance of H while ensuring that the other criteria are acceptable.

194

Impact of the constraint: Toy model

195

The constrained NMF algorithm as well as an unconstrained version were implemented on a toy model in order to illustrate the improvement in terms of CAFT analysis brought by the constraints. We considered a microbial ecosystem in which a metabolic process characterized by a simple 5 reactions graph takes place, each reaction being associated to a functional marker. The metabolic process was assumed to be structured in two CAFTs. The first CAFT characterizes a subpopulation in the ecosystem in which all the functional markers are present in equal proportion, while the second CAFT characterizes a subpopulation in which only three functional markers are present, in equal proportion. The metabolic graph and the true CAFTs are displayed in Fig. 3A. Thus, the true CAFT composition matrix was taken equal to

$$\bar{H} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \end{pmatrix}$$

We built a matrix A for 80 samples assumed to be mixtures of CAFT 1 and CAFT 2. We first generated a matrix \bar{W} of abundances of CAFT 1 and CAFT 2 for the 80 samples, with entries randomly drawn from a uniform distribution, and computed the corresponding marker count matrix $B = \bar{W}\bar{H}$. Noisy counts were generated from B according to the formula

$$a_{ij} = b_{ij}(1 + 0.2\epsilon_{ij})$$

where the ϵ_{ij} are independent, centered Gaussian variables with unit variance.

196

Moreover, we considered the same constraints for each intracellular metabolite V , X and Y , and imposed that the ratio between producing and consuming functional markers lies in $[1/2, 2]$. We compared the CAFTs reconstructed using the constrained (Fig. 3B) and unconstrained (Fig. 3C) NMF inference. The NMF inference with constraints recovers the true CAFTs, while the unconstrained inference fails to provide relevant biological results. Besides, both constrained and unconstrained NMF provide a similar reconstruction of the data, as shown in Fig. 3D. Indeed, from a mathematical point of view, the reconstructed CAFTs with and without constraints span the same

197

198

199

200

201

202

203

204

linear space.

Therefore, while both constrained and unconstrained NMF provide usable results for biological sample analysis, only the constrained version is relevant for CAFT inference.

Figure 3. A toy example to illustrate the impact of the constraints. We consider a metabolic process in a microbial ecosystem, which involves five reactions associated to functional markers **a**, **b**, **c**, **d** and **e**, three extracellular metabolites U , Z and T and three intracellular metabolites V , X and Y . The metabolic process is organized according two CAFTs, displayed in box A. The CAFTs were inferred from 80 samples of functional markers counts, generated from random mixtures of the two CAFTs. The estimated CAFTs are shown, both for the constrained (box B) and unconstrained NMF (box C). Box D shows the reconstructed counts in each sample $(WH)_{ij}$ with constrained (x-axis) and non-constrained (y-axis) NMF.

Application: fiber catabolism CAFT in the human gut microbiota

Data and metabolic process description We exploited gene frequencies in 1408 Whole Genome Shotgun metagenomic samples from 8 distinct studies [3, 20–26] focused on different health status and human populations (Europe, China, USA). For each sample, gene frequencies were obtained by counting sample reads over the Integrated Genome Catalog (IGC) [20] composed of 9.9 million non redundant genes. Functional markers of fiber catabolism were defined from Kegg Orthologies (KO) [27, 28] as well as Glycosides Hydrolases (GH) and Pectin Lyases (PL) families [29]. A total of 86 relevant markers (25 GH-PL and 61 KO) were carefully manually selected as being specifically associated to fiber catabolism in the human gut microbiota. Finally, a 1408×86 matrix of marker frequencies A was derived by summing the corresponding gene frequencies. The graph representing sugar fermentation is displayed on Fig.4. It comprises 43 major metabolites, among which 25 are intracellular and 18 are extracellular. It is oriented according to the biological catabolic pathway from the hydrolysis of fiber leading to simple sugars subsequently fermented to Short Chained Fatty Acid (SCFA) and gas.

For each of the 25 intracellular metabolites, the constraints were calibrated using 190 core gut microbial genomes (see Material and Methods). Both constraints (3) and (4) were implemented together on 17 intracellular metabolites, only one of them on 4 intracellular metabolites, resulting in 38 constraints. The remaining 4 metabolites were left unconstrained.

The overall data generation process and the graph building procedure are detailed in Section Material and Method.

Figure 4. Reaction graph of the biological catabolic pathway from simple sugars to Short Chained Fatty Acids (SCFA) and gas. The 61 edges of the graph correspond to the 61 selected functional markers listed in S1 Table with their associated reaction. The nodes of the graph are the metabolites, red nodes correspond to the 18 extracellular metabolites and blue nodes are the 25 extracellular, their full names are given in S2 Table.

Parameter selection. The concordance index, reconstruction error and 10-fold bi-cross-validation error were computed on a grid of parameters (see Material and Methods). The profiles of the three criteria as a function of the regularization parameter α are displayed in S4 Text, for various numbers of CAFTs k ; the value $\alpha = 0.031$ was selected. Fig. 5 displays the three criteria as a function of k for the selected value of α . The reconstruction error curve undergoes two changes of slope for $k = 4$ and 6. The profile of the concordance of H ensures a good reproducibility of the trait matrix H when less than 4 CAFTs are considered. Moreover, the bi-cross-validation error displays a significant decrease until 4 CAFTs, and a more moderate one for larger values of k ; besides, the variations as k increases are globally weak with respect to the range of values over the splits. In the context of our analysis, in which the main focus is the inference of metabolic traits shared by all samples, the value $k = 4$ was chosen.

Figure 5. Fiber digestion in the human gut: selection of the number of CAFTs k . The profiles of the three criteria as a function of k for the selected value of $\alpha = 0.031$ are displayed. (A) Reconstruction error; (B) 10-fold bi-cross validation error; (C) Concordance of H over repeated splits of the sets of samples. In (B) and (C) the value of the criterion for each split as well as the median are shown.

Analysis of the weight matrix W^* . The constrained NMF decomposition may be looked at a dimension reduction analysis, where biological samples originally characterized by their abundances in 86 functional markers are described by their abundances in the 4 CAFTs displayed in the matrix W^* . We checked that the information on biological samples carried out by A was well preserved in W^* . First of all, the row sums of A quantify the proportion of genes devoted to the 86 fiber digestion reactions in each biological sample; they vary between 0.8% and 1.9% across the 1408

samples. Fig. 6 (A) shows that this proportion is highly correlated to the the total abundance in the 4 CAFTs characterized as the row sums of W^*H^* (Pearson correlation 0.91). Thus, the individual variability of the proportion of fiber digestion related genes is properly recovered by the NMF decomposition.

Moreover, the two-by-two distances between the rows of A characterize the geometric structure of the biological sample population with respect to fiber digestion mechanisms. A similar geometric structure is defined by the rows of W^* . We compared the distances of each sample to the others, respectively in A and in W^* with Pearson correlation. The histogram of these correlations for the 1408 samples points out that all correlations are larger than 0.7, and that 99% are larger than 0.85, which indicates that the similarities between biological samples is well preserved in the reduced spaced induced by the NMF decomposition (Fig. 6 (B)).

Figure 6. Fiber digestion in the human gut: analysis of W^* . (A) For each biological sample, the total abundances in the 4 CAFTs (row sums of W^*H^*) is displayed as a function of the proportion of the 86 fiber digestion reactions in the metagenome (row sums of A). (B) For each sample, the vector of distances to the other samples is computed from the rows of A and of W^* respectively; the histogram of Pearson correlation between these two distance vectors for the 1408 biological samples is displayed.

The CAFTs form coherent pathways. The columns of H^* provide the composition of CAFTs. As an example, we detail here one of the CAFT presented in Fig.7. The 3 main genes encoding for enzyme involved in fibers hydrolysis are Glycosyl hydrolases GH13, GH16 and GH127 (Fig. 7, GH table). The GH13 family is an alpha-amylase, widely represented in human intestinal microorganisms, processing resistant starch and glucose-based polysaccharides. GH16 family members are involved in the breakdown of galactose and glucose based polymers. GH127 (EC 3.2.1.185) family are arabinofuranosidases. A less abundant gene encodes for GH3 enzyme family, broadly present in micro-organisms and performing diversity of functions such as xylanase, glucosidase and arabinofuranosidase. Consistently with the above sugars released from the fiber breakdown (glucose, xylose, galactose, arabinose), the main fermentation substrates are glucose (reactions 4, 3, 31), fucose (22), D-xylose (30) and to a less extent, galactose (19) (Fig. 7). Fermentation further proceeds through the Emben-Meyerhoff-Parnas pathway to pyruvate, yielding the end products short chain

fatty acids, mostly acetate, butyrate (which are the main SCFA measured in healthy human fecal samples [30]). Another major end product is lactaldehyde, resulting from the fermentation of fucose.

The hydrogenotrophic production of methane from H_2 and CO_2 is obtained (reactions 52 to 57). This production of methane from CO_2 (KEGG pathway module M00567) is only restricted to hydrogenotrophic methanogens, known to be present in more than 50% of human adult population [31].

Figure 7. An example of CAFT. The first 61 coordinates of the functional marker frequency vector given by the first column of H^* , associated to simple sugar fermentation, is represented on the reaction graph. The color scale represents percentages of the maximum coordinate among the 61 (reaction 7). The reactions form coherent pathways. The 25 coordinates associated with hydrolysis are presented in the table on the right. The numbers indicate GH families the color are scaled as percentages of the maximum coordinate among the 25 ($GH13$).

282

Discussion

We presented a new NMF based approach to investigate metabolic processes in microbial ecosystems using quantitative metagenomic data. As mentioned in the introduction, NMF techniques were previously used as a "soft" clustering tool to compare samples from various ecosystems, ranging from marine environment to human body ([15], [16] , [17]). However, our work focused on different aspects of NMF. Rather than clustering biological samples, we aimed at extracting biologically relevant features associated to a metabolic process, based on ecological modelling. Therefore our main interest, from the mathematical point of view, was the inference of a biologically interpretable trait matrix H rather than sample analysis through the weight matrix W .

We proposed to model the functional diversity of an habitat-specific microbial ecosystem as a mixture of contributions from subpopulations, each characterized by a profile of functional marker frequencies. These profiles were named Combined Aggregated Functional Traits in reference to [10], since they are measured at the community level from a random sample of microorganisms, regardless of their taxonomic identities. More precisely they result from both a combination of functional markers to form coherent and operative metabolic pathways, and an aggregation of individuals to form a subpopulation characterized by this trait.

