# Combining knowledge-based and sequence comparison approaches to elucidate metabolic functions, from pathways to communities

Arnaud Belcour

DOCTORAT BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

# THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES 1

ÉCOLE DOCTORALE N° 601
*Mathématiques et Sciences et Technologies de l'Information et de la Communication*
Spécialité : *Informatique*

Par
**Arnaud Belcour**

## « Combining knowledge-based and sequence comparison approaches to elucidate metabolic functions, from pathways to communities »

**Thèse présentée et soutenue à Rennes, le 21 octobre 2022**
**Unité de recherche : Institut de Recherche en Informatique et Sytèmes aléatoires (IRISA), UMR 6074**

### Rapportrice et rapporteur avant soutenance :

| | |
|---|---|
| Delphine Ropers | Directrice de recherche, INRIA Grenoble |
| David Vallenet | Chercheur, CEA Genoscope Évry-Courcouronnes |

### Composition du Jury :

| | | |
|---|---|---|
| Président : | Cédric Lhoussaine | Professeur Université de Lille, Lille |
| Examinateurs et examinatrice : | Olivier Dameron | Professeur, Université de Rennes 1 Rennes |
| | Karoline Faust | Professeure associée, KU Leuven Leuven |
| | Fabien Jourdan | Directeur de recherche, INRAE Toulouse |
| Directrice de thèse : | Anne Siegel | Directrice de recherche, CNRS, IRISA Rennes |
| Encadrant de thèse : | Samuel Blanquart | Chargé de recherche, INRIA, IRISA Rennes |

# Page with pretentious quotes

Tee-Tee-Too Tee-Tee-Too

*A Great Tit in a backyard in Corrèze*

**Treatle**:
I hadn't looked at it like that, but you're absolutely right.
He's really pushed back the boundaries of ignorance.

They both savoured the strange warm glow of
being much more ignorant than ordinary people,
who were only ignorant of ordinary things.

*Terry Pratchett, Equal rites*

They say that life's a carousel
Spinning fast, you've got to ride it well
The world is full of kings and queens
Who blind your eyes and steal your dreams
It's heaven and hell, oh well

And they'll tell you black is really white
The moon is just the sun at night
And when you walk in golden halls
You get to keep the gold that falls
It's heaven and hell, oh no

Fool, fool
You've got to bleed for the dancer
Fool, fool
Look for the answer
Fool, fool, fool

*Black Sabbath, Heaven And Hell*

Did I bleed for the dancer?

# Long list of Acknowledgements

(As a first side note, the acknowledgments are used by the appointed PhD Student to decompress after such experience, so it contains some jokes, thoughts and acknowledgements (and are potentially a little weird as I am bad for this practice (as for all other social practices)). Do not take it too seriously but at some time, yes take it seriously.)

So here are the acknowledgements in this horrible language that is English (surely created by terrible people, I am sure it is by the British people!). I mean in France we have a very go..., a decent langua... Well at least it is French!

To reassure people reading this document, no animals or PhD student were harmed during this thesis (the latter was quite stressed but it's okay). But in most of this thesis bacteria and archaea were numerically harmed and I am sure nobody care about their well-being! Whereas they have metabolisms and very interesting ones!

So I would like to thank:

— all the members of the jury for accepting to be in this jury. At this time, I do not have suffered their questions so I have quite a positive view of them (I think because I do not know them). Many thoughts and thanks to the thesis rapporteurs. I thank them for engaging a reading fight against this long, monstrous, hideous document (it is so big and ugly that some people say there exists community of microorganisms in it).

— Anne and Samuel for the supervision of this thesis. It was quite a long travel. But as the famous expression says: "it is not the goal that is important, it is the number of encounters that you successfully avoided that is important' (surely a phrase that will be accepted by Rincewind). I thank them for all their works, advices, long meeting during more than 2 hours on a minimal point of a subsection of this thesis. Without them, this thesis will not be the same. What? Ah yes, without them there will be no thesis.

— Anne (second round of acknowledgements) for her availability despite all of work, all her advice on how to do a presentation or on how to structure my thoughts. And I will never forget to add "Wahou" effect in my articles (easily done by adding chocolate crepe in them).

— Samuel (second round of acknowledgements), for all his work and help. The discussions we have were very interesting and informative. And many thanks for proofreading this manuscript, it was an enormous work!
— the INRIA (or the IRISA or both, damned I am still confused about them :b), the MathSTIC Doctoral School (before it changes to MATISSE) and the Université de Rennes 1 to give me the opportunity to do this PhD thesis. A special thanks to Amandine Seigneur for her work and answers to help a very stressed PhD student.
— the two members of my CSID (Philippe Vandenkoornhuyse and David Sherman). The two CSIDs in the first and second years were really helpful and I was glad to discuss the thesis subject with you.
— Alexandra Elbakyan for her work and for creating Sci-Hub. Without her, lots of knowledge would have been inaccessible (and this thesis would have contained less citations). Even if her statements on Staline are weird.
— all people contributing to open source software. I try to do my best to make all my developed tools open source. especially many thanks to all people who create and maintain a documentation of their tools. You are truly heroes.
— the team of Pathway Tools for their work and their answers to my questions.
— I would never be able to make this thesis without music so here is (a very tiny) list of artists/groups that I would like to thank: Aephanemer, Rainbow, Black Sabbath, Dio, Iron Maiden, B.B. King, R.L. Burnside, Roy Buchanan, Eric Clapton, Blind Guardian, Damjan Mravunac, John Williams, Stupeflip, Howard Shore, Clamavi De Profundis, Sofi Tukker, Banshee (the band led by Rachel Knight behind Fairy Metal to be more precise as there is multiple Banshee band), Bear McCreary, Nemesis (the serbian band), Samael, Plasmatics. And so many more, but it is not reasonable to cite all of them.
— special thanks and thoughts to 3 bands (Ignea, Jinjer and Go_A) as your country endures this war.
— the writers and their wonderful (or not) worlds: J. R. R. Tolkien, Terry Pratchett, Ursula K. Le Guin, Steven Erikson and so many more. Too many books and so little time. And also to the video game developers Creative Assembly, From Software, Spiders, Ninja Theory, Funcom and many others.

Et maintenant, les remerciements en français, merci:

— à toutes les personnes ayant travaillé sur le template LaTeX de l'école doctorale MathSTIC. L'écriture de cette thèse en LaTeX a simplifié tellement de chose, je

n'imagine pas l'avoir fait autrement (oui Word/LibreOffice, c'est VOUS que je regarde). Au passage, merci à Overleaf pour être un super éditeur LateX en ligne.
— aux équipes de la DSI pour m'avoir sorti du pétrin un certain nombre de fois. Un merci spécial à Philippe, Philippe et Bruno. Même si je crois que la source des pépins se trouve encore à l'interface siège/clavier.
— à l'environnement symbiose (les équipes Dyliss, Genscale et Genouest) qui est un environnement super dans lequel j'ai adoré travaillé. Ces années ont été super. Merci pour les discussions passionnantes que cela soit à la cafet', sur discord, lors de pot de stagiaires. Et je ne peux pas ne pas parler des séminaires au vert dont un certain en particulier qui a tant fait pour souder les équipes.
— à l'équipe Genouest pour leurs aides et leurs conseils sur les utilisations du cluster de calcul. Les trois quart des travaux de cette thèse repose sur des expériences qui ont pu être réalisées grâce à vous, merci! Et bien sur le petit message: "We acknowledge the GenOuest bioinformatics core facility (https://www.genouest.org) for providing the computing infrastructure." :b
— au sein de l'équpie, des mercis infinis à Marie. Merci pour ton travail surtout quand on voit l'imbroglio adminstratif qu'est la recherche française et les politiques des différents instituts.
— à toutes les personnes de l'équipe avec qui j'ai discuté pendant les repas du midi et les pauses à la cafet' (Olivier(s), Jacques, Nicolas, François(s), Emmanuelle, Pierre(s), Roland, Lucas, Clara, Anthony, Matéo, Fabrice, Victor, Thomas(s), Emeline, Camille(s), Khodor, Baptiste, Kerian, Téo, Garance, Meven). Un merci spécifique aussi au groupe "Midi les doctorants" (Méline, Lolita, Wesley, Hugo, Maël, Marine, Lucas, Grégoire, Anne) c'était génial de suivre et participer à ces élucubrations partant dans tous les sens. Merci à Maël pour les parties de Magic après le repas, elles étaient bien fun.
— à Victor pour la magnifique affiche de thèse.
— à Olivier, c'est grâce à toi que j'ai découvert l'équipe au travers de mon stage sur AskOmics. Merci pour tes explications, ton travail. Toutes les conversations sur les technologies du Web sémantique m'ont bien marqué. J'attends la publication de la chanson d'AskOmics avec impatience.
— à Jacques, ce fut enrichissant de travailler avec toi sur PathModel.
— à tout le sous-groupe avec qui j'ai travaillé sur le métabolisme dans l'équipe Dyliss (Mézaine, Clémence, Jeanne, Anne).

— à Clémence, travailler avec toi est un plaisir, je suis fier de ce que l'on a réussi à faire avec Metage2Metabo.
— à Lucas pour les discussions, les échanges sur clyngor/powergrasp.
— aux stagiaires (Adriana, Rania, Baptiste, Jacky, Moana, Pauline) pour leurs questions et leurs crash tests de mes outils.
— aux français et françaises qui graĉe à leurs impôts ont financé cette thèse, étant donné qu'elle dépend d'un institut publique. Enfin pour ceux qui payent leurs impôts en France (è-é).
— à tous les gens de la Station de Biologique de Roscoff avec qui j'ai eu la chance de collaborer: Gabriel, Simon, Catherine, Jean, Erwan, Jonas, Ludovic. Cela a été très enrichissant de travailler avec vous. Merci aussi au projet IDEALG et à toutes les personnes impliquées et notamment Philippe Potin. Cela a été un peu un cliché de travailler sur les algues en Bretagne mais j'ai appris plein de choses (et ma famille aussi par ricochet). Et puis bon, ça a été super de tester des chips d'algues (si vous vous posez la question c'est super bon avec un goût iodé). Un merci specifique à Gabriel, le travail sur Pathmodel a été un labeur long mais stimulant et voir les résultats prédits fut une belle récompense.
— à Patrick. Cela a été passionnant de travailler avec toi et de parler de méthanisation. J'espère qu'on va réussir à publier l'article présentant cette collaboration mais ça me semble faisable en mettant les gaz. Merci pour la relecture de la thèse sur les parties concernées.

Ah et aussi parmi tous ces mercis, un désolé pour tous les gens que j'ai bassiné avec ma thèse pendant ces dernières années (cela inclus les trois quart des gens rencontrés).

Et bien évidément, je termine par remercier ma famille sans qui je ne serais pas là. Merci pour le soutien, pour les vacances passées ensemble et qui étaient primordiales pour la réussite de la thèse. Merci à mes parents d'avoir cru en moi pendant toutes ces années d'études et de m'avoir laissé choisir cette voie (si on peut appeler cela une voie).

Chère lectrice, cher lecteur, merci d'avoir parcouru cette liste, en récompense voici l'une de mes blagues préférées:

Savez-vous ce que dit un requin mangeant un mérou?

"Aie, je suis tombé sur un os!"

Car c'est un Chondrichtyen (poisson à squelette cartilagineux) qui vient de manger un Ostéichtyen (poisson à squelette osseux).

# Table of Contents

# Résumé en français

La bioinformatique est un domaine scientifique visant à expliquer des phénomènes biologiques grâce à des approches informatiques. Elle est à l'interface entre biologie et mathématique et applique différentes méthodes et techniques (programmation par contraintes, comparaison de séquences, ingénierie des connaissances, etc) à des données provenant de sciences biologiques, chimiques ou physiques. L'une des applications de la bioinformatique (en relation avec la biologie des systèmes) est le développement de modèles visant à prédire le fonctionnement de phénomènes biologiques.

Parmi ces phétnomènes biologiques, il y a le métabolisme, l'ensemble des transformations chimiques menant à la modification de composés pour assurer la vie de la cellule (en chargeant ou déchargeant l'énergie, en recréant des composés nécessaires au fonctionnement de la cellule, etc). Certaines des transformations chimiques sont catalysées par des enzymes. Les enzymes sont issues de la traduction des ARN messagers qui eux-même sont transcrits à partir des gènes, qui sont des régions précises présentes sur le génome d'un organisme. Grâce aux nouvelles méthodes de séquençage, il est possible de connaître la séquence des génomes et ainsi des gènes. De nombreuses analyses sur ces séquences, couplées à des expériences biologiques ont permis d'associer à certains gènes une fonction précise et parfois d'identifier quels gènes vont être transcrits puis traduits en enzymes.

La thèse s'inscrit dans ce domaine de la bioinformatique en se mettant à l'interface de la comparaison des séquences génomiques, de l'ingénierie des connaissances et de la programmation par contraintes. Elle a permis de développer des méthodes pour intégrer et analyser des données et connaissances relatives au fonctionnement de la cellule, d'organismes et d'interactions entre organismes.

## Prédiction de voies métaboliques alternatives à partir de données métabolomiques

Au sein des bases de données sur le métabolisme, les réactions biochimiques sont souvent regroupées en des ensembles produisant des composés d'intérêts qui sont appelés

voies métaboliques. Ces voies métaboliques ont été décrites à partir d'organismes modèles. Mais elles ne sont parfois pas adaptées à d'autres organismes (à cause d'un manque de certains composés). Pour essayer de prédire des voies alternatives pour des organismes non-modèles, le Chapitre 2 décrit une méthode se basant sur la dérive de voie métabolique et propose un formalisme de cette dérive en l'implémentant en utilisant une programmation par contrainte. Cette méthode a été développée en collaboration avec Jacques Nicolas (INRIA, Rennes) et Gabriel Markov (Station Biologique de Roscoff). La dérive de voie métabolique est définie comme similaire à *la dérive du système de développement* (True et al. 2001). Chez des organismes présentant des similarités en terme de morphologie, des analyses comparatives ont montré que les systèmes amenant au développement de ces caractéristiques similaires divergeaient. C'est ainsi qu'a été proposée la dérive du système de développement à savoir que des changements peuvent être appliqués aux systèmes gérant le développement de phénotypes sans que le phénotype n'en soit modifié (True et al. 2001). Il n'y a ainsi pas d'effet de la sélection naturelle mais à la manière de la dérive génétique, des changements apparaissant par hasard. Une dérive similaire pourrait être appliquée au métabolisme. Ainsi des changements pourraient avoir lieu au sein d'une voie métabolique sans que cela n'entraîne un changement dans les composés finaux produits par la voie métabolique. C'est en se basant sur ce concept que nous avons développé un prototype visant à prédire des voies métaboliques alternatives à partir d'une voie métabolique de référence et des données métabolomiques (qui permettent de mesurer la présence de composés à travers leur masses et leurs charges chez un organisme).

En partant d'une voie métabolique de référence (obtenue depuis des bases de données métaboliques et/ou de la littérature), des données métabolomiques indiquant la présence de composés (et permettant de détecter des composés non encore connus et donc seulement identifiés par un rapport masse-sur-charge) et de la structure des composés métaboliques, la méthode PathModel va essayer d'inférer de possibles voies alternatives. Pour se faire, une base de connaissance est créée manuellement. Cette base contient un encodage simplifié des réactions enzymatiques où ne sera encodé que le réactant et le produit particulier en enlevant les cofacteurs de la réaction. De plus, les structures chimiques des composés sont encodées de manière consistante, c'est-à-dire que l'atome 1 du composé A correspond à l'atome 1 du composé B. Cette consistance dans l'écriture permet la comparaison des structures et l'identification de transformations chimiques qui sont modélisées comme les changements possibles de structures entre le réactant et le produit. De plus, les ratios masse-sur-charge des molécules non identifiées sont ajoutés dans la base de connaissance.

Les voies métaboliques sont enfin décrites avec des composés d'entrée et des composés de sortie. L'ensemble de cet encodage est fait dans le langage de programmation par contraintes Answer Set Programming (ASP). De cette manière, il sera possible d'appliquer des règles logiques pour explorer la combinatoire des changements possibles amenant à la production des composés de sortie de la voie métabolique.

La base de connaissance est donnéee comme entrée à une méthode se reposant sur de la programmation logique (Lifschitz 2008; Gebser et al. 2019) qui va appliquer itérativement plusieurs raisonnements. Un premier raisonnement vise à prédire de possibles transformations chimiques entre des composés connus en utilisant des réactions déjà connues provenant de bases de données métaboliques ou de la littérature. Cela permet de proposer des ensembles de transformations ne se reposant pas uniquement sur les composés présents dans notre voie métabolique. Il sera ainsi possible de rattacher à la voie des composés connus chez notre organisme d'intérêt. Le second raisonnement vise à chercher des composés dans la base de connaissance dont la transformation chimique amènerait à la production d'un composé dont le ratio masse-sur-charge est égal à celui d'un ratio masse-sur-charge n'ayant pas été identifié. Cela permet de valider par un ratio mesuré un composé intermédiaire (prédit) à notre voie métabolique. Ces deux raisonnement sont appliqués pour prédire l'ensemble des transformations alternatives permettant de passer du composé d'entrée aux composés de sortie.

PathModel a été appliqué sur deux voies métaboliques de l'algue rouge *Chondrus crispus*, la voie des stérols et la voie des acides-aminés analogues de la mycosporine. Pour la voie des stérols cela a permis de prédire une voie alternative avec un renversement dans l'ordre des enzymes utilisées chez *C. crispus*. Et pour les acides-aminés analogues de la mycosporine, il a été possible de rattacher deux ratios masse-sur-charge inconnus à la voie connue. C'est ainsi que nous avons pu proposer des voies alternatives pour deux voies présentes chez *C. crispus*. Cela permet ainsi de compléter les connaissances sur le métabolisme de l'algue, en ajoutant ces voies aux centaines de voies composant le métabolisme de l'organisme.

## Inférence de réseaux métaboliques comparables à partir de données hétérogènes

La seconde échelle d'analyse examinée dans mes travaux se concentre, non plus une voie métabolique contenant des dizaines de réactions mais sur l'ensemble du métabolisme d'un

organisme étant représenté par un graphe contenant des milliers de réactions et de composés. Plus particulièrement, le Chapitre 3 a visé à développer une méthode permettant d'homogénéiser la reconstruction de graphes métaboliques pour un groupe d'organismes en vue de comparer leurs métabolismes. Cette comparaison pourra se faire, par exemple, en comparant la présence ou l'absence de réactions biochimiques dans les réseaux reconstruits. Notamment, l'approche vise à être capable de reconstruire des graphes métaboliques à partir de génomes présents dans des bases de données publiques. Le problème rencontré est que les génomes présents dans ces bases de données ont souvent été annotés avec différents outils amenant à une hétérogénéité dans leurs annotations. Il peut y avoir des problèmes avec l'annotation structurelle, c'est-à-dire la prédiction de gènes au sein du génome pouvant amener à l'absence de certains gènes (Ejigu et al. 2020), dont certains pouvant réaliser une activité enzymatique. Et il peut aussi y avoir des problèmes avec l'annotation fonctionnelle, qui est l'étape visant à l'association d'un gène avec une fonction et donc une possible activité enzymatique. Une erreur lors de ces étapes peut empêcher l'annotation d'une enzyme et ainsi indiquer comme manquante une réaction pourtant bien présente dans le métabolisme de l'organisme. Ainsi, ces analyses détectent des différences entre les réseaux métaboliques de plusieurs organismes provenant de biais d'annotations et non de différences biologiques (Nègre et al. 2019).

Pour répondre à ce problème, nous avons développé AuCoMe, un outil visant à reconstruire de manière homogène le métabolisme de plusieurs organismes en vue de leur comparaison. AuCoMe est un workflow enchaînant des méthodes de comparaison de séquences pour réaliser l'homogénéisation des réseaux. La méthode prend en entrée des génomes provenant de bases de données publiques et pouvant donc être annotés de manière hétérogène. Puis, une première étape va reconstruire des réseaux métaboliques en se basant sur les annotations présentes dans les génomes en utilisant l'outil Pathway Tools (Karp et al. 2002a; Karp et al. 2019; Karp et al. 2021). Une recherche de gènes orthologues est réalisée avec OrthoFinder (Emms et al. 2015; Emms et al. 2019) pour trouver des groupes d'orthologues entre les différents génomes étudiés. En se basant sur ces groupes d'orthologues, AuCoMe va propager les réactions prédites avec Pathway Tools suivant un score de robustesse. Le score dépend du nombre d'orthologues dans le groupe et des orthologues au sein du groupe ayant la réaction selon Pathway Tools. Une réaction est propagée à un groupe d'orthologues si le gène ayant la réaction selon Pathway Tools est aussi orthologue à un autre gène ayant cette réaction ou si le nombre d'orthologues présent dans le groupe est inférieur à un seuil. Cela vise à limiter les propagations de réactions pour

les groupes contenant de nombreux orthologues et peu de réactions prédites par Pathway Tools. Puis, une troisième étape va comparer les réseaux deux à deux pour identifier les réactions absentes dans les réseaux. Pour ces réactions manquantes, une recherche dans le génome de l'organisme concerné est réalisée avec les outils blast (Altschul et al. 1990; Camacho et al. 2009) et exonerate (Slater et al. 2005), avec en entrée les séquences des protéines étant associées à cette réaction. De cette façon, AuCoMe vérifie s'il n'y a pas eu un problème lors de la prédiction structurelle pour le gène associé à cette réaction dans l'organisme d'intérêt. Une dernière étape est réalisée durant laquelle AuCoMe va chercher à compléter les voies métaboliques en ajoutant les réactions spontanées. Après ces différentes étapes, des réseaux homogénéisés à partir de leurs annotations respectives sont reconstruits.

AuCoMe a été appliqué à 3 jeux de données sur différents groupes taxonomiques (29 bactéries, 36 algues et 74 champignons). Les réseaux reconstruits avec la première étape sont extrêmement hétérogènes dus à des différences dans leurs annotations fonctionnelles. Certains réseaux sont vides alors que d'autres contiennent des milliers de réactions. La seconde étape en propageant les réactions parmi les orthologues permet une homogénéisation des contenus en réactions et de créer des réseaux ayant des nombres similaires de réactions (autour de quelques milliers). La troisième étape cherchant à vérifier si l'absence de réactions n'est pas due à une erreur dans la prédiction structurelle n'a eu un effet significatif que pour 2 génomes (un du jeu de données algue et un du jeu de données champignon). L'effet de cette troisième étape est ainsi limité sur les jeux de données testés. Puis l'étape finale de l'outil a permis de compléter une vingtaine de voies métaboliques en ajoutant des réactions spontanées.

Pour valider la méthode une seconde expérience a été réalisée sur le jeu de données bactériens. Le génome de la souche *Escherichia coli* K–12 MG1655 a été dégradé de plusieurs façons, soit en enlevant les annotations fonctionnelles des gènes soit en supprimant complètement les gènes du génome. De cette façon, 31 réplicats ont été créés. Chaque réplicat a été associé à 28 autres génomes (non dégradés) de bactéries et donné en entrée à AuCoMe. Cela a permis de montrer que si l'annotation fonctionnelle est dégradée (les annotations associées aux gènes) l'étape de propagation à travers les orthologues permet de récupérer la majorité des réactions. Si les gènes sont supprimés du génome c'est alors la troisième étape vérifiant l'absence des réactions en alignant des protéines aux génomes qui permet de récupérer les réactions. Et si les deux dégradations sont appliquées, la combinaison des deux étapes permet de récupérer les réactions perdues lors de la dégradation. De

cette manière, l'expérience indique la complémentarité des deux étapes suivant l'état des génomes donnés en entrée.

Finalement, une comparaison des réseaux métaboliques des algues a été réalisée. En réalisant un clustering des algues suivant l'absence ou la présence de réactions, il a été possible d'étudier la similarité des réseaux métaboliques. Les réseaux reconstruits après la première étape présentent une grande variabilité et ne sont pas regroupés suivant les groupes phylogénétiques auxquels ils appartiennent mais plutôt suivant leurs niveaux d'annotations. C'est notamment le cas pour les génomes peu annotés qui se retrouvent regroupés ensembles. Après l'ensemble des étapes d'AuCoMe, le clustering sur les réseaux métaboliques crée des groupes qui sont cohérents avec les groupes connus de la phylogénie. Une analyse plus fine compare l'arbre phylogénétique des algues et l'arbre obtenu à partir du clustering suivant la présence ou l'absence de réactions. Cette analyse montre de nouveau la cohérence des groupes suivant la phylogénie ou le métabolisme. Par contre sur des échelles plus fines des divergences entre les arbres sont présentes notamment pour des espèces. Ces résultats sont assez proches d'autres résultats montrant une cohérence entre les groupes connus phylogénétiquement et des regroupements suivant le métabolisme et plus de divergences dans des granularités plus fines comme le placement des espèces.

AuCoMe permets ainsi l'étude du métabolisme à partir de génomes hétérogènes pour permettre la comparaison à grande échelle de réseaux métaboliques. Il a été appliqué et validé sur 3 jeux de données et a permis l'exploration de la diversité du métabolisme chez les algues. Cette exploration a permis d'explorer le métabolisme d'un groupe d'organismes à partir de leurs génomes.

## Estimation des capacités métaboliques à partir d'une affiliation taxonomique

Les analyses réalisées par l'outil précédemment décrit s'appuient sur les génomes des organismes. Mais pour de nombreux organismes, il n'y a pas de génome disponible. Pour la majorité des expériences en métagénomique seuls certains gènes marqueurs sont séquencés. Ces séquences peuvent être associées à des affiliations taxonomiques, c'est à dire l'appartenance du gène marqueur à un groupe d'organismes apparentés (taxon). Pour analyser ces données d'un point de vue métabolique, il faudrait pouvoir estimer le métabolisme, non pas d'un organisme mais d'un groupe d'organisme (taxon). Des méthodes existent déjà pour prédire des profils fonctionnels à partir de séquences de gènes

marqueurs (Langille et al. 2013; Douglas et al. 2020; Bowman et al. 2015; Aßhauer et al. 2015; Wemheuer et al. 2020). Mais les profils fonctionnels ne permettent pas d'obtenir des réseaux métaboliques. D'autres méthodes ont été développées qui vont chercher les réseaux métaboliques des espèces les plus proches des séquences des gènes marqueurs pour avoir des réseaux métaboliques mais cela peut être problématique si le marqueur diverge fortement des plus proches espèces connues (Cardona Uribe 2019; Mendes-Soares et al. 2016). De plus, l'affiliation taxonomique ne correspond généralement pas au niveau de l'espèce mais à des rangs taxonomiques plus élevés (genre ou famille)à des rangs taxonomiques plus élevés (genre ou famille). Pour cela il y a besoin d'une méthode pour prédire les capacités métaboliques à partir d'une affiliation taxonomique et de proposer des réseaux métaboliques associés à des taxons.

C'est pour répondre à ce besoin qu'une nouvelle méthode (EsMeCaTa) se basant sur l'ingénierie des connaissances et la comparaison des séquences a été développée et est présentée dans le Chapitre 4. Cette méthode vise à prédire un ensemble de protéines annotées à partir d'une affiliation taxonomique. La première étape parcourt l'affiliation taxonomique et les taxa qu'elle contient. En partant du rang taxonomique le plus bas possible (par exemple le genre) et en remontant jusqu'au rang taxonomique le plus élevé (par exemple le royaume), EsMeCaTa va interroger la base de données de protéines UniProt (The UniProt Consortium 2021) pour trouver s'il existe des protéomes associés à chacun des taxa. Puis, le taxon associé à au moins un protéome (ou N protéomes en fonction d'une option) et ayant le rang taxonomique le plus bas est sélectionné. Durant la deuxième étape, un clustering est réalisé sur les protéines contenues dans l'ensemble des protéomes associés à un taxon en utilisant MMseqs2 (Steinegger et al. 2017). Cela permet de créer des groupes de protéines homologues provenant des différents protéomes. Ces groupes de protéines sont filtrés suivant un seuil de représentativité des protéomes. Ainsi pour chaque groupe de protéines, EsMeCaTa compte combien des protéomes trouvés pour le taxon sont représentés par des protéines dans le groupe et calcule un ratio de représentativité. Puis suivant le ratio sélectionné par l'utilisateur (de base il est à 0.95 ce qui implique que 95% des protéomes du taxon ont une protéine dans le groupe) les groupes de protéines ayant un ratio de représentativité supérieur à ce seuil sont conservés. EsMeCaTa va ensuite annoter ces groupes de protéines aves des requêtes sur la base de données UniProt. Ces annotations permettent de connaître les fonctions réalisées par les groupes de protéines. Au sein de ces fonctions il est possible de connaître les fonctions enzymatiques. Cela permet de repérer les enzymes associées à un taxon et d'ainsi avoir une estimation des capacités

métaboliques.

EsMeCaTa a été utilisée dans le cadre d'une expérience de métagénomique réalisée par Patrick Dabert (UR OPAALE, INRAE) visant à comprendre le fonctionnement d'un méthaniseur au cours du temps et suivant le traitement de différents déchets (lisier, pomme, beurre). Une première utilisation d'EsMeCaTa sur le jeu de données est réalisée en utilisant les options par défaut de l'outil. La première étape d'EsMeCaTa a montré que le nombre de protéomes trouvé par l'outil dépend du rang taxonomique sélectionné par EsMeCaTa. Ainsi, les rangs taxonomiques bas (genre, famille) sont associés à des petits nombres de protéomes associés à un taxon (10-20) alors que des rangs taxonomiques plus élevés (ordre, phylum) amènent à trouver plus de protéomes (entre 80 et 120). Cela a un impact sur les étapes ultérieures car avec une petite diversité de protéomes, il y aura plus de protéines partagées et donc conservées après l'étape de clustering quand on se base sur le seuil de représentativité par défaut à 0.95. Ce qui amène à avoir plus d'annotations quand le rang taxonomique est au niveau du genre ou de la famille et à peu d'annotations quand le rang taxonomique est élevé. Ces résultats ont ensuite été donnés à des méthodes de Machine Learning pour essayer de comprendre les différences entre les phases dans lequel le méthaniseur était considéré comme étant fonctionnel et les phases où le méthaniseur n'était plus fonctionnel. La méthode de Machine Learning appliquée aux abondances des taxa a permis de trouver les groupes microbiens discriminant un méthaniseur fonctionnel et inversement ceux discriminant un méthaniseur non fonctionnel. Pour les méthaniseurs fonctionnels, il y a un groupe contenant à la fois des bactéries et des archées méthanogènes. Alors que pour les méthaniseurs non fonctionnels il n'y a pas d'archées méthanogènes et que des bactéries. Une seconde approche de Machine Learning a été appliquée pour trouver les annotations fonctionnelles différenciant le fonctionnement des méthaniseurs. Cette approche a permis de séparer les 2 états des méthaniseurs et de détecter les annotations associées à la méthanisation.

Une seconde expérience a été réalisée sur le jeu de données. Cette expérience a pour but d'étudier l'impact du seuil de représentativité des protéomes sur les prédictions. Pour cela, 5 seuils ont été sélectionnés (0, 0.25, 0.5, 0.75 et 0.95) et différentes options de sélection des protéomes par EsMeCaTa ont été utilisées (au moins 5 protéomes par taxon et un rang taxonomique inférieur ou égal à la famille). En moyenne 33 protéomes ont été retrouvés par taxon. Plus le seuil de représentativité des protéomes est bas et plus le nombre de groupes de protéines sélectionnés est grand. Inversement, plus ce seuil augmente et plus le nombre de groupes conservés diminue. La même dynamique est présente pour les

annotations fonctionnelles retrouvées par EsMeCaTa suivant les seuils de représentativité. A partir de ces données des réseaux métaboliques associés aux taxa ont été reconstruits et le nombre de réactions dans les réseaux suit une dynamique similaire.

Ainsi, EsMeCaTa permet de prédire des fonctions métaboliques associées à des taxa. Cette méthode propose des estimations des capacités métaboliques pouvant être utilisées pour étudier le métabolisme d'organismes n'ayant été identifiés que par des assignations taxonomiques. Obtenir des réseaux métaboliques pour ce type de données permet d'essayer d'identifier des coopérations métaboliques au sein des communautés étudiées.

## Identification d'espèces clés au travers de la complémentarité métabolique

Ainsi, avec les résultats d'EsMeCaTa il est possible de prédire le métabolisme de taxa associés aux organismes détectés dans le milieu. Cela ouvre la voie à l'analyse de complémentarité métabolique possible entre groupes d'organismes identifiés qui est le quatrième niveau du métabolisme étudié dans cette thèse, la communauté. Un problème rencontré dans les simulations du métabolisme de communauté est en particulier le passage à l'échelle pour tester les interactions possibles au sein de communautés à grande échelle.

Pour résoudre ce problème, une méthode (présentée dans le Chapitre 5) a été développée en collaboration avec Clémence Frioux (INRIA, Bordeaux) pour prédire les interactions métaboliques potentielles au sein de communautés contenant des centaines d'organismes. Cette méthode prend en entrée des génomes ou des réseaux déjà reconstruits puis prédit des complémentarités métaboliques en résolvant des problèmes d'optimisation combinatoire avec des approches de programmation par contraintes. La première étape (pouvant être optionnelle) vise à reconstruire des réseaux métaboliques à grande échelle. Pour cela un outil a été développé pour utiliser plusieurs processus de Pathway Tools en parallèle sur différents coeurs d'un processeur. Cela permet un passage à l'échelle pour l'obtention des réseaux métaboliques. Ensuite, les capacités de production individuelles des réseaux métaboliques reconstruits sont calculées à partir d'un ensemble de métabolites appelé graines (qui représentent les nutriments présents dans le milieu de culture). Pour ce faire l'outil se repose sur une analyse topologique des graphes métaboliques grâce à de la programmation par contraintes avec l'outil MeneTools (Aite et al. 2018). En partant des métabolites graines, les réactions utilisant les graines comme réactant sont activées amenant à la production de leurs produits. Ces produits deviennent ainsi utilisables pour

d'autres réactions qui peuvent être activées à leur tour. Cela amène à une expansion des métabolites productibles et permet de voir les métabolites productibles dans le graphe du réseau métabolique. Puis, une méthode similaire est appliquée pour trouver les métabolites productibles par l'ensemble de la communauté. Cette étape se repose sur l'outil MiSCoTo (Frioux et al. 2018a) pour calculer la production de la communauté en autorisant les échanges entre organismes sans coût. Il est ainsi possible de voir les métabolites productibles par la communauté. Ensuite, Metage2Metabo utilise MiSCoTo (Frioux et al. 2018a) pour calculer les communautés minimales (c'est-à-dire les plus petites communautés possibles) permettant de produire soit des métabolites cibles soit les métabolites uniquement productibles par des échanges entre membres de la communauté. L'analyse des communautés minimales ainsi produites permet de détecter des espèces clés dans les interactions métaboliques avec des organismes apparaissant dans toutes les communautés minimales (appelés symbiote essentiel) et des organismes apparaissant dans certaines des communautés mais pas toutes (symbiote alternatif). L'ensemble de ce workflow permet de détecter les interactions métaboliques potentielles entre organismes et d'identifier les acteurs clés de ces interactions. La méthode est présentée avec un exemple réalisé sur un microbiote de l'intestin humain contenant 1520 génomes pour montrer le potentiel de coopération métabolique entre ces organismes.

La méthode a ensuite été appliquée aux données du méthaniseur et notamment aux réseaux prédits. Une première étape a été de tester les réseaux obtenus sans les filtres sur les protéomes pour identifier les espèces clés pour la production du méthane et de quelques autres métabolites clés (acétate, dioxyde de carbone, etc). Cette première analyse a donné peu de résultats car les communautés minimales n'étaient composées que de 2 organismes une archée méthanogène et un taxon bactérien (le genre *Alcaligenes*). L'un des problèmes étant que le taxon *Alcaligenes* était associé à un unique protéome et avait donc accès à toutes les protéines et les annotations de ce protéome. Pour résoudre ce problème, nous avons ensuite utilisé les résultats de la seconde expérience d'EsMeCaTa avec les filtres demandant au moins 5 protéomes par taxon et un rang taxonomique inférieur ou égal à la famille. De plus, nous avons testé les 5 seuils de représentativité. Plus le seuil de représentativité diminue et plus le nombre de métabolites productibles par la communauté augmente. La même tendance s'observe pour les productions individuelles. Par contre le nombre de métabolites uniquement productibles par la communauté ne varie pas. Les tailles des communautés minimales diminuent avec la diminution des seuils. Mais le nombre d'espèces clés est variable avec plus de 20 espèces clés pour les seuils à 0.95 et

0 alors qu'il y a environ 5 espèces clés sélectionnées pour les autres seuils. L'analyse des communautés minimales permet d'identifier des espèces clés de la méthanogénèse comme des archées méthanogènes (*Methanosarcina*, Methanobacterium) et des bactéries connues pour leurs implications dans les méthaniseurs (comme les *Ruminiclostridium*, *Enterococcus*). Cela permet aussi d'observer l'évolution au cours du temps de ces communautés.

# Conclusion

Cette thèse a permis le développement d'un ensemble de méthodes pour avoir une meilleure compréhension du métabolisme à plusieurs échelles : les voies métaboliques, les génomes, les taxa et les communautés d'organismes. Les données mesurant le métabolisme provenant de différentes approches (génomique, métabolomique) et en grande quantité (génomes séquencés, données de métagénomiques) peuvent rendre la compréhension de celui-ci complexe. Pour cela, la thèse a visé à développer des approches d'aide à la décision pour l'analyse du métabolisme en proposant des candidats (voies métaboliques alternatives, métabolisme de taxon, espèces clés) qui pourraient être analysés plus précisément. Cela est possible en combinant plusieurs domaines. L'informatique en proposant des méthodes de passage à l'échelle au travers de programmation par contraintes, de parallélisation et d'ingénierie des connaissances. La bioinformatique en développant des méthodes de filtrage des données et de workflow. Cela a permis de proposer des candidats potentiels pour chaque échelle analysée: des voies métaboliques alternatives chez *Chondrus crispus*, une comparaison du métabolisme des groupes d'algues et l'identification d'espèces clés dans une communauté microbiennes d'un méthaniseur.

# Publications

Cette thèse est basée en partie sur des travaux déjà publiés. Ainsi le Chapitre 2 est associé à une publication réalisée en collaboration avec la Station Biologique De Roscoff (abrégée en SBR, LBI2M UMR8227) et publiée dans le journal *iScience* (Belcour et al. 2020b). Le Chapitre 5 a été créé à partir d'une publication du journal *eLife* (Belcour et al. 2020a). De plus des parties de la thèse correspondent à des articles en cours de soumission. Le Chapitre 3 est fait à partir d'un article écrit en collaboration avec une équipe de la SBR et en cours de soumission. Le texte associé est déposé sur un serveur de preprint [1]. La

1. https://www.biorxiv.org/content/10.1101/2022.06.14.496215v1

méthode présentée dans le Chapitre 4 est décrite dans un article et le texte est disponible sur un serveur de preprint [2]. Et les sous-sections 4.3 et 5.3 correspondent à un article en cours d'écriture en collaboration avec Patrick Dabert (UR OPAALE, INRAE).

De plus, durant ma thèse des articles dont je suis co-auteur ont été publiés. Parmi ces articles, certains ont un lien avec le sujet de la thèse comme un article publié dans *Antioxidants* (Nègre et al. 2019) écrit en collaboration avec une équipe de la SBR, un article publié dans *PeerJ* (Karimi et al. 2021) écrit en collaboration avec une équipe de la SBR et un dernier publié dans *Frontiers in Plant Science* (Girard et al. 2021). D'autres ne sont pas cités comme un article publié dans *Journal of Phycology* (Xing et al. 2021), un article publié dans *Genomics* (Daval et al. 2019) et un article publié dans *Microbial Biotechnology* (Daval et al. 2020).

## Logiciels

Tous les logiciels développés durant cette thèse sont accessibles en open-source. Les codes sont déposés sur le site GitHub dans différents répertoires accompagnés de readme et de documentation. Pour le chapitre 2, le code de la méthode développée (PathModel) est présent sur le github pathmodel/pathmodel. Pour le Chapitre 3, le code d'AuCoMe est accessible dans le répertoire github AuReMe/aucome. Le code de la méthode présentée dans le Chapitre 4 (EsMeCaTa) est accessible dans le répertoire AuReMe/esmecata. Pour le Chapitre 5, plusieurs outils ont été développés. Metage2Metabo se repose sur plusieurs autres outils qui ont été dévelopés comme le package mpwt ou optimisés comme MeneTools et MiSCoTo. Le code du workflow Metage2Metabo est accessible à AuReMe/metage2metabo.

2. https://www.biorxiv.org/content/10.1101/2022.03.16.484574v1

# INTRODUCTION

Metabolism, the set of biochemical reactions converting compounds for multiple goals, is a vital process for the cell. Increasing the knowledge of metabolism is very important as numerous applications (such as drug design, organism comparison, or disease understanding) can benefit from it. Several scientific domains study metabolism with various methods, from direct measures to predictive methods.

Among these domains, Bioinformatics is a scientific field aiming at explaining biological phenomena through computational approaches. It is at the interface between biology and mathematics, seeking to apply different methods and techniques (constraint programming, sequence comparison, knowledge engineering, etc.) to biological, chemical or physical science data. One of the applications of bioinformatics (concerning systems biology) is the development of models to predict the functioning of biological phenomena.

These predictions often rely on biological experiments. With the rise of the -omics methods, it is possible to study the global content of the cell. Numerous methods can be used to measure the metabolism activity by identifying genes (through genomics and transcriptomics), enzymes (with proteomics), metabolites (with metabolomics) and the reaction rate (with fluxomics). These methods create numerous complementary data allowing the study of the metabolism at different levels.

This thesis aims at developing a set of methods to better understand metabolism at multiple levels (metabolic pathways, genome, taxon and community). It is at the interface of sequence comparison, knowledge engineering and constraint programming. It shows the development of decision support approaches for metabolic analysis by proposing candidates (alternative metabolic pathways, taxon metabolism, symbiotic species) that could be analysed more precisely. Chapter 1 presents the modelling of metabolism and the level of metabolism studied in the following chapters.

A first level can be the synthesis or degradation of specific metabolites, essential for designing metabolic pathways. Metabolomics allows for the identification of the metabolites present in a sample. Knowing the metabolites present in an organism can give insight into how the metabolites are processed and the underlying reactions. Numerous methods have been developed to predict metabolic pathways from known metabolism. But there

are issues as several of these methods rely on metabolic modelling but do not consider metabolite identification from metabolomics. Furthermore, they often proposed the path of molecules predicted from mathematical modelling, reflecting the rules behind such approaches. But metabolic pathway inference could benefit from adding knowledge such as biological rules. In Chapter 2, we propose a method to predict alternative metabolic pathways by combining metabolomics data and metabolic modelling under the metabolic pathway drift assumption.

A second level is the metabolic network associated with all the enzymes contained in a genome. This network allows for studying all the metabolic capabilities of an organism. Comparing these sets of enzymes between different organisms could get insight into the organism's metabolic differences. But there are issues due to how the genomes are annotated, especially in public databases. Genomes are annotated by various tools leading to heterogeneity in genome annotation. The metabolic comparison of these heterogeneously annotated genomes will identify differences resulting not from biological signals but heterogeneous annotations. To avoid this issue, we have developed a method (presented in Chapter 3) to homogenise the annotation by using sequence comparison (with orthology propagation and structural verification). In this way, it is possible to compare the metabolism of these homogenised metabolic networks.

A third level is the metabolism associated with groups of related organisms (taxon). Numerous approaches do not give access to genome information; this is often the case in metagenomics which studies the genetic elements in environmental samples. For example, in metabarcoding, only the gene markers' sequences are retrieved. Aligning this sequence to specific databases allows for identifying the corresponding organism. But there is an uncertainty in the assignation of the gene marker, preventing the identification of the lowest taxonomic rank (such as the species). Methods exist to find the function associated with gene markers (especially the 16S ribosomal RNA gene). But there is a lack of a method to estimate the metabolic capabilities of the associated organisms from any gene markers. To solve this issue, we present in Chapter 4 a method to estimate the metabolic capabilities of a taxonomic affiliation.

The last level this thesis studies is the community level, where we try to understand the metabolic interactions between several organisms. Exploring these interactions is of great interest for environmental study as it allows for creating hypotheses on the symbiosis between the organisms. But there is a significant issue, the community identified in metagenomics often contains hundreds or thousands of organisms. There is a need for

tractability to handle these large-scale. Furthermore, there is multiple possible input, either genome (complete genome or Metagenome-Assembled genome) or metabolic models such as the one produced in the previous paragraph. To handle this, we have developed a workflow to reconstruct the metabolism from genomes and compute the individual and community production capabilities to infer the possible metabolic interactions for the production of metabolites of interest.

Chapter 1

# State of Art

Bioinformatics is a scientific field developing computational methods for manipulating biological data to solve biological problems. Furthermore, it can rely on knowledge that can be represented in multiple ways (such as databases). Bioinformatics is also highly connected with systems biology as it tries to combine different information to model complex biological systems. Among these systems, metabolism is essential as it is used to produce and store energy and handle metabolic wastes.

The metabolism consists of the set of biochemical reactions occurring in a cell to perform various tasks such as stocking and producing energy and modifying compounds necessary for the cell. A biochemical reaction transforms a set of reactants into a set of products. The biochemical reactions are divided into enzymatic and spontaneous reactions. First reactions are the enzymatic reactions which are catalysed by an enzyme. And the second type is the spontaneous reaction which occurs without an enzyme.

Multiple methods have been developed to study metabolism, especially by representing it as a graph (where nodes or edges represent metabolites and biochemical reactions). This representation can be used for different purposes. The first purpose is to describe and explore the knowledge of an organism's metabolism. Using methods from the mathematical field of graph theory, it is also possible to analyse these graphs. Multiple algorithms are available to perform these tasks. A third purpose is to simulate the dynamics of the metabolism with methods such as Ordinary Differential Equation (ODE) or Flux Balance Analysis (FBA). Furthermore, new research opportunities in system biology rise with the latest sequencing technologies allowing for sequencing a high number of genes and genomes. Developments in other fields, such as metabolomics (allowing to identify metabolites), also help to model the metabolism better.

This thesis combines logic programming and knowledge engineering to answer multiple problems associated with the modelling of metabolism. These contributions occur at different levels of metabolism. The lowest level studied is the metabolic pathway, where the thesis explores the possible alternative pathways from a drift perspective in Chapter 2.

Then, it compares the metabolism at the genome level for multiple organisms in Chapter 3. Chapter 4 predicts the metabolism associated with a group of akin organisms (taxon). And in the final Chapter 5 identifies key species among a community using metabolic complementarity.

## 1.1 Modelling the metabolism in systems biology

Metabolism is defined as "the chemical reactions in living organisms by which energy is provided for vital processes and activities, and new material is assimilated" (PubMed MeSH Term 68008660 on 19 July 2022). The biochemical reactions are either in the catabolism or anabolism (Judge et al. 2020). Catabolism is the degradation of complex molecules into smaller molecules. Anabolism is the biosynthesis of complex molecules (such as nucleic acids). Some of the chemical reactions are catalysed by enzymes, which are specific proteins having the ability to bind to substrates and decompose them into products. These reactions are often called enzymatic reactions. Other reactions can occur without enzymes, and they are called spontaneous reactions.

The metabolism is a vital system of the cell. It is featured among the three subsystems used in minimal cell experiments (with informational and compartment-forming subsystems). The metabolism seems so essential that it was hypothesised (for a time) that it could be the first system to emerge from abiogenesis (the process from which life emerged from non-living matter). But this claim is controversial as it was shown that the three subsystems could emerge at the same time (Patel et al. 2015). Studying metabolism allows us to understand better life, the processes occurring in a cell, and the exchanges within and between organisms. The application of the metabolism is detailed in subsection 1.1.5. Before this, we will look into more details on how we can study metabolism and especially how we can measure it.

Multiple fields arose to study metabolism in the era of omics approaches (Figure 1.1). Each of these approaches examines specific parts of the metabolism, making them complementary to the global understanding of metabolism.

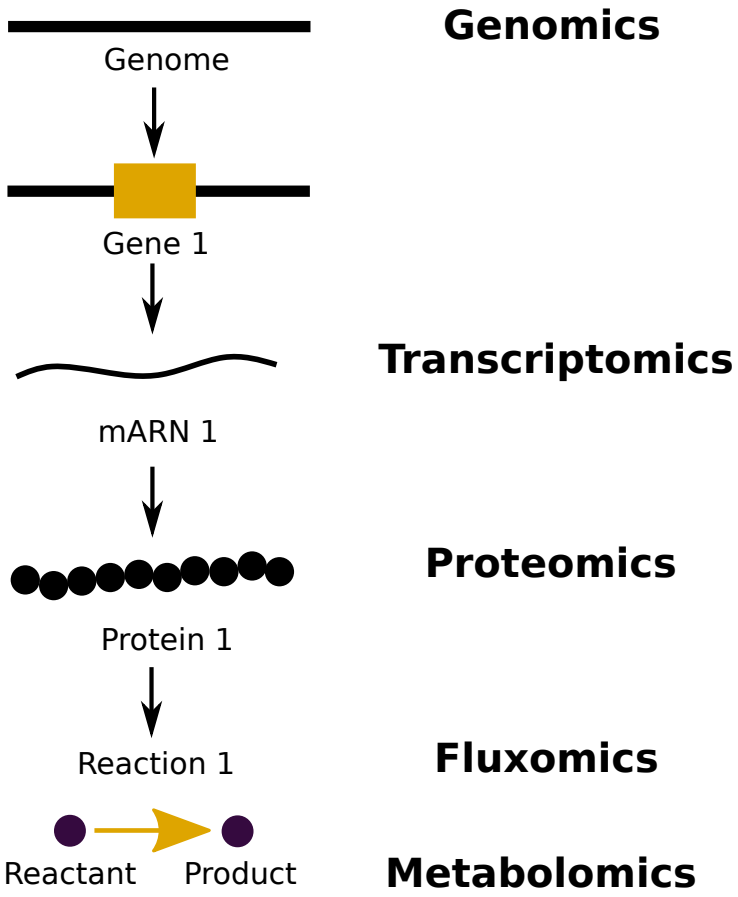



Figure 1.1 – Different -omics approaches.

— **Genomics** sequences the genomes of an organism (Giani et al. 2020) and can help to find gene sequence coding for specific enzymes. The results from these analyses make it possible to decipher which enzyme is present in a genome.
— **Transcriptomics** reveals the expression of genes, which can be an indicator of the production of enzymes (Z. Wang et al. 2009). But it has to be taken into account that gene expression can not be considered a one-to-one prediction of enzyme production as other processes are involved in the production of a protein. The genomics approaches indicate which enzyme is present, and transcriptomics identifies which enzyme-coding genes are active in a given condition.
— **Proteomics** identifies and quantifies the protein in a sample thus allowing to find enzyme (Aslam et al. 2017). Proteomics can identify the presence and quantity of

an enzyme and then indicates if an enzyme-coding gene is really translated into a protein.

— **Metabolomics** estimates the metabolites present in the sample. It can be untargeted to quantify the maximum of metabolites (thus allowing to detect new metabolites), semi-targeted to get metabolite assignment and quantification or targeted to have quantitative analysis of known metabolite (Liu et al. 2017). It is then possible to identify new metabolites.

— **Fluxomics** measures the metabolite rates of biochemical reactions in the cell (Winter et al. 2013).

These fields produce different types of data, which will be analysed differently and in complementarity. With genomics, it is possible to reconstruct the genome sequence of an organism and detect genes. Then by mapping data from transcriptomics it is possible to identify the expression of the genes. The protein from the translation can be identified through Proteomics. Among these proteins, enzymes catalyse reactions that convert metabolites into other metabolites. Metabolomics finds the metabolites present in an organism. Lastly, Fluxomics can compute the metabolite rates of the conversion between the metabolites performed by the enzyme.

These data acquisition can be used for different goals from a metabolic perspective (description, representation, exploration or prediction). First, the knowledge inferred from the genes identified with genomics and transcriptomics can help to describe the function of a group of genes. Secondly, the knowledge of all the genes of an organism can be used to describe or visualise all the possible functions present in the organism. It can also be explored to identify new functions. Lastly, it is possible to create dynamic modelling of the metabolism at different scales (compartment, cell or a community of organisms) to make predictions. Then depending on these goals, multiple ways to model the metabolism exists. In the following section, we will see the possible graph representations of the metabolism.

### 1.1.1 Graph representation

Metabolism can be modelled as a graph-based network. There is multiple ways to represent metabolism as a graph (Figure 1.2).

— A compound graph, where compounds are nodes, and there is an edge between two nodes if a biochemical reaction occurs between the compounds. Then the same biochemical reaction will occur multiple times in the network if there are multiple possible reactants or products for that reaction. These graphs are of interest to

link metabolic networks with metabolomics data. Indeed in this field, the focus is more on the metabolite than the biochemical reactions. Then these compound graphs can be associated with experimental networks constructed by looking at the metabolic relatedness found during metabolomics experiment (Amara et al. 2022). In the cited review, the compound graph is called a knowledge network. We will use this graph in Chapter 2 to associate metabolomics data with a metabolic network.

— A reaction graph contains nodes representing reactions, and an edge links them if they share metabolites. These graphs can be used to identify critical enzymes. Indeed as biochemical reactions are represented as nodes, the effect of an enzyme loss can be tested by deleting the reactions catalysed by the corresponding enzymes (Kim et al. 2019). In the previously cited article, they are called reaction-centric graphs.

— A bipartite graph contains reactions and metabolites. Nodes represent both reactions and metabolites. An edge is drawn between a metabolite node and a reaction node if the metabolites are the reactants of the reaction. And an edge is drawn between a reaction node and a metabolite node if the metabolites are the products of the reaction. This representation will be used in Chapter 5 to estimate the metabolic production of organisms.

— An hypergraph contains edges linking more than two nodes. The hypergraph can contain compounds as nodes (and reactions as hyperedges) or reactions as nodes (and then compounds are hyperedges). This method is used to compute network measures for the metabolism (Yeung et al. 2007; Pearcy et al. 2014; Pearcy et al. 2016; Klamt et al. 2009).

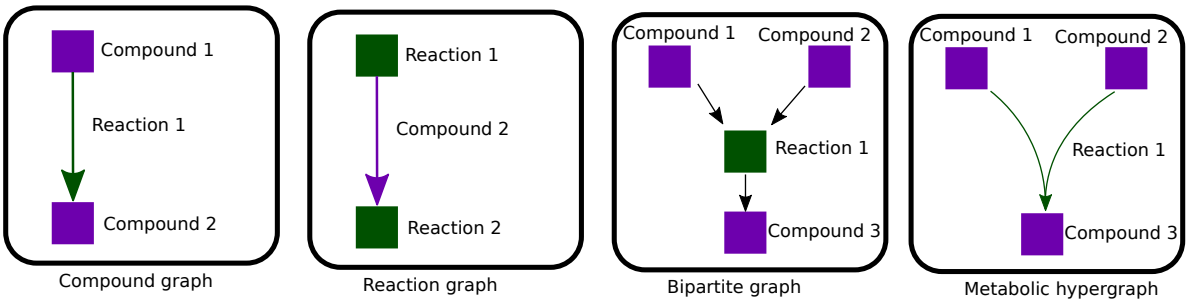



Figure 1.2 – Different representation of metabolism as a graph.

The graphs are directed as the reactions are directed (from the reactant to the product). But a reaction can be reversible, and then both directions are possible (from reactant to product and from product to reactant). The reversibility of all reactions is unknown, so some databases and models use non-directed representation or assume that all reactions are reversible.

These representations have different roles in the study of metabolism. All of these representations can be used to compute graph metrics (such as centrality or connectivity (Wagner et al. 2001)). The hypergraph with its hyperedges is interesting for that purpose. These representations can also be used to visualize the metabolism of an organism (Paley et al. 2021; Cottret et al. 2018; King et al. 2015; Hari et al. 2020). But among the graph representation of the metabolism, the bipartite graph is the most used, especially in dynamic modelling of the metabolism (presented in Subsection 1.1.4). Indeed, the bipartite graph allows to identify which reaction nodes can be activated by available metabolite nodes, thus allowing to compute the metabolic producibility.

This thesis will mainly focus on the bipartite graph to represent the metabolism (in Chapters 3, 4 and 5). A simplified version of this type of graph is used in Chapter 2. Also, the term metabolic network will be used to describe the graph-based representation of the metabolism.

### 1.1.2 Metabolic databases

Knowledge of the metabolism are stored inside databases. In these databases, reactions and metabolites are labelled by a set of identifiers. Most often, the metabolic databases are constructed from the literature. Multiple databases have been created such as BioCyc (Caspi et al. 2016), KEGG (Kanehisa et al. 2022), BiGG (King et al. 2016), ModelSeed (Seaver et al. 2021), MetaNetx (Moretti et al. 2021), Rhea (Bansal et al. 2022).

BioCyc (Caspi et al. 2016) is a collection of databases which contains 20,005 Pathway/Genome DataBases (PGDB) at the date of the 18 July 2022. A PGDB contains an organism's genes, proteins, and metabolic network information. Among these databases, there is MetaCyc (Caspi et al. 2020), a universal curated database containing compounds, enzymes, biochemical reactions across all domains of life. Reaction direction is indicated if there is information about it in the literature. Other databases in BioCyc contain metabolism associated with specific Archaea (470 databases), Bacteria (19,416 databases) or Eukaryota (37 databases). It is associated with Pathway Tools, a tool to reconstruct draft metabolic networks from annotated genomes (Karp et al. 2002a). This method al-

lows for the reconstruction of the PGDB of an organism using its genome. It has been widely used to create metabolic networks for multiple species with different levels of manual curation. It is divided into 3 tier databases, the first one being curated for at least one year, the second tier undergoing between 1-4 months of manual curation and the third tier having no manual curation.

KEGG (for Kyoto Encyclopedia of Genes and Genomes) is a knowledge base with multiples databases containing genomes, genes, orthologs, functional annotation, biochemical reactions and metabolic maps for 392 Archaea, 7,072 Bacteria and 770 Eukaryotes (at the 18 July 2022). It is associated with a suite of tools to visualise and map metabolic information for organisms. Despite showing reaction direction in their representation, KEGG does not give reaction direction in its file. The direction needs to be extracted from the map file.

Bigg is a database containing metabolic networks for 108 models (as the 18 July 2022) for 87 Bacteria, 1 Archaea and 20 Eukaryota. These models have been manually created mainly by the group of Dr B. Palsson at the University of California San Diego (Schellenberger et al. 2010; King et al. 2016). Being manually reconstructed, the models are of high quality. They are used as references to reconstruct other metabolic networks (Machado et al. 2018).

ModelSeed is a metabolic database built by integrating and merging reactions and compounds from MetaCyc, KEGG and internal models.

MetaNetx is a database trying to reconcile information from multiple other databases (such as KEGG, Bigg, MetaCyc). This is performed by an automated reconciliation procedure and manual curation.

Rhea is a knowledge base using the ChEBI (Chemical Entities of Biological Interest) database to specify the metabolites and is associated with the UniProt database.

These databases contain knowledge that can be queried to extract information for research purposes. But despite the effort made by MetaNetx, mapping the data between these multiple databases remains difficult. Thus combined analysis can be performed but is at a high cost. This mapping difficulty has an impact on the study of metabolism. A model created with a given metabolic database could not easily be used with models from other databases. This difficulty is especially an issue when comparing metabolic networks such as what is presented in Chapter 3 and is often corrected by working on the same metabolic databases or by performing mapping (by mixing automatic and manual procedures).

These multiple databases can also question how the metabolic community work and lack sufficient interaction to develop a general database. Indeed, now sequenced genome can be deposited on the International Nucleotide Sequence Database Collaboration (INSDC), handled by three partners (NCBI, ENA and DDBJ). A general database combining metabolic networks and storing data from different experiments (such as proteomes or metabolomics) could increase the prediction possibilities.

Combining information is growing as presented in the development of MetaNetx or Rhea. By using semantic web technologies, the latter gives more possibility to connect data from multiple databases (ChEBI or UniProt). This could lead to combining the results from numerous analyses (metabolomics, proteomics, transcriptomics) and their incorporation.

### 1.1.3 Inferring metabolism for organism

In the modelling of metabolism, these metabolic databases serve as a reference to estimate the metabolism of organisms. Different methods can perform this task. Most of them use as input the annotated genome of the organism to reconstruct the metabolic network. More details on the process of reconstructing an organism's metabolism will be presented in the subsection 1.3.3.

Combined with the sequencing methods, the reconstruction methods open the door to the functional analysis of less studied organisms, such as non-model organisms. Using genome sequences from the environment makes it possible to reconstruct draft metabolic networks of such organisms (often by relying on the homology of sequences, which is that two sequences are similar and can have similar functions).

This analysis can be extended to the environment with environmental research. It is possible to study the functions present in an environmental sample. This is, for example, the case of the functions identified in the microbiome of different ocean stations (Sunagawa et al. 2015). The metabolism can help to describe the functions present in unknown organisms. But the metabolic networks can go further than being descriptive as they can simulate the activity of the metabolism.

### 1.1.4 Dynamic modelling of metabolism

The metabolic network created for an organism can be used in multiple ways, such as describing the function of the genes and exploring the metabolic capabilities. Thanks to

numerous methods, it is also possible to model the activity of the organism's metabolism to make predictions.

**Kinetics Model.** A first one is the kinetic model. These modelling use thermodynamics, reaction stoichiometry and enzyme kinetics to elucidate the reaction fluxes (Srinivasan et al. 2015; Kumar et al. 2019). Thus these models need precise knowledge of the enzymatic rate equation. Enzymatic rate is associated with the formation of an enzyme-substrate complex, the decomposition of the substrate into a complex and the enzyme release. Kinetic models were developed with different assumptions. One of the most used is the Michaelis–Menten kinetics with assumptions such as the enzyme concentration being lesser than the substrate concentration.

Ordinary Differential Equation (ODE) is used to compute the parameter. For the Michaelis–Menten kinetics, there are two parameters to fit (Km and vmax) for each reaction. This lead to computational tractability in identifying and estimating the equation parameters. It is also complicated as there is a need for experimentally measured parameters. These tractability issues make these models unusable, with metabolic networks containing thousands of reactions. Other models have been used to handle such metabolic networks.

**Constraint-Based Modelling.** To scale with the thousands of reactions in an organism's metabolic network, methods such as Constraint-Based Modelling, also called stochiometric models, have been developed (Nielsen 2017). Compared to the kinetic models, it will use the steady-state assumption (metabolites are produced and consumed in mass balance, such as the metabolite concentration does not change over time) and constraints (Bordbar et al. 2014). Optimality will be searched in the solution space for the organism's growth (with the production of defined biomass).

Numerous methods have been developed, such as Flux Balance Analysis (FBA). These methods are applied in multiple situations, for example, to predict the growth and product secretion of a strain of *Escherichia coli* (Varma et al. 1994). It is possible to reconstruct a multi-organ metabolic model from the metabolic model reconstructed from a genome to study organism responses to perturbations (Gerlin et al. 2022).

But this modelling approach has several issues. It often relies on biomass production, but it can be challenging to estimate for wild organisms which are not cultivable (Frioux et al. 2020a). Furthermore, the gap-filling method used to complete metabolic networks

to ensure biomass production can add enzymes, whereas no gene encodes these enzymes.

**Topological Analysis.** A third method relies on the producibility of a metabolic network by using a boolean abstraction of the network. Compared to the previous methods, this one does not use stochiometry or the enzymatic rate of the reaction. This method relies on the connectivity of the metabolic network. From a set of initial metabolites called the seeds (which simulates the growth medium), an expansion algorithm will search for all the producible compounds. A compound is produced if it is the product of a reaction in which substrates are either produced or belong to the seeds. All the producible compounds and the seeds are called the scope (Ebenhöh et al. 2004). This method has been demonstrated to produce the same set of compounds as the one found with FBA method (Kruse et al. 2008). Thus it can be used to estimate the production of some metabolites qualitatively. But it can not quantify these productions.

**Conclusion.** As presented above, these methods have benefits and caveats. Kinetic models allow a precise metabolism analysis on a specific model with a low number of reactions. Constraint-Based Modelling allows quantification analysis for GSMN and community interactions (especially in pair comparison). The topological analysis permits qualitatively analysing metabolite production for large-scale metabolic networks. Following these indications, part two of this thesis will study a large-scale microbiota to estimate the producibility of a specific set of metabolites. It will use a topological analysis approach (more details in Chapter 5).

### 1.1.5 Current challenges in metabolism

The reconstructed metabolisms can be used in various applications such as (1) predicting the production of metabolites, (2) effect of drug target on pathogens, (3) discovering enzyme functions, (4) comparing the metabolism of organisms, (5) predicting interactions between organisms and (6) study disease (Gu et al. 2019). With each of these applications, different challenges arose, needing different levels of study of the metabolism (Figure 1.3).

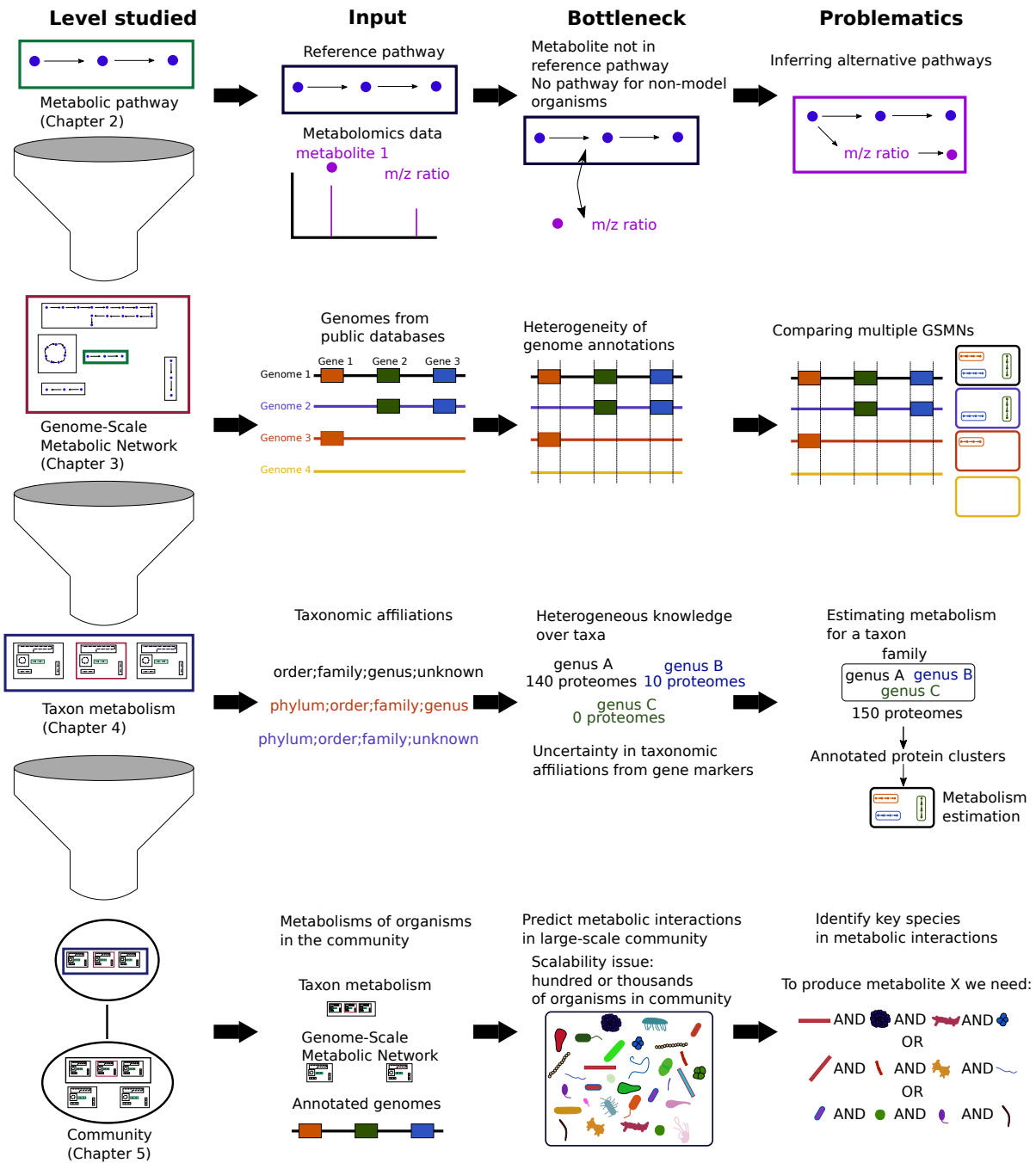



Figure 1.3 – The different levels of metabolism studied in this thesis.

