HAL
open science

Exploration de l'évolution du réseau métabolique chez les champignons

Vahiniaina Herinjiva Andriamanga

To cite this version:

Vahiniaina Herinjiva Andriamanga. Exploration de l'évolution du réseau métabolique chez les champignons. Réseaux moléculaires [q-bio.MN]. Université Paris-Saclay, 2023. Français. NNT : 2023UPASL150 . tel-04449819

HAL Id: tel-04449819

https://theses.hal.science/tel-04449819

Submitted on 9 Feb 2024

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

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

université PARIS-SACLAY

Exploration de l'évolution du réseau métabolique chez les champignons

Exploring the evolution of metabolic networks in fungi

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 577, Structure et dynamique des systèmes du vivant (SDSV)
Spécialité de doctorat : Evolution
Graduate School : Sciences de la vie et santé
Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **I2BC** (Université Paris-Saclay, CEA, CNRS), sous la direction de **Olivier LESPINET**, professeur.

Thèse soutenue à Paris-Saclay, le 18 décembre 2023, par

Vahiniaina Herinjiva ANDRIAMANGA

Composition du Jury

Membres du jury avec voix délibérative

Christine DILLMANN Professeure, Université Paris-Saclay	Présidente
Fabien JOURDAN Directeur de recherche, INRAE, Université de Toulouse	Rapporteur & Examinateur
Odile LECOMPTE Professeure, Université de Strasbourg	Rapporteure & Examinatrice
Marc-Henri LEBRUN Directeur de recherche, CNRS, Université Paris-Saclay	Examinateur
Philippe SILAR Professeur, Université Paris Cité	Examinateur

THESE DE DOCTORAT

NNT : 2023UPASL150

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

université PARIS-SACLAY

Title : Exploring the evolution of metabolic networks in fungi

Keywords : fungi, metabolic network, comparative genomics, graph theory

Metabolism is the set of biochemical reactions that occur within an organism. The sequence of these reactions forms metabolic pathways, and their interconnections constitute the organism's complete metabolic network. Reactions within the metabolic network are mostly catalyzed by enzymes, which are classified based on the specific reaction they catalyze. This metabolic network determines the organism's metabolic capacities, including its ability to use chemical compounds found in the environment and to synthesize new products. These networks evolve, giving rise to new pathways capable of producing new products or utilizing new substrates. To unravel the evolution dynamics of the metabolic network, we investigated the evolution of 910 enzyme activities (identified by EC-number) in 174 fungal species. Fungi serve as ideal models for this study due to their diverse metabolic profiles. They catabolize a wide variety of substrates, including lignin and cellulose, which are the most abundant biopolymers on earth. Moreover, fungi synthesize a variety of molecules, such as antibiotics and toxins. They also have successfully colonized every corner of the earth.

The analysis of the conservation of the 910 enzyme activities across the studied species exhibited 454 enzyme activities being universally present in all the species, while the remaining 456 were associated with particular clades or specific species. By grouping enzyme activities according to their phylogenetic profiles' similarity, we can identify sets of enzyme activities specific to particular clades or species. Through a phylostratigraphy approach, we reconstructed the evolutionary history, encompassing both the losses and acquisitions of enzyme activities related to specific clades or species. Our study revealed that 860 of these enzyme activities were already present in fungal ancestors but half of them were subsequently lost during evolution, while 8 newly emerged as fungal-specific enzyme activities. The evolutionary origin of the remaining 42 enzyme activities could not be determined.

We subsequently mapped the evolutionary information onto the metabolic network. The core of this metabolic network is mainly composed of primary metabolic pathways, while secondary metabolic pathways predominantly occupy the periphery. Using graph theory tools , we assess the localization of enzyme activities within the metabolic network. Lineage-specific enzyme activities tend to occupy the network's periphery, demonstrating lower connectivity compared to common enzyme activities. Often, these lineage-specific activities serve as alternatives to the common ones. Furthermore, when we group enzyme activities based on the similarity of their phylogenetic profiles, we observe that those with similar profiles tend to cluster together within the network. Our observations suggest that the loss of network-disrupting enzyme activity is tolerated for two reasons: either the affected portion of the subnetwork becomes dispensable, considering it as an accessory, or there exists an alternative enzyme activity to bridge the two sections.

This study underscore the significant role of enzyme loss in driving fungal metabolic network evolution, revealing a noteworthy constraint on the emergence of specific enzyme activities. Enzyme activity losses played a pivotal role in delineating specific taxonomic groups during the course of evolution. Importantly, the metabolic network constrains the evolution of enzyme activities, with certain network positions being prone to losses.

Titre : Exploration de l'évolution du réseau métabolique chez les champignons

Mots clés : champignon, réseau métabolique, génomique comparée, théorie des graphes.

Le métabolisme est l'ensemble des réactions biochimiques qui ont lieu dans un organisme. La séquence de ces réactions forme les voies métaboliques et leurs interconnexions constituent le réseau métabolique d'un organisme. Les réactions dans le réseau métabolique sont principalement catalysées par des enzymes qui sont classées en fonction de la nature de la réaction qu'elles catalysent. Ce réseau métabolique définit les capacités métaboliques de l'organisme, en particulier sa capacité à utiliser les composés présents dans son milieu et sa capacité à synthétiser de nouveaux produits. Les évolutions de ce réseau conduisent à l'émergence de nouvelles voies capables de développer de nouveaux produits ou d'utiliser de nouveaux substrats. Pour explorer l'évolution du réseau métabolique, nous avons analysé les activités enzymatiques de 174 espèces de champignons. En effet, les champignons sont d'excellents modèles, car ils présentent une grande diversité de profils métaboliques. Ils sont capables de cataboliser une grande variété de substrats, tels que la lignine et la cellulose, et de synthétiser une grande variété de molécules, telles que des antibiotiques et des toxines. Ils ont également colonisé une grande variété de niches écologiques auxquelles leur métabolisme s'est adapté.

L'analyse de la conservation de 910 activités enzymatiques (identifiées par leur EC-number) chez les espèces que nous avons étudiées a montré que 454 activités enzymatiques sont présentes chez toutes les espèces, tandis que les 456 autres sont associées à des clades particuliers ou à des espèces spécifiques. En utilisant une approche phylostratigraphique, nous avons reconstruit l'histoire évolutive, englobant à la fois les pertes et les acquisitions, de ces activités enzymatiques. Notre étude a révélé que 406 de ces activités enzymatiques étaient déjà présentes chez les ancêtres des champignons, mais perdues par certaines espèces au cours de l'évolution, tandis que 8 sont des nouveautés spécifiques des champignons. L'origine évolutive des 42 activités enzymatiques restant n'a pas pu être déterminée. De plus, la classification des activités enzymatiques par similarité de profil phylogénétique nous indique que certains groupes d'activités enzymatiques ont co évolué au cours de l'évolution.

Ces informations évolutives ont été cartographiées sur le réseau métabolique global. À l'aide d'outils issus de la théorie des graphes, nous avons alors analysé la position de ces activités enzymatique dans le réseau. Le centre du réseau métabolique est principalement composé de voies métaboliques primaires alors que, et les voies métaboliques secondaires sont principalement en périphérie. Nous avons également montré que les activités enzymatiques spécifiques de certaines espèces ont tendance à occuper la périphérie du réseau, et présentent une connectivité plus faible au sein du réseau métabolique. Elles servent souvent d'alternatives aux activités communes. Lorsque nous regroupons les activités enzymatiques en fonction de leur similarité de profils, nous observons que ceux ayant des profils similaires ont tendance à se regrouper au sein du réseau. Nos observations indiquent également que la perte d'activités enzymatiques qui connectent deux parties du réseau est tolérée pour deux raisons : soit une partie affectée du sous-réseau est considérée comme dispensable, ou une activité enzymatique alternative existe pour lier les deux parties.

Les résultats de cette étude indiquent que beaucoup de pertes enzymatiques ont principalement affecté l'évolution du réseau métabolique des champignons, ces pertes ont joué un rôle dans la divergence de certains groupes taxonomiques chez les champignons. Nous montrons qu'il semble également y avoir une contrainte sur l'émergence d'activités enzymatiques spécifiques. Enfin, le réseau métabolique exerce une pression sur la conservation des activités enzymatiques où certaines positions dans le réseau sont plus sujettes aux pertes.

Remerciements

Pour commencer ce manuscrit, j'aimerais d'abord remercier Olivier Lespinet et Anne Lopes, sans qui ce projet n'aurait pas été possible. Je tiens à les remercier du fond du cœur de m'avoir accueilli pour mener ce projet de thèse qui a été très passionnant et enrichissant. Olivier a toujours été à l'écoute et m'a prodigué de précieux conseils en dépit de ses nombreuses responsabilités. Au cours des 4 dernières années, puisque j'ai effectué mon stage de M2 dans son équipe avant de continuer en thèse, il a toujours fait son maximum pour m'encadrer, faire avancer mes travaux, et pour que tout se passe dans les meilleures conditions possibles. Les encouragements, même si ce n'était pas toujours évident, et les critiques qu'il a apportés m'ont beaucoup aidé à grandir scientifiquement.
Je tiens aussi à remercier Anne, ses conseils et ses encouragements ont été des plus précieux. Les discussions tard le soir et les échanges de 5mn autour d'un café (du thé pour moi) ou dans le couloir ont été de véritables moteurs pour ce projet. D'un point de vue personnel et scientifique, elle m'a toujours encouragé (forcé) à donner le meilleur de moi-même. C'est une personne qui m'a vraiment fait apprécier la recherche et a rendu ces 3 années tellement plus enrichissantes sur tous les points. Je n'oublie pas non plus les petits gâteaux qu'elle ramenait, essentiels pour pouvoir réfléchir efficacement.

Je voudrais aussi remercier mon comité de thèse composé de Arnaud Le-Rouzic, David Moreira et Anne Lopes pour le temps qu'ils ont consacré pour suivre l'avancement de ce projet et l'attention qu'ils ont apportée. Leurs conseils et les discussions, même si ce n'était pas fréquent, ont été particulièrement importants pour faire avancer ce projet dans la bonne direction. Je n'oublierai pas non plus les encouragements qu'ils m'ont donnés et qui m'ont permis d'accomplir ce qu'ils attendaient de moi après ces 3 années de thèse.

Je tiens aussi à remercier tous les membres du jury d'avoir accepté notre invitation et participé à l'évaluation de ces travaux, mes rapporteurs, Odile Lecompte et Fabien Jourdan, ainsi que mes examinateurs Christine Dillmann, Philippe Silar et Jean Marc-Lebrun.

Un grand merci aussi aux personnes que j'ai côtoyées au laboratoire et à l'institut, qu'ils aient été de passage ou permanents. Une petite dédicace à mon voisin de bureau et binôme de thèse, Paul Roginsky, qui m'a beaucoup aidé sur plusieurs aspects durant ces travaux et qui m'a aussi beaucoup fait souffrir (durant les séances de sports). Son aide et les discussions au quotidien, scientifiques ou non, sur tout et rien, ont été des plus précieux.
À Christine Drevet, qui m'a accueilli à bras ouverts lors de mon arrivée et qui a mis à ma disposition toutes les données nécessaires pour démarrer ce projet.
Je remercie aussi ceux qui ont travaillé en amont de ce projet car sans eux ces travaux n'auraient pas été possibles.
Un grand merci aussi aux autres membres des équipes BIM et SSFA et les autres personnes

que j'ai rencontrées, j'espère n'avoir oublié personne (vous pouvez rajouter votre nom à la fin), à Mélina, Jean-Christophe, Léonor, Gilles, Christine, Fabrice, Claire, Daniel, [.....] et aux doctorants Christos, Ambre, Simon, Taher [....]. Merci à vous tous pour les discussions que ce soit plus formel pendant les réunions d'équipe, ou pendant les pauses déjeuner et les TGIF. Ça a été un plaisir de vous avoir côtoyés et d'avoir partagé ces 3 années avec vous. Vous avez tous un peu contribué dans l'accomplissement de cette thèse. Et comme Anne disait : « un peu de beaucoup, ça commence à bien faire pas mal ».

Pour terminer, un grand merci à ma famille et mes amis, qu'ils soient en France ou à Madagascar de m'avoir toujours soutenu, et plus particulièrement à ma sœur qui a vécu pleinement ces 3 années de thèses avec moi, même si jusqu'à présent elle ne comprend pas en quoi mon sujet de thèse consiste. Elle a toujours fait en sorte que j'aie à manger quand je rentrais tard le soir après de longues journées au labo.

Ces 3 dernières années ont été faites de haut et de bas, mais sans toutes ces personnes, il y aurait eu plus de bas que de haut. « Misaotra betsaka » pour toutes vos contributions.

Table des matières

I. Introduction 4

1 Le métabolisme 5
1.1 Définition du métabolisme 6
1.2 Les enzymes et leur classification 8
1.3 Les voies et les réseaux métaboliques 12
1.3.1 Métabolisme primaire 13
1.3.2 Métabolisme secondaire 13
1.3.3 Représentation du réseau métabolique 17
1.3.4 Base de données des voies métaboliques 18

2 Evolution du métabolisme : état de l'art 28
2.1 Origine des activités enzymatiques d'un point de vue génomique 29
2.1.1 Emergence d'une nouvelle activité enzymatique 29
2.1.2 Perte d'une activité enzymatique 31
2.2 Mécanisme à l'origine du métabolisme 32
2.3 Approche par génomique comparée 34
2.4 Analyse sous forme de graphes 37
2.4.1 Quelques définitions liées à la théorie des graphes 38
2.4.2 Caractéristiques des réseaux biologiques 39

3 Les champignons 45
3.1 Taxonomies et diversités 47
3.2 Projet de séquençage des champignons 53

4 Objectifs de la thèse 55

II. Conservation et évolutions des activités enzymatiques chez les champignons 59

5 Profils phylogénétiques des activités enzymatiques 59

6 Construction de l'arbre phylogénétique des espèces étudiées 65
6.1 Choix du marqueur phylogénétique 67
6.2 Détection du core protéome 68
6.3 Inférence de l'arbre phylogénétique à partir du core protéome 70
6.4 Constructions de l'arbre phylogénétique sur les espèces étudiées 72

7 Répartition des activités enzymatiques 77
7.1 Activités enzymatiques associées au métabolisme 78
7.2 Conservation des activités enzymatiques chez les champignons. 80
7.3 Répartition des activités enzymatiques par classe taxonomique 81
7.4 Répartition des activités enzymatiques par classe enzymatique 86
7.5 Détection des activités enzymatiques co-évoluantes 86
7.5.1 Différentes méthodes de classification 87
7.5.2 Une méthode de classification : Cluster AGgregation (CLAG) 89
7.5.3 Paramétrage de CLAG 90
7.5.4 Choix des paramètres 93
7.5.5 Description des résultats de la classification et visualisation des groupes obtenus 103

8 Origines évolutives des activités enzymatiques 112
8.1 La phylostratigraphie 113
8.2 Méthodes pour la phylostratigraphie 115
8.2.1 Détection des homologues en dehors des champignons 115
8.2.2 Annotation des homologues en EC-number 117
8.3 Résultat de la phylostratigraphie 119
8.4 Limite de l'approche phylostratigraphique 120

III. Construction du réseau métabolique 123

9 Comparaison entre KEGG pathways et MetaCyc d'un point de vue computationnel 130

10 Identification des voies présentes chez les champignons _____ 137
10.1 Filtre topologique _____ 138
10.2 Classification topologique des voies _____ 142

11 Construction du réseau métabolique _____ 152
11.1 Connexion des voies de KEGG _____ 153
11.2 Caractéristique du réseau métabolique global _____ 155

IV. Exploration de l'évolution du réseau métabolique _____ 159

12 Localisation des activités enzymatiques dans le réseau en fonction de leur histoire évolutive _ 159
12.1 En fonction de la conservation _____ 160
12.2 En fonction de la similarité de profil _____ 163

13 Caractéristiques évolutives des voies métaboliques _____ 167
13.1 Les voies métaboliques essentielles et accessoires _____ 168
13.2 Différence d'un point de vue topologique entre voies communes et accessoires _____ 173
13.2.1 Position des voies dans le réseau en fonction de la conservation _____ 173
13.2.2 Localisation des voies dans le réseau en fonction des super groupes _____ 176
13.2.3 Analyse des liens entre les voie métaboliques _____ 178

14 Rôles des activités spécifiques dans les voies. _____ 186
14.1 Définitions des différents rôles possibles _____ 187
14.2 Résultats de l'étude des rôles des sommets spécifiques _____ 207

V. Discussion _____ 209

VI. Projet d'article _____ 220

VII. Annexes _____ 268

VIII. Bibliographie _____ 275

I. Introduction

Chapitre :

1 Le métabolisme

1.1 Définition du métabolisme

Le mot métabolisme est apparu pour la première fois dans la 11ème édition du dictionnaire médical du physiologiste et pédiatre Pierre-Hubert Nysten en 1858 pour désigner le « changement de nature moléculaire des corps ».

Le mot métabolisme vient du grec *métabolé* qui signifie « changement » . Il est issu du verbe *métabollô* qui veut dire « se déplacer, changer, se transformer ». Les deux mots sont constitués du radical *bol* , du préfixe *méta-* qui signifie « au milieu de, à la suite de, avec » et exprime aussi « un changement de place » et du suffixe *-isme* qui se rattache à une catégorie de verbes grecs désignant un processus, une action en cours de réalisation ou se répétant.

Selon le Centre National de Ressources Textuelles et Lexicales (CNRTL) le métabolisme se définit comme : « ensemble des réactions de synthèse, génératrices de matériaux (anabolisme), et de dégradation, génératrices d'énergie (catabolisme), qui s'effectuent au sein de la matière vivante à partir des constituants chimiques fournis à l'organisme par l'alimentation et sous l'action de catalyseurs spécifiques ».
Le métabolisme est donc l'ensemble des processus chimiques d'un être vivant qui lui permettent de se maintenir en vie, de se développer et de se reproduire. Le philosophe allemand Hans Jonas considère même le métabolisme comme un quasi-synonyme de la vie qui détermine le vivant. L'absence ou l'arrêt du métabolisme revient à mourir. « Son pouvoir est un devoir, puisque son exécution est identique à son être. Elle peut, mais elle ne peut cesser de faire ce qu'elle peut sans cesser d'être » (*Le Phénomène de la vie*, 1966).

Comme nous l'avons vu dans les définitions présentées ci-dessus, il y a deux types de processus métaboliques : l'**anabolisme** et le **catabolisme**.
Le catabolisme est l'ensemble de processus métaboliques qui impliquent la dégradation de molécules plus complexes. C'est l'opposé de l'anabolisme, qui se réfère à la synthèse de molécules plus complexes.

L'une des principales fonctions des processus cataboliques est de fournir l'énergie nécessaire aux différentes activités cellulaires. Par exemple, la dégradation du glucose lors de la respiration cellulaire est un processus catabolique qui produit de l'ATP (adénosine triphosphate), la principale molécule énergétique des cellules (Figure 1.1).

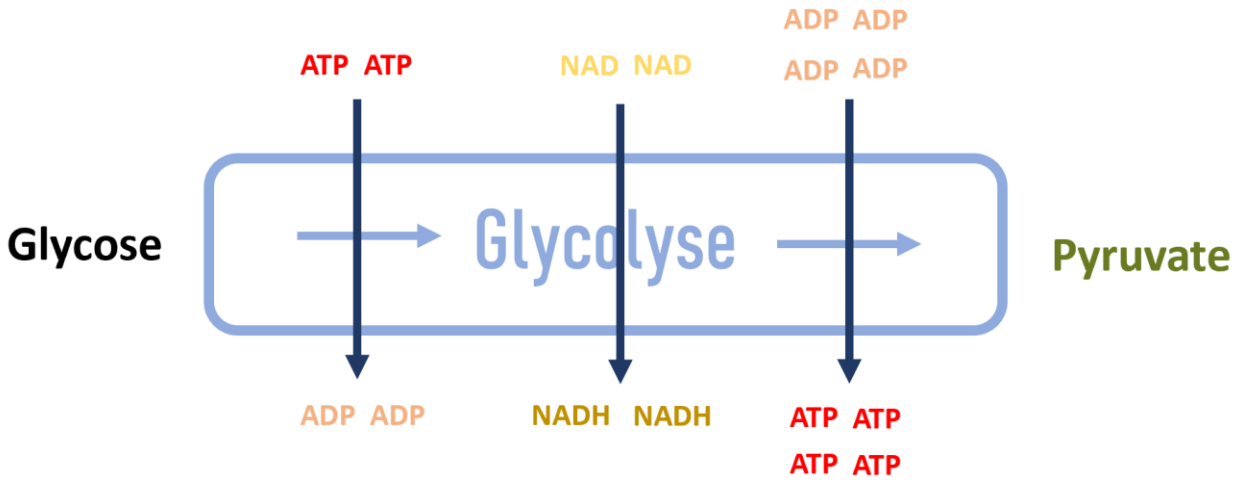


Figure 1.1 : Bilan énergétique de la glycolyse qui est une voie catabolique. La phase préparatoire consomme deux molécules d'ATP. La phase finale permet de produire 4 molécules d'ATP. Le bilan de la glycolyse est un gain de 2 molécules d'ATP.

Pendant le catabolisme, des molécules complexes, qui peuvent provenir d'un apport exogène (nutrition ou environnement) ou d'une synthèse endogène (anabolisme), telles que les glucides, les protéines et les graisses sont décomposées en composés plus simples. Ces composés plus petits peuvent ensuite être traités et utilisés par les cellules pour produire de l'énergie, construire et réparer les structures cellulaires ou servir de précurseurs aux réactions anaboliques.

L'anabolisme est l'ensemble des processus biologiques qui impliquent la synthèse des molécules plus complexes à partir de molécules simples.
Les réactions anaboliques nécessitent un apport d'énergie pour construire ces molécules complexes, et elles impliquent souvent la formation de nouvelles liaisons chimiques entre de plus petits éléments constitutifs.
Un exemple est la biosynthèse de l'ergostérol, qui est un composant essentiel de la membrane cellulaire des champignons et qui joue de ce fait un rôle crucial dans le maintien de l'intégrité et de la fluidité de la membrane cellulaire (Douglas and Konopka, 2014). La voie de synthèse de l'ergostérol illustre la façon dont les champignons utilisent des molécules simples comme l'acétyl-CoA pour synthétiser des stérols complexes comme l'ergostérol (Figure 1.2).

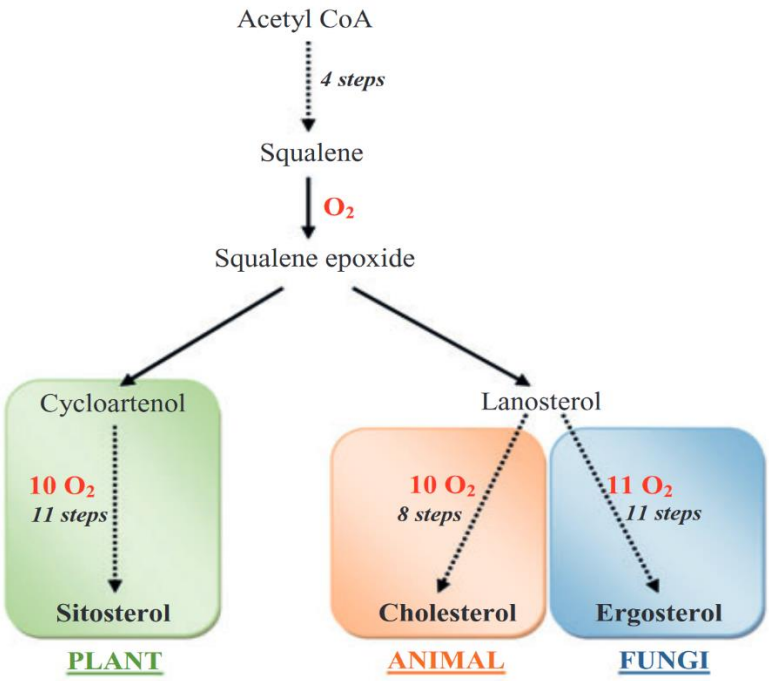


Figure 1.2 : Voie de synthèse généralisée et simplifiée des stérols, une voie anabolique, chez les plantes, les animaux et les champignons. Les trois règnes ont une voie commune à partir de l'Acetyl CoA jusqu'à l'époxyde squalène. Figure tirée de (Dupont *et al.*, 2012).

Ensemble, anabolisme et catabolisme permettent aux organismes de maintenir un équilibre dynamique, en décomposant et en construisant continuellement des molécules pour soutenir la vie et s'adapter aux évolutions constantes de leur environnement.

Chaque organisme vivant utilise son environnement pour survivre en y puisant les nutriments et les substances qui vont lui servir d'éléments de base pour sa croissance, son développement et sa reproduction. Ces nutriments et substances seront transformés en d'autres molécules à la suite de plusieurs réactions biochimiques. Les réactions chimiques anaboliques ou cataboliques peuvent être spontanées ou catalysées par une **enzyme**.

1.2 Les enzymes et leur classification

Le mot « enzyme » a été employé pour la première fois par le physiologiste allemand Wilhelm Kühne en 1878 quand il a décrit la capacité de la levure à transformer le sucre en alcool. Le mot enzyme vient de deux mots grecs *en* qui signifie « dans » ou « à l'intérieur de » et de *zumê* qui signifie « levain ».

À la fin du XIXe siècle et au début du XXe siècle, d'importants progrès ont été réalisés dans l'extraction, la caractérisation et l'exploitation commerciale de nombreux enzymes. Mais ce n'est que dans les années 1920 que les enzymes ont été cristallisées, révélant que l'activité

catalytique est associée aux **protéines**.
Pendant les années suivantes, seules des protéines ont été admises comme ayant une activité catalytique, mais dans les années 1980, il a été démontré que des molécules d'acide ribonucléique (ARN) sont également capables d'exercer une activité catalytique (Cech *et al.*, 1981). Ces ARNs, appelés ribozymes, jouent un rôle important dans l'expression des gènes. Par exemple, le ribosome est un complexe composé d'ARN et de protéines dont le site actif qui catalyse la synthèse de liaison peptidique entre les acides aminés pendant la traduction des ARNm messager en protéines est composé d'ARN (Brimacombe and Stiege, 1985). La ribonucléase P est responsable de la maturation des ARN de transfert (ARNt) (Frank and Pace, 1998). Les Riboswitches sont des éléments régulateurs dans les ARNm qui par liaison avec un ligand vont altérer la structure secondaire de l'ARNm et ce qui peut affecter l'expression du gène (Mironov *et al.*, 2002).
La découverte des ribozymes a révolutionné la recherche sur les origines de la vie en introduisant l'idée que l'ARN pouvait remplir à la fois un rôle catalytique et servir de support à l'information génétique. Cette avancée a ouvert la possibilité d'envisager un monde prébiotique où l'ARN aurait pu être à l'origine de toutes les fonctions biologiques, donnant ainsi naissance à l'hypothèse du « RNA world » (Gilbert, 1986).

Une enzyme, une protéine ou une molécule d'ARN, est un catalyseur biologique qui accélère les réactions biochimiques dans les organismes vivants à un rythme compatible avec la vie.

Les enzymes fonctionnent en réduisant l'énergie d'activation nécessaire pour qu'une réaction chimique se produise. L'énergie d'activation étant la barrière énergétique qui doit être surmontée pour qu'une réaction ait lieu. En abaissant cette barrière, les enzymes facilitent la conversion des réactifs (substrats) en produits, rendant la réaction plus efficace et plus rapide.

En tant que catalyseurs puissants, les enzymes accélèrent les réactions biochimiques sans qu'elles ne soient consommées pendant la réaction. Communément, une enzyme est capable de catalyser la conversion d'une molécule (le substrat) en une autre molécule (le produit) comme suit :

$$\text{Substrat} \xrightarrow{\text{Enzyme}} \text{Produit}$$

Figure 1.3 : une réaction enzymatique. L'enzyme catalyse la transformation du substrat en produit. Dans le cas d'une enzyme capable de catalyser la réaction dans les deux sens (par exemple l'aldolase (Pirovich *et al.*, 2021)) , le substrat peut être le produit et le produit un substrat.

Les enzymes, en tant que remarquables catalyseurs, se caractérisent aussi par une très forte spécificité. Les enzymes catalysent en général la conversion d'un seul type de substrat ou une gamme de substrats similaire en un produit. C'est le modèle « hand-in-glove » (Daniel Koshland 1958). Par exemple, la phosphatase alcaline peut éliminer un groupe de phosphate d'une large variété de substrat (Millán, 2006). C'est une enzyme avec une spécificité pour un groupe de molécules.
D'autres enzymes au contraire, ont une spécificité absolue pour une molécule de substrat (Figure 1.4). C'est le modèle « clé-serrure » (Emil Fisher 1894). Par exemple, le glucose oxydase présente une spécificité presque totale pour son substrat, le beta-D-glucose, et aucune activité avec les autres monosaccharides (Wilson and Turner, 1992).

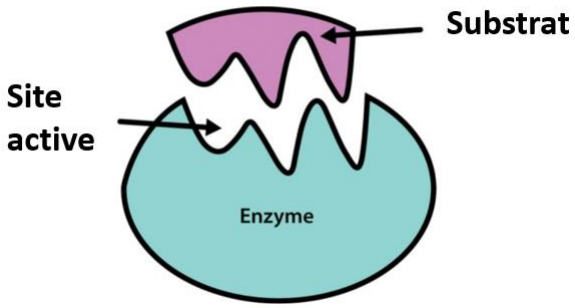


Figure 1.4 : Représentation d'une enzyme avec son site actif et son substrat spécifique.

Dans les exemples précédents, les enzymes ont généralement des noms communs qui se réfèrent le plus souvent à la réaction qu'ils catalysent en ajoutant le suffixe -ase (oxydase, déshydrogénase, carboxylase...). Les enzymes protéolytiques, enzymes qui cassent les liaisons peptidiques des protéines, ont pour suffixe -in (trypsine, chymotrypsine…).
Souvent le nom trivial indique le substrat sur lequel l'enzyme agit (glucose oxydase, alcool déshydrogénase, pyruvate décarboxylase…). Pourtant, quelques noms triviaux ne donnent aucune information sur le substrat, le produit ou la réaction impliqué (invertase, aldolase…).

En raison de la complexité croissante et de l'incohérence dans la dénomination des enzymes, le Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB) (anciennement Nomenclature Committee of the International Union of Biochemistry) a mis en place l'**Enzyme Commission** pour résoudre ce problème (Enzyme Nomenclature. Recommendations 1992, 1994).

Dans ce système, toutes les enzymes sont décrites par un code à 4 nombres séparé par des points qui est appelé **Enzyme Commission (EC) number** (Figure 1.5). Le premier nombre de l'EC-number définit la classe de l'activité enzymatique et fait référence à la réaction catalysée par l'enzyme (table 1.1). Les nombres restants définissent les sous-classes et ont des significations différentes selon la nature de la réaction identifiée par le premier chiffre. Le dernier nombre fait référence au substrat de la réaction.

Par exemple, dans la catégorie des oxydoréductases, le second nombre indique le donneur d'hydrogène, le troisième nombre indique l'accepteur d'hydrogène et le quatrième nombre indique le substrat (Figure 1.5).

1 : Oxidoreductases
1.2 : Acting on the aldehyde or oxo group of donors
1.2.1 : With NAD^+ or $NADP^+$ as an acceptor
1.2.1.12 : glyceraldehyde-3-phosphate dehydrogenase (phosphorylating)

Figure 1.5 : un exemple d'EC-number : 1.2.1.12.