Our main contribution was the design of a novel constrained NMF model for CAFT inference. Standard NMF approaches were popularised for biological data analysis mainly because usual dimension reduction techniques such as truncated Singular Value Decomposition (SVD), or the closely related Principal Component Analysis (PCA) often lead to negative coefficients, which do not offer straightforward interpretations. However, while SVD or PCA provide a unique and well defined reduced dataset, optimal in the ℓ^2 -error sense, NMF is an intrinsically ill-posed problem with multiple solutions. Imposing constraints reflecting biological knowledge greatly helps to select a relevant solution. This was illustrated through the toy example in Fig. 3, where both constrained and unconstrained NMF provided solutions leading to the same reconstructed data. However, only the constraints allow the solution that makes biological sense to be picked.

Furthermore, we proposed a method that exploit genomic structure to build relevant

constraints. Genes frequencies are not randomly distributed in the metagenome and result from metabolic association encoded in microbial genomes. Thus, constraints resulting from this metabolic structuration scale up at the community level and should be accounted for in the determination of CAFT. We designed local constraints, only involving adjacent functional markers in the graph representation of the metabolic process. Hence the constraints are not strongly restrictive and still leave a considerable freedom in assembling reactions to build the CAFT matrix, without ever forcing hard reconstruction of pathways. Moreover, our method is flexible since it allows imposing constraints only on parts of the metabolic process where biological knowledge is available and considered reliable.

Moreover, we considered a new approach for selecting the number of CAFTs, based on biological reproducibility instead of the numerical stability usually considered in literature. We proposed a criterion which evaluates the concordance between CAFTs computed on independent data sets, and combined it with classical procedures.

We implemented our approach on metagenes abundances from 1408 samples of the human gut microbiota, in order to determine CAFTs associated to fiber degradation in the distal gut. Fiber degradation was chosen because it is a major function of the gut microbial ecosystem. It involves the anaerobic fermentation of simple sugar which has been largely documented in the literature, allowing to impose constraints related to this step. The constraints were built by selecting 190 reference microbial genomes from prevalent species of the human gut microbiome, for which marker genes of anaerobic fermentation were annotated. This involved manual curation and relied on both state of the art knowledge of metabolism and accuracy of available annotations. We proposed an approach to account for missing information. In particular, genes were missing in pathways where they were expected to be present, which could be explained by annotation errors, or by the presence of alternative pathways undocumented in the literature. It could also come from unmodelled pathways, not related to fermentation, but sharing some common steps. Then, constraining missing genes markers would be an error. Our approach carefully took into account annotation, modelling and knowledge based errors by adjusting the constraints.

We found that our ecological model of CAFTs was consistent with the data, and that the metagenomic potential for fiber degradation by the human gut microbiota

could be interpreted as a mixture of 4 CAFTs, shared by all samples in variable 346
proportions. This data reduction highlighted biologically consistent functional patterns 347
while providing a high quality preservation of the original data. 348

Provided that prior knowledge is available, the generic framework we designed can 349
be applied to other metabolic processes in the gut or in other ecosystems. By 350
acknowledging fundamental data structuration our approach enables the inference 351
meaningful functional traits. 352

Materials and Methods

Metagenomic data processing

Data and gene frequencies. 1408 human faeces samples originating from 8 studies [3, 20–26] were exploited, covering several health status (Healthy, Crohn Disease, Ulcerative Colitis, Diabetes, Obesity) in donor groups from USA, Denmark, Spain, China and France.

The publicly released Integrated Gene Catalog (IGC) of 9.9 million genes from [20] was used to obtain counts, abundances and frequencies following the procedure described in [20] and [23].

Annotation. In order to obtain Glycosides Hydrolases (GH) and Pectines Lyases (PL) frequencies, we used Hmmer tool (<http://hmmer.org/>) and dbCan version3 [32] with default parameters. We first assessed dbCan annotation quality on 145 protein sequences for which a cazy annotation was available (<http://www.cazy.org/>) [33]. Similarity between annotation was evaluated by considering annotation from [33] as true and calculating accuracy. False positives mainly corresponded to Carbohydrate-binding modules, which we did not consider in this study. By ignoring them, we found an accuracy greater than 98%. As the results were good, the whole IGC was annotated, using the following rules in case of overlapping annotations, that is regions in an IGC gene sequence matching several GH or PL families: (1) if the size of the overlapping region did not exceed 10% of the smallest matched module size the overlap was ignored (2) otherwise, if the worst e-value between two overlapping annotations was 100 times smaller than the other, the corresponding annotation was rejected (3) in any other overlap situation both annotations were considered as erroneous and discarded. Regarding KO, we used the pre-existing KO annotation of the IGC, so that in the absence of any information on overlaps, the frequencies of genes with multiple KO annotations were dispatched in equal parts on each KO.

For each GH, PL and KO, the frequencies of all genes with the corresponding annotation were summed to obtain a global frequency.

Markers of fiber degradation in the human gut microbiota. We selected 25 GH and PL relevant to the human gut microbiota, involved in the degradation of dietary fiber such as cellulose, hemicellulose, starch and pectin [34]. The pathways involved in the anaerobic degradation and fermentation of sugars resulting from the hydrolysis of fiber are well characterized. They were identified using bibliographic resources [35] and Metacyc database [36] guided by expertise. Pathways exclusively associated to microorganisms unlikely present in the gut (e.g. soil microorganisms or extremophils) were excluded. A list of putative reactions and associated KO was compiled from the selected pathways using KEGG database, filtered according the following rules : (1) when a KO was not found in the IGC annotation it was ignored, (2) when a reaction could not be linked to a KO, it was ignored, (3) when the enzyme from a KO could catalyze more than one reaction, preventing from accurately linking the KO frequency to a unique target reaction, this KO was discarded as well as any other KO that could catalyze this reaction. Since rule (3) is stringent, a few KO able to catalyze several reactions were not discarded, but only when these additional reactions (and the associated pathway) were highly unlikely present in the gut microbiota (see S1 Table).

Some reactions were associated with multiple KO corresponding to either different enzymes catalyzing the same chemical reaction or parts of the same enzyme. In the first situation, each KO was considered as a functional marker, in the second situation the corresponding KO were merged into a unique marker whose frequency was set to the mean of the KO frequencies.

A final list of 86 functional markers characterizing the fiber degradation process in the human gut microbiome was obtained, comprising 25 GH and PL and 61 KO or KO aggregations (see S1 Table). The data matrix A was built from those 86 markers frequencies for the 1408 samples (see S1 tsv File).

Building the linear constraint from pathway prior knowledge

General principles for constraint building. Considering a microorganism x in the ecosystem, we denote n_{xj} the number of functional markers j in its genome. Moreover, for each intracellular metabolite m , we denote by P_m the set of indexes of functional makers involved in reactions producing m , and C_m the set of indexes of

functional makers involved in reactions consuming m . Then $N_{x,m,P} = \sum_{j \in P_m} n_{xj}$ is the number of markers associated to the production of metabolite m in the genome of x , and $N_{x,m,C} = \sum_{j' \in C_m} n_{xj'}$ denotes the number of markers associated to the consumption of m .

If $N_{x,m,C}$ is positive, we can define the ratio $\delta_{1,x,m} = N_{x,m,P}/N_{x,m,C}$, so that the following equality holds

$$N_{x,m,P} = \delta_{1,x,m} N_{x,m,C} \tag{8}$$

and in the same way, if $N_{x,m,P}$ is positive, we can define $\delta_{2,x,m} = N_{x,m,C}/N_{x,m,P}$, and write

$$N_{x,m,C} = \delta_{2,x,m} N_{x,m,P}. \tag{9}$$

If $N_{x,m,C}$ and $N_{x,m,P}$ are both zero, we arbitrarily set $\delta_{1,x,m} = \delta_{2,x,m} = 0$ and both (8) and (9) hold.

By definition, a CAFT l characterizes a gene pattern associated in each sample i to a subpopulation E_{il} in which the total abundance of any functional marker j is proportional to h_{lj} , namely

$$\sum_{x \in E_{il}} n_{xj} = w_{il} h_{lj}. \tag{10}$$

Therefore, for all m , $\sum_{x \in E_{i,l}} N_{x,m,P} = w_{il} \sum_{j \in P_m} h_{lj}$ and $\sum_{x \in E_{i,l}} N_{x,m,C} = w_{il} \sum_{j \in C_m} h_{lj}$.

Consider now a metabolite m for which (8) holds for all microorganisms x in the ecosystem, and define $\delta_{1,m} = \max_x \delta_{1,x,m}$. Therefore $N_{x,m,P} \leq \delta_{1,m} N_{x,m,C}$ and for each CAFT l , h_l satisfies

$$\sum_{j \in P_m} h_{lj} \leq \delta_{1,m} \sum_{j' \in C_m} h_{lj'} \tag{11}$$

In the same way, if we consider a metabolite m for which (9) holds for all microorganisms x in the ecosystem, each CAFT l satisfies

$$\sum_{j' \in C_m} h_{lj'} \leq \delta_{2,m} \sum_{j \in P_m} h_{lj} \tag{12}$$

where $\delta_{2,m} = \max_x \delta_{2,x,m}$.

In practice, for complex microbial ecosystems, it is not possible to compute the values of $\delta_{1,m}$ and $\delta_{2,m}$, since no exhaustive knowledge about the microorganisms that

are present and their genomes is available. However it is possible to propose sensible values that can be used in (11) and (12), as shown below in the case of fiber degradation by the gut ecosystem.

For a convenient implementation of the constraints, we built a graph representing the structure of the reactions under consideration, in which each vertex is a metabolite and each edge is directed and corresponds to a functional marker associated to a reaction. We built a $m_{int} \times r$ matrix Q derived from the incidence matrix of the graph. For each intracellular metabolite m and each functional marker j , $q_{mj} = 1$ (resp. $q_{mj} = -1$) if the reaction associated with functional marker j produces (resp. consumes) metabolite m , and $q_{mj} = 0$ otherwise. Thus, columns corresponding to functional markers not included in the graph are equal to zero. The matrix Q was split as $Q = Q^+ - Q^-$, where Q^+ contains the positive terms in Q and Q^- the absolute value of the negative ones. Note that

$$\sum_{j \in P_m} h_{lj} = q_m^+ h_l \quad \text{and} \quad \sum_{j \in C_m} h_{lj} = q_m^- h_l$$

where q_m^+ and q_m^- respectively denote the m^{th} line of Q^+ and Q^- .