There is a growing interest in predicting the production of metabolites (such as a drug) by modifying known organisms. Indeed metabolic engineering has been applied to make some organisms able to produce compounds. For example, yeast has been modified

to produce opioids (Galanie et al. 2015). This organism modification was performed by engineering a biosynthetic pathway thanks to enzyme discovery and pathway optimisation. In this goal, metabolic pathways can be tested *in silico* by using metabolic pathways prediction. Furthermore, with the rise of -omics approaches, there is a high number of genomes and metabolites identified for non-model organisms, which are organisms less studied than model organisms but with features or adaptations of interest (Russell et al. 2017). Identifying metabolic pathways in these organisms by taking into account their metabolites could be useful (Figure 1.3 level **Pathway**). This is developed in section 1.2 and Chapter 2.

The metabolic pathway level gives an insight into the set of reactions for metabolites production. With the metabolic networks reconstructed, it is also possible to compare the entire metabolism of multiple organisms (Vieira et al. 2011; Bauer et al. 2015; Prigent et al. 2017) to identify metabolic changes according to phylogeny and environmental conditions. The comparison of metabolisms brings several issues, especially when using genomes from public databases as input (Figure 1.3 level **Genome-Scale Metabolic Network**). This question is discussed in subsection 1.3 and in the Chapter 3.

As discussed in the previous paragraph, the study of organism metabolisms is eased when genomes are available. But without sequenced genomes, there is a need to estimate the metabolism. This need is often observed in metagenomics, the study of genetic elements in an environmental sample. Indeed several methods in metagenomics output only taxonomic information (which taxonomic groups are present in the sample). Then there is a need to estimate the metabolism of these taxonomic groups. This could help study the microbiota in different environments such as ocean (Sunagawa et al. 2015), soil (Wei et al. 2019) or humans (Cho et al. 2012). As most of these results from metagenomics indicate the composition of functions associated with a microbiota, the research on the metabolic potential of these microbiotas could help decipher their metabolic capabilities (Figure 1.3 level **Taxon metabolism**). This is studied in the section 1.4 and in Chapter 4.

Then having the metabolism of multiple organisms can help decipher the metabolic interactions between these organisms, giving insight into the environment and diseases. Indeed gut microbiota can impact the host metabolism (Martin et al. 2019), then investigating the microbiota metabolism can help elucidate these relations. Numerous tools have been developed to study their metabolisms (Kumar et al. 2019) but there is still issue especially with large-scale microbiota (Figure 1.3 level **Community**). This will be

looked in more details in section 1.5 and Chapter 5.

## 1.2 Pathway level: inferring alternative metabolic pathways for non-model organisms

In Subsection 1.1.5, we saw that metabolism representation could be used to identify drug targets or predict metabolite production. Multiple approaches allow for making those predictions, and one of these methods uses the metabolic pathway level. These methods often rely on metabolic databases to make new inferences or compare their predictions.

### 1.2.1 Definition: metabolic pathways

**Metabolic pathways.** Multiple definitions have been proposed for a metabolic pathway. A simple definition can be described as a series of biochemical reactions leading to the production or degradation of metabolites of interest. But this definition excludes some known metabolic pathways (Figure 1.4) such as branched pathways (a pathway in which multiple branches of reactions are split from an intermediate metabolite) or cyclic pathways (where each metabolite in the pathway is a substrate for a reaction of the pathway). Another possible definition of a metabolic pathway is a subgraph inside the metabolic graph (Faust et al. 2011).

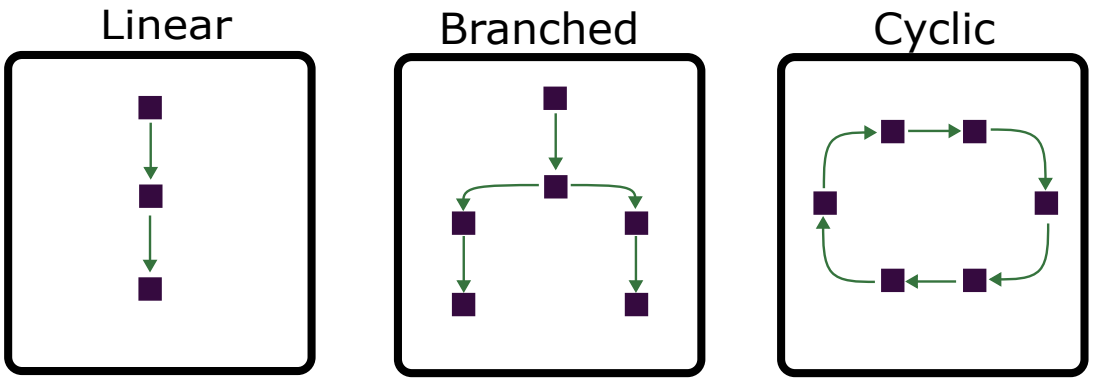



Figure 1.4 – Different metabolic pathway structures.

Some metabolic databases conceptualized literature knowledge and biologically inspired rules into metabolic pathways (Caspi et al. 2020; Kanehisa et al. 2022). A second

definition uses the properties of the metabolic network graph to identify metabolic pathways (Papin et al. 2003).

**Reference metabolic pathways.** The metabolic pathways are differentially conceptualized depending on the metabolic databases. Indeed, a comparison between KEGG and MetaCyc showed a vast difference in how the metabolic pathways are conceptualized (Altman et al. 2013; M. L. Green et al. 2006). MetaCyc and KEGG rely on their set of *reference pathways* and apply their own definitions to infer metabolic pathways for an organism.

— **MetaCyc** metabolic base pathways are designed from a set of rules: (1) the series of reactions in the pathway participates in a common biological process, (2) pathway input and output correspond to high-connectivity metabolites, (3) input and output of pathways are stable metabolites, (4) enzymes in the pathway have a common regulation, (5) pathways are evolutionary conserved thus often being specific to a group of organisms (M. L. Green et al. 2006). MetaCyc base pathways are also regrouped in super-pathways.

— **KEGG** reference map incorporates reactions from multiple organisms and can contain multiple biological processes (Altman et al. 2013; M. L. Green et al. 2006). The KEGG maps are also divided into KEGG module (Kanehisa et al. 2008), which are consecutive reactions. These consecutive reactions were obtained with genome comparison. The KEGG modules are closer to the MetaCyc base pathway (Altman et al. 2013).

— **ModelSeed** Subsystems contain collections of functionally related proteins that are created by scientific expert (Overbeek et al. 2014). These subsystems can rely on other databases' metabolic pathways, such as the pathways from MetaCyc.

These differences in definition lead to differences in the metabolic pathways and the database content, thus impacting the possible predictions. MetaCyc base pathways are more numerous than KEGG maps but have fewer reactions (Altman et al. 2013). For example, the representation of pathways associated with methane metabolism in KEGG and MetaCyc pathways are shown in Figure 1.5.

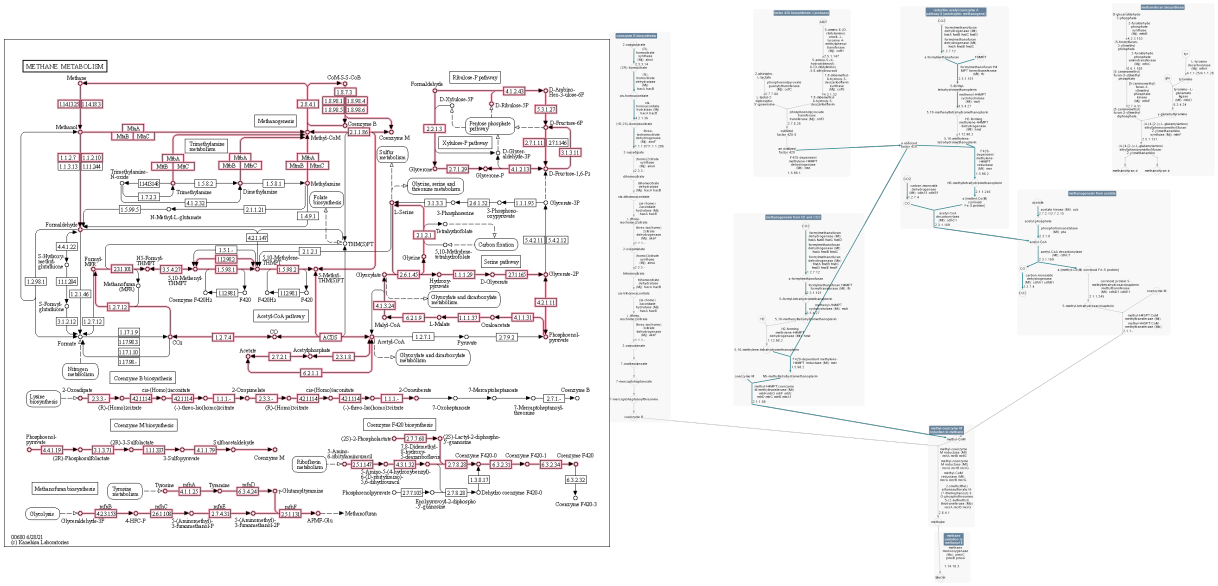


Figure 1.5 – Metabolic pathways for methane metabolism. KEGG methane metabolism map (left), KEGG modules are shown in red. MetaCyc group of pathways manually selected as associated with methane metabolism (right). MetaCyc pathway names are shown in blue.

### 1.2.2 Available data: reference pathways, metabolomics data

Multiple methods rely on the reference pathways from the metabolic databases (such as KEGG map or MetaCyc base pathways) to infer metabolic pathways. The knowledge in these databases are often used either as an input for prediction or as a gold standard for comparison.

These reference pathways can be inferred for an organism from its genome. But by using reference pathways from model organisms, it is possible that these pathways are not a good representation of the metabolic pathway present in the studied organism. Thus there is a need to find more accurate metabolic pathways for these organisms, possibly with new predictions.

Another input from the metabolic pathways inference is the metabolites found in an organism. This can be achieved thanks to metabolomics, allowing to identify (1) known metabolites by using analytical standards, which are compounds added in the analysis to identify their corresponding peak and (2) unknown metabolites associated with a Mass-To-Charge ratio. With this direct measure of the metabolite presence, it is possible to identify which metabolic pathways can occur in an organism.

### 1.2.3 Inferring metabolic pathways

There are multiple ways to perform metabolic pathway prediction. Some methods try to identify new metabolic pathways among a known metabolic network, thus creating new series of linked biochemical reactions. Other approaches try to predict the presence of a biochemical structure in a metabolic pathway, thus allowing to link unidentified molecules to a known metabolic pathway.

They can rely on graph-based methods, stoichiometry-based (constraint-based modelling) and retrosynthesis-based methods (L. Wang et al. 2017). Another way to identify metabolic pathways is emerging, relying on machine learning and especially deep learning.

The first type of methods is the subgraph extraction from a graph network. Numerous methods rely on graph traversal to identify metabolic pathways (Faust et al. 2011). This is performed using topological algorithms such as Depth-first search or Breadth-first search.

The second set of methods has applied Constraint-Based Modelling to search for optimal pathways that produce specific compounds such as OptStrain (Pharkya et al. 2004), RetSynth (Whitmore et al. 2019). These methods try to infer possible reactions to produce targeted compounds.

Retrosynthesis-based methods use target metabolites and try to infer the metabolites and the reactions that could lead to the production of these target metabolites, such as RetroPath2 (Delépine et al. 2018). These tools propose novel reactions.

Machine learning and deep learning have arisen in the prediction of metabolic pathways. For example, a graph-convolutional network (GCN) and random forest (RF) classifiers were used to classify metabolites from a SMILES representation to identify KEGG pathway class associated with the metabolite (Baranwal et al. 2020). These methods allow for predicting the presence of a metabolite in a metabolic pathway. Other methods rely on linking GSMN and structural biology (Calhoun et al. 2018) to infer pathways with candidate enzymes.

Some methods try to infer the potential transformation of a metabolite into other metabolites by creating metabolic maps. For example, a workflow has been developed to predict the metabolic map of the degradation of xenobiotic (Conan et al. 2021) by searching for the potential reaction on a metabolite structure.

Other methods can be used to infer a series of reactions between observed metabolites from metabolomics. A majority of the masses identified in metabolomics can not be mapped to known metabolites due to a lack of knowledge. To resolve this, an *ab initio* method has been developed to reconstruct a metabolic network using the observed masses

(Breitling et al. 2006). Potential reactions between the compounds were predicted according to frequently observed differences in masses between pairs of compounds. But these metabolic networks are isolated from the ones created by metabolic databases.

### 1.2.4 Limit: inferring metabolic pathways for non-model organisms

Known metabolic pathways from databases are created from model organisms, but the metabolic pathways can undergo variations due to changes occurring over time. Due to this, it can be challenging to study non-model organisms as their metabolic pathways can vary from model organisms present in metabolic databases. The inference of metabolic pathways for non-model organisms is complex as the non-model organisms are associated with very little data. Indeed, the data-driven methods will be limited when applied on non-model organisms.

A possibility to solve this issue is by adding knowledge and associating it with the known data. Most of the presented methods relied on the homology, metabolic pathways are conserved other time. Multiple hypotheses have been proposed to explain the evolution of pathways overtime (Scossa et al. 2020). A first one is the retrograde hypothesis, where a pathway is the result of the duplication of the genes catalysing its final step. This could be the result of the depletion of compounds in the primordial soup. The Granick's hypothesis stipulates that a metabolic pathway originates from its first and simpler metabolites and then expands to more complex molecules. A third hypothesis, the patchwork hypothesis proposes that ancestral genes were associated with promiscuous enzymes. Then with duplication and divergence, the activity performed by the promiscuous enzyme could be performed by different paralogs. The shell hypothesis proposes that metabolic pathways emerged from a consecutive addition of pathways from core central pathways to more secondary pathway. So biological knowledge and hypothesis could be used for metabolic pathways inference. Taking this information into account would require using Knowledge Representation and Reasoning (Levesque 1986) on the biological knowledge in hypothesis-driven methods.

We could combine the measured data (reference metabolic pathways and metabolomics data) and the hypothesis-driven approach to elucidate these limits. This could be performed by applying reasoning based on evolutionary hypotheses on data. This is explored in Chapter 2, which tries to infer metabolic pathways for a non-model organism.

### 1.2.5 Contribution: inferring alternative pathways from metabolic pathway drift

In Chapter 2 we propose to combine metabolic network and metabolomics data to infer alternative metabolic pathways resulting from metabolic pathway drift. For this, we offer a formalism of the metabolic pathway drift in the PathModel method. The idea of the metabolic pathway drift is that even if pathways are evolutionarily conserved, they can undergo changes that do not modify the pathway input and output. To this end, we used as a reference MetaCyc metabolic pathway as the input and output of these pathways are stable metabolites (both in terms of connectivity and decay). Then this formalism is implemented as a constraint problem and is used to explore the possible alternative pathways for two metabolic pathways in an alga.

## 1.3 GSMN level: Annotation heterogeneity in public databases impacts GSMNs reconstruction and comparison

As we have seen in the previous section 1.2, the metabolism can be analysed under the scale of metabolic pathways. Metabolic pathways are a subnetwork inside the metabolic network. The method presented in the previous section 1.2 (PathModel) makes it possible to infer alternative pathways, but it needs a reference pathway. Reference pathways can be inferred from metabolic databases during the metabolic network reconstruction. From a genome, multiple methods exist to reconstruct metabolic networks, called Genome-Scale Metabolic Network (GSMN), giving insight into the metabolism of an organism. And among them, several approaches also infer the presence of reference metabolic pathways during this step. With the rise of sequencing technologies, it is easier to sequence an organism's genome. With these sequences, it is possible to assess the metabolic potential of an organism according to its genome annotation. And as we have more and more genomes, comparing these organisms' metabolism could help us understand their differences and equivalences.

### 1.3.1 Definition: genome annotations

**Genome-Scale Metabolic Networks.** Representation of the metabolism can be reconstructed from the genome annotations. This genome-centric model of metabolism creates a **Genome-Scale Metabolic Network** (GSMN). In the first reconstruction of a GSMN (*Haemophilus influenzae*), the genome annotation was used to associate genes to enzymes catalysing biochemical reactions (Edwards et al. 1999). In further reconstruction, genes are connected to proteins associated with biochemical reactions. The combination of these three entities have been called a **Gene-Protein-Reaction** (GPR) association (Reed et al. 2003).

**Gene-Protein-Reaction association.** As explained in the previous paragraph, the association between gene, protein and reaction is defined as a GPR. Multiple types of GPRs are possible according to the number of genes, proteins and reactions involved (Machado et al. 2016). The different types are presented in Figure 1.6. The first type of GPR is the single enzyme, a single gene encodes for a protein, which performs a single reaction. A second type is the isozymes, where two genes encodes for two different proteins that catalyse the same reaction. A third type is associated with promiscuous enzyme, a single gene encodes for a protein that catalyses multiple reactions. Finally, an enzyme complex which is a set of genes encoding for multiple proteins that assemble into a complex which catalyses a reaction.

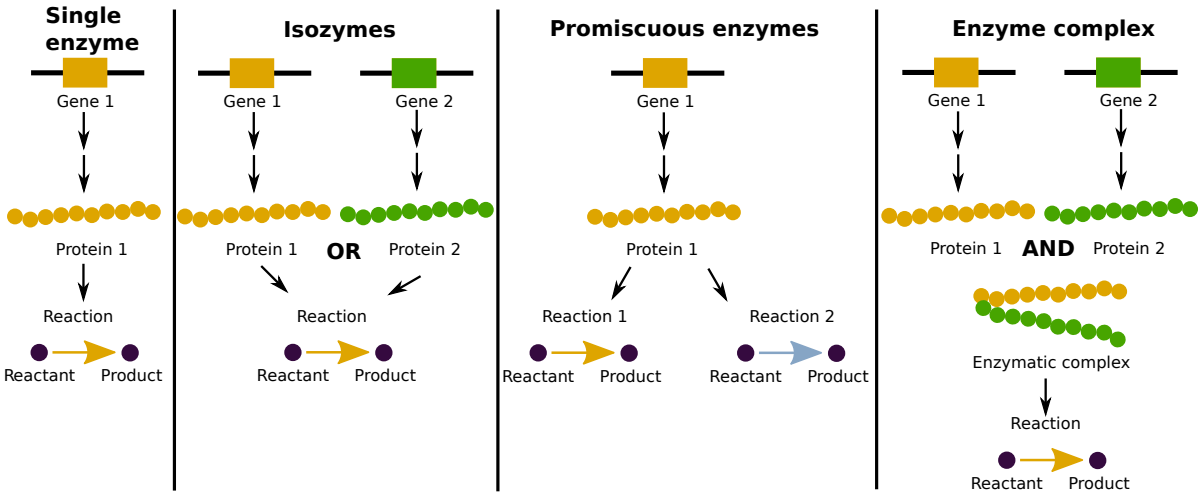



Figure 1.6 – Different types of Gene-Protein-Reaction associations.

**Genome annotation.** To link a genome to metabolism, a standard method is to find the enzyme in the genome. An enzyme is translated from a messenger RNA transcribed from genes located on the genome. With the rise of genome sequencing, it has been possible to obtain complete genome sequences. A **structural annotation** can be performed from these sequences. The structural annotation will search for specific DNA regions, such as the genes' introns and exons (Ejigu et al. 2020). It allows for the identification of gene sequences. Then a second step named **functional annotation** can associate biological information with the genes.

**Functional annotation.** To annotate the function of a gene, the associated biological information is often a set of terms from a function database. Multiple function databases exist such as the Gene Ontology, the Enzyme Commission number and the protein domains (Pfam (Mistry et al. 2021), InterPro (Blum et al. 2021)). As they will often be used in the thesis, Gene Ontology and Enzyme commission are described in more detail in the following paragraphs.

— **Gene Ontology.** One of the most known annotations is the Gene Ontology (GO) an ontology describing the function associated with a gene and a gene product (Ashburner et al. 2000; Gene Ontology Consortium 2021). This ontology is described with a structure describing the relationships between Gene Ontology Terms. And a specific Gene Ontology Term is associated with a specific gene, describing the function performed by the gene. A Gene Ontology Term is identified by a prefix 'GO:' associated with seven numbers (such as GO:0016860 for 'intramolecular oxidoreductase activity').

— **Enzyme Commission.** Another functional annotation is the Enzyme Commission number (EC), defining the enzymatic activity associated with the product of a gene. An EC number is composed of fourth digits (such as 1.1.1.1 for alcohol dehydrogenase) describing the corresponding enzymatic activity (McDonald et al. 2014). An enzymatic activity can catalyse multiple reactions. The first digit indicates the general class of chemical transformation performed (such as Oxidoreductases for 1.x.x.x meaning a transfer of proton between a reductant and an oxidant). The second and third digits (subclass and sub-subclass) describe the group, the reaction centre, or the chemical bond modified in the reaction. The third digit (sub-subclass) details the reaction, and a fourth digit is a serial number identifying the substrate specificity. Numerous databases stored the nomenclature

of the Enzyme Commission numbers such as ExploEnz (McDonald et al. 2009), ExPASy ENZYME database (Bairoch 2000). Metabolic databases link the Enzyme Commission number and their reactions.

### 1.3.2 Available data: annotated genome

The input data to reconstruct a GSMN is the sequenced genome of an organism (Figure 1.7). It consists of the nucleic acid sequences of the genomes. Often the sequences have been assembled into scaffolds (and chromosomes, if possible). Then using **structural annotation**, gene positions are predicted on the sequences to find the gene structure and predict the protein sequence. This prediction creates a set of protein sequences that can be used for the **functional annotation**. The functions associated with the genes are then used to estimate the enzymatic functions present in the organism.

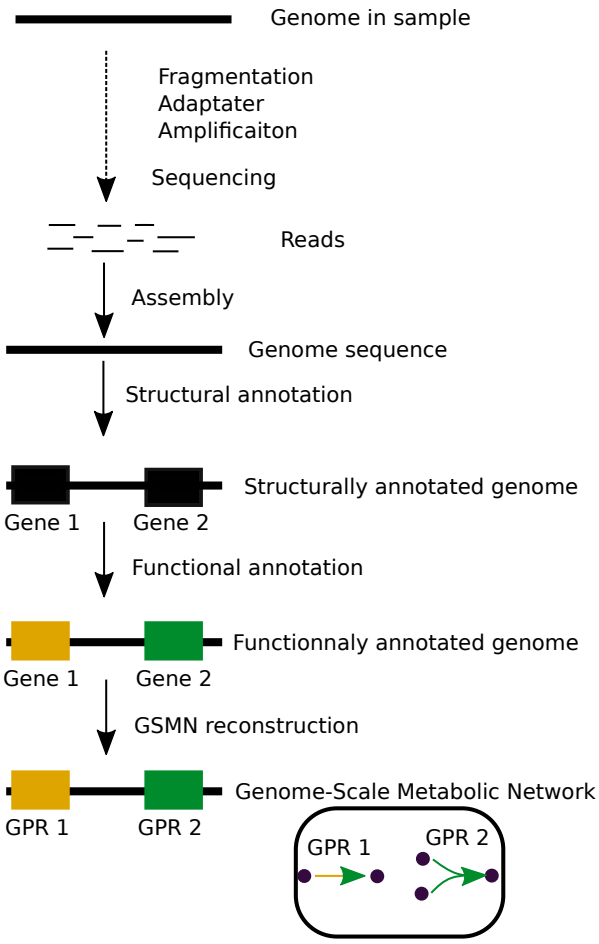



Figure 1.7 – Simplified genome assembly and reconstruction step.

### 1.3.3 GSMN reconstruction

A protocol has been defined to reconstruct a GSMN from a genome (Thiele et al. 2010a) focusing on creating functional GSMN, meaning that they can produce biomass in FBA experiment. First, a draft network is reconstructed from the genome annotation. Then the network is refined through manual curation (verification of the automatic prediction, adding multiple reactions (such as biomass, transport), compartment location, etc). The curated network is then converted into a computable format. The model is tested to ensure its growth capability; if not, missing reactions are added through gap-filling.

Multiple methods have been developed to perform this first step automatically. Two main approaches have emerged, a first called bottom-up, where metabolic networks are reconstructed from a genome, and a second called top-down, where the metabolic networks

are reconstructed from a universal metabolic network (Figure 1.8).

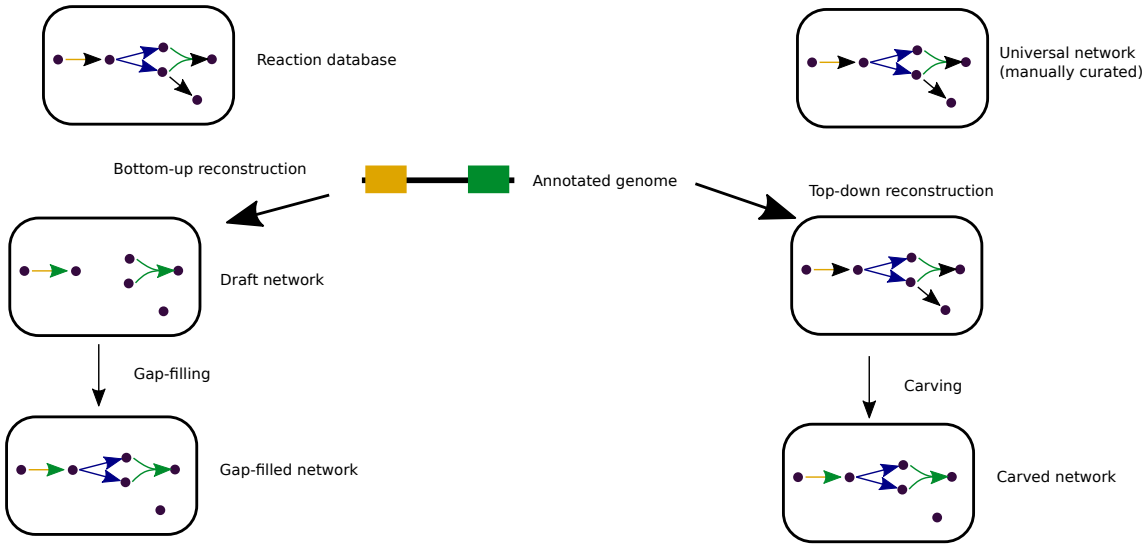



Figure 1.8 – The two approaches in GSMN reconstruction, bottom-up reconstruction or top-down reconstruction.

The bottom-up methods use the genome and the annotation to reconstruct a metabolic network. These methods are, for examples, Pathway Tools (Karp et al. 2002a; Karp et al. 2019; Karp et al. 2021), RAVEN (Agren et al. 2013; H. Wang et al. 2018), merlin (Dias et al. 2010; Dias et al. 2015; Capela et al. 2022), modelSEED (Henry et al. 2010; Seaver et al. 2021) in the Kbase framework (Arkin et al. 2018). These methods are impacted by the available knowledge associated with these annotations. For example, the Gene Ontology terms are incomplete and biased (Gaudet et al. 2017). This can impact the GSMN reconstruction by avoiding the prediction of reactions. Indeed, some reactions can not be retrieved from the annotations. But there is some possible gap in these metabolic networks due to missing reactions. These methods rely on applying gap-filling to solve these issues. Gap-filling is defined as the completion of missing reactions limiting the biomass production after the reconstruction, such as fastGapFill (Thiele et al. 2014) or meneco (Prigent et al. 2017).

The top-down method uses genome annotation and taxonomic information to remove reactions from a universal metabolic network (this step is called carving). The gap-filling part is also included in these tools. This has been designed especially for Bacteria and Archaea such as CarveMe (Machado et al. 2018) or gapseq (Zimmermann et al. 2021).

Toolboxes were also designed to ease this step, such as AuReMe (Aite et al. 2018).

Some of these toolboxes try to help the curation by providing a set of possible candidate reactions using a reference metabolic network template.

It has been shown that these tools do not outperform one another; instead, they have different strengths or weaknesses. CarveMe, gapseq or ModelSeed reconstruct functional metabolic networks ready to be used in Constraint-Based modelling. Still, there is a cost as gap-filling methods are used to ensure biomass production, possibly introducing false positive reactions within the organism's metabolism. Pathway Tools uses genome annotation and performs a gap-filling process to fill incomplete metabolic pathways. Thus, the predicted GSMNs are not functional and can need manual curation after their prediction. For more details, the benchmark in Mendoza et al. 2019 gives more information.

### 1.3.4 Limit: annotation heterogeneity biases GSMN comparison

Comparison of the metabolism of multiple organisms can be helpful as it can give insight into the phenotypes and lifestyles of organisms (Gu et al. 2019). This can be achieved by reconstructing GSMNs from genomes. But using different tools to annotate the genomes or to reconstruct GSMN will lead to differences in the metabolic networks produced. Various annotation tools for the genomes will induce differences between the GSMN reconstructed (Karimi et al. 2021).

These differences are one of the issues when comparing GSMNs, as the comparison will identify differences due to annotation and not real biological differences. During a collaboration with the Station Biologique de Roscoff, the comparison between the GSMN of *Saccharina japonica* (Nègre et al. 2019) and the GSMN of *Ectocarpus siliculosus* (from Prigent et al. 2014) highlighted differences only due to annotation issues (Nègre et al. 2019).

These differences are increased with genomes from public databases as they can display heterogeneity in their annotations (Lobb et al. 2020) because they were annotated by different teams and with various methods. This heterogeneity is often solved by re-annotating the genomes with the same tools. But this method can lead to the loss of annotations from the genomes.

### 1.3.5 Contribution: homogenisation of metabolic networks from public genomes

To solve these limits, we propose a new workflow called AuCoMe that propagates annotations from heterogeneously annotated genomes, thus creating homogeneous annotations and GSMN without information loss. This method allows for comparing the metabolic networks with fewer biases in their annotations. It has been tested on three groups (one bacterial, one fungal and one algal) to test the method and search for metabolic differences in algal groups. These issues and the proposed solution are presented in Chapter 3.

## 1.4 Taxon metabolism level: estimate the metabolism of a taxonomic affiliation

The previous section presents a method to compare the metabolism of organisms by reconstructing GSMN from their genomes. But how can we define an organism's metabolism when no genomes are available? A possible solution could be to look at the metabolism of closely related organisms. The same issue arises when the only information about a specific organism is that it belongs to a given group of akin organisms. This issue is often encountered in **metagenomics** (analysis of the genetic elements present in an environmental sample). These issues put us at a new level compared to the metabolism of a single organism. How can we estimate the metabolism of a taxonomic group, and what can it represent?

### 1.4.1 Definition: taxonomics, cladistics and metagenomics

We first need to define how organisms are classified to answer these questions. The classification of organisms is called **taxonomy**. It is the first attempt to name and group organisms according to their similarity.

This first approach comes from the **Linnean taxonomy** developed by Car Linnaeus. It regroups organisms into **taxon** (a set of organisms sharing similarities and presenting differences with other organisms). This approach relies on comparing morphological characters. Organisms are classified into embedded groups: species, genus, families, order, class, phylum and kingdom. These terms are still used for convenience and are largely

extended in modern taxonomies.

A second approach is to group organisms according to their common ancestors by regrouping them into **clade** (a set of organisms sharing a common ancestor). This approach relies on phylogenetic analysis, especially by comparing the genome sequences of the organisms. In this manuscript, the three terms **taxon**, **clade** or **taxonomic group** will be used interchangeably to describe group of related organisms.

With the sequencing of numerous genomes, there was an increasing need to classify these genomes. Taxonomic databases have been developed such as the NCBI (National Center for Biotechnology Information) Taxonomy database (Federhen 2012; Schoch et al. 2020). The structure of the NCBI Taxonomy database tries to follow a phylogenetic taxonomy by classifying organisms according to the evolutionary history of life. But to achieve this, it relies on expert knowledge from the literature and not on the phylogenetic comparison of sequences of organisms. Other taxonomic databases, such as GTDB (Genome Taxonomy Database) relies on phylogenetic comparison of marker proteins for Bacteria and Archaea (Parks et al. 2018; Parks et al. 2020; Rinke et al. 2021; Parks et al. 2022).

The taxonomies regroup organisms into clades from a lower taxonomic rank, such as the species, to a higher taxonomic rank, such as the phylum.

In metagenomics, it is possible to extract the sequence (either genome or gene marker sequence) of the organism in a sample. To identify these organisms, these sequences are compared to known sequences in databases such as SILVA (Quast et al. 2013), RDP (Cole et al. 2014) for ribosomal RNA (rRNA) sequences. This method called **taxonomic assignment** allows to classify sequences into taxonomic group (for example the genus *Escherichia*). Furthermore, it permits obtaining the classification of a sequence from the higher taxonomic rank to the lower taxonomic rank, which will be called **taxonomic affiliation** in this thesis. For example, it is 'cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae; Escherichia' for the genus *Escherichia*. Following a tree representation, the **taxonomic rank** is the depth of a taxon in the taxonomy, with species and genus corresponding to the lowest ranks and kingdom and phylum to the highest. In the example, the higher taxonomic rank is 'cellular organisms', and the lower taxonomic rank is 'Escherichia'. There is a different **taxonomic diversity** according to the taxonomic rank. Higher taxonomic ranks correspond to older common ancestors and more diverged and diverse organisms. Taxonomic affiliation may not have identified lowest ranks (such as species or genus) for wild organisms highly

diverging from any known species (this is referenced as **uncertainty**). Instead, it could only have identified high taxonomic ranks (such as the order). This limitation reflects uncertainty in the assignation of a gene marker.

### 1.4.2 Available data: metagenomics and taxonomic affiliation

Multiple methods have been developed to study the organisms present in environmental samples. The genetic analysis of such samples is performed by analysing the environmental DNA (eDNA). This DNA from an environmental sample (containing cell DNA and exterior DNA) is sequenced in a process called metabarcoding (Ruppert et al. 2019). The idea is to use a primer (barcode or gene marker) to identify the taxa present in the environmental sample. The primer used will differ according to the coverage or the taxa studied. For example, the primer associated with the gene Cytochrome oxidase-I (*COI*) is most used for metazoans, whereas 16S ribosomal RNA is used for Bacteria and Achaea (Ruppert et al. 2019). These methods are cost-effective and often used.

Other methods are used, such as shotgun metagenomics. This method retrieves the sequences in the environment without any target (Quince et al. 2017). This can be used to obtain the organism's genomes in the sample with methods such as Deep whole-metagenome shotgun. By sequencing at a shallower depth (Shallow whole-metagenome shotgun), it is impossible to obtain genomes, but it is possible to identify taxa in the sample with a less expensive method (Hillmann et al. 2018).

Metabarcoding and Shallow whole-metagenome shotgun produce different data but are often used to produce similar results, the taxonomic affiliation of the organisms in the sample. These are also often referred to as OTUs (operational taxonomic units), a cluster of gene markers with high identity identified with a representative sequence. Thus a possible input for studying the results of these methods is to work on these taxonomic affiliations.

### 1.4.3 Functional profile and metabolism estimation

From the metabarcoding data, numerous methods have been developed to study the function present in an environmental sample. Methods have been developed to produce functional profiles to give insight into the metabolic pathways that can be produced from the sequence of a specific gene marker (Bowman et al. 2015; Douglas et al. 2020; Wemheuer et al. 2020). They rely on databases, either their own or reference databases

(such as SILVA (Quast et al. 2013) or RDP (Cole et al. 2014)) to align the gene marker sequence to known sequences. These methods provide information on what function is present in an environmental sample. Still, they cannot produce metabolic networks that can be used for dynamic modelling of the community's metabolism.

Then different methods have been developed to create metabolic networks for metabarcoding results. For example, such methods searches for the closest genome to infer GSMN (Mendes-Soares et al. 2016) or use already inferred GSMN (Patumcharoenpol et al. 2021).

### 1.4.4 Limit: gene marker specificity and metabolism estimation

Functional profile prediction and metabolism inference methods have some caveats.

First, most rely only on metabarcoding data as inputs and sometimes are only associated with one specific gene marker (such as the 16S rRNA gene). Thus, if an analysis is performed on another gene marker or with another technology, they can be more complex to use. For example, the gene *ITS* for fungi (Schoch et al. 2012) or other gene markers for bacteria such as *rpob* gene (Ogier et al. 2019).

A second issue is that the metabolism estimation associated with a single organism can be biased due to the lack of knowledge of a taxonomic group. This bias is especially prevalent if the chosen organism is highly diverging from the wild organism associated with the gene marker sequence. Then a better estimation could be through comparative genomics of the closest known organisms (such as the closest taxonomic group).

Furthermore, from these comparisons, it could be of interest to produce several estimations of the various metabolisms within a taxonomic group according to the available knowledge. Among these estimations, the first could be the most conserved functions in the taxon, a method used by numerous functional profile prediction methods. But it could also be of interest to estimate all the possible functions the taxonomic group can achieve. In this way, we could have an idea of all the functions present in the taxon, which the wild organism associated with the taxon could perform.

### 1.4.5 Contribution: metabolism estimation through shared proteins

I have developed a method to estimate the metabolic capacities from a taxonomic affiliation to resolve this issue. Using taxonomic affiliations makes it possible to use input results from different sequencing methods. The method predicts a set of protein clus-

ters and their associated functions. The protein clusters can be filtered according to the conservation of the proteins from a given cluster. This is performed by looking at the representation of the member of the protein clusters among the known proteomes of the taxonomic group. This threshold can create different estimations of the metabolism and different view of what an unknown species from a taxonomic group can perform. This method is presented in Chapter 4.

## 1.5 Community level: Investigating metabolic interactions in community

From the two previous sections 1.3 and 1.4, there is a possibility to reconstruct metabolic networks from sequenced genomes or from data of metagenomics. So with these methods, it is possible to study the metabolism of a group of organisms living in an environment. This section explores the estimations of the metabolic interactions between community members. The metabolic networks of the community members can be created from genomes or taxonomic affiliations. As hundreds of organisms can be found in an environmental sample, there is a need for a large-scale method to identify key species involved in metabolic interactions.

### 1.5.1 Definition: microbiota and interactions

Metagenomics provides data to study a **microbial community**, which can be defined as a group of microorganisms living in the same space. The term **microbiota** designs the range of microorganisms in this community. And the set of genomes associated with these microbiotas is called the **microbiome**. Among these organisms, numerous possible interactions can be analysed by looking at the metabolites exchanged. It is possible to rely on known basic ecological interactions (García-Jiménez et al. 2021) to identify these interactions.

— **Neutralism** is a neutral interaction between organisms where no negative or positive outcomes happen between the two organisms.

— In **Commensalism** interaction, one organism benefits from the interaction, whereas the second organism has no positive or negative impact from the interactions.

— For **Amensalism**, one species has a negative impact on the other species in the

interaction. Still, the second organism has no negative or positive impact on the first organism.

— In **Mutualism**, there is a positive impact for the two organisms in interaction.

— In **Predation**, one organism benefits from the interaction, whereas the other has a negative outcome. Due to the analogy of the impact of predation and parasitism, it has been proposed to unify these two notions (Raffel et al. 2008).

— For the **Competition** both organisms in interaction suffer from a negative outcome. Among competition, it is possible to separate two types of competitions: direct and indirect (Birch 1957). In a direct competition for a resource (**interference competition**, the two organisms compete directly for a resource leading to aggression displays between them. An indirect competition (called **exploitation competition**) impacts the organisms as they use the same resource leading to a depletion of the resource without interactions between the organisms.

Among these interactions, there is the **syntrophy** (or cross-feeding), defined as an 'obligately mutualistic metabolism'. In this type of interaction, the product of the metabolism of a microorganism can be used by another (B. E. Morris et al. 2013). Multiple types of syntrophy have been identified according to the processes used to share the metabolites and depending on the organism benefiting from it (Smith et al. 2019). Thus depending on these criteria, syntrophy can be either mutualism, commensalism or parasitism (B. E. Morris et al. 2013).

To investigate the ecological interactions within an environmental sample, it could be of interest to use the metabolic networks and method relying on dynamic modelling of the metabolism.

### 1.5.2 Available data: metagenomics genome and metabolic networks

It is possible to extract either complete genomes or Metagenome-Assembled Genomes (MAG) through metagenomics. A MAG is a genome reconstructed by binding one or more metagenome sequences that are supposed to represent an individual genome. From these genomes, it is possible to predict a corresponding GSMN.

But as explain in section 1.4, in some metagenomics experiments, only taxonomic affiliations are available as no genomes were sequenced. But thanks to methods such as the one described in Chapter 4, it is possible to estimate the metabolic capabilities of the

wild organism identified by the taxonomic affiliation. Then we have metabolic networks for our organisms.

In this way, there are two types of data when studying metagenomics: genomes or already reconstructed metabolic networks.

### 1.5.3 Estimating metabolic interactions in community

Numerous Constraint-Based modelling methods have been developed to study the metabolic interactions between species and especially cross-feeding (Chan et al. 2017; Zomorrodi et al. 2012; Khandelwal et al. 2013; Bauer et al. 2017; Mendes-Soares et al. 2016). But most of these methods work for small communities as they do not scale with the hundreds of organisms in a microbiota. More recent methods scale to bigger community (Zelezniak et al. 2015; Diener et al. 2020; Baldini et al. 2019).

But as these methods rely on Constraint-Based modelling, they require functional metabolic networks. These networks need time to be created depending on the organism studied. For Bacteria and Archaea, fast reconstruction methods have been developed (such as CarveMe or gapseq), but they rely on a top-down reconstruction and gap-filling.

A method relying on topological analysis has been developed to predict metabolic interactions in large-scale communities (Frioux et al. 2018b). From the metabolic networks of the community members, it selects minimal community producing metabolites of interest. It has been applied to the microbiota of the brown algae *Ectocarpus siliculosus* to establish consortia among ten Bacteria. The selected consortia provide an increase in the growth of the algae shown by co-culture experiments (Burgunter-Delamare et al. 2020).

### 1.5.4 Limit: scalability for predicting metabolic interactions in community

Multiple scalability issues can be identified when inferring metabolic interactions in communities from metagenomics data.

The first issue lies in the size of the GSMNs used for metabolic modelling. For Bacteria, GSMNs can contain up to thousands of reactions. Then, interactions between multiple sets of thousand reactions can be challenging to compute. But the Constraint-Based modelling methods have been able to cope with this issue and predict interactions between pairs of GSMNs.

A second scalability issue is the size of the community from metagenomics experiments. Indeed such community can contain hundreds or thousands of organisms (Pasolli et al. 2019; Forster et al. 2019; Zou et al. 2019; Stewart et al. 2018; Almeida et al. 2021). With this community size, most Constraint-Based modelling methods have a scalability issue. Furthermore, as these methods rely on testing metabolic interactions two by two, this limits the potential metabolic interactions as more complex interactions can be missed. But methods have addressed this issue, such as one using topological modelling (Frioux et al. 2018b). This method can predict metabolic interactions for communities with hundreds or more organisms. But it requires an already reconstructed metabolic network.

This reconstruction issue is the third scalability issue. As we have thousands of genome sequences, there is a need to reconstruct the corresponding GSMNs. Some tools have been developed to scale with these numbers, but they have often been designed for Bacteria and Archaea and use gap-filling to create functional GSMN. So there is a need for a method to reconstruct GSMN at a large scale for different organisms. With such methods, it could be possible to estimate the metabolic interactions between organisms using already developed methods.

But these methods produce the fourth issue of scalability; the results from such analyses can be difficult to analyse. For example, an analysis between the human metabolic network and the 773 human gut microbiota networks produced 381 minimal communities (Frioux et al. 2018b). These minimal communities were the group of interacting organisms that could produce metabolites of interest. But analysing this number of minimal communities can be difficult, so there is a need for a method to help understand these results.

### 1.5.5 Contribution: identification of key species in large-scale community

To answer these issues, we proposed a workflow in Chapter 5 called Metage2Metabo. This method reconstructs draft metabolic networks from genomes (either Metagenome Assembled Genomes or complete genomes). It can also use already metabolic networks. Then it computes individual and community production to estimate the potential of metabolic cooperation. Finally, it identifies key species thanks to metabolic complementarity among large-scale microbiota.

## 1.6 Thesis contribution

Metabolism is a central biological process and, as such, of fundamental interest in life science. Metabolic functions can be studied at different levels raising specific problems and investigation methods. Considered at the metabolic pathway level, a series of enzymatic functions lead to the production of a metabolite of interest. The metabolic functions at the cell or organism level allow the organisms to grow in specific ecological conditions. At a taxonomic level, akin organisms share functions inherited during their evolution, and some of the metabolic functions appear redundant within a taxonomic group. Finally, at a community level, organisms interact, and the metabolic functions of all the individual organisms can achieve individual and collective production of metabolites.

This thesis will present four contributions to studying metabolic functions and their diversity among organisms at all four levels. The metabolic diversity corresponds to all the different metabolites, reactions and pathways of the organisms and communities maintaining their activities along their evolution.

First, the pathways shared among related organisms could undergo neutral variations, or pathway drift, which can be predicted to explore the diversity of metabolism in non-model organisms. I present a method to predict the alternative pathways which might have diverged through a pathway drift evolution (Chapter 2).

Second, understanding how species metabolisms have diverged over time is a crucial problem in evolutionary biology. I present a method to homogenise the annotations of multiple input genomes from public databases when reconstructing GSMNs. Then it is possible to compare those GSMNs to describe their differences and equivalences (Chapter 2).

Third, wild species are described only with a taxonomic affiliation (and without sequence data) in several metagenomics approaches. I present a method to estimate their metabolic capabilities based on the proteins shared among their taxonomic groups (Chapter 4).

Fourth, synthetic community design selects several candidate organisms according to their predicted interactions in their environment. I present a method to estimate the metabolic interactions within an organism community and to predict the minimal communities that possibly collectively achieve some targeted metabolite productions (Chapter 5).

These approaches aim to predict candidates of interest, alternative metabolic path-

ways, comparable GSMNs, taxonomic-based metabolic networks and key species in community interactions. Biological questions drive the predictions, and they could be helpful in testing hypotheses. Further biological experiments could be performed to test the proposed candidates.

**Publications** This thesis is based in part on previously published work. The Chapter 2 is associated with a publication written in collaboration with a team from the Station Biologique de Roscoff (abbreviated in SBR, LBI2M UMR8227) and published in the journal *iScience* (Belcour et al. 2020b). The Chapter 5 was created from a publication in the journal *eLife* (Belcour et al. 2020a). Furthermore, the methods presented in Chapter 3 and 4 are extracted from submitted articles. Chapter 3 is created from an article written in collaboration with a team from SBR (LBI2M UMR8227) and has been submitted. The corresponding text is available in a preprint [1]. The method presented in the Chapter 4 is in part coming from an article which has been submitted and the corresponding text is available as a preprint [2]. The subsections 4.3 and 5.3 correspond to an article being written with Patrick Dabert (UR OPAALE, INRAE).

Furthermore, multiple articles in which I was a co-author have been published during this thesis. Among these articles, some are linked to the subject of this manuscript such as the article published in *Antioxidants* (Nègre et al. 2019) written in collaboration with SBR, the article published in *PeerJ* (Karimi et al. 2021) written in collaboration with SBR and the article published in *Frontiers in Plant Science* (Girard et al. 2021). Other articles are not cited such as an article published in *Journal of Phycology* (Xing et al. 2021), an article published in *Genomics* (Daval et al. 2019) and a last article published in *Microbial Biotechnology* (Daval et al. 2020).

**Software** All the software developed during this thesis are open sources. Their codes (with readme and documentation) have been released on the GitHub website in different repertories. For Chapter 2, PathModel's source code is available at pathmodel/pathmodel. The AuCoMe tool presented in Chapter 3 is in the repository AuReMe/aucome. The code of EsMeCata, the method presented in Chapter 4 is available at AuReMe/esmecata. Finally, the method developed in Chapter 5 (Metage2Metabo) used multiple tools that were developed such as AuReMe/mpwt or optimised such as MeneTools and MiSCoTo. The code of the workflow of Metage2Metabo is available at AuReMe/metage2metabo.

1. https://www.biorxiv.org/content/10.1101/2022.06.14.496215v1
2. https://www.biorxiv.org/content/10.1101/2022.03.16.484574v1

Part I

# Predicting metabolic diversity from heterogeneous data

---

I present in this part, the exploration on the metabolic diversity of non-model organisms. As explained in the introduction, a growing number of genomes are sequenced for non-model organism and the analysis of their metabolism is a key to better understand them. But these analyses are complicated by the annotation of the genomes and the lack of knowledge about these organisms.

Chapter 2 will focus on the diversity of metabolic pathways in non-model organisms. I will present a method to model the evolutionary drift of a metabolic pathway and to infer the possible alternative pathways. This work has been achieved in collaboration with the Station Biologique de Roscoff.

Chapter 3 will analyse the diversity of metabolisms inside group of organisms. In this chapter, I will describe my work on AuCoMe a method to create homogenized metabolic networks of a set of genomes in order to compare the metabolism of different taxonomic groups. The method was developed as a follow-up of the collaboration with the Station Biologique de Roscoff performed in Chapter 2.

Chapter 2

# Alternative pathways prediction from GSMN and metabolomics

In this chapter, I will present PathModel, a prototype to predict alternative metabolic pathways from a reference pathway, genomics and metabolomics data. It has been developed in collaboration with Jacques Nicolas (INRIA, Rennes) and Gabriel Markov (Station Biologique de Roscoff, Roscoff). PathModel implements a method predicting alternative metabolic pathways resulting from **Metabolic Pathway Drift**. From a data knowledge base, Pathmodel will predict alternative metabolic pathways with new reaction orders and possibly include data from metabolomics. The knowledge base contains metabolic pathways, biochemical reactions and metabolomics data. By using incremental logic programming, PathModel will propose a set of biochemical reactions and metabolite structures associated with the metabolomics data. PathModel was applied on two metabolic pathways (sterol and Mycosporine-like Amino Acid (MAA)) of the red alga *Chondrus crispus*.

This chapter has been extracted from the article published in the journal *iScience* (Belcour et al. 2020b).

## 2.1 Metabolic Pathway Drift

### 2.1.1 Developmental System Drift

An important research topic in developmental and evolutionary biology was understanding the homology of morphological features between species. Especially how conserved were the molecular pathways leading to the development of these features? Indeed, one could expect that similar molecular pathways produce similar morphological features. But it has been shown that developmental pathways could diverge through time without impact on the outcome of the morphological feature. These changes seem to be determined

not by natural selection (as the phenotypic outcome is not modified) but by chance. This led to the definition of the **Developmental System Drift** (True et al. 2001; Haag et al. 2018). This drift explains how similar morphological similar structures in different species can stay similar even if the molecular mechanisms underlying their formations undergo variations.

Similar drift phenomenons have been described in other fields, such as protein evolution (Hart et al. 2014). In this article the author explored the **thermodynamic system drift** of the protein ribonuclease H1 from *Thermus thermophilus* and *Escherichia coli*. Especially they found that the melting temperature is stable over time, but the mechanism underlying this fluctuates over time.

Another description of Developmental system drift was also made in gene expression of molar development in rodents (Sémon et al. 2020). Upper molars in rodents show a drastic change in morphology, whereas the lower molars display little morphological variations. But the authors found that gene expressions display a lot of variation across rodents in both upper and lower molars during molar development. So despite a similar structure in lower molar shared across rodents, developmental gene expression of these teeth shows variation.

In metabolism, multiple metabolic pathways are shared between species. Many variations can occur in these metabolic pathways. Thus we proposed that drift can also occur in metabolic pathways. Metabolic pathways of related organisms can produce the same output metabolite from the same input metabolite. This "phenotype" is under selective pressure and remains conserved. However, changes in the cascade of biochemical reactions and intermediate products can be neutral and subject to evolutionary drift.

### 2.1.2 Metabolic Pathway Drift

Starting from an ancestral promiscuous pathway (Figure 2.1 main pathway in teal; the upper part, alternative pathway in olive green), changes can occur either by non-orthologous gene displacement (in orange, left side) or by changes in reaction order, leading to a different intermediate metabolite (in olive green, right side). Substrate promiscuity enables the same molecular transformation on different molecules, enabling the enzyme to catalyze two different reactions. Promiscuity can be secondarily lost, as shown on the left side, leading to the impossibility of observing the star-shaped metabolite in contemporary metabolic pathway 1.

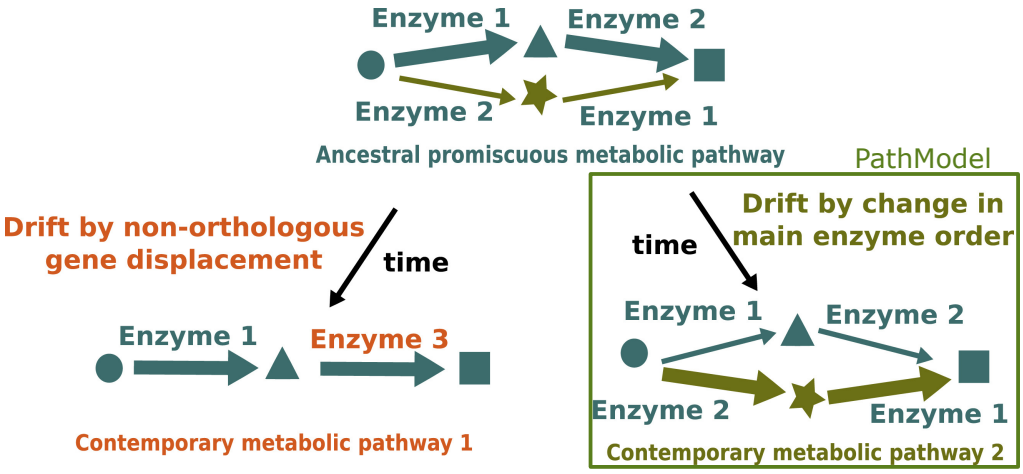



Figure 2.1 – The hypothesis of metabolic pathway drift is based on two possible elementary mechanisms. The first possible drift is the displacement of a gene by a non-orthologous one. The second drift is a change in the enzyme order, and it it is this context that is used for the alternative pathway prediction by PathModel.

### 2.1.3 Problems related to the formalism of the Metabolic Pathway Drift

As explained in the previous subsection, a Metabolic Pathway Drift can occur for a metabolic pathway by changing the reaction order. This change creates the context of our formalism: a metabolic pathway conserved in multiple species that can undergo a drift. We will also need additional information to propose this drift for a species of interest by combining different knowledge (metabolites present in the organism). Then we must create a method to explore the possibility of drift from these data. This raises four problems (Figure 2.2).

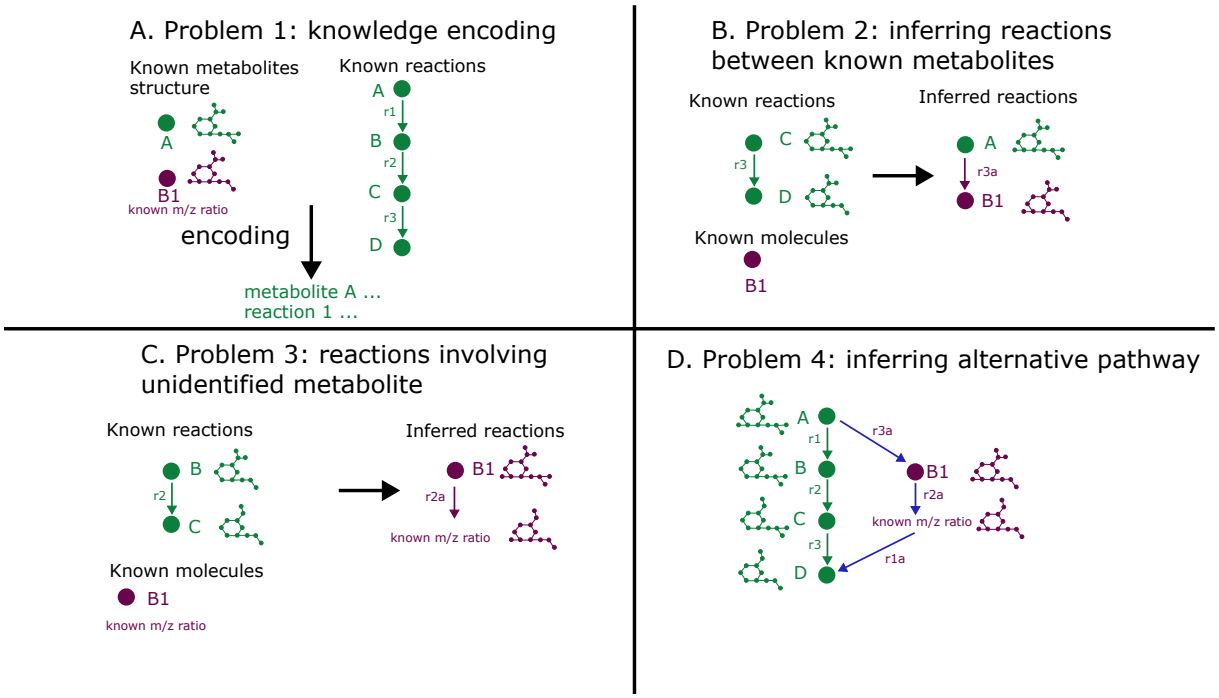



Figure 2.2 – **PathModel problems**. A. Encoding of the reference pathway and the known metabolites. B. Inference of reactions according to metabolites with known structures. C. Inference of reactions according to metabolites with only Mass-To-Charge ratio. D. Inference of new enzyme orders by using the inferred reactions. Metabolites and reactions in green are from the reference pathway, and the metabolites in purple are metabolites identified through metabolomics.

**Knowledge representation.** This is the first element of our formalism, how the metabolic pathway will be encoded (Figure 2.2 A). This way, we can rely on expert knowledge from metabolic databases (such as KEGG or MetaCyc), which often contain reference metabolic pathways from model species. And as metabolic pathways are sets of biochemical reactions, we also need enzyme information to know which biochemical reactions occur in the organism. It is possible to infer this information from genomics data by reconstructing GSMN. Biochemical reactions transform metabolites (reactants) into other metabolites (products). We also need to identify metabolites known to be present in our organism that could be substrates or products. Metabolomics can identify the presence of these metabolites in the organism. When using metabolomics, it is possible to detect known molecules using analytical standards. Analytical standards are known metabolites created with high purity that can be used to measure the presence of the corresponding metabolite in an organism. But not all metabolites have available standards and can not

be identified and measured with them. Then it is only possible to identify these molecules with a Mass-to-Charge ratio (m/z ratio).

**Problem 1: Knowledge encoding**

To test Metabolic Pathway Drift, we will have to handle data coming from different inputs such as **metabolic databases**, **metabolite structure**, **identified or unknown metabolite**.

Using the previous information, we will have to assess if drifts are possible, meaning if there are possible changes in enzyme order according to the known metabolites in the organism. This test can be decomposed into three problems.

**Inferring reactions between known metabolites.** We have known reactions inside our metabolic pathways with known metabolites. Before we search for change in enzyme order (meaning how the pathway is structured), we need to test if these reactions can occur on other metabolites known in our organism but not present in the reference metabolic pathway (Figure 2.2 B).

**Problem 2: inferring reactions between known metabolites**

From known reactions and metabolites, we have to check if **known biochemical reactions** can occur on **known metabolites of our organism** identified thanks to metabolomics approaches.

**Inferring reactions with unidentified metabolites.** As explained earlier, thanks to analytical agents, metabolomics can identify known metabolites. But it is also possible that unassigned peaks were found during the analysis leading to unknown m/z ratios. We have a third problem: we have to handle these data and map them to the reference metabolic pathway (Figure 2.2 C).

**Problem 3: reactions involving unidentified metabolite**

From known reactions and metabolites, we have to formulate a reasoning checking if **known biochemical reactions** can occur between a known metabolite and a **m/z ratio** identified through metabolomics.

**Inferring alternative pathways.** Finally, we have to combine the two previous methods to infer a possible alternative pathway from the reference metabolic pathways. From the input to the output of the metabolic pathway, we will have to see if a change in enzyme order using known metabolites from our organism can occur (Figure 2.2 D). Combining and solving these problems make it possible to infer alternative metabolic pathways satisfying the metabolic pathway drift. Problems 2 and 3 infer the set of alternative reactions that can be used, and Problem 4 identifies the order of these reactions to produce the target metabolites.

**Problem 4: inferring alternative pathway**

From the encoded knowledge of Problem 1 and the reasoning of Problems 2 and 3, we need a method to infer possible **alternative pathways**. These alternative pathways have the same input and output metabolites as the reference pathway but have different reaction orders inside them.

**Section summary**

From a **reference metabolic pathway**, we have to **encode** multiple information: **metabolite structure**, **biochemical reactions**, **m/z ratio of unknown metabolite**. Using these knowledge, we have to apply reasoning to test if **changes in enzyme order** are possible. So we need to predict reactions involving known metabolites or congruent measured m/z ratio for the species of interest. The goal is to infer a set of reactions composing the possible **alternative pathways** according to the **Metabolic Pathway Drift**.

## 2.2 PathModel implementation

### 2.2.1 PathModel formalism

**Pathway** We define a pathway as a directed compound graph $P = (M, R)$, where $M$ stands for metabolite nodes. A reaction $R$ links two metabolites nodes. When $(m1, m2) \in r$ with $m1 \in M$, $m2 \in M$ and $r \in R$, the metabolite $m1$ is called the *reactant* and the metabolite $m2$ is called the *product* of the reaction $r$.

**Metabolites** Metabolites are defined by a name, atoms (which have a number and a type $type(a)$ such as carbon, oxygen, etc.) and bonds linking atoms (we also know the bond type $bond(r)$ such as single or double bond). To handle the stereochemistry, it is possible to add to the bond type the corresponding configuration (R or S) such as $singleS$. The set of atoms and bonds describes the structure of a metabolite. This structure makes it possible to compute the m/z ratio associated with the metabolite $m/z\ ratio(m)$.

**Reaction** Reactions are defined by a name, their reactant ($reac(r)$) and their product ($prod(r)$). In this formalism, reactions have been simplified as the cofactors of the reaction have been removed and focusing on a pair of reactant and product in the reaction. This way, a reaction is associated with only one reactant and one product. This formalism limits the use of this approach on reaction associated with one reactant and one product. But it is not possible to use it on reactions with multiple reactants and multiple products.

**Molecular transformation** A molecular transformation $mt$ is inferred from a reaction. It is defined by the substructures modified by the reaction. To identify the substructures $sub$, PathModel compares the atoms and bounds of the reactant to the ones of the product. The differences between the two are identified as the substructure impacted by the reaction. For a molecular transformation, two substructures $sub$ are identified and consist of a set of atoms and bonds (either added, removed or modified by the molecular transformation). From these substructures, it is possible to compute the m/z ratio difference $diff(mt)$ between the reactant and the product.

**Known molecules** We defined a set of known molecules $K$ that are (1) molecules (with the same definition of a metabolite $m$ of the pathway) or (2) molecules only known from a $m/z\ ratio$ as they have not been identified.

**Metabolite production** In a pathway $P = (M, R)$ and with a set of seed metabolites $S \in M$, a reaction $r \in R$ is reachable from the seeds $S$ if its reactant in $reac(r)$ is reachable from $S$. Pushing forward this definition, a metabolite $m \in M$ is reachable from the set of seeds $S$ if $m \in S$ or if $m \in prod(r)$ is the product of a reaction $r \in R$ that is itself reachable from $S$.

**Pathway production** There is a path in a pathway $P$ from the seed metabolites $S$ to target metabolites $Tp$, which are the outputs of the pathway.

**Alternative pathway inference** From a set of seeds metabolites $S$, a set of target metabolites $Tap$ (for which we have $Tp \in Tap$), a reference pathway $P$ containing $M$ metabolites and $R$ reaction, a set of known molecules $K$ and $m/z$ *ratio*, PathModel will perform the following verification.

It will infer the molecular transformation $mt$ from the set of reactions by comparing the reactant and product of the corresponding reactions.

Then, it will infer a new reaction $r$ according to the molecular transformations found that satisfy one of the following reasoning: (1) we have a pair of molecules $\in (M \cup K)$ for which the difference of substructure *sub* between them correspond to a known molecular transformation $mt$ (solving Problem 2) or (2) from the m/z ratio difference $diff(r)$ of a molecular transformation $mt$, there exists one molecule $\in (M \cup K)$ which $m/z$ *ratio* is equal to the addition between the $m/z$ *ratio* of an unknown molecule and the m/z ratio difference $diff(mt)$ (solving Problem 3).

PathModel will iteratively apply these rules. Starting from the seeds, it will search for the metabolite products that can be created from the seed metabolites according to the known and inferred reactions. Then PathModel will iterate until it reaches all the target metabolites $Tap$ (solving Problem 4).

### 2.2.2 PathModel workflow

PathModel will try to infer alternative metabolic pathways from a reference metabolic pathway according to the possible drift for an organism of interest (Figure 2.3). To achieve this, Pathmodel relies on a **knowledge base** which needs to be **manually** created by the user. This knowledge base contains an encoding of the **reference metabolic pathway**, the **molecules** known in the organism of interest, the **structure** of the molecules (bond and atom) and **unassigned m/z ratio**.

PathModel will incrementally use two reasoning (Problem 2 and 3) to infer alternative pathways (Problem 4).

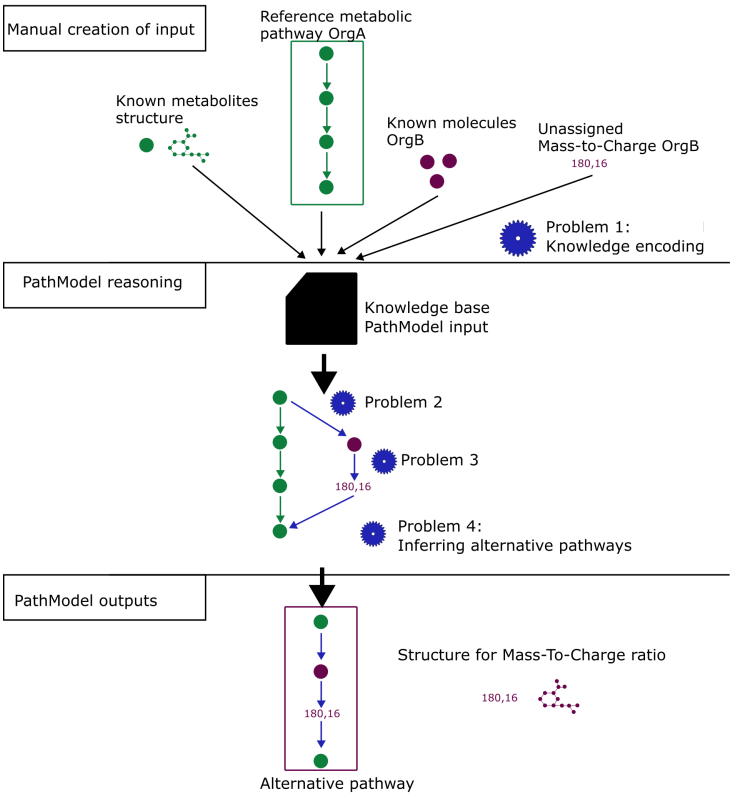



Figure 2.3 – PathModel workflow. The first step consists of manually encoding the data into a knowledge base. The knowledge base is a logic format written in ASP. Then the PathModel reasoning are used to infer new reactions. Then it will output alternative pathways and a possible structure for Mass-To-Charge ratios.

The output of PathModel is the alternative metabolic pathways and the possible structures associated with the m/z ratio of unknown molecules.

### 2.2.3 PathModel workflow application on a pathway in algae

To infer alternative metabolic pathways, we represent the different knowledge to be interpreted by the reasoning of PathModel. First, we select a metabolic pathway of interest from a metabolic database. Then the biochemical reactions and the molecules associated with this pathway are manually encoded. The available metabolomics data, such as known molecules in the organism and m/z ratios, are also added.

**Application to Mycosporine-like Amino Acids pathway in *Chondrus crispus*.** In this experiment, we want to study the possible pathway associated with the Mycosporine-like Amino Acids (MAA) for the red algae *Chondrus crispus*. This study was performed in collaboration with the Station Biologique de Roscoff (especially Gabriel Markov (LBI2M, Roscoff)). The tool AuReMe (Aite et al. 2018) was used to reconstruct a GSMN of *C. crispus* and meneco (Prigent et al. 2017) was applied to gap-fill the GSMN.