Premier nombre	Classe	Type de réaction
1	Oxydoréductases	Catalysent les réactions d'oxydations et de réduction
2	Transférases	Transfert d'un groupe
3	Hydrolases	Catalysent l'hydrolyse de diverses liaisons
4	Lyases	Clive C-C, C-O et C-N et autres liaisons par des moyens autres que l'hydrolyse ou l'oxydation
5	Isomérases	Catalysent les changements conformationnels au sein d'une molécule
6	Ligases	Catalysent la liaison entre deux molécules par liaison covalente
7	Translocases	Catalysent le déplacement d'ions ou de molécules à travers la membrane ou leur séparation de la membrane

Table 1.1 : Les principales classes d'enzymes.

L'EC-number décrit donc **l'activité enzymatique** de l'enzyme. Deux enzymes avec deux EC-number qui diffèrent seulement par le dernier nombre, ont une activité enzymatique fortement similaire.
Il est à noter que deux enzymes avec des séquences d'acides aminés différentes, peuvent catalyser la même réaction. On parle de isoenzymes. D'un point de vue évolutif, les deux enzymes peuvent être homologues dont les séquences ont beaucoup divergé ou alors il y a eu convergence de deux séquences indépendantes vers la même fonction.
Une enzyme peut aussi catalyser plusieurs réactions dans le cas d'une protéine avec plusieurs domaines ou d'une protéine capable de reconnaître plusieurs substrats de la même famille. Dans ce cas, plusieurs EC-number peuvent être associés à l'enzyme.

Actuellement, dans la base de données ExplorEnz (McDonald *et al.*, 2009) qui recense toutes les activités enzymatiques connues à ce jour et approuvées par l'IUBMB, il existe environ 8000 activités enzymatiques (dernier accès 08/2023) mais seulement environ 6700 EC-numbers sont actuellement utilisés. La table 1.2 montre le caractère dynamique de cette classification depuis sa création et en fonction des nouvelles découvertes. Par exemple pour la classe 1, 1973 sont actuellement approuvés, 426 ont été transférés vers un nouvel EC-number et 98 ont été supprimés. La classification des activités enzymatiques ne cesse d'évoluer en fonction des nouvelles connaissances apportées sur les enzymes et la réaction qu'elles catalysent.

	Class 1 (Oxidoreductases)	Class 2 (Transferases)	Class 3 (Hydrolases)	Class 4 (Lyases)	Class 5 (Isomerases)	Class 6 (Ligases)	Class 7 (Translocases)	All classes
Current:	1973	2026	1332	744	312	244	97	6728
Transferred:	426	100	374	83	13	6	1	1003
Deleted:	98	85	113	31	10	9	0	346
Total:	2497	2211	1819	858	335	259	98	8077

Table 1.2 : Dynamique et nombre d'activités enzymatiques (EC-number) par classe. La première ligne (current) indique le nombre d'EC-number actuellement utilisés. La deuxième ligne (transferred) indique le nombre d'EC-number transférés vers une autre EC-number. La troisième ligne indique le nombre d'EC number qui a été supprimé et la dernière ligne le nombre total. Table tirée de ExplorEnz.

1.3 Les voies et les réseaux métaboliques

Le réseau métabolique représente l'ensemble de toutes les réactions chimiques interconnectées d'un organisme. Une voie métabolique est une série de réactions chimiques pour effectuer un processus métabolique spécifique. Dans cette série de réactions, les métabolites sont transformés en produits qui pourront être utilisés dans les différentes activités cellulaires afin de permettre à l'organisme de s'adapter à son environnement.

La représentation du métabolisme en différentes voies métaboliques par les biochimistes a été toujours largement privilégiée pour permettre une délimitation des différents processus biologiques à décrire et éviter de surcharger les représentations. Cette délimitation peut se faire de manière arbitraire en fonction des processus biologiques d'intérêts (par exemple le métabolisme des acides aminés peut être décomposé en plusieurs petites voies en fonction de chaque acide aminé ou en fonction de sa dégradation ou de sa synthèse). Ainsi la délimitation d'une voie dépend du processus biologique et de l'information que l'auteur veut transmettre mais aussi du public visé (Michal, 1998).

Les voies métaboliques jouent un rôle fondamental dans le métabolisme de tous les organismes vivants, des organismes unicellulaires simples aux organismes multicellulaires complexes.
Le métabolisme est l'un des systèmes biologiques les mieux décrits. Les différentes voies métaboliques pouvant être catégorisées en deux groupes : les voies métaboliques **primaires** et les voies métaboliques **secondaires**. Elles sont responsables de divers processus vitaux, tels que la production d'énergie, la synthèse de nutriments et l'élimination des déchets pour assurer la survie et la croissance de l'organisme. Mais aussi des processus accessoires pour s'adapter à des conditions environnementales spécifiques.

1.3.1 Métabolisme primaire

Le **métabolisme primaire** est impliqué dans des processus cellulaires fondamentaux nécessaires à la survie et à la croissance de l'organisme. Ces voies sont responsables de fonctions cellulaires de base, telles que la production d'énergie, le métabolisme des nutriments et la synthèse des éléments constitutifs essentiels des structures cellulaires.
Les voies métaboliques primaires sont conservées chez beaucoup d'espèces et sont considérées comme universelles chez tous les organismes vivants. Elles sont essentielles pour maintenir la vie et sont généralement partagées entre tous les organismes.
La glycolyse (Figure 1.6) (dégradation du glucose) et le cycle de l'acide citrique (cycle de Krebs) sont des exemples de voies métaboliques primaires.

Mais certains organismes, en particulier ceux vivant dans des environnements extrêmes, ont développé des stratégies métaboliques alternatives et peuvent ne pas posséder certaines parties d'une voie métabolique primaire. C'est le cas par exemple des microsporidies (Dean *et al.*, 2016), des parabasalides (Vickers and Beverley, 2011) ou des organismes méthanogènes.
Par exemple, les archées méthanogènes sont des micro-organismes qui produisent du méthane comme sous-produit métabolique. Beaucoup d'entre eux ne possèdent pas une voie glycolytique complète. À la place, ils utilisent des voies alternatives, telles que la voie Wood-Ljungdahl ou la voie méthylotrophique, pour produire de l'énergie et du méthane (Müller *et al.*, 2008).
Ces organismes ont développé des adaptations uniques pour prospérer dans leurs environnements spécifiques, et leur métabolisme reflète ces stratégies spécialisées.

1.3.2 Métabolisme secondaire

Les voies métaboliques **secondaires**, également appelées métabolisme spécialisé ou accessoire, sont des voies qui ne sont pas présentes universellement chez tous les organismes et qui sont plus spécifiques à certains groupes d'organismes ou d'espèces.

Contrairement aux voies métaboliques primaires, les voies métaboliques secondaires ne sont pas directement impliquées dans des processus cellulaires essentiels. Elles sont plutôt responsables de la production de composés spécialisés qui jouent souvent des rôles écologiques, tels que la défense contre les prédateurs ou des compétiteurs, l'adaptation à des conditions environnementales spécifiques mais aussi un rôle de signalisation dans l'interaction cellulaire ou avec d'autres espèces (Keller, 2019).

Par exemple, l'aflatoxine (Figure 1.6), principalement synthétisée par les champignons du genre *Aspergillus : A.flavus* et *A.parasiticus*, est un métabolite secondaire avec des propriétés toxiques pour les insectes, et une étude récente a mis en évidence que la production d'aflatoxine confère un avantage sélectif en présence d'insecte prédateur (Drott *et al.*, 2017).

Mais les voies métaboliques accessoires peuvent aussi jouer un rôle essentiel dans le développement et la reproduction de certaines espèces.
Leonard (K.J Leonard 1977) a démontré que les spores mutantes de *B.maydis* (pathogène du maïs) déplétées de mélanine ne sont pas capables de survivre sur la plante. La mélanine semblant jouer un rôle essentiel contre les défenses de l'hôte.
Chez *Penicillium,* une espèce est capable de sécréter un métabolite capable d'inhiber la germination du champignon pathogène du riz *Magnaporthe oryzae* (Becker *et al.*, 2012).
En revanche, les imizoquines sont des métabolites endogènes produits par des bactéries phytopathogènes qui sont nécessaires pour la germination normale de l'*A.flavus* , l'absence du métabolite retarde la germination et la surproduction l'accélère (Khalid *et al.*, 2018).

Les métabolites secondaires sont principalement dérivés des voies métaboliques primaires, par exemple les acyl-Coas sont les blocs de construction initiaux qui alimentent la synthèse des métabolites secondaires de polykétides (par exemple l'aflatoxine) et de terpène (par exemple le carotène). Les acides aminés sont principalement les précurseurs des métabolites secondaires issues des peptides non ribosomaux (par exemple la pénicilline) (Figure 1.7).

Les bases de construction des métabolites secondaires mettent aussi en évidence un autre rôle de métabolisme secondaire. En effet, Malik (Malik, 1980) a émis l'hypothèse que les réactions successives qui mènent à un produit final sont toutes aussi importantes que le produit final en lui-même. Dans les conditions de stress, il se peut qu'il y ait une accumulation des métabolites primaires et intermédiaires (potentiellement toxiques). La synthèse des métabolites secondaires est un moyen d'éliminer l'excédent mais aussi de synthétiser un métabolite pour mieux s'adapter à l'environnement.

Même si les voies métaboliques secondaires ne jouent pas un rôle direct dans les processus vitaux, elles jouent un rôle essentiel dans la survie et la croissance de l'organisme.

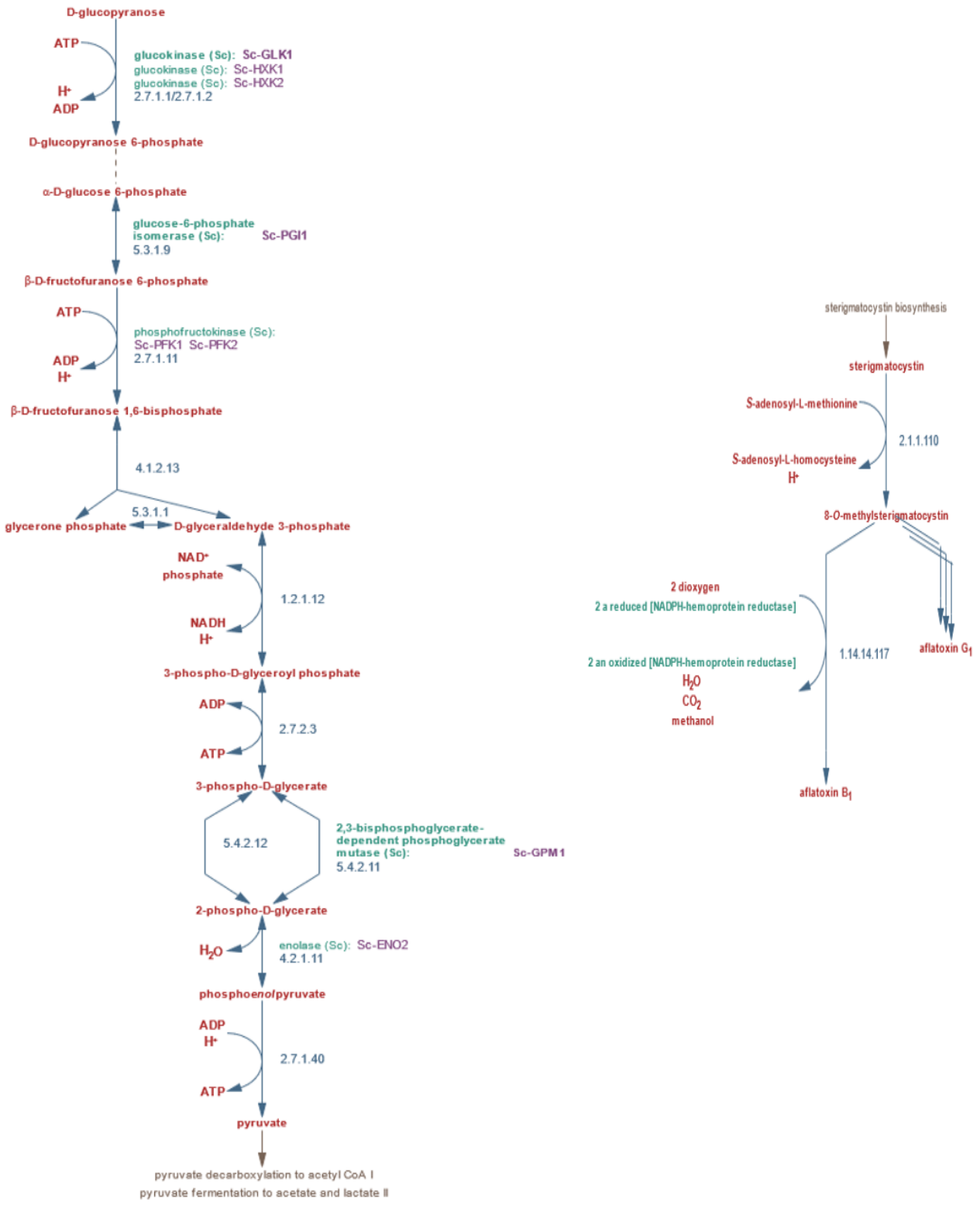


Figure 1.6 : A gauche. Voie de la glycolyse. **A droite.** Voie de synthèse de l'Aflatoxine B1 et G1. Les 2 représentations sont issues de MetaCyc (Karp *et al.*, 2000)

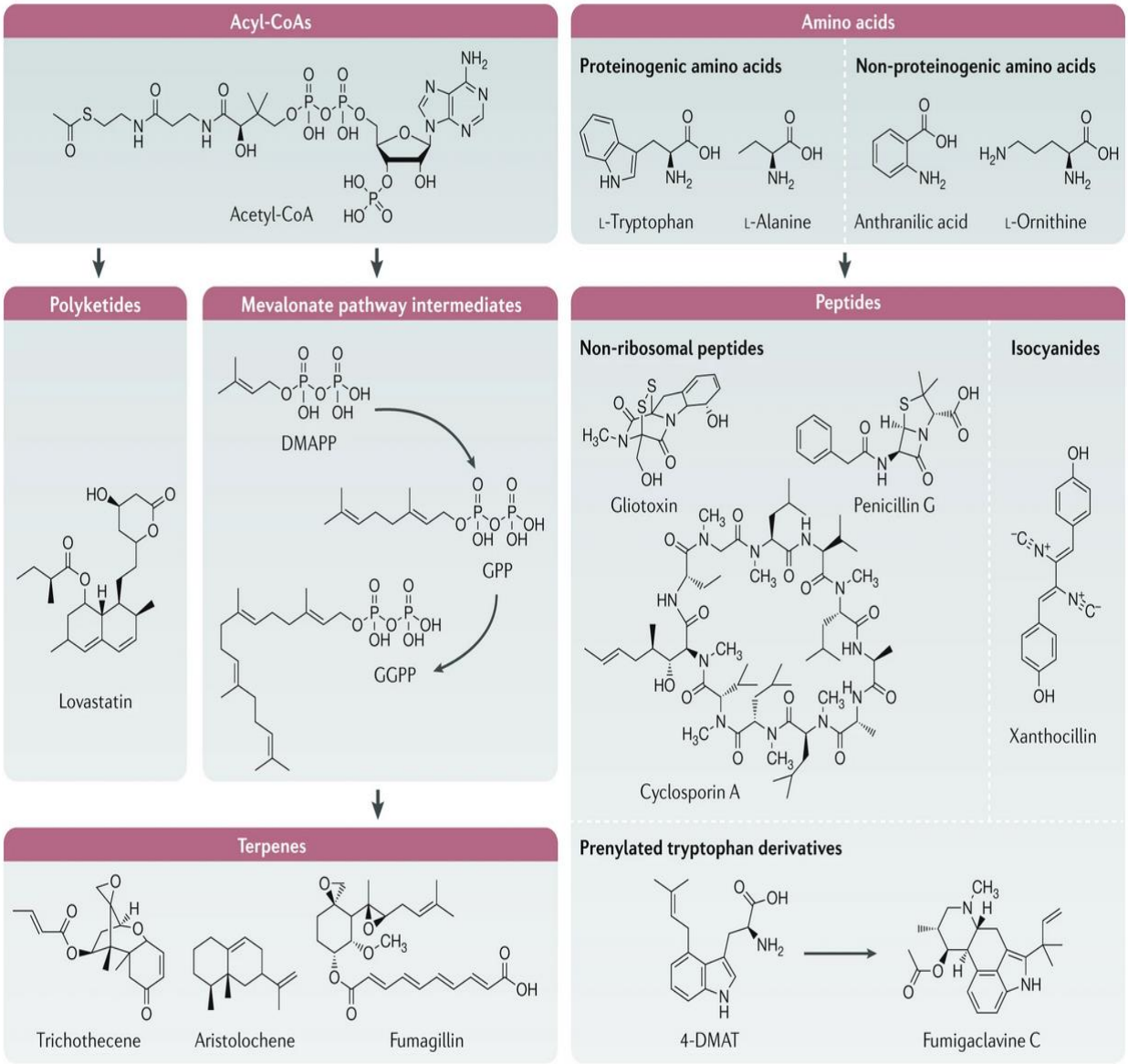


Figure 1.7 : Les précurseurs principaux des métabolites secondaires. La plupart des métabolites secondaires peuvent être regroupés en trois catégories : les polycétides dérivés d'acyl-CoAs, les terpènes dérivés d'acyl-CoAs et les petits peptides dérivés des acides aminés. Figure tirée de (Keller, 2019)

1.3.3 Représentation du réseau métabolique

Le métabolisme est l'ensemble de toutes les réactions biochimiques qui se déroulent au sein d'un organisme. Une réaction biochimique est composée de 3 éléments principaux : l'enzyme qui assure l'activité enzymatique, le substrat et le produit. Dans le cas d'une réaction bidirectionnelle, le produit peut être substrat et vice versa.

Le produit d'une réaction sert de substrat à une autre réaction. Une voie métabolique est une succession de réactions qui vont être utilisées pour effectuer un processus biologique défini.
Le métabolisme peut être représenté sous forme de réseau biologique (Fell and Wagner 2000), dont l'interaction entre les différents éléments est connue.

Pour être manipulable computationnellement, le réseau métabolique est généralement représenté sous forme de **graphe.** Un graphe permet de modéliser les réseaux en se ramenant à l'étude des sommets et des liens entre ces sommets.
Un graphe peut se définir comme une représentation de la relation entre différents éléments. Les interactions sont représentées par des lignes appelées arêtes et les éléments par des points appelés sommets (Figure 1.8).

Dans un réseau métabolique, les nœuds (ou sommets) sont les métabolites (graphe centré sur les métabolites) (Figure 1.8 B) ou les activités enzymatiques (graphe centré sur les activités enzymatiques) (Figure 1.8 C). Les nœuds peuvent aussi représenter les deux entités ensemble (graphe biparti) (Figure 1.8 A).

Dans un graphe biparti, il y a deux types de sommets : les métabolites et les réactions enzymatiques. Dans ce type de graphe, deux sommets de même type ne peuvent pas être connectés. Un sommet de type métabolite est lié à un sommet de type réaction si le métabolite est un substrat ou un produit de la réaction (Figure 1.8 A). Ce type de représentation est la représentation utilisée dans les bases de données KEGG (Kanehisa and Goto, 2000) et MetaCyc (Karp *et al.*, 2000) pour représenter les voies métaboliques car il permet à la fois de représenter les métabolites et les réactions enzymatiques.
Ce type de représentation peut être simplifié en ne gardant qu'un seul type de sommet et l'autre type de sommet devient un lien entre les sommets : graphe centré sur les métabolites (Figure 1.8 B) ou graphe centré sur les réactions enzymatiques (Figure 1.8 C).

Dans un graphe centré sur les métabolites, les nœuds représentent les métabolites et les arêtes les réactions enzymatiques qui permettent la transformation d'un métabolite en une autre (Figure 1.8 B).

Dans un graphe centré sur les réactions, les nœuds représentent les réactions. Les réactions peuvent être labellisées par le nom de l'enzyme, le nom du gène ou bien par l'EC-number. Deux nœuds sont reliés si les deux réactions partagent un composé en commun (Figure 1.8 C), c'est-à-dire le produit de la réaction en amont est le substrat de la réaction suivante.

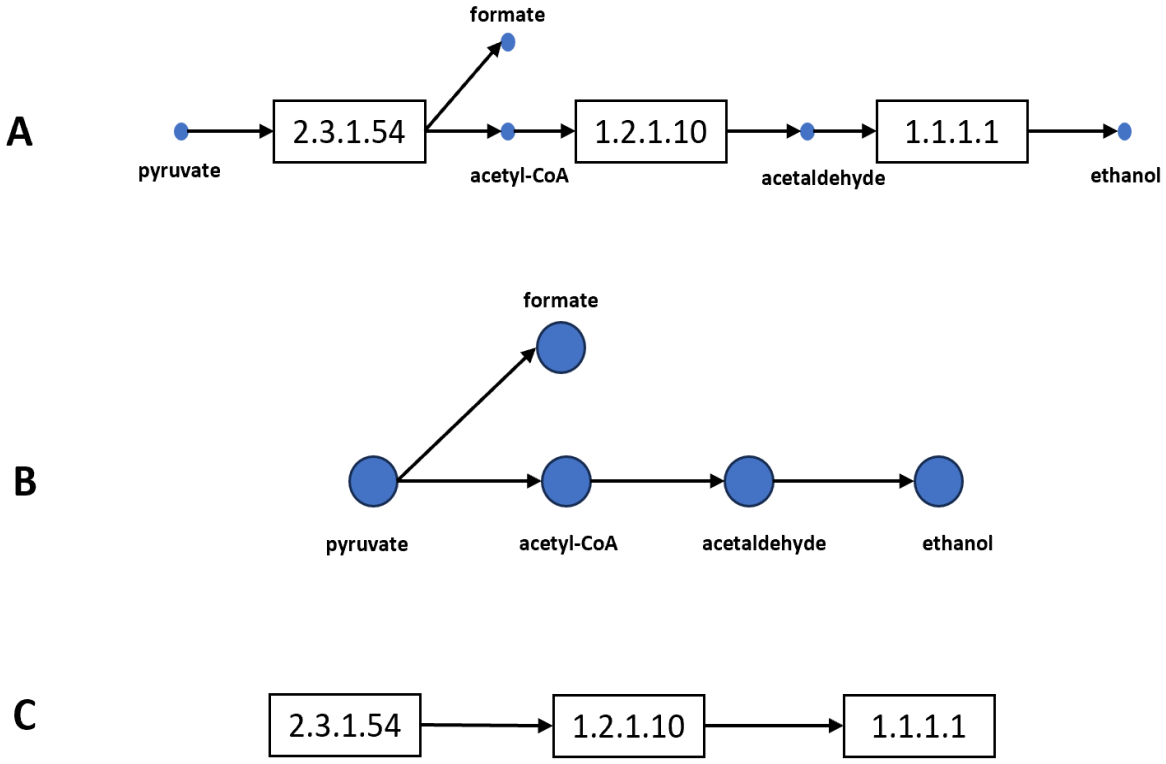


Figure 1.8 : Extrait de la voie de la fermentation du pyruvate en éthanol. Les rectangles représentent l'activité enzymatique et les cercles bleus les métabolites. **A :** Graphe biparti. **B** : Graphe centré sur les métabolites. **C :** Graphe centré sur les réactions représenté par les EC-number.

1.3.4 Base de données des voies métaboliques

Récemment, le nombre de bases de données biologiques pour stocker les connaissances acquises sur les voies métaboliques a augmenté rapidement (Labena *et al.*, 2018). Ces ressources visent à rassembler les informations sur les voies métaboliques et les mettre à disposition de la communauté scientifique pour améliorer et accélérer la recherche sur la biologie des systèmes.

La présence de plusieurs bases de données sur les voies métaboliques comme listé dans la table 1.3 signifie qu'il y a une diversité sur les informations/données qui y sont stockées, les outils mis à disposition pour exploiter ces données et les objectifs (Stobbe *et al.*, 2011). Dans cette table, il y a des bases de données qui sont spécifiques d'espèces comme Malaria (Ginsburg, 2006), maizeDB (Woodhouse *et al.*, 2021) ou HumanCyc (Romero *et al.*, 2004). D'un autre côté, certaines mettent à disposition une collection de voies métaboliques de différentes espèces comme KEGG (Kanehisa and Goto, 2000) ou MetaCyc (Karp *et al.*, 2000).

KEGG et Metacyc font partie des bases de données les plus populaires et les plus diverses

en termes de contenu. Plusieurs auteurs ont comparé le contenu de ces deux bases de données ainsi que d'autres bases bien connues : (Ooi *et al.*, 2010), (Wittig and De Beuckelaer, 2001), (Jing *et al.*, 2014) et (Stobbe *et al.*, 2014). Ces auteurs ont montré la différence en termes de contenu et la façon de mettre à disposition les données (plus de détails seront donnés dans la suite du manuscrit). Bien qu'elles soient différentes dans leur contenu et dans la façon de mettre à disposition les données, une base de données métabolique doit contenir les 4 éléments essentiels constituant le réseau métabolique, à savoir : les réactions biochimiques, les enzymes catalysant ces réactions, les métabolites et les voies métaboliques.

Name	Description	Website
Arabidopsis Reactome	Curated knowledge base of plant biological pathways	http://www.arabidopsisreactome.org
AtIPD	*Arabidopsis thaliana* isoprenoid pathway database	http://www.atipd.ethz.ch
BiGG	Knowledge base of genome-scale metabolic network models	http://bigg.ucsd.edu
BioCyc	A collection of pathway/genome databases (PGDBs) and software tools for understanding their data	http://biocyc.org
BioModels	Online reference repository for quantitative, dynamic models of biological network models	https://www.ebi.ac.uk/biomodels-main
BioPath	Database on biochemical pathways	https://www.mn-am.com/databases/biopath
BioSilico	A web-based database system that facilitates the search and analysis of metabolic pathways	http://biosilico.kaist.ac.kr
BRENDA	Comprehensive enzyme information system	http://www.brenda-enzymes.info
BsubCyc	Database of the bacterium *Bacillus subtilis* and is based on the updated *B. subtilis* 168 genome sequence and annotation	https://bsubcyc.org
CATHACyc	Metabolic pathway database of *Catharanthus roseus*	http://www.cathacyc.org
ECMDB	*Escherichia coli* metabolome database	http://ecmdb.ca
EcoCyc	Encyclopedia of *E.coli* genes and metabolism	https://ecocyc.org
EcoCyc	Scientific database for the bacterium *E. coli* K-12 MG1655	https://ecocyc.org
ENZYME	A repository of information relative to the nomenclature of enzymes	http://enzyme.expasy.org
ExPASy	Biochemical pathway maps	http://web.expasy.org/pathways
FlyReactome	A curated knowledgebase of *Drosophila melanogaster* pathways	http://fly.reactome.org
HMDB	The human metabolome database	http://www.hmdb.ca
HPD	An integrated human pathway database	http://discern.uits.iu.edu:8340/HPD
HUMANCyc	An encyclopedic reference on human metabolic pathways	https://humancyc.org
KaPPA-View4	Kazusa Plant Pathway Viewer	http://kpv2.kazusa.or.jp/kpv4
KEGG	Kyoto Encyclopedia of Genes and Genomes	http://www.kegg.jp
LAMP	Library of Apicomplexan metabolic pathways	http://www.llamp.net
LIPID MAPS	LIPID metabolites and pathways strategy	http://www.lipidmaps.org/resources/resources.html
MaizeGDB	Metabolic pathways in maize	http://maizecyc.maizegdb.org
Malaria	Malaria parasite metabolic pathways	http://mpmp.huji.ac.il
MedicCyc	A biochemical pathway database for *Medicago truncatula*	http://mediccyc.noble.org
MetaCyc	Knowledge of experimentally validated metabolic pathways	https://metacyc.org
MetaNetX/MetanetX.org	Repository and webserver for genome-scale metabolic network Models	http://www.metanetx.org/
MetNetDB	Contains information on networks of metabolic and regulatory and interactions in *Arabidopsis*	http://metnetweb.gdcb.iastate.edu/MetNet_db.htm
MMMDB	Mouse multiple tissue metabolome dataBase	http://mmdb.iab.keio.ac.jp
Model SEED	Web-based resource for high-throughput generation, optimization and analysis of genome-scale metabolic pathway models	http://modelseed.org
MouseCyc	Manually curated database of both known and predicted metabolic pathways for the laboratory mouse	http://mousecyc.jax.org
PathCase	Pathways database system	http://nashua.cwru.edu/PathwaysWeb
PC2	Pathway Commons 2 (integrates a number of pathway and molecular interaction databases supporting BioPAX and PSI-MI formats into one large BioPAX model)	http://www.pathwaycommons.org/pc2
PMN	Plant metabolic network	http://www.plantcyc.org
Reactome	A free, open-source, curated and peer-reviewed pathway database	http://www.reactome.org
RGB	Rat resource center	http://rgd.mcw.edu/wg/home/pathway2/mc cular-pathways2
SABIO-RK	Biochemical reaction kinetics database	http://sabio.villa-bosch.de
SSER	Species specific essential reactions database	http://cefg.cn/sser/index.html.
UniPathWay	Metabolic pathways database	http://www.grenoble.prabi.fr/obiwarehous unipathway
YEASTNET	A consensus reconstruction of yeast metabolism	http://www.comp-sys-bio.org/yeastnet
YMDB	Yeast metabolome database	http://www.ymdb.ca

Table 1.3 : Liste des bases de données de voies métaboliques. Table tirée de (Labena *et al.*, 2018)

KEGG

KEGG (Kyoto Encyclopedia of Genes and Genomes) est une base de données créée en 1995 (Kanehisa and Goto, 2000) pour permettre la compréhension des systèmes biologiques à partir des informations moléculaires. KEGG est devenu une base de connaissance de référence pour l'analyse et l'intégration d'ensemble de données moléculaires issues des technologies expérimentales à haut débit.

Le système de KEGG est composé de 16 bases de données fortement connecté qui peuvent être regroupées en trois blocs (table 1.4):

- « Chemicals information » avec les informations sur les composés (métabolites), les réactions et les enzymes.
- « Systems information » qui contient la représentation graphique des voies métaboliques qui ont été manuellement vérifiés avec la liste des ligands qui le composent.
- « Genomic information » contient des informations sur le génome et les gènes ainsi que la liste des gènes associés aux organismes et aux voies.

Category	Database	Content
Systems information	KEGG PATHWAY	KEGG pathway maps
	KEGG BRITE	BRITE hierarchies and tables
	KEGG MODULE	KEGG modules and reaction modules
Genomic information	KEGG ORTHOLOGY (KO)	Functional orthologs
	KEGG GENES	Genes and proteins
	KEGG GENOME	KEGG organisms and viruses
Chemical information	KEGG COMPOUND	Metabolites and other chemical substances
	KEGG GLYCAN	Glycans
	KEGG REACTION KEGG RCLASS	Biochemical reactions Reaction class
	KEGG ENZYME	Enzyme nomenclature
Health information	KEGG NETWORK	Disease-related network variations
	KEGG VARIANT	Human gene variants
	KEGG DISEASE	Human diseases
	KEGG DRUG KEGG DGROUP	Drugs Drug groups

Table 1.4 : Organisation des informations dans KEGG. Table tirée du site web de KEGG.