Two sets of intracellular metabolites M_1 and M_2 were defined, for which constraints (11) and (12) respectively apply. The rows corresponding to sets M_1 and M_2 were extracted from Q^+ and Q^- and four sub-matrices Q_i^+ and Q_i^- for $i = 1, 2$ were built. Then constraints (11) and (12) were formulated as

$$F_{\Delta} H^T = \begin{pmatrix} Q_1^+ - \Delta_1 Q_1^- \\ Q_2^- - \Delta_2 Q_2^+ \end{pmatrix} H^T \leq 0 \quad (13)$$

where Δ_1 and Δ_2 were diagonal matrices with diagonal entries $\delta_{1,m}$ and $\delta_{2,m}$ for m in M_1 and M_2 respectively.

Constraint building for fiber degradation by the gut microbiota. We considered two types of functional markers, GH/PL for fiber degradation into simple sugars and KO for anaerobic fermentation of simple sugars. While precise information is available for anaerobic fermentation, the GH/PL families are not substrate specific, thus although they are essential in the fiber degradation process, they could not be included

in a reaction graph. Based on the KO only, we built an oriented graph from simple sugars to Short Chained Fatty Acid (SCFA). Edge orientation represents the biological catabolic pathway from simple sugars to Short Chained Fatty Acid (SCFA) and gas. During the KO selection process, several KO did not pass the specificity criterion and were discarded, preventing fully continuous pathways from sugars to short chained fatty acids and simplifications were introduced in order to build a connected graph. We finally obtained a graph with 43 nodes corresponding to metabolites (see S2 Table for their complete and abbreviated names) and $r_1 = 61$ edges corresponding to the selected KO. The metabolites were divided into $m_{int} = 25$ intracellular and $m_{ext} = 18$ extracellular ones. A summary of simplification, selected metabolites and KO is available in S1 Table.

Approximate values for the entries of Δ_1 and Δ_2 were computed by screening 190 core genomes chosen from [21] and [37], which were annotated using blast on Kegg Orthologies with a bitscore threshold of 60. The number of copies of the r selected markers for each genome were gathered in a $190 \times r$ matrix denoted G . We denote n_x the x^{th} row of G containing the number of copies of the functional markers in genome x , so that according to the notations defined above, $N_{x,m,C} = q_m^- n_x$ and $N_{x,m,P} = q_m^+ n_x$. A proxy for M_1 was defined as the set of metabolites m for which, for a fraction ranging from 85% to 95% of the genomes x , $N_{x,m,C} > 0$. In order to account for possible errors in using 190 core bacterial genomes instead of all the genomes and relaxing the definition of M_1 , $\delta_{1,m}$ was defined as an upper bound of $\max_{x \in \{1, \dots, 190\}} \left(\frac{N_{x,m,P}}{N_{x,m,C}} \right)$. In the same way, M_2 was defined as the set of metabolites m for which, for a fraction ranging from 85% to 95% of the genomes x , $N_{x,m,P} > 0$ and $\delta_{2,m}$ was defined as an upper bound of $\max_{x \in \{1, \dots, 190\}} \left(\frac{N_{x,m,C}}{N_{x,m,P}} \right)$. Further details are given in S1 Text.

Finally, equation (13) represents a set of 38 constraints with entries of Δ_1 and Δ_2 ranging from 5 to 30. The resulting matrix F_Δ is given in S3 tsv File.

Inference

Constrained NMF problem formulation and inference procedure. To take into account discrepancies among reactions, the $(n \times r)$ matrix A is scaled per reaction according to the formula

$$\tilde{a}_{ij} = \frac{a_{ij}}{\sqrt{\sum_{p=1}^n (a_{pj}^2)}} \quad (14)$$

were the scaled abundance matrix is denoted \tilde{A} . Defining the scaling matrix D as the diagonal matrix whose j^{th} diagonal entry is $\sqrt{\sum_{p=1}^n (a_{pj}^2)}$, it follows that $\tilde{A} = AD^{-1}$, and the trait matrix has to be scaled accordingly as $\tilde{H} = HD^{-1}$. The approximation problem is therefore formulated and solved in terms of matrix W and \tilde{H} , after which H is recovered as $H = \tilde{H}D$.

The approximation error is defined using the Frobenius norm as given in (6). Following [12] we use a regularizing term aimed at encouraging sparsity over the columns of \tilde{H} , formulated as $\alpha(\|W\|_F^2 + \|\mathbb{1}^T \tilde{H}\|_2^2)$, where $\mathbb{1}$ is defined as

$$\mathbb{1} = \underbrace{(1, 1, \dots, 1)^T}_{\text{size } k} \tag{15}$$

More details about this regularization term is given in S2 Text.

The matrix $\tilde{F}_\Delta = F_\Delta M^T$ such that $F_\Delta H^T = \tilde{F}_\Delta \tilde{H}^T$ is defined, in order to translate the constraints in terms of \tilde{H} . Finally, for a given value of (k, α) the NMF problem to be solved is formulated as:

$$(W^*, \tilde{H}^*) = \underset{W, \tilde{H}}{\operatorname{argmin}} \|\tilde{A} - W\tilde{H}\|_F^2 + \alpha(\|W\|_F^2 + \|\tilde{H}\|_{1,2}^2) \tag{16}$$

$$s.t. W \geq 0, \tilde{H} \geq 0, \tilde{F}_\Delta \tilde{H}^T \leq 0$$

where \tilde{H} is a $(k \times r)$ matrix and W is a $(n \times r)$ matrix. The NMF decomposition of A is (W^*, H^*) with $H^* = \tilde{H}^* M$.

Numerical implementation. Algorithm 1 details the inference process for the determination of W^* and \tilde{H}^* . The minimization step (17) is carried out using a block-coordinate descent algorithm detailed in Algorithm 2, in which W and \tilde{H} are alternatively updated.

Algorithm 2 was implemented in C language. The minimization step (19) is performed using Nesterov’s first order method (see [38] and [39]). Step (18) is solved using a standard Augmented Lagrangian Method [40], in which Uzawa iterations are combined with Nesterov’s first order method (see S3 Text for pseudo-code). Finally, upon convergence, the NMF approximation of \tilde{A} is (W^*, \tilde{H}^*) defined in Algorithm 1.

Algorithm 1 NMF resolution

$$\tilde{A} \leftarrow AD^{-1}$$

$$\tilde{F}_\Delta \leftarrow F_\Delta D^T$$

for Iter= 1 to 5 **do**

$$(W_{0,Iter}, \tilde{H}_{0,Iter}) \sim \mathcal{U}(0, 1)$$

$$W_{ini,Iter} \leftarrow W_{0,Iter} \sqrt{\frac{\|\tilde{A}\|_F}{\|W_{0,Iter} \tilde{H}_{0,Iter}\|_F}}$$

$$\tilde{H}_{ini,Iter} \leftarrow \tilde{H}_{0,Iter} \sqrt{\frac{\|\tilde{A}\|_F}{\|W_{0,Iter} \tilde{H}_{0,Iter}\|_F}}$$

$$(W_{Iter}^*, \tilde{H}_{Iter}^*) = \underset{\mathbf{W}, \tilde{\mathbf{H}}}{\operatorname{argmin}} \|\tilde{A} - \mathbf{W}\tilde{\mathbf{H}}\|_F^2 + \alpha(\|\mathbf{W}\|_F^2 + \|\tilde{\mathbf{H}}\|_{1,2}^2) \quad (17)$$

$$s.t. \mathbf{W} \geq 0, \tilde{\mathbf{H}} \geq 0, \tilde{F}_\Delta \tilde{\mathbf{H}}^T \leq 0$$

end for

$$(W^*, \tilde{H}^*) \leftarrow \underset{\text{Iter} \in [1,5]}{\operatorname{argmin}} \|\tilde{A} - W_{Iter}^* \tilde{H}_{Iter}^*\|_F^2$$

$$H^* \leftarrow \tilde{H}^* D$$

Algorithm 2 Block-Coordinate-Descent for NMF

while Convergence \neq True **do**

$$\tilde{H}^{(t+1)} \leftarrow \underset{\mathbf{H} \geq 0, F_\Delta \mathbf{H}^T \leq 0}{\operatorname{argmin}} \|\tilde{A} - W^t \mathbf{H}\|_F^2 + \alpha \|\mathbf{H}\|_{1,2}^2 \quad (18)$$

$$W^{(t+1)} \leftarrow \underset{\mathbf{W} \geq 0}{\operatorname{argmin}} \|\tilde{A} - \mathbf{W} \tilde{H}^{(t+1)}\|_F^2 + \alpha \|\mathbf{W}\|_F^2 \quad (19)$$

end while

Criteria for parameter selection *Bi-cross-validation.* The bi-cross-validation procedure developed in [19] is an adaptation of the classic cross-validation procedure to the NMF framework, in which both samples (rows of A) and functional markers (columns of A) are split into training and validation sets. Bi-cross-validation error evaluates the prediction ability of the NMF decomposition. Nevertheless, it can not be computed for constrained NMF inference since the column split breaks up the structure of the metabolic relationships on which the constraints are based. Therefore, we consider this criterion as an indication of the performances of the NMF decomposition computed without constraints.

Reconstruction error. The reconstruction error $\|A - W^*H^*\|_F/\|A\|_F$ indicates the proportion of information in A recovered in the NMF approximation W^*H^* . For a fixed α , this quantity automatically decreases as the number of profiles k increases, and the decreasing rate when a profile is added indicates the significance of this profile. Thus, a slope discontinuity on the graph of the reconstruction error as a function of k indicates that additional profiles are less significant. Similarly, for a given k the reconstruction error increases as α increases; we considered the optimal value to be the largest α before the reconstruction error significantly increases.

Concordance of CAFTs from independent data sets. Let M be a matrix of dimension (n_1, r_1) , the similarity matrix S^M of M is the squared matrix with r_1 columns whose (j, j') entry is the cosine of the angle between the column vectors j and j' of M :

$$s_{jj'}^M = \frac{\sum_{l=1}^{n_1} m_{lj}m_{lj'}}{\left(\sum_{l=1}^{n_1} m_{lj}^2\right)^{1/2} \left(\sum_{l=1}^{n_1} m_{lj'}^2\right)^{1/2}}.$$

For repetitions $a = 1, \dots, 40$, the set of rows of \tilde{A} is randomly split in two independent sets of equal length, the corresponding submatrices $\tilde{A}_1^{(a)}$ and $\tilde{A}_2^{(a)}$ are extracted and their NMF decompositions $(W_1^{(a)}, \tilde{H}_1^{(a)})$ and $(W_2^{(a)}, \tilde{H}_2^{(a)})$ are inferred using the constrained NMF algorithm. Then, the concordance index is defined as

$$1 - \frac{1}{40 \times r} \sum_{a=1}^{40} \|S_1^{(a)} - S_2^{(a)}\|_F$$

with $S_1^{(a)}$ and $S_2^{(a)}$ the similarity matrices of $\tilde{H}_1^{(a)}$ and $\tilde{H}_2^{(a)}$.