A reference pathway was created containing the MetaCyc pathway called PWY-7751 'shinorine biosynthesis' and reactions from Brawley et al. 2017; Carreto et al. 2011 (Figure 2.4).

Metabolomics analysis performed at the Station Biologique de Roscoff allowed the identification of MAA molecules present in *Chondrus crispus*. Using liquid chromatography-MS (LC-MS) profiling, they confirmed the presence of six MAA metabolites in *C. crispus*: asterina-330, palythene, palythine, palythinol, porphyra-334, and shinorine. In addition, they identified mycosporine-glycine in *C. crispus*. They also noticed that two unknown peaks potentially corresponding to MAAs were detected in the samples. These peaks exhibit m/z ratios consistent with peaks reported in a study on 40 red algae by Lalegerie et al. 2019. Specifically, they found a peak at 270,2720 that does not match any already identified candidate MAA, which they named MAA1, and a second one at 302.3117, which they called MAA2.

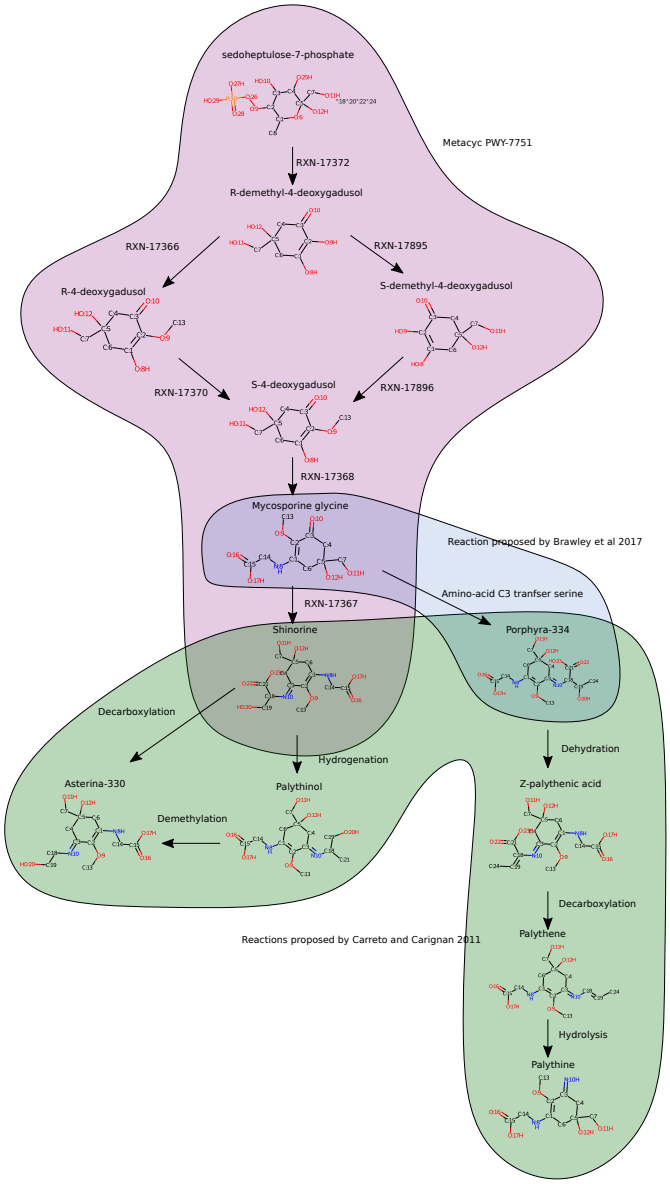



Figure 2.4 – Reference pathway for Mycosporine-like Amino Acids contains pathway from MetaCyc database (purple) and from Brawley et al. 2017; Carreto et al. 2011 (blue and green).

### 2.2.4 Problem 1: Knowledge representation

**Answer Set Programming.** To encode these knowledge, PathModel uses the Answer Set Programming language (abbreviated ASP, Lifschitz 2008). ASP is a declarative programming language. Declarative programming describes the problem instead of the

solution of the problem (compared to imperative programming). Furthermore, ASP is a logic programming as the problem is described with logical statements.

The base of a logical statement in ASP is the rule:

```
head :- body.
```

A rule is separated into a head and a body. The ':-' element corresponds to an 'if', implying that if all the elements in the body are satisfied, then the elements in the head are true.

If there is no body, such as in:

```
head.
```

This is called a fact, as it is always considered true.

On contrary, if there is no head:

```
:- body.
```

This is called a constraint. If the body is true, then nothing can be true as the head is empty. So to find a solution, the body must be false.

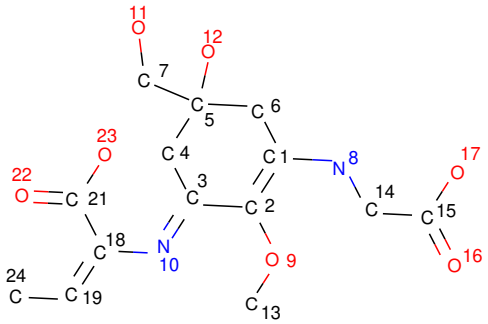
Head and Body are composed of *atoms* (in italic to differentiate it from the atom in molecules). An *atom* is a predicate followed by a set of terms in parenthesis. For example we represent the reaction between two molecules with $reaction(reaction\_1, "reactant", "product")$. The predicate $reaction$ is associated to the terms $reaction\_1$, "$reactant$" and "$product$". It is also possible to use variables (beginning with an uppercase) that can take different values.

A logic program is created to describe the problem by defining rules, facts, and constraints. In our case, the problem consists of a way to find alternative metabolic pathways from known metabolic pathways and metabolomics data. So we encode molecules and reactions associated with the pathway of interest. Metabolomics data are defined by the predicate $mzfiltering$ associated with a $mzratio$ ($mzfiltering(mzratio)$), which is an integer corresponding to the m/z ratio. All these encoded data consist of the instance of our logic program.

Then using this instance and the encoding rules (ASP scripts containing the reasoning of PathModel), a grounder will replace all variables with instantiated terms. Then a solver will use these terms to compute the answer sets (the stable model satisfying the rules written in the program).

The input molecules and reactions are encoded as facts on which the reasoning will be applied.

**Molecules encoding.** The chemical formula of the molecules is described by their atoms (identified by a number and atom types) and bonds (identified by atom numbers and bond type). Atom numbers are assigned manually to ensure consistency between molecules from the same family. We tried to follow the IUPAC conventions (such as Biochemical Nomenclature 1989) when existing. The same numbers are associated with the similar atom in different molecules (**consistent numbering across molecules**). This consistency of atom numbering between molecules allows the structure comparison performed by PathModel. Molecules are automatically associated with a theoretical m/z ratio, calculated using their chemical formula. For example, Figure 2.5 presents the encoding of the Z-palythenic acid from the MAA pathway.




atom("z-palythenic acid",1..7,carb). atom("z-palythenic acid",8,nitr). atom("z-palythenic acid",9,oxyg).
atom("z-palythenic acid",10,nitr). atom("z-palythenic acid",11..12,oxyg). atom("z-palythenic acid",13..15,carb).
atom("z-palythenic acid",16..17,oxyg). atom("z-palythenic acid",18..19,carb). atom("z-palythenic acid",21,carb).
atom("z-palythenic acid",22..23,oxyg). atom("z-palythenic acid",24,carb).

bond("z-palythenic acid",double,1,2). bond("z-palythenic acid",single,2,3). bond("z-palythenic acid",single,3,4).
bond("z-palythenic acid",single,4,5). bond("z-palythenic acid",single,5,6). bond("z-palythenic acid",single,1,6).
bond("z-palythenic acid",single,5,7). bond("z-palythenic acid",double,3,10). bond("z-palythenic acid",single,1,8).
bond("z-palythenic acid",singleS,5,12). bond("z-palythenic acid",single,7,11). bond("z-palythenic acid",single,2,9).
bond("z-palythenic acid",single,9,13). bond("z-palythenic acid",single,8,14). bond("z-palythenic acid",single,14,15).
bond("z-palythenic acid",double,15,16). bond("z-palythenic acid",single,15,17). bond("z-palythenic acid",single,10,18).
bond("z-palythenic acid",double,18,19). bond("z-palythenic acid",single,18,21). bond("z-palythenic acid",double,21,22).
bond("z-palythenic acid",single,21,23). bond("z-palythenic acid",single,19,24).

Figure 2.5 – The upper image is a 2D representation of the Z-palythenic molecule (containing 24 atoms without hydrogens and 23 bonds without counting bonds between atoms and hydrogens). Atom types are described (C: Carbon (black), O: Oxygen (red) and N: Nitrogen (blue)) with their numbers. The lower text corresponds to the molecule description in ASP.

**Common substructure.** PathModel will apply its reasoning only to metabolites sharing a similar substructure. So we encode a substructure that PathModel will use to check that two molecules are similar enough to use its reasoning on them. A substructure is encoded similarly to a molecule with atoms and bonds. And similarly, we keep the consistency of the atom numbering across the molecules and the substructures.

**Absent molecules.** It is also possible to encode absent molecules, meaning molecules that Pathmodel will ignore during its reasoning. Because we use reference pathways from other species, they can contain metabolites that are not present in our organism of interest; thus, we need to find alternative pathways without using these metabolites.

```
absentmolecules("z-palythenic acid").
```

**M/z ratio of unknown metabolite.** The m/z ratios associated with unknown metabolites from metabolomics data are encoded with the predicate $mzfiltering$ (and are multiplied by 10,0000 as ASP does not have float):

```
% For the mass-to-charge ratio 272.2720
mzfiltering(2702720).
% For the mass-to-charge ratio 302.3117
mzfiltering(3023117).
```

**Enzymatic reactions encoding.** The known enzymatic reactions of the reference pathways are simplified into one reactant/one product reactions. This is performed by removing the side metabolites (or cofactors) of the reactions (such as energy carrier, acceptor or donor of protons). Their IDs correspond to the reaction ID in the associated metabolic database or literature. Then they are encoded by using the predicate reaction, and it models the link between its reactant and its product:

```
reaction(decarboxylation,"z-palythenic acid","palythene").
```

**Pathway encoding.** The encoding of the pathway is performed using two predicates. A first predicate (source) defines the input of the pathway (either the reference or the alternative pathway). Then a goal is defined, corresponding to the source molecule and the output molecule expected by the pathway.

```
% Source molecule for inference.
source("sedoheptulose-7-phosphate").
% Initiation of incremental grounding.
goal(pathway("sedoheptulose-7-phosphate","palythine")).
```

**Subsection summary**

Using ASP, we manually encoded the information from different sources: metabolic databases, literature, Metabolomics data. In this way, we created a knowledge base on which PathModel can reason to infer alternative pathways for the reference pathway.

### 2.2.5 Problem 2 and 3: Reasoning on reactions

Using these encoded information, PathModel will use two reasoning to infer possible reaction between metabolites.

**Problem 2: inferring reactions between known metabolites.** From known reactions between known molecules, molecular transformations are inferred to identify the impacted substructure. Then this reasoning searches for molecules to which the molecular transformation can apply. The reactant and product are compared for each known reaction to identify the molecular transformation. The substructures impacted by the transformation are identified (for instance, the elimination of the carbon dioxide molecule as shown in green and purple in Figure 2.6).

PathModel will search for pair of molecules whose structures diverged only by the molecular transformation. For these pairs, PathModel will infer that this possible molecular transformation can happen between the molecules. An example is shown in Figure 2.6: from a known decarboxylation_1 between Z-palythenic acid and Palythene a pair of reactant/product is found in the knowledge base. Indeed there is also an elimination of a dioxide carbon between shinorine and asterina-330. PathModel assumes that a putative reaction exists between shinorine and asterina-330, similar to the decarboxylation between Z-palythenic acid and palythene.

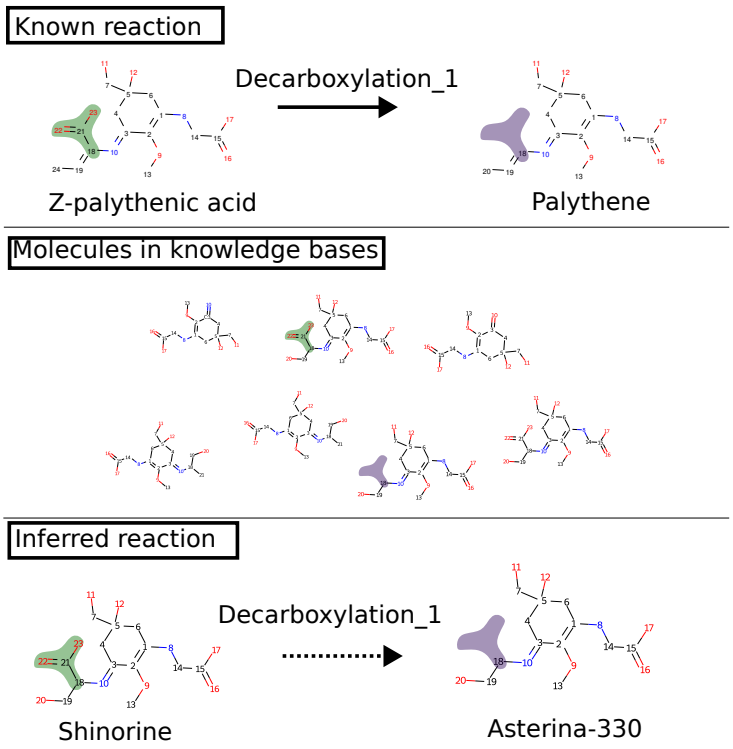



Figure 2.6 – Deductive reasoning from reaction Decarboxylation_1 converting Z-palythenic into Palythene. PathModel identifies molecular transformation and its associated substructures (green and purple). Then it searches for a pair of molecules having these substructures in the molecule knowledge base. If a pair is found, a reaction is inferred between the two molecules (here Shinorine and Asterina-330).

**Problem 3: reactions with unidentified metabolite.** In this reasoning, one of the molecules involved in the potential molecular transformation has no structure as it is only known with a m/z ratio. So by analogy to a known molecule, the m/z ratio will be used to infer the transformation.

First, PathModel will compute the difference in m/z ratio between the product and the reactant for each known reaction. Using this difference, PathModel will apply it to all the known molecules to compute a m/z ratio corresponding to the selected molecule after the molecular transformation implied by the known reaction. Then it will compare this computed m/z ratio to the m/z ratio given as input (representing unassigned peak in metabolomics). If there is a match, a possible new molecular transformation will be proposed using the molecule from which the m/z ratio was computed and the new molecule whose structure is newly created. In Figure 2.7, we search for possible decarboxylation

in our knowledge database. The reference reaction is the decarboxylation between Z-palythenic acid and palythene. First, we compute the change in m/z ratio between the Z-palythenic acid (328.3070) and the palythene (284.2975), which is 44.0095. Then, we compute the m/z ratio that we will search for to test molecular transformation. Here we have one unassigned m/z ratio from our metabolomics data (302.3117). By adding 44.0095, we get 346.3212 which is the m/z ratio for the reactant if the product of the reaction was the unassigned m/z ratio. And in our knowledge base, this equals the m/z ratio of Porphyra-334. Furthermore, Porphyra-334 contains the carbon dioxide structure impacted by the reaction. So PathModel infers a decarboxylation between Porphyra-334 and MAA2 (let us recall that it is the name associated with the m/z ratio 302.3117). Furthermore, PathModel creates the structure for the possible MAA2. Then the MAA2 becomes a molecule in the knowledge base and can be used to infer new reactions.

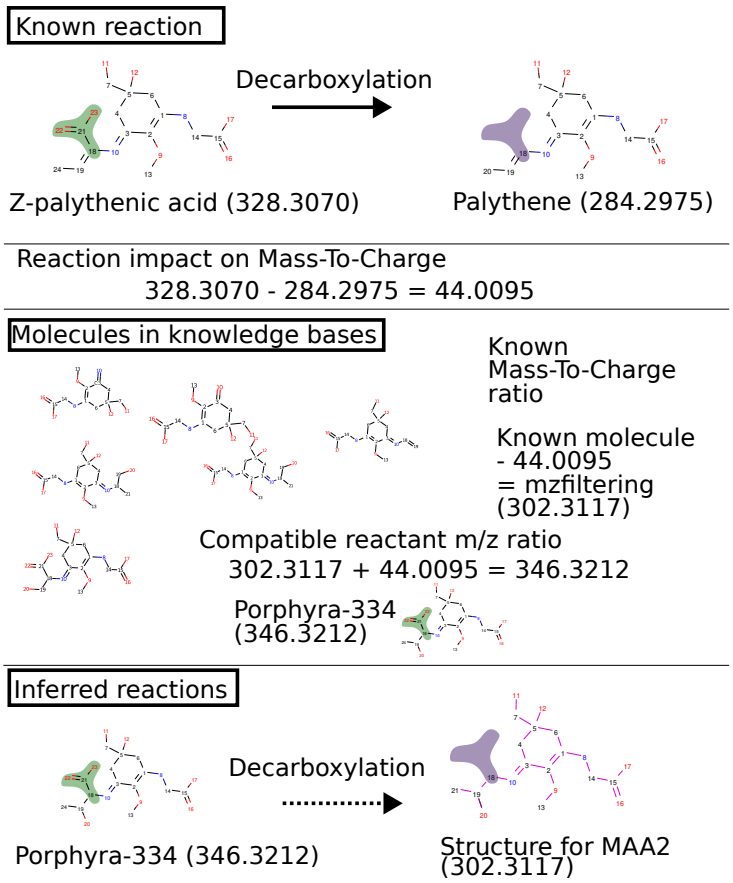



Figure 2.7 – Inference of reaction and metabolite structure corresponding to a m/z ratio of an unknown metabolite. First, PathModel computes the change of m/z ratio made by the molecular transformation by comparing the reactant and the product. Then, this number is added to m/z ratio measured by metabolomics to find the potential reactant associated with the m/z ratio. If this reactant contained the substructure impacted by the molecular transformation (in green), then PathModel will infer a reaction between the reactant identified and the m/z ratio. A structure is inferred for the m/z ratio (in purple) by using the structure of the reactant and applying the molecular transformation on it.

**Solving Problem 2 and 3**

By applying a set of rules to the knowledge base's encoded molecules, we could infer reactions between known metabolites and between a known metabolite and a m/z ratio of an unknown metabolite.

### 2.2.6 Problem 4: Inferring alternative pathway

For the encoding of the reasoning of PahtModel, among the ASP suite, we used clingo (Gebser et al. 2016). Furthermore, PahtModel uses the incremental mode of clingo (Gebser et al. 2019), a mode that will repeat grounding and solving until a solution is found. The idea behind the incremental mode used by PahtModel is to iterate through a metabolic pathway from a source molecule. Each iteration moves to new possible molecules from this source molecule according to the known reactions and the possible alternative transformations inferred from the reasoning (Fig. 2.8). By using this incremental mode, PathModel is able to monitor the new molecular transformations and the order in which molecules are reached, allowing for the reconstruction of alternative pathways. This mode uses 3 subprograms **base**, **step** and **check**.

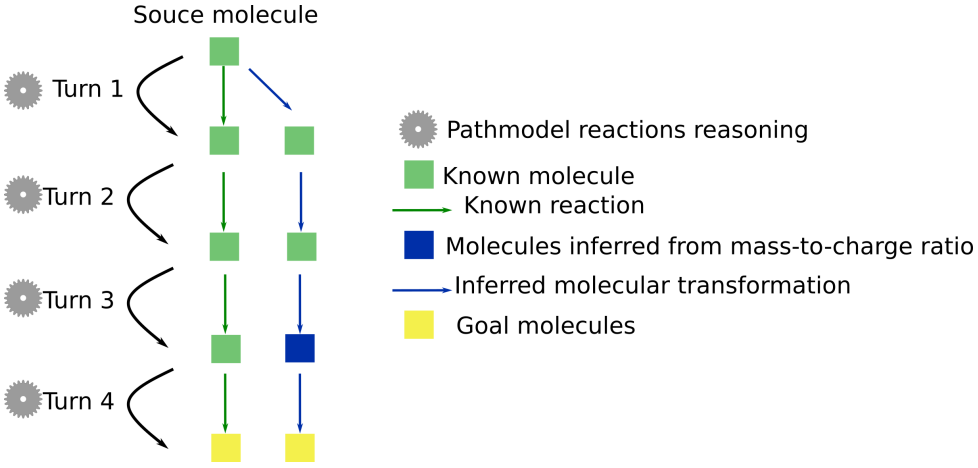



Figure 2.8 – Alternative pathway prediction by PathModel.

The subprogram **base** corresponds to the static knowledge that is not impacted by the iteration process. In our case, it is mainly associated with molecules' atom valence, input reactions, and metabolites.

The subprogram **step** defines the cumulative part. In PathModel, this subprogram contains the main reasoning rules that propose new transformations on molecules according to the known reactions. The cumulative part here is the advance in a known metabolic pathway. During this, the reasoning tries to infer alternative transformations according to the set of base reactions and the newly inferred transformations from the previous step.

The subprogram **check** consists of verifying if the goal is reached (the problem to solve). In our case, it corresponds to a couple of molecules. The first molecule corresponds to the beginning of the alternative pathway, and the second molecule corresponds to the

expected output metabolite of this pathway. We describe the beginning and the end of known metabolic pathways (described by the known reactions) and alternative pathways.

**Application to Mycosporine-like Amino Acids pathway in *Chondrus crispus*.** PathModel was applied to the reference pathway of MAA (Figure 2.4) by giving as input the encoded pathway, the sedoheptulose-7-phosphate as an input molecule for the pathway, the palythine as a goal and the two m/z ratios of unknown metabolites (MAA1 and MAA2). We were able to infer a possible alternative path in the reference pathway (Figure 2.9). From the sedoheptulose-7-phosphate to the mycosporine-glycine the structure of the pathway is similar to the reference. Then by using the m/z ratio of MAA 2, PathModel was able to infer a reaction between porphyra-334 and MAA2 (using the reasoning solving Problem 3). Then a reaction between MAA 2 and Palythene was inferred by using both known molecules (the reasoning solving Problem 2) as MAA2 structure was known from the previous incremental step. At the same time, a reaction was predicted between asterina-330 and MAA1.

For MAA1 and MAA2 PathModel predicted molecule structures according to the reactant that has been found by the reasoning of Problem 3 (respectively asterina-330 and porphyra-334). Interestingly, the structure of MAA 2 was identified as the Aplysiapalytine A after a search in literature (Kamio et al. 2011). And during the writing of the article of PathModel, an article was published demonstrating the presence of Aplysiapalytine A in red algae (Orfanoudaki et al. 2019).

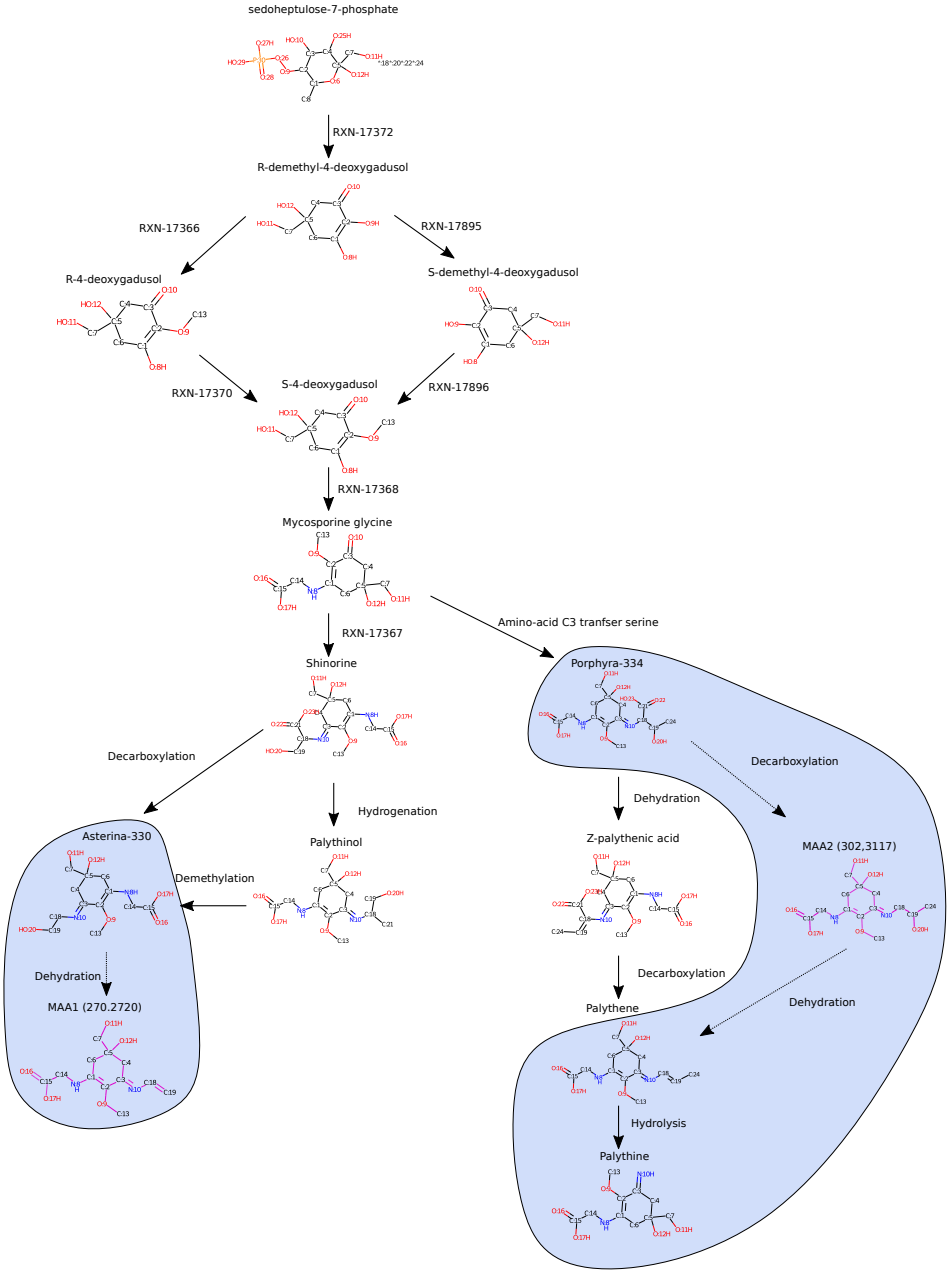



Figure 2.9 – Alternative pathway prediction by PathModel for the Mycosporine-like Amino Acids. The reactions inferred by PathModel are shown circled in blue. The new structures for the m/z ratio of unknown molecules (MAA1 and MAA2) appear in purple. By searching for producible metabolites, PathModel found a new way to produce Palythine (through MAA2) and a way to produce MAA1 (even if it was not a goal metabolite).

Manual searches performed by Gabriel Markov allow for the identification of a candidate gene (CHC_T00008892001) performing decarboxylation and dehydration transformations in MAA biosynthesis pathways. Further analysis and literature research (Heikinheimo et al. 2002), permits to hypothesize that this candidate enzyme may also perform serine/threonine decarboxylation on a serine/threonine linked to a mycosporine-glycin thus performing the reaction between porphyra-334 and MAA2. In the possible absence of Z-palythenic acid in *C. cripus*, the path using MAA2 instead could be a consistent alternative.

**Section summary**

To combine the input information, we manually created a **knowledge base** in ASP. Then we used the **incremental mode** of ASP to apply the reasoning iteratively on reactions to infer new reactions either on (i) **pair of known metabolites** or (ii) **a known metabolite and a m/z ratio of an unknown metabolite**. This was done by comparing the structure of metabolites. PathModel was applied on the **MAA pathway** to identify an alternative pathway for the red algae *C. crispus* and to link two m/z ratios of unknown molecules to the pathway (and predict their potential structures). It also proposed an alternative route using the m/z ratio to handle the possible absence of Z-palythenic acid in the studied organism.

## 2.3 Application on sterol pathway in algae

In this section, I will present the application of PathModel on the metabolic pathway of cholesterol biosynthesis in *C. crispus* (using the same GSMN as the one used with MAA).

**Sterol detection in *C. crispus*.** The presence of molecules was assessed by researchers from the Station Biologique de Roscoff. They performed gas chromatography-mass spectrometry for sterol detection. Standards were used to detect known molecules. Furthermore, literature analyses were performed to complete this detection of metabolites. Metabolomics analysis conducted by SBR researchers confirmed the presence of eight previously identified sterols (brassicasterol, campesterol, cholesterol, 7-dehydrocholesterol,

desmosterol, lathosterol, b-sitosterol, and stigmasterol). They identified an immediate precursor of sterols, i.e., squalene. However, they did not find evidence for cycloeucalenol, ergosterol, fucosterol, and zymosterol, which are intermediates present in other eukaryotes (Desmond et al. 2009; Sonawane et al. 2016). They also did not find cycloartenol, contrary to a previous report in *C. crispus* using thin-layer chromatography (Alcaide et al. 1968). This negative finding is strengthened by the fact that they are able to identify the cycloartenol standard when added to algal extract.

**Data encoding.** The reference pathway is the cholesterogenesis pathway (called "early side-chain reductase" or early SSR) based on the model previously published for tomato (Sonawane et al. 2016), which was added in MetaCyc (MetaCyc: PWY18C3-1). It also included portions of the canonical plant sterol biosynthesis pathway (MetaCyc: PWY-2541) and portions of the animal sterol synthesis pathway (MetaCyc: PWY66-4). These pathways can be seen in Figure 2.10 marked by a white box. The sterol knowledge base contained 15 enzymatic reactions involving 24 molecules from the cycloartenol to the stigmasterol and brassicasterol. This knowledge base included the eight unproducible sterols by the GSMN of *C. crispus*. It also contained the orphan molecule 22-dehydrocholesterol, which was not linked to any reaction in the *C. crispus* GSMN.

The source metabolite was cycloartenol. The goal metabolites were 22-dehydrocholesterol, brassicasterol, and stigmasterol. In addition, absent metabolites were ergosterol, fucosterol, and zymosterol, as these compounds were not detected using targeted profiling with analytical standards. We decided to use cycloartenol even if we did not find it by gas chromatography (GC)-MS for the following reasons. First, a cycloartenol synthase from the red alga *Laurencia dendroidea* was cloned and expressed in yeast cells, where it can transform squalene into cycloartenol, even if the authors did not report cycloartenol identification in the whole alga by GC-MS (Calegario et al. 2016). Second, unambiguous cycloartenol derivatives are known in another florideophyte red alga, *Tricleocarpa fragilis* (Horgen et al. 2000). Therefore, we considered it more parsimonious to hypothesise that cycloartenol is present and below the experimental detection limit rather than considering that this step is performed via an unknown intermediate.

**Alternative pathway prediction.** For this experiments, we focused on the identified sterol molecules so only the reasoning for Problem 2 was used by PathModel to infer the new reactions. PathModel reached all the goal metabolites linking the orphan molecule

22-dehydrocholesterol to the sterol pathway. PathModel proposed an alternative pathway from cycloartenol to cholesterol (in blue box in Figure 2.10). So two sterol synthesis pathways are possible, from cycloartenol to cholesterol, depending on when the side-chain reductase (SSR) enzyme is acting (Figure 2.10). If *C. crispus* uses the early SSR pathway (Sonawane et al. 2016), the metabolic intermediates would be identical to tomato, but there would be an important difference concerning the enzymes. Indeed, the genes encoding SSR are duplicated in Solanaceae (tomato and potato) but not in the *C. crispus* genome or any red algal and plant genomes, analysed so far. The unduplicated SSR from nonsolanaceous plants is catalytically promiscuous, and Pathmodel suggested that SSR could act on all possible intermediates.

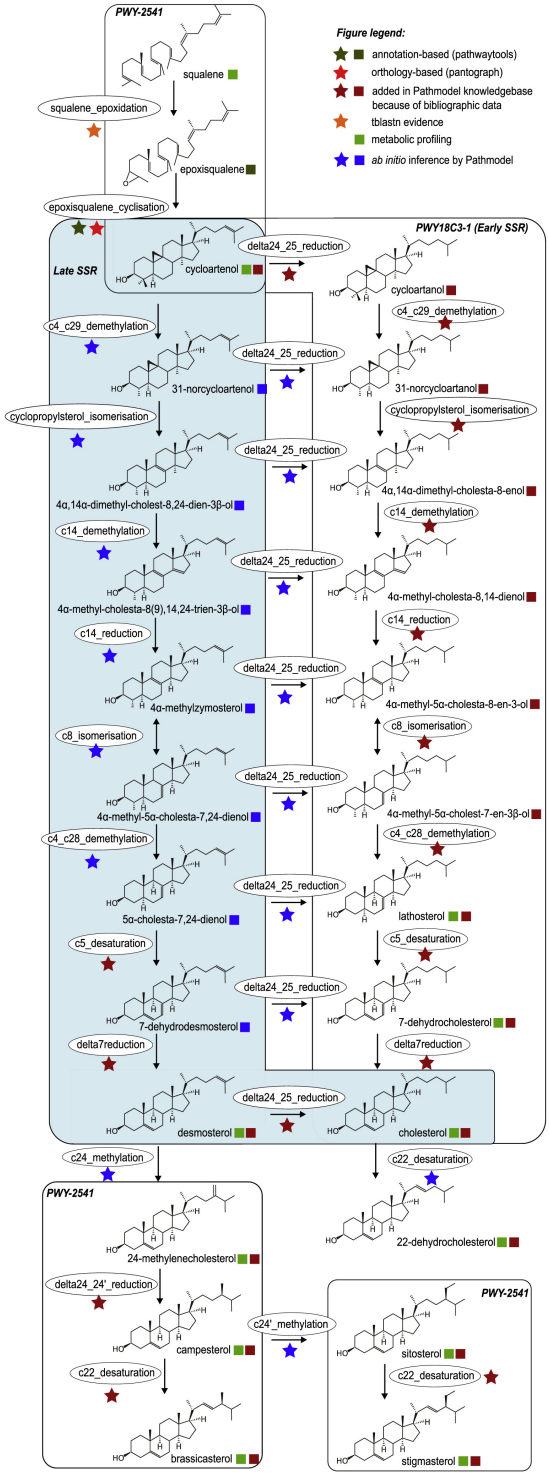



Figure 2.10 – A Model for Sterol Biosynthesis in *C. crispus*
Early SSR: pathway involving an early sterol side-chain reduction (SSR), also present in solanacean plants (PWY18C3-1). In the light blue box: late SSR pathway, involving a late sterol SSR, is only described in *C. crispus*. Portions identical to the plant sterol biosynthesis pathway (PWY-2541) are also boxed. Ovals indicate molecular transformations.

Another inference by Pathmodel consists in producing methylated sterols through C24-methylation on desmosterol (Figure 2.10). This is in agreement with the identification of a methylated sterol, 24-methylenecholesterol, in *C. crispus* (Tasende 2000) and builds on with other reports about methyltransferase catalytic promiscuity across land plants and green algae (Neelakandan et al. 2009; Haubrich et al. 2015). This option highly reduces the number of non-identified methylated intermediates. Indeed, in land plants, a first methylation, involving methyltransferases, occurs directly on cycloartenol. And the second one occurs later on 24-methylenelophenol, to produce methylated sterols like campesterol or brassicasterol (Benveniste 2004) with intermediates such as cycloeucalenol or fucosterol, both compounds for which we did not find any evidence of presence. By specifying the absence of fucosterol in our model, we naturally omitted this possibility. This model also seems more relevant from a quantitative viewpoint regarding the formation of cholesterol as the main sterol because this late methylation step would enable the production of methylated sterols using the late SSR pathway.

**Section summary**

PathModel was applied to the sterol biosynthesis pathway by combining knowledge from literature, metabolomics data and GSMN. PathModel inferred an alternative pathway from the one known in tomatoes. The significant change is a modification of the enzyme order, especially with the enzyme side-chain reductase being at the end of the pathway instead of at the beginning. Thus PathModel proposed an alternative pathway for sterol biosynthesis in alga according to the Metabolic Pathway Drift.

## 2.4 Conclusion

### Contribution

With PathModel, we developed a prototype predicting alternative pathways resulting from the Metabolic Pathway Drift. From a reference metabolic pathway containing known reactions and metabolites, known molecules and m/z ratio from metabolomics data, this method predicts reactions of an alternative metabolic pathway. The predicted alternative metabolic pathways help map metabolomics data to known pathways, allowing for the expansion of the possible knowledge on the metabolism of the studied organism.

It has been used on the GSMN of *C. cripus* and were able to predict two alternative metabolic pathways (for sterol and MAA). Combining chemoinformatics, metabolomics and system biology with logic programming made these results possible. This combination allowed us to map m/z ratios on metabolic pathways from GSMN and to associate m/z ratios of unknown molecules to known metabolites and metabolic pathways. Furthermore, we learned that despite conservative sterol and MMA production in organisms, the metabolic pathways could undergo substantial modification of their structure.

This highlight the utility of this approach as it can allow for the exploration of alternative metabolic pathways and increase the available knowledge. Indeed the pathways predicted by PathModel and curated by an expert were incorporated into the MetaCyc database.

## Limits and improvements

Further development can be made. Indeed, as it is a prototype, the creation of the input is manually made from the metabolomics data and the reference pathway. It is a time-consuming task which can be difficult. Thus, creating an automated method to create the input could improve the use of the tool. But this needs to tackle the issue of atom mapping as PathModel needs that each atom number of the different molecules corresponds to the same atom. A way to automatise this could be by using atom mapping method such as RDT (Rahman et al. 2016) to create automatically the input for PathModel.

Also, PathModel defines reaction as a simple transformation from one molecule to another. But this is a simplification compared to the reality where reactions often have cofactors, so improvements could be made to handle multiple products and reactants. These additions could also check that no atoms are lost in the process (between atoms from reactants and atoms from products).

## Perspectives

As a predicting prototype for the Metabolic Pathway Drift, PathModel allowed the incorporation of unknown molecules into known metabolic pathways. This project highlights the importance of relying on biological rules for bioinformatics methods. A perspective could be to expand such rules and take them into account directly when reconstructing a GSMN. In our experiment, we reconstructed the GSMN of *C. crispus* using AuReMe, a method relying on model organisms to transfer the annotation. PathModel

could become a step of such tool to automatically propose alternative pathways from these model organisms and their reference pathways. It could automatise the increase of metabolic knowledge. This could benefit from combining the metabolic database, chemical and metabolomics database to explore a wider universe of metabolites and propose possible new metabolic pathways.

The second perspective could be on the "Drift by non-orthologous gene displacement" presented in Figure 2.1. In the current chapter, we focused on the change in main enzyme order, but the method developed could handle the other type of drift. Indeed, by adding reactions that non-orthologous genes can catalyse, it could be possible for PathModel to infer this type of drift. But this requires the ability to identify such displacement of genes, which could be performed with comparative genomics.

Chapter 3

# Inferring comparable GSMN from heterogeneously annotated genomes

In this chapter, I will present AuCoMe, a method to compare the metabolism within a taxonomic group using publicly available genomes. Several issues arose when trying to compare the metabolic networks of multiple organisms, especially due to the heterogeneity of the tools that have been used to annotate their genomes. This heterogeneity can lead to biases when comparing them, where the main differences correspond to annotation differences, not biological ones.

With AuCoMe, we developed a method to fix this issue by propagating the annotation of multiple organisms to other ones divided into four steps. During the first step, draft Genome-Scale Metabolic Networks (GSMNs) are reconstructed using publicly available genomes and their corresponding annotations. In the second step, a search of orthologs is performed. Then using the orthologs, a propagation of reactions is performed to homogenise the networks using the functional annotations of the genomes. The third step consists of a structural verification of specific reactions. From the pool of GSMNs created in the previous step, we compare pairs of GSMNs to ensure that the absence of a particular reaction is not due to a missing gene prediction. In this way, we homogenise the reactions according to the structural annotation. Lastly, a final step adds spontaneous reactions to complement the incomplete metabolic pathways.

To test AuCoMe we have applied it to 3 datasets (bacterial, fungal and algal), having different taxonomic diversities. We show that using the various steps of AuCoMe; we are able to obtain a knowledge database of metabolic networks for a taxonomic group that can be analysed to find differences between species from a metabolic perspective.

## 3.1 Comparison of Genome-Scale Metabolic Networks

### 3.1.1 Comparison and curation of already reconstructed GSMN

More and more GSMNs are reconstructed, leading to the possibility of comparing them. Multiple applications can be found for the comparisons of GSMNs. The first is creating a consensus network when multiple metabolic networks exist for the same organism.

In this scenario, a proposed strategy is the "reconstruction annotation jamboree" (Thiele et al. 2010b), a community effort to curate pathway discrepancies by examining reactions, Gene-Protein-Reaction (GPR) associations and metabolites in GSMNs to create a consensus GSMN for an organism. This strategy is relevant for organisms with multiple GSMNs to establish a reference. This strategy was successfully applied to *Saccharomyces cerevisiae* (Herrgård et al. 2008) as well as *Salmonella Typhimurium LT2* (Thiele et al. 2011). It was recently extended to multiple organisms to create the pan-metabolism of 33 fungi from the Dikarya subkingdom (Correia et al. 2020). Although platforms now facilitate such community efforts (Cottret et al. 2018), these methods are costly in terms of time and human resources. Another approach proposes to explore variety among multiple GSMNs.

One of these approaches is the "metabolic network reconciliation", which consists of comparing multiple GSMNs from different organisms to eliminate errors (Oberhardt et al. 2011). One key element is finding "reciprocal gene pairs", which are pairs of genes such that each gene is the best hit of the other gene according to a sequence alignment tool. This search is similar to the concept of Best-Bidirectional Hit (BBH) presented in Overbeek et al. 1999. A search for BBH was also used in (Hamilton et al. 2012) to find ortholog differences between GSMN and then search for differences in metabolic networks by testing changes in reaction flux between the GSMNs. These two methods have been applied to a few organisms. But other approaches to GSMN comparison try to explain the different phenotypes between numerous organisms according to the differences in their GSMNs.

For example, metabolic networks of 975 organisms from the three domains of life were created from the KEGG metabolic database with AutoKEGGRec (Karlsen et al. 2018). Then a binary metabolic network comparison was performed by computing the Jaccard Index on a matrix containing the reactions for each organism. This comparison showed a

strong similarity between the metabolic distances and the Tree Of Life, especially domains that were well-separated in terms of metabolic distance despite some organisms being misplaced (Schulz et al. 2020).

These methods apply to already reconstructed GSMNs. This approach had advantages as metabolic networks from the same databases will be homogeneously annotated. But it is limited by the available GSMNs. So numerous articles performed comparisons from genomes. These methods often re-annotate the genome before the GSMNs reconstruction and the comparison.

### 3.1.2 Re-annotating and propagating annotations to compare reconstructed GSMN

Reconstructing multiple GSMNs from genomes has been performed to understand the difference between organisms. But there are numerous caveats in these comparisons. For example, after the reconstruction of the GSMN of *Saccharina japonica* and *Cladosiphon okamuranus* (Nègre et al. 2019), they were compared to the GSMN of *Ectocarpus siliculosus* (from Prigent et al. 2014): a thorough analysis of the differences between their genomic contents revealed that heterogeneity of genome annotations may be more important than genuine biological differences. This heterogeneity showed the impact of the genome annotation tools used.

**Heterogeneous annotation problem**

How do we discriminate differences resulting from annotation issues from fundamental biological differences?

Numerous articles try to solve this issue by re-annotating genomes before the GSMN reconstruction to compare organisms' metabolism.

For example, an experiment was performed on the metabolic diversity on *Escherichia coli* (Vieira et al. 2011). Genomes were re-annotated using tools from the platform MicroScope (Vallenet et al. 2009). Then the GSMN were reconstructed using Pathway Tools (Karp et al. 2010). Furthermore, GPRs from EcoCyc (Keseler et al. 2009) were transferred to the organisms by finding BBH between EcoCyc and the organisms. They identified the core metabolism (set of reactions common to all strains) and the pan-metabolism (all the

reactions present in all the strains). Then using these GSMNs, they computed metabolic distances (computed with the Manhattan distances on reaction vectors) and phylogenetic distances between *Escherichia coli* and *Shigella* strains. The differences were better explained by the phylogeny than the pathogenic phenotypes except for the *Shigella* strains. Indeed these strains were far more distant in terms of metabolic distance than the phylogenetic distance. This incongruence could be explained by the parasitic lifestyle of the *Shigella* strains and the loss of metabolic functions (Vieira et al. 2011).

Another study on 301 genomes from the gut microbiota used RAST (Aziz et al. 2008; Brettin et al. 2015; Overbeek et al. 2014) to re-annotate the genomes and reconstructed the GSMNs with modelSEED (Overbeek et al. 2014; Henry et al. 2010; Seaver et al. 2021). Then they compared the GSMNs and found an overall congruence between the phylogenetic tree and the metabolic distances tree (Bauer et al. 2015).

A study on 24 *Penicillium* species used HMM model constructed from MetaCyc to associate reactions to the proteomes of each organism (Prigent et al. 2018). Then the reactions from an already known GSMN were propagated to the 24 organisms. After this, the networks were gap-filled using Meneco (Prigent et al. 2017) and a manual curation for reactions occurring in specific cell compartment locations (such as the mitochondrial reactions). This study found a similarity between phylogenetic clades and metabolic clades, but the phylogenetic signal was insufficient to explain the differences between the GSMNs of all the species. Furthermore, no connection was found between the metabolic distances and the species' habitat.

This strategy was pushed further and automatised in the tool CoReCo, which was developed to reconstruct gapless metabolic networks from several non-annotated genomes (Pitkänen et al. 2014; Castillo et al. 2016). The tool is separated into two phases. The first phase annotates the protein of the organisms with multiple methods. Three annotation methods are used to associate EC number and sequence, one by using Blast on Swissprot, a second with Global Trace Graph (Heger et al. 2007) and a third using InterProScan (Zdobnov et al. 2001). A Bayesian network model is created using all these results to score the enzymes probabilities for the input organism and hypothetical ancestral species. Also, the ECs are associated with KEGG reaction (Kanehisa et al. 2008). The second phase will reconstruct metabolic networks by using the enzyme probabilities and nutrients to create biosynthetic pathways validated by the transfer of the atoms in the nutrients among the metabolite of the pathway. CoReCo permits studying the evolution of metabolic networks by predicting enzyme probabilities in hypothetical ancestral species.

These methods rely on the re-annotation of the genomes. Still, as genomes in public databases are already annotated, one could want to use these annotations to see how these predictions could be analysed for metabolic network comparisons.

### 3.1.3 Is it possible to find relevant information from GSMN comparisons using available knowledge from public genomes?

As we have presented in the previous subsection 3.1.2, numerous studies compared GSMNs, especially by re-annotating genomes. In this chapter, I propose to explore the question of using the annotation of genomes already present in public databases to reconstruct GSMNs and compare them. But annotations of genomes in public databases can show disparities. For example, an analysis of the annotation coverage of bacterial genomes in the Genome Taxonomy Database indicates a variation of coverage ranging from 14% to 98% (Lobb et al. 2020). Thus, directly using these annotated genomes can cause issues, as some differences observed will only be associated with the annotation disparity.

A wide variety of methods exists to perform the structural (DNA features prediction such as a gene) and functional (association of biological information to genes) annotation steps (Ejigu et al. 2020). It has previously been shown to induce direct effects on the reconstructed GSMNs of bacterial symbionts (Karimi et al. 2021). Then comparing GSMNs reconstructed from different tools will introduce biases in the comparison.

The structural annotation can lead to differences if genes' structures are not correctly predicted or if different prediction tools are used. For example, let's look at a group of 40 public genomes, among them 36 being associated with species related to algae. We have to deal with high heterogeneity of the number of genes (Figure 3.1). How can we ensure that this heterogeneity comes from biological signals?

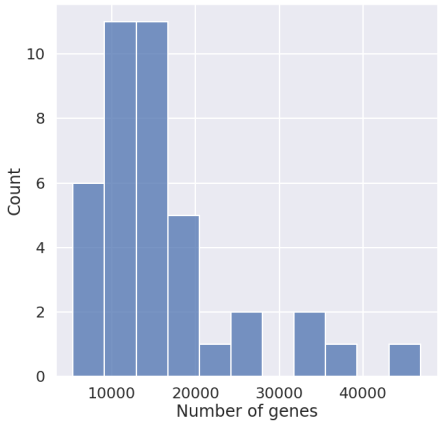



Figure 3.1 – Distribution of the number of genes (representing structural annotation results) across 40 genomes (with 36 organisms closely related to algae).

**Structural annotation problem**

Biological signals can explain differences in the number of predicted genes in the genomes, but they can also result from issues during the genome annotation. Contamination can impact the prediction by adding unrelated genes. Also, it is possible that some genes were not predicted, thus leading to missing annotations and, as the genes are absent to missing functions associated with the genes. This impacts organism comparison as we have artefactual differences.

Once the genes are predicted, the functional annotation will associate biological information with the genes by predicting the functions they perform. Such function can describe, for example, the role of an enzyme for a protein-coding gene. The function can be represented by specific terms described in numerous databases (such as the GO Term from the Gene Ontology (Ashburner et al. 2000) or the EC number from the Expasy ENZYME (Bairoch 2000)). This step can introduce another bias with a disparity in the functional annotations among the organisms. This can be illustrated for the group of 40 organisms by looking at the Enzyme Commission numbers (Figure 3.2). There are a lot of variations between genomes, some having 0 EC numbers and others thousands of EC.

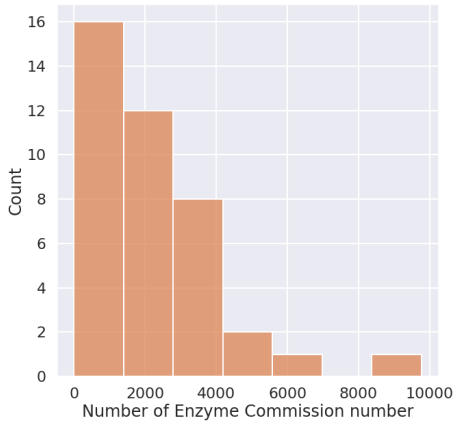



Figure 3.2 – Distribution of the number of redundant Enzyme Commission numbers giving an example of the available functional annotation across 40 genomes (with 36 organisms closely related to algae).

**Functional annotation problem**

Disparities in functional annotation of the genomes can lead to issues when comparing organisms. Indeed, if a gene encoding a known enzyme is not (correctly) annotated, we could miss this annotation for the GSMN, thus leading to an artefactual difference.

We propose a method to reconstruct comparable GSMNs from genome annotations by using the annotations of the publicly available genomes. To achieve this, we must decipher the biological signal (some organisms have more genes than others) from the noise (issues with annotations and GSMN reconstruction methods).

**Section summary**

From the **available public genomes** having potential heterogeneous structural and functional annotations, is it possible to create **comparable** GSMN suitable to identify differences of metabolism across the studied organisms?

## 3.2 AuCoMe: Create comparable GSMN from heterogeneously annotated genomes

AuCoMe is a python package that aims to build homogeneous metabolic networks starting from genomes with heterogeneous functional and/or structural annotations levels. AuCoMe propagates annotation information among organisms through a four-step pipeline (Fig.3.3).

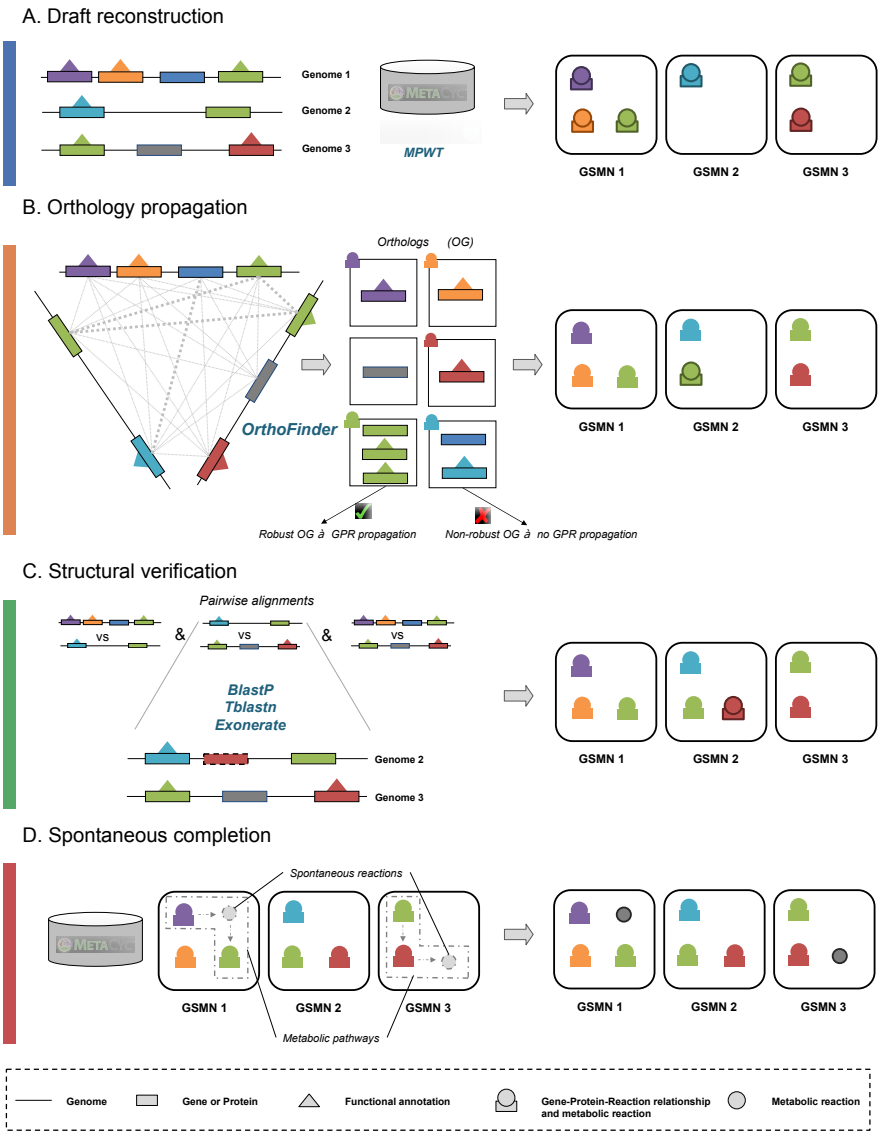



Figure 3.3 – **Reconstruction and homogenisation of metabolisms with AuCoMe.** Starting from a dataset of available public genomes, the AuCoMe pipeline performs the following four steps. **A. Draft reconstruction**. The reconstruction of draft genome-scale metabolic networks (GSMNs) is performed using Pathway Tools in a parallel implementation. **B. Orthology propagation**. OrthoFinder predicts orthologs by aligning protein sequences of all genomes. The robustness of orthology relationships is evaluated, and GPRs of robust orthologs are propagated. **C. Structural verification**. The absence of a GPR in genomes is verified through pairwise alignments of the GPR-associated sequence to all genomes where it is missing. If the GPR-associated sequence is identified in other genomes, the gene is annotated, and the GPR is propagated. **D. Spontaneous completion**. Missing spontaneous reactions enabling the completion of metabolic pathways are added to the GSMNs. GSMN: Genome-scale metabolic network. OG: orthologs. GPR: Gene-protein-reaction relationship.

### 3.2.1 Inferring pan-metabolism from the genomes functional annotations

The input files of AuCoMe are GenBank files. It is a file format that can contain the nucleic sequence of the genome and the results of the structural annotation (DNA features such as genes with their positions and the amino-acid sequence of the protein associated with the gene). It also contains the results of the functional annotations (such as GO Terms associated with genes).

The first step, **the draft reconstruction step**, consists in reconstructing draft GSMNs according to the set of available genome annotations. During this step, the pipeline first checks the input GenBank files using Biopython (Cock et al. 2009). Then using the mpwt package (Belcour et al. 2020a), AuCoMe launches parallel processes of the Patho-Logic algorithm of Pathway Tools (Karp et al. 2019). Pathway Tools creates one "Pathway/Genome Database" (PGDB) for each genome. The reaction inference of Pathway Tools relies mainly on Gene Ontology Terms, Enzyme Commission numbers and descriptions of genes (such as gene name and gene product). The inference of Pathway Tools is divided into several steps: it begins with the reaction inference, then performs a pathway inference where Pathway Tools filters the pathway according to the number of reactions inferred in them, the taxonomic ID of the input organism and the threshold given by the user. In all the experiments, we used the default threshold. During the pathway inference, if the organism seems to lack an enzyme for a pathway (if the pathway score is above the rejection threshold), then this reaction is kept and is called a pathway hole reaction.

The resulting PGDBs are converted into PADMet and SBML files (Hucka et al. 2003; Hucka et al. 2019) using the PADMet package (Aite et al. 2018). During this conversion, pathway hole reactions predicted by Pathway Tools are removed as they are not associated with a gene and are not spontaneous reactions. For example, in Fig. 3.3 A, the draft reconstruction step generates 6 GPRs in total for the 3 considered genomes.

This step will produce the significant part of reactions inferred by AuCoMe, the other being the spontaneous completion step. Indeed, Pathway Tools will predict for each organism a set of reactions. And by taking all of these reactions, we get all the enzymatic reactions (reactions catalysed by an enzyme different from the spontaneous reaction, which is a reaction occurring without an enzyme) inferred for our group of organisms. This set of reactions is defined as the pan-metabolism of the group.

**Definition 3.2.1 (Pan-metabolism)** *Following the definition of pan-metabolism pro-*

*posed in Vieira et al. 2011, we denote as pan-metabolism the set of all reactions of all the organisms of a taxonomic group.*

**Subsection summary**

The genomes input (in **GenBank file** format) were processed by a **parallel implementation** of Pathway Tools, we were able to reconstruct **draft metabolic networks** of multiple organisms thus creating the **pan-metabolism** of the studied group.

### 3.2.2 Propagating Gene-Protein-Reaction associations using orthology

The second step, the **orthology propagation step**, complements the previous draft GSMNs with GPR associations whose genes are predicted by propagating reaction between orthologs (Fig. 3.3 B). To that purpose, the pipeline relies on OrthoFinder (Emms et al. 2015; Emms et al. 2019) for the inference of *orthologs*. To identify orthologs, OrthoFinder first searches for the orthogroup (group of genes descending from a single gene in the last common ancestor of the studied organisms). Among the orthogroup, it is possible to find either paralog (gene emerging from a duplication event) or ortholog (gene emerging from a speciation event). For each orthogroup, an unrooted gene tree is reconstructed. From all the gene trees, a species tree is inferred, and then it will be used to infer a rooted gene tree and identify duplication events and orthologs.

For each orthologous gene shared between species, the pipeline checks whether one of the genes is associated with an existing GPR association (named *annotated GPR*), a GPR inferred during the draft reconstruction step. In that case, a putative GPR association with the orthologous genes is added to the GSMN (named *ortholog GPR*). At the end of the analysis of all genomes, a robustness score is computed to assess the confidence of each ortholog GPR association based on the number of annotated GPRs associations in the group of orthologous genes. If the group of orthologs contains two or more annotated GPRs then the annotated GPRs in the group are considered robust (Figure 3.4 A). If there is only one annotated GPR in the orthologous genes, then we will check the number of orthologs to which the annotated GPR propagates the reaction (Figure 3.4 B and C). If this number exceeds a robustness threshold, we will consider the ortholog GPR non-robust (Figure 3.4 B).

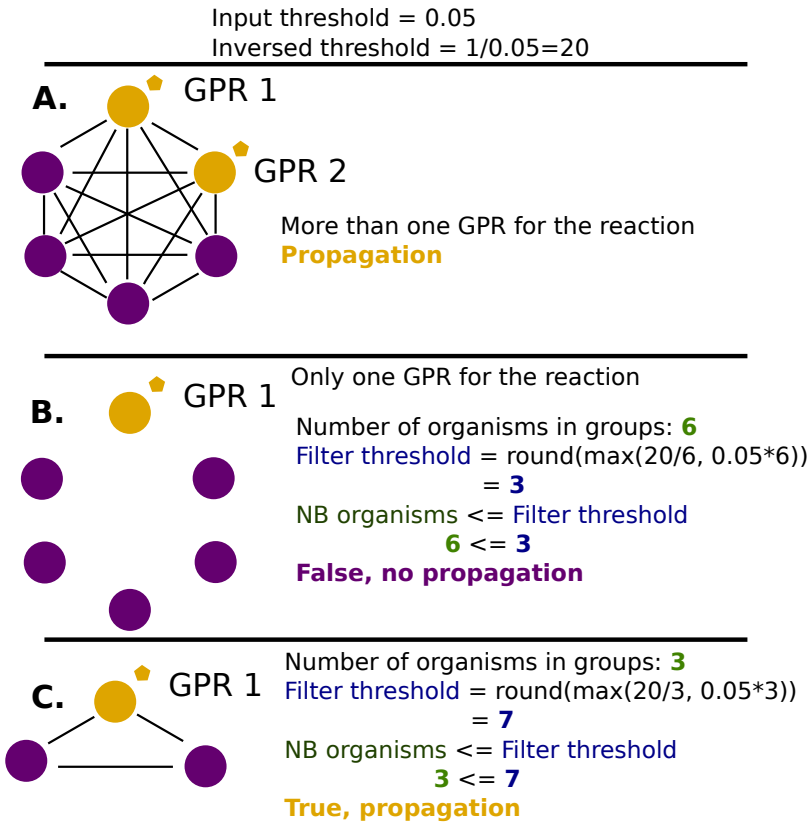



Figure 3.4 – Example of the filtering of method of AuCoMe on group of orthologs.

The robustness threshold is computed as the maximum between two formulas. The first took the inverse of an input threshold (by default, it is 5%) and divided it by the number of organisms involved in the orthologous genes. The second formula multiplies the input threshold (5% by default) by the number of organisms involved in the orthologous genes. The maximum between these two formulas defines the robustness threshold. If the total number of orthologs GPR associated with the reaction is superior to this robustness threshold, the ortholog GPR is considered non-robust. Having only one gene annotated for a large group of orthologs could be hazardous. Non-robust putative GPR associations are not integrated into the final GSMNs. If there are multiple reactions in the orthologous genes, each of them will be tested according to the filtering procedures separatly with the idea of function promiscuity.

In the example shown in Fig. 3.3 B, applying the robustness criteria leads to generating a putative new GPR association to the GSMN 2 (see the green orthologs as there is two annotated GPR among the orthologous genes). In this example, the pipeline does

not validate the GPR association related to the blue orthologs because of insufficient annotation support (one annotated GPR in the orthologous genes).

The goal of this step is to resolve the issue with the missing functional annotation (see the **Functional annotation problem**). To achieve this, we propagate reactions from the different organisms over the group of orthologs, and we filter the propagated GPR according to robustness criteria.

**Subsection summary**

In this section, we have seen that from the pan-metabolism associated with the draft GSMNs produced in the **Draft reconstruction step**, AuCoMe propagated the annotated GPRs by using **orthology**. This step complemented the GSMN by adding reactions that could be absent due to missing functional annotation. This results in the homogenisation of the GSMN and solves the **functional annotation problem**.

### 3.2.3 Structural verification of the absence of GPRs

The third step, **the structural verification step**, identifies GPRs associated with missing structural annotations in the input genomes. Suppose an enzymatic reaction is present in the pan-metabolism but absent in a GSMN. In that case, this pipeline step tries to complement the GSMN according to protein-against-genome alignment criteria using the sequences from the GSMNs containing the reaction. This search enables the identification of reactions associated with gene sequences absent from the initial structural annotations of the input genomes.

The GSMNs created in the previous step are taken as input. Pairwise comparison of the reactions in these GSMNs is performed (Fig. 3.3 C). In this comparison, if a reaction is missing in an organism (named *missing GSMN*), a structural verification is performed using the organism having that reaction (named *reference GSMN*). For each protein sequence associated with a GPR relation in a *reference GSMN* and if that GPR is absent in the missing GSMN, a blastp (Altschul et al. 1990) with Biopython (Cock et al. 2009) is performed on the proteome of the *missing GSMN*. The goal is to ensure whether or not a related protein could be found in the proteome of the *missing GSMN*. If no match with an evalue inferior to 1e-20 is found, no protein in the proteome of the *missing GSMN* can be associated with the reaction according to sequence alignment. Then AuCoMe searches

for possible gene candidates in the genome of the *missing GSMN*. This is performed by using a tblastn search (Altschul et al. 1990; Camacho et al. 2009) with Biopython (Cock et al. 2009) against the genome of the *missing GSMN*. If a match (evalue inferior to 1e-20) is found, the gene prediction tool Exonerate (Slater et al. 2005) is run on the region linked to the best match (region +- 10 KB). If Exonerate finds a match, then the reaction associated with the protein sequence is assigned to the *missing GSMN*. In Fig. 3.3 C, one reaction is added to the GSMN 2.

This step aims to solve the structural annotation issues, defined as the **Structural annotation problem**. This problem is solved by searching for each missing reaction if there is a homologous location on the organism genome that can perform the reaction.

**Subsection summary**

From the GPRs found during both the Draft reconstruction and the Orthology propagation, the **Structural verification** searched for missing reactions in organisms that resulted from **structural annotation problem**. By using protein-against-genome alignment between the **protein sequences** associated with the reaction and a **genome sequence**, AuCoMe tried to identify a homologous sequence corresponding to a missing gene prediction.

### 3.2.4 Spontaneous completion

The last step, the **spontaneous completion step** fills incomplete metabolic pathways with spontaneous reactions to complement each GSMN obtained after the structural-completion step with spontaneous reactions, i.e. reactions that do not need enzymes to occur. For each pathway of the MetaCyc database (Caspi et al. 2020), which was found incomplete in a GSMN, AuCoMe checks whether adding spontaneous reactions of the MetaCyc database could complete the pathway. When this is the case, spontaneous reactions are added to the GSMN. In Fig. 3.3 D, two spontaneous reactions are added to the GSMN 1 and GSMN 3. Then the final PADMet and SBML files are created for each studied organism.

**Subsection summary**

Using the metabolic pathways of MetaCyc, AuCoMe searched for **spontaneous reactions** that could fill **incomplete metabolic pathways** obtained after the structural verification.

**Section summary**

From **heterogeneously annotated genomes** (in GenBank format), AuCoMe reconstructed draft GSMNs with Pathway Tools. To solve the **issues in functional annotation**, reactions were propagated according to **orthology predictions** made by OrthoFinder. Then to cope with missing genes that could be omitted in the previous step, a **structural verification** was performed to search for genome positions matching known enzymes using blast and exonerate. After resolving these issues, the metabolic networks were complemented with **spontaneous reactions**.

## 3.3 Validation of AuCoMe on public datasets

The AuCoMe pipeline was validated on three datasets composed of genomes offering different levels of phylogenetic diversity and genome annotation heterogeneity. The datasets were used to assess the efficiency of the steps of AuCoMe to resolve the various annotation issues.

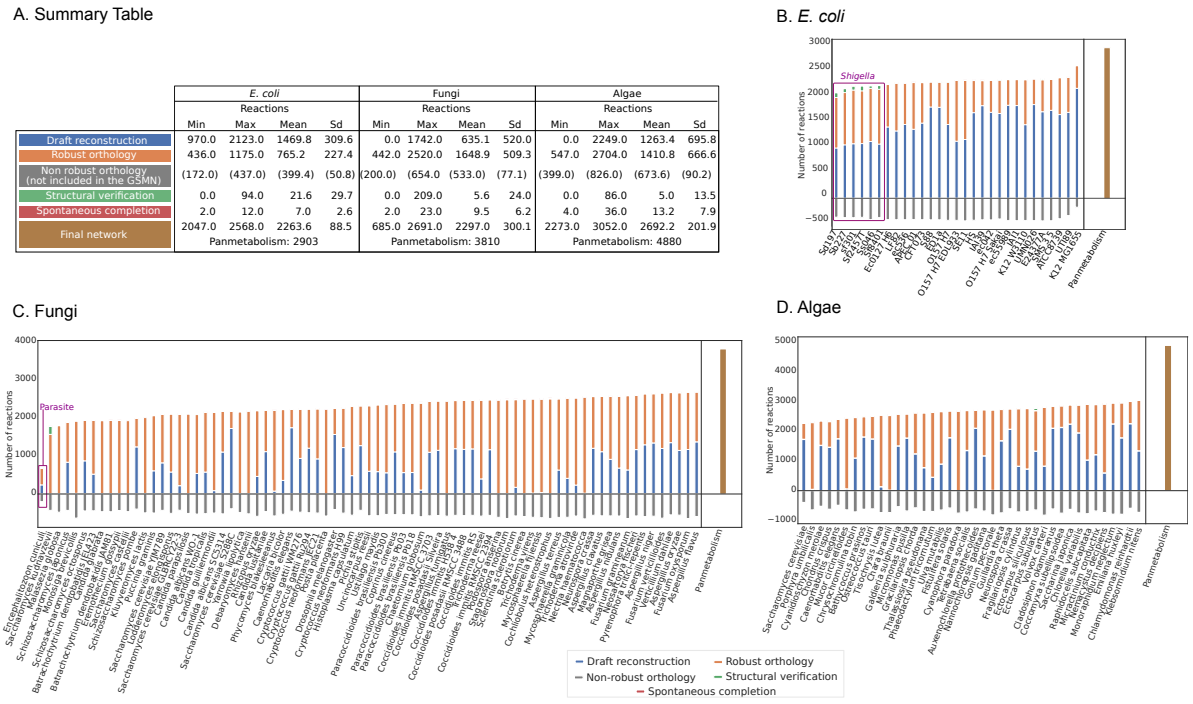
### 3.3.1 GSMN from available public genomes

**Input dataset.** The *bacterial dataset* includes the 29 bacterial *Escherichia coli* and *Shigella* strains studied in Vieira et al. 2011, downloaded from the NCBI GenBank Database (Sayers et al. 2019).