Une voie métabolique dans KEGG (stockée dans la base de données KEGG PATHWAY) est une collection/union de réactions de la même voie métabolique de plusieurs espèces. C'est une voie métabolique chimère dont l'ensemble n'est pas présent dans une seule espèce.
Les modules KEGG (KEGG module) sont plus proches de la réalité biologique que les KEGG

PATHWAY parce qu'ils tentent de modéliser les voies métaboliques à partir des réactions des organismes individuellement.

Les KEGG MODULE sont des unités fonctionnelles de gènes ou de réactions vérifiées manuellement. Ils sont divisés en trois modules : les modules de voie (incluant les complexes moléculaires), les modules caractérisant des phénotypes et les modules de réactions.
Ces modules sont construits avec les informations issues des groupes d'orthologues de KEGG (KEGG Orthologue). Ces groupes d'orthologues permettent d'établir la présence/absence des gènes associés à une activité enzymatique ce qui permet d'évaluer la présence/absence et d'évaluer automatiquement si l'ensemble génétique est complet et si l'unité fonctionnelle est présente dans un génome donné. Par exemple, le module M00002 représente le cœur de la glycolyse (Figure 1.9) et la liste des organismes possédant le module complet.

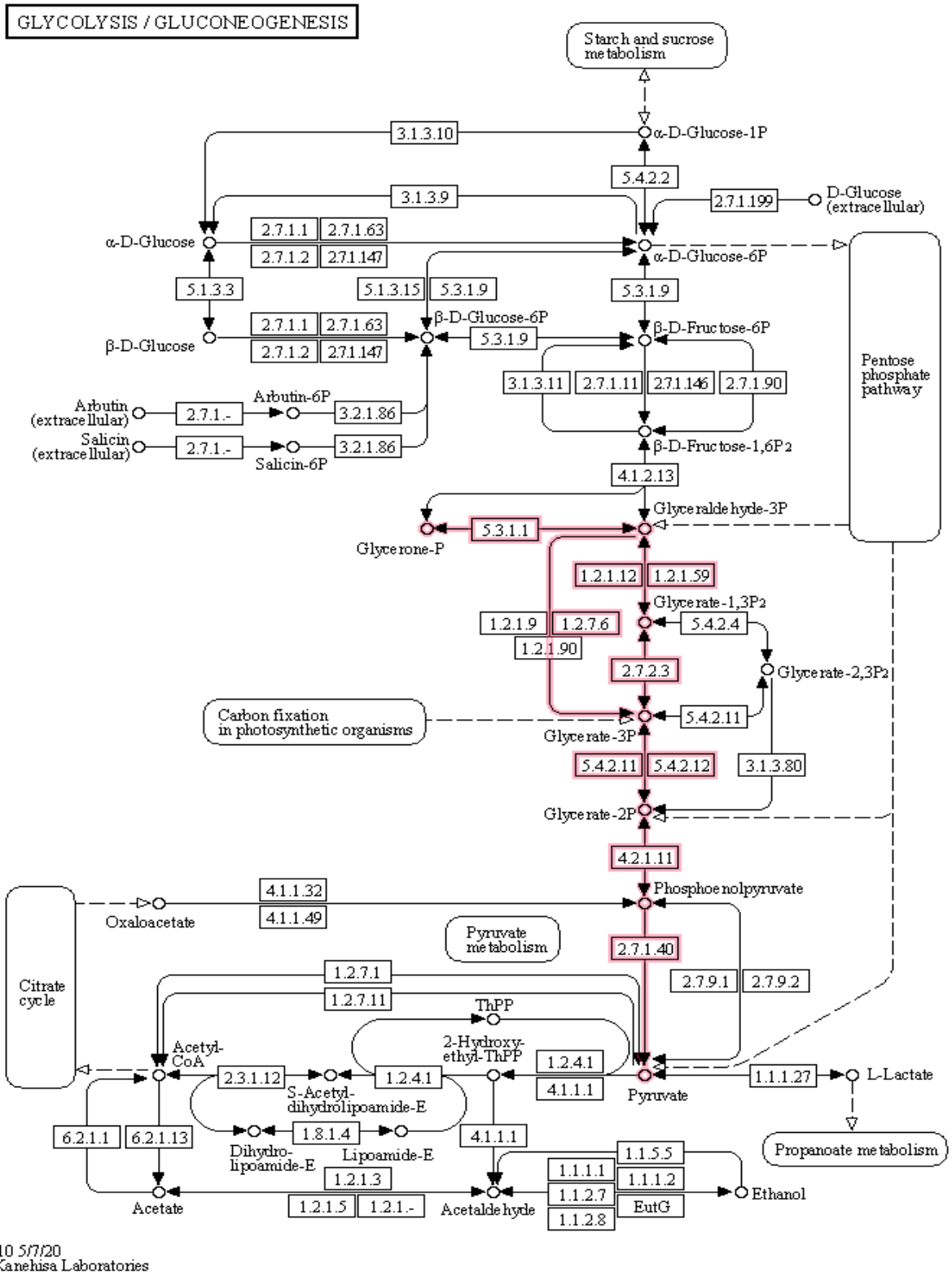


Figure 1.9 : La voie de la glycolyse/glycogenèse dans KEGG. Le module M0002 est coloré en rouge. C'est un module commun à toutes les espèces présentes dans KEGG.

MetaCyc

MetaCyc (Karp *et al.*, 2000) est une base de données de la collection BioCyc (Pathways/Genome Databases). À la différence des autres bases de données de la collection qui sont des bases de données organisme spécifique, MetaCyc est une base de données multi-organismes. Les voies métaboliques dans les bases de données d'organisme spécifique tel que EcoCyc (dédiée à *Escherichia Coli*) sont un mélange de voies vérifiées expérimentalement et prédites computationnellement. MetaCyc ne contient que des voies métaboliques validées expérimentalement.
MetaCyc contient des voies impliquées dans le métabolisme primaire et secondaire, ainsi que les métabolites, les réactions, les enzymes et les gènes associés.
MetaCyc relie les informations sur les voies, les réactions, composés, protéines et les gènes.

Aujourd'hui, KEGG et MetaCyc offrent un accès gratuit à une partie de ses données mais propose un service payant pour avoir accès a des fonctionnalités et des données plus avancées.

Comparaison entre KEGG et MetaCyc

La table 1.5 compare le nombre d'éléments entre les deux bases de données.
Le nombre de voies dans MetaCyc est largement supérieur au nombre de voies métaboliques dans KEGG. KEGG représente l'union des voies métaboliques de plusieurs espèces sous une seule représentation, alors que MetaCyc représente une voie qui a été vérifiée expérimentalement chez plusieurs organismes. Si une variante de la voie existe dans d'autres organismes, elle sera représentée avec une autre présentation dans MetaCyc.
Dans KEGG, les voies sont centrées sur les synthèses ou la dégradation d'un ou plusieurs métabolites reliés à un processus biologique. Par exemple, le métabolisme des purines contient à la fois les voies de dégradation et les voies de synthèse à partir de différents substrats.
Dans MetaCyc les voies sont délimitées par le processus biologique et la conservation des réactions dans l'évolution. Dans le métabolisme des purines, MetaCyc propose 11 entrées en fonction des processus biologiques (synthèse ou dégradation) et des points de départs du processus.
Cette délimitation fait que les voies dans KEGG sont plus grandes que les voies dans MetaCyc (Altman *et al.*, 2013).

MetaCyc met aussi à disposition des « superpathways » (Figure 1.10 B). Les superpathways sont plus proches des voies de KEGG alors que les KEGG MODULE sont plus proches des voies de MetaCyc car ils essaient de capturer la voie chez les organismes individuellement (Figure 1.10).
Les voies globales (KEGG pathways et MetaCyc superpathways) sont utiles pour montrer un contexte biologique plus large et avoir un aperçu global de la voie dans laquelle elle fonctionne.

Les voies plus petites sont avantageuses parce qu'elles correspondent à une seule fonction biologique, les enzymes qui les composent sont généralement régulées et conservées ensemble dans l'évolution.

Le format de fichier utilisé pour stocker les voies dans les deux bases de données est très différent. Dans KEGG, le format de fichier utilisé est le format KEGG Markup Langage (KGML) alors que dans MetaCyc les voies sont stockées au format Biopax.
Les différences entre ces deux formats seront présentées dans la partie III.9 de ce manuscrit.
Une différence majeure sur la représentation des voies métaboliques se situe sur les détails des réactions dans les voies métaboliques. KEGG ne représente dans ses voies que les métabolites principaux, alors que dans MetaCyc nous pouvons observer la présence des molécules telles que l'ATP, ADP , H+…. La présence de ces molécules ubiquitaires peut avoir des impacts sur les types d'analyses qui seront réalisées à partir de ces voies.

	KEGG	**MetaCyc**
Nombre de voie	160	3006
Nombre de réactions	11941	17780
Nombre d'enzymes	8056	14250
Nombre de composés	19119	18801
Nombre d'organismes	9187	3350
Format de fichier et méthode d'accès	KGML, API, dbget, flat files	Biopax, API, Pathways tools (for flat file)
Ontologie	Non applicable	Composante cellulaire
Détails des réactions	Seulement les métabolites principaux	Complètes

Table 1.5 : Comparaison entre KEGG et MetaCyc. Les valeurs pour chaque voie proviennent des versions du 08/2023 de chaque base de données.

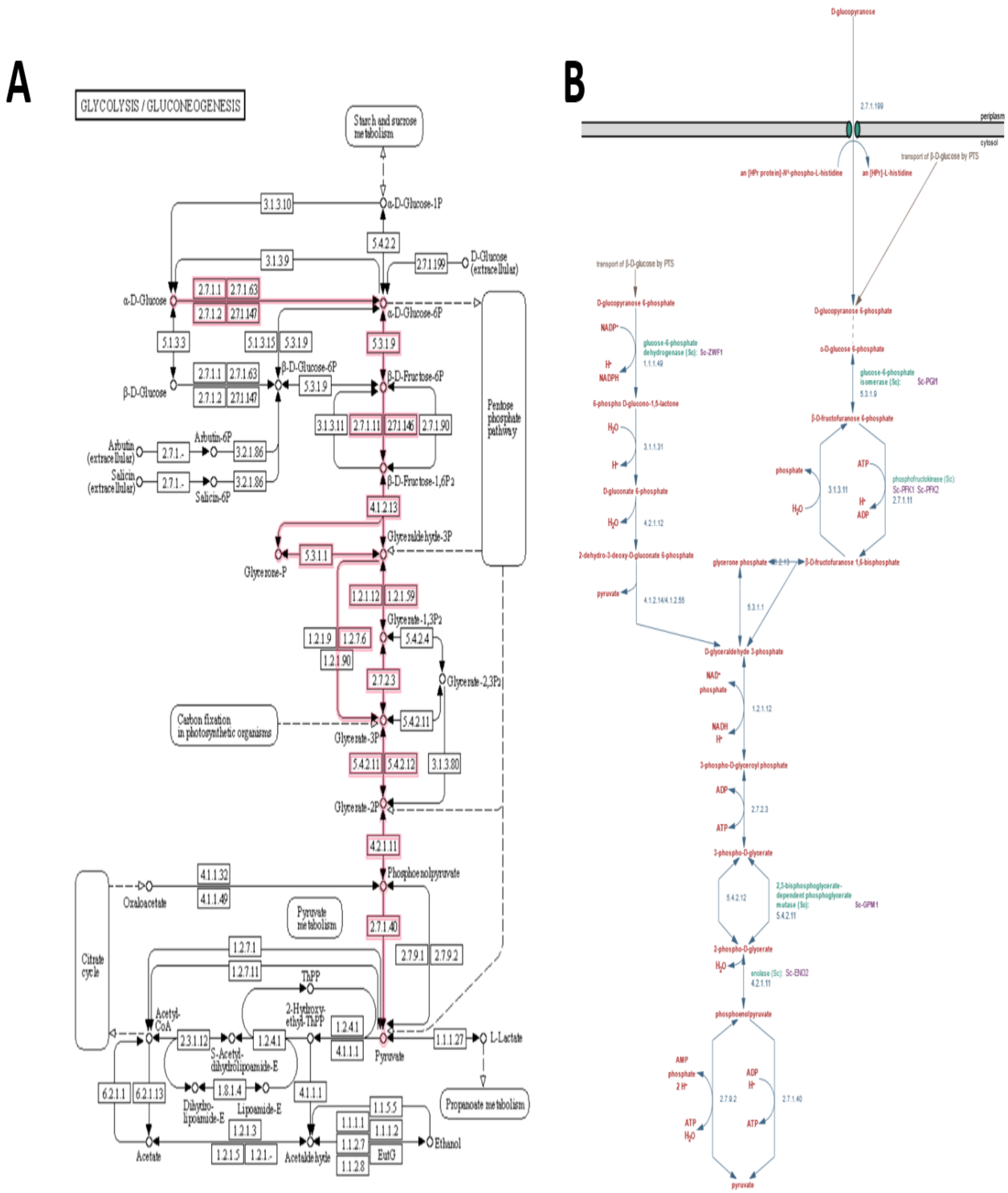


Figure 1.10 : A Voie de la glycolyse/Glycogenèse dans KEGG. Le module qui correspond à Embden-Meyerhof pathway est coloré en rouge. **B** Superpathway de la voie Embden-Meyerhof dans MetaCyc. Les Figures ont été extraites des sites web de KEGG et de MetaCyc.

Chapitre:

2 Evolution du métabolisme : état de l’art

2.1 Origine des activités enzymatiques d'un point de vue génomique

Les éléments principaux du métabolisme sont les enzymes qui catalysent les réactions et les métabolites qui sont interconvertis entre eux.
Le réseau métabolique évolue donc en fonction de ces deux éléments. Tout d'abord en fonction de la disponibilité des métabolites (substrats) dans le milieu pour synthétiser les métabolites essentiels et produire de l'énergie, mais aussi pour éliminer les métabolites intermédiaires en excès et qui peuvent être toxiques pour l'organisme. Pour s'adapter à ces changements de l'environnement, il est essentiel de pouvoir faire évoluer le métabolisme pour utiliser les substrats disponibles, éliminer les nouveaux métabolites qui peuvent être toxiques et maintenir l'équilibre du système. Cette adaptation passe par l'émergence d'une nouvelle activité enzymatique ou la réorganisation des activités enzymatiques déjà présentes.

2.1.1 Emergence d'une nouvelle activité enzymatique

L'un des principaux mécanismes de l'apparition d'un gène avec une nouvelle fonction est la **duplication-divergence** (Ohno, 2013). Un gène va se dupliquer et une des copies va conserver la fonction ancestrale qui généralement est essentielle pour l'organisme. L'autre copie ,qui est sous une pression de sélection moins forte, va accumuler les mutations qui seront à l'origine d'une nouvelle propriété fonctionnelle -néo fonctionnalisation (Figure 2.1). L'acquisition d'une nouvelle fonction peut parfois entraîner une perte de la fonction initiale pour la seconde copie du gène dupliqué (Haldane, 1933).
Plusieurs processus peuvent être à l'origine de la duplication d'un gène.
Il peut s'agir d'une duplication du génome entier, qui va entraîner une duplication de tous les chromosomes et de tous les gènes. Crow et Wagner ont par exemple confirmé que le genre *Saccharomyces* a connu une ancienne duplication de son génome (Crow and Wagner, 2006).
À une plus petite échelle, des processus locaux peuvent entraîner une duplication du gène. Le premier est le fruit d'un crossing-over inégal pendant la méiose. Ce mécanisme est favorisé par la présence d'éléments répétés ou dupliqués dans le génome qui augmente la fréquence du mauvais alignement du chromosome pendant la méiose (Kaessmann, 2010). Le mauvais alignement peut entraîner des crossing-over inégaux qui augmentent le nombre de copies de gènes sur un chromosome recombinant tout en diminuant le nombre de copies de gènes sur l'autre chromosome recombinant.
L'action des éléments transposables peut aussi avoir comme conséquence une duplication de gène. Ce sont des éléments qui ont la capacité de se déplacer d'une position à l'autre dans le génome. Lorsqu'ils se déplacent, certains d'entre eux (Pavlicek *et al.*, 2006; Kaessmann, 2010) sont capables de copier des éléments du génome et de les réinsérer ailleurs dans le génome, ayant pour conséquence la duplication de la séquence de l'hôte.
Un autre mécanisme, tout aussi important, à l'origine d'un gène est l'émergence d'un gène *de novo* à partir du génome non codant (Carvunis *et al.*, 2012). L'émergence d'un nouveau

gène à partir du génome non-codant est aujourd'hui considérée comme un réservoir additionnel à la création de nouveauté génétique. De nouvelles études rapportant des gènes de novo sont publiées chaque année (Knowles and McLysaght, 2009; Tautz and Domazet-Lošo, 2011; Wu *et al.*, 2011; Murphy and McLysaght, 2012; Li *et al.*, 2016; Vakirlis *et al.*, 2018; Zhang *et al.*, 2019; Heames *et al.*, 2020)

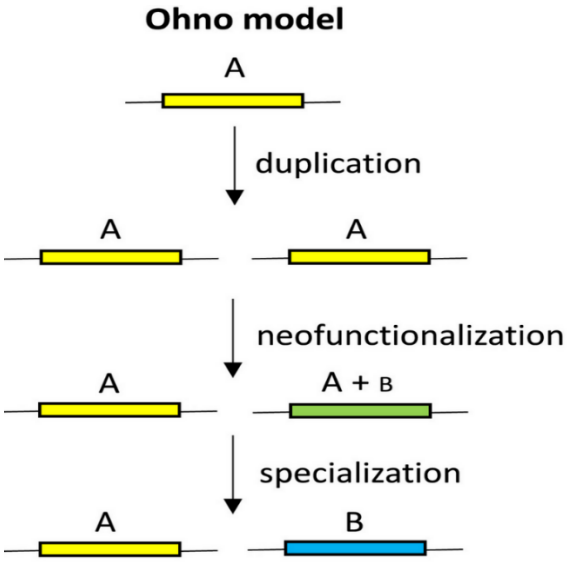


Figure 2.1 : Duplication du gène jaune avec la fonction A. Après la duplication l'une des copies acquiert une nouvelle fonction B puis seule la fonction B est conservée dans la copie. Figure tirée de (Copley, 2020)

Le transfert d'élément génétique entre les organismes (**Horizontal Gene Transfer**) peut avoir un effet direct ou à long terme dans l'évolution de l'organisme récepteur. Bien que les transferts horizontaux soient plus fréquents entre bactéries, le transfert de matériel génétique vers un organisme eucaryote se produit également. Dans ce cas les bactéries servent généralement de donneurs et les organismes tels que les champignons, les plantes et les animaux sont les récepteurs (Garcia-Vallvé *et al.*, 2000; Keeling and Palmer, 2008; Rancurel *et al.*, 2017). Mais des transferts de gènes d'eucaryotes à eucaryotes ont été aussi recensés (Emamalipour *et al.*, 2020) et plus particulièrement chez les champignons endophytes (se développant à l'intérieur des plantes) où un important échange de matériel génétique semble s'effectuer entre la plante et le champignon (Tiwari and Bae, 2020).

Chez les champignons, l'analyse des gènes associés au métabolisme (codant des protéines à activité catalytique) a montré que 3% semblent issus de transferts horizontaux (Wisecaver *et al.*, 2014) dont une grande majorité est associée aux gènes qui sont organisés en cluster. En effet, les gènes codant les activités enzymatiques associées au métabolisme accessoire sont la plupart associés dans le génome sous forme de cluster (Keller, 2019). Cette organisation particulière des gènes dans le génome facilite le transfert d'une voie métabolique secondaire vers un autre organisme par transfert de cluster. C'est par exemple le cas des

gènes impliqués dans la synthèse de l'aflatoxine (Figure 2.2). Le transfert horizontal d'une partie de ce cluster (enzymes synthétisant le sterigmatocystin) d'*Aspergillus* vers *Podosporina anserina* a été montré il y a quelques années (Slot and Rokas, 2011).

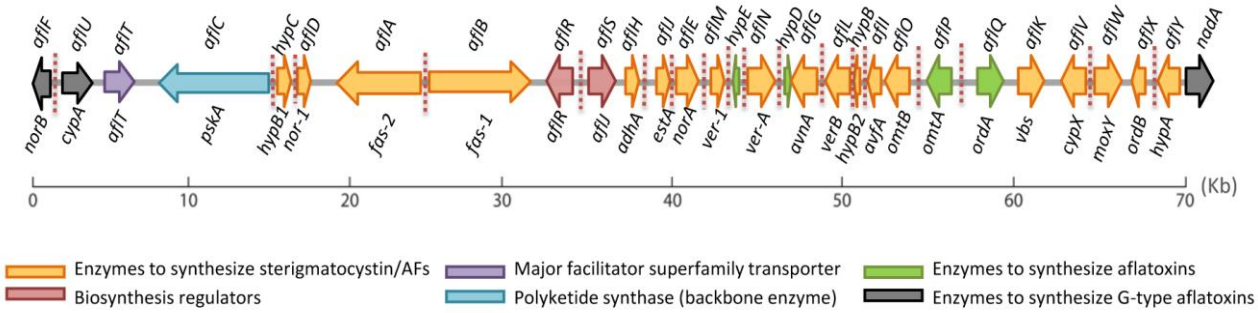


Figure 2.2 : Aflatoxine gene cluster. Figure tirée de (Caceres *et al.*, 2020).

2.1.2 Perte d'une activité enzymatique

La perte de gènes dans certaines espèces est l'un des principaux processus évolutifs qui ont été mis à jour par des analyses comparatives de gènes à partir de génomes entièrement séquencés (Aravind *et al.*, 2000; Moran, 2002). Beaucoup d'attention a été apportée au mécanisme d'évolution des gènes par duplication-divergence. À l'inverse, la perte de gène (et de la fonction associée) a toujours été associée à la perte d'un élément sans conséquence fonctionnelle apparente et qui n'est plus nécessaire pour l'organisme (Albalat and Cañestro, 2016).

Deux mécanismes moléculaires peuvent conduire à la perte d'un gène et à sa fonction : premièrement la perte physique du gène suite à un crossing-over inégal pendant la méiose ou la mobilisation d'un élément transposable.
Deuxièmement, l'accumulation de mutation qui va conduire à une perte de fonction (pseudogenisation). La perte de fonction sera due à une mutation non-sens, une délétion ou une insertion qui va toucher un acide aminé essentiel dans la fonction ou une région régulatrice qui va éteindre l'expression du gène.
La perte d'un gène dépend du caractère essentiel du gène. Cela dépend des conditions physiologiques de l'organisme affectées par la non-fonctionnalisation de ce gène. Par ailleurs, la fonction d'un gène n'est pas la même dans tous les environnements, par conséquent la variabilité de l'environnement modifie le caractère essentiel d'un gène (Albalat and Cañestro, 2016).

On parle de « **regressive evolution** » quand la perte implique des gènes qui ne sont plus nécessaires. La fonction n'apporte aucun avantage sélectif ni de désavantage dans un nouvel environnement (Albalat and Cañestro, 2016) .
The « **Less-is-more** » est proposé quand la perte de fonction confère un avantage sélectif

pour s'adapter à un nouvel environnement (Olson, 1999). Par exemple, des souches de *S.cerevisiae* prélevées dans un environnement chaud et riche en sucre ont subi des pertes de gènes par rapport aux souches vivant dans un environnement froid. Les gènes perdus par les souches issues d'un environnement chaud et riche en sucre a procuré un avantage pour survivre dans cet environnement. Mas les souches issues d'un environnement froid qui possèdent ces gènes sont sensibles à une forte température (Will *et al.*, 2010).

Bien que les pertes de gènes semblent être neutres, ce dernier exemple soutient l'idée que la perte de gènes peut être une force évolutive particulièrement efficace pour s'adapter à des changements de l'environnement.

2.2 Mécanisme à l'origine du métabolisme

La détermination de l'origine des voies métaboliques d'aujourd'hui et comment ces voies métaboliques ont évolué est toujours une question ouverte. Quel est par exemple l'apport des mécanismes qui ont été décrits précédemment (duplication, transfert horizontal, perte de gène..) dans la construction et l'évolution des voies métaboliques ?

Plusieurs théories ont tenté d'expliquer l'origine des voies métaboliques. Les deux principales théories sont basées sur la duplication des gènes.

Le modèle d'évolution rétrograde (Horowitz, 1945)

Dans cette première tentative d'explication, Horowitz a proposé que les enzymes sont acquises par la duplication de gènes et dans l'ordre inverse des voies actuelles pour répondre à une diminution dans l'environnement d'un substrat essentiel . C'est-à-dire si la biosynthèse d'un composé 1 nécessite la transformation des composés 4 en 3, 3 en 2 puis 2 en 1 avec leurs enzymes respectives, le produit final 1 est le composé le plus ancien.

Par exemple, dans la Figure 2.3, le composé 1 est essentiel pour la survie à l'état ancestral. Quand le composé 1 a commencé à diminuer dans l'environnement, cette condition a exercé une pression de sélection pour la survie et reproduction des cellules capables de biosynthétiser le composé 1 à partir d'un composé 2 à l'aide de l'enzyme E1.
Cette réaction et l'augmentation de la population entraînent une diminution de la concentration du composé 2 et une sélection va s'opérer sur les cellules capables de synthétiser le composé 2 à partir du composé 3 à l'aide de l'enzyme E2.
Sachant que l'enzyme E1 a déjà une affinité avec le composé 2 (on parle de promiscuité enzymatique), l'origine de l'enzyme E2 se fera le plus probablement à partir de la duplication puis la divergence de l'E1. Sur de nombreuses itérations, une voie peut être construite à partir du produit final jusqu'au précurseur initial.
Dans cet exemple, le produit final, qui est le composé 1, est le composé le plus ancien.

Dans ce modèle, les enzymes d'une même voie qui sont adjacentes sont forcément

homologues et organisées en cluster. Cette hypothèse a été vérifiée par (Alves *et al.*, 2002; Díaz-Mejía *et al.*, 2007) où ils ont démontré que quelques enzymes du réseau métabolique ont évolué en suivant ce modèle.

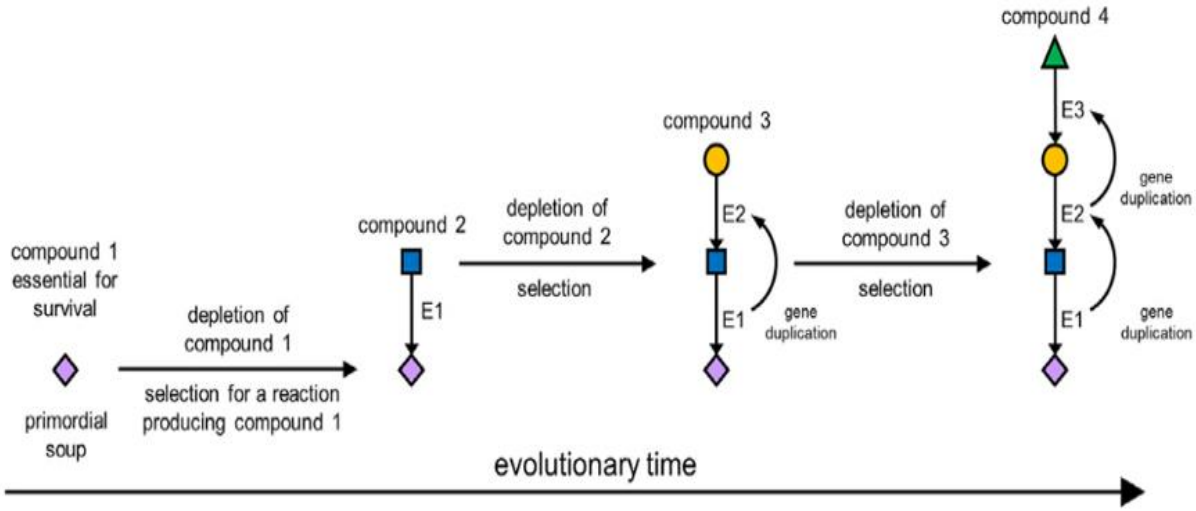


Figure 2.3 : Modèle d'évolution rétrograde proposé par Horowitz. Figure tirée de (Scossa and Fernie, 2020)

Le modèle d'évolution patchwork (Jensen, 1976)

Cette hypothèse en revanche, pose que les voies métaboliques ont été construites en recrutant des enzymes qui pouvaient initialement transformer un large éventail de substrats chimiquement proches (Figure 2.4). Cette théorie implique que la toute première enzyme était non-spécifique (enzyme E1 sur la Figure 2.4). À la suite d'une duplication, une des copies du gène codant cette enzyme va se spécifier pour un substrat particulier (sous-fonctionnalisation) et augmenter son rendement pour un substrat spécifique qui va lui conférer un avantage sélectif (l'enzyme E2 dans la Figure 2.4 est spécifique du substrat 3). L'autre copie (gène 1 codant pour l'enzyme E1) va garder la spécificité avec les autres substrats et peut devenir moins spécifique avec le substrat qui est spécifiquement catalysé par la nouvelle enzyme spécifique (L'enzyme E1 a perdu son efficacité de catalyser le substrat 3). La succession des évènements de duplications et spéciations va permettre la spécialisation de la voie.
Ce modèle a été utilisé pour expliquer l'évolution de plusieurs voies de biosynthèses des acides aminés et du cycle de KREBS (Fondi *et al.*, 2009)

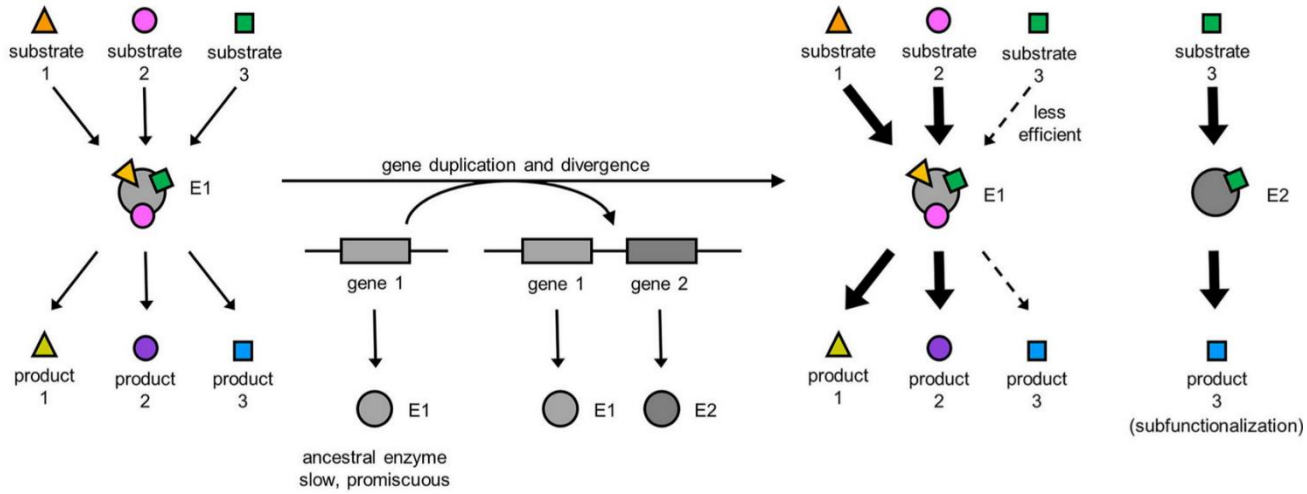


Figure 2.4 : Modèle d'évolution patchwork proposé par Jensen. Figure tirée de (Scossa and Fernie, 2020)

Même si les deux modèles sont totalement opposés, ils sont complémentaires (Lazcano and Miller, 1999). L'hypothèse actuelle suppose que le modèle rétrograde ait pu jouer un rôle sur l'évolution de la forme ancestrale du réseau métabolique et le modèle d'évolution patchwork est l'acteur prédominant dans l'évolution moderne des enzymes métaboliques. (Ralser *et al.*, 2021)

2.3 Approche par génomique comparée

Le génome est le support de l'information génétique de tous les organismes. C'est un élément commun qui unit tous les êtres vivants mais en même temps il est à l'origine de notre diversité. La diversité des espèces et des formes de vie que nous observons est le résultat de l'expression de ce génome. Au cours de l'évolution, ce génome est transmis le plus fidèlement possible aux descendants pour assurer la survie de l'espèce tout en autorisant une plasticité de celui-ci pour permettre aux espèces d'évoluer et de s'adapter.