Selection of optimal parameters for fiber digestion analysis. The optimal value of α

was selected by computing the three criteria for (α, k) in $(10^{-3}, 10^{-2}, 10^{-1.5}, 10^{-1}) \times (2, 4, 6, 8, 11, 13)$. Then, the optimal number of CAFTs k was selected by computing the three criteria for the optimal value of α and for k in the set of integers from 2 to 15. The same 10-fold split of rows and columns in the bi-cross-validation procedure, and the same row splits for concordance of H , were used for all values of k and α .

Preservation of the geometric structure

The columns of matrices A and W were ℓ^1 -normalized; the normalized matrices correspond to the repartition of the functional markers (resp. the CAFTs) abundancies among the samples. The distance matrix d^A (resp. d^W) of A (resp. W) defined as the 1408×1408 matrix whose element (i, j) is the euclidean distance between rows i and j of A (resp. W) was computed. For each biological sample i , the Pearson correlation between the distance of i to the other samples in X and in W was evaluated:

$$C_i = \text{cor} \left((d_{i,j}^A)_{j \neq i}, (d_{i,j}^W)_{j \neq i} \right)$$

and the histogram of $(C_i)_{i=1, \dots, 1408}$ was displayed.

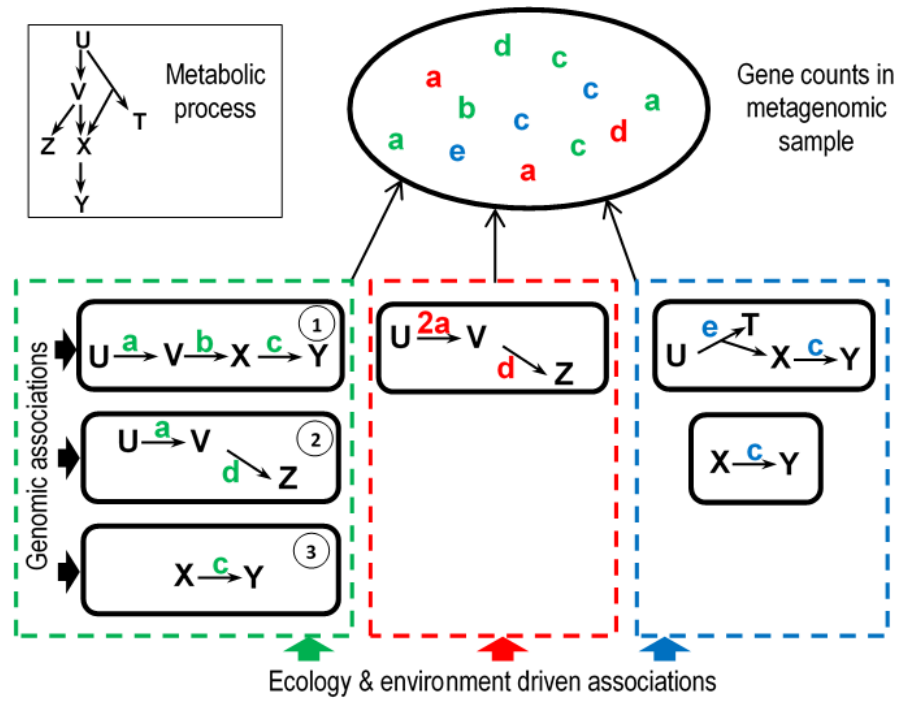


Figure 1. Hierarchical structure underlying metagenomic data.

525

526

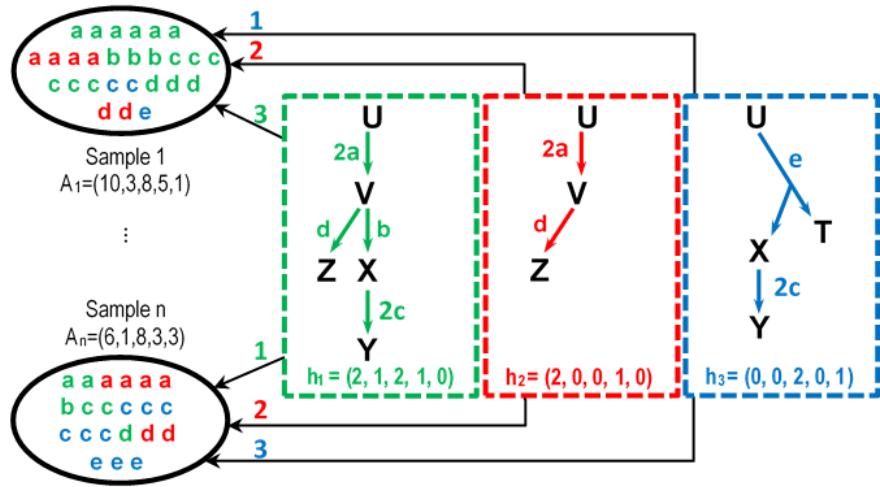


Figure 2. Combined Aggregated Functional Traits.

527

528

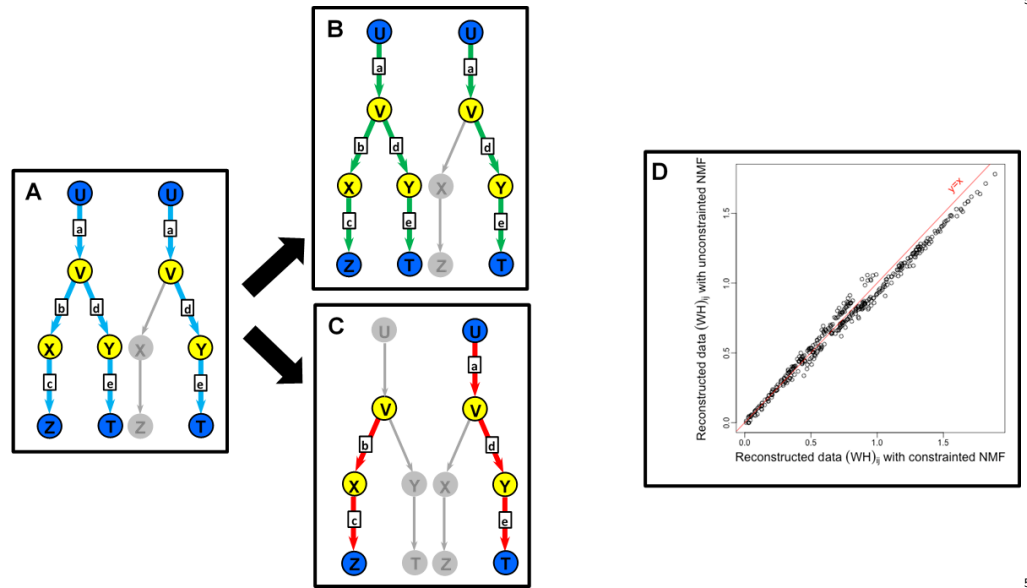


Figure 3. A toy example to illustrate the impact of the constraints.

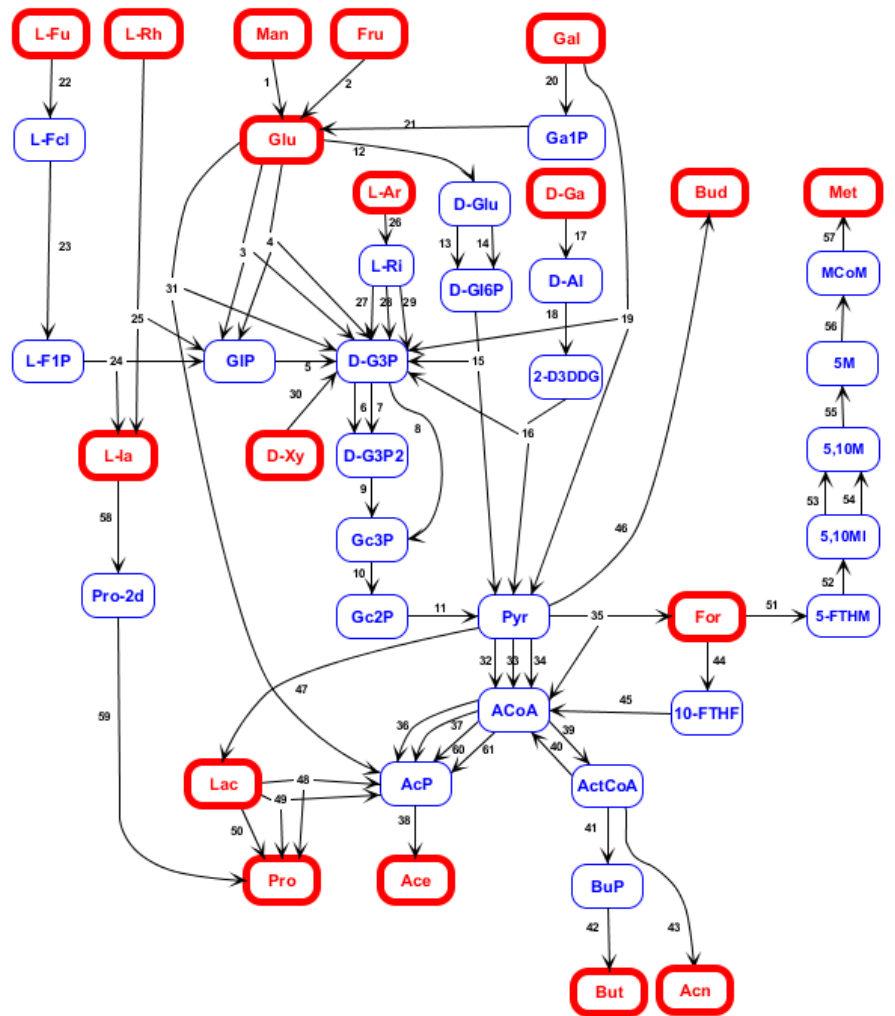


Figure 4. Reaction graph of the biological catabolic pathway from simple sugars to Short Chained Fatty Acids (SCFA) and gas.