The *fungal dataset* includes 74 fungal genomes which were selected according to H. Wang et al. 2009 as representative of the fungal diversity, together with 3 outgroup genomes: *C. elegans*, *D. melanogaster*, and *M. brevicollis*. The genomes were downloaded from the NCBI GenBank Database (Sayers et al. 2019).

The *algal dataset* contains 36 algal genomes. These genomes represented a wide diversity of photosynthetic eukaryotes and were downloaded from public databases. The dataset includes 16 viridiplantae (green algae), 5 phaeophyceae (brown algae), 5 rhodophyceae (red algae), 4 diatoms, 3 haptophytes, 1 cryptophyte (*Guillardia theta*), 1 eustigmatophyceae (*Nannochloropsis gaditana*), 1 glaucophyceae (*Cyanodophora paradoxa*). The genomes of *C. elegans* (Witting et al. 2018), *M. circinelloides* (Vongsangnak et al. 2016), *N. crassa* (Dreyfuss et al. 2013), and *S. cerevisiae* (Lu et al. 2019) were selected as outgroup genomes.

**Resulting GSMNs.** AuCoMe was used on the three datasets. The number of reactions in each GSMN can be seen in Figure 3.5.




| | *E. coli* Reactions | | | | Fungi Reactions | | | | Algae Reactions | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Min | Max | Mean | Sd | Min | Max | Mean | Sd | Min | Max | Mean | Sd |
| Draft reconstruction | 970.0 | 2123.0 | 1469.8 | 309.6 | 0.0 | 1742.0 | 635.1 | 520.0 | 0.0 | 2249.0 | 1263.4 | 695.8 |
| Robust orthology | 436.0 | 1175.0 | 765.2 | 227.4 | 442.0 | 2520.0 | 1648.9 | 509.3 | 547.0 | 2704.0 | 1410.8 | 666.6 |
| Non robust orthology (not included in the GSMN) | (172.0) | (437.0) | (399.4) | (50.8) | (200.0) | (654.0) | (533.0) | (77.1) | (399.0) | (826.0) | (673.6) | (90.2) |
| Structural verification | 0.0 | 94.0 | 21.6 | 29.7 | 0.0 | 209.0 | 5.6 | 24.0 | 0.0 | 86.0 | 5.0 | 13.5 |
| Spontaneous completion | 2.0 | 12.0 | 7.0 | 2.6 | 2.0 | 23.0 | 9.5 | 6.2 | 4.0 | 36.0 | 13.2 | 7.9 |
| Final network | 2047.0 | 2568.0 | 2263.6 | 88.5 | 685.0 | 2691.0 | 2297.0 | 300.1 | 2273.0 | 3052.0 | 2692.2 | 201.9 |
| | Panmetabolism: 2903 | | | | Panmetabolism: 3810 | | | | Panmetabolism: 4880 | | | |

Figure 3.5 – **Application of the AuCoMe pipeline to the *bacterial*, *fungal* and *algal* datasets of genomes.** The figure depicts the number of reactions identified for each species at each step of the AuCoMe pipeline: reactions recovered by the *draft reconstruction* step (blue), non-robust reactions predicted by orthology propagation and removed by the filter (gray), robust reactions predicted by *orthology propagation* that passed the filter (orange), additional reactions predicted by the *structural verification step* (green), and *spontaneous completion* (red). The final metabolic networks encompass all these reactions except the non-robust ones. The *pan-metabolism* (all the reactions occurring in all organisms after the final step of AuCoMe) is presented in brown.

After the *draft reconstruction* step, draft GSMNs from the three datasets exhibited a

highly heterogeneous range of reactions (blue in Fig. 3.5 A, B, C, D). This heterogeneity was particularly visible for the fungal dataset, demonstrating the variability of the original genome annotations. In this dataset, no reaction was inferred from annotations in seven species, and for 12 of them, draft GSMNs contained less than ten reactions. For the latter, their respective genome annotations included no EC number, and eleven genomes did not have any GO term. Pathway Tools relies mainly on these two annotations to infer reactions, so this absence impedes the reconstruction process. Similar observations were also made, although to a lesser extent, for the algal genome dataset, with seven genomes having more than 2,000 reactions and seven genomes with less than 500 reactions. At this step, high heterogeneity in the number of reactions could be attributed mainly to differences in the quality and quantity of the genome annotations provided, precluding biologically meaningful comparisons of the GSMNs obtained at the draft reconstruction step.

After the *orthology propagation* step, we observed an homogenisation of the number of reactions in the datasets (orange in Fig. 3.5). GSMNs with few reactions after the draft reconstruction recovered more reactions during the orthology propagation step than GSMNs with thousands of reactions. This observation was supported by the negative correlation between the number of reactions added at the draft reconstruction step and the orthology propagation step (Spearman's rs -0.82 $p<0.001$). The fungal dataset exhibited an outlier at this step. The GSMN of *Encephalitozoon cuniculi* contained only 681 reactions compared to the thousands of reactions in the other fungal GSMNs. This difference is consistent with the fact that this species is a microsporidian parasite with a strong genome and gene compaction (Katinka et al. 2001). Among the reactions propagated by the orthology, a few hundred were removed from all datasets because they did not fulfil the robustness score criterion. A striking result in the bacterial dataset was the difference in reactions removed by the robustness score criterion between the strain *E. coli* K–12 MG1655 and the other strains. Indeed for this strain, 172 reactions were removed by the robustness criterion, less than the 333-447 reactions removed for the other strains. A possible explanation could be that the strain *E. coli* K–12 MG1655 is one of the best-annotated organisms, so it had the most reactions in the bacterial dataset. But as it was the only one to propagate these reactions, the robustness criterion removed them. These results indicated that most of the heterogeneity presented in the previous step could be solved by propagating reactions among orthologs, thus fixing issues with functional annotation.

Compared to the orthology propagation, the *structural verification* step had a more

negligible impact on the size of the final networks (green in Fig. 3.5). Ninety-five per cent of the GSMNs received less than 28 reactions during this step, and the maximum was 209. In the bacterial dataset, the six *Shigella* received an average of 76.2 reactions which was more than ten times the average of the other strains (7.4). We manually examined these differences and found that most genes in the GPR associations added at this step in the *Shigella*'s GSMNs corresponded to pseudogenes. For the fungal dataset, 209 reactions were added to the species *Saccharomyces kudriavzevii*. These reactions were associated with 192 sequences recovered during the structural step. These sequences were linked to transcripts from a previously published transcriptome dataset (Blevins et al. 2021). As for the algal dataset, 86 reactions were added to *Ectocarpus subulatus*. Likewise, we validated the presence of 59 out of 65 genes (83 out of 86 reactions) by associating them with existing transcripts. The remaining six genes (three reactions) corresponded to valid plastid sequences that had remained in the nuclear genome assembly. These genes lacked transcription data because plastid mRNA lacks the PolyA tail used to prepare most RNAseq libraries. In both outlier cases, the structural completion step was, therefore, able to recover sequences translated into mRNA and thus likely correspond to functional genes. This step showed mixed results as there was the detection of false positives (pseudogenes) in the bacterial dataset but found probable GPRs for the algal and fungal datasets. Furthermore, results showed that issues with structural annotation were far less present in the three datasets than issues with the functional annotation.

Lastly, the *spontaneous completion* step added spontaneous reactions to each metabolic network if these reactions complete BioCyc pathways (red in Fig. 3.5). This step added between 2 and 23 spontaneous reactions for the fungal dataset, leading to 2 to 27 additional MetaCyc pathways that achieved a completion rate greater than 80%. As for the algae, the same step added between 4 and 36 spontaneous reactions, leading to 2 to 31 additional pathways to reach a completion rate greater than 80%. Generally, we observed that the fewer reactions were inferred at the draft reconstruction step, the more spontaneous reactions were added to complete pathways (Pearson R = -0.83 and -0.84 for the fungal and algal datasets, respectively). This difference in completion could be explained by the fact that the only other possibility to introduce spontaneous reactions in the GSMNs was the draft reconstruction step because Pathway Tools infers GSMN with both enzymatic and spontaneous reactions.

**Subsection summary**

Two steps (*draft reconstruction* and *orthology propagation*) created most of the GPRs in the networks. The *draft reconstruction* produced highly heterogeneous draft GSMN, which were mostly homogenised by the *orthology propagation* and to a lesser extent by the *structural verification*. Following these results, the datasets contained genomes with high functional annotation variability but little structural annotation variability.

### 3.3.2 Complementarity of the AuCoMe steps to recover reactions using a bacterial dataset

In the following section, we characterised the efficiency of the pipeline for recovering missing reactions from genomes with randomly degraded annotations.

**Input datasets.** The complementarity between the *orthology propagation* step and the *structural verification* step was tested using replicate genomes of *E. coli* K–12 MG1655 modified by a simulated degradation of its annotations. Degraded replicates were then provided as input to the AuCoMe method. The resulting GSMNs were finally compared to the non-degraded dataset and the ground truth EcoCyc metabolic network (Karp et al. 2002b; Karp et al. 2018; Keseler et al. 2021) to estimate the precision and recall of the method.

To that goal, the non-degraded GSMN for *E. coli* K–12 MG1655 obtained from the run of the previous subsection was used to find the genes associated with the reactions. It allowed us to detect 2267 genes that were the target of the degradation.

Then, the *E. coli* K–12 MG1655 genome was modified to generate replicates with randomly degraded annotations chosen among the GPR of the non-degraded *E. coli* K–12 MG1655 GSMN. Two degradation types were simulated, (i) degradation of the functional annotations of the genes, where all the annotations like GO Terms, EC numbers, gene names, etc. associated with a reaction were removed, and (ii) degradation of the structural annotation of the genes, where gene positions and functional annotations were removed from the genome annotations. The third type of replicate was considered, including the degradation of structural and functional annotations. Replicates with increasing percentages of degraded annotations were generated for each of the three types of degradation. Furthermore, the taxonomic ID associated with the *E. coli* K–12 MG1655 genome was

degraded to *cellular organism* to focus on the impact of genome annotations on GSMN reconstructions by AuCoMe, rather than on the effect of the automatic completion by the EcoCyc source performed by Pathway Tools when analysing *E. coli* K–12 MG1655 .

Each degraded replicate was associated with the 28 other unmodified *E coli* and *Shigella* genomes, producing a synthetic dataset which was analysed by the AuCoMe method. This procedure, therefore, generated 31 synthetic bacterial datasets, plus the dataset with non-degraded *E. coli* K–12 MG1655 genome, which was called dataset 0 and already commented on in the previous subsection.

**Impact of genome degradation on AuCoMe steps.** First, we looked at the number of reactions predicted at each step. Then we performed the first validation by comparing the different output GSMNs obtained from the degraded genomes of *E. coli* K–12 MG1655 to the GSMN of the non-degraded dataset 0 (Figure 3.6).

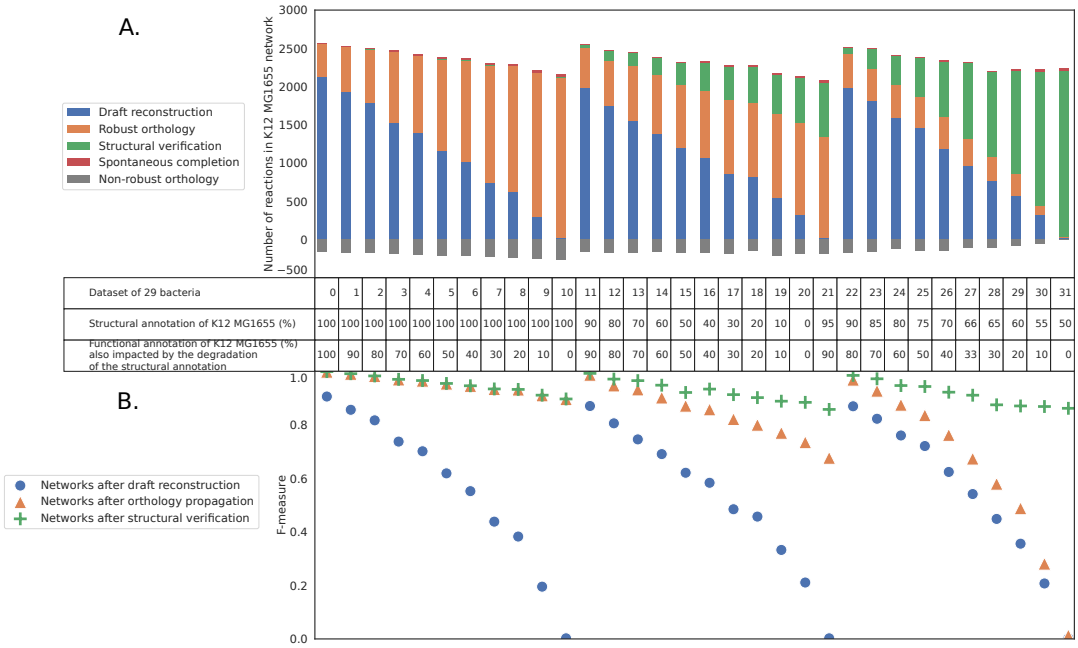



Figure 3.6 – **(A) Number of reactions in *E. coli* K–12 MG1655 degraded networks after application of AuCoMe to 32 synthetic bacterial datasets.** Each dataset consists of the genome of *E. coli* K–12 MG1655 to which degradation of the functional and/or structural annotations was applied, together with 28 bacterial genomes. Each vertical bar corresponds to the result on the *E. coli* K–12 MG1655 within a synthetic dataset, with the percentages of degraded annotations indicated below. The dataset labelled 0 was not subject to degradation of the *E. coli* K–12 MG1655 annotations. Three types of degradation were performed: functional annotation degradation only (left side, datasets labelled 1 to 10), structural annotation degradation only (right side, datasets labelled 22 to 31) and both degradation types (middle, dataset labelled 11 to 21). The coloured bars depict the number of reactions added to the degraded network at the different steps of the method (the blue, orange, green, grey, and red colour legends are as described in the figure 3.5). The table shown as an axis indicates the dataset number and the percentage of functional or structural annotation impacted by the degradation for the corresponding column in both subfigures. **(B) F-measures after comparison of the GSMNs recovered for each *E. coli* K–12 MG1655 genome replicate with the non-degraded dataset 0.** Reactions inferred by each AuCoMe step for each replicate were compared to the GSMN of the non-degraded dataset 0. For each comparison, the F-measure was computed. F-measures obtained after the draft reconstruction step, the orthology propagation step or the structural verification step are shown as blue circles, orange triangles, and green crosses, respectively.

Fig. 3.6 A illustrates the number of reactions predicted by the pipeline steps and each step's importance in the homogenisation of the GSMN sizes for each of the 32 synthetic

bacterial datasets. As expected, increasing degradation led to the recovery of fewer GPR associations in GSMN draft reconstructions (blue bars). When only functional annotations were degraded in the *E. coli* K–12 MG1655 genome (dataset labelled 1 to 10), the orthology propagation step enabled the recovery of a large number of reactions (orange bars) nearly up to the size of the GSMN associated with the non-degraded genome. In datasets where only structural annotations of the *E. coli* K–12 MG1655 genome were altered (datasets labelled 22 to 31), the structural verification step was the most important step to recover GPR associations (green bars). Finally, when both structural and functional annotations were degraded (datasets 11 to 21), combining the second and third AuCoMe steps recovered the discarded reactions. Notably, even when 100% of the *E. coli* K–12 MG1655 functional and structural annotations were degraded, the information from the other 28 non-altered genomes enabled the recovery of 2244 reactions (Fig. 3.6 A, dataset 31). These 2244 reactions represented 87% of the reactions in the non-degraded dataset 0.

According to the level of structural and functional annotations of the genomes, these results showed the complementarity between the *orthology propagation* step and the *structural verification* step.

**F-measure for internal and external validations.** These reactions were compared to two networks (1) in an internal validation, to the final GSMN of AuCoMe from the non-degraded dataset 0 and (2) in an external validation, to the literature-based curated network EcoCyc (Karp et al. 2002b; Karp et al. 2018; Keseler et al. 2021). In the following paragraph, we will refer to these networks as the *reference networks*.

We considered the reactions in a GSMN produced by AuCoMe and in a reference network as *True Positives (TPs)*. *False Positives (FPs)* were reactions that were present in the GSMN produced by AuCoMe but not present in a reference network, and *False Negatives (FN)* were reactions present in a reference network but not present in the GSMN produced by AuCoMe. There were no True Negative reactions because each considered reaction either belonged to the GSMNs produced by AuCoMe or to a reference network.

The F-measure of each AuCoMe dataset is defined as $F = \frac{2PR}{P+R}$, where $P = \frac{TP}{TP+FP}$ is the precision (number of reactions inferred by AuCoMe and present in a reference network among all the reactions predicted by AuCoMe) and $R = \frac{TP}{TP+FN}$ is the recall (number of reactions inferred by AuCoMe and present in a reference network among all the reactions in a reference network). The F-measure value is between 0 and 1. Values close to 1 indicate

that both precision and recall are high.

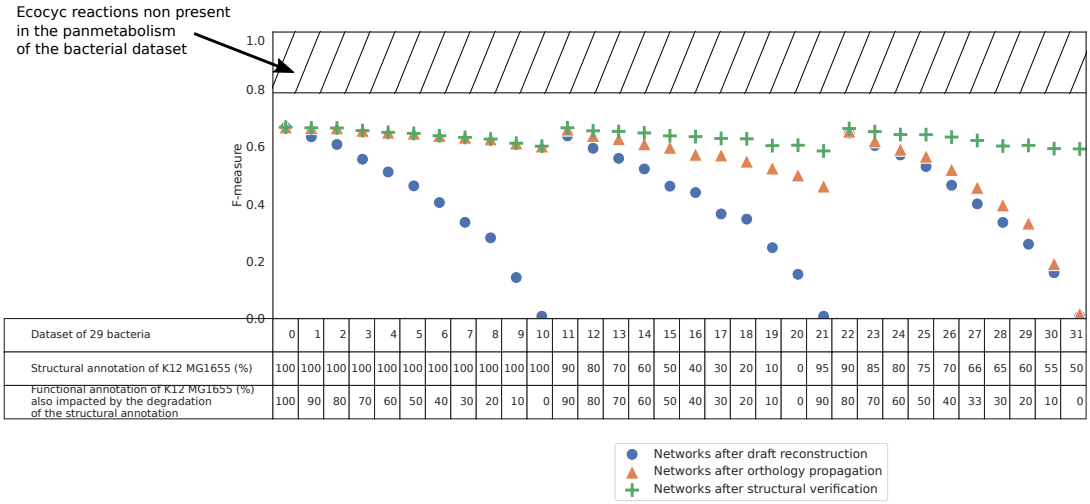
**Internal validation.** For the networks of the non-degraded dataset 0, the F-measure was 1 after the orthology propagation and structural verification steps. This was expected as the reference network here was the network from the final step of AuCoMe.

For the degradation on the functional annotations, the *orthology propagation* step helped to recover a F-measure greater than 0.9 (Fig. 3.6 B, dataset 1 to 10) after the degradation functional annotations.

The complementary of the two steps was demonstrated when both the functional and the structural annotations were degraded (Fig. 3.6 B, dataset 11 to 21). In these situations, the combination of the *orthology propagation* and the *structural verification* allowed to reach a F-measure around 0.9.

When degrading the structural annotation, it was the *structural verification* that helped to recover a F-measure around 0.9 (Fig. 3.6 B, dataset 22 to 31).

**External validation.** A padmet file containing the metabolic network of EcoCyc was created using the PGDB files from EcoCyc 23.5 and the padmet package. One thousand nineteen reactions contained only in EcoCyc and absented from all the metabolic networks of the 29 bacteria of dataset 0 were removed from the ground truth reactions as AuCoMe could never infer them. Indeed their absences from the pan-metabolism of the 29 Bacteria implies that Pathway Tools could not infer them from the genome annotations. Therefore they were not considered false negatives.




| Dataset of 29 bacteria | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Structural annotation of K12 MG1655 (%) | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 90 | 80 | 70 | 60 | 50 | 40 | 30 | 20 | 10 | 0 | 95 | 90 | 85 | 80 | 75 | 70 | 66 | 65 | 60 | 55 | 50 |
| Functional annotation of K12 MG1655 (%) also impacted by the degradation of the structural annotation | 100 | 90 | 80 | 70 | 60 | 50 | 40 | 30 | 20 | 10 | 0 | 90 | 80 | 70 | 60 | 50 | 40 | 30 | 20 | 10 | 0 | 90 | 80 | 70 | 60 | 50 | 40 | 33 | 30 | 20 | 10 | 0 |

Figure 3.7 – **F-measures after comparison of the GSMNs recovered for each *E. coli* K–12 MG1655 genome replicate with the gold-standard network EcoCyc 23.5.** Reactions inferred by each AuCoMe step for each replicate were compared to the gold-standard EcoCyc GSMN, allowing for F-measures' computation. F-measures obtained after the draft reconstruction step, the orthology propagation step, or the structural verification step are shown as blue circles, orange triangles, and green crosses, respectively. The hashed rectangle from F-measure 0.79 to 1 highlight the values of F-measure that are unreachable as 1019 reactions in EcoCyc were not present in the pan-metabolism of the 29 non-degraded Bacteria.

After the draft reconstruction step, the F-measure of non-degraded dataset 0 was 0.67. The F-measure remained at the same score for the other steps of AuCoMe. Therefore we considered 0.67 as the reference F-measure for this experiment.

As expected, increasing degradation of the genome annotations led to decreased F-measures after the draft reconstruction step (Fig. 3.7 B, blue circles). Furthermore, the F-measure consistently dropped to 0 when all functional annotations were degraded (datasets 10, 21 and 31). The F-measures then increased after the orthology propagation step in datasets where the degradation was performed on functional annotations (datasets 1 to 21, orange triangles). When only functional annotations were degraded, the orthology propagation step alone enabled the recovery of F-measures close to the reference F-measure of 0.67 (dataset 0). Regarding the datasets degraded in both functional and structural annotations, the structural verification step was additionally needed to reach F-measures close to 0.67 (datasets 11 to 31, green crosses). When all structural annotations of the genes associated with metabolism were depleted (dataset 31), the F-measure

obtained after the structural verification step was 0.60.

**Subsection summary**

Altogether, these results demonstrated that, by taking advantage of the annotations present in the other genomes of the considered dataset, AuCoMe built GSMNs with comparable amounts of reactions even for genomes completely missing functional and structural annotations. Furthermore, the internal and external validations showed the quality of the reactions inferred by AuCoMe.

**Section summary**

In this section, we have shown the results of AuCoMe on three **public datasets of genomes**. First, we highlighted a **high variability of annotations** in these genomes by looking at the results of the draft reconstruction. By combining the orthology propagation and the structural verification, we were able to **homogenise** the GSMN contents. Quality and quantity of recovered reactions were validated using multiple datasets with **degraded genome annotations**.

## 3.4 Application to algae

Using the algal dataset presented in the previous section 3.3, a more precise analysis was performed to compare the metabolism of the organisms in this group, yielding biological conclusions.

### 3.4.1 Phylogenetic consistency of the GSMNs clustering according to metabolic distances

To further assess the predictions of AuCoMe and to explore the distance between the metabolism of the algae, we performed a clustering. We clustered the GSMNs of all the organisms of the algal dataset according to the presence or absence of reactions in their GSMNs. Clustering obtained after the draft reconstruction and at the end of the

pipeline is firstly examined using multidimensional scaling (MDS, Fig. 3.8). The MDS was computed using the vegan package (Oksanen et al. 2020), and the distance matrix was created with the euclidian distance. We observed that the metabolism clusters obtained from the initial draft GSMNs produced from the annotations had a very low consistency with the phylogenetic relationships among the algae. Even well-established groups like red algae or brown algae were not recovered. An ANOSIM test supported this as it could not differentiate any known groups (R=0, P-value=0.4514). Furthermore, species from very different groups clustered together (circled in black in Figure 3.8) because they only had tens of reactions, compared to the hundred or thousand of reactions for the other organisms. So for GSMNs obtained after the draft reconstruction step, the principal factor leading to the GSMNs distribution in the MDS was the heterogeneity of genome annotations. However, in the GSMNs obtained after the final step of AuCoMe, we can see a clear separation of the organisms according to their known phylogenetic group (ANOSIM, R=0.811, P-value=1e-04).

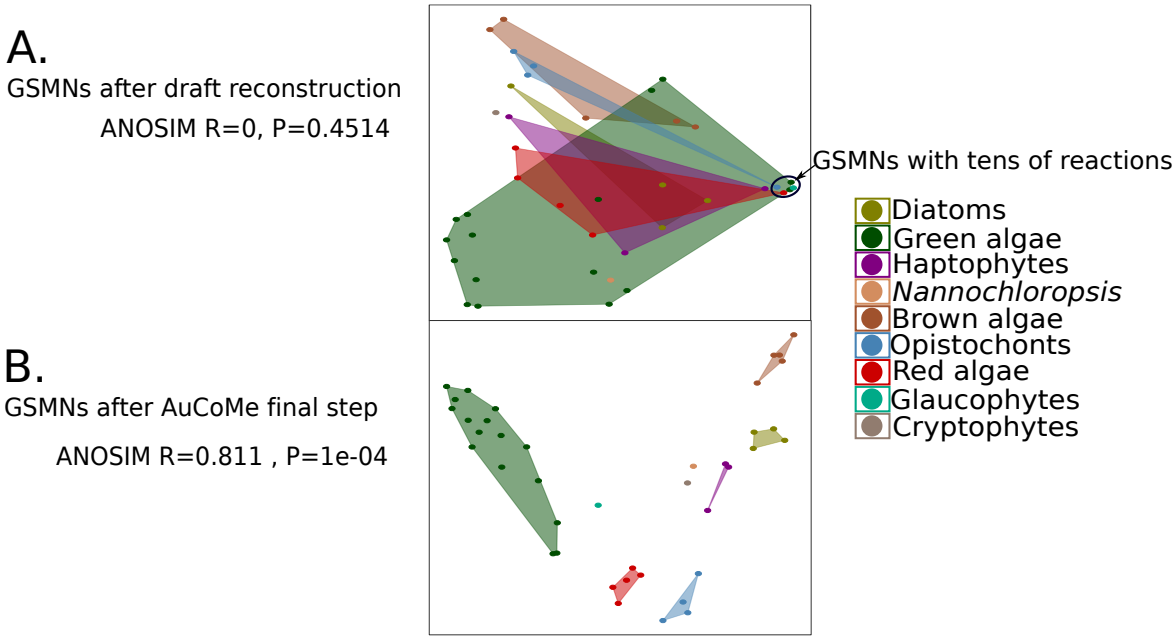



Figure 3.8 – **MDS showing the metabolic distance between GSMNs of the different groups of organisms in the algal dataset. Subplot A represents the GSMN distances after the draft reconstruction step, and subplot B shows the GSMN distances after the final step of AuCoMe.** The metabolic distances were computed using a presence-absence matrix of reactions among species in the algal dataset. ANOSIM values indicate the difference in variance between different groups. Close to 0, it is not possible to differentiate the group and close to 1, it is possible to differentiate the group. MDS and ANOSIM were computed using the vegan package (Oksanen et al. 2020).

**Subsection summary**

In summary, AuCoMe reconstructed GSMNs from heterogeneously annotated genomes which display evolutionary consistent reaction contents. The GSMN clustered according to the presence/absence of reactions regrouped into clusters corresponding to the major phylogenetic groups.

### 3.4.2 Comparison between algae phylogeny and algae metabolism

A tanglegram was created (Fig. 3.9) to compare the clustering of GSMNs to the known organisms' phylogeny. The tree on the left was made by Gabriel Markov (Station Biologique de Roscoff, Roscoff) using clades compiled from the literature (Strassert et al. 2021). The tree on the right was made using pvclust (Suzuki et al. 2006). To create the dendrogram, pvclust used a Jaccard distance on a matrix containing the presence or absence of reactions in the GSMNs obtained after the AuCoMe final step. The GSMNs generated using the complete AuCoMe pipeline were broadly consistent with the reference species phylogeny (Fig. 3.9). As illustrated in the previous section, the main phylogenetic groups were correctly clustered together. There were only three higher-order inconsistencies. It concerned the position of *Guillardia theta*, which is controversial (Strassert et al. 2021), the position of *Cyanidiophora paradoxa*, for which the genome version deposited in Genbank was lacking annotations (Price et al. 2012), and the position of *Nannochloropsis gaditana* which was the only representative of eustigmatophycean stramenopiles. The two other stramenopile groups (diatoms and brown algae) were represented by multiple species which likely minimises errors linked with single genome peculiarities. There were also minor inconsistencies in intra-group relationships in green algae, diatoms, brown algae, and opisthokonts.

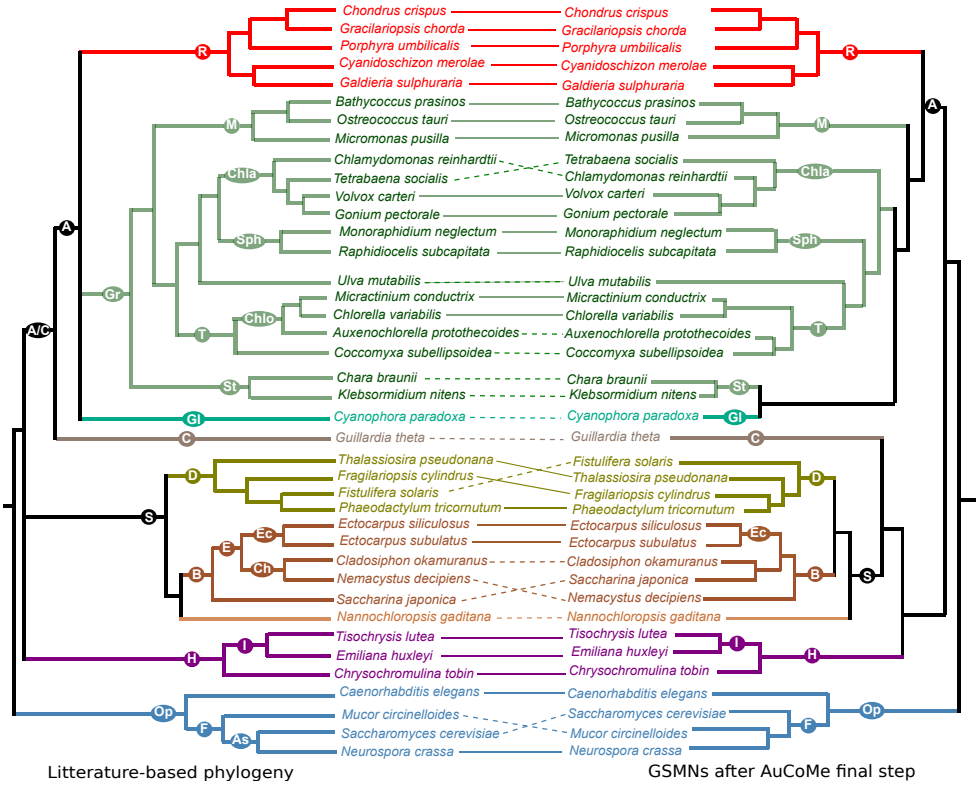



Figure 3.9 – **Tanglegram comparing phylogeny (left) and dendrogram of GSMNs reconstructed by AuCoMe (right)**. Tanglegram evaluating the taxonomic consistency of AuCoMe dendrogram based on metabolic distances created with the pvclust package (Suzuki et al. 2006) using the Jaccard distance (right side) in comparison with reference phylogeny (left side), compiled from literature (Strassert et al. 2021). Full lines join species for which the position in the AuCoMe dendrogram is consistent with the reference phylogeny. Dotted lines join species for which the metabolic dendrogram and the reference phylogeny diverge. A/C: Archeplastids/Cryptophytes, A: Archeplastids, R: Rodophytes, Gr: Green algae, M: Mamiellales, Chla: Chlamydomonadales, Sph: Sphaeropleales, T: Trebouxiophyceae, Chlo: Chlorellaceae, St: Streptophytes, Gl: Glaucophytes, C: Cryptophytes, H: Haptophytes, I: Isochrysida, D: Diatoms, S: Stramenopiles, B: Brown algae, E: Ectocarpales, Ec: Ectocarpaceae, Ch: Chordariaceae, Op: Opistochonts, F: Fungi, As: Ascomycetes.

A first illustration of the efficiency of AuCoMe was the *de novo* reannotation of the conserved enzyme-encoding genes from the glaucophyte *Cyanophora paradoxa*. To reconstruct this GSMN, we used the initially published genome sequence, which contained only two functionally-annotated genes (Price et al. 2012). The final GSMN created by AuCoMe enabled us to retrieve 2,664 reactions, a number within the same range as the

other species from the dataset. Accordingly, *C. paradoxa* branched at the basis of the dendrogram after the draft reconstruction step, whereas it clustered with the archeplastids after AuCoMe final step. Even if the grouping of *C. paradoxa* with the streptophytes *Chara brauni* and *Klebsormidium nitens* does not reflect the phylogenetic relationships, this shows that AuCoMe provides a reasonable proxy for handling almost fully unannotated genome sequences.

The consistency between the metabolic dendrogram computed from AuCoMe final step and the reference species phylogeny agrees with the literature. Indeed, numerous studies have compared GSMNs by computing a metabolic distance and clustering them into a dendrogram. These experiments allowed to cluster organisms into groups close to the ones known by phylogenetic analysis, but the position of species inside these groups were often different from the one of the phylogenetic groups (Vieira et al. 2011; Bauer et al. 2015; Prigent et al. 2018; Schulz et al. 2020). This difference could be explained by the large amount of the metabolism shared by most organisms. But for some specialised metabolic pathways, potential variations could occur. This chapter was focused on the analysis of genomics data, but combining other data (such as metabolomics) could help to explore this metabolic diversity, as we have shown in Chapter 2.

**Section summary**

We examined the impact of the GSMN homogenisation made by AuCoMe by clustering them according to their reaction contents. We have shown similarities between the metabolic clusters and the known phylogeny major groups. Inconsistencies in the metabolic dendrogram for some species might be investigated in further detail.

## 3.5 Conclusion

### Contribution

We have developed a method to compare GSMNs by homogenising their contents. This homogenisation is performed through orthology propagation and structural annotation verification. These methods allowed for the homogenisation of GSMNs according to the genome annotations so that comparable GSMNs can be reconstructed and compared. It

has been validated on a degraded dataset of *E. coli* with internal and external validations showing the quantity and quality of reactions retrieved by AuCoMe. Then we used it on an algal dataset to explore the relationship between phylogeny and distances between the inferred metabolisms. This comparison indicated an overall consistency between the phylogenetic groups and the clustered metabolisms, which was confirmed by the literature. We also explored the metabolism of algae and the position of organisms of interest.

This research aimed to explore the possibility of studying the metabolism from publicly available genomes. Especially to identify the impact of the heterogeneity in the genome annotations and how to solve it. Especially to determine which among the structural or functional annotations was the more impacting in such comparison. We have seen that the orthology propagation step seemed sufficient to retrieve most of the missing reactions resulting from the heterogeneity in original genome annotations. Thus, most issues in heterogeneity seemed to be associated with problems in functional annotations.

But for two eukaryotes genomes (*Saccharomyces kudriavzevii* and *Ectocarpus subulatus*), the structural verification allowed to retrieve probable reactions (respectively 209 and 86). This result implied that for some genomes, it could be helpful to handle the possibility of missing gene predictions. It is especially true for eukaryotes as the assembly of their genomes is more complex. For example, the NGS technologies have difficulties sequencing repeats regions, multicopy genes and other complex regions (Peona et al. 2021). So this advocates for caution when comparing the metabolism of eukaryotes if their genomes come from public databases.

## Limits and improvements

The structural annotation step could be improved: the annotation of pseudogenes in *Shigella* species would have been avoided by considering the annotations as pseudogenes available for the identified loci.

Running AuCoMe on the bacterial dataset highlighted the impact of a single highly-annotated genome on metabolic inference. This dataset included the reference genome of the *E. coli* K–12 MG1655 strain, which is better annotated than most of the other genomes considered. Consequently, a certain number of reactions initially propagated by orthology from the *E. coli* K–12 MG1655 genome to others were discarded by the AuCoMe filter due to a lack of support, following the rationale that a single genome supporting an annotation propagation is not robust enough. Reasoning on ortholog clusters, the filter implies that several congruent genome sources are mandatory to achieve annotation propagation

confidently. While the relevance of the filter was demonstrated on the algal dataset by avoiding the propagation of annotations related to photosynthesis to non-photosynthetic organisms, it may be too stringent in some applications and lead to discarding relevant reactions (such as the ones with the *E. coli* K–12 MG1655 strain).

The difference observed in the tanglegram 3.9 between phylogeny, and metabolic distances could be explored. One possibility could be to look at different similarity measures for the clustering. The Jaccard distance has been used in this work, but other measures could be used. For example, let's consider that an absence of reactions in two organisms could be considered to be a similarity (to represent the loss of a function). Then measures such as the Simple Matching Coefficient could be envisaged. Indeed this measure could count the missing reactions between organisms as a similarity (in a hypothesis of specific metabolic loss).

Another limitation is the use of a metabolic database. In this work we used MetaCyc (Caspi et al. 2020) as a reference metabolic database. But our results will also be limited by the knowledge present in the database. Despite a significant amount of work made on the manual curation of the database, it represents only what is known at a specific time.

## Perspectives

With the reconstruction of tens of GSMNs for eukaryotes, it is possible to explore the metabolism of taxonomic groups. This exploration permits the study of the diversity of the metabolism at the genome scale by comparing GSMNs. Furthermore, it gives a foundation to identify differences among organisms. And by combining with other data (such as metabolomics data), it could be used to increase the known diversity of metabolism. Indeed, by connecting this approach with the PathModel approach developed in Chapter 2, it could be possible to find the metabolic pathways associated with specific metabolites in a family. Using these pathways, we could find possible alternative pathways for each organism of the taxonomic groups. And by performing a comparison, it should be possible to explore the evolution of metabolic pathways in these groups.

Furthermore, by reconstructing the metabolisms of akin organisms, it becomes possible to study the evolution of metabolism. Indeed, the apparition or loss of specific metabolic pathways can be analysed by comparing these pathways' presence or absence in the GSMNs. This is similar, at a lower scale, to the analysis of the ancestral metabolic network of bacteria (Xavier et al. 2021). By comparing thousands of genomes, the authors identified a set of protein families that could be contained in the last bacterial common

ancestor. With the method presented in this chapter, the sequence comparisons and the inferred GSMN could be used to determine the evolution of the metabolic pathways in the group.

Part II

# Inferring metabolic complementarity between taxonomic groups

---

This part explores the methods developed during my thesis in order to study the metabolic complementarity between organisms from an environmental sample. One of the very general motivation behind these kind of analyses is the understanding of the ecosystem. In Chapters 4 and 5, I present the application of the developed methods to a biogas reactor ecosystem, in a collaboration with Patrick Dabert (UR OPAALE, INRAE). The diversity of the organisms herein was estimated using 16S rRNA gene sequencing and the corresponding taxonomic affiliations.

The Chapter 4 will present EsMeCaTa, the method developed to estimate metabolic capabilities from such taxonomic affiliations. The main issues lies in the uncertainty related to the taxonomic affiliations and the knowledge biases over taxa.

In Chapter 5, I will present Metage2Metabo, a method to identify the metabolic complementary between the organisms of the samples and the minimal communities which can collectively achieve the production of metabolites of interest.

Chapter 4

# Estimating metabolic capabilities from taxonomic affiliations

In Chapters 2 and 3, I have presented methods to analyse the metabolism of organisms by inferring metabolic networks and pathways from genomes to study the metabolic diversity of organisms. But genomes are not always available, especially when working on data from environmental samples. Thus, we need to use other methods to study the metabolic diversity within an environmental sample.

This issue is present in metabarcoding, where gene markers will be sequenced from an environmental sample to identify the taxonomic groups present in the sample. One of the most widely used gene markers is the 16S rRNA gene, which identifies taxonomic affiliations associated with a 16S amplicon using taxonomic assignment methods. It is possible to find functional profiles using already developed tools from the 16S rRNA gene sequence. But these tools do not provide metabolic networks that can be used to study the metabolic complementarity between the sample organisms.

That's why I developed a new method called EsMeCaTa (Estimating Metabolic Capabilities from Taxonomic affiliations) to fulfil this need. Several bottlenecks are solved by this method. First, the taxonomic affiliations of wild organisms taken as input are analysed to determine which clade from this taxonomic affiliation provides enough proteomes to be interpreted. Second, complete proteomes from the clade are clustered into groups of shared proteins, determining which enzymes are shared and thus likely present in the wild organisms. As a third bottleneck, one needs to find the functional annotations associated with the groups of clustered proteins. These annotated protein clusters are provided as outputs. Then it is possible to reconstruct the metabolic networks associated with the taxon.

## 4.1 Predicting metabolism from metagenomics and taxonomic affiliations

### 4.1.1 Inferring functional profiles from gene markers

**Sequencing community from the environment.** To study community from the environment, multiple methods have been developed, such as metabarcoding or shotgun sequencing. The first method uses a barcode, a region associated with a gene presenting variations. These genes are called gene markers. One of the most known gene markers is the 16S ribosomal RNA gene used for bacterial community analysis. It is one of the first markers used to analyse environmental community diversity (Giovannoni et al. 1990). Other gene markers have been analysed to study different groups, such as the *ITS* for fungi (Schoch et al. 2012). 16S rRNA gene sequencing is widespread as this method is cost-effective. The gene marker sequenced is then aligned to gene sequences in databases to find the closest sequences and the group of related organisms (**taxon**) associated with the gene marker. This way, it is possible to associate a taxonomic group to the sequenced gene marker (its **taxonomic affiliation**). But the gene marker sequence can highly diverge from known genes and then is assigned to a taxon with high taxonomic ranks (e.g. unknown bacteria). This divergent assignation is the **uncertainty** in the taxonomic affiliation. Until recently, it allowed a resolution of the taxonomic group present in the sample up to the genus level (Yarza et al. 2014). This taxonomic resolution is improved by analysing the entire gene thanks to advances in long-read sequencing (Johnson et al. 2019; Curry et al. 2022), allowing a resolution at the species level.

The second method, shotgun sequencing, consists of sequencing the genomes in the environment. Multiple analyses can be performed but will require different costs. It is possible to produce Metagenome-Assembled Genomes (MAG), but this can be quite expensive for large-scale microbiota analysis as it needs to sequence a genome region many times. The goal is to have a high depth ('average number of times a particular nucleotide is represented in a collection of random raw sequences' (Sims et al. 2014)), which allows assembly of the sequences into genomes. Sequencing with a low depth (reading the same sequence only a few times) does not allow reconstructing the genome. Still, it can be enough to identify the taxonomic groups of the organisms in the sample. This method (called **Shallow shotgun sequencing**) can be used to infer taxonomic groups in the environment at a lesser cost (Hillmann et al. 2018).

But the choice of the methods will impact the metabolism modelling (Frioux et al. 2020b). Indeed, by sequencing the genomes, it is possible to find metabolic functions present in the sample. Whereas disposing of only taxonomic affiliations, it is only possible to estimate the present function using available knowledge. In the following chapter, we will focus on the less cost-expensive sequencing methods: metabarcoding and Shallow shotgun sequencing.

**Predicting functional profiles.** Once the gene markers have been sequenced; methods are commonly applied to characterise the diversity of organisms in environmental samples (especially for 16S rRNA gene sequencing). From these data, taxonomic affiliations, such as Operational Taxonomic Units (OTUs), can be used to estimate the potential environmental functions. Multiple methods have been developed to predict functional profiles from 16S rRNA gene sequencing.

Some of these methods uses a phylogenetic placement of the sequence according to a database reference, such as PICRUSt and PICRUSt2 (Langille et al. 2013; Douglas et al. 2020), Paprica (Bowman et al. 2015).

Other methods uses a sequence alignment to find the closest sequences such as Tax4Fun and Tax4Fun2 (Aßhauer et al. 2015; Wemheuer et al. 2020), piphillin (Iwai et al. 2016; Narayan et al. 2020). Web servers using this method have also been developed with MicFunPred (Mongad et al. 2021) and MG-RAST (Keegan et al. 2016).

Another group of methods uses the taxonomic affiliations from taxonomic assignment methods to analyse them. Panfp (Jun et al. 2015) analyses the taxonomic affiliations to find the closest genomes and perform comparative genomics. FAPROTAX (Louca et al. 2016) maps prokaryotic taxon to its database to infer functional profiles.

But there are multiple issues with these methods. First, most of these tools support only 16S rRNA sequences. Whereas other gene markers can be considered for metabarcoding (such as *rpob* gene (Ogier et al. 2019)). Apart from gene markers, taxonomic affiliations can be obtained from shallow whole genome sequencing data, but the association of function with these data remains uneasy.

Secondly, they do not predict metabolic networks used in metabolism modelling. Thus system biologists working on these types of data have tested multiple methods to obtain metabolic networks associated with these data.

### 4.1.2 Metabolism and metabolic networks from gene markers and their taxonomic affiliations

Several works have been performed to predict metabolic networks from metabarcoding.

For example, MACADAM, a database of metabolic networks of microbial taxonomic groups (Le Boulch et al. 2019), has been constructed using PGDB reconstructed with Pathway Tools.

Other attempts to infer metabolic networks from metabarcoding data used either the closest known genomes and reconstruct the associated GSMN or already reconstructed GSMNs.

The first approach retrieves the closest known genomes associated with the taxon. Among the methods in this approach, the first one retrieves genomes associated with the taxa, then extracts the corresponding genes (Cardona Uribe 2019). The genes are then annotated with PATRIC (Wattam et al. 2017) and find enzymatic reactions. Then to produce a functional network, the networks are gap-filled in Kbase (Arkin et al. 2018). Another one, MMinte (Mendes-Soares et al. 2016) uses BLAST to map 16S rRNA sequences to a genome. Then using the matched genome, a metabolic network is reconstructed using modelSeed. Then it takes all the inferred networks and studies them two by two to evaluate the growth of interacting pairs of networks (Henry et al. 2010). But these approaches have issues as the closest available genome can diverge from the genome of the wild organism. We have here the first issue with a **limit in the available knowledge**.

The second method relying on already reconstructed metabolic networks was used in MetGEMS (Patumcharoenpol et al. 2021). Taxonomic affiliations of human gut microbiomes are used to predict networks with MetGEMs using a reference database of 818 GSMNs from AGORA database (Magnúsdóttir et al. 2017) with their associated genomes. Other methods such as MIMOSA and MIMOSA2 (Noecker et al. 2016; Noecker et al. 2022) linked OTU to three databases (KEGG by using output from PICRUSt, AGORA (Magnúsdóttir et al. 2017) and embl_gems (Machado et al. 2018)). Then for each metabolite, the tool predicts a score according to the abundance of the OTU for the possibility of having a reaction producing or using this metabolite. Directly linking taxon to already reconstructed metabolic networks can present issues. The wild organism associated with the gene marker can exhibit different metabolism than known organisms in the taxon. Here is a second issue: how can we **estimate the metabolic capabilities** of a wild taxon? The content of a GSMN could be an over (or under) approximation of

the metabolic capabilities of the taxon.

These methods rely on already reconstructed GSMN or closest genomes and have limits. We have issues with the available knowledge: how can we estimate the function of a taxon with few closest known organisms? And we have issues with estimating the metabolic capabilities once we have found these known organisms. Because the wild organism is one organism among the ones present in the taxon, with the proteome information of the known organisms in the taxon, we can have insight into what metabolic function can be performed, but it will only be an estimation.

### 4.1.3 How to reconstruct TSMN (Taxonomic-Scale Metabolic Network) from taxonomic affiliations

We have seen in the previous subsection that multiple methods have been developed to explore the functions and the metabolism associated with taxonomic affiliations. But most of the tools rely only on 16S rRNA, thus limiting the possible sequencing analysis performed. This limitation is the first issue to handle; we need to develop a tool that can be as flexible as possible and not related to a specific analysis.

**Issue: input data flexibility**

There is a potential diversity in gene markers or other sequencing methods to be used. It can be limiting to rely only on one gene marker. Furthermore, other metagenomics methods produce data that could also be used. Thus the developed method needs to be flexible in its input.

When searching for the metabolic capabilities from a taxonomic affiliation, we have to face the **uncertainty** of the taxonomic affiliation. It is the divergence between the gene marker sequenced and the known genes, leading to ambiguous taxonomic affiliations (such as unknown genus). This ambiguity will impact the method, forcing it to use higher taxonomic ranks. But due to the **taxonomic diversity**, higher taxonomic ranks correspond to older common ancestors and more diverged and diverse organisms. Then more metabolic functions will be available, but they may be more divergent from the wild organism.

Furthermore, some analyses and methods have been proposed to estimate the metabolism from input gene markers, but they face several issues. First, some methods rely on the closest known genomes and can be impacted by knowledge bias.

**Issue: Knowledge bias**

Making an estimation for an organism according to a taxon is biased towards known organisms, with some taxon being highly documented and others less or even not documented. The knowledge bias also applies according to the rank of the taxon considered, with higher taxonomic rank being associated with higher diversity, known organisms are more diverged from each other, and also most likely from the wild organism to model.

And other methods rely on known GSMNs to estimate the metabolic function of an unknown organism. But this closest metabolism can be an over or under-approximation of the metabolic capabilities of the unknown organism. Furthermore, depending on the metabolic network reconstruction tools, it can add metabolic functions based on gap-filling, thus overestimating what the unknown organism can do.

**Issue Estimating metabolic capabilities**

The putative metabolism of a wild organism (belonging to a given taxon) can be estimated according to the known proteomes of its relatives (belonging to that taxon). From these relatives' proteomes, the estimation of the metabolic capabilities of the wild organism can be achieved in multiple ways. First, a conservative option could be to only look at the more conserved functions in the relative organisms. A second option could be to find all the metabolic functions performed in the relative organisms to have a broader estimation.

The method needs to estimate the potential capabilities associated with a taxon according to the data from the database. So there is another step that is required, to use the estimated annotations to reconstruct a draft metabolic network associated with a given taxonomic group, thus leading to a Taxonomic-Scale Metabolic network.

**Section summary**

Tools predicting functional profiles do not produce **metabolic networks**, are often **limited to a gene marker**. Furthermore, the existing methods to create metabolic networks rely on selecting the closest genome or GSMNs from databases. This lead to two issues, a **knowledge bias** as more studied species will have more knowledge available. And an issue with the **metabolic capabilities estimation** from these biased knowledge. The developed method must provide input flexibility, considering taxonomic affiliation rather than being dedicated to a given type of raw sequence data. It should be possible to handle knowledge bias over more or less documented taxa and allow metabolic capabilities estimation.

## 4.2 *Es*timating *Me*tabolic *Ca*pabilities from *Ta*xonomic affiliations (EsMeCaTa)

To answer these issues, I have developed EsMeCaTa, a method to estimate metabolic capabilities from a taxonomic affiliation.

EsMeCaTa takes as input a tabulated file containing two columns (example in Table 4.1. The first column is an identifier, and the second contains a taxonomic affiliation (starting with the highest taxonomic rank, such as kingdom, to the lowest taxonomic rank, such as species) as defined by the NCBI Taxonomy database (Schoch et al. 2020). The inputs are processed using pandas (McKinney 2010).

The outputs of the workflow are, for each taxon (1) fasta files of all proteomes selected by EsMeCaTa, (2) a fasta file of the shared proteins clustered by MMseqs2 from these proteomes, and (3) a tabulated file containing the functional annotations associated with its proteins (Gene Ontology Terms (GO), Enzyme Commission (EC)). Furthermore, EsMeCaTa creates output annotated files that can be used as input to Pathway Tools, thus allowing to reconstruct draft Taxonomic-Scale Metabolic Network (TSMN).

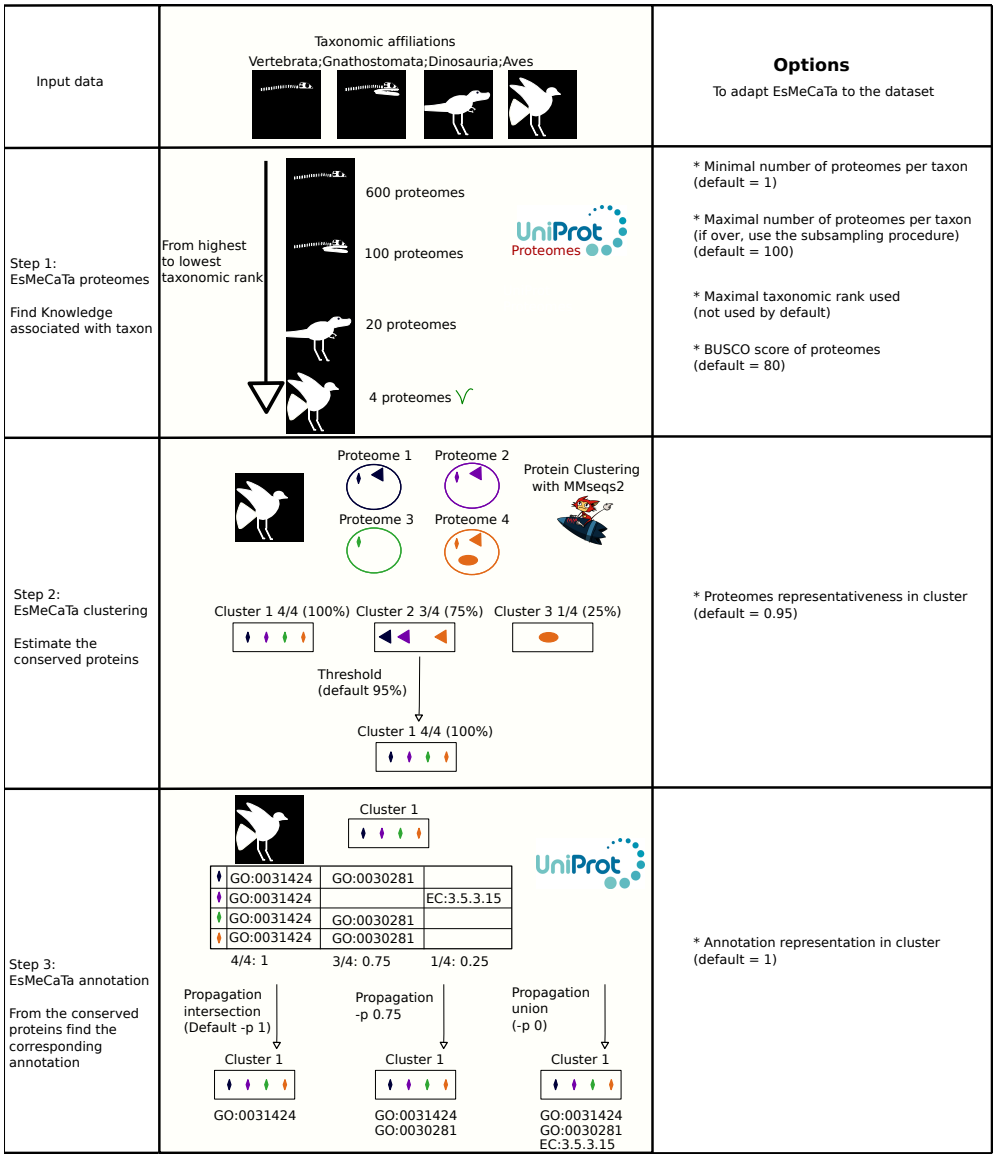



Figure 4.1 – EsMeCaTa workflow.
From a taxonomic affiliation, **esmecata proteomes** (step 1) searches for the lowest taxonomic rank associated with proteomes on UniProt. Then **esmecata clustering** (step 2) clusters the proteomes using MMseqs2. Then **esmecata annotation** (step 3) annotates the protein clusters by querying UniProt again.

To illustrate how the workflow works, EsMeCaTa was applied to the dataset presented in the example Table 4.1.

| observation_name | taxonomic_affiliation |
|---|---|
| Escherichia | cellular organisms;Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales;Enterobacteriaceae;Escherichia |
| Citrobacter | cellular organisms;Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales;Enterobacteriaceae;Citrobacter |
| Cronobacter | cellular organisms;Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales;Enterobacteriaceae;Cronobacter |
| Lelliottia | cellular organisms;Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales;Enterobacteriaceae;Lelliottia |
| Jejubacter | cellular organisms;Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales;Enterobacteriaceae;Jejubacter |
| Edaphovirga | cellular organisms;Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales;Enterobacteriaceae;Edaphovirga |
| Enterobacteriaceae | cellular organisms;Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales;Enterobacteriaceae |
| Enterobacterales | cellular organisms;Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales |
| Gammaproteobacteria | cellular organisms;Bacteria;Proteobacteria;Gammaproteobacteria |
| Plasmodium | cellular organisms;Eukaryota;Sar;Alveolata;Apicomplexa;Aconoidasida;Haemosporida;Plasmodiidae;Plasmodium |
| Leucocytozoon | cellular organisms;Eukaryota;Sar;Alveolata;Apicomplexa;Aconoidasida;Haemosporida;Leucocytozoidae;Leucocytozoon |
| Corallicola | cellular organisms;Eukaryota;Sar;Alveolata;Apicomplexa;Conoidasida;Corallicolida;Corallicolidae;Corallicola |
| Acavomonas | cellular organisms;Eukaryota;Sar;Alveolata;Colponemida;Acavomonidia;Acavomonas |

Table 4.1 – Example of input taxonomic affiliations provided to EsMeCaTa. The first column contains an observation name (for example, in metabarcoding data, it is often the OTU name, here, it corresponds to the lowest taxon name in the taxonomic affiliations). The second column contains the taxonomic affiliations (separated with ';' with the highest taxonomic rank at the left and the lowest taxonomic rank at the right).

The results are presented in the following subsections detailing each step of the method. They are also available in the Appendix Table 6.1.

### 4.2.1 Step 1: Retrieving proteomes from taxonomic affiliations

As we want to handle taxonomic affiliations from any gene marker and shallow shotgun sequencing, to have an **input data flexibility**, EsMeCaTa relies on taxonomic affiliations. Furthermore, this method can be used on a set of manually created taxonomic affiliations allowing the user to study taxonomic groups of interest.

A taxonomic affiliation indicates the evolutionary history of life of a taxon. Then it should be possible to identify the inherited functions that could be present in our wild organisms. By using the taxonomic affiliation as a query on a database, it is possible to find the set of knowledge associated with the taxon of interest. It is then possible to estimate the **knowledge bias** of the current taxon. EsMeCaTa can find the number of proteomes associated with the taxa described in the taxonomic affiliations by relying on UniProt. It can select a given number of proteomes associated with a taxon through different options. These proteomes will be used to estimate the capabilities of the wild organism behind the taxonomic affiliation.

#### Step description

The first step of EsMeCaTa focuses on parsing the taxonomic affiliations to find the proteomes associated with each taxon and then select the minimal taxonomic rank in the

taxonomic affiliations associated with at least a proteome (Figure 4.2). Multiple substeps perform this operation. The first converts taxon names in the taxonomic affiliations into taxon IDs. The second substep disambiguates the taxon IDs identified. The third substep applies the limit on the maximal taxonomic rank used (if the option has been selected, by default, it is not used). Then using the resulting taxon IDs, the fourth step queries UniProt to find the proteomes associated with each taxon ID. During the search of proteomes on UniProt, two options can be used. A first option specifies the **minimal number of proteomes per taxon**. Suppose there are fewer proteomes in the taxon than specified by the minimal number of proteomes option. In that case, the current taxon of the taxonomic affiliation is ignored, and the immediately higher taxonomic rank is considered. This operation is recursively achieved until a higher-ranking taxon with enough proteomes can be found. A second option regulates the **maximal number of proteomes per taxon**. A subsampling procedure is performed if there are more proteomes than the maximal number of proteomes per taxon option. These substeps are described in more detail below.

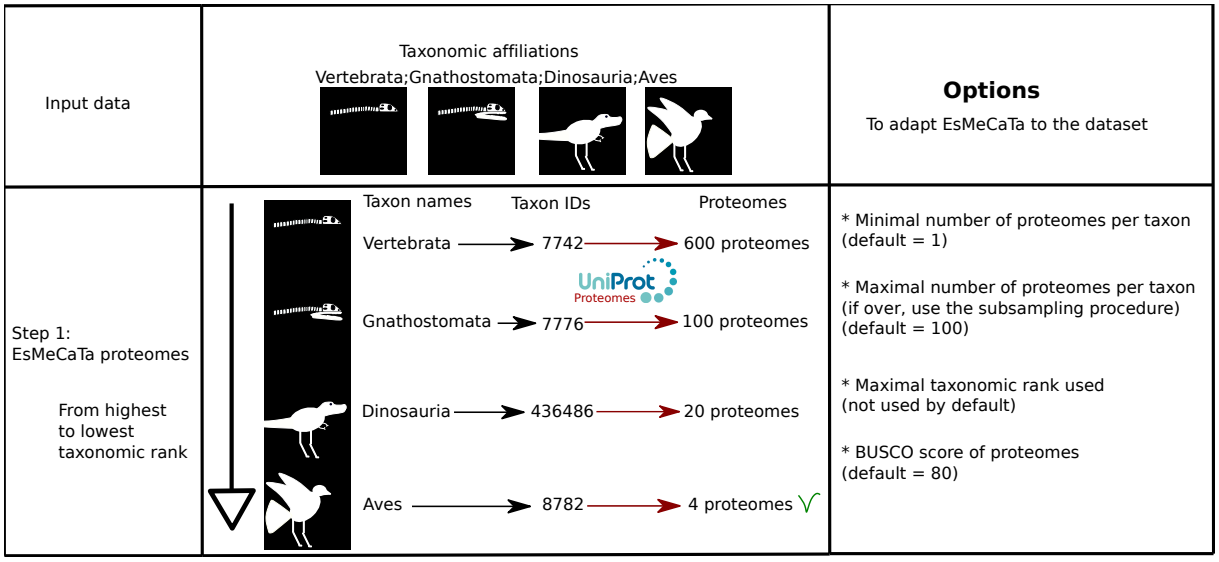



Figure 4.2 – EsMeCaTa annotation step workflow.

**Converting taxonomic affiliations.** The taxonomic affiliation is processed using the ete3 python package (Huerta-Cepas et al. 2016) to associate a taxon ID (from the NCBI taxonomy database) to each taxon from the affiliation.

**Taxon ID disambiguation.** A taxon name can be associated with more than one taxon ID (that we call ambiguous taxon). For example the taxon name *Yersinia* is associated with taxon IDs 444888 (*Yersinia* mantid) and taxon ID 629 (*Yersinia* bacteria). In this case, disambiguation will be performed by using the other taxon IDs present in the taxonomic affiliations and by comparing them to the lineage associated with the ambiguous taxon IDs.

If we take the example of the taxonomic affiliation 'Bacteria;Gammaproteobacteria;Yersinia', *Yersinia* will be first associated with two taxon IDs. Then EsMeCaTa will use ete3 to extract the lineage associated with these two taxon IDs:

— lineage of taxon ID 629 (*Yersinia* Bacteria): [1, 131567, 2, 1224, 1236, 91347, 1903411, 629]

— lineage of taxon ID 444888 (*Yersinia* mantid): [1, 131567, 2759, 33154, 33208, 6072, 33213, 33317, 1206794, 88770, 6656, 197563, 197562, 6960, 50557, 85512, 7496, 33340, 33341, 6970, 7504, 7505, 267071, 444888]

These lineages are then compared to the taxon IDs corresponding to the rest of the taxonomic affiliation provided as input. In the *Yersinia* example, the rest of the input is 'Bacteria; Gammaproteobacteria', which corresponds to the IDs [2, 1236]. The ambiguous taxon (here *Yersinia*) is ignored in this comparison. Then the lineages are compared to the IDs of the input taxonomic affiliation, looking for the best matching lineage. In the example, lineages for ambiguous *Yersinia* IDs 629 and 444888 are compared to the input affiliation. ID 629 is retained because its lineage contains IDs 2 and 1236 from the input taxonomic affiliation. Whereas ID 444888 lineage does not include IDs 2 and 1236.

**Limit highest taxonomic rank used.** The option 'rank limit' excludes the higher taxonomic rank at this step. For example, suppose the maximal taxonomic rank is set to 'family'. In that case, only the taxon with a taxonomic rank equal or inferior to the genus will be selected, the other will be ignored. This selection fits the following rationales. Higher taxonomic ranks are associated with more available proteomes to analyze, increasing calculations at each method step. Moreover, they are associated with a higher organism diversity to interpret, increasing the difficulty of inferring a consistent set of functions.

**Querying Uniprot to find proteomes.** Using these taxon IDs, queries against the UniProt Proteomes database find the lowest-ranking taxon for which there is at least one

proteome in the database. As the taxonomy is hierarchically classified in UniProt, when searching for a taxon in the proteome database, the query will return all the proteomes associated with an organism which is classified in this taxon.

Furthermore, to benefit from the work on UniProt and alleviate knowledge bias, EsMeCata uses the classification made by Uniprot on the proteomes between the reference and non-reference proteomes (see definition in Appendix 6.2.1). First, EsMeCaTa will prioritize the reference proteomes, but if none are available for the taxon, it will select the non-reference proteomes. A second filter is also applied by selecting only the proteomes with at least a BUSCO score of 80% to select proteomes with enough completeness. And a third filter excludes the proteomes tagged by Uniprot as being redundant (Appendix definition 6.2.2) and excluded (Appendix definition 6.2.3). For example, in figure 4.2, the lowest taxonomic rank with proteomes is 'Aves' with four proteomes.

Two options are applied during this substep, the minimal or maximal number of proteomes per taxon. When searching for proteomes, it is possible to specify a minimum number of proteomes associated with a taxon (by default, it is 1). For example, if five is given to this option, EsMeCaTa will search for a taxon with at least five proteomes. According to the example in the 4.2, this will not consider the taxon 'Aves' as it is only associated with four proteomes and selects the taxon Dinosauria, associated with 20 proteomes. For the maximal number of proteomes per taxon, this option is associated with the subsampling procedure presented below.

**Subsampling proteomes.** If the number of proteomes found is greater than the maximal number of proteomes per taxon (100 by default), only a subsample is downloaded. The subsampling procedure aims to decrease the number of proteomes used while keeping the taxonomic diversity presented in the taxon.

To achieve this, EsMeCaTa will look at the taxon selected. Then it will create a taxonomic tree using this taxon as a root and the organism ID associated with each proteome as a leaf. In this tree, EsMeCaTa will find all the direct child taxa of the root. For each of these child taxa, it will compute the number of proteomes contained in these taxa. This number of proteomes is divided by the total number of proteomes associated with the root. The resulting number corresponds to the ratio of proteomes brought by the child taxon to the root taxon (called *child taxon contribution ratio*). Using the threshold given by the user (by default, it is 100), EsMeCaTa will apply this ratio to the threshold to select a number of proteomes for each child taxon according to their computed ratio.