La génomique comparée est un domaine de la biologie moléculaire où on utilise différents outils pour comparer le matériel génétique de différentes espèces. Cette approche permet d'étudier les différences et les similarités dans le génome entre les différents organismes.
La génomique comparée est aussi un outil très puissant pour étudier l'évolution.
« Nothing makes sense in biology except in the light of evolution » Th. Dobzhansky, The American Biology Teacher, 1973
La génomique comparée permet en effet d'identifier les séquences d'ADN (gènes, séquences régulatrices…) qui sont conservées entre les espèces et les séquences d'ADN qui sont spécifiques de certaines espèces et qui sont responsables de leurs caractéristiques biologiques. L'hypothèse la plus parcimonieuse étant que les éléments qui sont conservés proviennent d'un ancêtre commun ce qui suggère l'importance de ces gènes pour le

fonctionnement de ces espèces.
En s'appuyant sur la relation évolutive entre les espèces (la phylogénie) et les différences dans leur génome, on peut comprendre comment les espèces et les éléments de leurs génomes ont évolué par rapport à leur ancêtre commun et par rapport aux autres espèces.

Avec l'évolution fulgurante des outils et des méthodes de séquençage de l'ADN qui ont permis de séquencer de plus en plus de génomes, mais aussi des outils informatiques, la génomique comparative est devenue une méthode très puissante pour comprendre les différences entre les différentes formes de vie.

Mais malgré ces avancées, la génomique comparée présente plusieurs défis et difficultés.
Tout d'abord l'assemblage précis d'un génome à partir des séquences brutes est une tâche complexe. Malgré la robustesse et la fiabilité des méthodes d'assemblages actuelles, des erreurs peuvent survenir (Lin *et al.*, 2011; Wick and Holt, 2019). Ces erreurs vont principalement impacter l'absence d'une séquence dans le génome. Mais les techniques de séquençage peuvent aussi introduire des biais tels que les erreurs de séquençage ou une mauvaise couverture de toute ou une partie du génome (Ross *et al.*, 2013). Ce qui peut rendre la comparaison des séquences moins fiable.
Une autre difficulté majeure en génomique comparée est l'évolution rapide de certains gènes et régions du génome, ce qui signifie qu'ils peuvent être différents entre les espèces proches. La détection de ces différences peut être difficile. Par ailleurs, l'annotation des gènes est cruciale en génomique comparée. L'annotation se base principalement sur un critère de similitude entre les séquences d'un gène non annoté et un gène annoté. Mais cela peut être difficile lorsque les gènes sont mal conservés.
Des espèces distantes dans l'arbre peuvent également développer des caractéristiques similaires (convergence) en réponse à une pression environnementale similaire. Cela peut conduire à une similitude dans les génomes qui ne reflètent pas un lien de parenté réel.

Une approche par génomique comparée pour détecter les séquences d'ADN (gènes, protéines…) similaires entre les espèces commence généralement par un alignement entre elles des séquences des gènes prédits afin d'identifier les séquences orthologues.
Par définition, deux gènes sont **homologues** si les deux gènes partagent un ancêtre commun. Deux gènes homologues sont **orthologues** si les deux gènes ont divergé par un événement de spéciation. Deux gènes homologues sont paralogues s'ils ont divergé par un évènement de duplication (Figure 2.5).

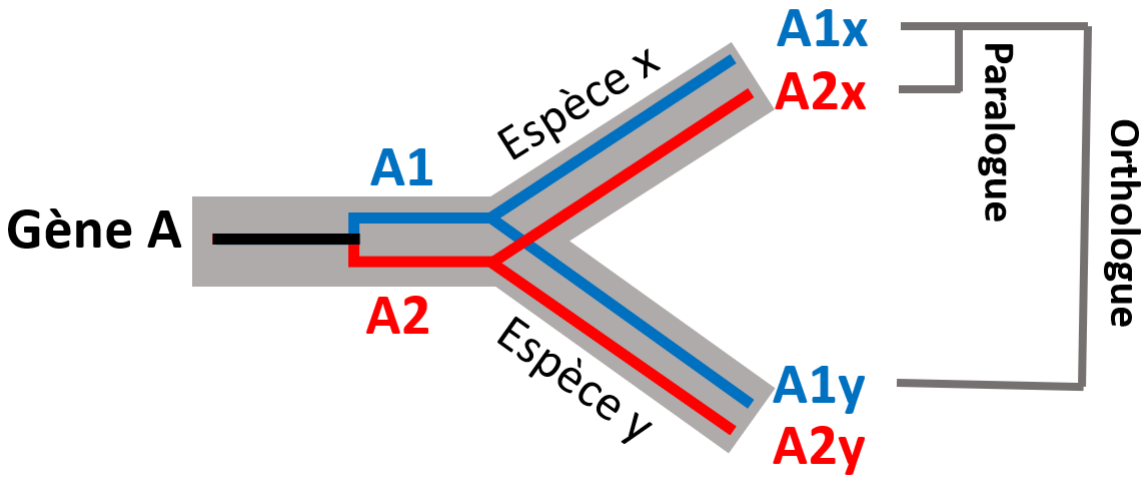


Figure 2.5 : Différence entre orthologie et paralogie. Le gène ancestral A s'est dupliqué en gène A1 (bleu) et A2 (rouge). Suite à un événement de spéciation, les gènes A1 et A2 sont présents dans les espèces x et y, (A1x,A2x) et (A1y,A2y) respectivement. A1x et A1y sont orthologues. A1x et A2x sont des paralogues. Figure adaptée de Wikipedia.

La détection des groupes d'orthologues (chaque groupe est constitué d'une liste de gènes orthologues) permet d'identifier les espèces qui partagent des gènes en commun. Ainsi on pourra déterminer l'occurrence d'un gène dans plusieurs espèces. Par exemple dans la Figure 2.5, A1x et A1y forment un groupe d'orthologue et A2x et A2y forment un autre groupe d'orthologue. Cette information sur les gènes orthologues peut être représentée sous forme d'un vecteur de présence (1) et d'absence (0) appelé **profil phylogénétique** (Pellegrini *et al.*, 1999). Si un orthologue est identifié dans une espèce, il sera annoté 1, si aucun orthologue de ce gène n'est retrouvé dans une espèce, il sera annoté 0 (Figure 2.6).
Un groupe d'orthologues de la protéine A de la Figure 2.6 est constitué des séquences protéiques provenant des génomes 1, 2 et 3.

Profil d'un gène

	Génome 1	Génome 2	Génome 3	Génome 4
Protéine A	1	1	1	0
Protéine B	1	1	1	0
Protéine C	1	0	1	1
Protéine D	1	1	0	1

Profil d'un organisme

Figure 2.6 : Profil phylogénétique de 4 protéines dans 4 génomes. Les lignes représentent les protéines (ou gènes) et les colonnes les génomes. Une valeur de 1 dans le tableau indique la présence de la protéine dans le génome et une valeur de 0 indique l'absence. Toute une ligne représente le profil d'un gène et toute une colonne représente le profil d'un organisme. Figure adaptée de (Bowers *et al.*, 2004)

La méthode par **profilage phylogénétique** a été initialement utilisée pour inférer la fonction de protéines en se basant sur l'hypothèse que les protéines fonctionnellement reliées (sous une même voie ou qui interagissent ensemble) évoluent de manière corrélée. Ainsi au cours de l'évolution, ces protéines fonctionnellement reliées sont probablement conservées ensemble ou éliminées ensemble. Cette méthode a permis d'inférer la fonction de protéines non caractérisées mais aussi de montrer que ces protéines opéraient dans la même voie métabolique (Pellegrini *et al.*, 1999).

La comparaison des profils phylogénétiques permet d'identifier les gènes communs à l'ensemble des espèces, mais également des gènes appartenant à des groupes d'espèces monophylétiques ainsi que des gènes appartenant à des groupes d'espèces non reliées taxonomiquement mais dont on peut imaginer qu'elles partagent une propriété biologique similaire. Ces ensembles de gènes qui co-évoluent forment un module évolutif.

Cette méthode a permis par exemple de montrer les voies métaboliques conservées dans les 3 domaines du vivant, mais aussi des voies ou modules évolutifs propres à un certain clade (Li *et al.*, 2014) . Peregrin-Alvarez et ses collaborateurs (Peregrín-Alvarez *et al.*, 2009) ont démontré avec une approche par profilage phylogénétique que les voies associées au métabolisme du Glycane sont très conservées chez les métazoaires et les voies de xénobiotiques sont très conservées chez les bactéries. Ces profils phylogénétiques ont été aussi associés à des phénotypes afin de comprendre le mécanisme sous-jacent (MacDonald and Beiko, 2010).

2.4 Analyse sous forme de graphes

Une autre approche pour analyser le métabolisme est d'analyser la structure du réseau métabolique. Le réseau métabolique pouvant être représenté sous forme de graphe.

En tirant parti de la théorie des graphes qui est une discipline mathématique et informatique qui étudie les graphes, il est possible d'analyser la structure du réseau métabolique.
La théorie des graphes permet une analyse topologique du réseau métabolique.
Outre le réseau métabolique (Fell and Wagner, 2000), une analyse topologique sur un large éventail de réseau biologique réel tel que le réseau d'interaction protéique (Jeong *et al.*, 2001) ou un réseau de régulation génique (Yook *et al.*, 2005; Bhan *et al.*, 2002) a été effectuée avec ce type d'approche.
L'analyse topologique de ces réseaux biologiques a démontré un ensemble de propriétés qui les distinguent d'un réseau aléatoire : la distribution des degrés des nœuds suit une loi de puissance (Barabási and Albert, 1999), les réseaux sont invariants d'échelle (ou « scale-freeness ») et enfin, ils sont de type « small-world ».
L'existence de ces propriétés conduit à l'idée que les réseaux biologiques sont régis par une loi universelle (Barabási and Oltvai, 2004) et permettent le développement de modèles

mathématiques pour générer un réseau capturant ces propriétés topologiques.

Avant de discuter des propriétés topologiques régissant les réseaux biologiques, nous allons d'abord définir quelques notions liées à la théorie des graphes. La définition de ces notions permettra une meilleure compréhension des propriétés topologiques des réseaux métaboliques mais aussi des différentes méthodes utilisées durant ce projet de thèse.

2.4.1 Quelques définitions liées à la théorie des graphes

Mathématiquement, un graphe G(V,E) se définit comme un ensemble de nœuds ou sommets V (V pour « vertice ») et un ensemble d'arêtes E (« edge »). Chaque élément de E est un pair de sommets (u,v) tel que u et v ϵ V(G) (Figure 2.7). Pour un graphe orienté, le sens de la relation entre une arête (u,v) va seulement de u vers v. Dans un graphe non-orienté, le sens de la relation entre une arête (u,v) va de u vers v et de v vers u (Figure 2.7).

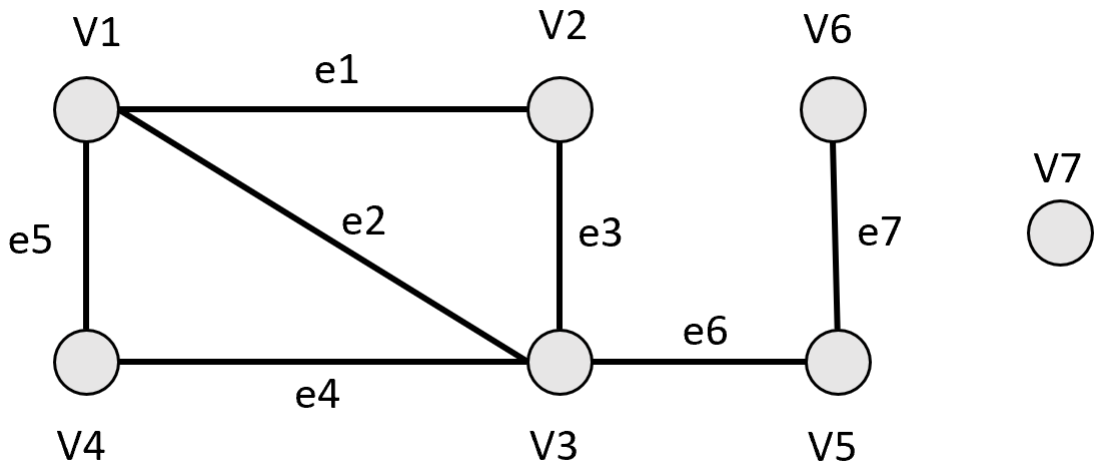


Figure 2.7 : Graphe non orienté avec sept sommets et sept arêtes. Les cercles représentent les sommets (V) et les traits la relation (e) entre les sommets. Deux sommets sont reliés par un trait s'ils ont une relation en commun.

Deux sommets u et v reliés par une arête sont appelés **adjacents** ou **voisins**. v est adjacent à u et u est adjacent à v. Dans la Figure 2.7, les adjacents de V1 sont V2, V3 et V4.

Le **degré** d'un sommet d(V) est le nombre de sommets reliés à ce sommet. C'est-à-dire son nombre de voisins. Dans la Figure 2.7, le degré de V1 est d(V1) = 3.

Un sommet **isolé** est un sommet de degré d(v)=0. Dans la Figure 2.7, V7 est un nœud isolé.

La **distance** d(u,v) entre deux sommets u et v est le nombre minimal d'arêtes à parcourir (le plus court chemin) pour aller de u a v. Dans la Figure 2.7, la distance d(V4,V2) est de 2, en effet pour aller de V4 à V2 le plus court chemin est de passer par e4 puis e3 (ou e5 puis e1).

Le **diamètre** d'un graphe est la distance entre le couple de sommets u et v dont la distance minimale est la plus grande. Le diamètre du graphe dans la Figure 2.7 est de 3. En effet les sommets les plus éloignés sont (V1,V6), (V4,V6) et (V2,V6) ayant tous comme distance 3.

Une **composante connexe** est un ensemble de sommets tel pour tout couple de sommets (u,v) ϵ C, il existe un chemin entre u et v. Dans la Figure 2.7, on a deux composantes connexes. La première composante qui contient les sommets (V1,V2,V3,V4,V5 et V6) et la deuxième composante contenant le sommet V7.

Une **clique** est un graphe G(V,E) tel que pour n'importe quel sommet V de G, il existe un lien (u,v) ϵ E. C'est-à-dire que tous les sommets V sont reliés entre eux. On parle également d'un graphe complet.

La centralité d'un sommet dans un réseau peut être interprétée comme étant la distance moyenne de ce sommet par rapport à tous les autres sommets appartenant à la même composante connexe de ce sommet. La centralité d'un sommet se calcule en prenant l'inverse de la somme des distances des chemins les plus courts entre ce sommet et tous les autres sommets de la même composante connexe que ce nœud.

2.4.2 Caractéristiques des réseaux biologiques

Comme décrit au début de la partie 2.4, les réseaux biologiques possèdent des propriétés topologiques qui les distinguent d'un réseau aléatoire.

La distribution des degrés suit une loi de puissance

Une loi de puissance est une relation mathématique entre deux quantités. Si une quantité est la valeur d'un évènement (axe des abscisses sur la Figure 2.8 a) et l'autre la fréquence des valeurs de cet évènement (axe des ordonnés sur la Figure 2.8 a), alors la relation est une distribution de la loi de puissance si les fréquences diminuent très lentement lorsque la valeur augmente (Figure 2.8 c). Une loi de puissance entre deux quantités x et y peut être décrite par l'équation $y = ax^{-k}$ où a est une constante de proportionnalité et k est le degré de puissance.

Dans un réseau métabolique, dans le cas où les sommets du graphe sont les activités enzymatiques et que deux activités sont connectées s'ils partagent un composé en commun (graphe centré sur les réactions), une distribution des degrés suivant une loi de puissance signifie que la majorité des nœuds sont de faibles de degrés (encadré en bleu dans la Figure 2.8 c) et une petite minorité sont très connectés (encadré en rouge dans la Figure 2.8

c). C'est-à-dire que ces sommets très connectés partagent leurs composés avec beaucoup d'activités enzymatiques (Figure 2.8).

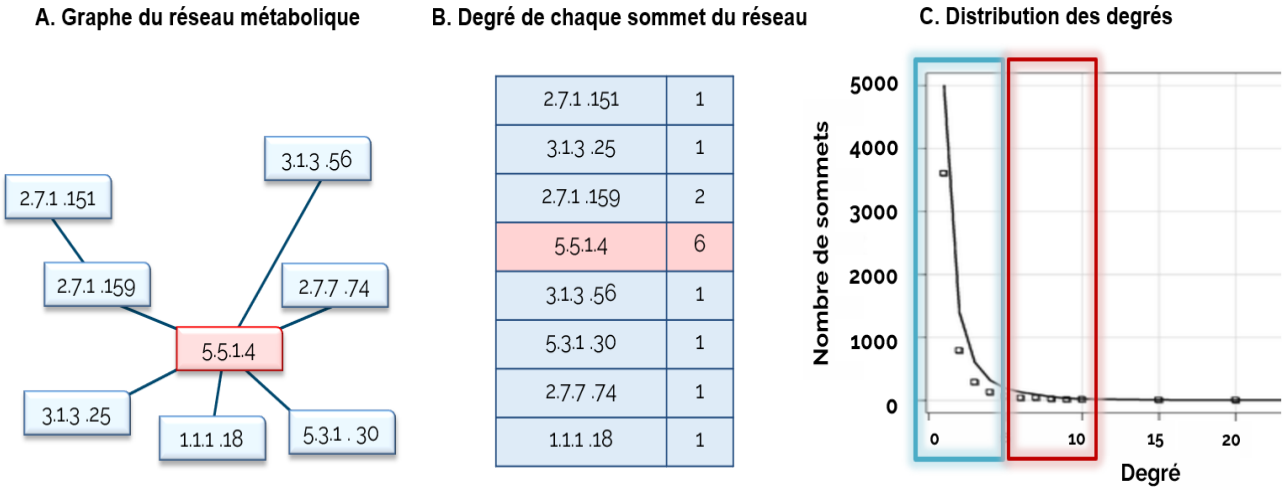


Figure 2.8 : Distribution des degrés suivant une loi de puissance. A. Un exemple de graphe du réseau métabolique. Le sommet en rouge indique un sommet très connecté. Les activités enzymatiques sont reliées si elles partagent un composé en commun dans leur réaction **B.** Degré de chaque sommet de la Figure a. **C.** Distribution des degrés d'un graphe biologique. L'axe des abscisses indique le degré et l'axe des ordonnées le nombre de sommets possédant un degré x. Figure adaptée de (Lima-Mendez and Helden, 2009)

Invariant d'échelle (scale-freeness)

Un réseau invariant d'échelle signifie que si on extrait un sous-réseau à partir d'un réseau initial, le sous-réseau gardera les propriétés du réseau complet. C'est-à-dire que les propriétés du graphe complet restent inchangées dans le sous-réseau mais l'échelle change. Dans le cas de la loi de puissance, il a été testé sur un réseau artificiel dont la distribution des degrés suit une loi de puissance que cette propriété est conservée dans les sous réseaux. Donc la propriété de loi de puissance est conservée d'une échelle à une autre (Stumpf *et al.*, 2005).

Dans la littérature, il a toujours été associé qu'un réseau dont le degré suit une loi de puissance est un réseau invariant d'échelle et vice-versa (Fox Keller, 2005; Arita, 2005). Or, ces deux propriétés ne sont pas bijectives car toute autre propriété du graphe peut être invariante d'échelle ou non. C'est la propriété de loi de puissance qui est invariante d'échelle, c'est une propriété conservée dans les sous-réseaux. Dire qu'un réseau est invariant d'échelle est incorrect à moins que toutes les propriétés du réseau le soient. Être invariant d'échelle s'applique à une propriété du réseau mais pas au réseau.

Petit monde (Small World)

Plusieurs auteurs (Fell and Wagner, 2000; Jeong *et al.*, 2000) ont démontré que le réseau métabolique est de type « small world ». Dans ce type de réseau, la plupart des sommets d'un réseau peuvent être atteints à partir de tous les autres sommets en un nombre relativement faible d'arêtes malgré la grande taille et la complexité du réseau. Plus particulièrement dans le cas du réseau métabolique, ils ont montré que la distance moyenne entre les composés dans le réseau est de 3, suggérant que les métabolites peuvent être interconvertis entre eux en seulement quelques étapes. Cette propriété semble suggérer l'efficacité et la robustesse du réseau pour transformer les métabolites efficacement malgré la complexité du réseau.

Vérifications de ces propriétés

L'analyse topologique du réseau métabolique a permis de mettre en lumière les caractéristiques des réseaux biologiques et du réseau métabolique. Des caractéristiques telles que la loi de puissance ont été observées pour tous les réseaux biologiques et ont été considérées comme une loi universelle régissant les réseaux biologiques.
Lima-Mendez et Helden (Lima-Mendez and Helden, 2009) ont soumis ces propriétés topologiques des réseaux à des tests statistiques pour discuter de leur validité ainsi que l'implication fonctionnelle et évolutive de ces propriétés.

Pour la loi de puissance, Lima-Mendez et Helden (Lima-Mendez and Helden, 2009) dans leur revue indiquent que la plupart des articles dans la littérature ne se basaient que sur un aspect visuel de la distribution pour confirmer si une distribution des degrés suit ou non une loi de puissance. Khanin et Wit (Khanin and Wit, 2006) ont montré par des tests statistiques sur 10 réseaux biologiques que la distribution des degrés ne suivait pas une loi de puissance.
Des analyses sur des réseaux de régulations chez *E.Coli* et des réseaux d'interactions protéiques ont montré que la distribution des degrés suivait une distribution asymétrique en forme de cloche. Le réseau d'interaction protéique obtenu à partir d'expériences à haut débit présente une forme de courbe qui peut difficilement être confondue avec la ligne droite attendue d'une loi de puissance (Lima-Mendez and Helden, 2009). Ces analyses plus poussées ont démontré que la loi de puissance n'est pas une loi universelle qui régit les réseaux biologiques. Cependant, pour les réseaux métaboliques, Noda-Garcia et ses collaborateurs (Noda-Garcia *et al.*, 2018) ont montré que la distribution des degrés suit effectivement une loi de puissance.

L'analyse approfondie des chemins pour interconvertir les métabolites dans un réseau métabolique « small-world » a montré que beaucoup des chemins proposés sont biologiquement et chimiquement impossibles (Lima-Mendez and Helden, 2009). La majorité des chemins contiennent des métabolites ubiquitaires à la plupart des réactions. La propriété « small world » du réseau métabolique est principalement due à la présence de

ces métabolites communs tel que H2O, O2, H+….. Ces métabolites étant présents dans la plupart des réactions, il en résulte que des réactions non adjacentes qui contiennent ces composés vont être connectées et créer des raccourcis erronés (Fell and Wagner, 2000). Ces métabolites communs vont créer des hubs artificiels qui relient beaucoup de réactions et confèrent artificiellement une propriété de « small-world » aux réseaux métaboliques. Arita a remis aussi en cause la notion de petit monde du métabolisme en analysant le réseau métabolique d'*E.coli* (Arita, 2004).

Modèle de construction d'un réseau biologique.

Ravasz et Barabási (Ravasz and Barabási, 2003) ont proposé un modèle très simple permettant de construire un réseau dont la distribution des degrés suit une loi de puissance. Dans leur algorithme, les sommets et les liens sont ajoutés progressivement. Les nouveaux sommets ont plus de probabilités de s'attacher sur les sommets déjà très connectés (modèle « Rich gets richer »).
Cet attachement préférentiel résulte surtout de la promiscuité des enzymes, une nouvelle enzyme émerge à partir de la spécification d'une enzyme capable de reconnaître d'autres substrats alternatifs et la catalyse des réactions alternatives. Les métabolites les plus utilisés sont reconnus par le plus d'enzymes, ainsi la probabilité qu'une nouvelle enzyme (ainsi que de nouvelles voies) émerge à partir de ces enzymes qui reconnaissent les métabolites les plus utilisés est plus probable.
En analysant le réseau métabolique, cet attachement préférentiel n'est pas suffisant pour expliquer l'évolution du réseau métabolique. En effet, ce modèle implique que les sommets les plus connectés sont les plus anciens. Ce scénario est raisonnable sur des composés très connectés tels que le pyruvate, le glutamate et quelques acides aminés mais est totalement incohérent par exemple avec l'ATP car l'ATP est largement plus connecté que l'adénosine (Eisenberg and Levanon, 2003), donc l'ATP serait plus ancien que l'adénosine ce qui est totalement aberrant. Ce modèle n'est pas non plus valable dans le cas des voies métaboliques linéaires. Dans les voies métaboliques linéaires, les enzymes ou les métabolites ont tous les mêmes connectivités (degré) sauf ceux qui se trouvent aux extrémités. Le modèle « rich gets richer » ne permet pas d'établir un ordre d'apparition des éléments dans une voie métabolique linéaire. Ce modèle semble donc rendre compte de façon imparfaite de la façon dont évoluent les réseaux biologiques.

Modularité et évolution du réseau métabolique

Les caractéristiques topologiques du réseau métabolique semblent indiquer une organisation modulaire du réseau métabolique (Hartwell *et al.*, 1999; Ravasz *et al.*, 2002). Un module se compose d'un ensemble d'éléments (par exemple réactions ou activités enzymatiques) qui forment un sous-système avec une fonction bien distincte (Figure 2.9). Les modules sont considérés comme des éléments fondamentaux de l'organisation des réseaux biologiques (Hartwell *et al.*, 1999). Diverses méthodes ont été utilisées et

développées pour identifier des modules dans divers systèmes biologiques. Dans un réseau d'interaction protéique, un module peut être défini comme un ensemble de protéines qui interagit formant un complexe protéique (Dezső *et al.*, 2003; Pereira-Leal and Teichmann, 2005). Dans un réseau de gènes, un module peut se traduire comme un ensemble de gènes régulés ensemble par le même mécanisme (Segal *et al.*, 2003).

Concernant le métabolisme, des données de profil phylogénétique ont été utilisées pour détecter des modules évolutifs (Yamada *et al.*, 2006; Zhao *et al.*, 2007). C'est-à-dire un ensemble d'éléments qui ont co-évolué ensemble. D'autres auteurs ont mis en évidence la nature modulaire de l'évolution du réseau métabolique (Kim *et al.*, 2006; Rives and Galitski, 2003).
Enfin, d'autres études ont montré le rôle de l'environnement et des relations évolutives dans l'évolution modulaire du réseau métabolique (Kreimer *et al.*, 2008; Zhao *et al.*, 2007). Plus particulièrement, ils ont montré une perte de modularité du réseau lors de la spécialisation d'une espèce à un environnement spécifique.
Ensemble, toutes ces études montrent l'importance de la modularité du réseau métabolique dans son évolution.

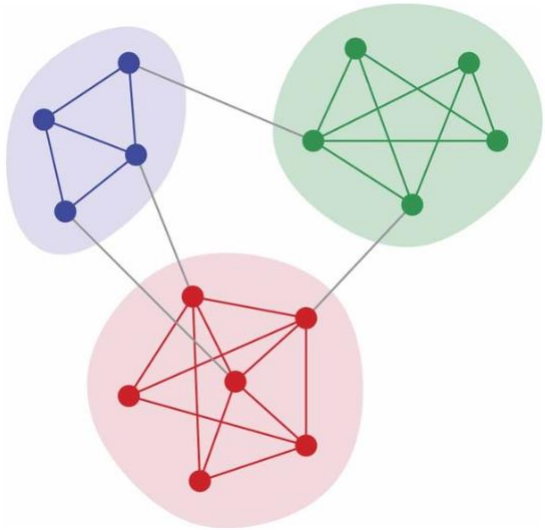


Figure 2.9 : Un réseau modulaire avec 3 modules. Chaque couleur indique une module.

Chapitre :

3 Les champignons

Les champignons sont connus depuis longtemps pour leur intérêt gastronomique et/ou industriel (consommation, fabrication du fromage, fermentation du vin…) mais aussi pour les dommages qu'ils causent à l'homme, aux animaux et aux plantes. Les champignons jouent également un rôle important dans la biosphère et dans notre vie de tous les jours (Chaudhary et al., 2022). Ils ont colonisé une grande diversité de niches écologiques. Les champignons peuvent être saprophytes, pathogènes, endophytes, symbiotiques, commensales et lichénisés sur un large éventail d'hôtes et d'environnement (Naranjo-Ortiz and Gabaldón, 2019).

De nombreux champignons vivent dans le sol mais aussi dans/sur les plantes sans leur causer de dommages apparents, en leur conférant même parfois des avantages (mutualistes). Par exemple, dans la symbiose mycorhizienne, le champignon fournit des nutriments à la plante et en retour ce dernier lui fournit des matières organiques (Huey *et al.*, 2020).

Les champignons sont les principaux décomposeurs dans l'environnement. Ils peuvent coloniser et décomposer les organismes morts. Le processus de décomposition est essentiel pour le cycle naturel des éléments chimiques, en particulier dans le cycle global du carbone où ils contribuent au renouvellement des approvisionnements en carbone dans l'environnement (Emilia Hannula and Morriën, 2022).
Les champignons présentent un large éventail de processus métaboliques et de biotransformations, y compris la sécrétion d'enzymes extracellulaires ciblant la décomposition de restes indigestes d'autres entités biologiques composés de la lignocellulose et de la chitine, ce qui les rendent centraux dans le recyclage du carbone (Hartl *et al.*, 2012).

Les animaux ne peuvent pas digérer la lignine ni la cellulose pourtant il existe des animaux herbivores. Pour digérer ces nutriments indigestes, l'estomac d'un ruminant contient un véritable écosystème composé de bactéries, protozoaires et de champignons capables de digérer la cellulose (Huws *et al.*, 2018).

Les champignons sont aussi d'excellents chimistes. Ils sont capables de synthétiser une myriade de substances secondaires qui ne servent pas directement à la croissance ou au maintien de la structure cellulaire et peuvent transformer une grande variété de substrats avec leur métabolisme accessoire. On retrouve des molécules d'intérêts thérapeutiques tels que des antibiotiques (Dutta *et al.*, 2022) comme la pénicilline, des anticancéreux (How *et al.*, 2022), et des immunosuppresseurs comme la ciclosporine (Wong *et al.*, 2017).
Les champignons produisent aussi des enzymes ayant un intérêt industriel comme la cellulase et la lipase qui ont des applications dans l'industrie textile, la fabrication de cosmétiques, la production de papier, la production de détergent….(Ejaz *et al.*, 2021; Kumar *et al.*, 2023).

3.1 Taxonomies et diversités

Les champignons forment un groupe hétérogène d'organismes qui partagent des caractéristiques communes.