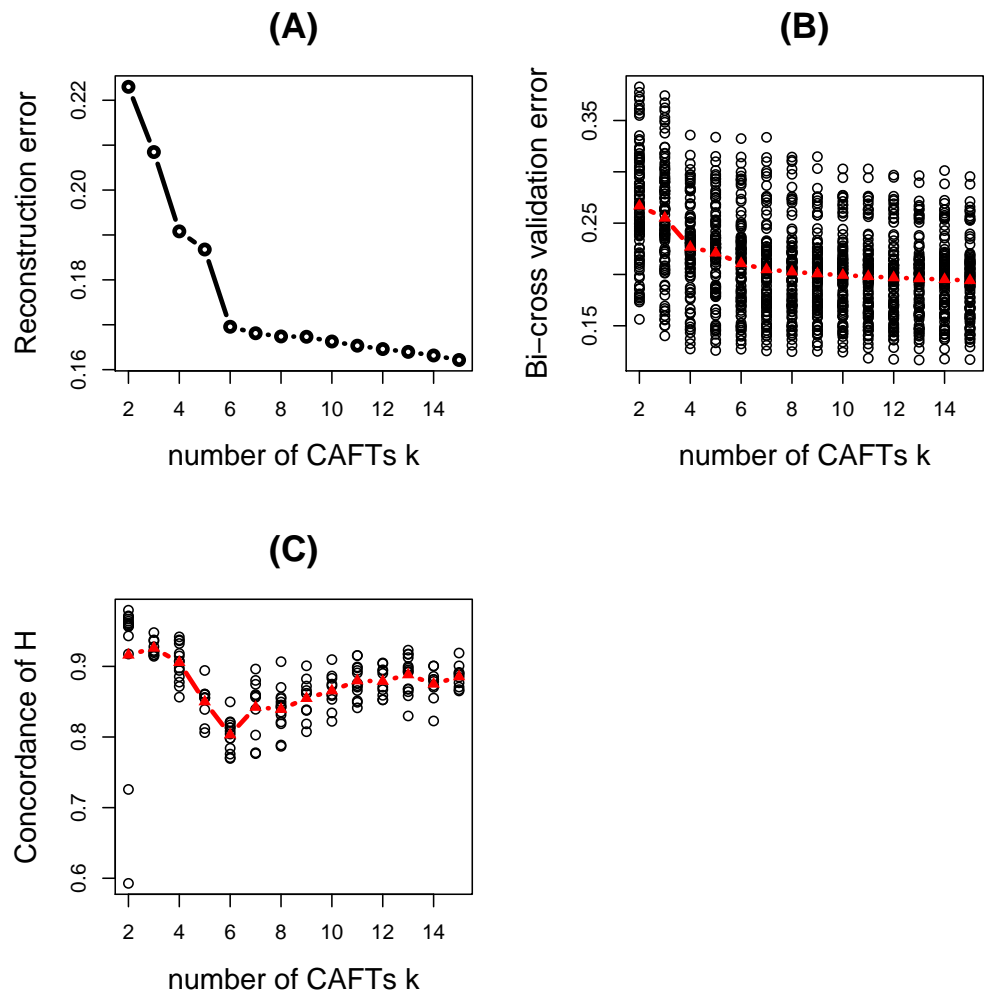


Figure 5. Fiber digestion in the human gut: selection of the number of CAFTs k .

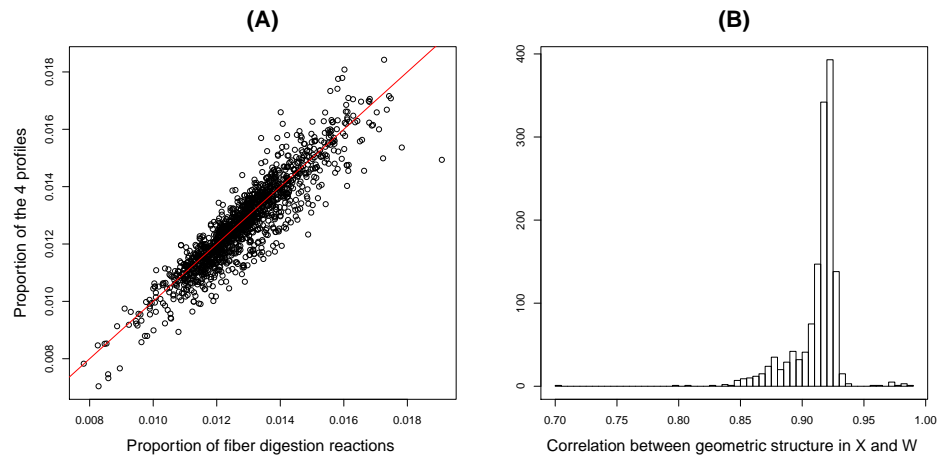


Figure 6. Fiber digestion in the human gut: analysis of W^* .

540

541

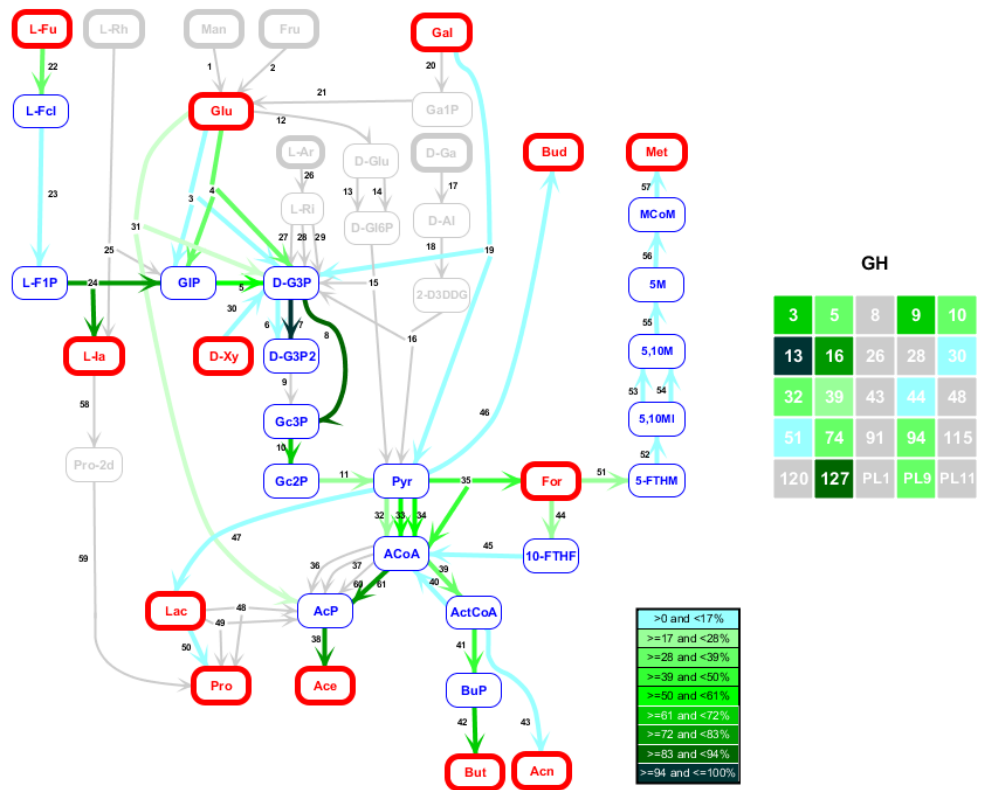


Figure 7. An example of CAFT.

Supporting Information 545

S1 Text 546

Choice of Δ . This file details the process of choosing the values in Δ from 190 core gut reference genomes. 547
548

S2 Text 549

Regularization term. This file details the mathematical motivation for using a unique regularization parameter. 550
551

S3 Text 552

Pseudo-code of the minimization Algorithm. This file provides details about optimization steps (18) and (19). 553
554

S4 Text 555

Tuning of the regularisation parameter α . This file details the selection of the regularization parameter α based on the three criteria. 556
557

S1 tsv File 558

Functional markers abundances. This file contains the abundance matrix of the 86 functional markers of fiber degradation in 1408 samples. 559
560

S2 tsv File 561

List of 190 core genomes used for constraint building. 562

S3 tsv File 563

Constraint matrix. This file contains the constraint matrix F_{Δ} , storing the coefficients of the 38 constraints on the 86 entries of the CAFTs. 564
565

S1 Table

566

Description of fiber hydrolysis. This file details the hypotheses underlying the selection of functional markers and metabolites representing simple sugar anaerobic fermentation in the human gut (sheet Reactions and KOs) and provides the list of the GH and PL selected in the model (sheet GH and PL).

567

568

569

570

S2 Table

571

List of metabolites. This file gives the full and abbreviated names of the metabolites selected for the description of simple sugar anaerobic fermentation in the human gut.

572

573

References

1. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, et al. Metagenomic species profiling using universal phylogenetic marker genes. *Nature methods*. 2013;10(12):1196–1199.
2. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic contigs by coverage and composition. *Nature methods*. 2014;11(11):1144–1146.
3. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature biotechnology*. 2014;32(8):822–828.
4. Kultima JR, Coelho LP, Forslund K, Huerta-Cepas J, Li SS, Driessen M, et al. MOCAT2: a metagenomic assembly, annotation and profiling framework. *Bioinformatics*. 2016; Available from: <http://bioinformatics.oxfordjournals.org/content/early/2016/04/08/bioinformatics.btw183.abstract>.
5. Tringe SG, Von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, et al. Comparative metagenomics of microbial communities. *Science*. 2005;308(5721):554–557.

6. Segata N, Boernigen D, Tickle TL, Morgan XC, Garrett WS, Huttenhower C. Computational meta'omics for microbial community studies. *Molecular Systems Biology*. 2013;9(1):1–15.
7. Sharpton TJ. An Introduction to the Analysis of Shotgun Metagenomic Data. *Frontiers in Plant Science*. 2014;5(209).
8. Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol*. 2012;8(6):e1002358.
9. Moreno García CA, Schellberg J, Ewert F, Brüser K, Canales-Prati P, Linstädter A, et al. Response of community-aggregated plant functional traits along grazing gradients: insights from African semi-arid grasslands. *Applied vegetation science*. 2014;17(3):470–481.
10. Fierer N, Barberan A, Laughlin DC. Seeing the forest for the genes: using metagenomics to infer the aggregated traits of microbial communities. *Frontiers in Microbiology*. 2014;5(November):1–6.
11. Raes J, Letunic I, Yamada T, Jensen LJ, Bork P. Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Molecular Systems Biology*. 2011;7(1):473–473.
12. Kim H, Park H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*. 2007;23(12):1495–1502.
13. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A*. 2004;101(12):4164–4169.
14. Baran Y, Halperin E. Joint analysis of multiple metagenomic samples. *PLoS computational biology*. 2012;8(2):e1002373.
15. Jiang X, Weitz JS, Dushoff J. A non-negative matrix factorization framework for identifying modular patterns in metagenomic profile data. *Journal of Mathematical Biology*. 2012;64(4):697–711.