Let's take an example: taxon 1 is associated with 200 proteomes (which is above the default threshold of 100). When looking at the child taxa of taxon 1 we have 4 direct child taxa (taxon 2, taxon 3, taxon 4 and taxon 5). Taxon 2 is associated with 100 proteomes, taxon 3 with 50 proteomes, taxon 4 with 40 proteomes and taxon 5 with 10 proteomes. By computing the *child taxon contribution ratio*, we have 0.5 for taxon 2 (100/200), 0.25 for taxon 3 (50/200), 0.2 for taxon 4 (50/200) and 0.05 for taxon 5 (10/200). By default, we will select 100 proteomes so that each ratio will be multiplied by 100. This results in a random subsampling of 50 proteomes for taxon 2 (100*0.5), 25 proteomes for taxon 3 (100*0.25), 20 proteomes for taxon 4 (100*0.2) and 5 proteomes for taxon 5 (100*0.05). This procedure ensures a proteome sampling consistent with the taxonomic diversity.

After these steps, the selected proteomes are downloaded.

### Application to the example dataset

EsMeCata first step was applied to the dataset of Table 4.1. The results are shown in Table 4.2. The column 'Input' shows the lowest taxon names in the taxonomic affiliations and the corresponding taxon ranks. This taxa selection aims to illustrate the uncertainty in taxonomic assignments, as taxon assigned to the class, order or family ranks can be considered input. Then the column 'Taxa selected by EsMeCaTa' presents the taxon selected by EsMeCaTa according to the available proteomes. And the column 'Proteomes selection (Busco $\geq$ 0.8)' indicates the total number of proteomes in UniProt ('UniProt total'), the number of reference proteomes in UniProt ('UniProt references') and the number of proteomes selected by EsMeCaTa ('EsMeCaTa proteomes').

| **Input** | | **Taxa selected by EsMeCaTa** | | **Proteomes selection (Busco ≥ 0.8)** | | |
|---|---|---|---|---|---|---|
| Lowest taxon name | Taxon rank | Taxon rank used | Taxon name used | UniProt total | UniProt references | EsMeCaTa proteomes |
| *Escherichia* | Genus | Genus | *Escherichia* | 1,506 | 3 | 3 |
| *Citrobacter* | Genus | Genus | *Citrobacter* | 138 | 2 | 2 |
| *Cronobacter* | Genus | Genus | *Cronobacter* | 15 | 0 | 15 |
| *Lelliottia* | Genus | Genus | *Lelliottia* | 5 | 0 | 5 |
| *Jejubacter* | Genus | Genus | *Jejubacter* | 1 | 1 | 1 |
| *Edaphovirga* | Genus | **Family** | Enterobacteriaceae | 2,435 | 42 | 42 |
| Enterobacteriaceae | Family | Family | Enterobacteriaceae | 2,435 | 42 | 42 |
| Enterobacterales | Order | Order | Enterobacterales | 3,028 | 129 | 96 |
| Gammaproteobacteria | Class | Class | Gammaproteobacteria | 8,271 | 911 | 96 |
| *Plasmodium* | Genus | Genus | *Plasmodium* | 67 | 17 | 17 |
| *Leucocytozoon* | Genus | **Order** | Haemosporida | 68 | 18 | 18 |
| *Corallicola* | Genus | **Class** | Conoidasida | 30 | 10 | 10 |
| *Acavomonas* | Genus | **Clade** | Alveolata | 124 | 48 | 48 |

Table 4.2 – Proteomes found by the first step of EsMeCaTa on the example data of Table 4.1.

The selected taxonomic rank can be seen in the column 'Taxa selected by EsMeCaTa' (Table 4.2). For 9 taxonomic affiliations, the selected taxon was the lowest possible taxon, but for 4 taxonomic affiliations (in bold), EsMeCaTa had to select a higher taxonomic rank as the lowest was not associated with a proteome respecting the filter criteria. This difference illustrates the method behaviour concerning knowledge bias. Given an input taxon without sufficient knowledge, it considers a taxon of higher rank instead. Using higher rank taxon will increase the prediction uncertainty as more proteomes with more diversity will have to be considered (see Proteomes selection for the Enterobacteriaceae, Enterobacterales, and Gammaproteobacteria selected taxa)

The proteomes available per taxa are presented in the column 'Proteomes selection (Busco ≥ 0.8)' (Table 4.2). For some taxa, there were reference proteomes that could be used and were selected (such as *Escherichia*). But for other taxa (such as *Cronobacter*), no reference proteomes were available, so the downloaded proteomes corresponded to non-reference proteomes. For 2 taxa (Enterobacterales and Gammaproteobacteria), there were more than 100 proteomes, so the subsampling procedure was applied to select 96 proteomes. This result illustrates another outcome of the knowledge bias, genome quality and quantity differ even considering genus as fixed taxonomic rank. While *Escherichia* are the most documented genus, the *Edaphovirga* genus has no available proteome, and modelling its metabolism will require considering more diverse proteomes from the family Enterobacteriaceae. In other words, a taxon with less available knowledge will provide much more uncertain inferences than a well-documented one.

**Subsection summary**

In this subsection, I have presented how EsMeCaTa retrieves proteomes by parsing the **taxonomic affiliations** and performing knowledge engineering on **UniProt** to associate proteomes with the taxon. This selection allowed us to find a **set of proteomes** associated with a taxon that will be used in the next step.

### 4.2.2 Step 2: Clustering proteins from the proteomes

To handle the issue with the **estimation of metabolic capabilities**, EsMeCaTa will apply a clustering method on the proteomes. Then homologous protein clusters can be sorted according to a threshold specifying the number of proteomes sharing the proteins from the cluster. The threshold could be used to keep only the most conservative proteins (proteins that are shared by most of the proteomes).

#### Step description

In the previous step 4.2.1, EsMeCaTa selects a taxon among the taxonomic affiliations and retrieves the associated proteomes. Step 2 takes these proteomes as input and will use a clustering method to find homologous protein clusters and then estimates how these proteins are shared among the proteomes of the taxon, denoted as the representativeness of the proteomes in the protein cluster. (Figure 4.3). The proteome representativeness of each cluster indicates the proportion of the proteomes with at least one protein represented in the cluster. This ratio provides a threshold for selecting only a subset of supposedly well-conserved protein clusters or extracting a larger set of protein clusters. Thus proteome representativeness close to 1 indicates a cluster in which proteins are shared in almost all the known proteomes and are thus likely to be shared in any wild organism of the taxon.

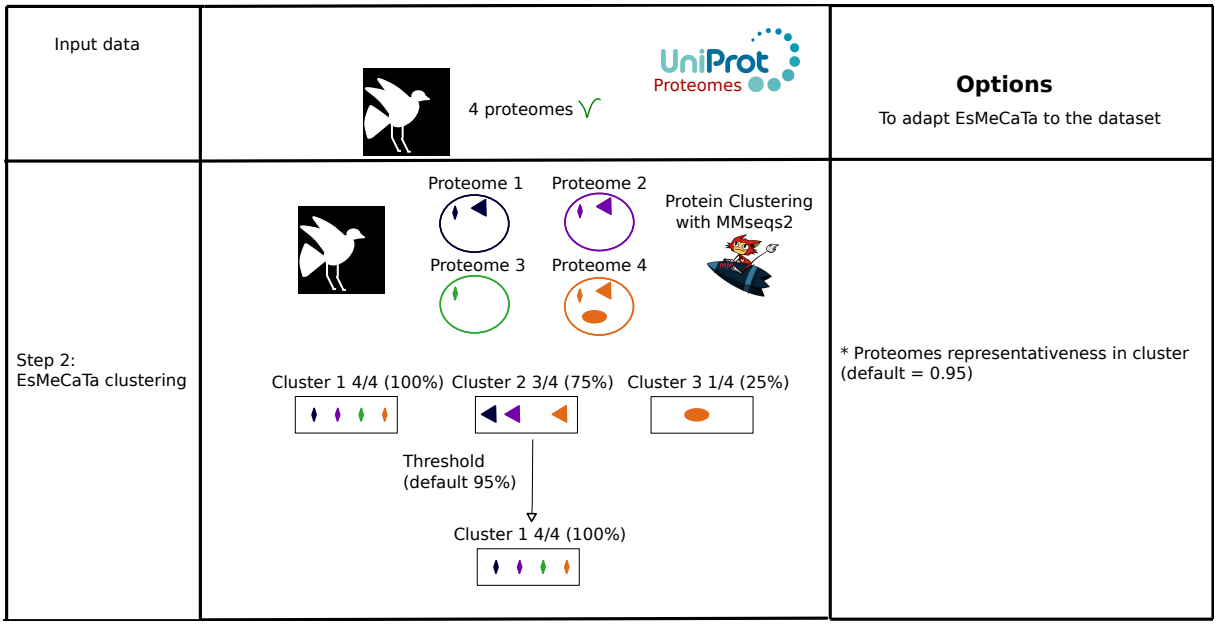



Figure 4.3 – EsMeCaTa clustering step.

**Clustering methods.** The workflow's second part aims to identify proteins shared by proteomes associated with a taxon. With the downloaded proteomes, EsMeCaTa performs protein clustering using MMseqs2 (Steinegger et al. 2017) and selects clusters such that proteins are shared by at least $Pr\%$ of the proteomes (this ratio is called the *proteomes representativeness*). MMseqs2 was selected as it is one of the fastest tools to align huge protein sequence sets.

The clustering uses as options a minimal sequence identity of 30 % and a coverage of 80%. The 30% of minimal sequence identity was chosen to search for homologs even if this threshold can miss known homologs (Pearson 2013). For each cluster of proteins, a consensus sequence is created.

**Filtering protein clusters according to proteome representativeness.** With the previous step, we obtained numerous protein clusters. This step aims to filter these clusters according to the proteome represented in each cluster. It is possible because we know for each protein inside a cluster from which proteome it was extracted. Thus we can know how many of the proteomes are represented in a given protein cluster. And by using this, we can estimate if the protein cluster is present in all the proteomes or only one. To represent this information, we computed the proteome representativeness $Pr$, which is the ratio between the number of proteomes represented in the protein clusters and the

total number of proteomes found by EsMeCaTa for the taxon.

For example, in Figure 4.3, the ratio for cluster 1 is 1 because it contains one protein from each of the 4 proteomes associated with the taxon. For cluster 2, the ratio of proteome representativeness is at 0.75 (as only 3 of the 4 proteomes are represented) and is 0.25 for cluster 3 (only 1 proteome is represented on 4). This ratio allows us to filter the protein clusters according to the number of proteomes represented. Especially we can use a clustering threshold that will select all the protein clusters with a ratio superior to or equal to it. In Figure 4.3, we used the default threshold at $Pr = 0.95$ on the 3 clusters; this led to the selection of cluster 1 as it is the only higher than 0.95.

If the threshold is at $Pr = 0$, it corresponds to the case where all protein clusters are selected (called 'Pan-proteome', abbreviated Pan-P, in reference to pan-genome). A second threshold at $Pr = 0.95$ retains only the clusters containing a protein originating from at least 95% of the proteomes (called 'Soft-core proteome', abbreviated Soft-P). The third threshold at $Pr = 0.5$ retains cluster containing proteins occurring in at least 50% of the proteomes (called 'Shell-core proteome', abbreviated Shell-P). For each protein cluster, EsMeCaTa selects the representative protein (first sequence in the alignment made by MMseqs2) to represent the cluster. The selected sequences of the representative proteins and the consensus sequences are stored in a fasta file using biopython (Cock et al. 2009).

This ratio is computed for each protein cluster. We represent the number of protein clusters associated with these ratios for the data of Table 4.1 in Figure 4.4.

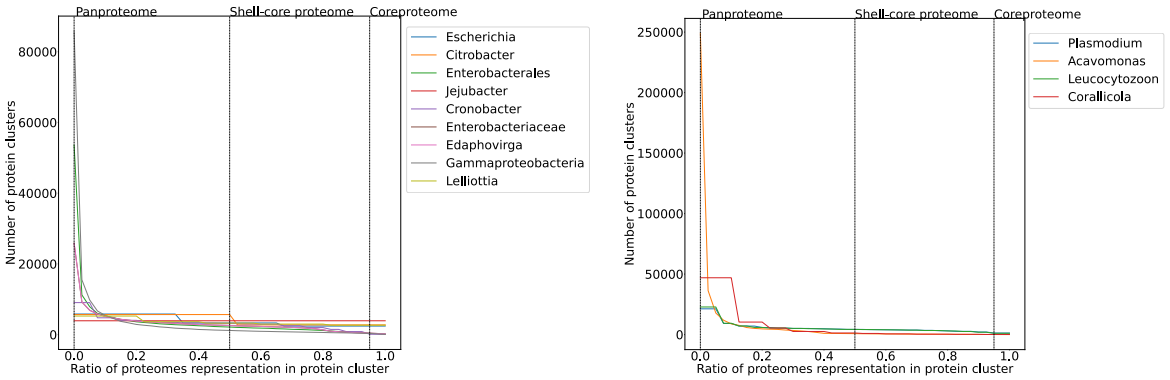



Figure 4.4 – Number of protein clusters according to proteome representativeness (ratio of proteomes representation in protein cluster) for Bacteria (left) and eukaryotes (right). The number of protein clusters is additive as we get closer to 0; this means that all the protein clusters at 0.95 are also contained for the same taxon at 0. For example, for the Shell-core proteome, we have a number of protein clusters having at least a ratio of 0.5, meaning at least 50% of the input proteomes are represented in the protein cluster.

### Application to the example dataset

The clustering step of EsMeCaTa was applied to the proteomes found in Table 4.2 (column 'Proteomes selection'). The result can be seen in Table 4.3. The column 'Taxa selected by EsMeCaTa' is the same as in the previous tables. The column 'Protein clusters (MMseqs2)' shows the number of protein clusters kept according to the proteome representativeness ratio (0% for Pan-Proteome, 95% for Soft-core Proteome and 50% for Shell-Core proteome).

| **Taxa selected by EsMeCaTa** | | **Proteomes selection** | **Protein clusters (MMseqs2)** | | |
|---|---|---|---|---|---|
| Taxon rank used | Taxon name used | EsMeCaTa proteomes | Pan-P | Soft-P | Shell-P |
| Genus | *Escherichia* | 3 | 5,821 | 2,421 | 3,298 |
| Genus | *Citrobacter* | 2 | 5,674 | 2,753 | 5,674 |
| Genus | *Cronobacter* | 15 | 9,057 | 101 | 3,128 |
| Genus | *Lelliottia* | 5 | 5,252 | 2,651 | 3,245 |
| Genus | *Jejubacter* | 1 | 3,915 | 3,915 | 3,915 |
| **Family** | Enterobacteriaceae | 42 | 25,822 | 415 | 2,581 |
| Family | Enterobacteriaceae | 42 | 25,822 | 415 | 2,581 |
| Order | Enterobacterales | 96 | 53,617 | 375 | 2,145 |
| Class | Gammaproteobacteria | 96 | 85,797 | 329 | 1,183 |
| Genus | *Plasmodium* | 17 | 21,287 | 1,276 | 4,263 |
| **Order** | Haemosporida | 18 | 22,813 | 1,076 | 4,313 |
| **Class** | Conoidasida | 10 | 46,959 | 76 | 1,326 |
| **Clade** | Alveolata | 48 | 248,878 | 50 | 785 |

Table 4.3 – Protein clusters selected according to the proteome representativeness 0 (Pan-P) 0.5 (Shell-P) and 0.95 (Soft-P), and using the proteomes found in the previous step 4.2.1.

In the column 'Protein clusters (MMseqs2)' of Table 4.3, we observe that in general, the size of the Pan-P increases with the number of selected proteomes, while the size of the Soft-P decreases. The size of the Shell-P appears to be much more stable. In Figure 4.4, we can see that the number of protein clusters kept increasing as the clustering threshold decreased for both the Bacteria and eukaryotes taxa.

This result is consistent with the definitions of Pan and Soft-Core proteomes. However, note that the low numbers of clusters selected in the Soft-Core proteome result from the difficulty of finding shared proteins as the number of proteomes increases. Because with more proteomes, it is possible that some of them lost some of the proteins during evolution, or there is also a risk of increasing alignment errors. The high number of selected clusters for the Pan proteome might reflect the diversity of the protein content within the whole taxon, presumably resulting from specific adaptations and the misalignment of highly diverged sequences.

**Subsection summary**

In this subsection, for each taxonomic affiliation, EsMeCaTa clustered the proteins from the selected proteomes using **MMseqs2**. Then each cluster is analysed to compute the **proteome representativeness ratio** $Pr$ showing how many proteomes are represented in each cluster compared to the total number of proteomes. These ratios are used to estimate the protein clusters representative of the taxon as they are shared in at least $Pr\%$ of the selected proteomes.

### 4.2.3 Step 3: Associating annotation to protein clusters

EsMeCaTa has identified the proteins shared in the taxon with the protein conservation threshold. Then it could be possible to find the functions associated with the protein clusters. The combination of these functions leads to an estimation of the metabolic potential of the wild organism.

#### Step description

The final step of the workflow annotates the protein clusters by querying the UniProt database (GO, EC). In this step, annotations (protein name, gene name, Gene Ontology terms, Enzyme Commission number, InterPro domain and Rhea reaction ID) associated with all the proteins inside the selected proteomes are extracted from UniProt. This extraction is represented in Figure 4.5 by the GO and EC associated with each protein from cluster 1.

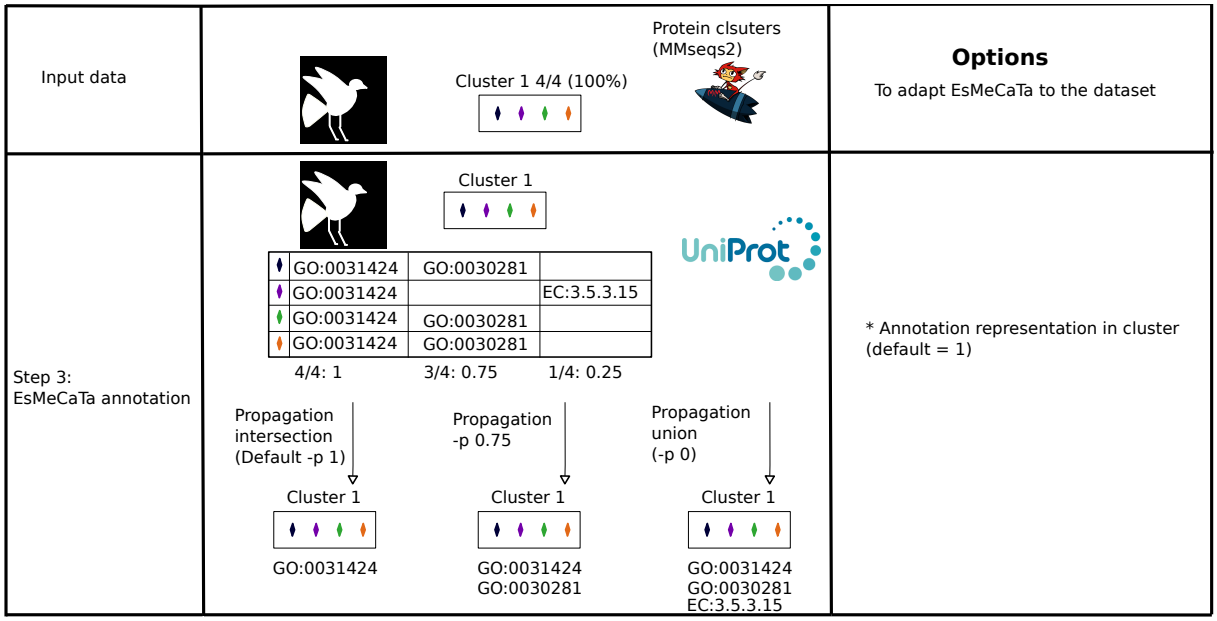



Figure 4.5 – EsMeCaTa annotation step.

Using these annotations, EsMeCaTa will choose which annotations to propagate to a given cluster according to a propagation threshold. An annotation is associated with a cluster if the ratio of proteins associated with this annotation divided by the number of proteins in the cluster is greater or equal to the propagation threshold ($Pt\%$). For example, the GO Term GO:0031424 is associated with 4 proteins from cluster 1 (all the proteins of this cluster), so its ratio is 1. The GO Term GO:00302841 is assigned to 3 proteins, so its ratio is 3/4, 0.75. And finally, the EC number EC:3.5.3.15 is assigned to 1 protein on the 4 of cluster 1, so its ratio is 0.25.

By default, the propagation threshold is at 1, meaning an annotation is kept if it occurs in all the cluster's proteins. This filtering is a very conservative threshold, as any missing functional annotation among the protein clusters will prevent the whole cluster from being annotated. The propagation threshold can be set to 0, meaning that it will use annotations from any proteins of the cluster (union of the annotations of a protein cluster).

### Application to the example dataset

The protein clusters found in the previous step (Table 4.3) have been annotated by EsMeCaTa third step and the result can be seen in Table 4.4 using the default propagation

filter of 1. The column 'Functional annotation of clusters' presents the number of unique GO Terms and EC numbers found for each taxonomic affiliation.

| **Taxa selected by EsMeCaTa** | | **Functional annotation of clusters** | | | | | |
|---|---|---|---|---|---|---|---|
| Taxon rank used | Taxon name used | Pan-P | | Soft-P | | Shell-P | |
| | | GO | EC | GO | EC | GO | EC |
| Genus | *Escherichia* | 2,183 | 866 | 1,661 | 679 | 1,906 | 792 |
| Genus | *Citrobacter* | 2,013 | 772 | 1,835 | 708 | 2,013 | 772 |
| Genus | *Cronobacter* | 970 | 677 | 0 | 12 | 600 | 603 |
| Genus | *Lelliottia* | 1,993 | 756 | 1,784 | 687 | 1,884 | 718 |
| Genus | *Jejubacter* | 1,983 | 837 | 1,983 | 837 | 1,983 | 837 |
| **Family** | Enterobacteriaceae | 2,253 | 867 | 514 | 193 | 1,560 | 595 |
| Family | Enterobacteriaceae | 2,253 | 867 | 514 | 193 | 1,560 | 595 |
| Order | Enterobacterales | 2,475 | 1,010 | 487 | 175 | 1,383 | 512 |
| Class | Gammaproteobacteria | 2,650 | 1,040 | 387 | 123 | 924 | 327 |
| Genus | *Plasmodium* | 1,305 | 225 | 611 | 104 | 1,103 | 200 |
| **Order** | Haemosporida | 1,327 | 259 | 546 | 95 | 1,090 | 199 |
| **Class** | Conoidasida | 1,919 | 530 | 94 | 14 | 717 | 121 |
| **Clade** | Alveolata | 3,746 | 924 | 42 | 7 | 418 | 76 |

Table 4.4 – Number of GO Terms and EC numbers retrieved for the protein clusters found in previous step 4.2.2 and according to the proteome representativeness 0 (Pan-P) 0.5 (Shell-P) and 0.95 (Soft-P).

In the column 'Functional annotation of clusters' of Table 4.4, the numbers of GOs and ECs recovered follow the same trends and are systematically lower than the Pan-P, Shell-P and Soft-P sizes.

**Taxonomic-Scale Metabolic Network.** By the end of the last method step, we have functionally annotated sets of protein clusters. Among these proteins, we could have enzymes that catalyse biochemical reactions, thus allowing us to study the metabolism associated with the taxon. We propose to reconstruct draft Taxonomic-Scale Metabolic Networks (TSMN), in reference to Genome-Scale Metabolic Network (GSMN). Notably, in difference to GSMN, TSMNs are not coming from all the genes of a genome but from the protein clusters found by EsMeCaTa (which represent proteins shared in $Pr\%$ of the proteomes associated with the taxon). Thus, the network's content will change according to the chosen proteome representativeness ratio $Pr$. Indeed using the Soft-P($Pr$=95%), we will only extract the functions associated with enzymes significantly shared in the

taxon. By using the Pan-P we can explore the panmetabolism of the taxon. The step 'EsMeCaTa annotation' creates PathoLogic files that can be used as input to Pathway Tools to reconstruct draft metabolic networks.

**Subsection summary**

For each protein cluster, EsMeCaTa retrieves the **functional annotations** associated with these proteins. Then for each annotation associated with the cluster proteins, we compute the proportion of protein associated with it. Then the annotations occurring among the proteins of the cluster with a greater proportion than a propagation threshold $Pt\%$ are kept.

**Section summary**

To describe the metabolism of metagenomics data, EsMeCata takes as input **taxonomic affiliations**. Thus it can be used in a wide range of situations, resulting from taxonomic assignment following metagenomics and analysis of specific taxon. By relying on **UniProt**, EsMeCaTa uses a reference protein database that allows to cover numerous species and their **proteomes**. The taxonomic affiliation is browsed to find the taxon with the lowest rank providing sufficient knowledge to interpret (default 1 proteome). Then EsMeCaTa uses **MMseqs2** to create **cluster of protein sequences** using sequence similarity to identify homologs sequences. The method then selects protein clusters where $Pr\%$ of the proteomes are represented (**proteome representativeness** ratio $Pr$). This way, EsMeCaTa estimates how representative is a protein cluster according to the proteomes of relative organisms for the taxon. Then by querying UniProt, EsMeCaTa retrieves **functional annotations** associated with the proteins and propagates them to the clusters (according to a $Pt\%$ propagation threshold). This last step creates tabulated files (containing annotations) and PathoLogic files. With the PathoLogic files, it is possible to reconstruct **draft Taxonomic-Scale Metabolic Network** with Pathway Tools.

## 4.3 Application to metagenomics data from a biogas reactor

**Input dataset.** The input dataset consists of data from an experiment on a biogas reactor performed by collaborators from the UR OPAALE (INRAE, Rennes). This experiment monitors the production of biogas (especially methane) in the reactor according to variations in the organic matter (manure, apple, butter, casein) provided to the community (Awhangbo et al. 2020). An illustration of the experiment can be seen in Figure 4.6. The abscissa corresponds to time points, and the ordinate axis represents the quantity of biogas produced. The inputs of the biogas reactor are described under the abscissa according to the time point (manure, apple, butter, or casein). Furthermore, each time point has been labelled as functional (in green) or non-functional (in red) according to the reactor's biogas production and environmental measures, such as the pH in the process.

We also have the result from an analysis of 16S rRNA gene sequencing for each time point. So we have OTUs with taxonomic affiliations (made by FROGS (Escudié et al. 2018)) and the corresponding relative abundances of the OTU for each time point.

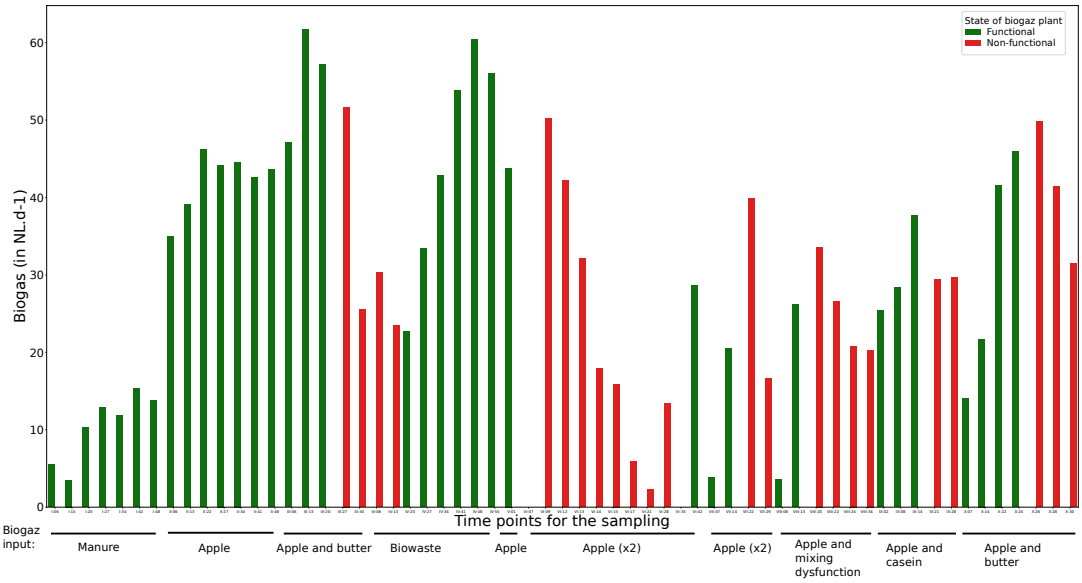



Figure 4.6 – Measure of the biogas production in the biogas reactor experiment (Awhangbo et al. 2020). The abscissa axis corresponds to the time points of the measure of the biogas. Below the abscissa axis, the inputs of the biogas reactor over time are described. The ordinate axis indicates the biogas production in NL.d-1, Normal litre per day, a unit used to measure the gas production on a normal temperature and pressure (in our case 20° C and 1 standard atmosphere), as the volume taken by the same quantity of gas can change according to these two parameters. The bars in green correspond to time points where the biogas reactor was labelled as functional, and the bars in red correspond to time points where the biogas reactor was described as non-functional.

### 4.3.1 Predicting annotated protein clusters from taxonomic affiliations

**Proteomes.** We applied EsMeCaTa on the 587 OTUs found by FROGS (Escudié et al. 2018). First, we got for each OTU the lowest taxonomic rank associated with at least one proteome. Among the 587 OTUs, 12 were associated with Archaea and 575 were associated with Bacteria.

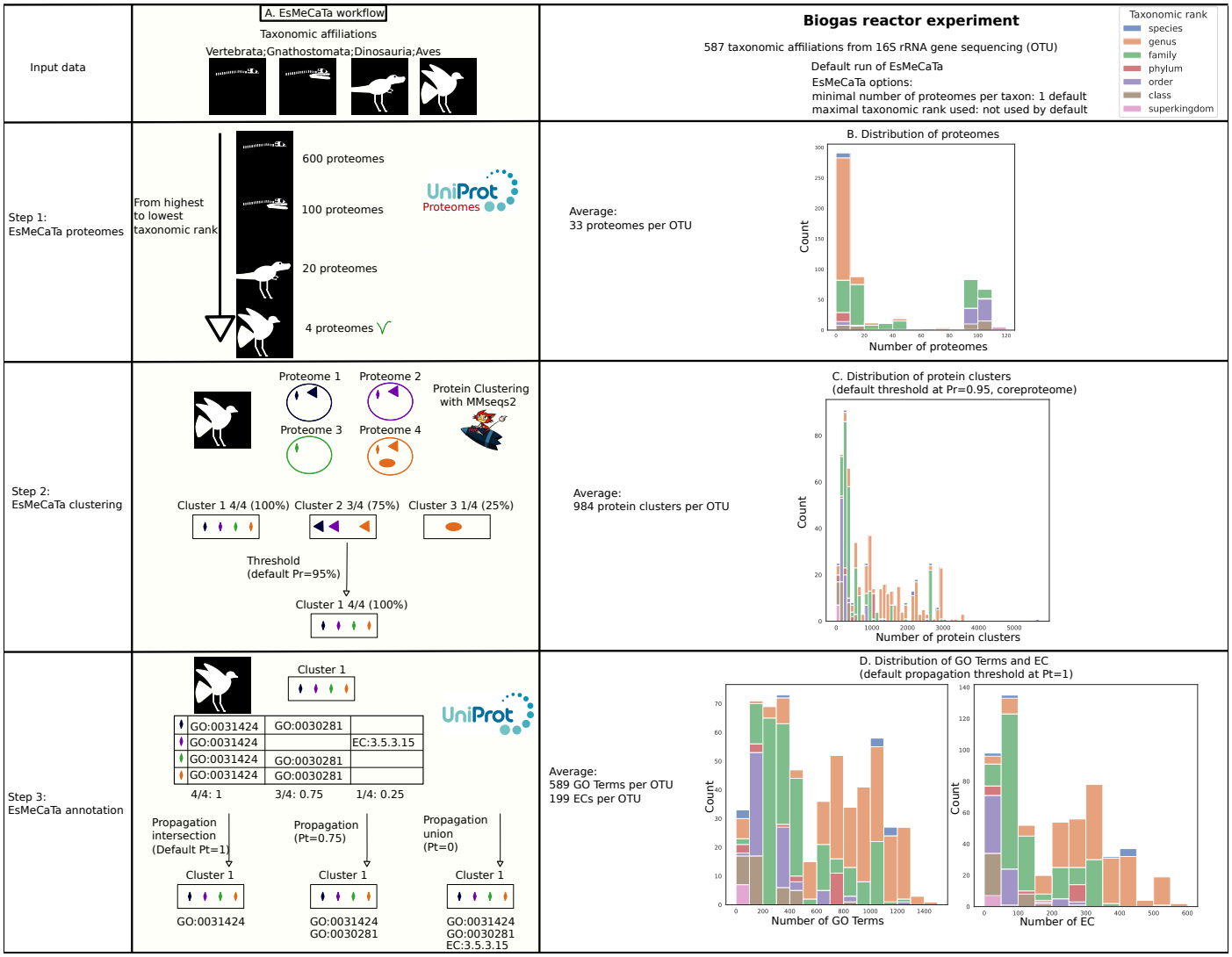



Figure 4.7 – **Results of EsMeCaTa using default options on the biogas dataset.** A. EsMeCaTa workflow description. B. Distribution of the number of proteomes per OTU found by EsMeCaTa according to the taxonomic rank used by EsMeCaTa. The default subsampling at maximum 100 proteomes explains the distribution pick around 100 proteomes. C. Distribution of the number of protein clusters with the default proteome representativeness ratio $Pr = 0.95\%$. D. Distribution of the annotation found by EsMeCaTa for the OTU with default propagation threshold at $Pt = 1$. At the left the distribution of the number of unique GO Terms according to the OTU, and at the right the distribution of unique EC. The taxonomic rank is shown in colour in all the different graphs.

**Proteomes selection.** On average 33 proteomes per OTU were found. The distribution of the results can be seen in Figure 4.7 B. The taxonomic ranks mostly selected by

EsMeCaTa were genus, family and order. Also, among the OTU with 0-10 proteomes, 141 were associated with 1 proteome. The lowest the taxonomic rank (species (in blue) or genus (in orange)) is, the least proteomes are available. When EsMeCaTa selects the genus for a taxonomic affiliation, it often obtains less than 10 proteomes. But when higher taxonomic ranks are selected, there is a greater number of proteomes retrieved. Most taxa associated with family (in green) are associated with more than 10 proteomes. Finally, the maximal number of proteomes cutoff was applied when more than 100 proteomes were available. This is applied for the highest taxonomic rank (order in purple, class in brown) when the taxonomic affiliation is linked to a poorly documented clade or has high taxonomic assignment uncertainty.

**Protein clustering.** Then, EsMeCaTa used MMseqs2 to cluster the proteins contained in these proteomes. For this experiment, we applied the threshold associated with the Soft-Core proteome (proteome representativeness $Pr = 0.95$). On average, 984 protein clusters were found by OTU. The distributions of protein cluster sizes is shown in the Figure 4.7 C. For the lowest taxonomic rank, the proteome diversity was expected to be low (proteomes in a genus are not expected to be highly diverging); thus, more protein clusters are expected to be found (more proteins are likely shared by the genus). And this was observed as the highest number of protein clusters per OTU (between 1,000 and 3,000) were mostly associated with genus taxonomic ranks (orange in Figure 4.7 C). Contrarily, for higher taxonomic rank, proteome diversity was expected to be more important. Consequently, it was less likely that all the proteomes shared a protein. The number of protein clusters was then lower (between 0 and 500) for the taxon associated with higher taxonomic rank (order in purple, class in brown and superkingdom in rose in Figure 4.7 C).

**Protein cluster annotation.** For each of the protein clusters selected in the previous step, EsMeCaTa searched for functional annotations in UniProt (especially GO Terms and EC numbers). We have an average of 589 unique GO Terms per OTUs and 199 unique ECs per OTUs (Figure 4.7 D). Such as, in the previous paragraph, when EsMeCaTa selects lower taxonomic ranks, there was less diversity among the selected proteomes. This selection leads to more protein clusters and annotations in the Soft-Core proteome. For example, genera were often associated with 500 to 1400 GO Terms and 150 to 600 EC numbers. And for higher taxonomic rank, more proteomes with more diversity led to

fewer protein clusters in the Soft-Core proteome. Then fewer annotations were retrieved, between 0 and 500 GO Terms and 0 and 200 ECs for the OTUs having the rank order (brown in Figure 4.7 D).

**Draft metabolic network.** During the previous step, EsMeCaTa created PathoLogic files for each OTU that were given as input to Pathway Tools. In this way, we were able to reconstruct draft metabolic networks for each OTU with an average of 598 reactions per OTU. The distribution of the number of reactions inferred by this step can be seen in Figure 4.8. Following the annotation prediction of the Soft-Core proteomes, lower taxonomic ranks were associated with higher numbers of reactions (genus in orange). Higher taxonomic ranks were linked to fewer reactions.

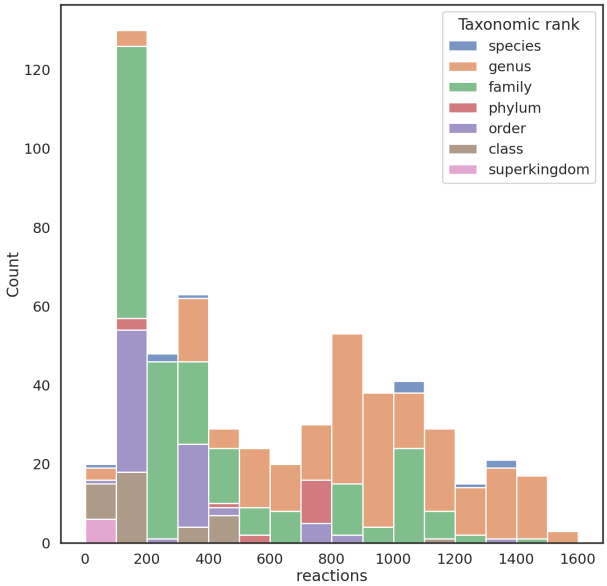



Figure 4.8 – Distribution of the number of reactions in each TSMN estimated for each OTU. Colour corresponds to the taxonomic rank used by EsMeCaTa to retrieve the proteome from a taxonomic affiliation.

Then we analysed the TSMN produced with the default option of EsMeCaTa. We summed the abundance of OTUs having the reaction producing the methane (in MetaCyc: 'METHYL-COM-HTP-RXN') for each time point, and we compare these production predictions to the variation of the measured biogas production (Figure 4.9).

We observe that the biogas production can be divided into two parts, also seen in Figure 4.6. The first half from the first time points to time points VI, where the biogas reactor shows a long period of stability with only two dysfunctions. In this first half, the

methane predictions agree with measured productions. This suggests that methanogenesis reactions are present in the TSMNs and that the methanogenic population size (sum of OTUs abundances) might be a sufficient proxy to estimate the biogas reactor production.

The second half (after time points VI) is less stable, with more variations in biogas production. The function of the biogas reactor was more chaotic as it was emptied and reinoculated several times. From these times, the predictions and the measured productions are no more congruent.

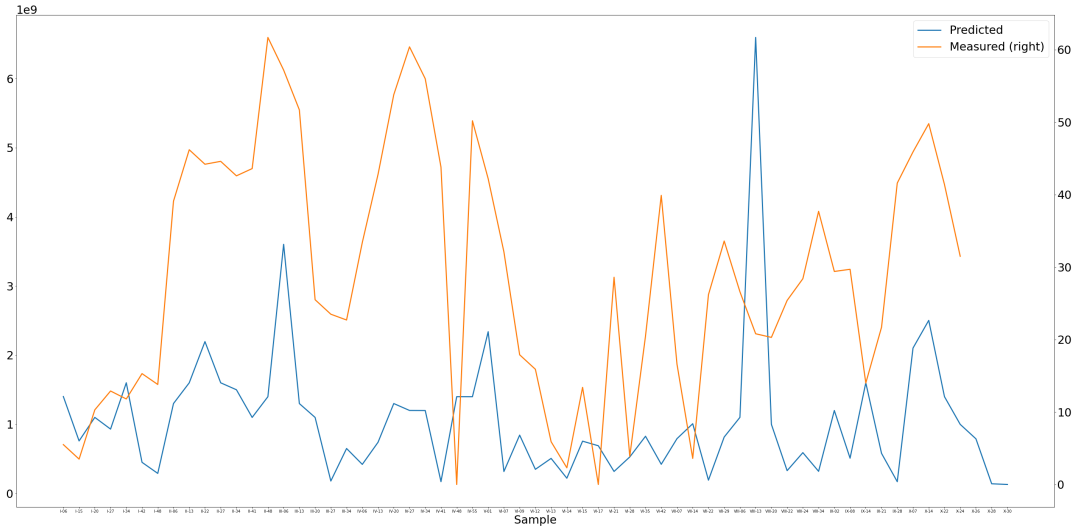



Figure 4.9 – Summed relative abundance of the OTUs having methanogenesis reaction in their draft TSMN (blue) and the measured production of biogas (orange) over time points.

These TSMN will be used in the next Chapter 5 to study metabolic complementarity in the biogas reactor community. Furthermore, EsMeCaTa predictions (with default option) and OTUs abundances will be used in the next subsection 4.3.3 to compare the functional and non-functional time points identified in the Figure 4.6.

**Subsection summary**

We have seen in this subsection the results of the prediction of EsMeCaTa on a dataset from the taxonomic assignments of 16S rRNA gene sequences. We have seen the impact of **knowledge bias**. Some OTUs were associated with lower taxonomic rank, thus having more protein clusters and annotations in their Soft-Core proteomes. Inversely, other OTUs were associated with higher taxonomic ranks and had more proteomes but fewer protein clusters and annotations in their Soft-Core proteomes. This pattern reversed when we decreased the proteome representativeness ratio to select the Pan-proteome. Finally, an analysis of the abundance of TSMN having a methane-producing reaction was performed, suggesting that methanogenesis reactions are present in the reconstructed TSMNs.

### 4.3.2 EsMeCaTa predictions according to different options

To explore the metabolism associated with the taxon of the biogas dataset, we investigated the effect of selecting different proteome representativeness ratios. Also, to avoid the fact that some OTUs were associated with only one proteome or with taxonomic rank too high to have enough protein clusters (subsection 4.3.1) several constraints were imposed on the proteome selection of EsMeCaTA.

First, the minimal number of proteomes for a taxon was increased from 1 to 5. In this way, EsMeCaTa will only select taxon when they are associated with at least 5 proteomes. If they are associated with fewer proteomes, EsMeCaTa will search for the proteomes of a higher taxon in the taxonomic affiliation.

A second option was applied, the maximal taxonomic rank selected by EsMeCaTa. By default, this option was not used. In the runs presented in this subsection, we tuned this option to choose only three taxonomic ranks: species, genus and family. This means that EsMeCaTa will search for proteomes associated with the OTU up to the family. If no proteomes validating the other options are found, the OTU will be ignored. This procedure will remove OTUs that belong to poorly documented taxa or have high uncertainty in their taxonomic affiliations. One possible issue of keeping them will be having very few protein clusters due to the high diversity among the selected proteomes, as EsMeCaTa had to use higher taxonomic ranks.

These constraints led to a decrease in the OTU used from 587 to 362 as EsMeCaTa could not find taxon with proteomes satisfying the constraints for 225 taxonomic affilia-

tions.

Due to the combination of options, the major taxonomic rank used by EsMeCaTa to find proteomes was the family (Figure 4.10 A). On average, 39 proteomes were associated with an OTU using these options. This average is superior to the average number of proteomes (33 proteomes per OTU) found with the default option. Still, the same patterns that were identified with the default options were retrieved here. The higher taxonomic rank (here, the family) was associated with more proteomes, and the lower taxonomic ranks (species and genus) were linked to fewer proteomes (Figure 4.10 A).

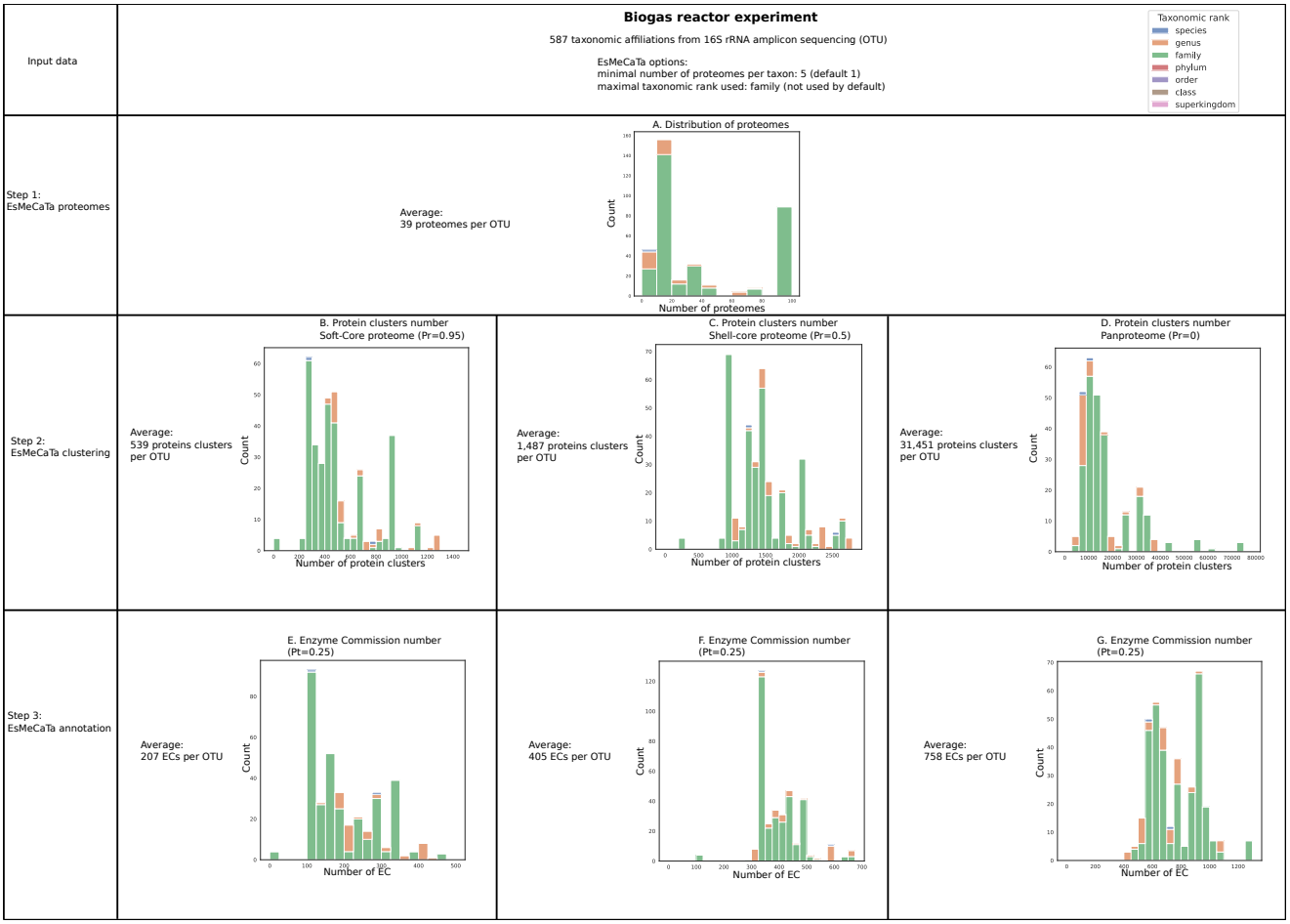



Figure 4.10 – **Results of EsMeCaTa on the biogas dataset with specific options**. A. Proteomes distribution with maximal taxonomic ranked limit at family and a minimal number of proteomes at 5. Distribution of protein clusters for: B. Pan-proteome ($Pr = 0$), C. Shell-Core Proteome ($Pr = 0.5$) and D. Soft-Core proteome ($Pr = 0.95$). Distribution of EC numbers retrieved for: E. Pan-proteome ($Pr = 0$), F. Shell-Core Proteome ($Pr = 0.5$) and G. Soft-Core proteome ($Pr = 0.95$).

The clustering of the proteins was performed with 5 proteome representativeness ratios ($Pr = 0$, $Pr = 0.25$, $Pr = 0.5$, $pr = 0.75$ and $Pr = 0.95$). Three are shown in the Figure 4.10 (B $Pr = 0$, C $Pr = 0.5$ and D $Pr = 0.95$) as the ratios $Pr = 0.25$ and $Pr = 0.75$ presented similar distribution than the ratio $Pr = 0.5$ . The proteome representativeness ratio for the Soft-core proteome ($Pr = 0.95$) presented a similar pattern to the one analysed with the default options. Higher taxonomic ranks (family in green) were associated with fewer protein clusters than lower taxonomic ranks (genus in orange). For the Shell-Core proteome, this was still true. But as the proteome representativeness ratio decreased, this pattern reversed. And with the Pan-proteome, this was completely reversed. Higher taxonomic ranks were associated with more protein clusters, and lower taxonomic ranks were associated with fewer protein clusters. This result was expected as the Pan-proteome keeps all the protein clusters, and as higher taxonomic ranks are associated with more proteomes and high proteome diversity, they have more protein clusters. For example, in the Pan-proteome, all the OTUs with more than 40,000 protein clusters were associated with the family taxonomic rank.

As we have more proteomes and protein clusters than the experiment with proteome representativeness $Pr = 0.95$, we used a lower annotation propagation threshold $Pt$. By default, it is at $Pt = 1$, meaning that annotation associated with a protein cluster must be present in all the proteins of the protein cluster. It has been lowered to $Pt = 0.25$, meaning that an annotation is associated with a protein cluster when 25% of the protein in the cluster have this annotation.

Then EsMeCaTa annotates the protein clusters found for each proteome representativeness ratio. The same pattern as the one observed with protein clusters was observed for the number of GO Terms and the number of EC (but only EC numbers are shown in Figure 4.10 E, F and G).

Then we reconstructed for each proteome representativeness ratio the corresponding draft TSMN. We used the output annotations of EsMeCaTa and gave them as input to Pathway Tools.

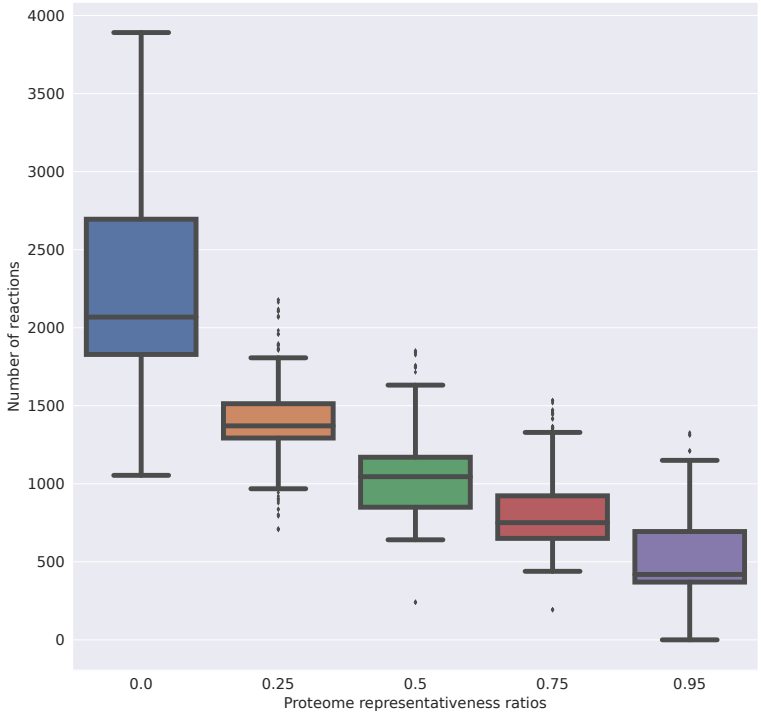



Figure 4.11 – Distribution of the number of reactions inferred by Pathway Tools for the 362 OTUs by using the annotations found by EsMeCaTa according to the proteome representativeness ratio $Pr$.

The number of reactions increased with the decrease of the proteome representativeness ratio 4.11. As we can see, the number of reactions for the Soft-Core proteome ($Pr = 0.95$) is around 500, whereas the number of reactions in the metabolism is around 2,000 for the Pan-proteome. In this way, we have different sets of reactions, one representing the metabolic potential that is shared by most proteomes of the taxon ($Pr = 0.95$, the Soft-core metabolism). And another represents most of the known metabolic potential of the taxon ($Pr = 0.0$, the Pan-metabolism).

**Subsection summary**

In this section, we have explored some EsMeCaTa options and their impacts. The minimal number of proteomes per taxon and the limit on maximal taxonomic rank constrained the taxon used. These options impact the predictions by decreasing the number of OTUs for which EsMeCata found proteomes. Less OTUs were selected and the taxonomic affiliations kept were the ones associated with more knowledge on UniProt. The proteome representativeness ratio allowed the selection of different numbers of protein clusters to model either the Pan, Shell-Core or Soft-Core proteomes. The Pan-proteome was associated with a higher number of reactions as most of the available metabolic knowledge was associated with a taxon. With the Soft-Core proteome, fewer reactions were associated following the number of protein clusters well represented in the different proteomes.

### 4.3.3 Prospectives: Applying Machine Learning on EsMeCaTa results

This subsection presents preliminary results on the EsMeCaTa outputs. It especially uses a method still in development by Baptiste Ruiz, a PhD student at the IRISA (supervised by Yann Le Cunff) and not published. So it must be considered as a prospective analysis to test the utility of EsMeCaTa results for Machine Learning classification.

**Analysing OTU abundance.** A first analysis based on the abundance of OTUs was performed using the classification developed in DeepMicro (Oh et al. 2020), especially the random forests. The random forest goal was to classify the samples (here, time points of the experiment) according to the state of the biogas reactor (either functional or non-functional). Then the classification results were analysed with metacoder (Foster et al. 2017) to visualise the taxonomic classification of the discriminant OTUs of the classification. To differentiate the discriminant OTUs in functional or non-functional biogas reactor states, we compared the average abundance of the OTUs between the two sets of time points. We separated the OTU depending on whether they were more abundant in the functional or the non-functional biogas reactor.

The Random Forest classification was run 20 times. It achieved the classification of the samples with an average AUC of 0.82. First, two separated taxonomic groups have been identified for the functional biogas reactor (Figure 4.12 A), one associated with

Bacteria and the other with Archaea. This first result is expected as both Bacteria and Archaea are needed to produce methane (Bacteria performed the first steps of organic matter degradation, and Archaea performed the last step of methanogenesis). We can see that the most important taxonomic groups according to the Random Forests are the Clostridiales (among them the Ruminococcaceae), the Bacteroidales, the Synergistaceae, the Lactobacillales and the Methanomicrobia.

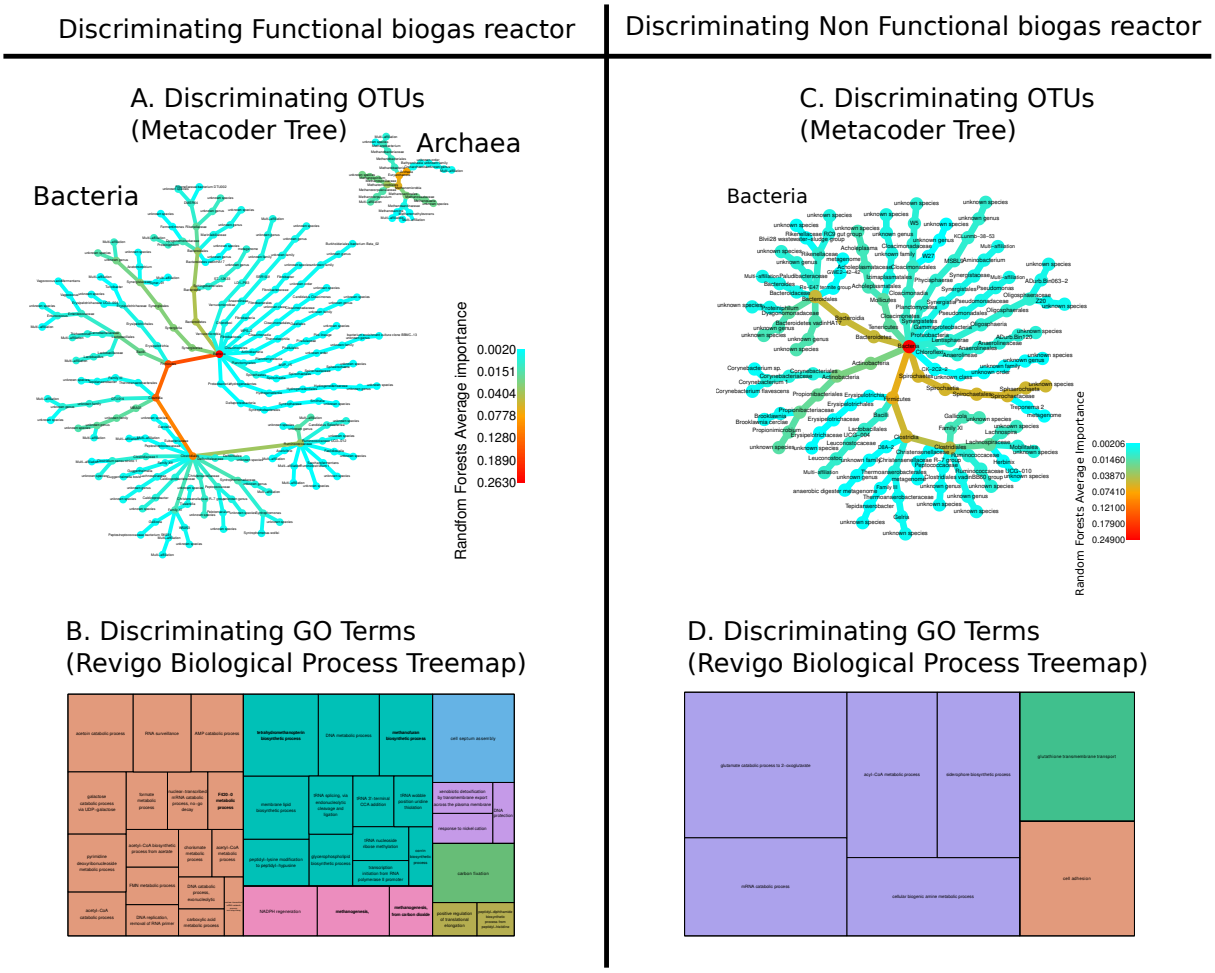



Figure 4.12 – A. Metacoder tree representation of the taxonomic classification of the 84 most important OTUs according to the Random Forests classifiers for **functional biogas reactor** according to the average importance outputted by the classifier. These 84 OTUs are also more abundant on average in the samples from the functional biogas reactor than in the non-functional biogas reactor. B. Revigo Tree map on the 110 GO Terms determined as most important annotations by the Random Forests for **functional biogas reactor** according to the average importance outputted by the classifier. The 110 GO Terms are, on average, more abundant in the functional biogas reactor than in the non-functional biogas reactor. C. Metacoder tree representation of the taxonomic classification of the 56 most important OTUs for the Random Forests classifiers for **non-functional biogas reactor** according to the average importance outputted by the classifier. These 56 OTUs are also more abundant on average in the samples from non-functional biogas reactors than in the functional biogas reactor. D. Revigo Tree map on the 20 GO Terms determined as the most important annotations by the Random Forests for **non-functional biogas reactor** according to the average importance outputted by the classifier. The 110 GO Terms are, on average, more abundant in the non-functional biogas reactor than in the functional biogas reactor.

For the non-functional biogas reactor (Figure 4.12 C), the most striking result was the absence of Archaea (let us recall that Archaea are needed to produce methane). The most important groups discriminating non-functional reactors against the functional ones were the Sphaerochaeta, some Clostridiales and some Bacteroidales.

By comparing the two figures, we could see differences in the presence of some taxonomic groups. First, there was a higher diversity of discriminating OTUs in the functional compared to the non-functional biogas reactor. For example, the Clostridiales were important for both functional and non-functional biogas reactors. Still, there was a higher number of Clostridiales ranked as a discriminant in the functional biogas reactor, especially the Ruminococcaceae.

**Analysing annotations abundance.** In a second analysis, we used another classification method currently developed by Baptiste Ruiz, supervised by Yann le Cunff (IRISA, Rennes). So the details of the method will not be discussed in this manuscript. Instead of classifying the OTU according to their abundance, it classifies the annotations found by EsMeCaTa by specifying a score combining OTU abundance and annotation occurrences. This classification was performed on the GO Terms and the EC numbers. Then the discriminant GO Terms separating functional and non-functional biogas reactors were visualised using Revigo (Supek et al. 2011). Again we separated them according to the difference between the average abundance of the annotation in the sample associated with a functional reactor compared to the non-functional reactor.

The Random Forest classification was run 20 times. It achieves the classification of the samples with an average AUC of 0.92, suggesting that a more discriminant signal can be found when comparing the samples' annotations than when comparing the samples' OTUs. For the GO terms classified as important by the classifier and more abundant in the samples associated with functional biogas reactor (Figure 4.12 B), we can see that some GO Terms are directly related to methanogenesis. Indeed we have the GO Term 'methanogenesis' and other GO Terms associated with metabolites close to the methanogenesis pathway shown in bold ('F420-0 metabolic process', 'tetrahydromethanopterin biosynthetic process' and 'methanofuran biosynthetic process'). This validated the results found from the OTU abundance because the discriminant annotations selected here were directly associated with methane production.

We had very few discriminating GO Terms (20) for the GO terms classified as important by the classifier and more abundant in the samples associated with non-functional

biogas reactor (Figure 4.12 D). These GO Terms are still analysed.

**Subsection summary**

From the annotation predicted by EsMeCaTa for 587 OTUs and the OTUs abundances, we could discriminate OTUs and annotations according to the state of the biogas reactor (functional or not). Using the OTUs, we were able to identify a subset of OTUs more present and discriminant for a functional biogas reactor which seemed plausible as it contains Bacteria associated with methanogenic Archaea. Then we could also identify a sub-set of Bacteria which seemed more present in a non-functional biogas reactor. For the annotations, the separation between functional and non-functional biogas reactors was associated with annotations linked to methanogenesis and methane synthesis. Furthermore, we were able to identify other functional annotations that could help to understand better the difference between a functional and a non-functional biogas reactor. These results also showed that the Soft-Core proteome estimated by EsMeCaTa provided meaningful biological signals.

**Section summary**

In this application, we have shown the result of EsMeCaTa on an experiment using metabarcoding to study the functionality of a biogas reactor. First, EsMeCaTa retrieved the set of **protein cluster** and **functional annotations** associated with the different taxonomic affiliations from the metabarcoding. We also performed an experiment to show the impact of EsMeCaTa options on its predictions. Then by using machine learning methods on these results, we were able to classify the samples from functional to non-functional biogas reactors. Looking at the OTU abundance, one of the main differences was that **methanogenic Archaea** were more abundant in the functional reactor than in the non-functional one. By examining the annotations produced by EsMeCaTa, we found that methanogenesis **GO terms** discriminated the functional biogas reactor, suggesting that EsMeCaTa produced relevant annotations.

## 4.4 Conclusion

### Contribution

This chapter presents a new method to estimate metabolic capabilities from taxonomic affiliations. The goal of this method was to satisfy the need to infer draft metabolic networks from metabarcoding data. The **taxonomic affiliations** were selected as input because they can be created from numerous sequencing methods (metabarcoding from any gene markers, shallow shotgun sequencing) but can also be manually created to work on taxon of interest. The method searches for the lowest possible taxonomic rank in the taxonomic affiliations to associate proteomes with the corresponding taxon. Then it creates, by clustering the proteomes, a set of protein clusters that can be described as representative of the taxon (according to the representation of the proteomes in the protein clusters). By finding these protein clusters and their annotations, we could identify the enzymes among them.

To that end, I implemented this method in the python package EsMeCaTa. From a taxonomic affiliation, EsMeCaTa finds the **lowest taxonomic rank** associated with **proteomes** in **UniProt**. Then a clustering is performed on the proteins in these proteomes with **MMseqs2**. This step creates a set of **protein clusters** for each taxonomic affiliation. Then these protein clusters are filtered according to the number of input **proteomes represented** in them. In this way, detecting proteins conserved in numerous proteomes is possible. Then these protein clusters are **functionally annotated** by querying UniProt. With the output of EsMeCaTa, it is possible to reconstruct draft Taxonomic-Scale Metabolic Networks.

This method was applied to a biogas reactor dataset. In this experiment, the biologists attempted to separate functional from non-functional biogas reactors to identify the organism associated with these differences. To that end, abundances of OTUs were available for different time points (labelled as being associated with a functional or non-functional biogas reactor). Using the taxonomic affiliations of the OTUs, EsMeCaTa was able to predict functional annotations related to each OTU. We saw that the taxonomic rank selected by EsMeCaTa impacted the following predictions. With the default option and the search of the Soft-Core proteomes, lower taxonomic ranks were associated with fewer proteomes and a higher number of annotated protein clusters. In contrast, higher taxonomic ranks were associated with higher numbers of proteomes and less annotated protein clusters. Furthermore, we were able to reconstruct draft Taxonomic-Scale Metabolic Net-

works that allow studying metabolisms of taxa. This will be used in the next Chapter 5 to study the metabolic interactions in the community.

Following this run of EsMeCaTa with default options, a second experiment was performed with different options to constrain the number of selected proteomes. This change led to similar results to the ones of the Soft-Core proteome. We also studied the Pan-Proteome and found that it produces opposite results to the Soft-Core proteomes. Indeed with the Pan-Proteome, lower taxonomic ranks were associated with fewer protein clusters than higher taxonomic ranks. This observation can be interpreted in terms of taxa diversity, which is more significant for higher taxonomic ranks, and functional diversity, which is wider for the Pan-proteome.

Then using machine learning, we were able to separate samples between functional or non-functional biogas reactors according to the OTU abundance in the first experiment and according to the annotations found by EsMeCaTa in the second experiment. Discriminating OTUs and annotations were linked to methane production, highlighting the identification of meaningful biological signals.

## Limits and improvements

Despite the validation presented in the previous paragraph, it is important to keep this method's limits in mind. Indeed this method is sensitive to numerous biases.

The first is that methods estimating function for unknown wild organisms often use known cultivated species. But these species, despite being in a similar taxonomic group, can have different functions and adaptations. We try to limit this by considering the proteins shared inside a taxon, but this does not allow us to study the specific functions of an unknown species.

A second one is a limit to the knowledge available. EsMeCaTa relies on several databases (NCBI taxonomy, UniProt); thus, it will only be able to predict what is known and what is present in these databases. Also, some species are far more studied than others; then, some groups will be more annotated than others. A possible way to fix this issue is to propagate the known annotations (like the one performed in Chapter 3). A possible way of propagating such annotations could be to created a set of annotated sequences that could be given as input to the clustering step. It is possible to create such data using the Enzyme Expasy and the UniProt databases. With the first one, we have access to Enzyme Commission number and with the second one, we have access to the associated protein sequences. It is possible to create a consensus sequence for each Enzyme Commis-

sion number. Then during the clustering step on the proteome the consensus sequences could be given as a supplementary input to identify sequence. Then all protein clusters containing one consensus sequence will be annotated by this method.

A third issue with the method is the time needed. Indeed this method takes more than a day for a dataset of 500 OTUs. The bottlenecks are the multiple queries to UniProt and the clustering of sequences. One way to fix this issue could be to create a database for each taxonomic rank, but with this, we lose the advantage of using the newest version of the different databases.

A fourth issue is the databases used. Indeed we rely on the NCBI Taxonomy database, but the user could prefer to provide taxonomic affiliations from a different database (for example, the GBIF taxonomy (GBIF Secretariat 2021)). If this is the case, EsMeCaTa will likely be unable to work as it will not recognize the taxa. An improvement would be to take this information into account. One possible way could be to use Taxonbridge (Veldsman et al. 2022) as it tries to combine the two databases.

A way to improve EsMeCaTa could be to identify the set of EsMeCaTa options that allows for finding a better intersection between its prediction and the proteins present in a MAG or a complete genome. This comparison could select the best parameters to extract the estimated proteins when working on environmental samples.

## Perspectives

EsMeCaTa allows working on new datasets to perform metabolic analysis by estimating the metabolic capacities from a taxonomic affiliation. But it also permits to study of the set of shared proteins among large taxonomic groups. It could permit analysis of the known and unknown functions of organisms. The clustering method identifies groups of probable homologous proteins. Multiple analyses could be performed on these groups. A perspective could be exploring the shared domain among these sequences using multiple alignment tools. But this surely needs to tune the parameters used by MMseqs2 on the clustering to identify relevant domains. But such analysis could propose exploring the protein domain diversity among taxonomic groups and refining the protein cluster annotations. The current perspective of the method relies finally on the generated TSMNs, which will be used to test metabolic interactions in the community. In Chapter5 the EsMeCaTa outputs will be used to identify key species in the biogas reactor according to metabolic complementarity.