Intuitivement, la première image qui nous vient à l'esprit quand on parle de champignon est la forme du champignon de Paris ou de l'Amanite tue-mouches (Figure 3.1). Cette image que nous avons des champignons n'est autre que l'appareil reproducteur utilisé pour la dispersion des spores appelé carpophore ou sporophore par les mycologues. Les champignons qui possèdent cette forme sont appelés agaricoïdes.

Figure 3.1 : A gauche le champignon de Paris (*Agaricus bisporus*). **A droite** l'Amanite tue-mouches (*Amanita Muscaria*).

Le sporophore est seulement visible une partie de l'année. La grande partie du champignon est invisible car cachée sous terre. La partie invisible est appelée **mycélium** (Figure 3.2). Les cellules des champignons s'alignent et forment des hyphes (Figure 3.3). La ramification des hyphes forme le mycélium. C'est la partie végétative qui permet au champignon d'absorber les nutriments.
Cette forme de champignons n'est qu'une partie infime de la diversité des champignons.

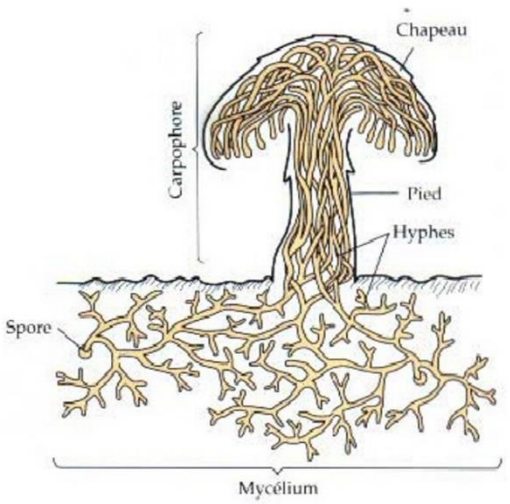


Figure 3.2 : Structure d'un champignon « agaricoïde ». Les agaricoïdes sont les champignons qui développent un sporophore avec un chapeau. La ramification des hyphes forme le mycélium qui est en grande partie caché sous terre.

Il existe aussi des formes unicellulaires appelées levures (par exemple *Saccharomyces cerevisiae*). Bien que les formes précédentes soient les plus répandues, Il existe aussi d'autres formes comme les rhizoïdes, les plasmodes ou les protoplastes.

Si le sporophore est le mode d'appareil reproducteur chez les agaricoïdes, chez les levures la reproduction se matérialise par un simple renforcement de la paroi cellulaire.
Le cycle de vie d'un champignon se résume en une succession d'étapes qui sont : germination de la spore pour former le mycélium, la croissance et l'agrégation du mycélium pour développer la structure de reproduction puis la sporulation et la dispersion des spores.

La cellule chez les champignons présente une caractéristique unique. Si les cellules animales et végétales se divisent, chez les champignons la croissance des hyphes est polarisée. C'est-à-dire que la croissance d'hyphe ne se fait que par son apex et se cloisonne. Une partie délimitée par deux cloisons est appelée article, qui structurellement ressemble à une cellule (Figure 3.3).
Le mode de nutrition des champignons est l'osmotrophie, qui consiste en l'absorption des nutriments depuis leur milieu par des transporteurs. Ils sont incapables de phagotrophie (phagocyter des proies) contrairement aux cellules animales. Une grande majorité des champignons est saprophyte**.**
Le saprophytisme est une variante de l'osmotrophie par laquelle des enzymes sont sécrétées dans le milieu pour digérer des matières solides externes qui ne peuvent pas être transportées par les transporteurs directement.

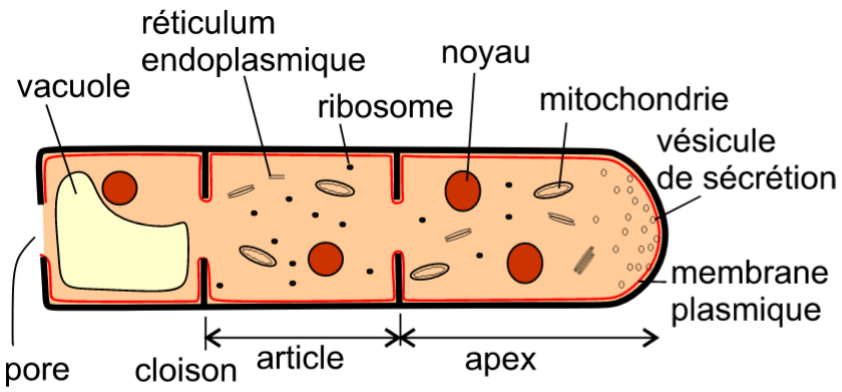


Figure 3.3 : Structure d'un hyphe. L'hyphe est un appareil végétatif filamenteux caractéristique des champignons. Deux cloisons délimitent une structure qui ressemble à une cellule appelée article. Les noyaux sont dispersés dans le cytoplasme avec d'autres organites cellulaires tels que le réticulum endoplasmique, la mitochondrie et les ribosomes. L'hyphe ne s'allonge que par son sommet appelé apex. Chez certaines espèces, ces cloisons sont trouées par des pores. Figure tirée de « les champignons redécouverts » (Fabienne Malagnac et Philippe Silar 2013).

Pour remonter à l'origine des champignons et leur divergence des animaux et des plantes, il faut remonter à l'origine des cellules eucaryotes et procaryotes.

La cellule eucaryote (du Grec « eu- » qui veut dire « bien » et « caryon », noyau) se distingue de la cellule procaryote (ex : bactérie) par la présence d'un noyau qui contient la majorité de son matériel génétique. Beaucoup admettent aujourd'hui que la cellule eucaryote a acquis la capacité de phagocyter d'autres cellules (la présence d'organelles comme la mitochondrie en est la preuve), étape décisive dans son évolution (Vosseberg *et al.*, 2021; Gabaldón, 2021). La phagotrophie était le mode de nutrition primitif des cellules eucaryotes.
La phagocytose d'une cyanobactérie qui n'a pas été digérée et qui est restée dans la cellule puis au cours de l'évolution s'est transformé en plaste a donné la possibilité à certaine cellule de devenir autotrophe en fournissant des nutriments à son hôte grâce à la photosynthèse. Cette phagocytose serait à l'origine des plantes (Margulis, 1971).
Certaines cellules eucaryotes ont perdu la phagotrophie pour revenir à un mode de vie osmotrophe surtout quand la nourriture est trop grosse pour être phagocytée (comme la cellulose), ces cellules ont acquis la capacité de diffuser des enzymes dans le milieu pour digérer ces nourritures indigestes. C'est la saprotrophie et c'est le mode de nutrition de la majorité des champignons (Naranjo-Ortiz and Gabaldón, 2019).
Les champignons sont actuellement des saprotrophes ou sont des parasites des plantes ou des animaux ayant perdu la capacité de phagocyter de la cellule animale.

Les pseudomycètes (pseudo-champignon) sont aussi osmotrophes et différencient des hyphes mais leur paroi est composée de cellulose et non de chitine (principal composant de

la paroi cellulaire des champignons). Leur spore contient deux flagelles (bicontes) contrairement aux champignons (unicontes comme les cellules animales) (Davison, 1998).
De ce fait, les champignons (eumycetes) sont des proches parents des animaux.

Les vrais champignons (eumycètes) sont divisés en deux groupes : les **champignons inférieurs** et les **champignons supérieurs** (Figure 3.4).

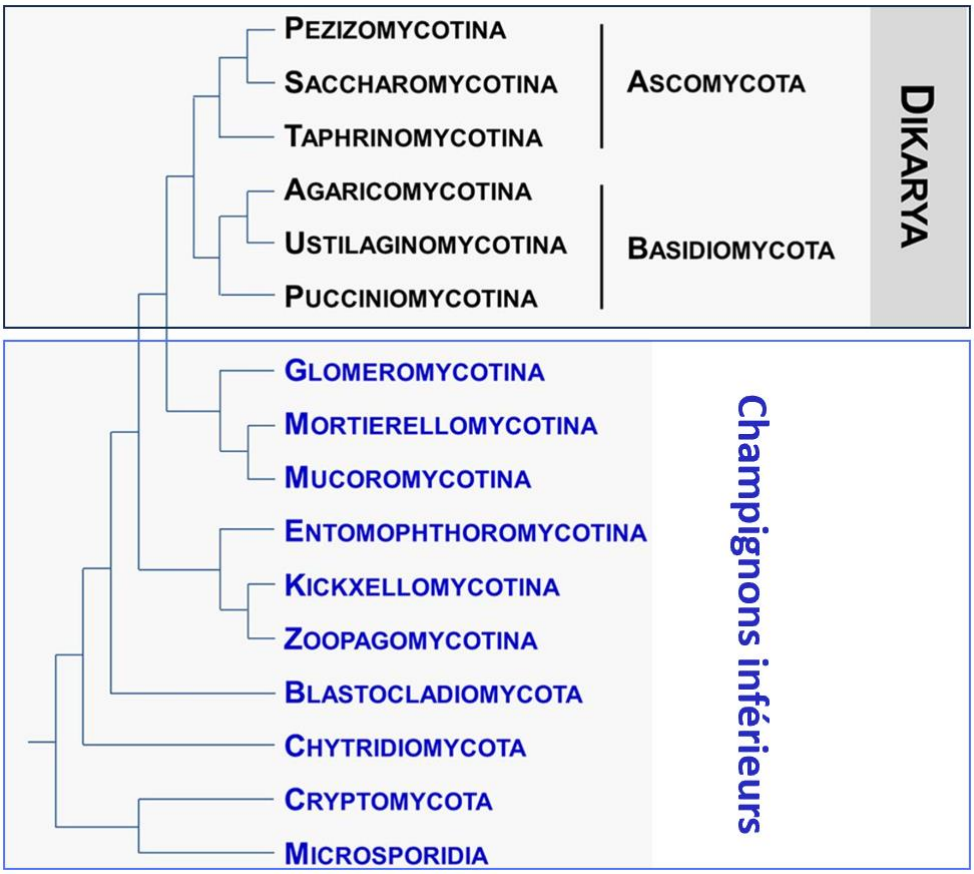


Figure 3.4 : Arbre phylogénétique des principaux groupes de champignons (Eumycètes). En bleu les champignons inférieurs et en noir les *Dikarya* (Champignons supérieurs). Figure adaptées de (Hérivaux *et al.*, 2017).

Les champignons inférieurs

Les champignons inférieurs ne comptent que quelques milliers d'espèces. Ils forment un groupe paraphylétique (n'incluant pas tous les organismes appartenant aux eumycètes) constitué des organismes qui n'appartiennent pas au groupe taxonomique des Dikarya (Figure 3.4 A).
Ce groupe informel se caractérise par des organismes incapables de développer des structures de dispersion pluricellulaires même si quelques espèces forment des truffes

simples. Ils sont majoritairement aquatiques et nécessitent beaucoup d'humidité (Berbee *et al.*, 2017).

Les *Chytridiomycota, Neocallimastigomycota* et *Blastocladiomycota* produisent des spores flagellées et mobiles (zoospore) et sont majoritairement saprophytes des eaux douces et du sol.

Les microsporidies sont des parasites intracellulaires obligatoires. Ce sont des organismes avec un métabolisme très réduit (plus de détails dans la partie II.7.3).

Les *Entomophthoromycotina, Zoopagomycotina et Kickxellomycotina* sont essentiellement des parasites et sont très rarement saprophytes.
Les *Mucoromycotina* sont pour la plupart des saprophytes présents dans le sol. On les retrouve souvent impliqués dans la dégradation des denrées alimentaires. Ce sont les seuls champignons inférieurs capables de différenciation et qui ressemblent à des truffes.

Les *Glomeromycota* sont des champignons en symbioses avec la racine des plantes. Ils possèdent des hyphes capables de fusionner, une caractéristique seulement retrouvée chez les champignons supérieurs.

Les champignons supérieurs

Les champignons supérieurs (Dikarya) dérivent d'un même ancêtre commun ayant acquis des cellules à deux noyaux. En plus de cette capacité à avoir deux noyaux, les Dikarya peuvent différencier des hyphes à cloisons simples et ont la possibilité de fusionner deux cellules.
En revanche, la production de sporophore pluricellulaire semble être acquise par convergence évolutive dans les deux sous-embranchements une fois chez les basidiomycètes et deux fois chez les ascomycètes.
La différence biologique entre les deux groupes se situe au niveau des spores (Figure 3.5). Les spores des ascomycètes sont souvent symétriques alors que les spores des basidiomycètes sont plus ou moins asymétriques.

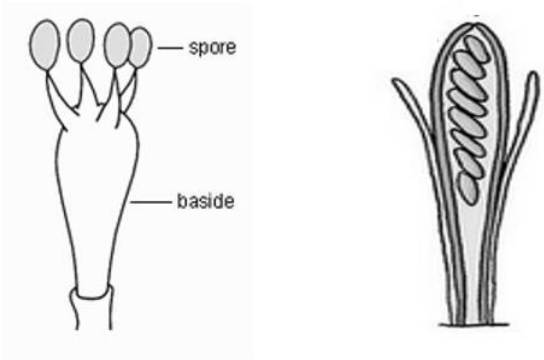


Figure 3.5 : A gauche : Baside typique des basidiomycètes. **À droite** : Asque typique des ascomycètes.

Chez les **basidiomycètes**, il y a 3 sous-embranchements.
Le groupe *Pucciniomycotina* contient des champignons avec des modes de vies très variés : saprophytes, pathogènes des plantes, parasites… Parmi les parasites redoutables il y a les rouilles qui affectent le blé. En plus, elles présentent des cycles biologiques parmi les plus complexes du monde vivant.
Les Ustilaginomucotina contiennent principalement des pathogènes des plantes comme *Ustilago maydis.*
Les Agaromycotina contiennent 2/3 des organismes des basidiomycètes et contiennent des espèces de biologies diverses (saprophytes, parasites, levures, mutualistes…). Ils contiennent la majorité des espèces de gros champignons : agarics (champignons de paris), bolets, girolles, vesse-de-loup …

Les **ascomycètes** constituent à elles seules 60% des champignons dont la moitié vit avec des algues et sont donc des lichens.
Les Taphrinomycotina sont constitués d'espèces à la biologie diverse : parasites, levures comme *S.Pombe*, symbiotes.
Le groupe *des Saccharomycotina* ne contient qu'une seule classe les *Saccharomycetes*. Les quelques centaines d'espèces connues passent une grande partie de leur vie sous forme de levures mais quelques-unes peuvent différencier des hyphes. Ce sont des organismes qui ont envahi de nombreuses niches écologiques.
Le groupe des *Pezizomycotina* regroupe des espèces aux biologies diverses comme chez les *Agaromycotina.* Les convergences évolutives sont nombreuses chez ce groupe. Ce groupe contient les gros champignons comestibles : morilles et truffes.

La description des différents groupes de champignons montre une très grande diversité au niveau des modes de vie des champignons. Les champignons ont colonisé une très grande variété de niches écologiques. Les champignons peuvent être saprophytes, pathogènes, endophytes, symbiotiques, commensales et lichénisés sur une large gamme d'hôtes, tandis que l'interaction avec l'hôte peut varier.

3.2 Projet de séquençage des champignons

S.cerevisiae a été le premier génome eucaryote entièrement séquencé (Goffeau *et al.*, 1996). Les informations génétiques acquises avec d'autres champignons ont permis de comparer les séquences entre elles (Krantz *et al.*, 2006). La génomique comparée a permis de reconstruire l'histoire évolutive (phylogénies moléculaires) (Taylor, 1995) des champignons, de déterminer de nouveaux gènes et des gènes spécifiques qui ont permis de comprendre une partie des raisons de la diversité des champignons. Le succès des méthodes de génomique comparative et la disponibilité d'un grand nombre d'espèces ont conduit plusieurs centres de recherche à se focaliser sur le séquençage et l'annotation des génomes de champignons.

Parmi ceux-ci, le Fungal Genome Initiative du Fungal Genomics Group du Broad Institute a été initié en 2000 pour accélérer le séquençage de génome des champignons ayant un intérêt particulier pour la médecine (pathogène de l'homme), l'industrie (champignons capables de synthétiser des molécules d'intérêts thérapeutiques) et l'agriculture (pathogène des plantes). Actuellement, plus de 200 génomes sont accessibles dans leur base de données (https://www.broadinstitute.org/fungal-genome-initiative/history-fungal-genome-initiative) .

Le projet « 1000 fungal genomes project » est une initiative de plusieurs chercheurs internationaux en collaboration avec le Joint Genome Institute of the Department of Energy sur un projet de 5 ans afin de séquencer 1000 génomes de champignons qui couvrent l'ensemble de l'arbre phylogénétique des champignons. L'objectif est de comprendre la diversité des champignons et de séquencer au moins 2 génomes par famille pour chacune des 500 familles de champignons connues. L'ensemble des génomes séquencés est déposé dans une base de données accessible à partir du portail Mycocosm (Grigoriev *et al.*, 2014).
Initialement prévu pour 1000 génomes, le nombre de génomes disponible sur le portail Mycocosm est actuellement de 2427 (dernier accès 08/2023).
Ce portail met aussi à disposition des outils de visualisations et des outils de génomique comparative pour l'analyse des génomes. Aujourd'hui, la base de données répertorie également des séquences issues d'autres projets comme ceux du Fungal genomes Initiative.

Chapitre :

4 Objectifs de la thèse

La vie est intrinsèquement instable, et le métabolisme cellulaire agit comme son avant-garde, s'adaptant continuellement pour maintenir l'équilibre et assurer la survie de l'organisme. La capacité des organismes à s'adapter aux conditions changeantes de leur habitat est essentielle pour garantir leur survie et leur reproduction. L'adaptation aux variations de l'environnement stimule l'innovation, l'échange et la disparition des activités enzymatiques. Au niveau métabolique, ce processus d'adaptation se manifeste par la capacité des enzymes à faire évoluer des fonctions bénéfiques et à se délester des fonctions inutiles dans un environnement aux conditions changeantes.
Cependant une activité enzymatique ne confère pas un avantage physiologique à elle seule. Chaque activité enzymatique est le maillon d'une chaîne qui assure la formation ou la dégradation de métabolites. Elle fait partie d'une voie métabolique ou un réseau plus large pour travailler de manière coordonnée avec d'autres activités enzymatiques.

L'objectif de cette thèse est de comprendre comment ce réseau métabolique a évolué chez les champignons.
Pour répondre à cette question, deux grandes approches ont été principalement utilisées.
Une première approche repose sur l'utilisation des méthodes de génomique comparée. En effet, avec le grand nombre de données de séquençage à disposition aujourd'hui, il est devenu possible de comparer les génomes d'un grand nombre d'espèces. La comparaison des profils phylogénétiques enzymatiques entre les différentes espèces permettra ainsi de comprendre quelle est la partie du métabolisme commune à toutes ces espèces. Cette comparaison permettra également d'identifier les activités enzymatiques qui sont propres à un groupe taxonomique ou à des espèces non reliées taxonomiquement mais qui partagent une propriété biologique commune.
Une seconde approche repose sur l'analyse du réseau métabolique complet. En effet, l'identification de l'ensemble des activités enzymatiques permet de modéliser le réseau métabolique global d'une espèce. Avec la théorie des graphes, il est désormais possible d'analyser la structure du réseau métabolique et de comparer le réseau métabolique de plusieurs espèces.

L'approche par génomique comparée a comme principal inconvénient une analyse des éléments individuellement sans tenir compte des relations dans le réseau métabolique. Dans une approche par analyse de graphe, nous n'avons pas l'information évolutive des activités enzymatiques. En intégrant ces deux approches, il nous sera possible de résoudre la problématique suivante qui sera au cœur de ce travail :

Est-ce que le réseau métabolique exerce une pression sur l'évolution des activités enzymatiques chez les champignons ?

Yamada et ses collaborateurs (Yamada *et al.*, 2006) ont initié une intégration des deux approches en montrant une corrélation entre les profils phylogénétiques similaires et leur proximité dans le réseau. D'autres auteurs (Vitkup *et al.*, 2006) ont montré sur un réseau de

gènes, que les gènes qui évoluent plus lentement sont les plus connectés dans le réseau métabolique. À l'aide des données de génomiques, plusieurs auteurs (Yamada *et al.*, 2006; Zhao *et al.*, 2007; Peregrín-Alvarez *et al.*, 2009) ont identifié des modules évolutifs dans le réseau métabolique, c'est-à-dire un ensemble d'éléments qui co-évoluent. Les réseaux sont également utilisés dans les études comparatives pour mettre en évidence les différences et les similitudes existant dans l'organisation des mécanismes intracellulaires de plusieurs espèces. Des études qui comparent les caractéristiques topologiques des réseaux métaboliques pour les taxons échantillonnés des trois règnes du vivant ont été publiées (Jeong *et al.*, 2000; Zhu and Qin, 2005; Peregrín-Alvarez *et al.*, 2009). Ces recherches mettent en lumière comment ces réseaux métaboliques diffèrent entre les archées, les bactéries et les eucaryotes à la suite de processus de sélection naturel.
Plus récemment, Montanucci et ses collaborateurs (Montanucci *et al.*, 2018) ont montré que les gènes les plus conservés sont situés dans les dernières étapes de transformations dans les voies métaboliques et que les gènes qui initient les voies métaboliques sont ouverts aux changements évolutifs.
Ces études montrent une relation entre la structure du réseau métabolique et l'évolution des éléments le constituant.

Dans ce projet de thèse nous avons intégré ces deux grandes approches ensembles pour comprendre les contraintes imposées par le réseau métabolique sur l'évolution des activités enzymatiques chez les champignons. Les champignons sont d'excellents modèles pour explorer l'évolution du réseau métabolique du fait de leur diversité exceptionnelle.
Plus précisément nous avons travaillé avec 174 génomes de champignons, dont l'arbre phylogénétique a été construit et qui va permettre de faire une comparaison à grande échelle mais aussi de retracer l'histoire évolutive de chaque activité enzymatique (détection des gains et des pertes).
Les informations évolutives ont été cartographiées sur le réseau métabolique et les voies métaboliques. Cette approche intégrative vise à détecter le module commun à toutes les espèces et les modules spécifiques (taxonomique ou environnemental), et les différentes propriétés topologiques des activités enzymatiques en fonction de leur information évolutive pour comprendre les contraintes du réseau métabolique sur l'évolution des activités enzymatiques.

II. Conservation et évolutions des activités enzymatiques chez les champignons

Chapitre :

5 Profils phylogénétiques des activités enzymatiques

Dans le cadre de cette thèse, nous avons travaillé avec 174 espèces de champignons qui couvrent la totalité de l'arbre des champignons (Figure 5.1) dont le génome a été complètement séquencé et annoté. Ces données ont été téléchargées à partir de plusieurs bases de données, majoritairement issues du JGI et du Broad Institute .

Ces 174 espèces de champignons correspondent à l'ensemble des champignons dont le génome a été complètement séquencé au 1er janvier 2011. En effet, mes travaux de thèse se situent dans la continuité des travaux effectués par Cécile Pereira (Pereira, 2015).
Une grande partie de ses travaux était dédiée à l'annotation fonctionnelle automatique et homogène des protéomes de champignons afin de pouvoir comparer le métabolisme de différentes espèces. Pour pouvoir les comparer, il fallait identifier les éléments métaboliques (en l'occurrence les enzymes) de chaque espèce.
Les protéomes complets sur lesquels nous avons travaillé proviennent de différentes bases de données, séquencées et annotées par différents groupes. Par conséquent, les 174 protéomes de champignons n'ont pas été annotés de la même manière ni par les mêmes méthodes.
Afin de pouvoir comparer le métabolisme des champignons d'un point de vue enzymatique, les protéomes ont donc été réannotés fonctionnellement de manière standardisée et homogène par Cécile Pereira (Pereira, 2015).

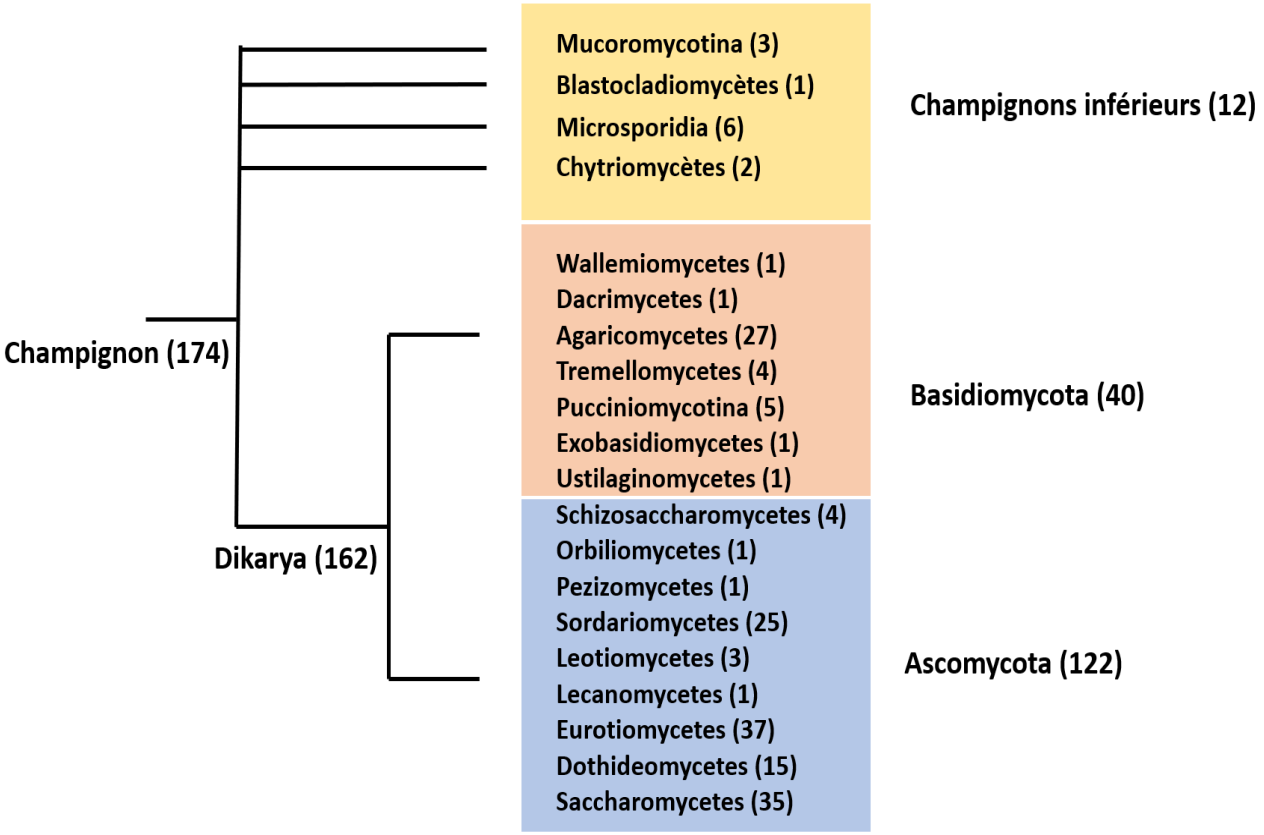


Figure 5.1 : Répartition taxonomique des espèces étudiées. Pour chaque groupe taxonomique le nombre d'espèces appartenant à ce groupe est indiqué à droite.

Dans ce court chapitre, je détaillerai brièvement la méthode utilisée pour annoter ces protéomes car cela s'avère nécessaire pour une bonne compréhension des choix des approches et des méthodes utilisées durant mes travaux de thèses.

Pour chaque espèce, les séquences protéiques ont été annotées par transfert d'annotation entre séquences homologues. Deux séquences orthologues auront donc la même annotation. Diverses bases de données répertorient des orthologues prédits à l'aide de différentes méthodes, et il est bien établi que ces méthodes produisent des prédictions partiellement divergentes. Pour incorporer les prédictions actuelles tout en ajoutant des informations pertinentes, Cécile Pereira a développé la méta approche **MARIO** (Pereira, 2015)**.** Cette approche combine les intersections des résultats de plusieurs méthodes de détection de groupes d'orthologues (BRH, Inparanoid, orthoMCL et Phylogeny) et les enrichit en utilisant des profils HMM.

MARIO (Pereira *et al.*, 2014) est une version améliorée de FungiPath (Grossetête *et al.*, 2010) qui était la première méta-approche développée pour la détection d'orthologues au sein de l'équipe de Bioinformatique Moléculaire. MARIO est aussi basé sur l'utilisation des intersections des groupes d'orthologues inférés à partir des quatre méthodes énumérées précédemment. Le but de FungiPath/MARIO est surtout d'annoter de façon pertinente et homogène un ensemble de protéomes de champignons.
MARIO a permis d'identifier 68182 groupes d'orthologues dans 174 protéomes de champignons.

Une fois les groupes d'orthologues définis, chaque groupe d'orthologues a été annoté par transfert d'annotation (Figure 5.2).
Pour chaque groupe d'orthologues, les séquences protéiques ont été alignées avec MUSCLE (Edgar, 2004) puis un profil HMM a été construit avec HMMER (Finn *et al.*, 2011). Les séquences annotées qui possèdent une annotation de type EC-number dans la base de données SwissProt (Bairoch and Apweiler, 2000) et MetaCyc (Karp *et al.*, 2000) ont toutes été comparées avec les profils HMM de tous les groupes d'orthologues. S'il y a une similarité entre le profil et une séquence annotée, l'annotation de la séquence annotée est transférée au groupe d'orthologues.
Plus précisément, si la similarité entre les séquences annotées et la base de données de profils HMM présente une E-value inférieure à 10^{-80} , l'annotation de son EC-number est transférée au groupe d'orthologues.
Si la similarité est comprise entre 10^{-20} et 10^{-80}, l'annotation est transférée si une des séquences dans le groupe d'orthologue est présente dans SwissProt ou MetaCyc et porte déjà la même annotation. Les séquences sans orthologues (orphelines) sont annotées si la séquence est déjà annotée dans SwissProt ou MetaCyc . Sinon, le groupe n'est pas annoté.
Environ 4500 groupes d'orthologues ont été annotés avec 968 EC-numbers différents. Ce qui signifie que plusieurs groupes d'orthologues ont été annotées avec le même EC-number.