16. Jiang X, Langille MGI, Neches RY, Elliot M, Levin SA, Eisen JA, et al. Functional Biogeography of Ocean Microbes Revealed through Non-Negative Matrix Factorization. *PLoS ONE*. 2012;7(9):e43866.
17. Jiang X, Hu X, Xu W. Joint Analysis of Functional and Phylogenetic Composition for Human Microbiome Data. In: Basu M, Pan Y, Wang J, editors. *Bioinformatics Research and Applications: 10th International Symposium, ISBRA 2014, Zhangjiajie, China, June 28-30, 2014. Proceedings*. Springer International Publishing; 2014. p. 346–356.
18. Kim H, Park H. Nonnegative Matrix Factorization Based on Alternating Nonnegativity Constrained Least Squares and Active Set Method. *SIAM Journal on Matrix Analysis and Applications*. 2008;30(2):713–730.
19. Owen AB, Perry PO. Bi-cross-validation of the SVD and the nonnegative matrix factorization. *The Annals of Applied Statistics*. 2009;3(2):564–594.
20. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog of reference genes in the human gut microbiome. *Nature biotechnology*. 2014;32(8):834–841.
21. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing : Article : *Nature*. *Nature*. 2010;464(7285):59–65.
22. Methé BA, Nelson KE, Pop M, Creasy HH, Giglio MG, Huttenhower C, et al. A framework for human microbiome research. *Nature*. 2012;486(7402):215–221.
23. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*. 2012;490(7418):55–60.
24. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, et al. Richness of human gut microbiome correlates with metabolic markers. *Nature*. 2013;500(7464):541–6.
25. Erickson AR, Cantarel BL, Lamendella R, Darzi Y, Mongodin EF, Pan C, et al. Integrated Metagenomics/Metaproteomics Reveals Human Host-Microbiota Signatures of Crohn's Disease. *PLoS ONE*. 2012;7(11).

26. Cotillard A, Kennedy SP, Kong LC, Prifti E, Pons N, Le Chatelier E, et al. Dietary intervention impact on gut microbial gene richness. *Nature*. 2013;500(7464):585–588.
27. Kanehisa M, Goto S. KEGG: Kyoto Encyclopaedia of Genes and Genomes. *Nucl Acids Res*. 2000;28(1):27–30.
28. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic acids research*. 2014;42(Database issue):D199—205.
29. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Research*. 2009;37(Database):D233–D238.
30. Cummings JH, Macfarlane GT. The control and consequences of bacterial fermentation in the human colon. *Journal of Applied Bacteriology*. 1991;70(6):443–459.
31. Gaci N, Borrel G, Tottey W, O’Toole PW, Brugère JF. Archaea and the human gut: new beginning of an old story. *World J Gastroenterol*. 2014;20(43):16062–16078.
32. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic acids research*. 2012;40(Web Server issue):W445–51.
33. Tasse L, Bercovici J, Pizzut-Serin S, Robe P, Tap J, Klopp C, et al. Functional metagenomics to mine the human gut microbiome for dietary fiber catabolic enzymes. *Genome Research*. 2010;20(11):1605–1612.
34. El Kaoutari A, Armougom F, Gordon JI, Raoult D, Henrissat B. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nature reviews Microbiology*. 2013 jul;11(7):497–504.
35. Gottschalk G. *Bacterial Metabolism*. Springer Series in Microbiology. New York, NY: Springer New York; 1986.

36. Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic acids research*. 2014;42(Database issue):D459—71.
37. Tap J, Mondot S, Levenez F, Pelletier E, Caron C, Furet JP, et al. Towards the human intestinal microbiota phylogenetic core. *Environmental Microbiology*. 2009;11(10):2574–2584.
38. Nesterov Y. *Introductory Lectures on Convex Programming Volume I: Basic course*. vol. I; 1996.
39. Tseng P. On accelerated proximal gradient methods for convex-concave optimization. submitted to *SIAM Journal on Optimization*. 2008;.
40. Arrow KJ, Hurwicz L, Uzawa H. *Studies in linear and non-linear programming*. Stanford mathematical studies in the social sciences. Stanford University Press; 1958.

Glossaire

Notations

\mathbb{R}^+ est le corps des réels positifs.

\mathbb{R}^* est le corps des réels privés de 0.

$\mathcal{M}_{n,r}(\mathbb{R}^+)$ est l'ensemble des matrices à termes positifs de dimensions $n \times r$.

X_{ij} est le terme de la ligne i et de la colonne j de la matrice X .

$X_{(i,.)}$ désigne la ligne i de la matrice X .

$X_{(.,j)}$ désigne la colonne j de la matrice X .

$\|\cdot\|_F$ est la norme matricielle de Frobenius qui correspond à la racine carrée de la somme des carrés des termes de la matrice.

Plusieurs problèmes d'optimisations sont posés dans ce manuscrit, les variables optimisées sont par convention mises en gras.

Acronymes

ACP : Analyse en Composantes Principales, correspond à une décomposition en valeurs singulières de la matrice de variance/covariance d'une matrice de données.

AGCC : Acides Gras à Chaînes Courtes, composés chimiques composés d'une chaîne carbonée de quelques atomes carbones et dont une extrémité correspond à un acide carboxylique. Ils sont excrétés par le microbiote intestinal humain lors du catabolismes des polysaccharides.

ANLS : Alternative Non-negative Least Square. Approche de résolution approchée du problème de NMF, reposant sur la résolution d'un problème des moindres carrés suivis d'une projection sur l'orthant positif.

ATP : Adénosine Tri-Phosphate, molécule dont la conversion en Adénosine Di-Phosphate permet la catalyse de réaction endo-énergétiques.

BCD : Block Coordinate Descent, approche de résolution de la NMF consistant en la résolution alternée des problèmes de minimisation partiels en les variables.

BiCV : Bi-Cross Validation. Approche de validation croisée proposée dans le cadre de la NMF et des problème de décomposition en valeurs singulières.

BLAST : Basic Local Alignment Search Tool. Outils classique de bioinformatique, permettant la recherche de similarité entre séquences. Utilisée dans l'annotation de séquences de nucléotides ou d'acides aminés.

CAG : Co-Abundance genes Group. Groupes de gènes dont l'abondance co-varie construits et utilisés par Nielsen et al. (2014).

CAZY : Carbohydrate-Active enZYmes. Base de donnée de familles d'enzymes relative à la dégradation des polysaccharides, crée et entretenue par Henrissat et al. (1991).

CBM : Carbohydrate Binding Module. Modules protéiques permettant l'adsorption sur le sucre ou le polysaccharide cible. Est présent des protéines possédant par ailleurs une fonction hydrolitique des polysaccharides.

CTFA : Combinaison de Traits Fonctionnels Agrégés. Vecteurs réunissant différents TFA permettant de caractériser un assemblage fonctionnel. Il est par convention à valeur sur le simplex.

EC : Enzyme Commission number. Classification numérique et hiérarchique des enzymes selon leur activité chimique.

EMP : Embden-Meyehof-Parnas. Aussi appelée glycolyse, associé à la consommation de sucres et la production d'ATP.

FAD : Flavine Adénine Dinucléotide. Cofacteur d'oxydo-réduction.

GH : Glycosides Hydrolases. Familles d'enzymes associées à l'hydrolyses de différentes liaisons des polysaccharides.

HGC : High Gene Count. Dénomination proposée par Le Chatelier et al. (2013) pour caractériser la quantité de gène d'un groupe d'échantillons.

MetaHIT : Metagenomics of the Human Intestinal Tract. Projet européen de grande envergure focalisé sur l'étude du microbiote intestinal humain.

HMP : Human Microbime Project. Projet américain de grande envergure focalisé sur l'étude du microbiote humain.

HUMANn : Processus de traitement des données métagénomiques proposé par Abubucker et al. (2012).

IGC : Integrated Genome Catalog. Catalogue de 9.9 millions de gènes assemblé *de novo* à partir de 1263 échantillons métagénomiques.

IMC : Indice de Masse Corporelle.

KEGG : Kyoto Encyclopedia of genes and genomes. Base de donnée faisant le liens entre réactions métaboliques, composés métaboliques, orthologies, gènes et organismes microbiens.

KKT : Karush-Kuhn-Tucker. Conditions nécessaire à l'obtention d'un point critique dans le cas d'un problème d'optimisation sous contraintes linéaires.

KLG : Kullback-Leibler Généralisée.

KO : KEGG Orthology. Orthologie construite manuellement par KEGG.

LGC : Low Gene Count. Dénomination proposée par Le Chatelier et al. (2013) pour caractériser la quantité de gène d'un groupe d'échantillons.

LOMD : Loss Of Microbiota Diversity. Perte de diversité taxonomique du microbiote intestinal humain, souvent en liens avec une pathologie.

MGS : MetaGenomic Species. Regroupement de CAG pouvant vraisemblablement correspondre à des génomes d'espèces microbiennes. Proposés par Nielsen et al. (2014).

MICI : Maladies Inflammatoires Chroniques Intestinales.

MOCAT : MetagenOmiC Analysis Toolkit. Logiciel développé par le consortium MetaHIT qui associe différents outils spécialisés afin de proposer un processus de traitement standardisé prenant en entrée des échantillons métagénomiques et produisant en sortie un catalogue de gènes assemblés *de novo*. Proposés par (Kultima et al. (2012)).

MUR : Multiplicative Update Rule. Règle multiplicative assurant la décroissance de la fonction de coût d'un problème de NMF (Lee and Seung, 1999).

NAD⁺NADH : Nicotinamide Adénine Dinucléotide. Cofacteur principal des réactions d'oxydation de la fermentation des monosaccharides.

NMF : Non-negative Matrix Factorisation. Problème où l'on recherche un couple de matrices positives permettant d'approcher une matrice de donnée.

NNLS : Non-Negative Least Square. Moindres carrés sous contrainte de positivité.

eggNOG : Orthologies construites automatiquement par Powell et al. (2012).

OTU : Operational Taxonomic Unit. Organismes pragmatiquement et taxonomiquement définis par leur séquence d'ARNr 16S.

PL : Polysaccharide Lysase. Famille d'enzyme associés à la lyse de liaisons entre acides uroniques.