A second perspective could be to add more information for EsMeCaTa during the

selection of proteomes. We rely on data from UniProt to select proteomes, but other knowledge could be used. We could improve the prediction by selecting proteomes of organisms living in the same condition as the conditions of the wild organisms. We need to add living conditions knowledge to this method to achieve this. This knowledge could be added using databases containing this information, such as BacDive (Reimer et al. 2022). By combining UniProt and BacDive, it could be possible to create a more realistic estimation of metabolism.

Chapter 5

# Identification of key species in microbiota using metabolic complementarity

In the previous Chapter 4, I presented a method to estimate metabolic capabilities from taxonomic affiliations to answer the need to predict metabolic networks from metagenomics data. I estimated the metabolism for a set of taxonomic affiliations from 16S rRNA gene sequencing on a community from a biogas reactor. Now, the issue is to be able to study the metabolic interactions between the members of the community.

Several methods exist to study the possible metabolic interactions between GSMN. Still, they often rely on functional GSMN (GSMN that can produce biomass), and they often have issue handling large-scale microbiota. To address these issues, I have participated in the development of a new method in collaboration with Clémence Frioux (INRIA, Bordeaux), called Metage2Metabo. The goal of this method is to identify key species for the production of target metabolites in a community. By using genomes as input, it is possible to study large-scale microbiota thanks to parallel computing. Then topological analysis finds the metabolites producible by the individual organisms and the community. Then it identifies the organisms that collectively achieve the target metabolite production.

I will present this method and its application to find the key taxa involved in the organic matter degradation in the biogas reactor experiment presented in the previous Chapter 4. The first two sections (which present the method) have been extracted from the article published in *eLife* (Belcour et al. 2020a).

## 5.1 Predicting metabolic interactions in community

### 5.1.1 Metabolic interactions between GSMNs

Studying the interactions inside a microbial community is possible thanks to the improvements in metagenomics. In the previous Chapter 4, I investigated the possibility to reconstruct hundreds or thousands of genomes from environmental samples (Pasolli et al. 2019; Forster et al. 2019; Zou et al. 2019; Stewart et al. 2018; Almeida et al. 2021). Also, it has been shown in the previous Chapter 4 that it is possible to reconstruct metabolic networks from taxonomic affiliations.

**Organisms interactions.** There are multiple possible interactions between organisms as seen in subsection 1.5.1. Some have a negative impact on the participants, such as competition with either direct competition for a resource (**interference competition**) or an indirect impact as the organisms use the same resource called **exploitation competition** (Birch 1957). There are also positive interactions for one or the two participants, such as **cross-feeding** where organisms exchange elements. Multiple cross-feeding types have been proposed (Smith et al. 2019) according to how and what is exchanged. This cross-feeding can become mandatory, such as some *syntrophic* interactions where two organisms are in a mandatory relationship to exchange compounds for their growth. The dependencies can occur after gene loss such as proposed in the **Black Queen Hypothesis** (J. J. Morris et al. 2012; Mas et al. 2016). In this hypothesis, some metabolic functions can be costly for an organism to maintain, and their loss thus provides a selective advantage. Organisms maintaining these functions provide a "public good" useful for a part of the community. Some organisms (called helpers) will keep these functions and help the community by providing these functions. These interactions are also defined as commensalism interactions. In this regard, these organisms can be seen as keystone species, meaning that the community will dramatically change if they are removed. The other organisms depending on these helpers are called *auxotrophs* as they can not synthesise metabolites essential for their growth and require other sources for these metabolites.

**Metabolic interactions.** To unravel such interactions between species, it is necessary to go beyond the functional annotation of individual genomes and interpret metagenomics data in terms of metabolic modelling. The main challenges impeding mathematical and computational analysis and simulation of metabolism in microbiomes are the scale of

metagenomic datasets and the incompleteness of their data. Indeed, reconstructing all the GSMNs corresponding to the hundreds or thousands of genomes from an environmental sample sequencing can be challenging. Some tools have been developed to fastly reconstruct metabolic networks, especially by using a top-down approach (from a universal model), such as CarveMe (Machado et al. 2018) or gapseq (Zimmermann et al. 2021). But they often rely on gap-filling approaches to produce functional GSMN, which can be an issue if we want to model auxotrophic organisms. Indeed, as mentioned above, such organisms took selective advantage of a function loss and relied on other organisms' products to grow. Gap-filling, in this case, could interpret the function loss as an artefact to be corrected. Furthermore, these methods work on Bacteria and Archaea; thus, such analysis of eukaryotes is challenging. So new methods to reconstruct large-scale metabolic networks could be used.

Reconstructed GSMNs are a resource to analyse the metabolic complementarity between species, which can be seen as a representation of the putative cooperation within communities (Opatovsky et al. 2018). Most of the methods to study metabolic interactions rely on constraint-based modelling and predict interactions from pairwise comparisons to associate GSMNs in small communities (Chan et al. 2017; Zomorrodi et al. 2012; Khandelwal et al. 2013; Bauer et al. 2017; Mendes-Soares et al. 2016). SMETANA (Zelezniak et al. 2015) estimates the cooperation potential and simulates flux exchanges within communities. It has been applied to communities with up to 40 members (Machado et al. 2021). Recently new methods were developed to study the metabolic interactions for bigger communities such as MICOM (Diener et al. 2020) or the "Microbiome Modeling Toolbox" (Baldini et al. 2019).

However, these tools can only be applied to communities with few members, as the computational cost scales exponentially with the number of included members (Kumar et al. 2019). Only recently has the computational bottleneck started to be addressed (Diener et al. 2020; Baldini et al. 2019). In addition, current methods require GSMNs of high quality to produce accurate mathematical predictions and quantitative simulations. Reaching this level of quality entails manual modifications to the models using human expertise. This curation is not feasible at a large scale in metagenomics. Automatic reconstruction of GSMNs scales to metagenomic datasets but comes with the cost of possible missing reactions and inaccurate stoichiometry that impede the use of constraint-based modelling (Bernstein et al. 2019). Therefore, the development of tools tailored to analyse large communities is needed.

### 5.1.2 Metabolic interactions in large-scale microbiota

Methods have been developed to reduce a large-scale community in a minimal set of organisms selected to produce specific metabolites. For example, (Eng et al. 2016) has developed CoMiDA to select a group of species producing the desired targets. This non-compartmentalised method allows the exchange of metabolites between organisms without any cost. This community formalism is called the mixed-bag (Henry et al. 2016).

Other methods were developed to study metabolic interactions in the community by using the network expansion algorithm (Ebenhöh et al. 2004). By using it, MiSCoTo (Frioux et al. 2018a) computes the metabolic potential of interacting species and performs community reduction. It has been applied to select symbionts in the bacterial community associated with the *Ectocarpus siliculosus* (Burgunter-Delamare et al. 2020). But this method requires GSMN associated with each organism and thus requires a way to reconstruct GSMNs on a large scale. Furthermore, it allows community reduction for a set of target metabolites. But it is possible that one does not have target metabolites and wants to study the metabolic potential of the community. A new method is required to perform such analyses.

**Issue on scalability for large-scale community**

To identify the interactions among the ecosystems, we need to study the metabolic potential at the organism communities' scale. Such analysis faces two **scalability** issues, inferring the organism GSMNs and modelling the **metabolic interactions** for large-scale microbiota.

As an increasing amount of complete genomes and Metagenome-Assembled genomes (MAG) are currently reconstructed, there is an emerging need to study the metabolic potential of the corresponding communities. In this chapter, I present a method allowing to scale up to the size of metagenomic communities and designing the putative minimal communities which can achieve targeted metabolite productions. Furthermore, it can also estimate the metabolic potential of the community.

**Section summary**

We need to find the **metabolic potential of the microbiota communities**. To this end, we first have to reconstruct organism GSMNs for a large number of input **genomes**, applying methods scaling up to the complexity of metagenomic data. Then we will need to assess the individual metabolic potential of each GSMN and examine the possible interactions between all the GSMNs. By combining these pieces of information, we will be able to identify the species of interest in the community and the metabolic potential of the community.

## 5.2 Metage2Metabo: identification of key species according to metabolic complementarity

Metage2Metabo is a solution that performs automatic GSMN reconstruction and systematic screening of metabolic capabilities for up to thousands of species for which an annotated genome is available. It can also directly take as input already reconstructed metabolic networks such as the one created by Pathway Tools on EsMeCaTa outputs (as done in section 4.3.1). The tool computes both the individual and collective metabolic capabilities to estimate the complementarity between the metabolisms of the species. Metabolism complementarity is determined to satisfy a particular metabolic objective, which typically is the production of targeted metabolites that need cooperation. Metage2Metabo performs a community reduction step that aims at identifying a minimal community fulfilling the metabolic objective and outputs the set of associated key species.

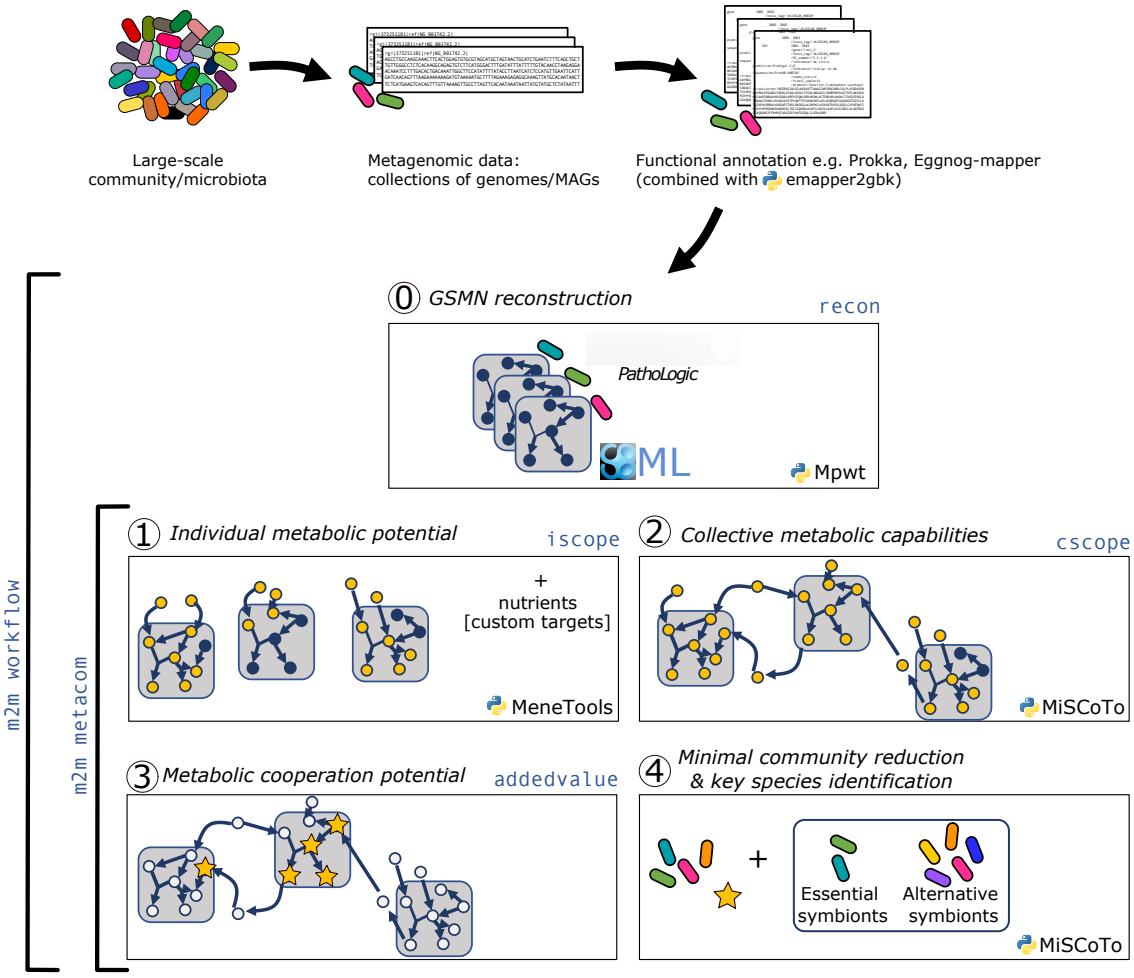



Figure 5.1 – **Overview of the Metage2Metabo pipeline.** Main steps of the M2M pipeline and associated tools. The software's main pipeline (`m2m workflow`) takes as inputs a collection of annotated genomes that can be reference genomes or metagenomics-assembled genomes. The first step of Metage2Metabo consists in reconstructing metabolic networks with Pathway Tools (step 0). This first step can be bypassed, and GSMNs (or TSMNs) can be directly loaded in Metage2Metabo. The resulting metabolic networks are analysed to identify the individual (step 1) and collective (step 2) metabolic capabilities. The addedvalue of cooperation is calculated (step 3) and used as a metabolic objective to compute a minimal community and key species (step 4). Optionally, one can customise the metabolic targets for community reduction.

Metage2Metabo's main pipeline (Figure 5.1) consists in five main steps that can be performed sequentially or independently: i) reconstruction of metabolic networks for all annotated genomes (step 0 in Figure 5.1), ii) computation of individual (step 1) and iii) collective metabolic capabilities (step 2), iv) calculation of the cooperation potential

(step 3) and v) identification of minimal communities and key species for a targeted set of compounds (step 4).

The inputs for the method are a set of annotated genomes, a list of nutrients representing a growth medium, and optionally a list of targeted compounds to be produced by selected communities that will bypass the default objective of ensuring the producibility of the cooperation potential. Users can use the annotation pipeline of their choice before running Metage2Metabo.

The outputs of the methods are (i) metabolites producible individually by each member of the community, (ii) metabolites producible by the community, (iii) metabolites producible only by cooperation, and (iv) minimal communities to produce specific metabolites (either targets or cooperation potential). Furthermore, it can create a visual representation of the minimal communities.

### 5.2.1 Step 0: Large-scale reconstruction of draft GSMN

Metage2Metabo was applied on 1,520 reference genomes from the human gut microbiota (Zou et al. 2019).

#### Step description

Metage2Metabo can process existing metabolic networks in SBML format or propose the automatic reconstruction of non-curated metabolic networks from genomes (Figure 5.1 step 0). As a multiprocessing solution, it facilitates the treatment of hundreds or thousands of genomes that can be retrieved from metagenomic experiments. The underlying GSMN reconstruction software is Pathway Tools (Karp et al. 2002a; Karp et al. 2021), a graphical user interface (GUI) based software suite for the generation of individual GSMNs, called Pathway/Genome Databases (PGDBs). Typically, a PGDB is obtained from an annotated genome using PathoLogic (Karp et al. 2011), the software prediction component of Pathway Tools, and curated afterwards.

We developed mpwt (Multiprocessing Pathway Tools), a multiprocessing wrapper for Pathway Tools. First, mpwt reads the input genomes (Genbank, Generic Feature Format (GFF) or PathoLogic format) to create the input files needed by PathoLogic (Pathway Tools prediction algorithm). The genomes must contain functional annotations (such as GO terms or EC numbers) necessary for Pathway Tools. Then mpwt runs the PathoLogic process in multiprocess. It is possible to use the Transport Inference Parser (Lee et al.

2008), operon predictor (Romero et al. 2004), Hole-Filler to retrieve missing enzymes (Michelle L. Green et al. 2004).

Then Metage2Metabo extracts and converts the resulting PGDB in SBML (Hucka et al. 2003; Hucka et al. 2019) formats using the PADMet library (Aite et al. 2018).

#### Application to the human gut microbiota

For the human gut microbiota (Zou et al. 2019), the 1,520 draft GSMNs were reconstructed in 155 minutes on a cluster with 72 CPUs and 144Gb of memory. In average they contain 1144 reactions and 1366 metabolites (Table 5.1).

| | Gut dataset |
|---|---|
| All reactions | 3932 |
| All metabolites | 4001 |
| Average reactions per GSMN | 1144 (±255) |
| Avg metabolites per GSMN | 1366 (±262) |
| Avg genes per GSMN | 596 (±150) |
| Percentage reactions associated to genes | 74.6 (±2.17) |
| Avg pathways per GSMN | 163 (±49) |

Table 5.1 – Statistics on GSMN reconstruction by Metage2Metabo using Pathway Tools.

**Subsection summary**

By using a **parallel implementation** of Pathway Tools, Metage2Metabo can reconstruct draft GSMNs from genomes (in GenBank format) at a large scale.

### 5.2.2 Step 1: GSMN individual production

#### Step description

With the reconstructed GSMNs by the previous steps or input GSMNs given by the user, Metage2Metabo will search for the individual production of each GSMN (Figure 5.1 step 1). This step is performed with the network expansion algorithm (Ebenhöh et al. 2004) implemented in MeneTools (Aite et al. 2018).

The network expansion algorithm computes the *scope* of a metabolic network from a description of the growth medium called *seeds*. The scope consists of the set of metabolic

compounds which are reachable or producible, according to a boolean abstraction of the network dynamics assuming that cycles cannot be self-activated. First, there is an initiation step with a set of seed nutrients. Then the algorithm recursively considers products of reactions to be producible if all reactants of the reactions are producible or among the seeds. The underlying implementation of the network expansion algorithm used in Metage2Metabo relies on Answer Set Programming (ASP) (Schaub et al. 2009) and its implementation in MeneTools (Aite et al. 2018).

We define a metabolic network as a bipartite graph $G = (R \cup M, E)$, where $R$ and $M$ stand for reaction and metabolite nodes. When $(m, r) \in E$ (respectively $(r, m) \in E$), with $m \in M$ and $r \in R$, the metabolite is called a *reactant* (respectively *product*) of the reaction $r$. The scope of a set of seed compounds $S$ according to a metabolic network $G$, denoted by $\text{scope}(G, S)$, is iteratively computed until it reaches a fixed point (Handorf et al. 2005). It is formally defined by:

$$\text{scope}(G, S) = \bigcup_i M_i, \text{ where } M_0 = S \text{ and } M_{i+1} = M_i \cup \mathit{products}(\{r \in R \mid \mathit{reactants}(r) \subseteq M_i\}).$$

Metage2Metabo predicts the set of reachable metabolites for each GSMN using the network expansion algorithm and the given nutrients as seeds.

### Application to human gut microbiota

The given nutrients were composed of 93 metabolites of a European average classical diet from the VMH (Virtual Metabolic Human) resource (Noronha et al. 2018), and a small number of currency metabolites (Schilling et al. 2000). The average scope of the GSMNs was 286 ±70, and the union contained 828 metabolites (among them the 93 seeds).

**Subsection summary**

For each draft GSMN, an **individual scope** (the set of producible metabolites computed from a group of seed metabolites) is computed using the network scope expansion implemented in ASP.

### 5.2.3 Step 2 and 3: Community production

#### Step description

Then Metage2Metabo computes the metabolic capabilities of the whole microbiota by taking into account the complementarity between GSMNs (Figure 5.1 step 2). This step simulates the sharing of metabolic biosynthesis through a meta-organism composed of all GSMNs and assesses the metabolic compounds that can be reached using network expansion. This calculation is an extension of the features of MiSCoTo (Frioux et al. 2018a) in which the community scope of a collection of metabolic networks $\{G_1, \dots G_N\}$ is introduced. Metage2Metabo relies on the mixed-bag modelling implemented in MiSCoTo. This modelling considers the community as a boundary-free meta-organism. So the metabolism of organisms is studied in a virtual compartment containing all the community members. Exchanges can be performed between the organisms without any costs. This method allows a scalable computation of the community scope for large-scale communities.

In this way, the community scope is defined as:

$$\text{collectiveScope}(G_1..G_N, S) = \text{scope}\left(\left(\bigcup_{i\in\{1..n\}} R_i, \bigcup_{i\in\{1..n\}} M_i, \bigcup_{i\in\{1..n\}} E_i\right), S\right).$$

**Cooperation potential.** Given individual and community metabolic potentials, the *cooperation potential* consists in the set of metabolites whose producibility can only occur if several organisms participate in the biosynthesis (Figure 5.1 step 3). Metage2Metabo computes the cooperation potential by performing a set difference between the community scope and the union of individual scopes and produces an SBML file with the resulting metabolites. This list of compounds is inclusive and could comprise false positives not necessitating cooperation for production but selected due to missing annotations in the initial genomes. One can modify the SBML file before the following Metage2Metabo community reduction step.

The $\text{cooperationPotential}(G_1, .., G_n, S)$ of a collection of metabolic networks $\{G_1..G_n\}$ is defined by:

$$\text{cooperationPotential}(G_1, .., G_n, S) = \text{collectiveScope}(G_1, .., G_n, S) \setminus \bigcup_{i\in\{1..n\}} \text{scope}(G_i, S).$$

#### Application to human gut microbiota

The community scope was composed of 984 metabolites. Among them, 156 were only producible by cooperation between members of the community.

In the 156 metabolites, we identified 6 groups of metabolites: sugar derivatives (58 metabolites), lipids (28 metabolites), carboxy acids (14 metabolites), aromatic compounds (11 metabolites), coA derivatives (10 metabolites) and amino acids and derivatives (5 metabolites).

**Subsection summary**

Taking into account **all the GSMN in the community**, Metage2Metabo computes the community's producible compounds, called **community scope**. Then it separates metabolites present in the community scope and the union of the individual scopes from metabolites needing the **cooperation** between community members. The latter is named **cooperation potential** (or also addedvalue).

### 5.2.4 Step 4: Identifying key species in the community

#### Step description

Metage2Metabo provides at the end of the pipeline a set of key species associated with a metabolic function together with one or more minimal communities predicted to satisfy this function. We define as *key species* organisms whose GSMNs are selected in at least one of the minimal communities predicted to fulfil the metabolic objective. Among key species, we distinguish those that occur in every minimal community, suggesting that they possess key functions associated with the objective, from those that occur only in some communities. We call the former *essential symbionts*, and the latter *alternative symbionts*. These terms were inspired by the terminology used in flux variability analysis (Orth et al. 2010) for describing reactions in all optimal flux distributions. If interested, one can compute the enumeration of all minimal communities with *m2m_analysis*, which will provide the total number of minimal communities as well as the composition of each (see following subsection 5.2.5).

Figure 5.2. illustrates these concepts. The initial community is formed of eight species. Four minimal communities satisfy the metabolic objective. Each includes three species; in

particular, the yellow one is systematically a member. Therefore the yellow species is an essential symbiont, whereas the four other species involved in minimal communities constitute the set of alternative symbionts. As key species represent the diversity associated with all minimal communities, their number is likely greater than the size of a minimal community, as this is the case in 5.2.

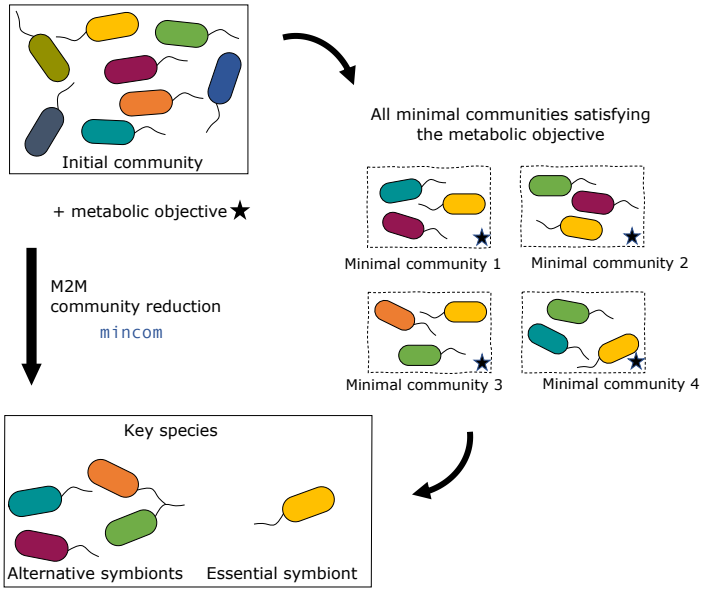



Figure 5.2 – Description of key species. Community reduction performed at step 4 can lead to multiple equivalent communities. Metage2Metabo provides one minimal community and efficiently computes the full set of species that occur in all minimal communities without the need for a full enumeration, thanks to solving heuristics. It is possible to distinguish the species occurring in every minimal community (essential symbionts) from those occurring in some (alternative symbionts). Altogether, these two groups form the key species.

Metage2Metabo computes the minimal communities to produce compounds of interest (Figure 5.1 step 4) by using MiSCoTo (Frioux et al. 2018a) with the mixed-bag modelling. A minimal community $\mathcal{C}$ enabling the producibility of a set of targets $T$ from the seeds $S$ is a sub-family of the community $G_1, \ldots, G_n$ which is the solution of the following optimisation problem:

$$\begin{aligned} &\underset{\{G_{i_1}..G_{i_L}\} \subset \{G_1..G_N\}}{\text{minimize}} && \text{size}(\{G_{i_1}..G_{i_L}\}) \\ &\text{subject to} && T \subset \text{collectiveScope}(G_{i_1}..G_{i_L}), S). \end{aligned}$$

Solutions to this optimisation problem are communities $\mathcal{C} = (G_{i_1} \ldots, G_{i_L})$ of minimal size. We define minimalCommunities$(G_1..G_n, S, T)$ as the set of all such minimal communities. The first output of the `m2m mincom` command is the (minimal) size $L$ of the communities solution to the optimisation problem. The composition of one optimal community is also provided. The targets are, by default, the components of the cooperation potential, $T =$ cooperationPotential$(G_1, .., G_n, S)$, but can also be a group of target metabolites defined by the user.

Many minimal communities are expected to be equivalent for a given metabolic objective, but their enumeration can be computationally costly. We define the *key species*, organisms occurring in at least one community among all the optimal ones. The *essential symbionts*) occur in every minimal community, whereas the *alternative symbionts* occur only in some minimal communities. More precisely, the key species keySpecies$(G_1..G_n, S, T)$, the essential symbionts essentialSymbionts$(G_1..G_n, S, T)$, and the alternative symbionts alternativeSymbionts$(G_1..G_n, S, T)$ associated to a set of metabolic networks, seeds $S$ and a set of target metabolites $T$ are defined by:

$$\text{keySpecies}(G_1..G_n, S, T) = \{G \mid \exists \mathcal{C} \in \text{minimalCommunities}(G_1..G_n, S, T),\ G \in \mathcal{C}\}.$$
$$\text{essentialSymbionts}(G_1..G_n, S, T) = \{G \mid \forall \mathcal{C} \in \text{minimalCommunities}(G_1..G_n, S, T),\ G \in \mathcal{C}\}.$$
$$\text{alternativeSymbionts}(G_1..G_n, S, T) = \text{keySpecies}(G_1..G_n, S, T) \setminus \text{essentialSymbionts}(G_1..G_n, S, T).$$

As a strategy layer over MiSCoTo, Metage2Metabo relies on the Clasp solver (Gebser et al. 2012) for efficient resolution of the underlying grounded ASP instances. Although this type of decision problem is NP-hard (Julien-Laferrière et al. 2016), as with many real-world optimisation problems, worst-case asymptomatic complexity is less informative for applications than practical performance using heuristic methods. The Clasp solver implements a robust collection of heuristics (Gebser et al. 2007; Andres et al. 2012) for core-guided weighted MaxSAT (Manquinho et al. 2009; Morgado et al. 2012). This implementation provides rapid set-based solutions to combinatorial optimisation problems, much like heuristic solvers (such as CPlex) provide rapid numerical solutions to mixed integer programming optimisation problems. The kinds of ASP instances constructed by MiSCoTo for Metage2Metabo are solved in a matter of minutes to identify key species and essential/alternative symbionts. Indeed the space of solutions is efficiently sampled using adequate projection modes in ASP, which enables the computation of these species

groups without the need for a full enumeration by taking advantage of the underlying ASP solver and associated projection modes.

Furthermore, we can analyse the interactions predicted by Metage2Metabo. First, the essential symbionts are symbionts with specific functions not present in any other symbiont of the community. In this way, they perform a role similar to one of the helpers proposed in the Black Queen Hypothesis. Furthermore, as they are providing a function not present in other organisms, they are providing metabolites not available to others, and the interactions between essential symbionts and alternative symbionts can be seen as an auxotrophic interaction. Secondly, the alternative symbionts can be viewed as organisms with redundant functions. Thus, as they perform the same function, they can rely on the same metabolites as input to participate in producing the metabolites of interest. In this way, their interactions may be close to an exploitation competition.

### Application to human gut microbiota

We computed for the 6 groups of metabolites found in the previous subsection the minimal communities producing them. These minimal communities are shown in Table 5.2.

| targets | | Firm. | Bact. | Acti. | Prot. | Fuso. | total |
|---|---|---|---|---|---|---|---|
| **aminoacids and derivatives** (5 targets) | KS | 142 | 52 | 0 | 27 | 6 | 227 |
| 4 bact. per community | ES | 0 | 0 | 0 | 0 | 0 | 0 |
| 120,329 communities | AS | 142 | 52 | 0 | 27 | 6 | 227 |
| **aromatic compounds** (11 targets) | KS | 52 | 0 | 0 | 20 | 0 | 72 |
| 5 bact. per community | ES | 2 | 0 | 0 | 1 | 0 | 3 |
| 950 communities | AS | 50 | 0 | 0 | 19 | 0 | 69 |
| **carboxyacids** (14 targets) | KS | 16 | 13 | 0 | 28 | 2 | 59 |
| 9 bact. per community | ES | 2 | 0 | 0 | 2 | 0 | 4 |
| 48,412 communities | AS | 14 | 13 | 0 | 26 | 2 | 55 |
| **coA derivatives** (10 targets) | KS | 106 | 0 | 50 | 17 | 1 | 174 |
| 5 bact. per community | ES | 0 | 0 | 0 | 0 | 1 | 1 |
| 95,256 communities | AS | 106 | 0 | 50 | 17 | 0 | 173 |
| **lipids** (28 targets) | KS | 3 | 140 | 22 | 20 | 0 | 185 |
| 7 bact. per community | ES | 3 | 0 | 0 | 1 | 0 | 4 |
| 58,520 communities | AS | 0 | 140 | 22 | 19 | 0 | 181 |
| **sugar derivatives** (58 targets) | KS | 11 | 30 | 78 | 23 | 0 | 142 |
| 11 bact. per community | ES | 5 | 0 | 0 | 0 | 0 | 5 |
| 7,860,528 communities | AS | 6 | 30 | 78 | 23 | 0 | 137 |

Table 5.2 – Community reduction analysis of the target categories in the gut. All minimal communities were enumerated, starting from the set of 1,520 GSMNs. KS: key species, ES: essential symbionts, AS: alternative symbionts, Firm.: Firmicutes, Bact.: Bacteroidetes, Acti.: Actinobacteria, Prot.: Proteobacteria, Fuso.: Fusobacteria.

Some metabolite targets were associated with many minimal communities (7,860,528 for the sugar derivatives) and composed of different phyla. To understand the interactions between these organisms, we need a visualisation method to represent such results.

**Subsection summary**

From a set of seed metabolites, Metage2Metabo computes the **minimal communities** that can produce either selected targets or the cooperation potential. Among the member of the minimal communities, called **key species**, the tool will identify two groups. The first group contains organisms occurring in all minimal communities; they are called **essential symbionts**. The second group contains organisms occurring in at least one minimal community but not in all minimal communities and are called **alternative symbionts**.

### 5.2.5 Visualisation of minimal communities

To represent the minimal communities, we first enumerated all the minimal communities producing a given group of targets. The number of optimal solutions is large, reaching more than 7 million equivalent minimal communities producing the sugar-derived metabolites presented in the previous subsection 5.2.4. Our analysis of key species indicates that many optimal communities are due to combinatorial choices among a relatively small number of Bacteria (137 organisms for the sugar-derived metabolites).

To visualise the association of GSMNs in minimal communities, we created a graph for each target whose nodes are the key species and whose edges represent the association between two species if they co-occur in at least one of the enumerated communities. Graphs were very dense, for example, 227 nodes and 8772 edges for the amino acids and derivatives (left Figure 5.3). This density is expected given the large number of optimal communities and the small number of key species.

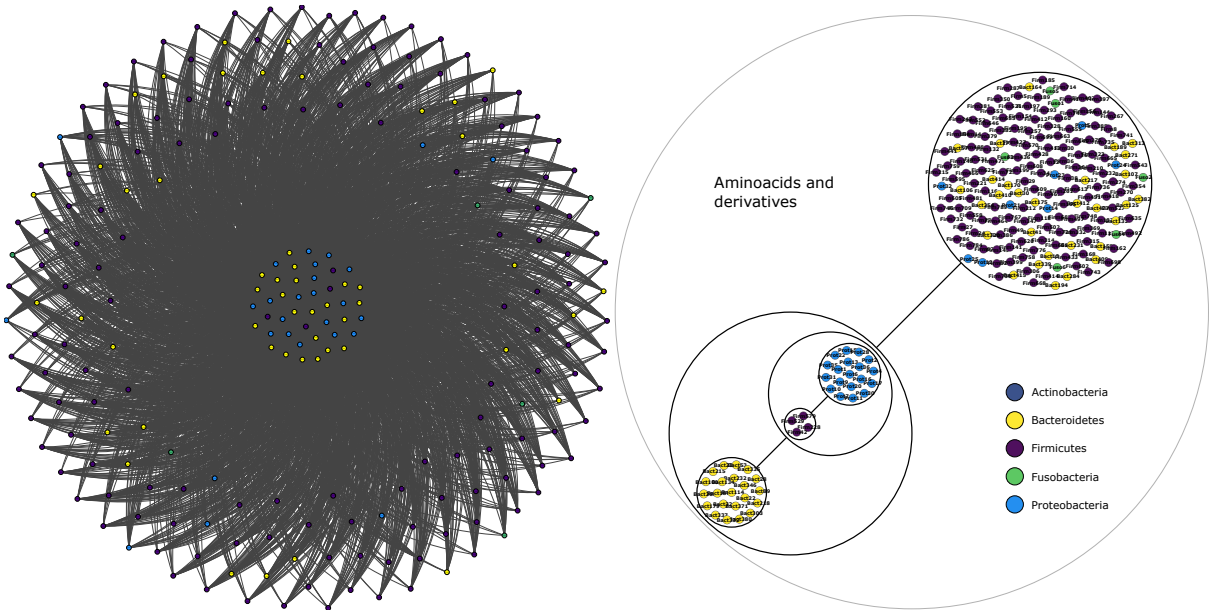



Figure 5.3 – At the left, the solution graph shows the minimal communities for the amino acids and derivatives targets. Nodes represent the organism in the minimal community. Edges are drawn between nodes in the same minimal communities. Right power graph compressing the information of the previous graph.

To allow the readability of these graphs, they were compressed into power graphs to capture the combinatorics of association within minimal communities by using Power-GrASP (Bourneuf et al. 2017). Power graphs enable a lossless compression of re-occurring

motifs within a graph: cliques, bicliques and star patterns (Royer et al. 2008). The increased readability of power graphs permits pinpointing metabolic equivalency between members of the key species for the target compound families. These equivalencies can be seen in Figure 5.3 (corresponding power graph at the right).

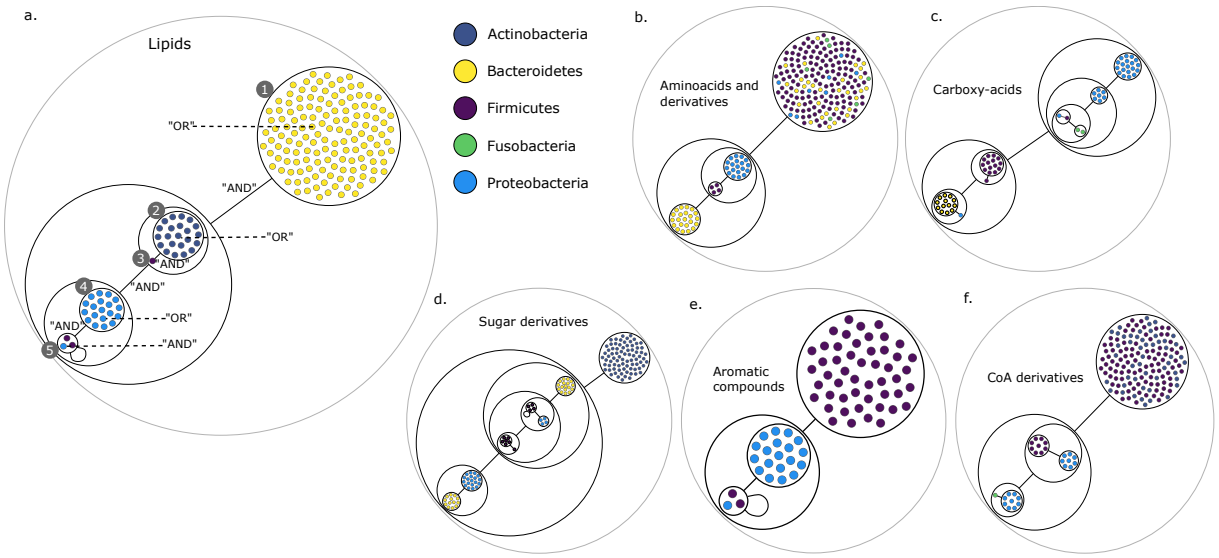



Figure 5.4 – **Power graph analysis of predicted microbial associations within communities for the human gut dataset.** Each category of metabolites predicted as newly producible in the gut was defined as a target set for community selection among the 1,520 GSMNs from the human gut dataset. Key species and the full enumeration of all minimal communities were computed for each metabolic group. Association graphs were built to associate members that are found together in at least one minimal community among the enumeration. These graphs were compressed as power graphs to identify patterns of associations and groups of equivalence within key species. Power graphs a., b., c., d., e., f., g. were generated for the sets of lipids, amino acids and derivatives, carboxy-acids, sugar derivatives, aromatic compounds, and coenzyme A derivative compounds respectively. Node colour describes the phylum associated with the GSMN. Figure a. has an additional description to ease readability. Edges symbolise conjunctions ("AND"), and the co-occurrences of nodes in regular power nodes (as in power node 1, 2, 4) symbolise disjunctions ("OR") related to alternative symbionts. Power nodes with a loop (e.g. power node 5) indicate conjunctions. Therefore, each enumerated minimal community for lipid production is composed of the two Firmicutes and the Proteobacteria from power node 5, the Firmicutes node 3 (the four of them being the essential symbionts), and one Proteobacteria from power node 4, one Actinobacteria from power node 2 and 1 Bacteroidetes from power node 1. Members from an inner power node are interchangeable with respect to the metabolic objective. The figures display one visual representation for each power graph, although such representations are not unique. The number of power edges is minimal, which leads to the nesting of (power) nodes.

Figure 5.4 presents the compressed graphs for each set of targets. Graph nodes are the key species, coloured by their phylum. Nodes are included in power nodes that are connected by power edges, illustrating the redundant metabolic function(s) that species provide to the community when considering particular end-products. GSMNs belonging to a power node play the same role in constructing the minimal communities. In this visualisation, essential symbionts are easily identifiable, either into power nodes with loops (Figure 5.4 a, e) or as individual nodes connected to power nodes (Figure 5.4 a, c, d, f).

We observe that power nodes often contain GSMNs from the same phylum, indicating that phylogenetic groups encode redundant functions. Figure 5.4 *a* has additional comments to guide the reader into analysing the community composition on one example. Each minimal community suitable for the production of the targeted lipids is composed of one Bacteroidetes from power node (PN) 1, one Actinobacteria from PN 2, the Firmicutes member 3, one Proteobacteria from PN 4 and finally the two Firmicutes and the Proteobacteria from PN 5. For all the target groups of this study, the large enumerations can be summarised with a boolean formula derived from the graph compressions. For instance, for the lipids of Figure 5.4 a, the community composition as described above is the following:

$$(\vee PN1) \wedge (\vee PN2) \wedge (PN3) \wedge (\vee PN4) \wedge (\wedge PN5).$$

Altogether, computation of key species coupled with the visualisation of community compositions enables a better understanding of the associations of organisms in the minimal communities. In this genome collection, groups of equivalent GSMNs allow us to identify genomes that provide specialised functions to the community, enabling metabolic pathways leading to specific end-products.

**Subsection summary**

From the enumerated minimal communities producing a set of metabolites, Metage2Metabo creates a graph solution which is compressed into a **powergraph** showing all the minimal communities that can produce the targets.

**Section summary**

By combining **multiprocessing** and **ASP formalism** of the network scope expansion, Metage2Metabo is able to identify **key species** by using the **metabolic complementarity** of the member in the community in large-scale community. Furthermore, it is possible to visualise the minimal community producing the compounds of interest using power graphs. The method was applied to a large-scale microbiome of the human gut.

## 5.3 Identification of key taxa in the biogas reactor experiments

### 5.3.1 Metage2Metabo inputs

**Metabolic networks.** The metabolic networks used for this step were the one predicted by EsMeCaTa in the section 4.3.

In subsection 5.3.2, we will use the TSMN created from the default run of EsMeCaTa (subection 4.3.1).

In subsection 5.3.3, we will use the TSMN created from the run of EsMeCaTa with constraints on the selected proteomes and with multiple proteome representativeness ratio (Pan-proteome, Shell-core proteome and Soft-core proteome) in subsection 4.3.2.

**Nutrients.** For the nutrients given to the community, we created a list of metabolites with Patrick Dabert according to the input provided to the biogas reactor.

To represent the pig manure, we used the metabolites from the human diet described in VMH (Noronha et al. 2018) and used in the previous subsection 5.2.2.

For the apple, we added multiple sugars (sucrose, fructose, glucose and sortibol), metabolites associated with the cellulose (hemicellulose, cellulose, glucopyranose, galactopyranose, xyloglucan and arabinoxylan) and pectin (arabinoses, galactoses, rhammoses and xyloses).

For the butter, we added multiple fatty acids (palmitate, oleate, stearic acid, mystiric acid, butyric acid, linoleic acid, linolenic acid, lauric acid, capric acid, caproic acid and caprylic acid).

As a protein, casein provided amino acids (even if most of them were already given by the VMH diet).

We also added two metabolites, Methylenetetrahydromethanopterin and Methanofurans, as both are needed to produce metabolites required for methane. These two metabolites are linked by cycle in the metabolic network: Methylenetetrahydromethanopterin is needed to produce the metabolite THMPT, but THMPT is needed to produce Methylenetetrahydromethanopterin.

Furthermore, we added a set of currency metabolites (NAD, NADP, Acceptor, Donor, etc.) to activate reactions.

These different seeds were applied differently to the GSMN according to what input was given to the biogas reactor during the experiment. For more details see Figure 4.6.

**Targets.** The targets consist of methane and intermediate metabolites of the organic matter degradation (acetate, propionate, succinate, formate, lactate, butyrate, ethanol, dihydrogen, carbon dioxide, and sulfide).

### 5.3.2 Key taxa involved in organic matter degradation

In this first experiment, we used the metabolic network from the default run of EsMeCaTa. For each time point, we selected the OTU if their abundance was greater than 0; otherwise, they were not used in the time point. As explained in the previous subsection 5.3.1, we used the nutrients according to what was given to the biogas reactor for each time point (shown in Figure 4.6). And we used the same targets at each time point. Metage2Metabo was run on each time point, and we extracted the key species found in the minimal communities.

The number and composition of the key species for each time point are represented in Figure 5.5. The abscissa shows the time point, and the ordinate axis indicates the number of key species found in the minimal communities producing all the targets. The colour identifies the taxa among the key species. The absence of key species in a column indicates that no minimal communities could produce all the targets (here, it was the methane that was not produced). To identify the composition and the relations between key species, we also created the power graphs for some specific time points.

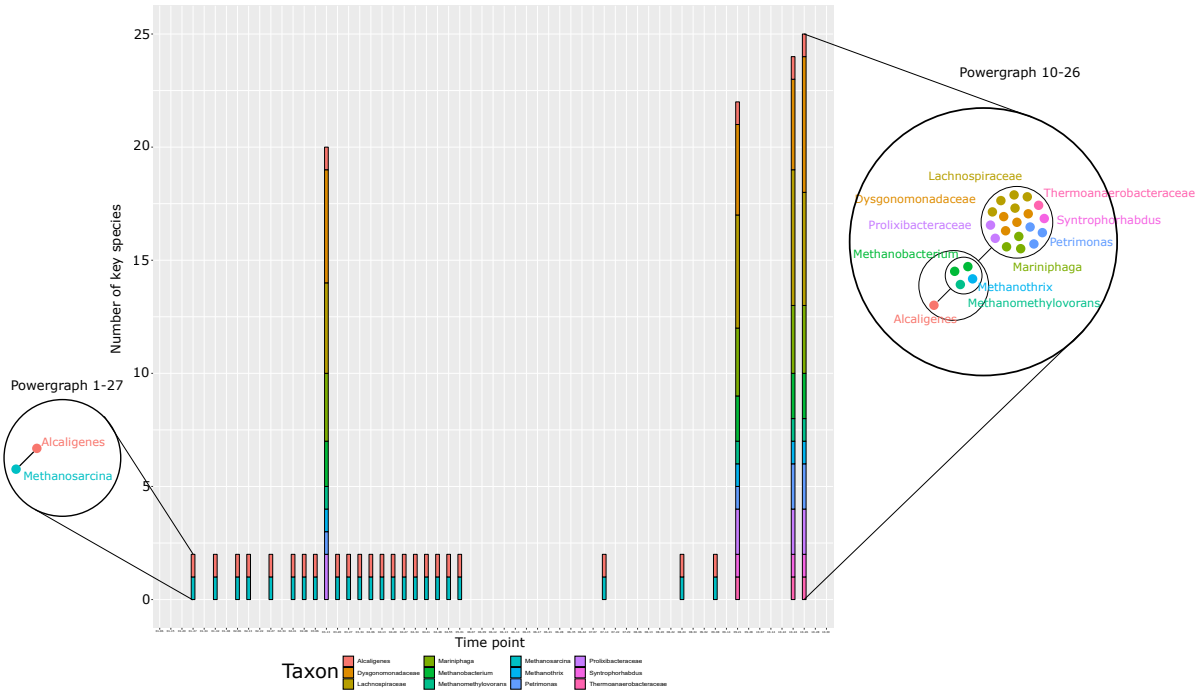



Figure 5.5 – Power graph showing the key species producing the targets for each time point, according to the sequenced OTUs.

An astonishing result was that most of the time points (34 on 61) were unable to produce the targets (especially methane). And among the time points producing the targets, we observed two sets of time points.

The first one (such as the time point 01-27) contained only two key species (*Methanosarcina* and *Alcaligenes*). The corresponding power graph for the time point 01-27 is shown at the left of Figure 5.5. As expected, we can see that the minimal community in the power graph also consists of only two members.

Minimal communities for the second group of time points (such as time point 10-28) were more diverse as they included around 20 key species. A striking feature was that these time points also contained *Alcaligenes* but were missing *Methanosarcina*. We showed the power graph for 10-28 at the right of Figure 5.5. The community was more complex than the one found for time point 01-27 as we had 3 members in each minimal community. Indeed, to produce all the targets, Metage2Metabo needed the following:

— (1) the OTU associated with *Alcaligenes*.

— (2) one OTU associated with methanogenic Archaea among: *Methanobacterium*, *Methanothrix* or *Methanomethylovorans*.

— (3) one OTU among: *Prolixibacteraceae*, *Dysgonomonadaceae*, *Thermoanaerobacteraceae*, *Syntrophorhabdus*, *Petrimonas* or *Mariniphaga*.

We first observed that all community producing methane and the other targets contained at least one Bacteria (*Alcaligenes*, *Prolixibacteraceae*, *Dysgonomonadaceae*, *Thermoanaerobacteraceae*, *Syntrophorhabdus*, *Petrimonas* or *Mariniphaga*) and one methanogenic Archaea (*Methanosarcina*, *Methanobacterium*, *Methanothrix* or *Methanomethylovorans*). But this was expected as methane production is performed by methanogenic Archaea using products of organic matter decomposition performed by Bacteria.

A second result was the importance of the OTU *Alcaligenes*. Indeed, in all minimal communities, *Alcaligenes* were an essential symbiont meaning that without them, it was impossible to produce the targets. One possible explanation was that EsMeCaTa associated this OTU with only one proteome. With one proteome, the reconstructed TSMN contained all the metabolic functions of this proteome. Indeed the TSMN contained 1,550 reactions, whereas the mean for the 587 OTUs was 598 reactions.

A third interesting result was that the absence of *Methanosarcina* led to more complex minimal communities (with 3 members). This observation could be explained by the fact that some functions performed by *Methanosarcina* were not all present in any other OTU and thus needed multiple OTUs. By examining the TSMN of *Methanosarcina*, this was supported by the size of the metabolic networks (916 reactions). But some of the methanogenic Archaea used in the 3 member minimal communities also had TSMN bigger than the mean (for example *Methanobacterium* with 734 reactions). *Methanosarcina* can perform acetoclastic and hydrogenotrophic methanogenesis (Thauer et al. 2008) whereas most other methanogenic Archaea perform only one.

As we had seen, *Alcaligenes* was an essential symbiont for methane production. This result raised several questions as it was not a Bacteria known to interact with methanogen Archaea. Further analysis revealed that its metabolic capacities came from a bias in EsMeCaTa. But multiple questions arose because other interactions between Bacteria and Archaea could produce methane and the other targets. Furthermore, it limited the number of time points producing targets which seemed quite different from the known production and state of the biogas reactor. So we explored how the limitation observed can be explained by different points.

These results raised several questions. First, some OTUs were associated with only one proteome, meaning they contain all the proteins in the proteome, thus overestimating

the OTU capacities compared to other OTUs. It is the case of the *Alcaligenes*. Similarly, EsMeCaTa associated with the OTU *Methanosarcina* a TSMN showing more reactions than other methanogenic Archaea. In both cases, the higher capacities associated with those OTUs led them to be preferentially selected in minimal communities. Whereas this is biologically relevant in the case of *Methanosarcina*, this is not the case of *Alcaligenes*. This result suggests a need to select more than one proteome to infer a TSMN. After a closer examination of the inferred TSMNs and the minimal communities metabolisms, we observed the second issue: some cofactors added to the seeds were directly used to produce metabolites. We discovered that cofactors were used to produce metabolites. This discovery was not expected as cofactors were added to participate in the production of metabolites by being combined with other metabolites. One cause of this issue was the use of the NAD/NADP degradation pathways to produce multiple metabolites. For example, 20 OTUs with only NAD, NADP, ATP and ADP as seeds had an individual scope superior to 200. These cofactors should not be available to produce these metabolites. They should only be usable by reactions where they are used as cofactors. Finally, the incapacity to generate minimal communities for most time points could be due to the relatively low average of 598 selected reactions per Soft-core TSMNs (proteome representativeness ratio of $Pr = 0.95$). The following section examines solutions to resolve these issues.

**Subsection summary**

Using the metabolic networks reconstructed with the default option of EsMeCaTa, we found some minimal communities able to produce methane. All contained interaction between Bacteria and methanogenic Archaea. For most time points, however, no community was able to produce methane. Several issues were identified: (1) OTU associated with one proteome, (2) cofactors used to produce targets and (3) proteome representativeness ratio too stringent.

### 5.3.3 Changes in metabolic complementarity according to EsMeCaTa parameters

#### Handling issues

To answer the issues found in the previous subsection 5.3.1 we used the metabolic networks created by using constraints on EsMeCaTa proteome selections through its parameters. First, the minimal number of proteomes associated with a taxon was set to at least 5 proteomes. Secondly, the maximal taxonomic rank was set to the family, and OTUs with higher ranks were ignored. The assumptions behind these constraints were that if we did not have enough information (here proteomes) for a taxon, it was safer not to use it. Third, five proteome representativeness ratio were considered ($Pr = 0.0$, $Pr = 0.25$, $Pr = 0.5$, $Pr = 0.75$ and $Pr = 0.95$). Behind these ratios, different hypotheses were assessed (see Chapter 4).

For the proteome representativeness ratio at $Pr = 0.95$ (Soft-Core metabolism), the assumption was that key functions of a taxon are coded by significantly shared genes among the organisms of a taxon. Thus, we selected fewer number proteins than with the other thresholds. These functions could be described as the most likely present functions in the uncultivated organism associated with the 16S rRNA gene sequences.

For the proteome representativeness ratios at $Pr = 0.25$, $Pr = 0.5$ and $Pr = 0.75$, as the ratio decreased compared to the Soft-Core metabolism, we increased the number of selected protein clusters (see Chapter 4, Figures 4.10, 4.11 and explanations). Besides the protein clusters also found in the Soft-Core metabolism, this allows for the consideration of clusters in which proteins are less represented among the taxon proteomes.

The last proteome representativeness ratio at $Pr = 0$ (Pan-metabolism) contained all the protein clusters identified within the taxon proteomes. This ratio represents all the possible functions yet identified in the proteomes of the taxon. The TSMN associated with this pan-metabolism of the taxon thus fits the following assumption: all its reactions were observed at least once in the taxon organisms. Thus these reactions are considered as possibly achieved by the wild organisms. Conclusions provided using these TSMNs should be regarded with caution, as any organism is unlikely to perform all the reactions observed in its taxon. In the context of a biogas reactor (with the known seeds and targets), the idea was to identify taxa having metabolic functions that could be involved in this metabolic process.

Finally, to solve the issue with the cofactors, we used the method implemented in

moped (Saadat et al. 2022) to handle cofactors. This method searches for reactions involving pair of cofactors (such as NAD/NADH). It will create new metabolites (with new identifiers) corresponding to the cofactors (with ___cof___ in their names) and new reactions from the reaction in which the cofactors are involved. For example, for the previous NAD/NADH cofactors, NAD___cof___ NADH___cof___ are created, and new reactions involving these two metabolites are also created. Then these metabolites will be added to the seeds. In this way, they will be explicitly used by the reaction involving the pair of cofactors, but they will not be used in other reactions (such as in the NAD/NADP pathway degradation), avoiding the issue identified in the previous subsection 5.3.2.

### Application to the biogas reactor experiment

In the following experiments, cofactors metabolites created by moped were added to the seeds (the nutrients). The targets to be reached were the same as in the previous section (methane, acetate, propionate, succinate, formate, lactate, butyrate, ethanol, dihydrogen, carbon dioxide, and sulfide). EsMeCaTa was run applying the parameters mentioned above (minimal number of proteome per taxon: 5, maximal taxonomic rank considered: family and with 5 proteome representativeness ratios). As in the subsection 5.3.2, for a given time point, we selected the TSMNs to be used as input to Metage2Metabo when their OTU abundances were superior to 0.

For the first experiment, we analysed the metabolic networks obtained with the same proteome representativeness ratio (Core-proteome at 0.95) as used in the previous subsection 5.3.2 experiments. Let's first consider the minimal communities found by Metage2Metabo for each time of point (Figure 5.6).

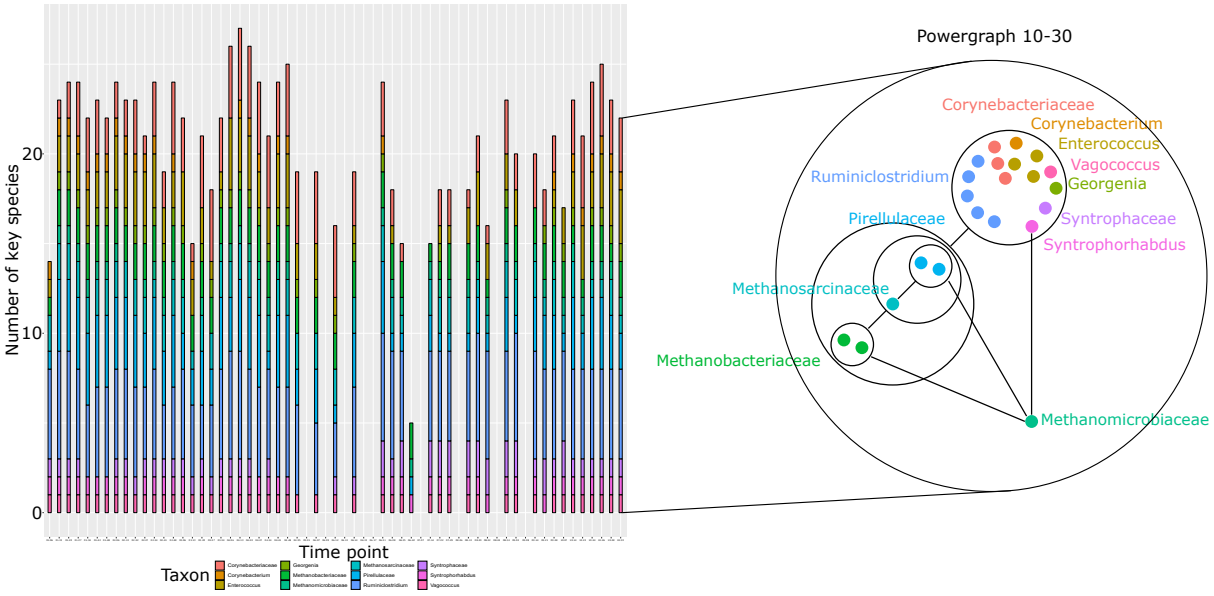



Figure 5.6 – Alluvial plot showing the key species for all the minimal communities predicted according to the targets for the experimental steps. The TSMNs corresponding to the Soft-core metabolisms are considered (EsMeCaTa proteome representativeness ratio $Pr = 0.95$). Minimal communities produce the targets methane, acetate, propionate, succinate, formate, lactate, butyrate, ethanol, dihydrogen, carbon dioxide, and sulfide. Power graph showing the key species for producing the targets from the community.

When looking at the key species of minimal communities in Figure 5.6, we observed that, for 52 of the 61 time points, the TSMNs inferred from the OTUs led to estimate minimal communities that can produce the methane and the other targets. Only 9 time points were unable to produce the targets compared to the 34 time points in subsection 5.3.2. Most of the time points were associated with more than 20 key species.

To illustrate the minimal communities, we created the power graph for the time point 10-30 (Figure 5.6). The minimal communities were composed of 4 members, and there were no essential symbionts. The enumerated minimal communities were composed of:

— (1) either *Methanosarcinaceae* or *Methanomicrobiaceae*. The only bacteria of group 4 associated with *Methanomicrobiaceae* is *Syntrophorhabdus*.
— (2) one of the *Pirellulaceae*.
— (3) one of the *Methanobacteriaceae*.
— (4) one Bacteria among: *Ruminiclostridium*, *Corynebacteriaceae*, *Corynebacterium*, *Enterococcus*, *Vagococcus*, *Georgenia*, *Syntrophaceae* or *Syntrophorhabdus*.

Thus, after applying new parameters and solving issues with the cofactors, the obtained

minimal communities were more complex than the one inferred in the previous section 5.3.2. Interestingly they still contained methanogenic Archaea (see above, members 1 and 3) associated with groups of Bacteria (2 and 4). The Bacteria of group 4 contained genus known to be present and important in a biogas reactor (such as *Ruminiclostridium*, *Enterococcus* and *Syntrophorhabdus* (Lim et al. 2020)).

Finally, to further investigate why no minimal communities producing methane can be found for nine time points, we examined the production of biogas and the reactor state according to the time point (Figure 4.6). Among those nine time points, five corresponded to time points where the biogas reactor was described as non-functional (6-09, 6-13, 6-15, 6-17, 8-24). Two time points (7-07, 8-06) corresponded to a time point where the biogas reactor was described as functional but with low biogas production. The two remaining time points (5-01, 9-08) were associated with a functional reactor producing more than 30 NL.j-1 of biogas. Thus of seven over nine time points where we inferred no methane production, no or low production of methane was obtained from the reactor.

**Impact of the proteome representativeness ratios**

In the following experiment, we examined the impact of the considered representativeness ratios on the obtained TSMNs and the obtained minimal communities. The other EsMeCaTa parameters, the seeds and targets are the same as in the previous experiment. We considered a hypothetical set of OTUs, which was not observed at any of the time points but corresponds to all of the 362 TSMNs that could be inferred while applying the EsMeCaTa parameters considered in this section (minimal number of proteome per taxon: 5, maximal taxonomic rank considered: family). Given each of the five proteome representativeness ratios considered, Metage2Metabo inferred minimal communities able to produce all the targets, including methane. The obtained minimal communities were enumerated, and their compositions are illustrated using power graphs (Figure 5.7). All these hypothetical communities were able to produce the targets, including methane. For all the power graphs, each minimal community contained at least one methanogenic Archaea (either *Methanomicrobiaceae*, or *Methanobacteriaceae* or *Methanosarcinaceae*) and a Bacteria.

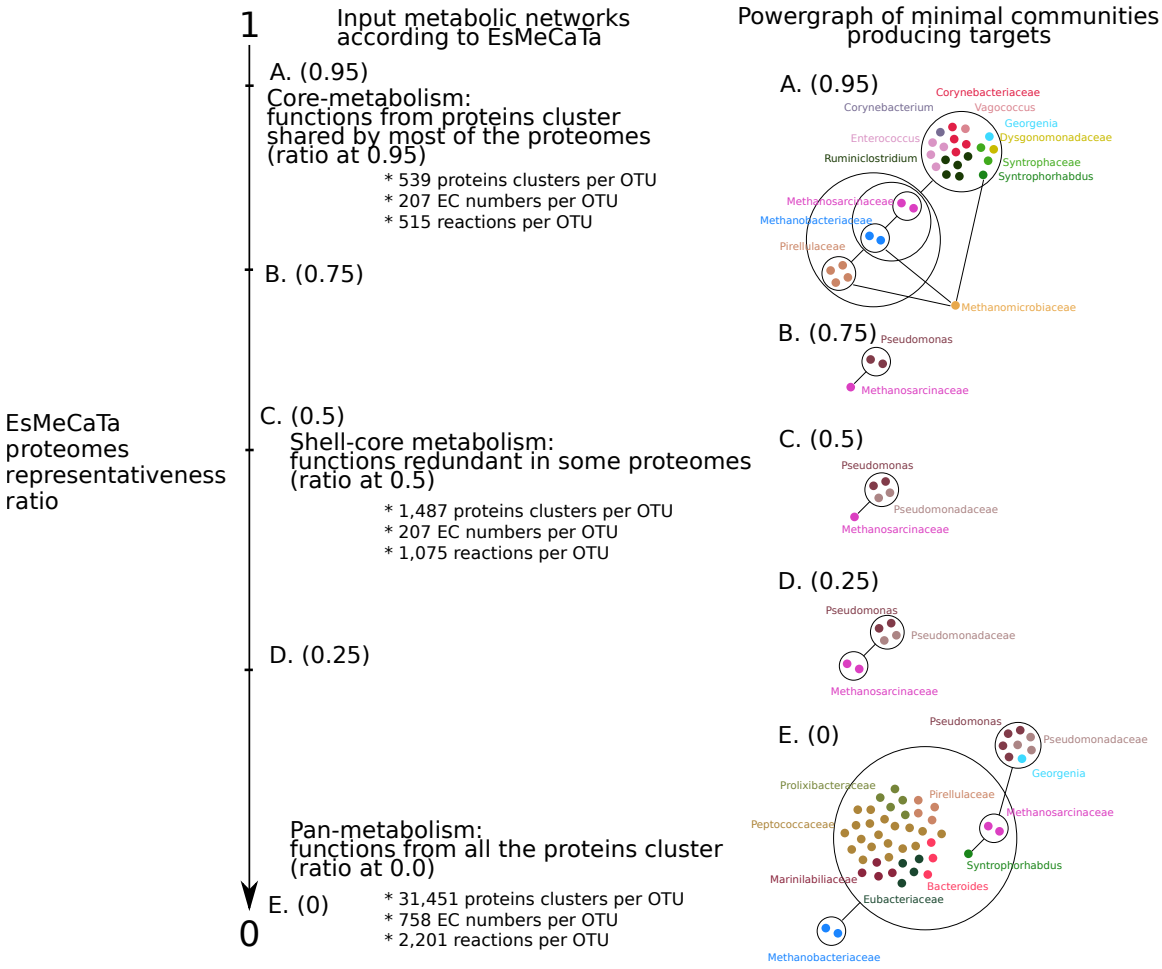



Figure 5.7 – Inferred minimal communities according to the proteome representativeness ratios applied to recover TSMNs. According to each of the five ratios ($Pr = 0$, $Pr = 0.25$, $Pr = 0.5$, $Pr = 0.75$ and $Pr = 0.95$) and to the other applied parameters (minimal number of proteome per taxon: 5, maximal taxonomic rank considered: family) EsMeCaTa reconstructed 362 TSMNs. These TSMNs were all provided as input to Metage2Metabo to infer minimal communities producing methane and several other targets. The power graphs showing the minimal community compositions are shown on the right. Label A, B, C, D and E correspond to the results obtained given the ratio applied by EsMeCaTa, respectively $Pr = 0.95$ (Soft-core proteome), $Pr = 0.75$, $Pr = 0.5$ (Shell-core proteome), $Pr = 0.25$ and $Pr = 0$ (Pan-proteome).

More key species were inferred by considering the Soft-core metabolisms ($Pr = 0.95$, 29 key species, Figures 5.7 A) and the Pan-metabolisms ($Pr = 0$, 52 key species, figure 5.7 E) then by considering TSMNs inferred using intermediary proteome representativeness ratios ($Pr$ in [0.25, 0.5, 0.75], 3 to 6 key species, Figures 5.7 B, C and D).

To investigate these results, we computed the minimal communities for each time point of the experiment. We took only the OTU with an abundance superior to 0. Then we computed the individual scope, the community scope, the cooperation potential and the minimal communities than can produce the targets.

The community scope increased as the proteome representativeness ratio decreased (Figure 5.8). This result is congruent with the conclusion of Subsection 4.3.2, where we showed that the size of TSMNs increased with the decrease of the proteome representativeness ratio. As more reactions are present in larger TSMNs, the number of metabolites produced is also greater. The community scope can be split into two parts: the metabolites produced by individual organisms and the metabolites that require cooperation between organisms to be produced. As it can be seen in Figure 5.8, the dynamic of the community scope according to the proteome representativeness ratio is mainly explained by the individual scope.

Indeed, the distributions of the individual scope sizes followed those of the community scope sizes, whereas the sizes of the cooperation potentials appeared much more stable, with 250 metabolites. The Soft-Core metabolism TSMNs produced individually fewer metabolites than the Pan-metabolism TSMNs.

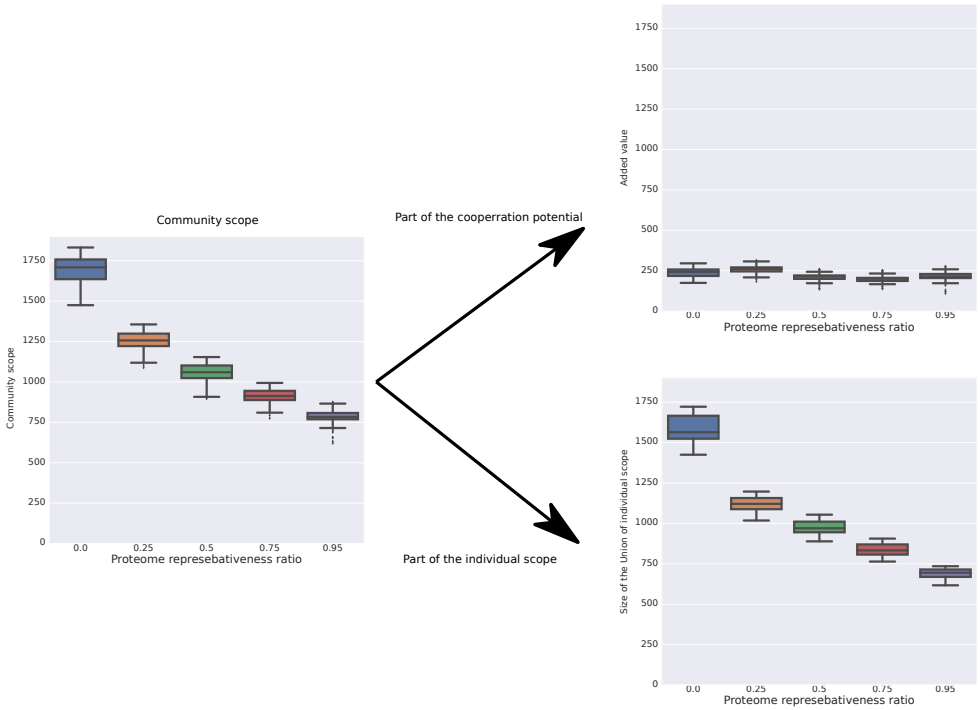



Figure 5.8 – Distribution of the community scope sizes over the 61 time points using five proteome representativeness ratios. The community scope (figure left) is composed of the union of the individual scope (figure bottom right) and the cooperation potential (figure upper right).