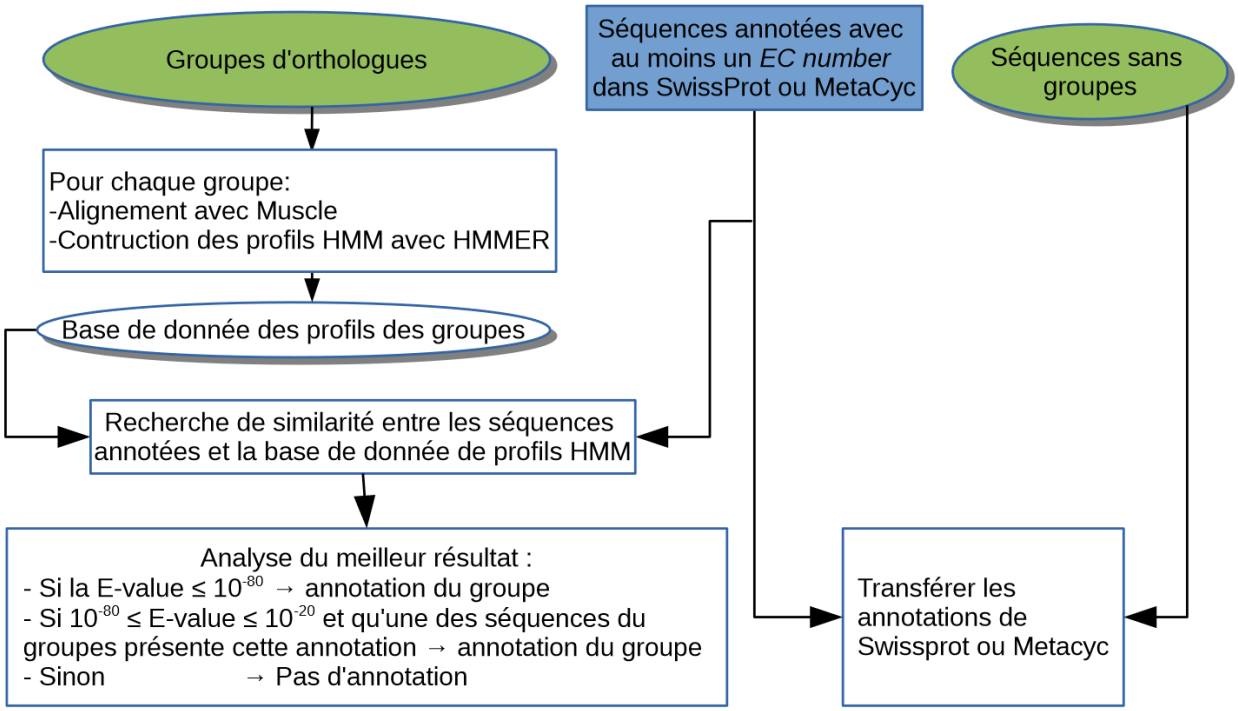


Figure 5.2 : Pipeline d'annotation des groupes d'orthologues. Les étapes en vert représentent les résultats obtenus avec la méta-approche MARIO et en bleu les données téléchargées sur les sites de SwissProt et MetaCyc. Figure tirée de (Pereira, 2015)

Chaque groupe d'orthologues constitue une liste de protéines assurant la même fonction enzymatique inférée par similarité de séquence avec une séquence déjà annotée.
L'annotation des groupes d'orthologues a permis la reconstruction du profil de chaque activité enzymatique. C'est-à-dire la présence ou l'absence de l'activité enzymatique à travers les 174 espèces de champignons étudiées.

Un groupe d'orthologues peut contenir des protéines qui sont spécifiques de quelques espèces. Ces espèces peuvent être taxonomiquement reliées ou non. Afin de déterminer les relations évolutives entre les espèces, nous avons construit l'arbre phylogénétique des 174 espèces de champignons.

Chapitre :

6 Construction de l’arbre phylogénétique des espèces étudiées

La phylogénie est l'étude des liens de parenté entre différents gènes, protéines ou organismes et permet de retracer leur histoire évolutive. Cette relation évolutive entre les espèces va permettre de retracer l'histoire évolutive du métabolisme mais aussi d'identifier des activités enzymatiques qui appartiennent à des espèces reliées taxonomiquement.

Cette relation est souvent représentée sous forme de diagramme en forme d'arbre ramifié connu sous le nom **d'arbre phylogénétique**. Cet arbre indique les relations évolutives entre les différentes espèces. Il indique les degrés de ressemblance entre les organismes en fonction des marqueurs utilisés (Figure 6.1).
Un arbre phylogénétique peut être construit en utilisant différents types de marqueurs. Avant les années 60, les données utilisées pour inférer la phylogénie proviennent de la comparaison de données morphologiques ou de caractéristiques métaboliques entre les espèces. Mais à partir des années 1960, les informations issues des données moléculaires (séquence protéique et nucléotidique) ont fourni des données précieuses aux études phylogénétiques en comparant les différences entre les différentes séquences (Woese and Fox, 1977).

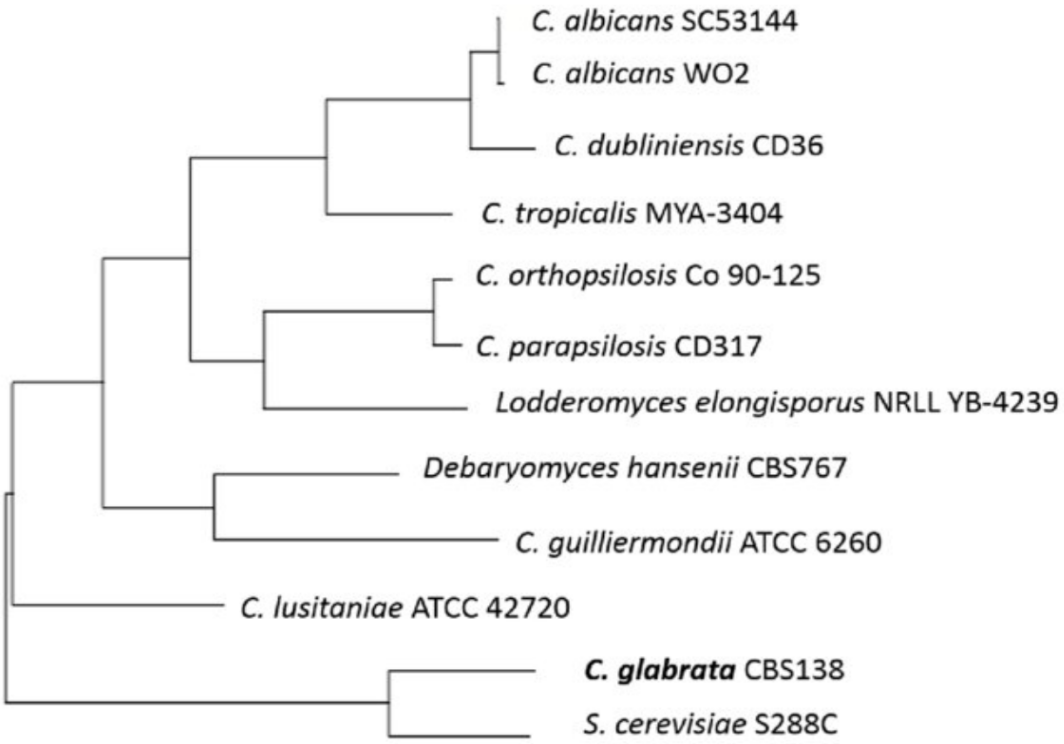


Figure 6.1 : Arbre phylogénétique de quelques levures. L'arbre a été construit à partir de la comparaison du gène ERG11 (une famille de cytochrome P450). Les branches entre les espèces indiquent la relation évolutive entre les espèces et la longueur des branches représente la distance évolutive entre les espèces. Figure tirée de (Tam *et al.*, 2015).

À partir des données moléculaires, il existe plusieurs méthodes d'inférence phylogénétique. La première étape consiste à aligner les séquences homologues entre elles. L'alignement de séquences permet d'identifier les similarités et les différences entre les séquences. Les régions similaires sont probablement conservées par les espèces étudiées et les différences sont des régions variables. L'arbre phylogénétique sera construit à partir de ces informations. Plusieurs programmes sont aujourd'hui disponibles pour faire des alignements multiples de séquence tels que Clustal Omega (Sievers and Higgins, 2018), MUSCLE (Edgar, 2004) et MAFFT (Katoh *et al.*, 2002). Une fois les séquences alignées, plusieurs méthodes d'inférence d'arbre phylogénétique peuvent être utilisées.

Parmi les méthodes de construction d'arbre phylogénétique, il existe des méthodes basées sur les distances telles que « neighbor joining (NJ)» (Saitou and Nei, 1987) et UPGMA (unweighted Pair Group Method with arithmetic mean) (Sokal R. and Michener C. 1958).
Il y a aussi les méthodes par parcimonie (Felsenstein, 1978) et les méthodes probabilistes par Maximum de vraisemblance (Neyman, 1971) ou par inférence bayésienne (Rannala and Yang, 1996).
Bien qu'elle soit relativement lente et coûteuse en calcul, le maximum de vraisemblance est la méthode phylogénétique la plus couramment utilisée car l'arbre inféré est souvent plus proche de la réalité et les comparaisons entre méthodes montrent qu'elle donne souvent de meilleurs résultats que les autres méthodes (Lees *et al.*, 2018).

Ces dernières années, d'important progrès ont été effectués pour améliorer la rapidité des méthodes basées sur le maximum de vraisemblance et des outils comme PhyML (Guindon *et al.*, 2010), IQTree (Nguyen *et al.*, 2015), RAxML (Stamatakis, 2014) et qui sont désormais capables de prendre en charge plus de 1000 séquences et d'inférer un arbre en un temps raisonnable.

6.1 Choix du marqueur phylogénétique

Les arbres phylogénétiques peuvent être construits en utilisant des séquences d'ADN ou de protéines. Une étape importante en utilisant ces données est de sélectionner le gène approprié pour inférer l'arbre. Ces gènes sont appelés des « marqueurs phylogénétiques ». Le choix de ces marqueurs dépend de la diversité et de la divergence des organismes et de la disponibilité des données (Choi *et al.*, 2019).

Chez les animaux, une région du gène mitochondrial codant pour une sous-unité du cytochrome C oxydases (CO1) est le marqueur le plus utilisé (Hebert *et al.*, 2003) et a été utilisée par défaut chez les champignons (Schindel and Miller, 2005). CO1 est un marqueur qui a été utilisé chez quelques genres de champignons tels que le *Penicillium* (Seifert *et al.*, 2007) mais le plus souvent peut construire des arbres inconsistant ou quasiment impossibles dus à l'absence de mitochondrie chez *Neocallimastigomycota* (Bullerwell and Lang, 2005). Chez les champignons il a donc été délaissé au profit d'un autre marqueur : « internal transcribed spacer » (ITS) (Dentinger *et al.*, 2011; Schoch *et al.*, 2012). L'ITS est une

région de l'ADN ribosomique situé entre les gènes de la petite et de la grande sous-unité du ribosome. La comparaison de l'ITS avec d'autres marqueurs pour inférer l'arbre chez les champignons a montré que l'arbre inféré est le plus souvent correct pour la plupart des groupes de champignons (Schoch *et al.*, 2012). Malgré cela, l'ITS peut être peu informatif ou même trompeur dans quelques groupes de champignons (Crouch *et al.*, 2009; Gazis *et al.*, 2011; Větrovský *et al.*, 2016).

Le choix du marqueur est une étape importante dans l'inférence phylogénétique. Une solution est de choisir un gène commun à tous, c'est-à-dire appartenant au core génome. Une difficulté réside cependant dans le fait que le taux d'accumulation des mutations de chaque gène au cours de l'évolution est différent. Il est ainsi possible d'obtenir deux arbres totalement différents en utilisant deux marqueurs différents pour les mêmes espèces. Les gènes qui évoluent rapidement vont permettre de séparer les individus étroitement liés mais risque d'être trompeur pour les individus éloignés. Par contre, les gènes qui évoluent lentement auront du mal à séparer les individus étroitement liés. Certains gènes peuvent même avoir une évolution non homogène selon les espèces. C'est-à-dire une évolution rapide dans certains groupes et une évolution très lente dans d'autres groupes.

Dans notre cas, où on effectue une comparaison à grande échelle de 174 espèces de champignons, se reposer sur un seul marqueur peut nous induire en erreur dans certains clades. Nous avons donc décidé d'inférer l'arbre phylogénétique à partir de l'ensemble des éléments communs entre toutes les espèces, plus précisément de l'ensemble du core protéome.
La concaténation du core protéome (Figure 6.2) va permettre d'augmenter le signal phylogénétique, augmenter la résolution pour obtenir un arbre le plus robuste possible (Choi and Kim, 2017; Chung *et al.*, 2018).

6.2 Détection du core protéome

Le core protéome est l'ensemble des protéines conservées par toutes les espèces. Ce core protéome a été identifié à partir des paires d'orthologues identifiées par BBH (Bidirectional Best Hit). Deux séquences protéiques A (chez l'espèces 1) et P (chez l'espèces 2) sont considérées comme orthologue si le meilleur score d'alignement de A contre le protéome de l'espèce 2 est l'alignement avec P, et réciproquement le meilleur score d'alignement de P contre le protéome de l'espèce 1 est A (Figure 6.2). En plus de la réciprocité des meilleurs hits, et afin d'obtenir le moins de faux positif possible, les séquences doivent partager au minimum 40% d'identité, s'aligner au moins sur 70% de la séquence la plus petite et présenter une E-value inférieure à 1^{-10}.

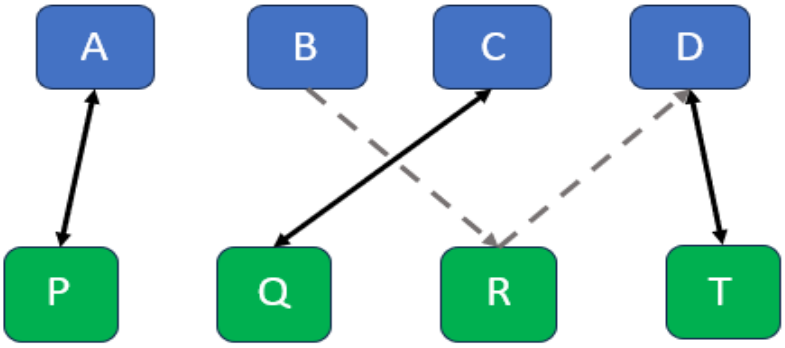


Figure 6.2 : Définition des paires d'orthologues entre deux espèces par la méthode BBH (Bidirectional Best Hit). Le protéome de l'espèce 1 est coloré en bleu et celui de l'espèce 2 en vert. Un rectangle représente une protéine. Les flèches bidirectionnelles en trait plein indiquent que les deux protéines présentent réciproquement le meilleur score d'alignement : ils sont orthologues. Les flèches grises discontinues signifient que le meilleur alignement n'est pas réciproque.

Le core protéome est identifié en se basant sur la méthode pour détecter l'orthologie. Une protéine fait partie du core protéome si elle possède un orthologue dans toutes les autres espèces et ses orthologues dans les autres espèces sont eux aussi reliés entre eux par une relation d'orthologie.
Cette définition se traduit par l'obtention d'un graphe complet (ou clique) si l'on représente les relations d'orthologie entre tous les membres du core génome (Figure 6.3).

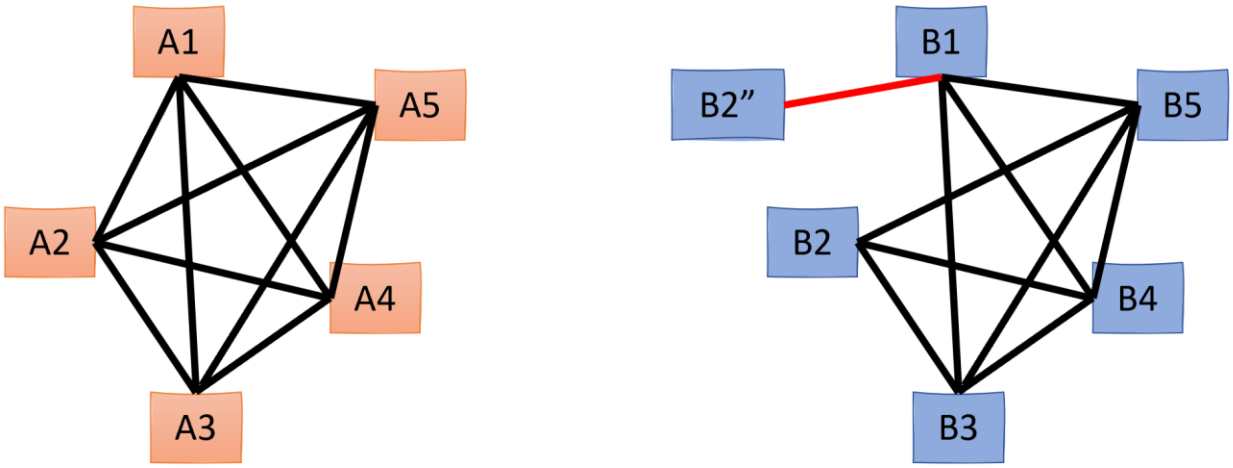


Figure 6.3 : Identification du core protéome par BBH. La lettre et la couleur du rectangle indiquent le nom de la protéine (A ou B). Le chiffre après la lettre indique l'espèce (1 et 2). Deux protéines sont reliées par une arête si elles sont orthologues. Dans la Figure de gauche, toutes les protéines sont orthologues par paires et forment un graphe complet (une clique) : la protéine A fait donc partie du core protéome. Dans la Figure de droite, B1 présente une relation d'orthologie avec une protéine B2'' qui n'a pas les mêmes orthologues que B1 : le graphe est incomplet. B ne fait donc pas partie du core protéome.

6.3 Inférence de l'arbre phylogénétique à partir du core protéome

Pour chaque protéine appartenant au core protéome, un alignement multiple entre les protéines de chaque espèce a été réalisé avec MUSCLE (Edgar, 2004). Une fois l'alignement effectué, toutes les protéines d'une même espèce sont concaténées pour former une seule super-séquence qui va servir pour construire un arbre phylogénétique (Figure 6.4).
L'arbre phylogénétique a été inféré avec IQTree (Nguyen *et al.*, 2015). C'est un outil qui utilise une méthode de maximum de vraisemblance pour inférer l'arbre et qui est adapté pour traiter des gros jeux de données. IQTree a été utilisé avec ses paramètres par défaut, plus particulièrement la recherche de l'arbre commence à partir de 100 arbres déterminés par une méthode par parcimonie et un arbre inféré par BIONJ (Gascuel, 1997) qui est une méthode basée sur les distances.

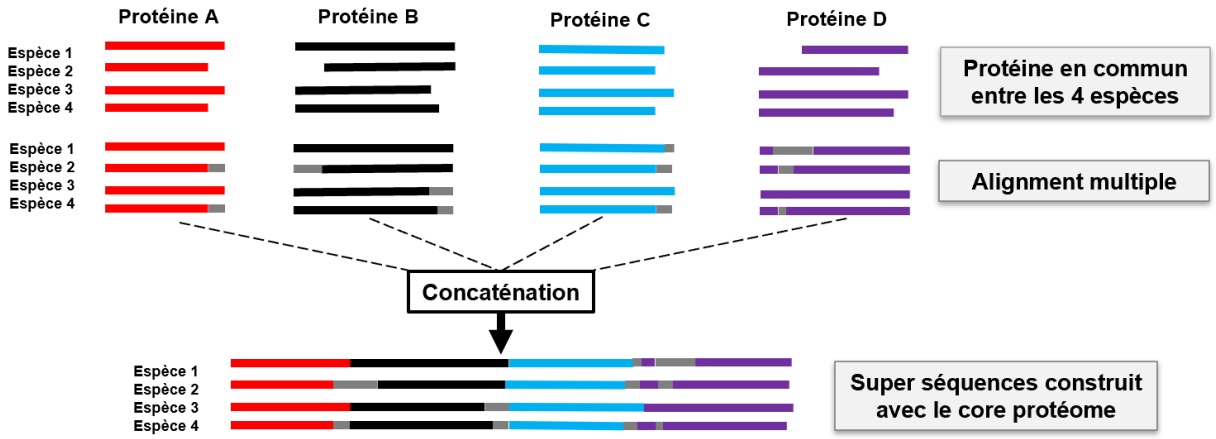


Figure 6.4 : Construction de la super séquence comme marqueur phylogénétique à partir du core protéome. Chaque protéine faisant partie du core protéine est aligné avec un outil d'alignement multiple (MUSCLE). Puis les protéines d'une même espèce faisant partie du core protéome sont concaténés pour former une super-séquence qui va servir de marqueur pour inférer l'arbre phylogénétique.

L'arbre produit par IQTree est non enraciné c'est-à-dire qu'aucun point n'a été déterminé comme étant le plus ancien (dernier ancêtre commun) et aucune direction évolutive n'a été déterminée (Baldauf, 2003). Plusieurs techniques permettent d'enraciner un arbre mais deux approches sont principalement utilisées : l'utilisation d'un groupe externe et le « midpoint rooting ».
L'utilisation d'un groupe externe est la plus populaire et la plus utilisée (Kinene *et al.*, 2016). Cette méthode suppose qu'un ou plusieurs taxons diffèrent du reste des taxons. Ce taxon est appelé groupe externe. L'estimation de la racine est tout simplement la branche où notre ou nos groupes(s) externe(s) rejoignent le reste de notre arbre d'intérêt. L'un des principaux inconvénients de cette méthode est la possession d'un groupe externe qui n'est pas toujours possible (Kinene *et al.*, 2016).
Le midpoint rooting place la racine au milieu des deux branches les plus longues (Swofford *et al.*, 1996). Cette méthode suppose que toutes les séquences ont évolué au même rythme.

Nous avons choisi comme méthode le « midpoint rooting » du fait que l'utilisation d'un groupe externe aurait nécessité de déterminer le core protéome de nos espèces d'intérêts avec le groupe externe. L'ajout du groupe externe aurait diminué le nombre d'éléments du core protéome et aurait entraîné une perte d'information pour inférer l'arbre de nos espèces.

6.4 Constructions de l'arbre phylogénétique sur les espèces étudiées

La taille du core protéome obtenue pour chaque classe taxonomique est indiquée dans la table 6.1. Le core protéome a été seulement identifié pour les classes possédant plus de 2 espèces. La lecture de cette table nous montre que la taille du core protéome des Microsporidies est très réduite. Il a été démontré que les Microsporidies ont subi la perte de nombreuses familles de protéines présentes dans d'autres eucaryotes (Nakjang *et al.*, 2013). Le nombre d'espèces influe aussi sur la taille du core protéome. Plus le nombre d'espèces est élevé, plus la taille du core protéome diminue.

Classe (nombre d'espèces)	**Taille du core protéome**
Microsporidia (6)	72
Mucoromycotina (3)	3702
Pucciniomycètes (3)	2651
Tremellomycetes (4)	3500
Agarycomycetes (27)	367
Eurotiomycetes (37)	564
Dothideomycetes (15)	1807
Saccharomycetes (35)	339
Schizosaccharomycetes (4)	2909
Sordariomycetes (25)	920
Leotiomycetes (3)	4126

Table 6.1 : Taille du core protéome par classe taxonomique. Le chiffre entre parenthèses indique le nombre d'espèces pour lesquelles le core protéome a été calculé. En jaune les champignons inférieurs, en orange les *Basidiomycota* et en bleu les *Ascomycota*.

A la lecture de la table 6.1, on voit que le calcul du core protéome de toutes les espèces sera fortement influencé par la classe taxonomique des microsporidies qui va conduire à réduire drastiquement la taille du core protéome. Par ailleurs, la comparaison des 174 protéomes de champignons peut être particulièrement longue.
Pour gagner en temps de calcul mais aussi pour limiter l'impact des Microsporidies dans

l'arbre, nous avons choisi de nous baser sur les connaissances déjà acquises pour l'arbre phylogénétique des champignons. Plus particulièrement, nous avons récupéré l'arbre phylogénétique des classes taxonomiques défini dans Mycocosm (Grigoriev *et al.*, 2014). Cet arbre de classes (Figure 6.5) va nous servir de base pour construire l'arbre phylogénétique de toutes nos espèces. En effet, les différentes classes de champignons sont aujourd'hui bien définies ainsi que leur position dans l'arbre (Spatafora *et al.*, 2017).
Il reste donc seulement à inférer la relation entre les espèces à l'intérieur de chaque classe taxonomique.
Les Figures 6.6 et 6.7 représentent l'arbre phylogénétique des 174 espèces de champignons obtenu avec l'approche décrite ci-dessus.

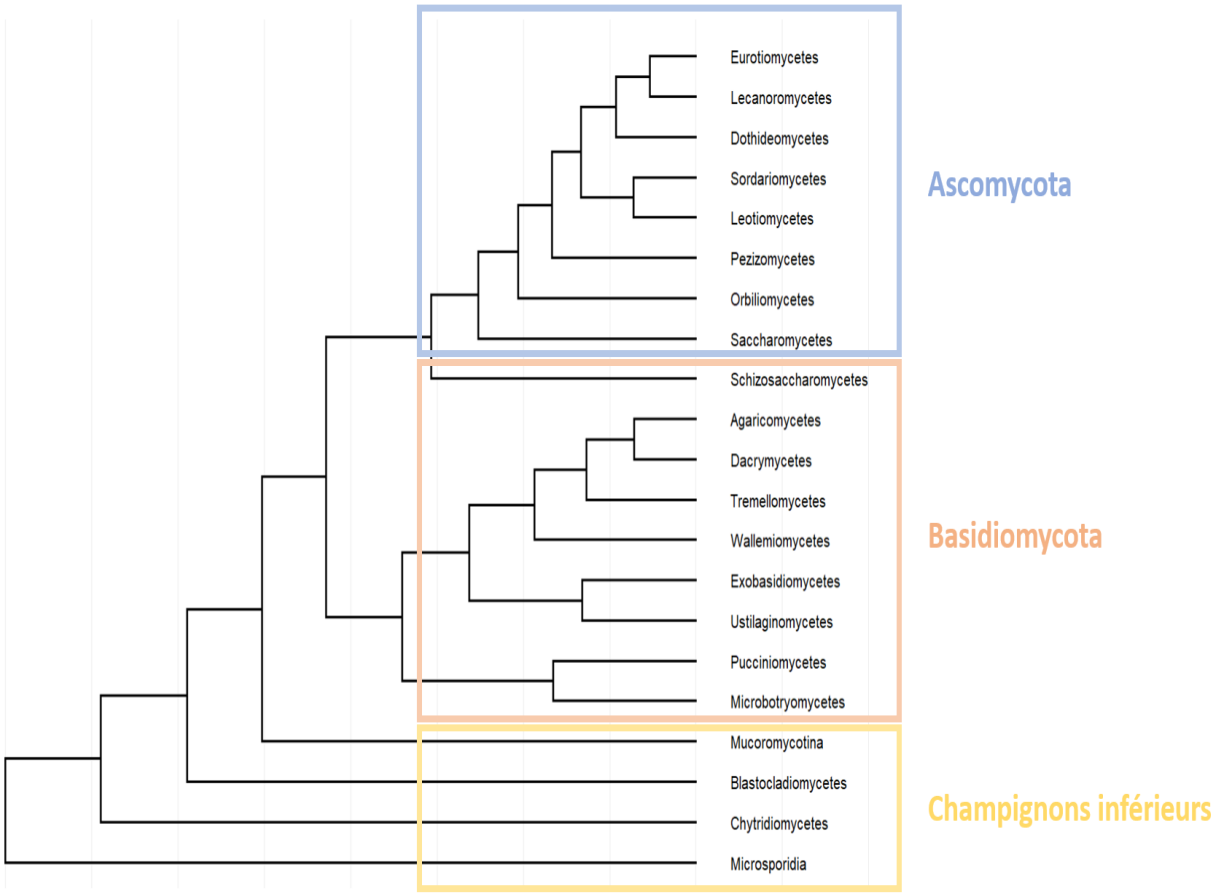


Figure 6.5 : Arbre des classes taxonomiques des champignons dans MycoCosm. Figure adaptée de MycoCosm (Grigoriev *et al.*, 2014)

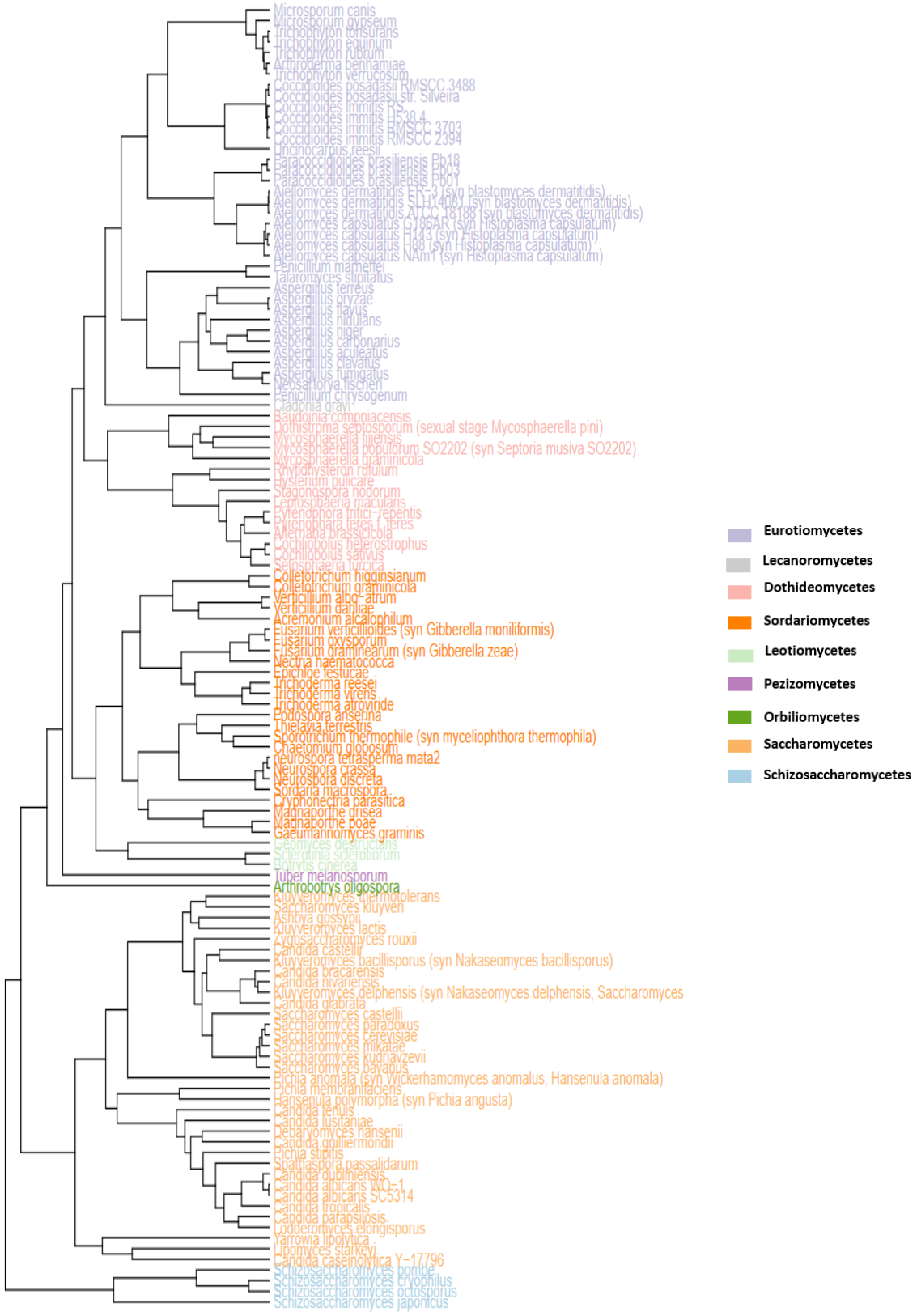


Figure 6.6. Arbre phylogénétique des champignons appartenant aux Ascomycota. Le nom des espèces est coloré en fonction de la classe taxonomique.