SOAP : Outils d'alignement rapide de reads courts.

SVD : Singular Value Decomposition.

TFA : Trait Fonctionnel Agrégés. Ce sont la valeur moyenne d'un trait fonctionnel sur l'ensemble des individus d'une communauté ou d'un écosystème.

WGS : Whole Genome Shotgun. Méthode de séquençage passant par une étape de fragmentation du matériel génétique en séquences de faibles tailles appelées "reads".

Bibliographie

Bibliographie

- Abubucker, S., et al. (2012). Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.*, 8(6) :e1002358.
- Albenberg, L., et al. (2014). Correlation between intraluminal oxygen gradient and radial partitioning of intestinal microbiota. *Gastroenterology*, 147(5) :1055–1063.e8.
- Antoulas, A. C. (2005). *Approximation of Large-Scale Dynamical Systems*, volume 6. Society for Industrial and Applied Mathematics.
- Arrow, K. J., Hurwicz, L., and Uzawa, H. (1958). *Studies in linear and non-linear programming*. Stanford University Press.
- Arumugam, M., et al. (2011). Enterotypes of the human gut microbiome. *Nature*, 473(7346) :174–180.
- Baran, Y. and Halperin, E. (2012). Joint analysis of multiple metagenomic samples. *PLoS Comput. Biol.*, 8(2).
- Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P., and Plemmons, R. J. (2007). Algorithms and applications for approximate nonnegative matrix factorization. *Comput. Stat. Data Anal.*, 52(1) :155–173.
- Bertsekas, D. P. (1999). *Nonlinear Programming. (01 Sept. 1999)*.
- Bik, E. M., et al. (2006). Molecular analysis of the bacterial microbiota in the human stomach. *Proc. Natl. Acad. Sci.*, 103(3) :732–737.
- Brunet, J. P., Tamayo, P., Golub, T. R., and Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A*, 101(12) :4164–4169.

- Caspi, R., et al. (2014). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, 42(Database issue) :D459—71.
- Cemgil, A. T. (2009). Bayesian inference for nonnegative matrix factorisation models. *Comput. Intell. Neurosci.*, 2009.
- Cho, Y. and Saul, L. (2011). Nonnegative Matrix Factorization for Semi-supervised Dimensionality Reduction. *arXiv Prepr. arXiv1112.3714*.
- Cichocki, A., Zdunek, R., and Amari, S.-I. (2006). New Algorithms for Non-Negative Matrix Factorization in Applications to Blind Source Separation. In *2006 IEEE Int. Conf. Acoust. Speed Signal Process.*, volume 5, pages V-621–V-624. IEEE.
- Cotillard, A., et al. (2013). Dietary intervention impact on gut microbial gene richness. *Nature*, 500(7464) :585–8.
- Cryan, J. F. and Dinan, T. G. (2012). Mind-altering microorganisms : the impact of the gut microbiota on brain and behaviour. *Nat. Rev. Neurosci.*, 13(10) :701–712.
- Cummings, J. H. and Macfarlane, G. T. (1997). Role of intestinal bacteria in nutrient metabolism. *JPEN. J. Parenter. Enteral Nutr.*, 21(6) :357–365.
- Davila, A. M., et al. (2013). Re-print of "intestinal luminal nitrogen metabolism : Role of the gut microbiota and consequences for the host". *Pharmacol. Res.*, 69(1) :114–126.
- Dhillon, I. S. and Sra, S. (2005). Generalized Nonnegative Matrix Approximations with Bregman Divergences. *Adv. Neural Inf. Process. Syst.*, 19 :283–290.
- Donaldson, G. P., Lee, S. M., and Mazmanian, S. K. (2015). Gut biogeography of the bacterial microbiota. *Nat. Rev. Microbiol.*, 14(1) :20–32.
- Eckburg, P. B., et al. (2005). Diversity of the Human Intestinal Microbial Flora. *Science (80-)*, 42(1) :84–94.
- Egert, M., De Graaf, A. A., Smidt, H., De Vos, W. M., and Venema, K. (2006). Beyond diversity : Functional microbiomics of the human colon. *Trends Microbiol.*, 14(2) :86–91.
- El Kaoutari, A., Armougom, F., Gordon, J. I., Raoult, D., and Henrissat, B. (2013). The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nat. Rev. Microbiol.*, 11(7) :497–504.

- Flint, H. J., Duncan, S. H., Scott, K. P., and Louis, P. (2007). Interactions and competition within the microbial community of the human colon : Links between diet and health : Minireview. *Environ. Microbiol.*, 9(5) :1101–1111.
- Fogel, P., Young, S. S., Hawkins, D. M., and Ledirac, N. (2007). Inferential, robust non-negative matrix factorization analysis of microarray data. *Bioinformatics*, 23(1) :44–49.
- Foster, J. A. and McVey Neufeld, K. A. (2013). Gut-brain axis : How the microbiome influences anxiety and depression. *Trends Neurosci.*, 36(5) :305–312.
- Franzosa, E. A., et al. (2015). Identifying personal microbiomes using metagenomic codes. *Proc. Natl. Acad. Sci.*, 112(22) :E2930–E2938.
- Gabert, L., et al. (2011). ¹³C tracer recovery in human stools after digestion of a fat-rich meal labelled with [1,1,1-¹³C₃]tripalmitin and [1,1,1-¹³C₃]triolein. *Rapid Commun. Mass Spectrom.*, 25(19) :2697–2703.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*.
- Glenn R Gibson, George T Macfarlane, J. C. (1993). Leading article Sulphate reducing bacteria and hydrogen metabolism in the human. *Gut*, 34 :437–439.
- Gonzalez, E. and Zhang, Y. (2005). Accelerating the Lee-Seung algorithm for non-negative matrix factorization. *Dept. Comput. Appl. Math., Rice Univ., Houston, TX, Tech. Rep. TR-05-02*, (TR05-02) :1–13.
- Gottschalk, G. (1986). Bacterial Fermentations. pages 208–282.
- Grippo, L. and Sciandrone, M. (2000). On the convergence of the block nonlinear Gauss-Seidel method under convex constraints. *Oper. Res. Lett.*, 26(3) :127–136.
- Guan, N., Tao, D., Luo, Z., and Yuan, B. (2012). NeNMF : An optimal gradient method for nonnegative matrix factorization. *IEEE Trans. Signal Process.*, 60(6) :2882–2898.
- Han, L., Neumann, M., and Prasad, U. (2009). Alternating projected Barzilai-Borwein methods for Nonnegative Matrix Factorization. *Electron. Trans. Numer. Anal.*, 36(06) :54–82.
- Henrissat, B., Vegetales, M., and Grenoble, F. (1991). A classification of glycosyl hydrolases based sequence similarities amino acid. *Biochem. J.*, 280((Pt 2)) :309–316.

- Hoyer, P. O. (2004). Non-negative Matrix Factorization with Sparseness. pages 1457–1459. *Journal of Machine Learning Research*.
- Huttenhower, C., et al. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402) :207–214.
- Jeraldo, P., et al. (2012). Quantification of the relative roles of niche and neutral processes in structuring gastrointestinal microbiomes. *Proc. Natl. Acad. Sci.*, 109(25) :9692–9698.
- Jiang, X., Hu, X., and Xu, W. (2014). Joint Analysis of Functional and Phylogenetic Composition for Human Microbiome Data. *Bioinforma. Res. Appl.*, pages 346–356.
- Jiang, X., et al. (2012a). Functional Biogeography of Ocean Microbes Revealed through Non-Negative Matrix Factorization. *PLoS One*, 7(9) :1–9.
- Jiang, X., Weitz, J. S., and Dushoff, J. (2012b). A non-negative matrix factorization framework for identifying modular patterns in metagenomic profile data. *J. Math. Biol.*, 64(4) :697–711.
- Johansson, M. E. V., Larsson, J. M. H., and Hansson, G. C. (2011). The two mucus layers of colon are organized by the MUC2 mucin, whereas the outer layer is a legislator of host-microbial interactions. *Proc. Natl. Acad. Sci. U. S. A.*, 108 Suppl(Supplement_1) :4659–65.
- Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2009). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, 38(SUPPL.1) :D355–D360.
- Kim, H. and Park, H. (2007). Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12) :1495–1502.
- Kim, H. and Park, H. (2008). Nonnegative Matrix Factorization Based on Alternating Non-negativity Constrained Least Squares and Active Set Method. *SIAM J. Matrix Anal. Appl.*, 30(2) :713–730.
- Kim, P. M. and Tidor, B. (2003). Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Res.*, 13(7) :1706–1718.
- Koenig, J. E., et al. (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl. Acad. Sci. U. S. A.*, 108 Suppl :4578–4585.

- Koren, O., et al. (2013). A Guide to Enterotypes across the Human Body : Meta-Analysis of Microbial Community Structures in Human Microbiome Datasets. *PLoS Comput. Biol.*, 9(1) :e1002863.
- Kultima, J. R., et al. (2012). MOCAT : A Metagenomics Assembly and Gene Prediction Toolkit. *PLoS One*, 7(10) :e47656.
- Kurtz, Z. D., et al. (2015). Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLoS Comput. Biol.*, 11(5) :e1004226.
- Le Chatelier, E., et al. (2013). Richness of human gut microbiome correlates with metabolic markers. *Nature*, 500(7464) :541–6.
- Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755) :788–91.
- Lee, D. D. and Seung, H. S. (2001). Algorithms for non-negative matrix factorization. *Adv. Neural Inf. Process. Syst.*, 13(1) :556–562.
- Lepage, P., et al. (2011). Twin study indicates loss of interaction between microbiota and mucosa of patients with ulcerative colitis. *Gastroenterology*, 141(1) :227–236.
- Lepage, P., et al. (2013). A metagenomic insight into our gut’s microbiome. *Gut*, 62(1) :146–58.
- Li, J., et al. (2014). An integrated catalog of reference genes in the human gut microbiome. *Nat Biotech*, advance on(8) :834–41.
- Li, R., et al. (2009). SOAP2 : An improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15) :1966–1967.
- Lin, C. (2007). Projected gradient methods for nonnegative matrix factorization. *Neural Comput.*, 19(10) :2756–2779.
- Lozupone, C. a., et al. (2008). The convergence of carbohydrate active gene repertoires in human gut microbes. *Proc. Natl. Acad. Sci. U. S. A.*, 105(39) :15076–15081.
- Lu, J., Tomfohr, J. K., and Kepler, T. B. (2005). Identifying differential expression in multiple SAGE libraries : an overdispersed log-linear model approach. *BMC Bioinformatics*, 6 :165.
- Martens, E. C., et al. (2011). Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts. *PLoS Biol.*, 9(12).