Most of the experiments' communities could produce the 12 targets (Figure 5.9 upper left). There were only some experiments with the proteome representativeness ratio of 0.95 where methane was not producible. The minimal community's size decreased with the proteome representativeness ratio (Figure 5.9 upper right).

This analysis confirmed the proposition according to the results presented in Figure 5.7. The TSMNs associated with Pan-metabolism required less cooperation and thus could constitute minimal communities of smaller sizes. And the Soft-core metabolism TSMNs required more cooperation and could constitute minimal communities of larger sizes to achieve a given target production.

The minimal communities inferred from Soft-core metabolisms were of the size of 4, higher than the ones of the other ratios, which were of size 2 (Figure 5.9). Minimal communities built with Pan-metabolisms seemed to require less taxa (2 in figure 5.7 F) than the minimal communities built with Soft-core metabolism (4 in figure 5.7 A). Because

Pan-metabolism TSMNs produced more metabolites individually, they required less cooperation. Reciprocally, because Soft-core metabolism TSMNs produced fewer metabolites individually, they required more cooperation.

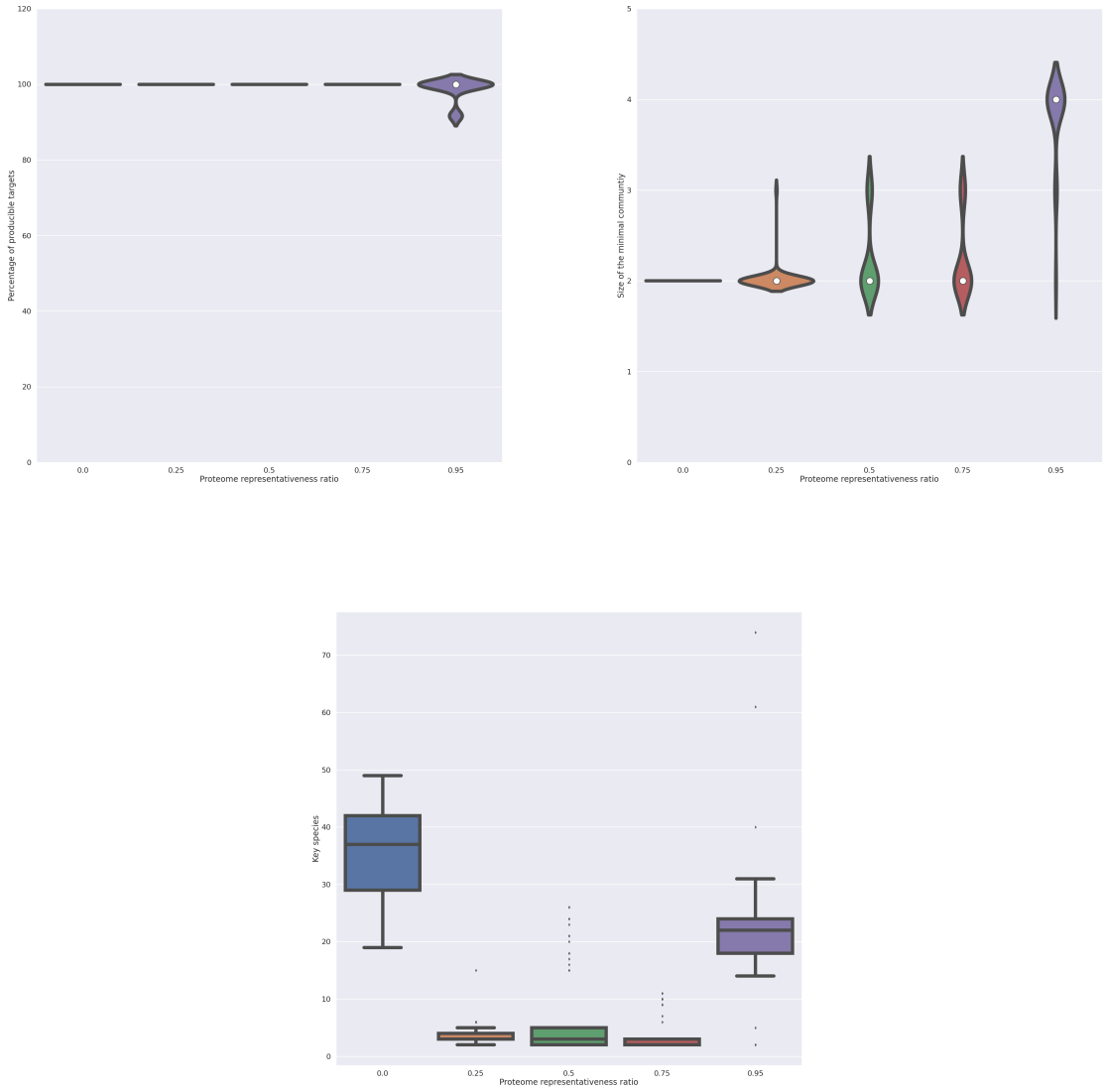



Figure 5.9 – Distributions over the 61 time points of the per cent of producible target metabolites (upper left), the minimal community sizes (upper right) and the key species numbers (bottom), according to varying proteome representativeness ratios.

The number of key species involved in the minimal communities was higher when these were inferred from Pan-metabolism ($Pr = 0$) TSMNs than from Soft-Core metabolism

($Pr = 0.95$) TSMNs. This was expected, as we have shown before; the TSMNs associated with Pan-metabolism contained more reactions and produced individually more metabolites. In this way, fewer TSMNs were required to produce the target metabolite. This could be explained by the fact that these TSMNs contained metabolic functions from all the proteomes in the taxon. In contrast, the TSMNs associated with Soft-core metabolism contained fewer reactions and produced fewer metabolites individually. Then to produce the same target metabolites, they could require more taxa.

The number of key species was low for the intermediate proteome representativeness ratios. One explanation could be that the taxa identified in the minimal communities with Soft-core metabolisms were too specialised and thus required more cooperation. But with the decrease in the proteome representativeness ratio, some taxa with more generalised metabolic functions were retrieved. Among them, a specific subset (*Pseudomonas* and *Methanosarcinaceae* achieved the production of all the targets. And when we looked at the Pan-metabolism, then, as all metabolic functions were available for all taxa, a broader range of cooperation between taxa was possible (with fewer members in the cooperation). To better understand these results, more investigations are required to decipher the combined effects of the possible EsMeCaTa parameters and Metage2Metabo.

**Subsection summary**

By looking at the multiple levels of metabolism for the taxon (Pan, Shell-core and Soft-core) we were able to identify key species in the methane production for the various time points. All minimal communities producing the targets (among them the methane) contained at least an association between a methanogenic Archaea and a Bacteria. We were able to identify multiple Bacteria that could be of interest in the understanding of the biogas reactor. We finally investigated the impact of the proteome representativeness ratio on the outputs of EsMeCaTa, the taxonomic scale metabolic networks (TSMNs) that can be used by Metage2Metabo to produce metabolites (individually and collectively) and predict minimal communities achieving targeted productions.

**Section summary**

By combining EsMeCaTa and Metage2Metabo, we identified multiple taxa that could be of interest in producing methane in a biogas reactor. In all the minimal communities found, we have seen the importance of interactions between methanogenic Archaea and Bacteria. We have also seen the effect of the proteome representativeness ratio of EsMeCaTa on prediction. The Soft-core proteomes were associated with minimal communities of greater sizes involving greater interactions between TSMNs. As Soft-core TSMNs were smaller, more cooperation was needed to achieve the target production. Reciprocally, the Pan-proteome led to the recovery of minimal communities of smaller sizes: as Pan-metabolism TSMNs were bigger, fewer interactions were required to achieve the target production.

## 5.4 Conclusion

### Contribution

In this chapter, I presented Metage2Metabo, a method to predict the metabolic interactions between community members. The method identifies key species producing metabolites of interest using a parallel implementation of Pathway Tools and constraint programming. It is possible to separate essential and alternative symbionts among the key species. This method allows a screening of large-scale communities to identify organisms with metabolic specificities achieving ecosystemic functions of interest. We have presented its predictions on a human gut microbial community.

Furthermore, we have combined the results of EsMeCaTa with the results of Metage2Metabo to study the metabolism of a community sequenced using metabarcoding technology. From all the minimal communities predicted by Metage2Metabo with EsMeCaTa results (with any options), we consistently observed at least a methanogenic Archaea and a Bacteria. This result was expected from a literature point of view as one organism can't perform the production of all the targets alone. By testing multiple EsMeCaTa options, it was possible to obtain meaningful minimal communities and relevant results.

## Limits and improvements

Metage2Metabo focused on searching for minimal communities, but this hypothesis impacts the results as minimisation of the community size selects the generalist organisms, which have more metabolic functions, instead of more specialised organisms. It is especially the case when working with the Pan-metabolism created by EsMeCaTa. Exploring communities of different sizes could be engaging, but this will be associated with issues in scalability, as increasing community size will increase combinatorial possibilities.

Also, several refinements could be applied to improve the results given by Metage2Metabo.

As shown in subsection 5.3.2 and 5.3.3, cofactors can impact the prediction. In that case, it allowed the production of many metabolites due to using the NAD degradation pathway. Thus a method avoiding these issues should be implemented into Metage2Metabo. To fix this, I used in subsection 5.3.3 the method proposed by moped, which creates new metabolites dedicated to the goal of being cofactors. One solution could be to add this method into Metage2Metabo.

Another issue is that we did not know which path the network expansion algorithm uses to reach the targets starting from the seed metabolites. Indeed, this path could not be a biologically meaningfully one. The first way to discover this issue could be to visualise the path taken by the network expansion scope. An implementation could be possible by using the incremental scope and counting the number of steps to produce the targets. This approach could give an idea of which path was used.

Following this idea of a biologically meaningful path, another improvement could be to force some specific paths. For example, in methanogenesis, we have two paths from organic matter to methane, one from acetate and another from carbon dioxide. Thus we could be interested in determining if these paths are indeed used by the minimal community and which paths. One possible way to implement this could be to force the production of intermediary metabolites (such as acetate or carbon dioxide) so that the path used by the minimal community goes through these intermediary metabolites.

A third refinement could be considering other data, such as gene expression or the abundance of organisms. Indeed in this experiment, we only used the abundance to filter whether the taxon was present or not. But it would be interesting to deal with population sizes as more abundant members could have more impact. Such improvements could be associated with other methods, such as co-occurrence methods.

## Perspectives

The study of metabolic interactions of large-scale microbiota allows the analysis of community. It provides the identification of key species. A first perspective could be to identify the metabolic functions that are performed by the key species. It corresponds to understanding why the topological analysis has chosen this key species. A second perspective could be to create the metabolic networks of all the known genus or family with EsMeCaTa and adds the possibility of using Metage2Metabo on any combination of organisms. This combination of EsMeCaTa and Metage2Metabo could allow the systematic exploration of metabolic complementarity among all the known organisms.

Part III

# Conclusions and perspectives

Chapter 6

# Conclusion and Perspectives

## 6.1 Conclusion

This thesis presented different methods to elucidate metabolic functions at various levels, from metabolic pathways to metabolic interactions in communities. As explained in Chapter 1, the metabolism is measured through different methods such as the -omics approaches. They produce multiple and complementary data. To achieve a wide understanding of the metabolism, combining the different data are needed when developing methods.

The methods presented in this thesis aimed at predicting interesting candidates for the studied levels (metabolic pathways, comparable GSMNs, flexible TSMN and key species in communities). To achieve this, we combine computer science (**constraint programming**, **knowledge engineering**), bioinformatics (**sequence analysis**) and biology (**evolution**). The predicted candidates were then discussed with experts and confronted with the literature.

### 6.1.1 Knowledge representation, querying and reasoning

Existing methods relied on already available knowledge to model the metabolism and make predictions. This central point occurs in all the chapters of this thesis. In Chapter 2, we based our predictions on reference metabolic pathways, which were knowledge representations of metabolism from metabolic databases (such as MetaCyc) and literature. The comparison of GSMNs in Chapter 3 was possible thanks to different knowledge sources. The first one comes from genomes in public databases (associated with heterogeneous annotations, which were another level of knowledge). A second source comes from the inference of metabolism from these annotations by relying on metabolic reconstruction tools (in this case, Pathway Tools and the MetaCyc database). To estimate the metabolism of taxa, we also combined multiple databases in Chapter 4. Indeed, we used

the knowledge from combined databases of UniProt. Lastly, the computation of minimal community relied on the knowledge from metabolism associated with the reconstructed metabolic networks in Chapter 5.

These multiple pieces of knowledge were described in different representations to be adapted to each analysis. This way, metabolic information was converted into logical statements in Chapter 2. The manual creation of these statements allowed us to reach the consistency required for the metabolic pathways prediction. As we went to a higher study level, we used different representations, from compound graphs to bipartite graphs. Indeed, in the three other chapters (3, 4 and 5), the graphical representation of metabolism used bipartite graphs to focus on information reification (especially by taking into account the origin of the information which were genes, proteins or organisms of the community).

Knowledge representation was performed to apply querying and reasoning adapted to each level. Constraint programming was used for the metabolic pathway level to implement logical rules relying on evolutionary biology hypotheses (the Metabolic Pathway Drift). These rules were then used on the logical statements to infer alternative pathways. Knowledge reasoning was used to compare GSMNs to study the equivalences and differences between organisms. Knowledge querying was performed in Chapter 4 to retrieve knowledge and infer metabolic network at the scale of the taxon, applying evolutionary hypotheses stating that inherited functions are shared in the taxon. Constraint programming was again used in Chapter 5 to handle the large-scale community to identify the key species which collectively contribute to producing some metabolites, hence investigating the applicability of the Black Queen hypothesis.

**Subsection summary**

In this thesis, we relied on Knowledge Representation and Reasoning to handle the data and combine it with knowledge for the four metabolism levels studied. We adapted the knowledge representation for each level. The reasoning and querying were used according to the level, from constraint programming to knowledge querying. With these methods, it was possible to represent, query and reason on knowledge. This allowed scaling for large-scale analysis, solving problems (combinatorics) and handling knowledge.

### 6.1.2 Homogenising and filtering data and predictions

The biological hypotheses, knowledge representation, querying, and reasoning presented in the previous subsection were combined with homogenising and filtering procedures. These procedures were developed by relying on different approaches.

A major approach used is the sequence analysis and comparison from Bioinformatics. This approach was used in all the chapters. In the Chapter 2, sequences of *C. crispus* were compared to sequences from model organisms with AuReMe to infer the GSMN. Orthology propagation and structural verification performed in Chapter 3 relied on sequence analysis to homogenise GSMNs, the first on comparing protein sequences and the second on searching for proteins in genomes. Furthermore, the robustness criterion filtered reactions according to groups of orthologous genes found by sequence analysis. To identify protein clusters among the proteomes of a taxon in Chapter 4, we clustered proteins according to their sequences, allowing us to create groups of shared proteins.

Other approaches of filtering were used, especially one relying on knowledge representation. Indeed, to identify candidate reactions for metabolic pathways inference, we relied on the filtering of candidates according to the comparison of the atoms and bonds in the metabolites. This identification was possible thanks to the knowledge representation, logical rules and manually created molecule structures. A second filtering strategy also relied on logical rules. It concerns identifying key species in Chapter 5, where organisms were selected according to their production capacities through logical rules based on the network expansion algorithm.

**Subsection summary**

To handle heterogeneous data, we had to develop methods to homogenise the content. It was specially performed with sequence comparison. Also, to handle the numerous predictions made by the methods, there was a need for filtering methods to identify the most relevant predictions. This was done thanks to the development of filtering approaches in the different methods.

### 6.1.3 Aiding experts by predicting candidates and testing hypothesis

In this thesis, two main contributions were made from a biological point of view.

First, by taking into account already performed experiments, we were able to produce a list of candidates for the different analyses. These candidates were multiples (genes, metabolic pathways, comparable GSMNs, metabolic interactions in a biogas reactor). In Chapter 2, we proposed two modified metabolic pathways with variations resulting from the Metabolic Pathway Drift in the red algae *C. crispus*. By comparing the metabolism of algae in Chapter 3, we proposed a set of comparable GSMNs. These GSMNs allowed for comparing the species' phylogeny and the distances between their metabolisms. Furthermore, it permits highlighting the metabolic specificities of organisms in the algal group. Estimating metabolic capabilities from taxonomic affiliations in Chapter 4 allowed for creating sets of candidate protein clusters for a taxon. Furthermore, it gave us insight into the metabolic functions of each taxon in a biogas reactor. With the results of Chapter 4 and the method of Chapter 5, we could propose a set of candidate taxa that could be of interest for the production of biogas in the reactor. These candidates were discussed with experts (such as the team from the Station Biologique de Roscoff or Patrick Dabert) to understand these results better and help them better understand their experiments.

In a second contribution, we helped test and propose hypotheses according to their experiments. Indeed, for Chapter 2, we tested the metabolic pathway drift through a set of logical rules that were able to predict alternative metabolic pathways of interest for the collaborator. In Chapter 3, the result of the comparison of GSMNs helped the experts into the metabolic distances between specific organisms and the place of interesting organisms. For Chapters 4 and 5, it was possible to test the Black Queen hypothesis. We especially studied the microbial community in a biogas reactor and the metabolic interactions between its members.

**Subsection summary**

In this thesis, I helped experts understand the data and test biological hypotheses through knowledge representation and reasoning. A first contribution was identifying genes, metabolic, and organisms of interest according to their biological context. And a second contribution was to test biological hypotheses by creating models working under these hypotheses.

## 6.2 Perspectives

I participated in the development of several methods in this thesis. I presented some of their applications. As hinted at in each chapter's conclusion, numerous perspectives can be proposed following these works.

### 6.2.1 Computer sciences

Knowledge representation and reasoning were used to make predictions on non-model organisms. The addition of knowledge for the case where we have little data, such as the one presented in Chapter 2 with non-model organisms, helped make predictions. Such knowledge-driven hypotheses could be applied to the network expansion algorithm. Indeed, a limitation of the analysis performed in Chapter 5 is that we do not know the path taken by the network expansion algorithm. Thus it is not possible to understand if the chosen paths (and the proposed organisms) corresponded to a meaningful biological path of reactions or not. A combination of adjustments could be performed on the logical rules used by the methods. First, the use of the incremental mode of ASP (such as used in Chapter 2) can help by separating step by step the network expansion. This mode could be applied to determine which metabolites were used to go from seed to target metabolites. A second improvement could be to add knowledge through the use of intermediary metabolites. It could require that a path between seeds and targets travel through known intermediate metabolites, thus enforcing a biologically meaningful path between seeds and targets.

The second perspective in computer science lies in the available knowledge and its querying. As presented in the previous section, the methods developed in this thesis relied on databases to extract knowledge. An improvement to several of the developed

methods could be to combine different databases. This combination could help to connect metabolomics data and metabolic network in Chapter 2). And in Chapter 4, this could permit to combine proteomes selection on Uniprot with known living conditions of the organism associated with the proteomes (such as BacDive (Reimer et al. 2022)). A possible perspective to solve these issues is with the Semantic Web technologies. Through these technologies, it is possible to describe the data and their representation inside a hierarchy in a machine-readable way. Furthermore, these methods ease the interoperability between databases. Some already use Semantic Web technologies (such as UniProt and Rhea). These technologies allow querying multiple databases at the same time. If more databases were available through these technologies, it should be possible to quickly request them and associate them with other databases. This could allow methods relying on multiple databases to increase their possibility of prediction.

### 6.2.2 Biological perspectives

As omics data are currently accumulated, new questions emerge in biology. Such questions concern understanding the genome dark matter, all the sequences and metabolites unrelated to anything known. It also concerns the understanding of the holobionts, defining an organism and its microbiota as a whole. Bioinformatics methods presented in this thesis contribute to exploring, testing or validating the hypotheses related to those emerging ideas. By drawing various predictions, the methods I contributed to assess the putative outcomes of the pathway drifts or of the Black Queen hypotheses. The modelling and predictions performed in this thesis were made to help experts. But several methods could need to be experimentally validated, increasing the need for collaboration with the experts. And the methods could also be used as tools for new research questions.

The first perspective would be the experimental validation of the different methods developed by our collaborators. For Chapter 2, the method predicts alternative metabolic pathways. There is a need to ensure that these pathways are present in the organisms. Several experiments could be performed to ensure good predictions of the pathways. A first way could be to silence the gene candidates associated with the pathway. Such an approach could be combined with metabolite quantification. This could show the accumulation of the substrate of the reaction catalysed by the silenced enzyme-coding gene and validate the intermediary metabolites. In Chapter 3, verification could be performed by comparing similar organisms living in different environments. For example, this could test the metabolic changes according to the change in ecological conditions. A test of Chapter

4 would be to perform an analysis by taking taxonomic affiliations and complete genomes from an environmental sample. We could predict with the method and compare it to the metabolism performed by the complete genomes. Comparing the two predictions could give insight into the most adapted options of EsMeCaTa. Finally, for the final Chapter 5, the difference between essential and alternative symbionts could be experimentally tested in a controlled environment. The possible role of essential symbionts as "helpers" could be tested using synthetic communities where species predicted to be essential symbionts are removed.

A second perspective could be to test and propose new hypotheses for the experts. In Chapter 2, we have tested a way to predict changes resulting from the metabolic pathway drift. By associating this method with the method presented in Chapter 3, an interesting perspective could be to research the modifications over time of metabolic pathways for groups of related organisms. This approach could lead to understanding the evolution of such pathways and finding the ancestral metabolic pathways. This is the subject of research of one of our collaborators, Gabriel Markov, from the Station Biologique de Roscoff. In a second time, the combination of methods developed in Chapter 4 and 5 led to the prediction of minimal communities. Such analysis allows us to test the Black Queen hypothesis, estimate the metabolic interactions between the organisms and identify if there are possible syntrophic organisms.

### 6.2.3 Integrated predictions of metabolism

A combination of the methods could permit easier predictions from users. For example, we could create a database containing EsMeCaTa predictions on all the known genera and families. Then it could be possible to use the predicted TSMNs for metabolic predictions. Indeed, by adding an overlay to this database containing Metage2Metabo, it should be possible to create a web service. This web service could allow any user to create a job selecting a set of genera or families (that could correspond to the taxonomic affiliations from a metagenomics sample or only taxa of interest), a group of seed metabolites and optionally a set of target metabolites. Then the overlay will launch the job with Metage2Metabo on the selected taxa. In this way, it could be possible to test metabolic complementarity between any known genus and/or family that had proteomes on UniProt as illustrated in Chapter 5.

The different levels of metabolism presented in this thesis could be combined to make better predictions. Considering the possibility of metabolic pathway drift in non-model

organisms could help the GSMN reconstruction. Such an addition could create more modular metabolic networks than those created only from reference metabolic networks of model organisms. Increasing the diversity of metabolic pathways makes it possible to have more diversity for GSMN. Then this could impact all the other levels as the metabolism of taxonomic groups and the metabolic interactions in communities could benefit from such addition. But there is also possible feedback as the metabolic interactions between organisms could help to better shape the metabolism of organisms by identifying likely auxotrophic and syntrophic organisms. Then associated with the comparison of metabolism at the organism level, it should be possible to study the evolution of auxotrophy and syntrophy in these organisms.

# BIBLIOGRAPHY


Agren, Rasmus, Liming Liu, Saeed Shoaie, Wanwipa Vongsangnak, Intawat Nookaew, and Jens Nielsen (2013), « The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for Penicillium chrysogenum », *in*: *PLOS Computational Biology* 9.*3*, e1002980, ISSN: 1553-7358, DOI: 10.1371/journal.pcbi.1002980.

Aite, Méziane, Marie Chevallier, Clémence Frioux, Camille Trottier, Jeanne Got, María Paz Cortés, Sebastián N. Mendoza, Grégory Carrier, Olivier Dameron, Nicolas Guillaudeux, Mauricio Latorre, Nicolás Loira, Gabriel V. Markov, Alejandro Maass, and Anne Siegel (2018), « Traceability, reproducibility and wiki-exploration for "à-la-carte" reconstructions of genome-scale metabolic models », *in*: *PLOS Computational Biology* 14.*5*, e1006146, ISSN: 1553-7358, DOI: 10.1371/journal.pcbi.1006146.

Alcaide, A., M. Devys, and M. Barbier (1968), « Remarques sur les sterols des algues rouges », *in*: *Phytochemistry* 7.*2*, pp. 329–330, ISSN: 0031-9422, DOI: 10.1016/S0031-9422(00)86332-3.

Almeida, Alexandre, Stephen Nayfach, Miguel Boland, Francesco Strozzi, Martin Beracochea, Zhou Jason Shi, Katherine S. Pollard, Ekaterina Sakharova, Donovan H. Parks, Philip Hugenholtz, Nicola Segata, Nikos C. Kyrpides, and Robert D. Finn (2021), « A unified catalog of 204,938 reference genomes from the human gut microbiome », *in*: *Nature Biotechnology* 39.*1*, pp. 105–114, ISSN: 1546-1696, DOI: 10.1038/s41587-020-0603-3.

Altman, Tomer, Michael Travers, Anamika Kothari, Ron Caspi, and Peter D. Karp (2013), « A systematic comparison of the MetaCyc and KEGG pathway databases », *in*: *BMC Bioinformatics* 14.*1*, p. 112, ISSN: 1471-2105, DOI: 10.1186/1471-2105-14-112.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman (1990), « Basic local alignment search tool », *in*: *Journal of Molecular Biology* 215.*3*, pp. 403–410, ISSN: 0022-2836, DOI: 10.1016/S0022-2836(05)80360-2.

Amara, Adam, Clément Frainay, Fabien Jourdan, Thomas Naake, Steffen Neumann, Elva María Novoa-del-Toro, Reza M Salek, Liesa Salzer, Sarah Scharfenberg, and Michael Witting (2022), « Networks and Graphs Discovery in Metabolomics Data Analysis and Interpretation », *in*: *Frontiers in Molecular Biosciences* 9, ISSN: 2296-889X, DOI: 10.3389/fmolb.2022.841373.

Andres, Benjamin, Benjamin Kaufmann, Oliver Matheis, and Torsten Schaub (2012), « Unsatisfiability-based optimization in clasp », *in*: *Technical Communications of the 28th International Conference on Logic Programming (ICLP'12)*, ed. by Agostino Dovier and

Vítor Santos Costa, vol. 17, Leibniz International Proceedings in Informatics (LIPIcs), Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, pp. 211–221, ISBN: 978-3-939897-43-9, DOI: 10.4230/LIPIcs.ICLP.2012.211.

Arkin, Adam P, Robert W Cottingham, Christopher S Henry, Nomi L Harris, Rick L Stevens, Sergei Maslov, Paramvir Dehal, Doreen Ware, Fernando Perez, Shane Canon, Michael W Sneddon, Matthew L Henderson, William J Riehl, Dan Murphy-Olson, Stephen Y Chan, Roy T Kamimura, Sunita Kumari, Meghan M Drake, Thomas S Brettin, Elizabeth M Glass, Dylan Chivian, Dan Gunter, David J Weston, Benjamin H Allen, Jason Baumohl, Aaron A Best, Ben Bowen, Steven E Brenner, Christopher C Bun, John-Marc Chandonia, Jer-Ming Chia, Ric Colasanti, Neal Conrad, James J Davis, Brian H Davison, Matthew DeJongh, Scott Devoid, Emily Dietrich, Inna Dubchak, Janaka N Edirisinghe, Gang Fang, José P Faria, Paul M Frybarger, Wolfgang Gerlach, Mark Gerstein, Annette Greiner, James Gurtowski, Holly L Haun, Fei He, Rashmi Jain, Marcin P Joachimiak, Kevin P Keegan, Shinnosuke Kondo, Vivek Kumar, Miriam L Land, Folker Meyer, Marissa Mills, Pavel S Novichkov, Taeyun Oh, Gary J Olsen, Robert Olson, Bruce Parrello, Shiran Pasternak, Erik Pearson, Sarah S Poon, Gavin A Price, Srividya Ramakrishnan, Priya Ranjan, Pamela C Ronald, Michael C Schatz, Samuel M D Seaver, Maulik Shukla, Roman A Sutormin, Mustafa H Syed, James Thomason, Nathan L Tintle, Daifeng Wang, Fangfang Xia, Hyunseung Yoo, Shinjae Yoo, and Dantong Yu (2018), « KBase: The United States Department of Energy Systems Biology Knowledgebase », *in*: *Nature Biotechnology* 36.*7*, pp. 566–569, ISSN: 1087-0156, DOI: 10.1038/nbt.4163.

Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock (2000), « Gene Ontology: tool for the unification of biology », *in*: *Nature genetics* 25.*1*, pp. 25–29, ISSN: 1061-4036, DOI: 10.1038/75556.

Aslam, Bilal, Madiha Basit, Muhammad Atif Nisar, Mohsin Khurshid, and Muhammad Hidayat Rasool (2017), « Proteomics: Technologies and Their Applications », *in*: *Journal of Chromatographic Science* 55.*2*, pp. 182–196, ISSN: 0021-9665, DOI: 10.1093/chromsci/bmw167.

Aßhauer, Kathrin P., Bernd Wemheuer, Rolf Daniel, and Peter Meinicke (2015), « Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data », *in*: *Bioinformatics* 31.*17*, pp. 2882–2884, ISSN: 1367-4811, DOI: 10.1093/bioinformatics/btv287.

Awhangbo, L., R. Bendoula, J. M. Roger, and F. Béline (2020), « Multi-block data analysis for online monitoring of anaerobic co-digestion process », *in*: *Chemometrics and Intelligent*

*Laboratory Systems* 205, p. 104120, ISSN: 0169-7439, DOI: 10.1016/j.chemolab.2020.104120.

Aziz, Ramy K., Daniela Bartels, Aaron A. Best, Matthew DeJongh, Terrence Disz, Robert A. Edwards, Kevin Formsma, Svetlana Gerdes, Elizabeth M. Glass, Michael Kubal, Folker Meyer, Gary J. Olsen, Robert Olson, Andrei L. Osterman, Ross A. Overbeek, Leslie K. McNeil, Daniel Paarmann, Tobias Paczian, Bruce Parrello, Gordon D. Pusch, Claudia Reich, Rick Stevens, Olga Vassieva, Veronika Vonstein, Andreas Wilke, and Olga Zagnitko (2008), « The RAST Server: Rapid Annotations using Subsystems Technology », *in*: *BMC Genomics* 9.*1*, p. 75, ISSN: 1471-2164, DOI: 10.1186/1471-2164-9-75.

Bairoch, A. (2000), « The ENZYME database in 2000 », *in*: *Nucleic Acids Research* 28.*1*, pp. 304–305, ISSN: 0305-1048, DOI: 10.1093/nar/28.1.304.

Baldini, Federico, Almut Heinken, Laurent Heirendt, Stefania Magnusdottir, Ronan M T Fleming, and Ines Thiele (2019), « The Microbiome Modeling Toolbox: from microbial interactions to personalized microbial communities », *in*: *Bioinformatics* 35.*13*, pp. 2332–2334, ISSN: 1367-4803, DOI: 10.1093/bioinformatics/bty941.

Bansal, Parit, Anne Morgat, Kristian B Axelsen, Venkatesh Muthukrishnan, Elisabeth Coudert, Lucila Aimo, Nevila Hyka-Nouspikel, Elisabeth Gasteiger, Arnaud Kerhornou, Teresa Batista Neto, Monica Pozzato, Marie-Claude Blatter, Alex Ignatchenko, Nicole Redaschi, and Alan Bridge (2022), « Rhea, the reaction knowledgebase in 2022 », *in*: *Nucleic Acids Research* 50 (D1), pp. D693–D700, ISSN: 0305-1048, DOI: 10.1093/nar/gkab1016.

Baranwal, Mayank, Abram Magner, Paolo Elvati, Jacob Saldinger, Angela Violi, and Alfred O Hero (2020), « A deep learning architecture for metabolic pathway prediction », *in*: *Bioinformatics* 36.*8*, pp. 2547–2553, ISSN: 1367-4803, DOI: 10.1093/bioinformatics/btz954.

Bauer, Eugen, Cedric Christian Laczny, Stefania Magnusdottir, Paul Wilmes, and Ines Thiele (2015), « Phenotypic differentiation of gastrointestinal microbes is reflected in their encoded metabolic repertoires », *in*: *Microbiome* 3.*1*, p. 55, ISSN: 2049-2618, DOI: 10.1186/s40168-015-0121-6.

Bauer, Eugen, Johannes Zimmermann, Federico Baldini, Ines Thiele, and Christoph Kaleta (2017), « BacArena: Individual-based metabolic modeling of heterogeneous microbes in complex communities », *in*: *PLOS Computational Biology* 13.*5*, e1005544, ISSN: 1553-7358, DOI: 10.1371/journal.pcbi.1005544.

Belcour, Arnaud, Clémence Frioux, Méziane Aite, Anthony Bretaudeau, Falk Hildebrand, and Anne Siegel (2020a), « Metage2Metabo, microbiota-scale metabolic complementarity for the identification of key species », *in*: *eLife* 9, ed. by María Mercedes Zambrano, Gisela Storz, Daniel Machado, and Oliver Ebenhoh, e61968, ISSN: 2050-084X, DOI: 10.7554/eLife.61968.

Belcour, Arnaud, Jean Girard, Méziane Aite, Ludovic Delage, Camille Trottier, Charlotte Marteau, Cédric Leroux, Simon M. Dittami, Pierre Sauleau, Erwan Corre, Jacques Nicolas, Catherine Boyen, Catherine Leblanc, Jonas Collén, Anne Siegel, and Gabriel V. Markov (2020b), « Inferring Biochemical Reactions and Metabolite Structures to Understand Metabolic Pathway Drift », *in*: *iScience* 23.*2*, p. 100849, ISSN: 2589-0042, DOI: 10.1016/j.isci.2020.100849.

Benveniste, Pierre (2004), « Biosynthesis and accumulation of sterols », *in*: *Annual Review of Plant Biology* 55, pp. 429–457, ISSN: 1543-5008, DOI: 10.1146/annurev.arplant.55.031903.141616.

Bernstein, David B, Floyd E Dewhirst, and Daniel Segrè (2019), « Metabolic network percolation quantifies biosynthetic capabilities across the human oral microbiome », *in*: *eLife* 8, ed. by Wenying Shou, Naama Barkai, Wenying Shou, and Christopher Quince, e39733, ISSN: 2050-084X, DOI: 10.7554/eLife.39733.

Biochemical Nomenclature, IUPAC-IUB Joint Commission on (1989), « The nomenclature of steroids », *in*: *European Journal of Biochemistry* 186.*3*, pp. 429–458, ISSN: 1432-1033, DOI: 10.1111/j.1432-1033.1989.tb15228.x.

Birch, L. C. (1957), « The Meanings of Competition », *in*: *The American Naturalist* 91, pp. 5–18, ISSN: 0003-0147, DOI: 10.1086/281957.

Blevins, William R., Jorge Ruiz-Orera, Xavier Messeguer, Bernat Blasco-Moreno, José Luis Villanueva-Cañas, Lorena Espinar, Juana Díez, Lucas B. Carey, and M. Mar Albà (2021), « Uncovering de novo gene birth in yeast using deep transcriptomics », *in*: *Nature Communications* 12.*1*, p. 604, ISSN: 2041-1723, DOI: 10.1038/s41467-021-20911-3.

Blum, Matthias, Hsin-Yu Chang, Sara Chuguransky, Tiago Grego, Swaathi Kandasaamy, Alex Mitchell, Gift Nuka, Typhaine Paysan-Lafosse, Matloob Qureshi, Shriya Raj, Lorna Richardson, Gustavo A Salazar, Lowri Williams, Peer Bork, Alan Bridge, Julian Gough, Daniel H Haft, Ivica Letunic, Aron Marchler-Bauer, Huaiyu Mi, Darren A Natale, Marco Necci, Christine A Orengo, Arun P Pandurangan, Catherine Rivoire, Christian J A Sigrist, Ian Sillitoe, Narmada Thanki, Paul D Thomas, Silvio C E Tosatto, Cathy H Wu, Alex Bateman, and Robert D Finn (2021), « The InterPro protein families and domains database: 20 years on », *in*: *Nucleic Acids Research* 49 (D1), pp. D344–D354, ISSN: 0305-1048, DOI: 10.1093/nar/gkaa977.

Bordbar, Aarash, Jonathan M. Monk, Zachary A. King, and Bernhard O. Palsson (2014), « Constraint-based models predict metabolic and associated cellular functions », *in*: *Nature Reviews Genetics* 15.*2*, pp. 107–120, ISSN: 1471-0064, DOI: 10.1038/nrg3643.

Bourneuf, Lucas and Jacques Nicolas (2017), « FCA in a Logical Programming Setting for Visualization-Oriented Graph Compression », *in*: *Formal Concept Analysis*, International

Conference on Formal Concept Analysis 2017, ed. by Karell Bertet, Daniel Borchmann, Peggy Cellier, and Sébastien Ferré, Cham: Springer International Publishing, pp. 89–105, ISBN: 978-3-319-59271-8, DOI: 10.1007/978-3-319-59271-8_6.

Bowman, Jeff S. and Hugh W. Ducklow (2015), « Microbial Communities Can Be Described by Metabolic Structure: A General Framework and Application to a Seasonally Variable, Depth-Stratified Microbial Community from the Coastal West Antarctic Peninsula », *in*: *PLOS ONE* 10.*8*, e0135868, ISSN: 1932-6203, DOI: 10.1371/journal.pone.0135868.

Brawley, Susan H., Nicolas A. Blouin, Elizabeth Ficko-Blean, Glen L. Wheeler, Martin Lohr, Holly V. Goodson, Jerry W. Jenkins, Crysten E. Blaby-Haas, Katherine E. Helliwell, Cheong Xin Chan, Tara N. Marriage, Debashish Bhattacharya, Anita S. Klein, Yacine Badis, Juliet Brodie, Yuanyu Cao, Jonas Collén, Simon M. Dittami, Claire M. M. Gachon, Beverley R. Green, Steven J. Karpowicz, Jay W. Kim, Ulrich Johan Kudahl, Senjie Lin, Gurvan Michel, Maria Mittag, Bradley J. S. C. Olson, Jasmyn L. Pangilinan, Yi Peng, Huan Qiu, Shengqiang Shu, John T. Singer, Alison G. Smith, Brittany N. Sprecher, Volker Wagner, Wenfei Wang, Zhi Yong Wang, Juying Yan, Charles Yarish, Simone Zäuner-Riek, Yunyun Zhuang, Yong Zou, Erika A. Lindquist, Jane Grimwood, Kerrie W. Barry, Daniel S. Rokhsar, Jeremy Schmutz, John W. Stiller, Arthur R. Grossman, and Simon E. Prochnik (2017), « Insights into the red algae and eukaryotic evolution from the genome of Porphyra umbilicalis (Bangiophyceae, Rhodophyta) », *in*: *Proceedings of the National Academy of Sciences of the United States of America* 114.*31*, E6361–E6370, ISSN: 0027-8424, DOI: 10.1073/pnas.1703088114.

Breitling, Rainer, Shawn Ritchie, Dayan Goodenowe, Mhairi L. Stewart, and Michael P. Barrett (2006), « Ab initio prediction of metabolic networks using Fourier transform mass spectrometry data », *in*: *Metabolomics* 2.*3*, pp. 155–164, DOI: 10.1007/s11306-006-0029-z.

Brettin, Thomas, James J. Davis, Terry Disz, Robert A. Edwards, Svetlana Gerdes, Gary J. Olsen, Robert Olson, Ross Overbeek, Bruce Parrello, Gordon D. Pusch, Maulik Shukla, James A. Thomason, Rick Stevens, Veronika Vonstein, Alice R. Wattam, and Fangfang Xia (2015), « RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes », *in*: *Scientific Reports* 5, p. 8365, ISSN: 2045-2322, DOI: 10.1038/srep08365.

Burgunter-Delamare, Bertille, Hetty KleinJan, Clémence Frioux, Enora Fremy, Margot Wagner, Erwan Corre, Alicia Le Salver, Cédric Leroux, Catherine Leblanc, Catherine Boyen, Anne Siegel, and Simon M. Dittami (2020), « Metabolic Complementarity Between a Brown Alga and Associated Cultivable Bacteria Provide Indications of Beneficial Interactions », *in*: *Frontiers in Marine Science* 7, ISSN: 2296-7745, DOI: 10.3389/fmars.2020.00085.

Calegario, Gabriela, Jacob Pollier, Philipp Arendt, Louisi Souza de Oliveira, Cristiane Thompson, Angélica Ribeiro Soares, Renato Crespo Pereira, Alain Goossens, and Fabiano L.

Thompson (2016), « Cloning and Functional Characterization of Cycloartenol Synthase from the Red Seaweed Laurencia dendroidea », *in*: *PLOS ONE* 11.*11*, e0165954, ISSN: 1932-6203, DOI: 10.1371/journal.pone.0165954.

Calhoun, Sara, Magdalena Korczynska, Daniel J Wichelecki, Brian San Francisco, Suwen Zhao, Dmitry A Rodionov, Matthew W Vetting, Nawar F Al-Obaidi, Henry Lin, Matthew J O'Meara, David A Scott, John H Morris, Daniel Russel, Steven C Almo, Andrei L Osterman, John A Gerlt, Matthew P Jacobson, Brian K Shoichet, and Andrej Sali (2018), « Prediction of enzymatic pathways by integrative pathway mapping », *in*: *eLife* 7, e31097, ISSN: 2050-084X, DOI: 10.7554/eLife.31097.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden (2009), « BLAST+: architecture and applications », *in*: *BMC bioinformatics* 10, p. 421, ISSN: 1471-2105, DOI: 10.1186/1471-2105-10-421.

Capela, João, Davide Lagoa, Ruben Rodrigues, Emanuel Cunha, Fernando Cruz, Ana Barbosa, José Bastos, Diogo Lima, Eugénio C Ferreira, Miguel Rocha, and Oscar Dias (2022), « merlin, an improved framework for the reconstruction of high-quality genome-scale metabolic models », *in*: *Nucleic Acids Research* 50.*11*, pp. 6052–6066, ISSN: 0305-1048, DOI: 10.1093/nar/gkac459.

Cardona Uribe, Cesar Emilio (2019), « Using Genomically Inferred Co-Occurrence and Metabolic Networks to Characterize the Built Environment Microbiome », PhD thesis, The University of Chicago: The University of Chicago, 158 pp., DOI: 10.6082/uchicago.1978.

Carreto, Jose I. and Mario O. Carignan (2011), « Mycosporine-like amino acids: relevant secondary metabolites. Chemical and ecological aspects », *in*: *Marine Drugs* 9.*3*, pp. 387–446, ISSN: 1660-3397, DOI: 10.3390/md9030387.

Caspi, Ron, Richard Billington, Luciana Ferrer, Hartmut Foerster, Carol A. Fulcher, Ingrid M. Keseler, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Lukas A. Mueller, Quang Ong, Suzanne Paley, Pallavi Subhraveti, Daniel S. Weaver, and Peter D. Karp (Jan. 4, 2016), « The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases », *in*: *Nucleic Acids Research* 44 (D1), pp. D471–D480, ISSN: 0305-1048, DOI: 10.1093/nar/gkv1164.

Caspi, Ron, Richard Billington, Ingrid M Keseler, Anamika Kothari, Markus Krummenacker, Peter E Midford, Wai Kit Ong, Suzanne Paley, Pallavi Subhraveti, and Peter D Karp (2020), « The MetaCyc database of metabolic pathways and enzymes - a 2019 update », *in*: *Nucleic Acids Research* 48 (D1), pp. D445–D453, ISSN: 0305-1048, DOI: 10.1093/nar/gkz862.

Castillo, Sandra, Dorothee Barth, Mikko Arvas, Tiina M. Pakula, Esa Pitkänen, Peter Blomberg, Tuulikki Seppanen-Laakso, Heli Nygren, Dhinakaran Sivasiddarthan, Merja Penttilä, and Merja Oja (2016), « Whole-genome metabolic model of Trichoderma reesei built by com-

parative reconstruction », *in*: *Biotechnology for Biofuels* 9.*1*, p. 252, ISSN: 1754-6834, DOI: 10.1186/s13068-016-0665-0.

Chan, Siu Hung Joshua, Margaret N. Simons, and Costas D. Maranas (2017), « SteadyCom: Predicting microbial abundances while ensuring community stability », *in*: *PLOS Computational Biology* 13.*5*, e1005539, ISSN: 1553-7358, DOI: 10.1371/journal.pcbi.1005539.

Cho, Ilseung and Martin J. Blaser (2012), « The human microbiome: at the interface of health and disease », *in*: *Nature Reviews Genetics* 13.*4*, pp. 260–270, ISSN: 1471-0064, DOI: 10.1038/nrg3182.

Cock, Peter J. A., Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon (2009), « Biopython: freely available Python tools for computational molecular biology and bioinformatics », *in*: *Bioinformatics* 25.*11*, pp. 1422–1423, ISSN: 1367-4803, DOI: 10.1093/bioinformatics/btp163.

Cole, James R., Qiong Wang, Jordan A. Fish, Benli Chai, Donna M. McGarrell, Yanni Sun, C. Titus Brown, Andrea Porras-Alfaro, Cheryl R. Kuske, and James M. Tiedje (2014), « Ribosomal Database Project: data and tools for high throughput rRNA analysis », *in*: *Nucleic Acids Research* 42 (Database issue), pp. D633–642, ISSN: 1362-4962, DOI: 10.1093/nar/gkt1244.

Conan, Mael, Nathalie Théret, Sophie Langouet, and Anne Siegel (2021), « Constructing xenobiotic maps of metabolism to predict enzymes catalyzing metabolites capable of binding to DNA », *in*: *BMC Bioinformatics* 22.*1*, p. 450, ISSN: 1471-2105, DOI: 10.1186/s12859-021-04363-6.

Correia, Kevin and Radhakrishnan Mahadevan (2020), « Pan-Genome-Scale Network Reconstruction: Harnessing Phylogenomics Increases the Quantity and Quality of Metabolic Models », *in*: *Biotechnology Journal* 15.*10*, p. 1900519, ISSN: 1860-7314, DOI: https://doi.org/10.1002/biot.201900519.

Cottret, Ludovic, Clément Frainay, Maxime Chazalviel, Floréal Cabanettes, Yoann Gloaguen, Etienne Camenen, Benjamin Merlet, Stéphanie Heux, Jean-Charles Portais, Nathalie Poupin, Florence Vinson, and Fabien Jourdan (2018), « MetExplore: collaborative edition and exploration of metabolic networks », *in*: *Nucleic Acids Research* 46 (W1), W495–W502, ISSN: 0305-1048, DOI: 10.1093/nar/gky301.

Curry, Kristen D., Qi Wang, Michael G. Nute, Alona Tyshaieva, Elizabeth Reeves, Sirena Soriano, Qinglong Wu, Enid Graeber, Patrick Finzer, Werner Mendling, Tor Savidge, Sonia Villapol, Alexander Dilthey, and Todd J. Treangen (2022), « Emu: species-level microbial community profiling of full-length 16S rRNA Oxford Nanopore sequencing data », *in*: *Nature Methods*, pp. 1–9, ISSN: 1548-7105, DOI: 10.1038/s41592-022-01520-4.

Daval, Stéphanie, Arnaud Belcour, Kévin Gazengel, Ludovic Legrand, Jérôme Gouzy, Ludovic Cottret, Lionel Lebreton, Yoann Aigu, Christophe Mougel, and Maria J. Manzanares-Dauleux (2019), « Computational analysis of the Plasmodiophora brassicae genome: mitochondrial sequence description and metabolic pathway database design », *in*: *Genomics* 111.*6*, pp. 1629–1640, ISSN: 08887543, DOI: 10.1016/j.ygeno.2018.11.013.

Daval, Stéphanie, Kévin Gazengel, Arnaud Belcour, Juliette Linglin, Anne-Yvonne Guillerm-Erckelboudt, Alain Sarniguet, Maria J. Manzanares-Dauleux, Lionel Lebreton, and Christophe Mougel (2020), « Soil microbiota influences clubroot disease by modulating Plasmodiophora brassicae and Brassica napus transcriptomes », *in*: *Microbial Biotechnology* 13.*5*, pp. 1648–1672, ISSN: 1751-7915, 1751-7915, DOI: 10.1111/1751-7915.13634.

Delépine, Baudoin, Thomas Duigou, Pablo Carbonell, and Jean-Loup Faulon (2018), « RetroPath2.0: A retrosynthesis workflow for metabolic engineers », *in*: *Metabolic Engineering* 45, pp. 158–170, ISSN: 1096-7176, DOI: 10.1016/j.ymben.2017.12.002.

Desmond, Elie and Simonetta Gribaldo (2009), « Phylogenomics of Sterol Synthesis: Insights into the Origin, Evolution, and Diversity of a Key Eukaryotic Feature », *in*: *Genome Biology and Evolution* 1, pp. 364–381, ISSN: 1759-6653, DOI: 10.1093/gbe/evp036.

Dias, Oscar, Miguel Rocha, Eugenio C. Ferreira, and Isabel Rocha (2010), « Merlin: Metabolic Models Reconstruction using Genome-Scale Information* », *in*: *IFAC Proceedings Volumes* 43.*6*, pp. 120–125, ISSN: 1474-6670, DOI: 10.3182/20100707-3-BE-2012.0076.

Dias, Oscar, Miguel Rocha, Eugénio C. Ferreira, and Isabel Rocha (2015), « Reconstructing genome-scale metabolic models with merlin », *in*: *Nucleic Acids Research* 43.*8*, pp. 3899–3910, ISSN: 0305-1048, DOI: 10.1093/nar/gkv294.

Diener, Christian, Sean M. Gibbons, and Osbaldo Resendis-Antonio (2020), « MICOM: Metagenome-Scale Modeling To Infer Metabolic Interactions in the Gut Microbiota », *in*: *mSystems* 5.*1*, e00606–19, DOI: 10.1128/mSystems.00606-19.

Douglas, Gavin M., Vincent J. Maffei, Jesse R. Zaneveld, Svetlana N. Yurgel, James R. Brown, Christopher M. Taylor, Curtis Huttenhower, and Morgan G. I. Langille (2020), « PICRUSt2 for prediction of metagenome functions », *in*: *Nature Biotechnology* 38.*6*, pp. 685–688, ISSN: 1546-1696, DOI: 10.1038/s41587-020-0548-6.

Dreyfuss, J. M., J. D. Zucker, H. M. Hood, L. R. Ocasio, M. S. Sachs, and J. E. Galagan (2013), « Reconstruction and validation of a genome-scale metabolic model for the filamentous fungus Neurospora crassa using FARM », *in*: *PLoS Comput Biol* 9.*7*, [PubMed Central:PMC3730674] [DOI:10.1371/journal.pcbi.1003126] [PubMed:16507154], e1003126.

Ebenhöh, Oliver, Thomas Handorf, and Reinhart Heinrich (2004), « Structural Analysis of Expanding Metabolic Networks », *in*: *Genome Informatics* 15.*1*, pp. 35–45, DOI: 10.11234/gi1990.15.35.

Edwards, Jeremy S. and Bernhard Ø. Palsson (1999), « Systems Properties of the Haemophilus influenzaeRd Metabolic Genotype * », *in*: *Journal of Biological Chemistry* 274.*25*, pp. 17410–17416, ISSN: 0021-9258, 1083-351X, DOI: 10.1074/jbc.274.25.17410.

Ejigu, Girum Fitihamlak and Jaehee Jung (2020), « Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing », *in*: *Biology* 9.*9*, p. 295, ISSN: 2079-7737, DOI: 10.3390/biology9090295.

Emms, David M. and Steven Kelly (2015), « OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy », *in*: *Genome Biology* 16.*1*, p. 157, ISSN: 1474-760X, DOI: 10.1186/s13059-015-0721-2.

— (2019), « OrthoFinder: phylogenetic orthology inference for comparative genomics », *in*: *Genome Biology* 20.*1*, p. 238, ISSN: 1474-760X, DOI: 10.1186/s13059-019-1832-y.

Eng, Alexander and Elhanan Borenstein (2016), « An algorithm for designing minimal microbial communities with desired metabolic capacities », *in*: *Bioinformatics* 32.*13*, pp. 2008–2016, ISSN: 1367-4811, DOI: 10.1093/bioinformatics/btw107.

Escudié, Frédéric, Lucas Auer, Maria Bernard, Mahendra Mariadassou, Laurent Cauquil, Katia Vidal, Sarah Maman, Guillermina Hernandez-Raquet, Sylvie Combes, and Géraldine Pascal (2018), « FROGS: Find, Rapidly, OTUs with Galaxy Solution », *in*: *Bioinformatics* 34.*8*, pp. 1287–1294, ISSN: 1367-4803, DOI: 10.1093/bioinformatics/btx791.

Faust, Karoline, Didier Croes, and Jacques van Helden (2011), « Prediction of metabolic pathways from genome-scale metabolic networks », *in*: *Biosystems* 105.*2*, pp. 109–121, ISSN: 0303-2647, DOI: 10.1016/j.biosystems.2011.05.004.

Federhen, Scott (2012), « The NCBI Taxonomy database », *in*: *Nucleic Acids Research* 40 (Database issue), pp. D136–D143, ISSN: 0305-1048, DOI: 10.1093/nar/gkr1178.

Forster, Samuel C., Nitin Kumar, Blessing O. Anonye, Alexandre Almeida, Elisa Viciani, Mark D. Stares, Matthew Dunn, Tapoka T. Mkandawire, Ana Zhu, Yan Shao, Lindsay J. Pike, Thomas Louie, Hilary P. Browne, Alex L. Mitchell, B. Anne Neville, Robert D. Finn, and Trevor D. Lawley (2019), « A human gut bacterial genome and culture collection for improved metagenomic analyses », *in*: *Nature Biotechnology* 37.*2*, pp. 186–192, ISSN: 1546-1696, DOI: 10.1038/s41587-018-0009-7.

Foster, Zachary S. L., Thomas J. Sharpton, and Niklaus J. Grünwald (2017), « Metacoder: An R package for visualization and manipulation of community taxonomic diversity data », *in*: *PLOS Computational Biology* 13.*2*, e1005404, ISSN: 1553-7358, DOI: 10.1371/journal.pcbi.1005404.

Frioux, Clémence, Simon M. Dittami, and Anne Siegel (2020a), « Using automated reasoning to explore the metabolism of unconventional organisms: a first step to explore host–microbial interactions », *in*: *Biochemical Society Transactions* 48.*3*, pp. 901–913, ISSN: 0300-5127.

Frioux, Clémence, Enora Fremy, Camille Trottier, and Anne Siegel (2018a), « Scalable and exhaustive screening of metabolic functions carried out by microbial consortia », *in*: *Bioinformatics* 34.*17*, pp. i934–i943, ISSN: 14602059, DOI: 10.1093/bioinformatics/bty588.

— (2018b), « Scalable and exhaustive screening of metabolic functions carried out by microbial consortia », *in*: *Bioinformatics* 34.*17*, pp. i934–i943, ISSN: 1367-4803, DOI: 10.1093/bioinformatics/bty588.

Frioux, Clémence, Dipali Singh, Tamas Korcsmaros, and Falk Hildebrand (2020b), « From bag-of-genes to bag-of-genomes: metabolic modelling of communities in the era of metagenome-assembled genomes », *in*: *Computational and Structural Biotechnology Journal* 18, pp. 1722–1734, ISSN: 2001-0370, DOI: 10.1016/j.csbj.2020.06.028.

Galanie, Stephanie, Kate Thodey, Isis J. Trenchard, Maria Filsinger Interrante, and Christina D. Smolke (2015), « Complete biosynthesis of opioids in yeast », *in*: *Science* 349.*6252*, pp. 1095–1100, DOI: 10.1126/science.aac9373.

García-Jiménez, Beatriz, Jesús Torres-Bacete, and Juan Nogales (2021), « Metabolic modelling approaches for describing and engineering microbial communities », *in*: *Computational and Structural Biotechnology Journal* 19, pp. 226–246, ISSN: 2001-0370, DOI: 10.1016/j.csbj.2020.12.003.

Gaudet, Pascale and Christophe Dessimoz (2017), « Gene Ontology: Pitfalls, Biases, and Remedies », *in*: *The Gene Ontology Handbook*, ed. by Christophe Dessimoz and Nives Škunca, New York, NY: Springer, pp. 189–205, DOI: 10.1007/978-1-4939-3743-1_14.

GBIF Secretariat (2021), « GBIF Backbone Taxonomy », *in*: DOI: 10.15468/39omei, URL: https://www.gbif.org/fr/dataset/d7dddbf4-2cf0-4f39-9b2a-bb099caae36c.

Gebser, Martin, Roland Kaminski, Benjamin Kaufmann, Max Ostrowski, Torsten Schaub, and Philipp Wanko (2016), « Theory Solving Made Easy with Clingo 5 », *in*: *Technical Communications of the 32nd International Conference on Logic Programming (ICLP 2016)*, ed. by Manuel Carro, Andy King, Neda Saeedloei, and Marina De Vos, vol. 52, OpenAccess Series in Informatics (OASIcs), ISSN: 2190-6807, Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2:1–2:15, ISBN: 978-3-95977-007-1, DOI: 10.4230/OASIcs.ICLP.2016.2.

Gebser, Martin, Roland Kaminski, Benjamin Kaufmann, and Torsten Schaub (2012), « Answer Set Solving in Practice », *in*: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6.*3*, pp. 1–238, ISSN: 1939-4608, DOI: 10.2200/S00457ED1V01Y201211AIM019.

— (Jan. 2019), « Multi-shot ASP solving with clingo », *in*: *Theory and Practice of Logic Programming* 19.*1*, Publisher: Cambridge University Press, pp. 27–82, ISSN: 1471-0684, 1475-3081, DOI: 10.1017/S1471068418000054.

Gebser, Martin, Benjamin Kaufmann, André Neumann, and Torsten Schaub (2007), « clasp: A Conflict-Driven Answer Set Solver », *in*: *Logic Programming and Nonmonotonic Reasoning*, ed. by Chitta Baral, Gerhard Brewka, and John Schlipf, Lecture Notes in Computer Science, Berlin, Heidelberg: Springer, pp. 260–265, ISBN: 978-3-540-72200-7, DOI: 10.1007/978-3-540-72200-7_23.

Gene Ontology Consortium (2021), « The Gene Ontology resource: enriching a GOld mine », *in*: *Nucleic Acids Research* 49 (D1), pp. D325–D334, ISSN: 1362-4962, DOI: 10.1093/nar/gkaa1113.

Gerlin, Léo, Ludovic Cottret, Antoine Escourrou, Stéphane Genin, and Caroline Baroukh (2022), « A multi-organ metabolic model of tomato predicts plant responses to nutritional and genetic perturbations », *in*: *Plant Physiology* 188.*3*, pp. 1709–1723, ISSN: 0032-0889, DOI: 10.1093/plphys/kiab548.

Giani, Alice Maria, Guido Roberto Gallo, Luca Gianfranceschi, and Giulio Formenti (2020), « Long walk to genomics: History and current approaches to genome sequencing and assembly », *in*: *Computational and Structural Biotechnology Journal* 18, pp. 9–19, ISSN: 2001-0370, DOI: 10.1016/j.csbj.2019.11.002.

Giovannoni, Stephen J., Theresa B. Britschgi, Craig L. Moyer, and Katharine G. Field (1990), « Genetic diversity in Sargasso Sea bacterioplankton », *in*: *Nature* 345.*6270*, pp. 60–63, ISSN: 1476-4687, DOI: 10.1038/345060a0.

Girard, Jean, Goulven Lanneau, Ludovic Delage, Cédric Leroux, Arnaud Belcour, Jeanne Got, Jonas Collén, Catherine Boyen, Anne Siegel, Simon M. Dittami, Catherine Leblanc, and Gabriel V. Markov (2021), « Semi-Quantitative Targeted Gas Chromatography-Mass Spectrometry Profiling Supports a Late Side-Chain Reductase Cycloartenol-to-Cholesterol Biosynthesis Pathway in Brown Algae », *in*: *Frontiers in Plant Science* 12, Publisher: Frontiers, p. 740, ISSN: 1664-462X, DOI: 10.3389/fpls.2021.648426.

Green, M. L. and Peter D. Karp (2006), « The outcomes of pathway database computations depend on pathway ontology », *in*: *Nucleic Acids Research* 34.*13*, pp. 3687–3697, ISSN: 1362-4962, DOI: 10.1093/nar/gkl438.

Green, Michelle L. and Peter D. Karp (2004), « A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases », *in*: *BMC Bioinformatics* 5.*1*, p. 76, ISSN: 1471-2105, DOI: 10.1186/1471-2105-5-76.

Gu, Changdai, Gi Bae Kim, Won Jun Kim, Hyun Uk Kim, and Sang Yup Lee (2019), « Current status and applications of genome-scale metabolic models », *in*: *Genome Biology* 20, ISSN: 1474-7596, DOI: 10.1186/s13059-019-1730-3.

Haag, Eric S. and John R. True (2018), « Developmental System Drift », *in*: *Evolutionary Developmental Biology: A Reference Guide*, ed. by Laura Nuno de la Rosa and Gerd Müller, Cham: Springer International Publishing, pp. 1–12, DOI: 10.1007/978-3-319-33038-9_83-1.

Hamilton, Joshua J. and Jennifer L. Reed (2012), « Identification of Functional Differences in Metabolic Networks Using Comparative Genomics and Constraint-Based Models », *in*: *PLoS ONE* 7.*4*, ISSN: 1932-6203, DOI: 10.1371/journal.pone.0034670.

Handorf, Thomas, Oliver Ebenhöh, and Reinhart Heinrich (2005), « Expanding metabolic networks: Scopes of compounds, robustness, and evolution », *in*: *Journal of Molecular Evolution* 61.*4*, pp. 498–512, ISSN: 00222844, DOI: 10.1007/s00239-005-0027-1.

Hari, Archana and Daniel Lobo (2020), « Fluxer: a web application to compute, analyze and visualize genome-scale metabolic flux networks », *in*: *Nucleic Acids Research* 48 (W1), W427–W435, ISSN: 1362-4962, DOI: 10.1093/nar/gkaa409.

Hart, Kathryn M., Michael J. Harms, Bryan H. Schmidt, Carolyn Elya, Joseph W. Thornton, and Susan Marqusee (Nov. 11, 2014), « Thermodynamic System Drift in Protein Evolution », *in*: *PLOS Biology* 12.*11*, e1001994, ISSN: 1545-7885, DOI: 10.1371/journal.pbio.1001994.

Haubrich, Brad A., Emily K. Collins, Alicia L. Howard, Qian Wang, William J. Snell, Matthew B. Miller, Crista D. Thomas, Stephanie K. Pleasant, and W. David Nes (2015), « Characterization, mutagenesis and mechanistic analysis of an ancient algal sterol C24-methyltransferase: Implications for understanding sterol evolution in the green lineage », *in*: *Phytochemistry*, Special Issue in honor of Professor Vincenzo De Luca's 60th birthday 113, pp. 64–72, ISSN: 0031-9422, DOI: 10.1016/j.phytochem.2014.07.019.

Heger, Andreas, Swapan Mallick, Christopher Wilton, and Liisa Holm (Sept. 15, 2007), « The global trace graph, a novel paradigm for searching protein sequence databases », *in*: *Bioinformatics* 23.*18*, pp. 2361–2367, ISSN: 1367-4811, DOI: 10.1093/bioinformatics/btm358.

Heikinheimo, Liisa and Pentti Somerharju (2002), « Translocation of Phosphatidylthreonine and -serine to Mitochondria Diminishes Exponentially with Increasing Molecular Hydrophobicity », *in*: *Traffic* 3.*5*, pp. 367–377, ISSN: 1600-0854, DOI: 10.1034/j.1600-0854.2002.30506.x.

Henry, Christopher S., Hans C. Bernstein, Pamela Weisenhorn, Ronald C. Taylor, Joon-Yong Lee, Jeremy Zucker, and Hyun-Seob Song (2016), « Microbial Community Metabolic Modeling: A Community Data-Driven Network Reconstruction », *in*: *Journal of Cellular Physiology* 231.*11*, pp. 2339–2345, ISSN: 1097-4652, DOI: 10.1002/jcp.25428.

Henry, Christopher S., Matthew DeJongh, Aaron A. Best, Paul M. Frybarger, Ben Linsay, and Rick L. Stevens (2010), « High-throughput generation, optimization and analysis of genome-scale metabolic models », *in*: *Nature Biotechnology* 28.*9*, pp. 977–982, ISSN: 1546-1696, DOI: 10.1038/nbt.1672.

Herrgård, Markus J., Neil Swainston, Paul Dobson, Warwick B. Dunn, K. Yalçin Arga, Mikko Arvas, Nils Blüthgen, Simon Borger, Roeland Costenoble, Matthias Heinemann, Michael Hucka, Nicolas Le Novère, Peter Li, Wolfram Liebermeister, Monica L. Mo, Ana Paula Oliveira, Dina Petranovic, Stephen Pettifer, Evangelos Simeonidis, Kieran Smallbone, Irena Spasié, Dieter Weichart, Roger Brent, David S. Broomhead, Hans V. Westerhoff, Betül Kürdar, Merja Penttilä, Edda Klipp, Bernhard Ø Palsson, Uwe Sauer, Stephen G. Oliver, Pedro Mendes, Jens Nielsen, and Douglas B. Kell (2008), « A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology », *in*: *Nature Biotechnology* 26.*10*, pp. 1155–1160, DOI: 10.1038/nbt1492.

Hillmann, Benjamin, Gabriel A. Al-Ghalith, Robin R. Shields-Cutler, Qiyun Zhu, Daryl M. Gohl, Kenneth B. Beckman, Rob Knight, and Dan Knights (2018), « Evaluating the Information Content of Shallow Shotgun Metagenomics », *in*: *mSystems* 3.*6*, e00069–18, DOI: 10.1128/mSystems.00069-18.

Horgen, F. David, Bryan Sakamoto, and Paul J. Scheuer (2000), « New Triterpenoid Sulfates from the Red Alga Tricleocarpa fragilis », *in*: *Journal of Natural Products* 63.*2*, pp. 210–216, ISSN: 0163-3864, DOI: 10.1021/np990448h.

Hucka, Michael, Frank T. Bergmann, Claudine Chaouiya, Andreas Dräger, Stefan Hoops, Sarah M. Keating, Matthias König, Nicolas Le Novère, Chris J. Myers, Brett G. Olivier, Sven Sahle, James C. Schaff, Rahuman Sheriff, Lucian P. Smith, Dagmar Waltemath, Darren J. Wilkinson, and Fengkai Zhang (2019), « The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 2 Core Release 2 », *in*: *Journal of Integrative Bioinformatics* 16.*2*, ISSN: 1613-4516, DOI: 10.1515/jib-2019-0021.

Hucka, Michael, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novère, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, J. Wang, and and the rest of the SBML Forum: (2003), « The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models », *in*: *Bioinformatics* 19.*4*, pp. 524–531, ISSN: 1367-4803, DOI: 10.1093/bioinformatics/btg015.

Huerta-Cepas, Jaime, François Serra, and Peer Bork (2016), « ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data », *in*: *Molecular Biology and Evolution* 33.*6*, pp. 1635–1638, ISSN: 0737-4038, DOI: 10.1093/molbev/msw046.

Iwai, Shoko, Thomas Weinmaier, Brian L. Schmidt, Donna G. Albertson, Neil J. Poloso, Karim Dabbagh, and Todd Z. DeSantis (2016), « Piphillin: Improved Prediction of Metagenomic Content by Direct Inference from Human Microbiomes », *in*: *PLOS ONE* 11.*11*, e0166104, ISSN: 1932-6203, DOI: 10.1371/journal.pone.0166104.

Johnson, Jethro S., Daniel J. Spakowicz, Bo-Young Hong, Lauren M. Petersen, Patrick Demkowicz, Lei Chen, Shana R. Leopold, Blake M. Hanson, Hanako O. Agresta, Mark Gerstein, Erica Sodergren, and George M. Weinstock (2019), « Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis », *in*: *Nature Communications* 10.*1*, p. 5029, ISSN: 2041-1723, DOI: 10.1038/s41467-019-13036-1.