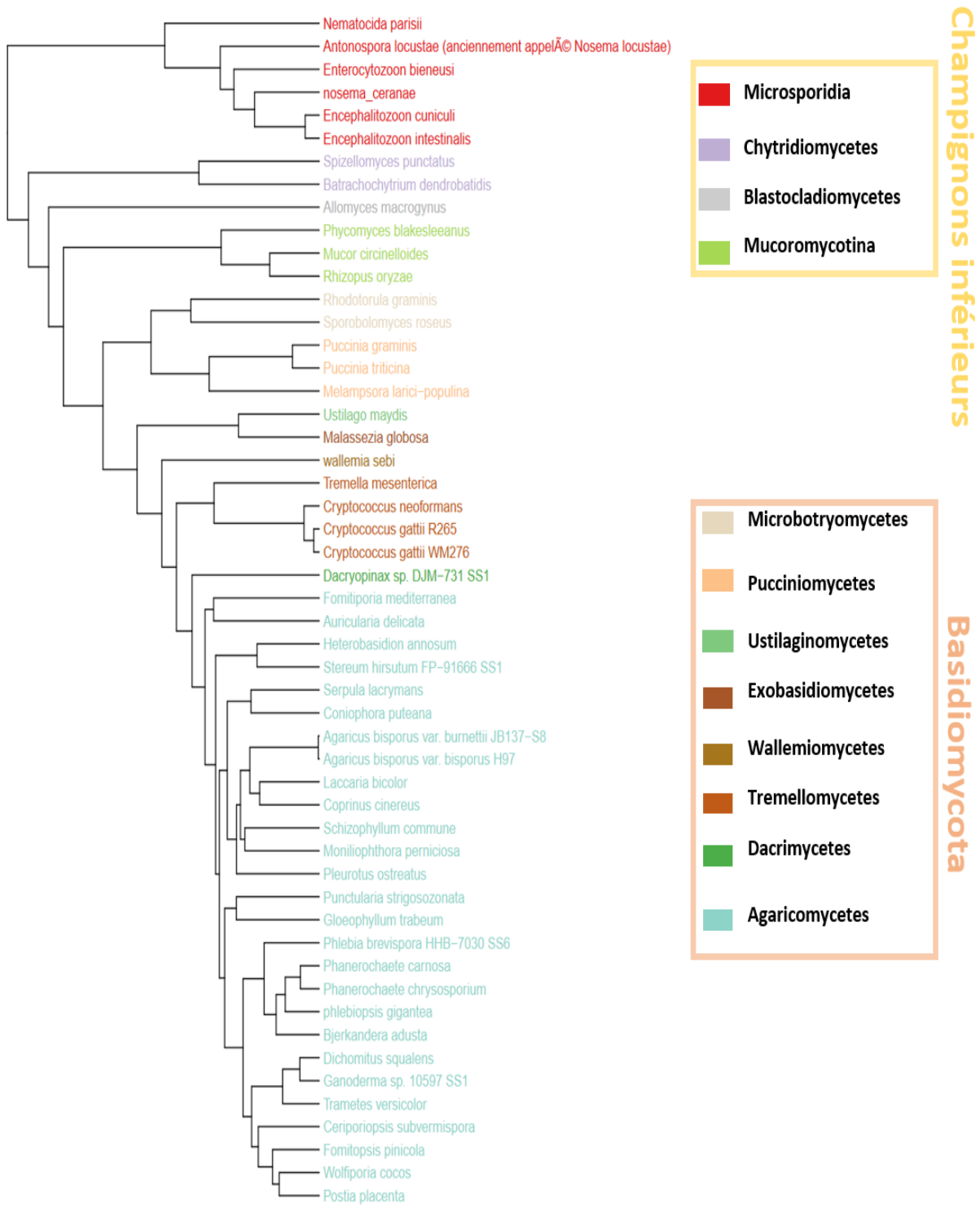


Figure 6.7. Arbre phylogénétique des champignons appartenant aux Basidiomycota et aux champignons inférieurs. Le nom des espèces est coloré en fonction de la classe taxonomique.

Chapitre :

7 Répartition des activités enzymatiques

En général, les activités enzymatiques associées au métabolisme sont très conservées dans les trois domaines de la vie. Peregrin-Alvarez et ses collaborateurs ont montré que sur 1474 activités enzymatiques associées au métabolisme, 1145 ont été détectées dans chacun des trois domaines de la vie (Archées, Bactéries et Eucaryotes) (Peregrín-Alvarez *et al.*, 2009).
Les champignons sont connus pour leur impressionnante diversité. Cette diversité se manifeste au niveau morphologique mais aussi au niveau de leur capacité métabolique. Ils sont présents dans différents types d'environnements, sont capables de cataboliser divers substrats et de synthétiser une grande diversité de composés.
Notre hypothèse est que cette diversité métabolique devrait forcément avoir un impact sur la distribution et la conservation des activités enzymatiques. Les activités enzymatiques essentielles à la vie seront probablement plus conservées alors que les activités enzymatiques à l'origine de la diversité le seront vraisemblablement moins.

7.1 Activités enzymatiques associées au métabolisme

Les enzymes associées au traitement de l'information génétique, par exemple l'ADN polymérase : 2.7.7.7 qui permet la duplication de l'ADN, sont généralement très conservées dans les trois domaines de la vie : les bactéries, les archées et les eucaryotes. Cette conservation élevée est due au rôle fondamental que jouent ces enzymes dans les processus cellulaires critiques, tels que la réplication de l'ADN, la transcription et la traduction, qui sont communs et essentiels pour la survie et la reproduction de tous les organismes vivants (Kanemaki, 2022). La conservation élevée de ces enzymes dans différents domaines de la vie reflète leur importance fondamentale dans le traitement de l'information génétique. Bien qu'il puisse y avoir des différences entre les différents organismes, leurs fonctions et mécanismes globaux sont remarquablement similaires dans tous les systèmes vivants (Marques and McCulloch, 2018).

Cependant, lors de l'analyse des réseaux métaboliques, les voies (et les activités enzymatiques) associées aux traitements de l'information génétique (ADN et ARN) sont souvent exclues. Cette exclusion se fait par commodité et pour une question pratique dans l'analyse et la construction du réseau métabolique. En effet, la nature et la fonction de ces voies sont très différentes des voies métaboliques. Ce dernier se focalise surtout sur les petites molécules telles que les carbohydrates, acides aminés, lipides et les réactions impliquées dans leur interconversion. Alors que les réactions associées au traitement de l'information génétique sont directement reliées au processus lié à l'ADN et l'ARN (réplication, réparation et modification) (Figure 7.1), leur exclusion permet de se focaliser sur les processus métaboliques qui contribuent directement à la production d'énergie et des autres molécules nécessaires pour l'organisme.

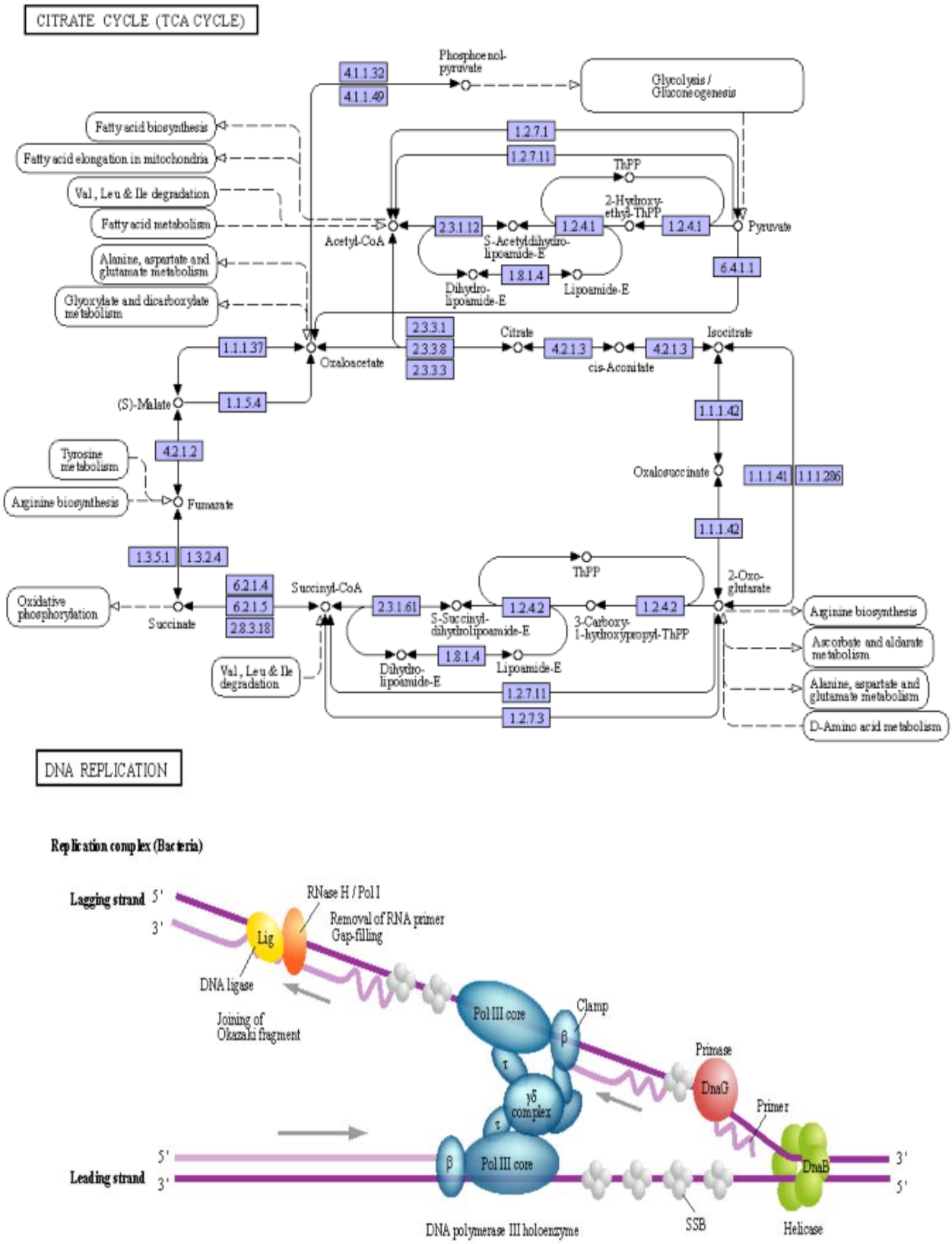


Figure 7.1 : En haut représentation des activités enzymatiques qui participent au cycle de KREBS. **En bas** représentation des enzymes liées à la réplication de l'ADN. Représentations tirées de KEGG.

Chez les 174 espèces de champignons, nous avons détecté 968 activités enzymatiques associées au métabolisme. Les groupes d'orthologues contenant les enzymes qui traitent l'information génétique n'ont pas été annotés. Sur les 968 activités enzymatiques, 942 activités enzymatiques sont présentes dans les 134 voies métaboliques de KEGG qui sont exploitables computationnellement (représenter sous forme de graphe) (version 05/2021). Certaines voies de KEGG ont été exclues du fait que ce sont des voies contenant des images figées non exploitables computationnellement.

La partie III de ce manuscrit portera sur la construction du réseau métabolique. Nous y expliquerons notre choix d'avoir utilisé KEGG comme base de données de voies métaboliques et la méthode de construction du réseau métabolique à partir des voies mises à disposition par KEGG.
Dans la partie III, des filtres topologiques et des vérifications dans la littérature ont été appliqués sur ces voies métaboliques pour identifier les voies métaboliques présentes chez les champignons. À la suite de ces filtres, nous avons identifié 102 voies présentes chez les champignons. La liste de ces voies métaboliques est disponible dans l'annexe 1. Les 102 voies métaboliques présentes chez les champignons contiennent 910 activités enzymatiques différentes**.** La liste des activités enzymatiques identifiées dans ces voies est disponible dans l'annexe 2.

7.2 Conservation des activités enzymatiques chez les champignons.

Le profil phylogénétique d'une activité enzymatique indique la présence et l'absence de cette activité enzymatique à travers les espèces étudiées. En utilisant ce profil phylogénétique, nous avons calculé le pourcentage de conservation de chaque activité enzymatique. Notre objectif est de déterminer quelles sont les activités enzymatiques qui sont partagées par presque toutes les espèces et donc que nous considérons comme essentielles pour le maintien de la vie. Nous avons également identifié les activités enzymatiques qui sont spécifiques de certains clades ou de certaines espèces (non conservées par toutes les espèces) et pourraient être à l'origine de la diversité métabolique.

Pour chaque activité enzymatique représentée par son EC-number, le pourcentage de conservation a été calculé en divisant le nombre d'espèces où l'activité enzymatique est présente par le nombre total d'espèces (174 espèces).
Les EC number sont ensuite regroupés en 20 classes en fonction de ces valeurs de façon à ce que chaque barre de l'histogramme présenté dans la Figure 7.2 corresponde à une classe de 5% de niveau de conservation. Les valeurs minimale et maximale de conservation de chacune des classes sont indiquées de part et d'autre de chaque barre. Nous observons ainsi que 450 des 910 activités enzymatiques ont un niveau de conservation supérieure à 85%.
Les classes entre 5% et 85% ont une valeur moyenne de 20 activités enzymatiques.

En fixant un *seuil* de 85% de conservation, nous avons défini que les activités enzymatiques avec au moins une conservation de 85% sont des activités enzymatiques fortement conservées, et partagées par toutes les espèces (Figure 7.2). Nous avons ainsi identifié 456 activités enzymatiques partagées par toutes les espèces et 464 activités enzymatiques qui sont moins conservées et participent à la diversité métabolique.

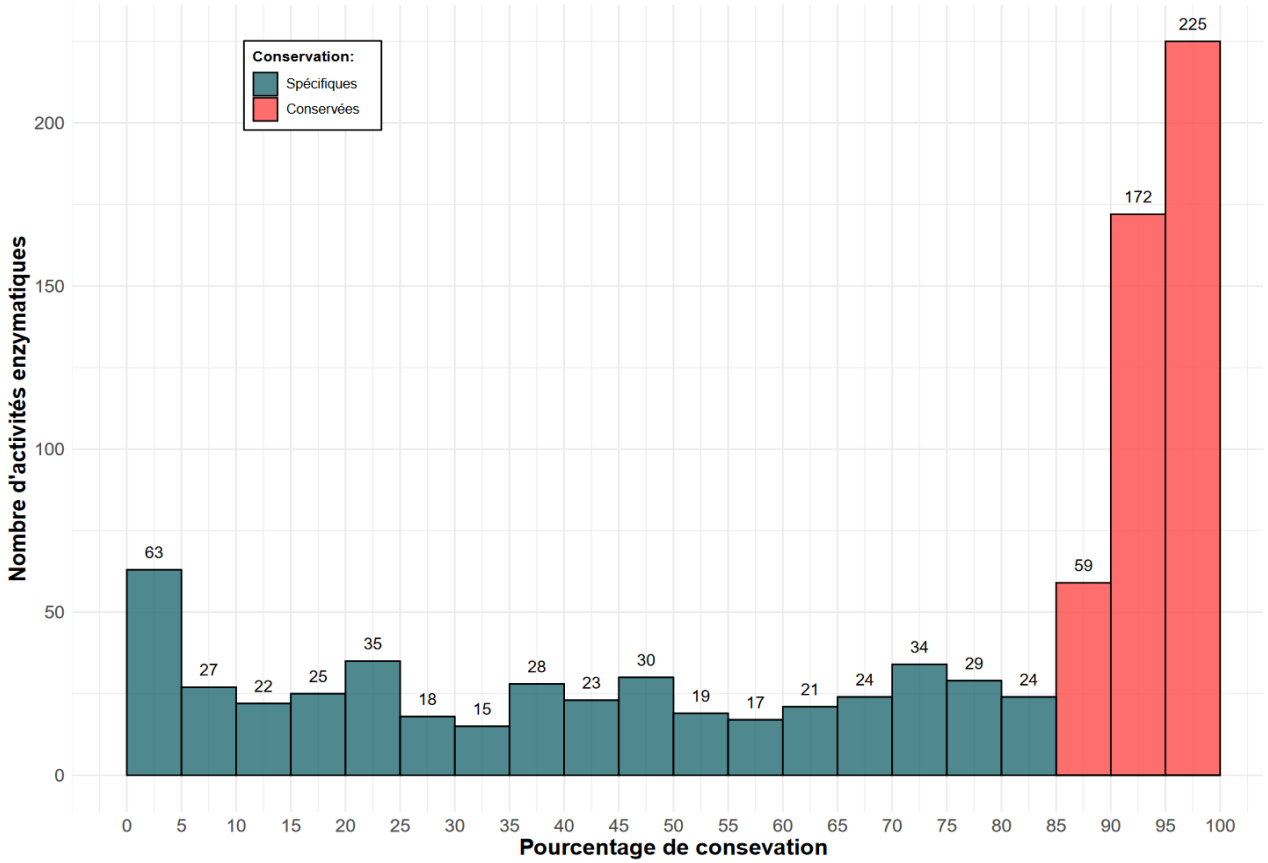


Figure 7.2 : Nombre d'activités enzymatiques regroupé par pourcentage de conservation dans les 174 espèces étudiées. Chaque barre correspond à une intervalle de 5%. Le nombre d'activités enzymatiques dans chaque barre est affiché au-dessus. Les barres rouges indiquent les activités enzymatiques ayant une conservation supérieure à 85%. Les barres bleues indiquent les activités enzymatiques ayant une conservation inférieure à 85%.

7.3 Répartition des activités enzymatiques par classe taxonomique

La répartition des activités enzymatiques par espèces (Figure 7.3) nous montre que les microsporidies ont un nombre d'activités enzymatiques moyenne de 100. Le nombre d'activités enzymatiques des espèces des autres groupes varient de 450 à 750.
Chez les Ascomycota, il est à noter que les *Schizosaccharomycetes* et les *Saccharomycetes* ont des activités plus restreintes par rapport aux autres classes du même groupe. La spécificité de ces deux classes est que la grande majorité des espèces de ces deux classes passent leur vie sous forme de levures.

Chez les *Eurotiomycetes*, il y a deux ordres qui se dégagent : les *eurotiales* avec 700 activités enzymatiques en moyenne et les *onygenales* avec 600 activités enzymatiques en moyenne. Les *eurotiales* sont majoritairement associés aux plantes tandis que les *onygenales* sont majoritairement des pathogènes de l'homme (Van Dyke *et al.*, 2019; Houbraken *et al.*, 2020). Wang et ses collaborateurs ont montré que la majorité des réductions géniques chez les onygenales sont des gènes associés à la dégradation de la cellulose et des gènes directement impliqués dans l'association avec les plantes (Wang *et al.*, 2022). La diversité enzymatique entre ces deux genres reflète vraisemblablement cette différence de mode de vie.

Une comparaison entre les 3 groupes de champignons et entre certaines classes ne peut pas être effectuée précisément du fait du déséquilibre dans les nombres d'espèces par classes et par groupes. Les Ascomycètes représentent une grande majorité de nos espèces (122/174) et certaines classes telles que les *Pezizomycetes* et les *Exobasidiomycetes* sont représentées que par une seule espèce. Une espèce seule ne pouvant pas être représentative de la classe.

On dénombre aussi 4 activités enzymatiques orphelines qui sont 3.5.2.7, 4.2.1.49, 5.1.1.7 et 1.17.4.2. Ces activités enzymatiques sont seulement retrouvées que chez *A.macorgynus (*classe des *Blastocladiomycetes)*. 3.5.27 et 4.2.1.49 sont des activités enzymatiques retrouvées chez presque tous les animaux mais pratiquement absentes chez les champignons. La seule présence de ces activités enzymatiques chez un champignon a été reportée par (Ribichich *et al.*, 2006) chez *B.emersonii* qui appartient à la classe des Chytridiomycetes, leur permettant la dégradation partielle de l'histidine. Mais cette espèce ne possède pas 4.2.1.49 ce qui ne lui permet pas d'utiliser l'histidine comme source de carbone. *A.macorgynus* , qui appartient au Blastocladiomycetes, est probablement capable de dégrader complètement l'histidine avec la présence de 4.2.1.49. Ces deux branches se situent à la racine de l'arbre phylogénétique des champignons (Figure 6.6). 3.5.2.7 et 4.2.1.49 semblent être une caractéristique ancestrale des champignons et des animaux mais qui n'est conservée que dans ces deux espèces de champignons.

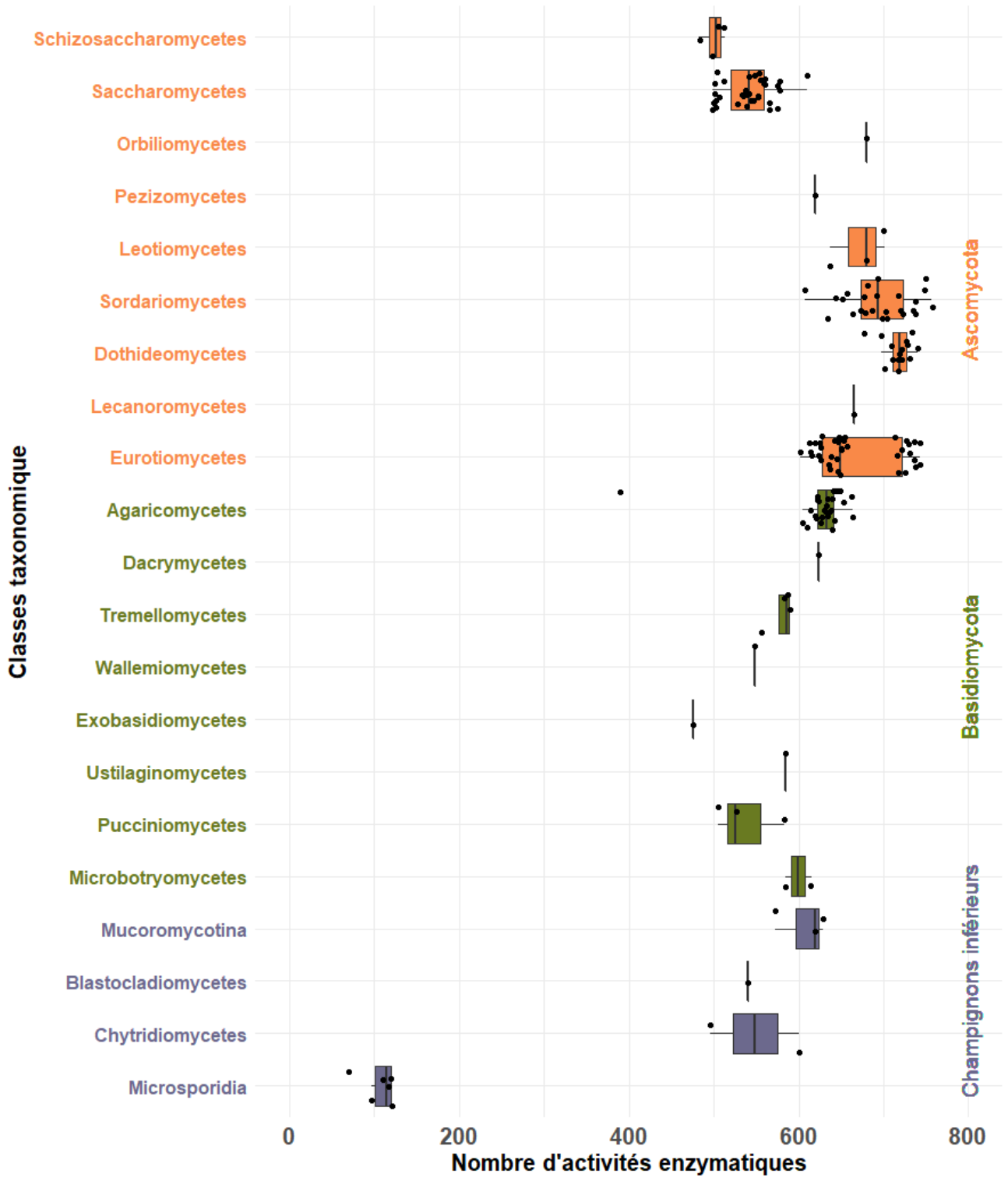


Figure 7.3 : Nombre d'activités enzymatiques par espèces regroupées par classe taxonomique. En abscisse, le nombre d'activités enzymatiques et en ordonnée la classe taxonomique de l'espèce. Les boîtes et les classes taxonomiques sont colorées en fonction du groupe : Ascomycota en orange, Basidiomycota en vert et les champignons inférieurs en violet.

Le cas des *microsporidies*

La répartition des activités enzymatiques par espèces nous montre que les microsporidies ont subi une importante réduction de leur métabolisme (Figure 7.3).

Les microsporidies sont des champignons unicellulaires totalement adaptés à la vie parasitaire. Ils parasitent un large spectre d'hôtes : mammifères, insectes, poissons…. (Stentiford *et al.*, 2016). Leur cycle de vie se déroule majoritairement à l'intérieur de cellule hôte. La seule forme de vie extracellulaire consiste en une forme de spore.

Ce sont des organismes qui ont poussé le parasitisme à son paroxysme et qui ont perdu de nombreuses voies métaboliques (Corradi, 2015), y compris la voie de synthèse de nombreux acides aminés et de nucléotides. Par conséquent ils dépendent fortement des métabolites de leur hôte (Dean *et al.*, 2016). L'un des exemples le plus marquant, est l'absence de mitochondrie canonique, permettant l'oxydation des composés organiques. Ainsi la voie de la glycolyse est la seule voie permettant de générer de l'ATP. Pendant le stade de développement intracellulaire, les Microsporidies n'utilisent pas leur propre métabolisme énergétique. A la place, ils utilisent l'ATP produit par leur hôte (Alexander *et al.*, 2016).

Le nombre d'activités enzymatiques témoigne de la réduction maximale de son métabolisme pour la vie parasitaire.

Ce processus de réduction est particulièrement visible dans leurs génomes. Ils sont en effet constitués de plusieurs chromosomes linéaires très petits avec l'un des plus petits génomes eucaryotes connus. Le génome des Encephalitozoon est seulement constitué de 2.3 Mbp (Keeling and Slamovits, 2004).

Son génome ne contient que les gènes essentiels à sa survie.

Répartition par classe taxonomique des activités enzymatiques spécifiques de certaines espèces

Les activités enzymatiques non conservées ne sont pas réparties de façon uniforme chez toutes les espèces et les différentes classes taxonomiques (Figure 7.4). Chez les Ascomycota les Sordariomycetes, les Dothideomycetes et le genre des *eurotiales* chez les Eurotiomycetes ont une répertoire enzymatiques non conservées pars toutes les espèces plus élevées que les autres classes d'Ascomycota (environ 300 activités enzymatiques non conservées). Les Saccharomycetes et Schizosaccharomycetes ont de leur part moins conservé ces activités enzymatiques, ce qui peut être associé à une duplication suivie par une perte massive de gène durant leurs évolutions (Cliften et al., 2006)

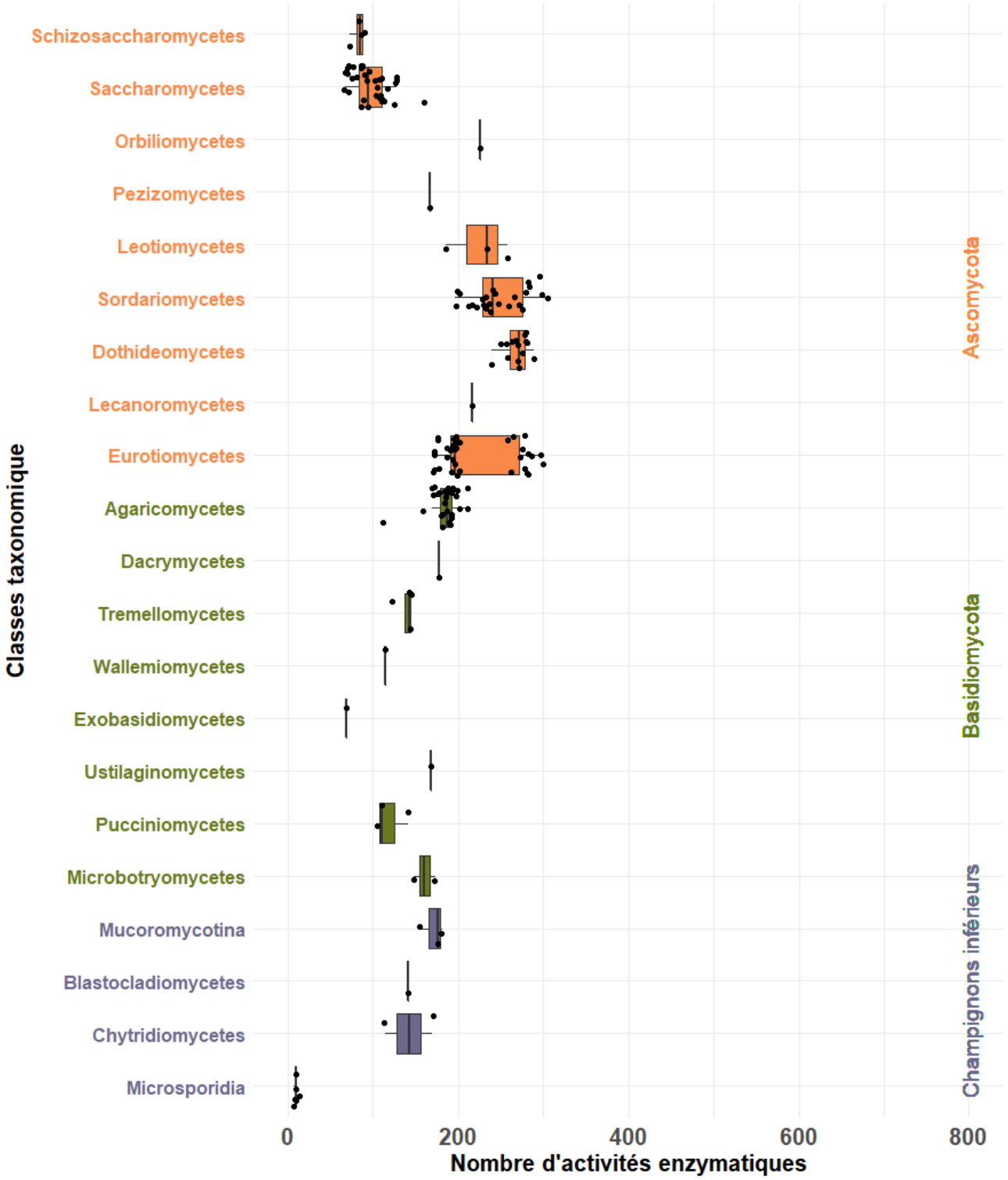


Figure 7.4 : Nombre d'activités enzymatiques non conservées par toutes les espèces par espèces regroupées par classe taxonomique. En abscisse, le nombre d'activités enzymatiques non conservées par toutes les espèces et en ordonné la classe taxonomique de l'espèce. Les boîtes et les classes taxonomiques sont colorées en fonction du groupe : Ascomycota en orange, Basidiomycota en vert et les champignons inférieurs en violet.

7.4 Répartition des activités enzymatiques par classe enzymatique

Parmi toutes les activités enzymatiques, les classes 1 : oxydoréductase (26,70%), et 2 : transférase (29,12%) sont les plus abondantes dans le réseau métabolique. Les isomérases et les ligases sont les moins représentées, 5,05% et 7,03% respectivement (table 7.1).
Au niveau de la conservation des activités enzymatiques dans les 174 espèces étudiées, les ligases sont les plus conservées (81,25%) suivies des transférases (64,52) et on observe beaucoup d'activités enzymatiques non conservées chez les oxydoréductases (64,19%). Cette classe participe beaucoup à la dynamique de l'évolution du réseau métabolique.
Les oxydoréductases sont probablement moins conservées que les transférases du fait de la nature des réactions qu'elles catalysent. En effet, une oxydoréductase transfert l'électron d'un donneur vers un accepteur. Ce qui varie ce sont le donneur et l'accepteur. Un donneur peut avoir plusieurs accepteurs et vice versa. Ainsi dans certaines espèces, une seule enzyme capable de transférer un électron à partir d'un seul donneur vers plusieurs accepteurs peut être suffisante et dans certaines conditions une enzyme plus spécifique peut être nécessaire. Par contre, dans le cas des transférases, le donneur avec un groupement fonctionnel (méthyl, phosphate…) peut être limité. Il est donc difficile de substituer une transférase avec une autre.