- Mazmanian, S. K., Cui, H. L., Tzianabos, A. O., and Kasper, D. L. (2005). An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. *Cell*, 122(1) :107–118.
- McHugh, P. R. and Moran, T. H. (1979). Calories and gastric emptying : a regulatory capacity with implications for feeding. *Am. J. Physiol.*, 236(10) :R254–R260.
- Méthé, B. A., et al. (2012). A framework for human microbiome research. *Nature*, 486(7402) :215–221.
- Moreno García, C. A., et al. (2014). Response of community-aggregated plant functional traits along grazing gradients : insights from African semi-arid grasslands. *Appl. Veg. Sci.*, 17(3) :470–481.
- Mosca, A., Leclerc, M., and Hugot, J. P. (2016). Gut microbiota diversity and human diseases : Should we reintroduce key predators in our ecosystem ? *Front. Microbiol.*, 7(MAR) :1–12.
- Muñoz-Tamayo, R. (2012). *Modélisation Mathématique de la Dégradation des Hydrates de Carbone dans le Côlon Humain*. PhD thesis.
- Nelson, K. E., et al. (2010). A catalog of reference genomes from the human microbiome. *Science*, 328(5981) :994–9.
- Nemergut, D. R., Cleveland, C. C., Wieder, W. R., Washenberger, C. L., and Townsend, A. R. (2010). Plot-scale manipulations of organic matter inputs to soils correlate with shifts in microbial community composition in a lowland tropical rain forest. *Soil Biol. Biochem.*, 42(12) :2153–2160.
- Nemergut, D. R., et al. (2013). Patterns and Processes of Microbial Community Assembly. *Microbiol. Mol. Biol. Rev.*, 77(3) :342–356.
- Nesterov, Y. (1998). *Introductory Lectures on Convex Programming Volume I : Basic course*, volume I.
- Nielsen, H. B., et al. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.*, 32(8) :822–828.
- O’Hara, A. M. and Shanahan, F. (2006). The gut flora as a forgotten organ. *EMBO Rep*, 7(7) :688–693.

- Owen, A. B. and Perry, P. O. (2009). Bi-cross-validation of the SVD and the nonnegative matrix factorization. *Ann. Appl. Stat.*, 3(2) :564–594.
- Paatero, P. and Tapper, U. (1994). Positive Matrix Factorization - A Nonnegative Factor Model with Optimal Utilization of Error Estimates of Data Values. *Environmetrics*, 5(2) :111–126.
- Powell, S., et al. (2012). eggNOG v3.0 : Orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.*, 40(D1) :D284–D289.
- Qin, J., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285) :59–65.
- Qin, J., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418) :55–60.
- Quince, C., et al. (2015). Extensive Modulation of the Fecal Metagenome in Children With Crohn’s Disease During Exclusive Enteral Nutrition. *Am. J. Gastroenterol.*, 110(12) :1718–29; quiz 1730.
- Raes, J., Letunic, I., Yamada, T., Jensen, L. J., and Bork, P. (2011). Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Mol. Syst. Biol.*, 7(1) :473.
- Robinson, M. D. and Smyth, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9(2) :321–332.
- Rondon, M. R., Goodman, R. M., and Handelsman, J. (1999). The Earth’s bounty : Assessing and accessing soil microbial diversity. *Trends Biotechnol.*, 17(10) :403–409.
- Ruas-Madiedo, P., Gueimonde, M., Fernandez-García, M., De Los Reyes-Gavilán, C. G., and Margolles, A. (2008). Mucin degradation by Bifidobacterium strains isolated from the human intestinal microbiota. *Appl. Environ. Microbiol.*, 74(6) :1936–1940.
- Scharlau, D., et al. (2009). Mechanisms of primary cancer prevention by butyrate and other products formed during gut flora-mediated fermentation of dietary fibre. *Mutat. Res. - Rev. Mutat. Res.*, 682(1) :39–53.
- Sokol, H., et al. (2008). Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc. Natl. Acad. Sci. U. S. A.*, 105(43) :16731–6.

- Stepankova, R., et al. (2010). Absence of microbiota (germ-free conditions) accelerates the atherosclerosis in ApoE-deficient mice fed standard low cholesterol diet. *J. Atheroscler. Thromb.*, 17(8) :796–804.
- Suau, A., et al. (1999). Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. *Appl Env. Microbiol.*, 65(11) :4799–4807.
- Tap, J., et al. (2009). Towards the human intestinal microbiota phylogenetic core. *Environ. Microbiol.*, 11(10) :2574–2584.
- Tasse, L., et al. (2010). Functional metagenomics to mine the human gut microbiome for dietary fiber catabolic enzymes. *Genome Res.*, 20(11) :1605–1612.
- Tseng, P. (2008). On accelerated proximal gradient methods for convex-concave optimization. *Submitt. to SIAM J. Optim.*
- Violle, C., et al. (2007). Let the concept of trait be functional! *Oikos*, 116(5) :882–892.
- Widder, S., et al. (2016). Challenges in microbial ecology : building predictive understanding of community function and dynamics. *ISME J.*, (November 2015) :1–12.
- Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms : proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U. S. A.*, 87(12) :4576–4579.
- Wu, G. D., et al. (2011). with Gut Microbial Enterotypes. *Science (80-.)*, 334(October) :105–109.
- Yang, Z., Rong, J., and Steven C., H. (2010). Exclusive Lasso for Multi-task Feature Selection. *Aistats*, 9 :988–995.
- Yatsunenko, T., et al. (2012). Human gut microbiome viewed across age and geography. *Nature*, 486(7402) :222–227.
- Yin, Y., et al. (2012). DbCAN : A web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.*, 40(W1) :445–451.
- Zhang, Q., et al. (2011). Understanding atmospheric organic aerosols via factor analysis of aerosol mass spectrometry : A review. *Anal. Bioanal. Chem.*, 401(10) :3045–3067.

Titre : Analyse de données de métagénomique fonctionnelle par NMF pour la modélisation de la dégradation des fibres par le microbiote intestinal humain.

Mots clés : Microbiome intestinal humain/NMF/Métagénomique/Modélisation/Réduction de données/Dégradations des Polysaccharides

Résumé : Ce travail de thèse a pour but de modéliser la capacité de dégradation des polysaccharides non digestibles par le microbiote intestinal humain. Nous exploitons pour cela des données métagénomiques. Il s'agit de données d'abondances de séquences de nucléotides dans 1408 échantillons dont les fonctions métaboliques sont assignées par annotation contre une base de données. Les séquences sont annotées par des marqueurs fonctionnels. Après une étape de sélection manuelle de 86 marqueurs fonctionnels pertinents à l'activité de métabolisation des polysaccharides, nous étudions leurs variations d'abondances parmi les échantillons métagénomiques.

Nous proposons une approche de modélisation écologique du microbiote intestinal humain et considérons principalement la sélection fonctionnelle intense de cet écosystème pour faire l'hypothèse que des regroupements identiques de fonctions métaboliques sont présents en proportions différentes dans tous les microbiotes intestinaux humains. Nous proposons le terme d'assemblage fonctionnel qui rend compte de la co-occurrence spatiale et temporelle d'un groupement de fonctions. Ces assemblages sont en pratique déterminés par leur composition en marqueurs fonctionnels, et peuvent s'interpréter comme une combinaison de traits fonctionnels agrégés au niveau des micro-organismes composant l'assemblage.

Les assemblages fonctionnels sont inférés par le biais d'une factorisation en matrice positive aussi nommée NMF de l'anglais Non-Negative Matrix Factorisation. Cette méthode permet de déterminer les assemblages fonctionnels, à la fois concernant leur composition et à la fois concernant leur abondance dans chacun des 1408 échantillons. Nous exploitons par ailleurs une information métabolique provenant de 190 génomes microbiens et de la bibliographie qui permet de préciser la composition de ces assemblages fonctionnels. Cette information se traduit sous forme d'une contrainte.

Nous trouvons 4 assemblages en considérant un consensus entre différents critères. L'utilisation de l'information métabolique nous permet d'interpréter biologiquement ces assemblages. Les métadonnées associées aux 1408 échantillons nous permettent d'observer un comportement différent pour les échantillons provenant d'individus atteints de la maladie de Crohn. Nous validons cette observation sur des données extérieures.

Nous avons proposé une approche réductionniste permettant de représenter un processus métabolique important à l'échelle du microbiote. Nous trouvons un nombre réduit de 4 assemblages fonctionnels qui sont biologiquement vraisemblables et permettent de bien approcher les 1408 échantillons métagénomiques.

Title : Modelling of fiber degradation by the human gut microbiota based on NMF analysis of functional metagenomic data

Keywords : Human Gut Microbiota/NMF/Metagenomic/Modeling/Data and model reduction/Polysaccharides degradation

Abstract : The purpose of this work of thesis is to model the capacity of degradation of non-digestible polysaccharides by the human intestinal microbiote. To this end we exploit metagenomic data. We use abundances of nucleotide sequences in 1408 samples whose metabolic function are assigned by annotation against a database. The sequences are annotated with functional markers. Upon manual selection of 86 functional markers relevant to the activity of metabolisation of polysaccharides, we their abundances variation among the metagenomic samples are studied.

We propose an ecological approach in modeling the human intestinal microbiote. We consider the intense functional selection of this ecosystem and assume that identical cluster of metabolic functions can be found in different proportions in every human gut microbiota. We propose the term of functional assembly as to account for spacial and temporal co-occurrence of functional cluster. In practice, these assemblies are determined by their composition and can be interpreted as combinations of functional traits aggregated at the levels of the cluster of micro-organisms composing each assembly. Functional assemblies are inferred by the means of Non-Negative Matrix Factorization (NMF). This method allows to determine the composition of functional assemblies and their abundance in each of the 1408 metagenomic sample.

Furthermore, we exploit metabolic information from bibliographic resources and 190 microbial genomes in order to specify the composition of these functional assemblies. This information is translated in the form of a constraint.

We find 4 assemblies by considering a consensus between various criteria. The use of metabolic information allow to interpret these assemblies biologically. By exploiting the meta-data of the 1408 samples, we observe a different behaviour for the samples coming from individuals suffering from Crohn disease. We validate this observation on external data.

We proposed a reductionistic approach allowing to represent an important metabolic process at the level of the microbiota. We find a small number of 4 functional assemblies which are biologically likely and approach well the 1408 metagenomic samples.