Judge, Ayesha and Michael S. Dodd (2020), « Metabolism », *in*: *Essays in Biochemistry* 64.*4*, pp. 607–647, ISSN: 0071-1365, DOI: 10.1042/EBC20190041.

Julien-Laferrière, Alice, Laurent Bulteau, Delphine Parrot, Alberto Marchetti-Spaccamela, Leen Stougie, Susana Vinga, Arnaud Mary, and Marie-France Sagot (2016), « A Combinatorial Algorithm for Microbial Consortia Synthetic Design », *in*: *Scientific Reports* 6, p. 29182, ISSN: 2045-2322, DOI: 10.1038/srep29182.

Jun, Se-Ran, Michael S. Robeson, Loren J. Hauser, Christopher W. Schadt, and Andrey A. Gorin (2015), « PanFP: pangenome-based functional profiles for microbial communities », *in*: *BMC Research Notes* 8.*1*, p. 479, ISSN: 1756-0500, DOI: 10.1186/s13104-015-1462-8.

Kamio, Michiya, Cynthia E. Kicklighter, Linh Nguyen, Markus W. Germann, and Charles D. Derby (2011), « Isolation and Structural Elucidation of Novel Mycosporine-Like Amino Acids as Alarm Cues in the Defensive Ink Secretion of the Sea Hare Aplysia californica », *in*: *Helvetica Chimica Acta* 94.*6*, pp. 1012–1018, ISSN: 1522-2675, DOI: 10.1002/hlca.201100117.

Kanehisa, Minoru, Michihiro Araki, Susumu Goto, Masahiro Hattori, Mika Hirakawa, Masumi Itoh, Toshiaki Katayama, Shuichi Kawashima, Shujiro Okuda, Toshiaki Tokimatsu, and Yoshihiro Yamanishi (2008), « KEGG for linking genomes to life and the environment », *in*: *Nucleic Acids Research* 36 (Database issue), pp. D480–484, ISSN: 1362-4962, DOI: 10.1093/nar/gkm882.

Kanehisa, Minoru, Yoko Sato, and Masayuki Kawashima (2022), « KEGG mapping tools for uncovering hidden features in biological data », *in*: *Protein Science* 31.*1*, pp. 47–53, ISSN: 0961-8368, 1469-896X, DOI: 10.1002/pro.4172.

Karimi, Elham, Enora Geslain, Arnaud Belcour, Clémence Frioux, Méziane Aïte, Anne Siegel, Erwan Corre, and Simon M. Dittami (2021), « Robustness analysis of metabolic predictions in algal microbial communities based on different annotation pipelines », *in*: *PeerJ* 9, e11344, ISSN: 2167-8359, DOI: 10.7717/peerj.11344.

Karlsen, Emil, Christian Schulz, and Eivind Almaas (2018), « Automated generation of genome-scale metabolic draft reconstructions based on KEGG », *in*: *BMC Bioinformatics* 19.*1*, p. 467, ISSN: 1471-2105, DOI: 10.1186/s12859-018-2472-z.

Karp, Peter D., Mario Latendresse, and Ron Caspi (2011), « The pathway tools pathway prediction algorithm », *in*: *Standards in Genomic Sciences* 5.*3*, pp. 424–429, ISSN: 1944-3277, DOI: 10.4056/sigs.1794338.

Karp, Peter D., Peter E Midford, Richard Billington, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Wai Kit Ong, Pallavi Subhraveti, Ron Caspi, Carol Fulcher, Ingrid M Keseler, and Suzanne M Paley (Dec. 2019), « Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology », *in*: *Briefings in Bioinformatics*, bbz104, ISSN: 1477-4054, DOI: 10.1093/bib/bbz104.

— (2021), « Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology », *in*: *Briefings in Bioinformatics* 22.*1*, pp. 109–126, ISSN: 1477-4054, DOI: 10.1093/bib/bbz104.

Karp, Peter D., Wai Kit Ong, Suzanne Paley, Richard Billington, Ron Caspi, Carol Fulcher, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Peter E. Midford, Pallavi Subhraveti, Socorro Gama-Castro, Luis Muñiz-Rascado, César Bonavides-Martinez, Alberto Santos-Zavaleta, Amanda Mackie, Julio Collado-Vides, Ingrid M. Keseler, and Ian Paulsen (2018), « The EcoCyc Database », *in*: *EcoSal Plus* 8.*1*, ISSN: 2324-6200, DOI: 10.1128/ecosalplus.ESP-0006-2018.

Karp, Peter D., Suzanne Paley, and Pedro Romero (2002a), « The Pathway Tools software », *in*: *Bioinformatics* 18 (suppl_1), S225–S232, ISSN: 1367-4803, DOI: 10.1093/bioinformatics/18.suppl_1.S225.

Karp, Peter D., Suzanne M. Paley, Markus Krummenacker, Mario Latendresse, Joseph M. Dale, Thomas J. Lee, Pallavi Kaipa, Fred Gilham, Aaron Spaulding, Liviu Popescu, Tomer Altman, Ian Paulsen, Ingrid M. Keseler, and Ron Caspi (2010), « Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology », *in*: *Briefings in Bioinformatics* 11.*1*, pp. 40–79, ISSN: 1467-5463, DOI: 10.1093/bib/bbp043.

Karp, Peter D., Monica Riley, Milton Saier, Ian T. Paulsen, Julio Collado-Vides, Suzanne M. Paley, Alida Pellegrini-Toole, César Bonavides, and Socorro Gama-Castro (2002b), « The EcoCyc Database », *in*: *Nucleic Acids Research* 30.*1*, pp. 56–58, ISSN: 0305-1048.

Katinka, Michaël D., Simone Duprat, Emmanuel Cornillot, Guy Méténier, Fabienne Thomarat, Gérard Prensier, Valérie Barbe, Eric Peyretaillade, Philippe Brottier, Patrick Wincker, Frédéric Delbac, Hicham El Alaoui, Pierre Peyret, William Saurin, Manolo Gouy, Jean Weissenbach, and Christian P. Vivarès (2001), « Genome sequence and gene compaction of

the eukaryote parasite Encephalitozoon cuniculi », *in*: *Nature* 414.*6862*, pp. 450–453, ISSN: 1476-4687, DOI: 10.1038/35106579.

Keegan, Kevin P., Elizabeth M. Glass, and Folker Meyer (2016), « MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function », *in*: *Methods in Molecular Biology* 1399, pp. 207–233, ISSN: 1940-6029, DOI: 10.1007/978-1-4939-3369-3_13.

Keseler, Ingrid M., César Bonavides-Martínez, Julio Collado-Vides, Socorro Gama-Castro, Robert P. Gunsalus, D. Aaron Johnson, Markus Krummenacker, Laura M. Nolan, Suzanne Paley, Ian T. Paulsen, Martin Peralta-Gil, Alberto Santos-Zavaleta, Alexander Glennon Shearer, and Peter D. Karp (2009), « EcoCyc: a comprehensive view of Escherichia coli biology », *in*: *Nucleic Acids Research* 37 (Database issue), pp. D464–470, ISSN: 1362-4962, DOI: 10.1093/nar/gkn751.

Keseler, Ingrid M., Socorro Gama-Castro, Amanda Mackie, Richard Billington, César Bonavides-Martínez, Ron Caspi, Anamika Kothari, Markus Krummenacker, Peter E. Midford, Luis Muñiz-Rascado, Wai Kit Ong, Suzanne Paley, Alberto Santos-Zavaleta, Pallavi Subhraveti, Víctor H. Tierrafría, Alan J. Wolfe, Julio Collado-Vides, Ian T. Paulsen, and Peter D. Karp (2021), « The EcoCyc Database in 2021 », *in*: *Frontiers in Microbiology* 12, p. 2098, ISSN: 1664-302X, DOI: 10.3389/fmicb.2021.711077.

Khandelwal, Ruchir A., Brett G. Olivier, Wilfred F. M. Röling, Bas Teusink, and Frank J. Bruggeman (2013), « Community Flux Balance Analysis for Microbial Consortia at Balanced Growth », *in*: *PLOS ONE* 8.*5*, e64567, ISSN: 1932-6203, DOI: 10.1371/journal.pone.0064567.

Kim, Eun-Youn, Daniel Ashlock, and Sung Ho Yoon (2019), « Identification of critical connectors in the directed reaction-centric graphs of microbial metabolic networks », *in*: *BMC Bioinformatics* 20.*1*, p. 328, ISSN: 1471-2105, DOI: 10.1186/s12859-019-2897-z.

King, Zachary A., Andreas Dräger, Ali Ebrahim, Nikolaus Sonnenschein, Nathan E. Lewis, and Bernhard O. Palsson (2015), « Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways », *in*: *PLOS Computational Biology* 11.*8*, e1004321, ISSN: 1553-7358, DOI: 10.1371/journal.pcbi.1004321.

King, Zachary A., Justin Lu, Andreas Dräger, Philip Miller, Stephen Federowicz, Joshua A. Lerman, Ali Ebrahim, Bernhard O. Palsson, and Nathan E. Lewis (2016), « BiGG Models: A platform for integrating, standardizing and sharing genome-scale models », *in*: *Nucleic Acids Research* 44 (D1), pp. D515–D522, ISSN: 0305-1048, DOI: 10.1093/nar/gkv1049.

Klamt, Steffen, Utz-Uwe Haus, and Fabian Theis (2009), « Hypergraphs and Cellular Networks », *in*: *PLOS Computational Biology* 5.*5*, e1000385, ISSN: 1553-7358, DOI: 10.1371/journal.pcbi.1000385.

Kruse, Kai and Oliver Ebenhöh (2008), « Comparing flux balance analysis to network expansion: producibility, sustainability and the scope of compounds », *in*: *Genome Informatics. International Conference on Genome Informatics* 20, pp. 91–101, ISSN: 0919-9454.

Kumar, Manish, Boyang Ji, Karsten Zengler, and Jens Nielsen (2019), « Modelling approaches for studying the microbiome », *in*: *Nature Microbiology* 4.*8*, pp. 1253–1267, ISSN: 2058-5276, DOI: 10.1038/s41564-019-0491-9.

Lalegerie, Fanny, Sirine Lajili, Gilles Bedoux, Laure Taupin, Valérie Stiger-Pouvreau, and Solène Connan (2019), « Photo-protective compounds in red macroalgae from Brittany: Considerable diversity in mycosporine-like amino acids (MAAs) », *in*: *Marine Environmental Research* 147, pp. 37–48, ISSN: 0141-1136, DOI: 10.1016/j.marenvres.2019.04.001.

Langille, Morgan G. I., Jesse Zaneveld, J. Gregory Caporaso, Daniel McDonald, Dan Knights, Joshua A. Reyes, Jose C. Clemente, Deron E. Burkepile, Rebecca L. Vega Thurber, Rob Knight, Robert G. Beiko, and Curtis Huttenhower (2013), « Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences », *in*: *Nature Biotechnology* 31.*9*, pp. 814–821, ISSN: 1546-1696, DOI: 10.1038/nbt.2676.

Le Boulch, Malo, Patrice Déhais, Sylvie Combes, and Géraldine Pascal (2019), « The MACADAM database: a MetAboliC pAthways DAtabase for Microbial taxonomic groups for mining potential metabolic capacities of archaeal and bacterial taxonomic groups », *in*: *Database* 2019 (baz049), ISSN: 1758-0463, DOI: 10.1093/database/baz049.

Lee, Thomas J., Ian Paulsen, and Peter D. Karp (2008), « Annotation-based inference of transporter function », *in*: *Bioinformatics* 24.*13*, pp. i259–i267, ISSN: 1367-4803, DOI: 10.1093/bioinformatics/btn180.

Levesque, H J (1986), « Knowledge Representation and Reasoning », *in*: *Annual Review of Computer Science* 1.*1*, pp. 255–287, DOI: 10.1146/annurev.cs.01.060186.001351.

Lifschitz, Vladimir (2008), « What Is Answer Set Programming? », *in*: *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, AAAI Conference on Artificial Intelligence, AAAI Press, pp. 1594–1597.

Lim, Jun Wei, Tansol Park, Yen Wah Tong, and Zhongtang Yu (2020), « Chapter One - The microbiome driving anaerobic digestion and microbial analysis », *in*: *Advances in Bioenergy*, ed. by Yebo Li and Samir Kumar Khanal, vol. 5, Elsevier, pp. 1–61, DOI: 10.1016/bs.aibe.2020.04.001.

Liu, Xiaojing and Jason W. Locasale (2017), « Metabolomics: A Primer », *in*: *Trends in Biochemical Sciences* 42.*4*, pp. 274–284, ISSN: 0968-0004, DOI: 10.1016/j.tibs.2017.01.004.

Lobb, Briallen, Benjamin Jean-Marie Tremblay, Gabriel Moreno-Hagelsieb, and Andrew C. Doxey (2020), « An assessment of genome annotation coverage across the bacterial tree of life », *in*: *Microbial Genomics* 6.*3*, e000341, ISSN: 2057-5858, DOI: 10.1099/mgen.0.000341.

Louca, Stilianos, Laura Wegener Parfrey, and Michael Doebeli (2016), « Decoupling function and taxonomy in the global ocean microbiome », *in*: *Science* 353.*6305*, pp. 1272–1277, DOI: 10.1126/science.aaf4507.

Lu, H., F. Li, B. J. Sánchez, Z. Zhu, G. Li, I. Domenzain, S. Marcišauskas, P. M. Anton, D. Lappa, C. Lieven, M. E. Beber, N. Sonnenschein, E. J. Kerkhoven, and J. Nielsen (Aug. 2019), « A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism », *in*: *Nat Commun* 10.*1*, [PubMed Central:PMC6687777] [DOI:10.1038/s41467-019-11581-3] [PubMed:26582926], p. 3586.

Machado, Daniel, Sergej Andrejev, Melanie Tramontano, and Kiran Raosaheb Patil (2018), « Fast automated reconstruction of genome-scale metabolic models for microbial species and communities », *in*: *Nucleic Acids Research* 46.*15*, pp. 7542–7553, ISSN: 0305-1048, DOI: 10.1093/nar/gky537.

Machado, Daniel, Markus J. Herrgård, and Isabel Rocha (2016), « Stoichiometric Representation of Gene–Protein–Reaction Associations Leverages Constraint-Based Analysis from Reaction to Gene-Level Phenotype Prediction », *in*: *PLoS Computational Biology* 12.*10*, e1005140, DOI: 10.1371/journal.pcbi.1005140.

Machado, Daniel, Oleksandr M. Maistrenko, Sergej Andrejev, Yongkyu Kim, Peer Bork, Kaustubh R. Patil, and Kiran R. Patil (2021), « Polarization of microbial communities between competitive and cooperative metabolism », *in*: *Nature Ecology & Evolution* 5.*2*, pp. 195–203, ISSN: 2397-334X, DOI: 10.1038/s41559-020-01353-4.

Magnúsdóttir, Stefanía, Almut Heinken, Laura Kutt, Dmitry A. Ravcheev, Eugen Bauer, Alberto Noronha, Kacy Greenhalgh, Christian Jäger, Joanna Baginska, Paul Wilmes, Ronan M. T. Fleming, and Ines Thiele (2017), « Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota », *in*: *Nature Biotechnology* 35.*1*, pp. 81–89, ISSN: 1546-1696, DOI: 10.1038/nbt.3703.

Manquinho, Vasco, Joao Marques-Silva, and Jordi Planes (2009), « Algorithms for Weighted Boolean Optimization », *in*: *Theory and Applications of Satisfiability Testing - SAT 2009*, ed. by Oliver Kullmann, Lecture Notes in Computer Science, Berlin, Heidelberg: Springer, pp. 495–508, ISBN: 978-3-642-02777-2, DOI: 10.1007/978-3-642-02777-2_45.

Martin, Alyce M., Emily W. Sun, Geraint B. Rogers, and Damien J. Keating (2019), « The Influence of the Gut Microbiome on Host Metabolism Through the Regulation of Gut Hormone Release », *in*: 10, ISSN: 1664-042X.

Mas, Alix, Shahrad Jamshidi, Yvan Lagadeuc, Damien Eveillard, and Philippe Vandenkoornhuyse (2016), « Beyond the Black Queen Hypothesis », *in*: *The ISME Journal* 10.*9*, pp. 2085–2091, ISSN: 1751-7370, DOI: 10.1038/ismej.2016.22.

McDonald, Andrew G., Sinéad Boyce, and Keith F. Tipton (2009), « ExplorEnz: the primary source of the IUBMB enzyme list », *in*: *Nucleic Acids Research* 37 (suppl_1), pp. D593–D597, ISSN: 0305-1048, DOI: 10.1093/nar/gkn582.

McDonald, Andrew G. and Keith F. Tipton (2014), « Fifty-five years of enzyme classification: advances and difficulties », *in*: *The FEBS Journal* 281.*2*, pp. 583–592, ISSN: 1742-4658, DOI: 10.1111/febs.12530.

McKinney, Wes (2010), « Data Structures for Statistical Computing in Python », *in*: *Proceedings of the 9th Python in Science Conference*, Python in Science Conference, ed. by Stéfan van der Walt and Jarrod Millman, Proceedings of the Python in Science Conference, SciPy, pp. 56–61, DOI: 10.25080/Majora-92bf1922-00a.

Mendes-Soares, Helena, Michael Mundy, Luis Mendes Soares, and Nicholas Chia (2016), « MMinte: an application for predicting metabolic interactions among the microbial species in a community », *in*: *BMC Bioinformatics* 17.*1*, p. 343, ISSN: 1471-2105, DOI: 10.1186/s12859-016-1230-3.

Mendoza, Sebastián N., Brett G. Olivier, Douwe Molenaar, and Bas Teusink (2019), « A systematic assessment of current genome-scale metabolic reconstruction tools », *in*: *Genome Biology* 20.*1*, p. 158, ISSN: 1474-760X, DOI: 10.1186/s13059-019-1769-1.

Mistry, Jaina, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik L L Sonnhammer, Silvio C E Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, Robert D Finn, and Alex Bateman (2021), « Pfam: The protein families database in 2021 », *in*: *Nucleic Acids Research* 49 (D1), pp. D412–D419, ISSN: 0305-1048, DOI: 10.1093/nar/gkaa913.

Mongad, Dattatray S., Nikeeta S. Chavan, Nitin P. Narwade, Kunal Dixit, Yogesh S. Shouche, and Dhiraj P. Dhotre (2021), « MicFunPred: A conserved approach to predict functional profiles from 16S rRNA gene sequence data », *in*: *Genomics* 113.*6*, pp. 3635–3643, ISSN: 08887543, DOI: 10.1016/j.ygeno.2021.08.016.

Moretti, Sébastien, Van Du T Tran, Florence Mehl, Mark Ibberson, and Marco Pagni (2021), « MetaNetX/MNXref: unified namespace for metabolites and biochemical reactions in the context of metabolic models », *in*: *Nucleic Acids Research* 49 (D1), pp. D570–D574, ISSN: 0305-1048, DOI: 10.1093/nar/gkaa992.

Morgado, Antonio, Federico Heras, and Joao Marques-Silva (2012), « Improvements to Core-Guided Binary Search for MaxSAT », *in*: *Theory and Applications of Satisfiability Testing – SAT 2012*, ed. by Alessandro Cimatti and Roberto Sebastiani, Lecture Notes in Computer Science, Berlin, Heidelberg: Springer, pp. 284–297, ISBN: 978-3-642-31612-8, DOI: 10.1007/978-3-642-31612-8_22.

Morris, Brandon E.L., Ruth Henneberger, Harald Huber, and Christine Moissl-Eichinger (2013), « Microbial syntrophy: interaction for the common good », *in*: *FEMS Microbiology Reviews* 37.*3*, pp. 384–406, ISSN: 0168-6445, DOI: 10.1111/1574-6976.12019.

Morris, J. Jeffrey, Richard E. Lenski, and Erik R. Zinser (2012), « The Black Queen Hypothesis: Evolution of Dependencies through Adaptive Gene Loss », *in*: *mBio* 3.*2*, e00036–12, DOI: 10.1128/mBio.00036-12.

Narayan, Nicole R., Thomas Weinmaier, Emilio J. Laserna-Mendieta, Marcus J. Claesson, Fergus Shanahan, Karim Dabbagh, Shoko Iwai, and Todd Z. DeSantis (2020), « Piphillin predicts metagenomic composition and dynamics from DADA2-corrected 16S rDNA sequences », *in*: *BMC Genomics* 21.*1*, p. 56, ISSN: 1471-2164, DOI: 10.1186/s12864-019-6427-1.

Neelakandan, Anjanasree K., Zhihong Song, Junqing Wang, Matthew H. Richards, Xiaolei Wu, Babu Valliyodan, Henry T. Nguyen, and W. David Nes (2009), « Cloning, functional expression and phylogenetic analysis of plant sterol 24C-methyltransferases involved in sitosterol biosynthesis », *in*: *Phytochemistry* 70.*17*, pp. 1982–1998, ISSN: 0031-9422, DOI: 10.1016/j.phytochem.2009.09.003.

Nègre, Delphine, Méziane Aite, Arnaud Belcour, Clémence Frioux, Loraine Brillet-Guéguen, Xi Liu, Philippe Bordron, Olivier Godfroy, Agnieszka P. Lipinska, Catherine Leblanc, Anne Siegel, Simon M. Dittami, Erwan Corre, and Gabriel V. Markov (2019), « Genome–Scale Metabolic Networks Shed Light on the Carotenoid Biosynthesis Pathway in the Brown Algae Saccharina japonica and Cladosiphon okamuranus », *in*: *Antioxidants* 8.*11*, p. 564, DOI: 10.3390/antiox8110564.

Nielsen, Jens (2017), « Systems Biology of Metabolism », *in*: *Annual Review of Biochemistry* 86.*1*, pp. 245–275, ISSN: 0066-4154, 1545-4509, DOI: 10.1146/annurev-biochem-061516-044757.

Noecker, Cecilia, Alexander Eng, Efrat Muller, and Elhanan Borenstein (2022), « MIMOSA2: a metabolic network-based tool for inferring mechanism-supported relationships in microbiome-metabolome data », *in*: *Bioinformatics* 38.*6*, pp. 1615–1623, ISSN: 1367-4803, DOI: 10.1093/bioinformatics/btac003.

Noecker, Cecilia, Alexander Eng, Sujatha Srinivasan, Casey M. Theriot, Vincent B. Young, Janet K. Jansson, David N. Fredricks, and Elhanan Borenstein (2016), « Metabolic Model-Based Integration of Microbiome Taxonomic and Metabolomic Profiles Elucidates Mechanistic Links between Ecological and Metabolic Variation », *in*: *mSystems* 1.*1*, e00013–15, DOI: 10.1128/mSystems.00013-15.

Noronha, Alberto, Jennifer Modamio, Yohan Jarosz, Elisabeth Guerard, Nicolas Sompairac, German Preciat, Anna Dröfn Daníelsdóttir, Max Krecke, Diane Merten, Hulda S Haraldsdóttir, Almut Heinken, Laurent Heirendt, Stefanía Magnúsdóttir, Dmitry A Ravcheev, Swagatika

Sahoo, Piotr Gawron, Lucia Friscioni, Beatriz Garcia, Mabel Prendergast, Alberto Puente, Mariana Rodrigues, Akansha Roy, Mouss Rouquaya, Luca Wiltgen, Alise Žagare, Elisabeth John, Maren Krueger, Inna Kuperstein, Andrei Zinovyev, Reinhard Schneider, Ronan M T Fleming, and Ines Thiele (2018), « The Virtual Metabolic Human database: integrating human and gut microbiome metabolism with nutrition and disease », *in*: *Nucleic Acids Research* 47.*D1*, pp. D614–D624, ISSN: 0305-1048, DOI: 10.1093/nar/gky992.

Oberhardt, Matthew A., Jacek Puchałka, Vítor A. P. Martins dos Santos, and Jason A. Papin (2011), « Reconciliation of Genome-Scale Metabolic Reconstructions for Comparative Systems Analysis », *in*: *PLOS Computational Biology* 7.*3*, e1001116, ISSN: 1553-7358, DOI: 10.1371/journal.pcbi.1001116.

Ogier, Jean-Claude, Sylvie Pagès, Maxime Galan, Matthieu Barret, and Sophie Gaudriault (2019), « rpoB, a promising marker for analyzing the diversity of bacterial communities by amplicon sequencing », *in*: *BMC Microbiology* 19.*1*, p. 171, ISSN: 1471-2180, DOI: 10.1186/s12866-019-1546-z.

Oh, Min and Liqing Zhang (2020), « DeepMicro: deep representation learning for disease prediction based on microbiome data », *in*: *Scientific Reports* 10.*1*, ISSN: 2045-2322, DOI: 10.1038/s41598-020-63159-5.

Oksanen, Jari, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs, and Helene Wagner (2020), *vegan: Community Ecology Package*, URL: https://CRAN.R-project.org/package=vegan.

Opatovsky, Itai, Diego Santos-Garcia, Zhepu Ruan, Tamar Lahav, Shany Ofaim, Laurence Mouton, Valérie Barbe, Jiandong Jiang, Einat Zchori-Fein, and Shiri Freilich (2018), « Modeling trophic dependencies and exchanges among insects' bacterial symbionts in a host-simulated environment », *in*: *BMC Genomics* 19.*1*, p. 402, ISSN: 1471-2164, DOI: 10.1186/s12864-018-4786-7.

Orfanoudaki, Maria, Anja Hartmann, Ulf Karsten, and Markus Ganzera (2019), « Chemical profiling of mycosporine-like amino acids in twenty-three red algal species », *in*: *Journal of Phycology* 55.*2*, pp. 393–403, ISSN: 1529-8817, DOI: 10.1111/jpy.12827.

Orth, Jeffrey D., Ines Thiele, and Bernhard Ø. Palsson (2010), « What is Flux Balance Analysis ? », *in*: *Nature biotechnology* 28.*3*, pp. 245–248, ISSN: 1546-1696, DOI: 10.1038/nbt.1614, arXiv: NIHMS150003.

Overbeek, Ross, Michael Fonstein, Mark D'Souza, Gordon D. Pusch, and Natalia Maltsev (1999), « The use of gene clusters to infer functional coupling », *in*: *Proceedings of the National Academy of Sciences of the United States of America* 96.*6*, DOI: 10.1073/pnas.96.6.2896.

Overbeek, Ross, Robert Olson, Gordon D. Pusch, Gary J. Olsen, James J. Davis, Terry Disz, Robert A. Edwards, Svetlana Gerdes, Bruce Parrello, Maulik Shukla, Veronika Vonstein, Alice R. Wattam, Fangfang Xia, and Rick Stevens (2014), « The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST) », *in*: *Nucleic Acids Research* 42 (Database issue), pp. D206–214, ISSN: 1362-4962, DOI: 10.1093/nar/gkt1226.

Paley, Suzanne, Richard Billington, James Herson, Markus Krummenacker, and Peter D. Karp (2021), « Pathway Tools Visualization of Organism-Scale Metabolic Networks », *in*: *Metabolites* 11.*2*, p. 64, ISSN: 2218-1989, DOI: 10.3390/metabo11020064.

Papin, Jason A., Nathan D. Price, Sharon J. Wiback, David A. Fell, and Bernhard O. Palsson (2003), « Metabolic pathways in the post-genome era », *in*: *Trends in Biochemical Sciences* 28.*5*, pp. 250–258, ISSN: 0968-0004, DOI: 10.1016/S0968-0004(03)00064-1.

Parks, Donovan H., Maria Chuvochina, Pierre-Alain Chaumeil, Christian Rinke, Aaron J. Mussig, and Philip Hugenholtz (2020), « A complete domain-to-species taxonomy for Bacteria and Archaea », *in*: *Nature Biotechnology* 38.*9*, pp. 1079–1086, ISSN: 1546-1696, DOI: 10.1038/s41587-020-0501-8.

Parks, Donovan H., Maria Chuvochina, Christian Rinke, Aaron J Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz (2022), « GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy », *in*: *Nucleic Acids Research* 50 (D1), pp. D785–D794, ISSN: 0305-1048, DOI: 10.1093/nar/gkab776.

Parks, Donovan H., Maria Chuvochina, David W. Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz (2018), « A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life », *in*: *Nature Biotechnology* 36.*10*, pp. 996–1004, DOI: 10.1038/nbt.4229.

Pasolli, Edoardo, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, Paolo Manghi, Adrian Tett, Paolo Ghensi, Maria Carmen Collado, Benjamin L. Rice, Casey DuLong, Xochitl C. Morgan, Christopher D. Golden, Christopher Quince, Curtis Huttenhower, and Nicola Segata (2019), « Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle », *in*: *Cell* 176.*3*, 649–662.e20, ISSN: 0092-8674, DOI: 10.1016/j.cell.2019.01.001.

Patel, Bhavesh H., Claudia Percivalle, Dougal J. Ritson, Colm D. Duffy, and John D. Sutherland (2015), « Common origins of RNA, protein and lipid precursors in a cyanosulfidic protometabolism », *in*: *Nature Chemistry* 7.*4*, pp. 301–307, ISSN: 1755-4349, DOI: 10.1038/nchem.2202.

Patumcharoenpol, Preecha, Massalin Nakphaichit, Gianni Panagiotou, Anchalee Senavonge, Narissara Suratannon, and Wanwipa Vongsangnak (2021), « MetGEMs Toolbox: Metagenome-scale models as integrative toolbox for uncovering metabolic functions and routes of human gut microbiome », *in*: *PLOS Computational Biology* 17.*1*, ed. by Sergei L. Kosakovsky Pond, e1008487, ISSN: 1553-7358, DOI: 10.1371/journal.pcbi.1008487.

Pearcy, Nicole, Nadia Chuzhanova, and Jonathan J. Crofts (2016), « Complexity and robustness in hypernetwork models of metabolism », *in*: *Journal of Theoretical Biology* 406, pp. 99–104, ISSN: 1095-8541, DOI: 10.1016/j.jtbi.2016.06.032.

Pearcy, Nicole, Jonathan J. Crofts, and Nadia Chuzhanova (2014), « Hypergraph Models of Metabolism », *in*: *International Journal of Bioengineering and Life Sciences* 8.*8*, pp. 829–833.

Pearson, William R. (2013), « An Introduction to Sequence Similarity ("Homology") Searching », *in*: *Current Protocols in Bioinformatics* 42.*1*, pp. 3.1.1–3.1.8, ISSN: 1934-340X.

Peona, Valentina, Mozes P. K. Blom, Luohao Xu, Reto Burri, Shawn Sullivan, Ignas Bunikis, Ivan Liachko, Tri Haryoko, Knud A. Jønsson, Qi Zhou, Martin Irestedt, and Alexander Suh (2021), « Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise », *in*: *Molecular Ecology Resources* 21.*1*, pp. 263–286, ISSN: 1755-0998, DOI: 10.1111/1755-0998.13252.

Pharkya, Priti, Anthony P. Burgard, and Costas D. Maranas (2004), « OptStrain: A computational framework for redesign of microbial production systems », *in*: *Genome Research* 14.*11*, pp. 2367–2376, ISSN: 1088-9051, 1549-5469, DOI: 10.1101/gr.2872004.

Pitkänen, Esa, Paula Jouhten, Jian Hou, Muhammad Fahad Syed, Peter Blomberg, Jana Kludas, Merja Oja, Liisa Holm, Merja Penttilä, Juho Rousu, and Mikko Arvas (2014), « Comparative Genome-Scale Reconstruction of Gapless Metabolic Networks for Present and Ancestral Species », *in*: *PLOS Computational Biology* 10.*2*, e1003465, ISSN: 1553-7358, DOI: 10.1371/journal.pcbi.1003465.

Price, Dana C., Cheong Xin Chan, Hwan Su Yoon, Eun Chan Yang, Huan Qiu, Andreas P. M. Weber, Rainer Schwacke, Jeferson Gross, Nicolas A. Blouin, Chris Lane, Adrián Reyes-Prieto, Dion G. Durnford, Jonathan A. D. Neilson, B. Franz Lang, Gertraud Burger, Jürgen M. Steiner, Wolfgang Löffelhardt, Jonathan E. Meuser, Matthew C. Posewitz, Steven Ball, Maria Cecilia Arias, Bernard Henrissat, Pedro M. Coutinho, Stefan A. Rensing, Aikaterini Symeonidi, Harshavardhan Doddapaneni, Beverley R. Green, Veeran D. Rajah, Jeffrey Boore, and Debashish Bhattacharya (2012), « Cyanophora paradoxa genome elucidates origin of photosynthesis in algae and plants. », *in*: *Science* 335.*6070*, pp. 843–847, ISSN: 1095-9203, DOI: 10.1126/science.1213561.

Prigent, Sylvain, Guillaume Collet, Simon M. Dittami, Ludovic Delage, Floriane Ethis de Corny, Olivier Dameron, Damien Eveillard, Sven Thiele, Jeanne Cambefort, Catherine Boyen, Anne Siegel, and Thierry Tonon (2014), « The genome-scale metabolic network of Ectocarpus siliculosus (EctoGEM): a resource to study brown algal physiology and beyond », *in*: *The Plant Journal* 80.*2*, pp. 367–381, ISSN: 1365-313X, DOI: 10.1111/tpj.12627.

Prigent, Sylvain, Clémence Frioux, Simon M. Dittami, Sven Thiele, Abdelhalim Larhlimi, Guillaume Collet, Fabien Gutknecht, Jeanne Got, Damien Eveillard, Jérémie Bourdon, Frédéric Plewniak, Thierry Tonon, and Anne Siegel (2017), « Meneco, a Topology-Based Gap-Filling Tool Applicable to Degraded Genome-Wide Metabolic Networks », *in*: *PLOS Computational Biology* 13.*1*, e1005276, ISSN: 1553-7358, DOI: 10.1371/journal.pcbi.1005276.

Prigent, Sylvain, Jens C. Nielsen, Jens C. Frisvad, and Jens Nielsen (2018), « Reconstruction of 24 Penicillium genome-scale metabolic models shows diversity based on their secondary metabolism », *in*: *Biotechnology and Bioengineering* 115.*10*, pp. 2604–2612, ISSN: 1097-0290, DOI: https://doi.org/10.1002/bit.26739.

Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner (2013), « The SILVA ribosomal RNA gene database project: improved data processing and web-based tools », *in*: *Nucleic Acids Research* 41 (D1), pp. D590–D596, ISSN: 0305-1048, DOI: 10.1093/nar/gks1219.

Quince, Christopher, Alan W. Walker, Jared T. Simpson, Nicholas J. Loman, and Nicola Segata (2017), « Shotgun metagenomics, from sampling to analysis », *in*: *Nature Biotechnology* 35.*9*, pp. 833–844, ISSN: 1546-1696, DOI: 10.1038/nbt.3935.

Raffel, Thomas R., Lynn B. Martin, and Jason R. Rohr (2008), « Parasites as predators: unifying natural enemy ecology », *in*: *Trends in Ecology & Evolution* 23.*11*, pp. 610–618, ISSN: 0169-5347, DOI: 10.1016/j.tree.2008.06.015.

Rahman, Syed Asad, Gilliean Torrance, Lorenzo Baldacci, Sergio Martínez Cuesta, Franz Fenninger, Nimish Gopal, Saket Choudhary, John W. May, Gemma L. Holliday, Christoph Steinbeck, and Janet M. Thornton (2016), « Reaction Decoder Tool (RDT): extracting features from chemical reactions », *in*: *Bioinformatics* 32.*13*, pp. 2065–2066, DOI: 10.1093/bioinformatics/btw096.

Reed, Jennifer L., Thuy D. Vo, Christophe H. Schilling, and Bernhard Ø. Palsson (2003), « An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR) », *in*: *Genome Biology* 4.*9*, R54, ISSN: 1474-760X, DOI: 10.1186/gb-2003-4-9-r54.

Reimer, Lorenz Christian, Joaquim Sardà Carbasse, Julia Koblitz, Christian Ebeling, Adam Podstawka, and Jörg Overmann (2022), « BacDive in 2022: the knowledge base for standardized bacterial and archaeal data », *in*: *Nucleic Acids Research* 50 (D1), pp. D741–D746, ISSN: 0305-1048, DOI: 10.1093/nar/gkab961.

Rinke, Christian, Maria Chuvochina, Aaron J. Mussig, Pierre-Alain Chaumeil, Adrián A. Davín, David W. Waite, William B. Whitman, Donovan H. Parks, and Philip Hugenholtz (2021), « A standardized archaeal taxonomy for the Genome Taxonomy Database », *in*: *Nature Microbiology* 6.7, pp. 946–959, ISSN: 2058-5276, DOI: 10.1038/s41564-021-00918-8.

Romero, P. R. and Peter D. Karp (2004), « Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway-genome databases », *in*: *Bioinformatics* 20.5, pp. 709–717, ISSN: 1367-4803, DOI: 10.1093/bioinformatics/btg471.

Royer, Loïc, Matthias Reimann, Bill Andreopoulos, and Michael Schroeder (2008), « Unraveling Protein Networks with Power Graph Analysis », *in*: *PLOS Computational Biology* 4.7, e1000108, ISSN: 1553-7358, DOI: 10.1371/journal.pcbi.1000108.

Ruppert, Krista M., Richard J. Kline, and Md Saydur Rahman (2019), « Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA », *in*: *Global Ecology and Conservation* 17, e00547, ISSN: 2351-9894, DOI: 10.1016/j.gecco.2019.e00547.

Russell, James J., Julie A. Theriot, Pranidhi Sood, Wallace F. Marshall, Laura F. Landweber, Lillian Fritz-Laylin, Jessica K. Polka, Snezhana Oliferenko, Therese Gerbich, Amy Gladfelter, James Umen, Magdalena Bezanilla, Madeline A. Lancaster, Shuonan He, Matthew C. Gibson, Bob Goldstein, Elly M. Tanaka, Chi-Kuo Hu, and Anne Brunet (2017), « Non-model model organisms », *in*: *BMC Biology* 15.1, p. 55, ISSN: 1741-7007, DOI: 10.1186/s12915-017-0391-5.

Saadat, Nima P., Marvin van Aalst, and Oliver Ebenhöh (2022), « Network Reconstruction and Modelling Made Reproducible with moped », *in*: *Metabolites* 12.4, p. 275, ISSN: 2218-1989, DOI: 10.3390/metabo12040275.

Sayers, Eric W, Mark Cavanaugh, Karen Clark, James Ostell, Kim D Pruitt, and Ilene Karsch-Mizrachi (2019), « GenBank », *in*: *Nucleic Acids Research* 47 (Database issue), pp. D94–D99, DOI: 10.1093/nar/gky989.

Schaub, Torsten and Sven Thiele (2009), « Metabolic network expansion with answer set programming », *in*: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5649 LNCS, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 312–326, ISBN: 3642028454, DOI: 10.1007/978-3-642-02846-5_27.

Schellenberger, Jan, Junyoung O. Park, Tom M. Conrad, and Bernhard Ø Palsson (2010), « BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions », *in*: *BMC Bioinformatics* 11.1, p. 213, ISSN: 1471-2105, DOI: 10.1186/1471-2105-11-213.

Schilling, Christophe H., David Letscher, and Bernhard O. Palsson (2000), « Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective », *in*: *Journal of Theoretical Biology* 203.*3*, pp. 229–248, ISSN: 00225193, DOI: 10.1006/jtbi.2000.1073.

Schoch, Conrad L., Stacy Ciufo, Mikhail Domrachev, Carol L. Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, Richard Mcveigh, Kathleen O'Neill, Barbara Robbertse, Shobha Sharma, Vladimir Soussov, John P. Sullivan, Lu Sun, Seán Turner, and Ilene Karsch-Mizrachi (2020), « NCBI Taxonomy: a comprehensive update on curation, resources and tools », *in*: *Database: The Journal of Biological Databases and Curation* 2020, baaa062, ISSN: 1758-0463, DOI: 10.1093/database/baaa062.

Schoch, Conrad L. et al. (2012), « Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi », *in*: *Proceedings of the National Academy of Sciences* 109.*16*, pp. 6241–6246, DOI: 10.1073/pnas.1117018109.

Schulz, Christian and Eivind Almaas (2020), « Genome-scale reconstructions to assess metabolic phylogeny and organism clustering », *in*: *PLOS ONE* 15.*12*, e0240953, ISSN: 1932-6203, DOI: 10.1371/journal.pone.0240953.

Scossa, Federico and Alisdair R. Fernie (2020), « The evolution of metabolism: How to test evolutionary hypotheses at the genomic level », *in*: *Computational and Structural Biotechnology Journal* 18, pp. 482–500, DOI: 10.1016/j.csbj.2020.02.009.

Seaver, Samuel M D, Filipe Liu, Qizhi Zhang, James Jeffryes, José P Faria, Janaka N Edirisinghe, Michael Mundy, Nicholas Chia, Elad Noor, Moritz E Beber, Aaron A Best, Matthew DeJongh, Jeffrey A Kimbrel, Patrik D'haeseleer, Sean R McCorkle, Jay R Bolton, Erik Pearson, Shane Canon, Elisha M Wood-Charlson, Robert W Cottingham, Adam P Arkin, and Christopher S Henry (2021), « The ModelSEED Biochemistry Database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes », *in*: *Nucleic Acids Research* 49 (D1), pp. D575–D588, ISSN: 0305-1048, DOI: 10.1093/nar/gkaa746.

Sémon, Marie, Laurent Guéguen, Coraline Petit, Carine Rey, Anne Lambert, Manon Peltier, and Sophie Pantalacci (2020), « Comparison of developmental genome expression in rodent molars reveals extensive developmental system drift », *in*: *bioRxiv*, DOI: 10.1101/2020.04.22.043422.

Sims, David, Ian Sudbery, Nicholas E. Ilott, Andreas Heger, and Chris P. Ponting (2014), « Sequencing depth and coverage: key considerations in genomic analyses », *in*: *Nature Reviews. Genetics* 15.*2*, pp. 121–132, ISSN: 1471-0064, DOI: 10.1038/nrg3642.

Slater, Guy St C and Erwan Birney (2005), « Automated generation of heuristics for biological sequence comparison. », *in*: *BMC Bioinformatics* 6, p. 31, ISSN: 1471-2105, DOI: 10.1186/1471-2105-6-31.

Smith, Nick W., Paul R. Shorten, Eric Altermann, Nicole C. Roy, and Warren C. McNabb (2019), « The Classification and Evolution of Bacterial Cross-Feeding », *in*: *Frontiers in Ecology and Evolution* 7, ISSN: 2296-701X, DOI: 10.3389/fevo.2019.00153.

Sonawane, Prashant D., Jacob Pollier, Sayantan Panda, Jedrzej Szymanski, Hassan Massalha, Meital Yona, Tamar Unger, Sergey Malitsky, Philipp Arendt, Laurens Pauwels, Efrat Almekias-Siegl, Ilana Rogachev, Sagit Meir, Pablo D. Cárdenas, Athar Masri, Marina Petrikov, Hubert Schaller, Arthur A. Schaffer, Avinash Kamble, Ashok P. Giri, Alain Goossens, and Asaph Aharoni (Dec. 22, 2016), « Plant cholesterol biosynthetic pathway overlaps with phytosterol metabolism », *in*: *Nature Plants* 3, p. 16205, ISSN: 2055-0278, DOI: 10.1038/nplants.2016.205.

Srinivasan, Shyam, William R. Cluett, and Radhakrishnan Mahadevan (2015), « Constructing kinetic models of metabolism at genome-scales: A review », *in*: *Biotechnology Journal* 10.*9*, pp. 1345–1359, ISSN: 1860-7314, DOI: 10.1002/biot.201400522.

Steinegger, Martin and Johannes Söding (2017), « MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets », *in*: *Nature Biotechnology* 35.*11*, pp. 1026–1028, ISSN: 1546-1696, DOI: 10.1038/nbt.3988.

Stewart, Robert D., Marc D. Auffret, Amanda Warr, Andrew H. Wiser, Maximilian O. Press, Kyle W. Langford, Ivan Liachko, Timothy J. Snelling, Richard J. Dewhurst, Alan W. Walker, Rainer Roehe, and Mick Watson (2018), « Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen », *in*: *Nature Communications* 9.*1*, p. 870, ISSN: 2041-1723, DOI: 10.1038/s41467-018-03317-6.

Strassert, Jürgen F. H., Iker Irisarri, Tom A. Williams, and Fabien Burki (2021), « A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids », *in*: *Nature Communications* 12.*1879*, pp. 1–13, ISSN: 2041-1723, DOI: 10.1038/s41467-021-22044.

Sunagawa, Shinichi, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, Georg Zeller, Daniel R. Mende, Adriana Alberti, Francisco M. Cornejo-Castillo, Paul I. Costea, Corinne Cruaud, Francesco d'Ovidio, Stefan Engelen, Isabel Ferrera, Josep M. Gasol, Lionel Guidi, Falk Hildebrand, Florian Kokoszka, Cyrille Lepoivre, Gipsi Lima-Mendez, Julie Poulain, Bonnie T. Poulos, Marta Royo-Llonch, Hugo Sarmento, Sara Vieira-Silva, Céline Dimier, Marc Picheral, Sarah Searson, Stefanie Kandels-Lewis, Tara Oceans coordinators, Chris Bowler, Colomban de Vargas, Gabriel Gorsky, Nigel Grimsley, Pascal Hingamp, Daniele Iudicone, Olivier Jaillon, Fabrice Not, Hiroyuki Ogata,

Stephane Pesant, Sabrina Speich, Lars Stemmann, Matthew B. Sullivan, Jean Weissenbach, Patrick Wincker, Eric Karsenti, Jeroen Raes, Silvia G. Acinas, and Peer Bork (2015), « Structure and function of the global ocean microbiome », *in*: *Science* 348.*6237*, p. 1261359, DOI: 10.1126/science.1261359.

Supek, Fran, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc (2011), « REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms », *in*: *PLOS ONE* 6.*7*, e21800, ISSN: 1932-6203, DOI: 10.1371/journal.pone.0021800.

Suzuki, Ryota and Hidetoshi Shimodaira (2006), « Pvclust: an R package for assessing the uncertainty in hierarchical clustering », *in*: *Bioinformatics* 22.*12*, pp. 1540–1542.

Tasende, M. G. (2000), « Fatty acid and sterol composition of gametophytes and sporophytes of *Chondrus crispus* (Gigartinaceae, Rhodophyta) », *in*: *Scientia Marina* 64.*4*, Number: 4, pp. 421–426, ISSN: 1886-8134, DOI: 10.3989/scimar.2000.64n4421.

Thauer, Rudolf K., Anne-Kristin Kaster, Henning Seedorf, Wolfgang Buckel, and Reiner Hedderich (2008), « Methanogenic archaea: ecologically relevant differences in energy conservation », *in*: *Nature Reviews. Microbiology* 6.*8*, pp. 579–591, ISSN: 1740-1534, DOI: 10.1038/nrmicro1931.

The UniProt Consortium (2021), « UniProt: the universal protein knowledgebase in 2021 », *in*: *Nucleic Acids Research* 49 (D1), pp. D480–D489, ISSN: 0305-1048, DOI: 10.1093/nar/gkaa1100.

Thiele, Ines, Daniel R. Hyduke, Benjamin Steeb, Guy Fankam, Douglas K. Allen, Susanna Bazzani, Pep Charusanti, Feng-Chi Chen, Ronan M. T. Fleming, Chao A. Hsiung, Sigrid C. J. De Keersmaecker, Yu-Chieh Liao, Kathleen Marchal, Monica L. Mo, Emre Özdemir, Anu Raghunathan, Jennifer L. Reed, Sook-il Shin, Sara Sigurbjörnsdóttir, Jonas Steinmann, Suresh Sudarsan, Neil Swainston, Inge M. Thijs, Karsten Zengler, Bernhard Ø. Palsson, Joshua N. Adkins, and Dirk Bumann (2011), « A community effort towards a knowledge-base and mathematical model of the human pathogen Salmonella Typhimurium LT2 », *in*: *BMC systems biology* 5, p. 8, ISSN: 1752-0509, DOI: 10.1186/1752-0509-5-8.

Thiele, Ines and Bernhard Ø. Palsson (2010a), « A protocol for generating a high-quality genome-scale metabolic reconstruction », *in*: *Nature Protocols* 5.*1*, pp. 93–121, ISSN: 1750-2799, DOI: 10.1038/nprot.2009.203.

— (2010b), « Reconstruction annotation jamborees: a community approach to systems biology », *in*: *Molecular Systems Biology* 6.*1*, p. 361, ISSN: 1744-4292, DOI: 10.1038/msb.2010.15.

Thiele, Ines, Nikos Vlassis, and Ronan M. T. Fleming (2014), « fastGapFill: efficient gap filling in metabolic networks », *in*: *Bioinformatics* 30.*17*, pp. 2529–2531, ISSN: 1367-4811, DOI: 10.1093/bioinformatics/btu321.

True, J. R. and E. S. Haag (2001), « Developmental system drift and flexibility in evolutionary trajectories », *in*: *Evolution & Development* 3.*2*, pp. 109–119, ISSN: 1520-541X, DOI: 10.1046/j.1525-142x.2001.003002109.x.

Vallenet, D., S. Engelen, D. Mornico, S. Cruveiller, L. Fleury, A. Lajus, Z. Rouy, D. Roche, G. Salvignol, C. Scarpelli, and C. Médigue (2009), « MicroScope: a platform for microbial genome annotation and comparative genomics », *in*: *Database* 2009, bap021, ISSN: 1758-0463, DOI: 10.1093/database/bap021.

Varma, A. and B. O. Palsson (1994), « Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110 », *in*: *Applied and Environmental Microbiology* 60.*10*, pp. 3724–3731, ISSN: 0099-2240, DOI: 10.1128/aem.60.10.3724-3731.1994.

Veldsman, Werner P., Giulia Campli, Sagane Dind, Valentine Rech de Laval, Harriet B. Drage, Robert M. Waterhouse, and Marc Robinson-Rechavi (2022), « Taxonbridge: an R package to create custom taxonomies based on the NCBI and GBIF taxonomies », *in*: *bioRxiv*, p. 2022.05.02.490269, DOI: 10.1101/2022.05.02.490269.

Vieira, Gilles, Victor Sabarly, Pierre-Yves Bourguignon, Maxime Durot, François Le Fèvre, Damien Mornico, David Vallenet, Odile Bouvet, Erick Denamur, Vincent Schachter, and Claudine Médigue (2011), « Core and Panmetabolism in Escherichia coli », *in*: *Journal of Bacteriology* 193.*6*, pp. 1461–1472, ISSN: 0021-9193, DOI: 10.1128/JB.01192-10.

Vongsangnak, Wanwipa, Amornpan Klanchui, Iyarest Tawornsamretkit, Witthawin Tatiyaborwornchai, Kobkul Laoteng, and Asawin Meechai (2016), « Genome-scale metabolic modeling of Mucor circinelloides and comparative analysis with other oleaginous species », *in*: *Gene* 583.*2*, pp. 121–129, ISSN: 0378-1119, DOI: 10.1016/j.gene.2016.02.028.

Wagner, A. and D. A. Fell (2001), « The small world inside large metabolic networks. », *in*: *Proceedings of the Royal Society B: Biological Sciences* 268.*1478*, pp. 1803–1810, ISSN: 0962-8452, DOI: 10.1098/rspb.2001.1711.

Wang, Hao, Simonas Marcišauskas, Benjamín J. Sánchez, Iván Domenzain, Daniel Hermansson, Rasmus Agren, Jens Nielsen, and Eduard J. Kerkhoven (2018), « RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor », *in*: *PLOS Computational Biology* 14.*10*, e1006541, ISSN: 1553-7358, DOI: 10.1371/journal.pcbi.1006541.

Wang, Hao, Zhao Xu, Lei Gao, and Bailin Hao (2009), « A fungal phylogeny based on 82 complete genomes using the composition vector method », *in*: *BMC evolutionary biology* 9, p. 195, DOI: 10.1186/1471-2148-9-195.

Wang, Lin, Satyakam Dash, Chiam Yu Ng, and Costas D. Maranas (2017), « A review of computational tools for design and reconstruction of metabolic pathways », *in*: *Synthetic and*

*Systems Biotechnology* 2.*4*, pp. 243–252, ISSN: 2405-805X, DOI: 10.1016/j.synbio.2017.11.002.

Wang, Zhong, Mark Gerstein, and Michael Snyder (2009), « RNA-Seq: a revolutionary tool for transcriptomics », *in*: *Nature reviews. Genetics* 10.*1*, pp. 57–63, ISSN: 1471-0056, DOI: 10.1038/nrg2484.

Wattam, Alice R., James J. Davis, Rida Assaf, Sébastien Boisvert, Thomas Brettin, Christopher Bun, Neal Conrad, Emily M. Dietrich, Terry Disz, Joseph L. Gabbard, Svetlana Gerdes, Christopher S. Henry, Ronald W. Kenyon, Dustin Machi, Chunhong Mao, Eric K. Nordberg, Gary J. Olsen, Daniel E. Murphy-Olson, Robert Olson, Ross Overbeek, Bruce Parrello, Gordon D. Pusch, Maulik Shukla, Veronika Vonstein, Andrew Warren, Fangfang Xia, Hyunseung Yoo, and Rick L. Stevens (2017), « Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center », *in*: *Nucleic Acids Research* 45 (D1), pp. D535–D542, ISSN: 1362-4962, DOI: 10.1093/nar/gkw1017.

Wei, Zhong, Yian Gu, Ville-Petri Friman, George A. Kowalchuk, Yangchun Xu, Qirong Shen, and Alexandre Jousset (2019), « Initial soil microbiome composition and functioning predetermine future plant health », *in*: *Science Advances* 5.*9*, eaaw0759, ISSN: 2375-2548, DOI: 10.1126/sciadv.aaw0759.

Wemheuer, Franziska, Jessica A. Taylor, Rolf Daniel, Emma Johnston, Peter Meinicke, Torsten Thomas, and Bernd Wemheuer (2020), « Tax4Fun2: prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences », *in*: *Environmental Microbiome* 15.*1*, p. 11, ISSN: 2524-6372, DOI: 10.1186/s40793-020-00358-7.

Whitmore, Leanne S., Bernard Nguyen, Ali Pinar, Anthe George, and Corey M. Hudson (2019), « RetSynth: determining all optimal and sub-optimal synthetic pathways that facilitate synthesis of target compounds in chassis organisms », *in*: *BMC Bioinformatics* 20.*1*, p. 461, ISSN: 1471-2105, DOI: 10.1186/s12859-019-3025-9.

Winter, Gal and Jens O. Krömer (2013), « Fluxomics – connecting 'omics analysis and phenotypes », *in*: *Environmental Microbiology* 15.*7*, pp. 1901–1916, ISSN: 1462-2920, DOI: 10.1111/1462-2920.12064.

Witting, M., J. Hastings, N. Rodriguez, C. J. Joshi, J. P. N. Hattwell, P. R. Ebert, M. van Weeghel, A. W. Gao, M. J. O. Wakelam, R. H. Houtkooper, A. Mains, N. Le Novère, S. Sadykoff, F. Schroeder, N. E. Lewis, H. J. Schirra, C. Kaleta, and O. Casanueva (2018), « » *in*: *Front Mol Biosci* 5, [PubMed Central:PMC6246695] [DOI:10.3389/fmolb.2018.00096] [PubMed:28431245], p. 96.

Xavier, Joana C., Rebecca E. Gerhards, Jessica L. E. Wimmer, Julia Brueckner, Fernando D. K. Tria, and William F. Martin (2021), « The metabolic network of the last bacterial

common ancestor », *in*: *Communications Biology* 4.*1*, pp. 1–10, ISSN: 2399-3642, DOI: 10.1038/s42003-021-01918-4.

Xing, Qikun, Guiqi Bi, Min Cao, Arnaud Belcour, Méziane Aite, Zhaolan Mo, and Yunxiang Mao (2021), « Comparative Transcriptome Analysis Provides Insights into Response of Ulva compressa to Fluctuating Salinity Conditions », *in*: *Journal of Phycology* 57.*4*, pp. 1295–1308, ISSN: 1529-8817, DOI: 10.1111/jpy.13167.

Yarza, Pablo, Pelin Yilmaz, Elmar Pruesse, Frank Oliver Glöckner, Wolfgang Ludwig, Karl-Heinz Schleifer, William B. Whitman, Jean Euzéby, Rudolf Amann, and Ramon Rosselló-Móra (2014), « Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences », *in*: *Nature Reviews. Microbiology* 12.*9*, pp. 635–645, ISSN: 1740-1534, DOI: 10.1038/nrmicro3330.

Yeung, Matthew, Ines Thiele, and Bernard Ø Palsson (2007), « Estimation of the number of extreme pathways for metabolic networks », *in*: *BMC Bioinformatics* 8.*1*, p. 363, ISSN: 1471-2105, DOI: 10.1186/1471-2105-8-363.

Zdobnov, E. M. and R. Apweiler (2001), « InterProScan–an integration platform for the signature-recognition methods in InterPro », *in*: *Bioinformatics* 17.*9*, pp. 847–848, ISSN: 1367-4803, DOI: 10.1093/bioinformatics/17.9.847.

Zelezniak, Aleksej, Sergej Andrejev, Olga Ponomarova, Daniel R. Mende, Peer Bork, and Kiran Raosaheb Patil (2015), « Metabolic dependencies drive species co-occurrence in diverse microbial communities », *in*: *Proceedings of the National Academy of Sciences* 112.*20*, pp. 6449–6454, ISSN: 0027-8424, 1091-6490, DOI: 10.1073/pnas.1421834112.

Zimmermann, Johannes, Christoph Kaleta, and Silvio Waschina (2021), « gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models », *in*: *Genome Biology* 22.*1*, p. 81, ISSN: 1474-760X, DOI: 10.1186/s13059-021-02295-1.

Zomorrodi, Ali R. and Costas D. Maranas (2012), « OptCom: A Multi-Level Optimization Framework for the Metabolic Modeling and Analysis of Microbial Communities », *in*: *PLOS Computational Biology* 8.*2*, e1002363, ISSN: 1553-7358, DOI: 10.1371/journal.pcbi.1002363.

Zou, Yuanqiang, Wenbin Xue, Guangwen Luo, Ziqing Deng, Panpan Qin, Ruijin Guo, Haipeng Sun, Yan Xia, Suisha Liang, Ying Dai, Daiwei Wan, Rongrong Jiang, Lili Su, Qiang Feng, Zhuye Jie, Tongkun Guo, Zhongkui Xia, Chuan Liu, Jinghong Yu, Yuxiang Lin, Shanmei Tang, Guicheng Huo, Xun Xu, Yong Hou, Xin Liu, Jian Wang, Huanming Yang, Karsten Kristiansen, Junhua Li, Huijue Jia, and Liang Xiao (2019), « 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses », *in*: *Nature Biotechnology* 37.*2*, pp. 179–185, ISSN: 1546-1696, DOI: 10.1038/s41587-018-0008-8.

# APPENDIX

## Uniprot definitions of proteomes

**Definition 6.2.1 (UniProt definition of reference proteomes)** *Reference proteomes are a subset of proteomes that have been selected either manually or algorithmically according to a number of criteria to provide a broad coverage of the tree of life and a representative cross-section of the taxonomic diversity found within UniProtKB, as well as the proteomes of well-studied model organisms and other species of interest for biomedical research.*

**Definition 6.2.2 (UniProt definition of redundant proteome)** *A redundant proteome is one in which all or nearly all protein sequences are highly similar or identical to an existing proteome from the same species. To reduce redundancy in proteomes and subsequently UniProtKB/TrEMBL, we have developed a procedure to identify highly redundant proteomes within species groups, using a combination of manual and automatic methods. Proteomes can only be redundant to other proteomes of the same taxonomy branch at species level or below (sub-species, strains, etc.). We use the CD-Hit 2D program for pairwise comparison of proteomes within each taxonomic group. Based on the results, we calculate the level of similarity between pairs of proteomes within the groups. Proteomes that rank lowest are the most redundant.*

**Definition 6.2.3 (UniProt definition of excluded proteome)** *UniProt excludes certain proteomes where the assembly has been excluded from the NCBI Reference Sequence (RefSeq) project for any of the reasons listed below. This list is a subset of the exclusion reasons used by RefSeq.*

| *Exclusion reason* | *Explanation* |
|---|---|
| *chimeric* | *Sequences from two different organisms are joined together.* |
| *contaminated* | *Sequences from another organism, cloning vectors, linkers, adapters or primers are present in the assembly.* |
| *hybrid* | *Sequences from a hybrid between different species, strains or isolates.* |
| *misassembled* | *Alignment to related genome assemblies or other evidence indicates the assembly is likely to have errors.* |
| *mixed culture* | *Sequences come from two or more organisms that were not cultured separately.* |
| *sequence duplications* | *Assembly has one or more large duplications.* |
| *unverified source organism* | *The origin of the assembly is misidentified.* |
| *genome length too large* | *Total non-gapped sequence length of the assembly is more than 1.5 times that of the average for the genomes in the Assembly resource from the same species more than 15 Mbp, or is otherwise suspiciously long.* |
| *genome length too small* | *Total non-gapped sequence length of the assembly is less than half that of the average for the genomes in the Assembly resource from the same species less than 300 Kbp, or is otherwise suspiciously short.* |

# Application of EsMeCaTa on bacteria and eucaryotes

| **Input** | | **Taxa selected by EsMeCaTa** | | **Proteomes selection (Busco ≥ 0.8)** | | | **Protein clusters (MMseqs2)** | | | **Functional annotation of clusters** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Lowest taxon name | Taxon rank | Taxon rank used | Taxon name used | UniProt total | UniProt references | EsMeCaTa proteomes | Pan-P | Soft-P | Shell-P | Pan-P | | Soft-P | | Shell-P | |
| | | | | | | | | | | GO | EC | GO | EC | GO | EC |
| *Escherichia* | Genus | Genus | *Escherichia* | 1,506 | 3 | 3 | 5,821 | 2,421 | 3,298 | 2,183 | 866 | 1,661 | 679 | 1,906 | 792 |
| *Citrobacter* | Genus | Genus | *Citrobacter* | 138 | 2 | 2 | 5,674 | 2,753 | 5,674 | 2,013 | 772 | 1,835 | 708 | 2,013 | 772 |
| *Cronobacter* | Genus | Genus | *Cronobacter* | 15 | 0 | 15 | 9,057 | 101 | 3,128 | 970 | 677 | 0 | 12 | 600 | 603 |
| *Lelliottia* | Genus | Genus | *Lelliottia* | 5 | 0 | 5 | 5,252 | 2,651 | 3,245 | 1,993 | 756 | 1,784 | 687 | 1,884 | 718 |
| *Jejubacter* | Genus | Genus | *Jejubacter* | 1 | 1 | 1 | 3,915 | 3,915 | 3,915 | 1,983 | 837 | 1,983 | 837 | 1,983 | 837 |
| *Edaphovirga* | Genus | **Family** | Enterobacteriaceae | 2,435 | 42 | 42 | 25,822 | 415 | 2,581 | 2,253 | 867 | 514 | 193 | 1,560 | 595 |
| Enterobacteriaceae | Family | Family | Enterobacteriaceae | 2,435 | 42 | 42 | 25,822 | 415 | 2,581 | 2,253 | 867 | 514 | 193 | 1,560 | 595 |
| Enterobacterales | Order | Order | Enterobacterales | 3,028 | 129 | 96 | 53,617 | 375 | 2,145 | 2,475 | 1,010 | 487 | 175 | 1,383 | 512 |
| Gammaproteobacteria | Class | Class | Gammaproteobacteria | 8,271 | 911 | 96 | 85,797 | 329 | 1,183 | 2,650 | 1,040 | 387 | 123 | 924 | 327 |
| *Plasmodium* | Genus | Genus | *Plasmodium* | 67 | 17 | 17 | 21,287 | 1,276 | 4,263 | 1,305 | 225 | 611 | 104 | 1,103 | 200 |
| *Leucocytozoon* | Genus | **Order** | Haemosporida | 68 | 18 | 18 | 22,813 | 1,076 | 4,313 | 1,327 | 259 | 546 | 95 | 1,090 | 199 |
| *Corallicola* | Genus | **Class** | Conoidasida | 30 | 10 | 10 | 46,959 | 76 | 1,326 | 1,919 | 530 | 94 | 14 | 717 | 121 |
| *Acavomonas* | Genus | **Clade** | Alveolata | 124 | 48 | 48 | 248,878 | 50 | 785 | 3,746 | 924 | 42 | 7 | 418 | 76 |

Table 6.1 – **Result of EsMeCaTa on 13 taxonomic affiliations from Table 4.1** (described by names and ranks). First EsMeCaTa proteomes identifies the **lowest taxonomic rank** associated with **proteomes** (column 'Taxa selected by EsMeCaTa'), then it selects the proteomes associated with the taxon (column 'Proteomes selection (Busco ≥ 0.8)'). The sub-column 'UniProt total' indicates the total number of proteomes associated with the taxon in UniProt (with BUSCO score ≥ 0.8). The sub-column 'Uniprot reference' shows the number of reference proteomes for the taxon. And the sub-column 'EsMeCaTa proteomes' presents the number of proteomes selected by EsMeCaTa (which will vary according to the presence of reference proteome or not and if the subsampling procedure ahs been applied). In a second step, EsMeCaTa clusters the protein of the proteomes into **protein clusters** using **MMseqs2** (column 'Protein clusters (MMseqs2)'). The sub-column indicates the number of protein clusters selected according to various filter. These filters corresponds to the **clustering threshold**, this means that for each cluster, EsMeCaTa will look for how many different proteomes are represented in the proteins of the cluster. It will divide this number by the total number of proteomes to find a ratio of representation of proteome for each cluster. Three filters are presented: Soft-P(Soft core proteome), which means that the filtered cluster contains at least proteins coming from 95% of the proteomes used for the clustering. Shell-P(Shell core ) meaning that 50% proteomes are represented in each cluster. The last filter Pan-P(Pan-proteome) indicates that all protein cluster found by MMseqs2 are considered. The last step presented in the table with the column 'Functional annotation of clusters' shows how many **GO Terms** and **EC numbers** are found for each clustering threshold.

DOCTORAT BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

**Titre :** Combiner approches basées sur la connaissance et sur des comparaisons de séquences pour élucider les fonctions métaboliques, des voies aux communautés

**Mot clés :** Bioinformatique, Représentation des connaissances, Génomique comparative, Métabolisme, Biologie des systèmes, Microbiote, Évolution

**Résumé :** Le métabolisme peut être modélisé et étudié à plusieurs niveaux. Un premier niveau étudié est celui des voies métaboliques qui correspondent à des enchaînements de transformations chimiques amenant à la production de composés d'intérêt. Et c'est au travers d'une formalisation de la dérive métabolique en programmation par contraintes, que des voies métaboliques alternatives ont pu être proposées chez une algue. Un second niveau du métabolisme rassemble l'ensemble des centaines de voies métaboliques contenu dans le métabolisme d'un organisme. Une méthode visant à créer des réseaux métaboliques homogènes à partir de données publiques hétérogènes est présentée et est appliquée sur trois jeux de données bactériens et eucaryotes. Le troisième niveau est le métabolisme d'un groupe d'organismes et permet d'étudier le fonctionnement d'un organisme non spécifiquement identifié. Pour cela, une méthode reposant sur l'ingénierie des connaissances et la comparaison des séquences a été développée et a permis d'étudier le métabolisme d'une communauté bactérienne. Le dernier niveau correspond au métabolisme d'une communauté et vise à comprendre les possibles interactions métaboliques entre ces organismes. Une méthode a été développée permettant l'identification d'espèces clés au travers de la complémentarité métabolique.

**Title:** Combining knowledge-based and sequence comparison approaches to elucidate metabolic functions, from pathways to communities

**Keywords:** Bioinformatics, Knowledge representation, Comparative genomics, Metabolism, Systems biology, Microbiota, Evolution

**Abstract:** Metabolism can be modelled and studied at many levels. The first level is the metabolic pathways, which contain a set of chemical transformations leading to the production of compounds of interest. Alternative metabolic pathways were predicted in an alga using a formalism of the metabolic pathway drift and its implementation with constraint programming. The second level is the organism metabolism which contains hundreds of metabolic pathways. A method has been developed to reconstruct homogeneous metabolic networks from heterogeneous public data. The third level is the metabolism of a group of organisms (or taxon) which can be useful to characterize an organism that has not been clearly identified. To achieve this, a method using knowledge engineering and sequence comparison has been created. Finally, the fourth level is the metabolism of a community and the metabolic interaction in this community. A method has been developed to identify the key species among a community.