	Conservée	**Non-conservée**	**Total**
1 (oxydoréductases)	87 (35,80%)	156 (64,19%)	243 (26,70%)
2 (transférases)	171 (64,52%)	94 (35,47%)	265 (29,12%)
3 (Hydrolases)	76 (43,67%)	98 (56,32%)	174 (19,12%)
4 (Lyases)	48 (40.67%)	70 (59,32%)	118 (12,96%)
5 (Isomérases)	22 (47,82%)	24 (52,17%)	46 (5,05%)
6 (Ligases)	52 (81,25%)	12 (18,75%)	64 (7,03%)

Table 7.1 : Répartition des classes enzymatiques en fonction de leur conservation dans les 174 espèces. Les valeurs entre parenthèses dans les colonnes « conservée » et « non conservée » indiquent le pourcentage par rapport au nombre total d'activité enzymatique de la classe. Les valeurs entre parenthèses dans la colonne « Total » indiquent le pourcentage par rapport au nombre total d'activité enzymatique (910).

7.5 Détection des activités enzymatiques co-évoluantes

Les espèces appartenant à un même groupe monophylétique peuvent partager des propriétés communes mais c'est aussi le cas des espèces non reliées taxonomiquement et qui partagent des propriétés biologiques similaires. Cette particularité et spécificité peuvent se traduire par un module (ensemble d'activité enzymatique) spécifique à un clade ou à un environnement particulier. Ces activités enzymatiques sont probablement fonctionnellement reliées et ont subi les mêmes pressions évolutives. Par conséquent, ces activités enzymatiques sont à la fois présentes et absentes dans les mêmes espèces (co

évolues) et forment un module évolutif.

Pour détecter ces activités enzymatiques co-évoluantes, une manière très simple est de regrouper les activités enzymatiques qui ont un profil identique. C'est-à-dire à la fois présent et absent dans les mêmes espèces (Figure 7.5). Certains profils sont similaires, c'est-à-dire qu'il y a une différence minime (par exemple absence de l'activité que dans une espèce, Figure 7.5 encadré en vert).

Cette différence peut être due au fait qu'aucun orthologue n'a été détecté dans une espèce (*seuil* de détection trop strict) car la séquence a trop divergé des autres. La différence entre les profils peut être aussi due à des biais de séquençage. Certaines régions spécifiques du génome peuvent très peu couvertes en read et en fonction des filtres utilisés n'ont pas été conservées.
Par conséquent détecter des profils similaires et prendre en compte des petites différences dues aux différents biais dans la détection d'une activité enzymatique n'est pas trivial (Figure 7.5). Il faut par exemple définir jusqu'à quel niveau deux profils sont similaires.

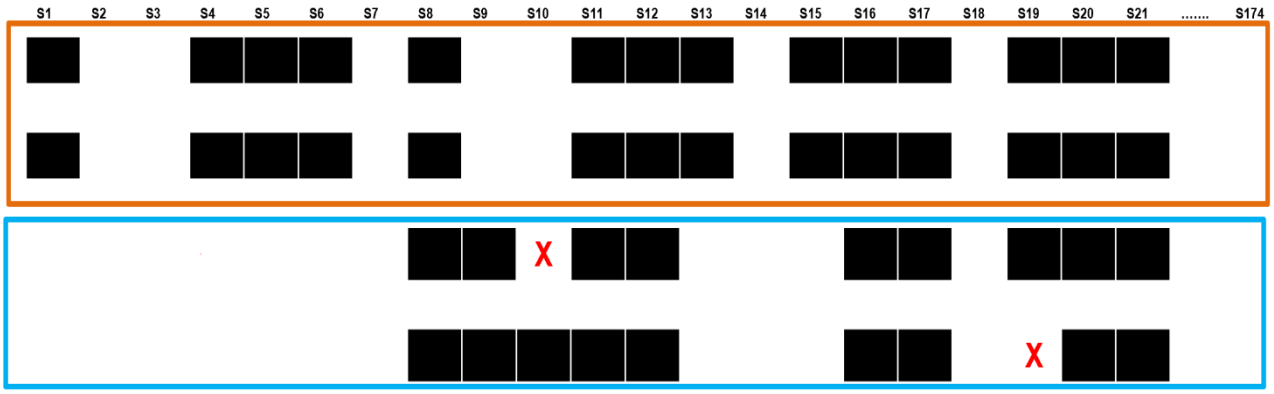


Figure 7.5 : Profils phylogénétiques similaires. Les lignes représentent les activités enzymatiques et les colonnes les différentes espèces. Les cellules en noir indiquent la présence de l'activité enzymatique et en blanc son absence. Les profils phylogénétiques en orange sont identiques et les profils phylogénétiques encadrés en bleu sont similaires. La différence entre les deux profils similaires se situe au niveau de l'espèce 10 et 19 (coché en rouge).

7.5.1 Différentes méthodes de classification

Des méthodes de classification qui permettent de regrouper des éléments similaires permettent de résoudre ce problème. Le profil phylogénétique d'une activité enzymatique peut être représenté sous forme de vecteur binaire (1 pour présence et 0 pour absence) dont la taille correspond au nombre total d'espèces. Plusieurs méthodes de classification existent et peuvent être appliquées à des vecteurs binaires. Parmi les plus courantes nous

pouvons citer les méthodes suivantes :

K-Means Clustering (MacQueen, 1967) est une méthode de classification très populaire. Dans cette méthode il faut spécifier le nombre de groupe K (cluster) souhaité. L'algorithme commence par une sélection aléatoire de K échantillons dans l'ensemble des données. Ces K premiers éléments vont servir de graines pour construire les groupes. Puis chaque vecteur binaire qui correspond au profil phylogénétique de chaque activité enzymatique va être partitionné dans un groupe dont le profil est le plus similaire.
La qualité du groupement final peut dépendre fortement des K échantillons initiaux. En d'autre termes, K-means va chercher le minimum local en fonction des K échantillons initiaux. Il faut relancer la méthode plusieurs fois pour trouver une solution satisfaisante. La meilleure approche pour utiliser cette méthode est d'identifier les meilleurs éléments k pour générer les groupes suffisamment différents en terme de profil (Frey and Dueck, 2007)

Hierarchical clustering est une méthode de classification hiérarchique. L'algorithme commence par considérer chaque observation comme un groupe avec un seul élément. Puis les groupes similaires sont fusionnés jusqu'à ce qu'un critère d'arrêt soit satisfait ou quand un seul groupe contient toutes les observations. C'est "*l'agglomerative clustering*".
L'inverse de l'« *agglomérative clustering* » est le « *divise clustering* ». Au début, il n'y a qu'un seul groupe qui contient toutes les observations. À chaque itération, le cluster le plus hétérogène est séparé en deux jusqu'à ce que chaque observation forme un cluster singleton. Dans les deux cas, le résultat est représenté sous forme d'arbre appelé dendrogramme. La similarité entre les éléments peut être mesurée en utilisant des métriques de distance appropriées. C'est une méthode peu pratique car il faut déterminer les nombres de groupes et les critères d'arrêt pour fusionner ou diviser les groupes (Fraley and Raftery, 1998).

Clustering supervisé est une technique d'apprentissage automatique. Dans le clustering supervisé, un ensemble d'éléments est étiqueté avec leur assignation dans leur groupe respectif tandis que d'autres éléments ne le sont pas. Cette approche est utilisée lorsque l'on souhaite exploiter les éléments étiquetés disponibles pour guider le clustering des éléments non étiquetés (Eick *et al.*, 2004).

Le choix d'une méthode de clustering dépend fortement du problème biologique à résoudre mais aussi des caractéristiques des données. Dans les jeux de données biologiques, tous les éléments ne devraient pas être classifiés si les contraintes ne sont pas respectées. Ces éléments uniques dans leur caractéristique peuvent avoir un sens biologique. Dans la plupart des cas biologiques, aucune connaissance ni a priori ne sont connus sur les groupes à obtenir, ainsi les méthodes de classifications supervisées ne sont pas adaptées. Par conséquent, nous avons choisi une méthode de classification qui nous semblait plus appropriée à notre problématique et aux données biologiques.

7.5.2 Une méthode de classification : Cluster AGgregation (CLAG)

CLAG (Dib and Carbone, 2012) est une méthode de classification non hiérarchique et non supervisée. CLAG a été spécifiquement développée pour classifier des jeux de données biologiques. Par exemple des résidus de protéines ayant subi les mêmes pressions évolutives, ou des données de RNA-seq pour détecter des gènes co-exprimés. CLAG est une méthode qui ne classe pas tous les éléments dans un groupe. Seuls les éléments qui ont une similarité satisfaisant un critère défini seront classés. À chaque utilisation, avec les mêmes paramètres et les mêmes données, CLAG renvoie les mêmes résultats. CLAG a été comparée avec d'autres méthodes de classification sur des jeux de données de référence. D'après ces comparaisons, CLAG est plus informatif et plus précis que les autres méthodes de classification connues (par exemple les méthodes K-means et hierachical clustering) et donne de meilleurs résultats sur les données multidimensionnelles (Dib and Carbone, 2012). L'algorithme CLAG se divise en deux parties. La première partie consiste à classer les vecteurs les plus similaires ensembles. La similarité entre deux vecteurs se base sur 3 scores : le score d'environnement, le score de différence et le score de symétrie. Le score d'environnement mesure le nombre de caractères similaires entre deux vecteurs. La différence est l'opposé du score d'environnement, c'est le nombre de caractères non similaires entre deux vecteurs. La similarité entre les caractères est modulée par la valeur de *delta*. Le score de symétrie mesure à quel point deux vecteurs (éléments) sont symétriques. Un groupe dans CLAG est défini par un score d'environnement, un score de différence et de symétrie identique entre les paires d'éléments du groupe. La seconde partie de l'algorithme correspond à une étape d'agrégation où les groupes définis précédemment sont fusionnés si un élément en commun est partagé et que le score d'environnement entre les deux groupes ne dépasse pas un *seuil* (N) (Figure 7.6).

CLAG est écrit en PERL mais une librairie pour utiliser CLAG est disponible sous R (https://clag.r-forge.r-project.org/). CLAG prend en entrée une matrice, les lignes de la matrice sont les éléments à classer en fonction des caractères dans les colonnes. Dans le cas d'une matrice de profils phylogénétiques, les éléments à classer sont les activités enzymatiques en fonction de leur présence et absence dans les 174 espèces. CLAG prend aussi en entrée 2 paramètres : le *seuil* et le *delta*, le *delta* va moduler le score de similarité entre les caractères. Un *delta* élevé signifie qu'il faut une similarité très élevée pour que deux caractères soient considérés comme similaires. Le *seuil* va avoir un impact sur les éléments qui seront classés ensemble. Par conséquent, il va falloir déterminer le *seuil* et le *delta* à utiliser pour obtenir des groupes informatifs et pertinents.

Pour déterminer les paramètres de *seuil* et de *delta* optimaux pour classer nos données avec CLAG, nous avons essayé de (1) maximiser la similarité intra-groupe et (2) minimiser la similarité inter-groupe, c'est-à-dire les éléments d'un même groupe doit être les plus proches possibles et les éléments entre différents groupes doivent être les plus distants possibles.

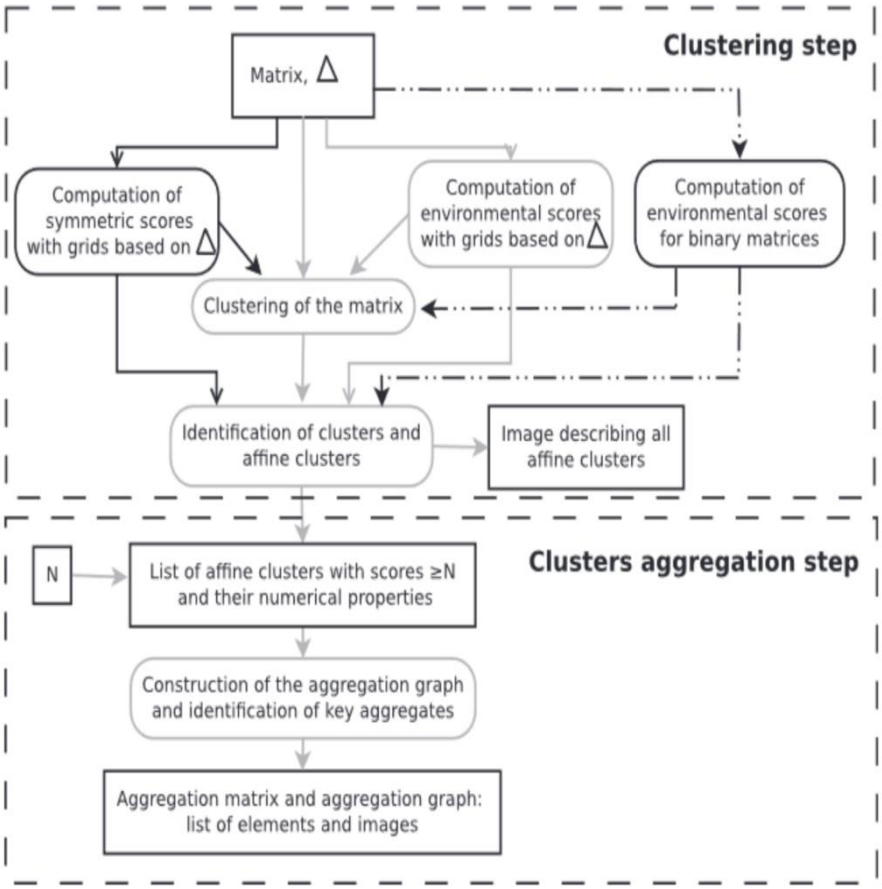


Figure 7.6 : Illustration des différentes étapes de la classification de CLAG. La première étape (clustering step) consiste d'abord au calcul des trois scores de similarités : le score d'environnement et le score de différence qui sont calculés simultanément, et le score symétrique. Les éléments sont classés en fonction de ces trois scores. Une matrice binaire sera classée en ne se basant que sur le score d'environnement et le score de différence. La deuxième étape consiste à fusionner les groupes qui ont des éléments identiques si leur score d'environnement ne dépasse pas un *seuil* N. Figure tirée de (Dib and Carbone, 2012)

7.5.3 Paramétrage de CLAG

Notre matrice de départ est la matrice de profils phylogénétiques des 910 activités enzymatiques qui est représentée sous forme de matrice binaire (Figure 7.7).

174 espèces

910 EC-number

1	1	1	1	0	...
0	1	1	0	1	
1	1	1	1	1	
1	1	0	1	1	
1	0	0	1	0	
...					

Figure 7.7 : Matrice binaire de profils phylogénétique. Chaque colonne représente une espèce et chaque ligne une activité enzymatique. Une cellule prend la valeur de 1 si l'activité enzymatique est présente dans la colonne (l'espèce) correspondante, sinon elle prend la valeur de 0.

Afin de déterminer les valeurs de *delta* et le *seuil* à utiliser, mais aussi quelle est la façon la plus efficace pour identifier les activités enzymatiques co-évoluantes et de respecter les contraintes (1) et (2) indiquées dans la partie 7.5.2, nous avons mis en place 2 protocoles de classification. Pour chaque protocole nous avons fait varier les valeurs de *delta* et de *seuil*. Plus précisément, nous avons testé les valeurs suivantes :

- *Delta* : 0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4
- *Seuil* : 0, 0.25, 0.5, 0.75

CLAG est capable de classer des éléments directement à partir d'une matrice binaire. Le premier protocole consiste à classer les activités enzymatiques directement en fonction de leur profil phylogénétique. Dans ce cas CLAG va classer les lignes (les activités enzymatiques) en fonction des colonnes (présence ou absence). Avec une matrice binaire, CLAG ne prend pas en compte le paramètre *delta* car les valeurs prises par la matrice ne sont que 1 et 0.

Dans la suite du manuscrit, ce protocole sera nommé matrice binaire.

Pour le second protocole, nous avons créé une matrice de distance entre les profils phylogénétiques. Comme métrique de distance, nous avons choisi deux métriques dont les idées sont totalement opposées :

- L'indice de Jaccard (Jaccard, 1901)
- la distance de Hamming (Hamming, 1950).

L'indice de Jaccard permet de mesurer la similarité et la diversité entre deux objets. L'indice ou la similarité de Jaccard entre deux objets A et B est défini comme étant le rapport de la

taille de leur intersection sur la taille de leur union :

$$J = \frac{(A \cup B)}{(A \cap B)}$$

De manière formelle, pour deux vecteurs binaires A et B, l'indice de Jaccard se calcule comme suit :

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

M_{11} : représente le nombre d'éléments qui valent 1 dans A et dans B
M_{01} : représente le nombre d'éléments qui valent 0 dans A et 1 dans B
M_{10} : représente le nombre d'éléments qui valent 1 dans A et 0 dans B

Contrairement à l'indice de Jaccard, la distance de Hamming permet de quantifier la différence entre deux séquences. Formellement la distance de Hamming entre deux vecteurs binaires A et B s'écrit :

$$H = M_{10} + M_{01}$$

M_{01} : représente le nombre d'éléments qui valent 0 dans A et 1 dans B
M_{10} : représente le nombre d'éléments qui valent 1 dans A et 0 dans B

Ainsi pour chaque paire de profils phylogénétiques nous avons calculé l'indice de Jaccard et la distance de Hamming pour construire deux matrices symétriques de taille 910 x 910 (Figure 7.8).

910 EC-number

910 EC-number

1.0	0.4	0.8	0.6	0.5	...
0.4	1	0.6	0.4	0	
0.8	0.6	1	0.8	0.4	
0.6	0.4	0.8	1	0.5	
0.5	0	0.4	0.5	1	
...					

Figure 7.8 : Transformation de la matrice binaire en une matrice de distance entre les activités enzymatiques en utilisant l'indice de Jaccard ou la distance de Hamming.

Dans la suite du manuscrit le protocole avec l'indice de Jaccard et la distance de Hamming sera nommé respectivement, matrice de Jaccard et matrice de Hamming.

Pour chaque protocole nous avons testé différents *seuils* et *delta*.
Puis après la classification obtenue, nous avons comparé les résultats des différents protocoles en fonction des valeurs du *seuil* et de *delta*, en s'intéressant d'abord au nombre d'activités enzymatiques classifiées, au nombre de groupes obtenus, puis aux deux critères que nous avons estimés être les plus importants, à savoir la similarité intra-groupe et la similarité inter-groupe.

La similarité intra-groupe pour un groupe C se calcule grâce à la formule suivante :

$$intra_C = moyenne(d(i,j)), i \neq j \ et \ i,j \in C$$

d(i,j) est la distance entre les profils phylogénétiques de deux activités enzymatiques i et j appartenant au même groupe C.

La similarité inter-groupe pour deux groupes C et K se calcule grâce à la formule suivante :

$$inter_{CK} = \ moyenne\big(d(i,j)\big), i \in C \ et \ j \in K$$

d(i,j) est la distance entre les profils phylogénétiques de deux activités enzymatiques i et j appartenant à deux groupes différents.

Dans le calcul de la similarité intra-groupe et inter-groupe, la distance d(i,j) utilisée est l'indice de Jaccard.

7.5.4 Choix des paramètres

Comparaison du nombre d'activités enzymatiques classées et du nombre de groupes obtenus

CLAG ne cherche pas à classer les éléments qui n'ont pas de similarités avec les autres éléments. Ainsi nous observons dans nos résultats que pour chaque protocole ou matrice utilisée, des profils phylogénétiques n'ont pas été classés dans un groupe (1ère colonne Figure 7.9). C'est-à-dire que le profil phylogénétique de ces activités enzymatiques n'a aucune similarité avec d'autres activités enzymatiques.
Nous observons aussi, que plus le *seuil* augmente plus le nombre d'activités enzymatiques classées diminue (1ère colonne Figure 7.9). Ce résultat est cohérent avec l'effet connu pour le *seuil*, à savoir qu'un *seuil* élevé signifie qu'il faut un score de similarité élevé pour regrouper ensemble des éléments.
Avec la matrice de Hamming, plus le *seuil* augmente plus le nombre d'activités enzymatiques classées diminue drastiquement. Par exemple, avec un *seuil* de 0.75 très peu d'activités enzymatiques sont classées avec la matrice de Hamming (Figure 7.9).
La différence principale entre les 3 méthodes est le nombre de groupes obtenus. Ainsi, le

nombre de groupes obtenus avec la matrice binaire est deux fois plus élevé que le nombre de groupes obtenus avec la matrice de Hamming ou de Jaccard pour un même nombre de profils phylogénétiques classés (2ème colonne Figure 7.9).

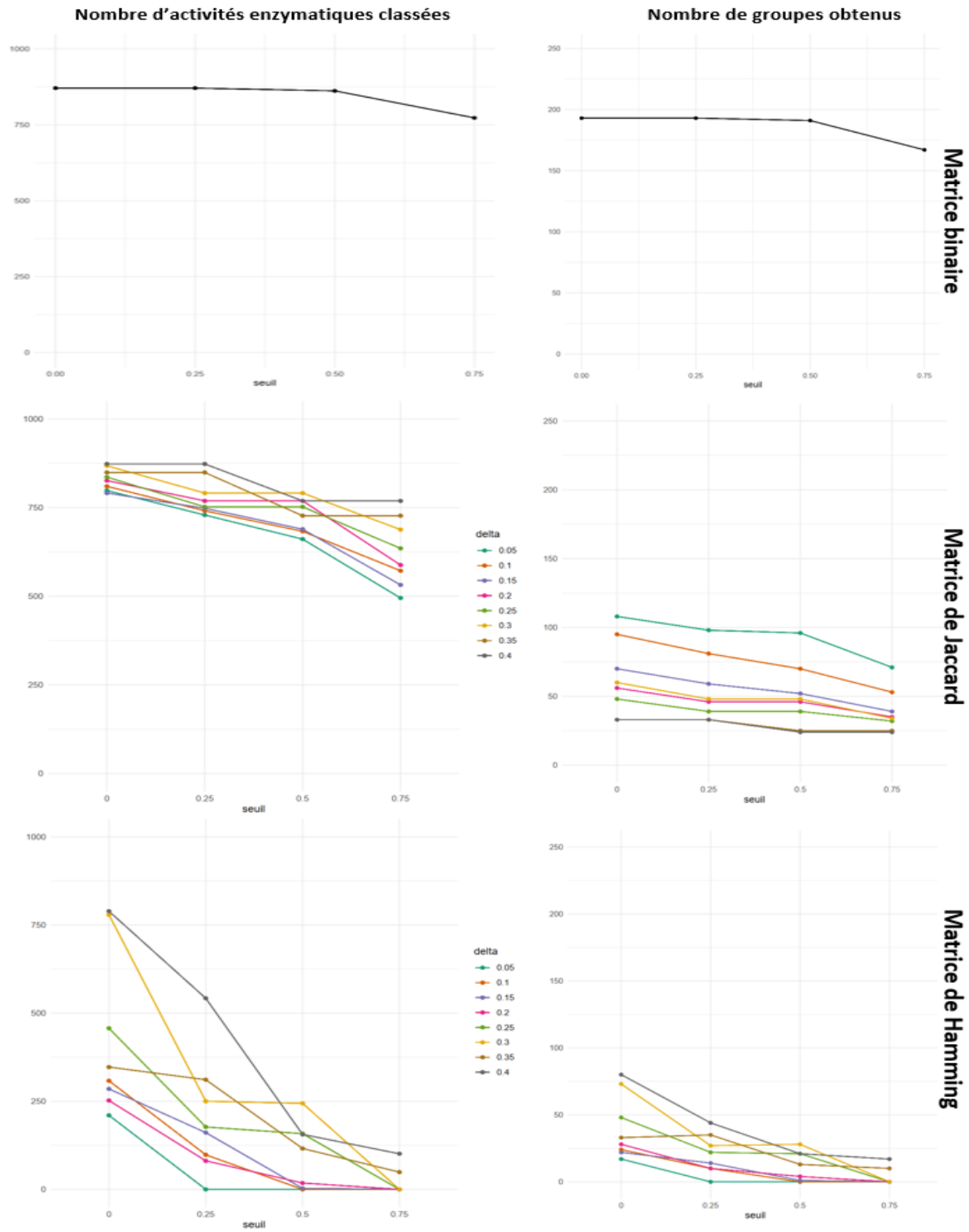


Figure 7.9 : Nombre d'activités enzymatiques classées et nombre de groupes obtenus à partir des trois matrices (binaire, jaccard ou hamming) en fonction des valeurs du *seuil* (en abscisse) et du *delta* (courbes colorées). La première colonne représente le nombre d'activités enzymatiques classées et la deuxième colonne représente le nombre de groupes obtenus. Chaque ligne représente l'une des trois matrices utilisées: matrice binaire, matrice de Jaccard et matrice de Hamming.

Comparaison de la similarité intra-groupe et inter-groupe

La première chose que l'on observe dans la similarité intra-groupe entre les 3 méthodes est que la similarité intra-groupe obtenue avec la matrice de Hamming est largement inférieure à celle deux autres matrices (Figure 7.10). Cela signifie que les profils phylogénétiques dans un groupe défini avec cette méthode sont moins similaires que les groupes définis avec les deux autres méthodes. La similarité intra-groupe est calculée si au moins un groupe de profils phylogénétiques similaires a été déterminé.
Avec la matrice binaire, la similarité intra-groupe reste quasiment identique pour les valeurs du *seuil* de 0, 0.25 et 0.5. Les groupes obtenus avec ces *seuils* sont quasiment les mêmes. Nous obtenons cependant la meilleure similarité intragroupe avec un *seuil* de 0.75 (Figure 7.10 A).
La matrice de Jaccard nous montre que plus nous augmentons le *seuil*, plus la similarité intra-groupe augmente. Nous obtenons les meilleures similarités intra-groupe avec un *seuil* de 0.75. Avec un *seuil* de 0.75, la similarité intra-groupe est très proche de 1 avec les *deltas* 0.05, 0.1, 0.15, 0.2, 0.25. Ce qui signifie que les profils au sein de ces groupes sont très similaires.

La similarité inter-groupe permet de mesurer la similarité entre des groupes distincts. Plus la similarité inter-groupe est proche de 0, plus les profils entre les groupes sont différents. La similarité inter-groupe est calculée sur les couples de paramètres où au moins deux groupes ont été identifiés.
Avec la matrice binaire, nous n'avons observé aucune évolution de la similarité inter-groupe que ce soit en modifiant la valeur de *delta* ou du *seuil* (Figure 7.11 A).
Avec la matrice de Hamming, les similarités inter-groupe se situent entre 0.5 et 0.25 mais sans grande variation entre les différents couples de paramètres.
Avec la matrice de Jaccard, nous observons que plus nous augmentons la valeur de *delta*, plus la valeur inter-groupe diminue. Pour chaque valeur du *seuil* testé, nous obtenons la plus petite similarité inter-groupe avec une valeur de *delta* égale à 0.4 (Figure 7.11 B). Cela signifie que plus on augmente la valeur de *delta,* plus les groupes définis avec la matrice de Jaccard deviennent plus distincts.

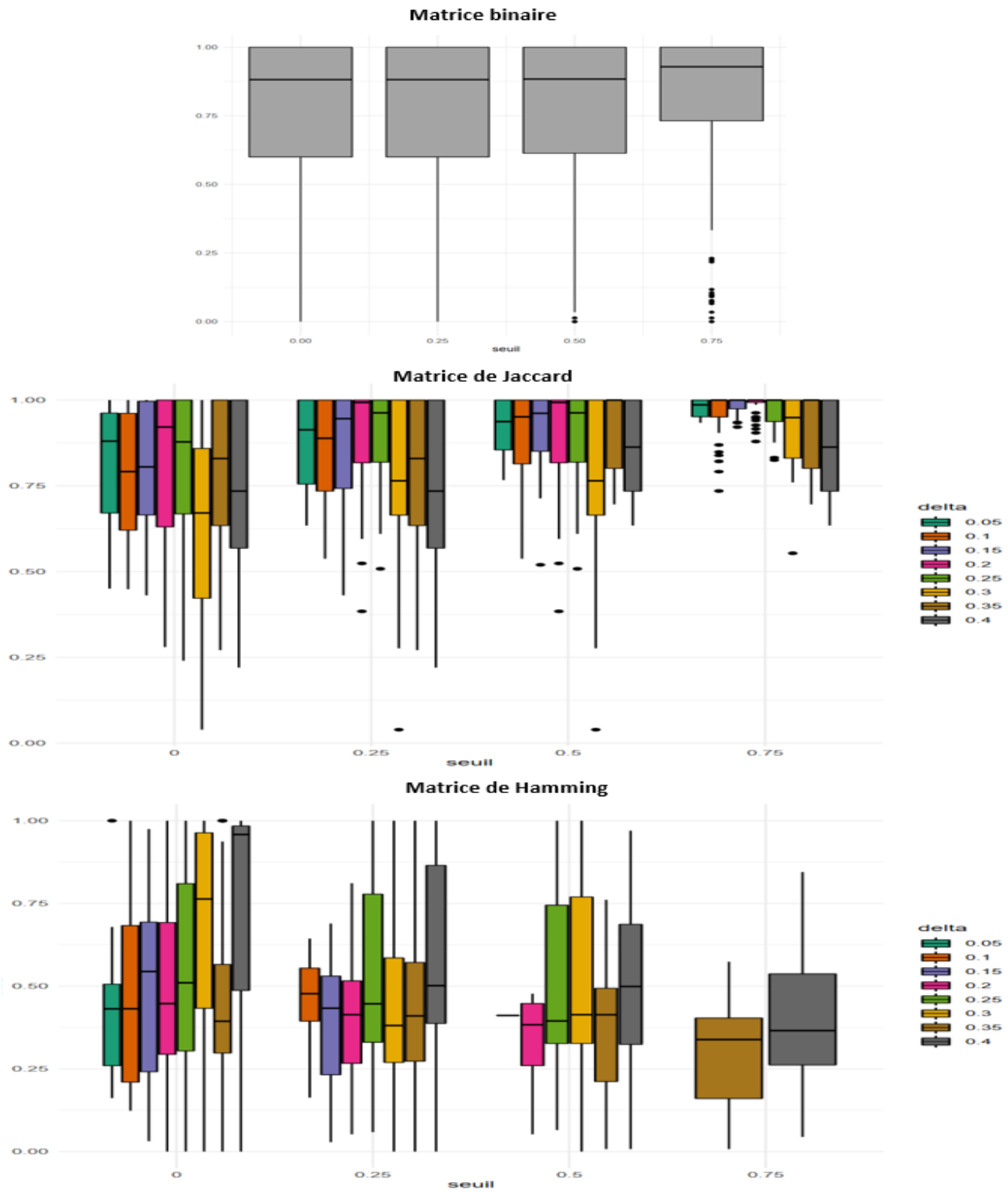


Figure 7.10 : Similarité intra-groupe en fonction des valeurs du *seuil* (en abscisse) et de *delta* (couleur de chaque boîte). De haut en bas : **A.** Classification à partir de la matrice binaire (ne prend pas en compte la valeur de *delta*), **B.** Classification à partir de la matrice de Jaccard et **C.** Classification à partir de la matrice de Hamming. Plus la valeur de la similarité est élevée, plus les profils phylogénétiques dans un même groupe sont similaires.

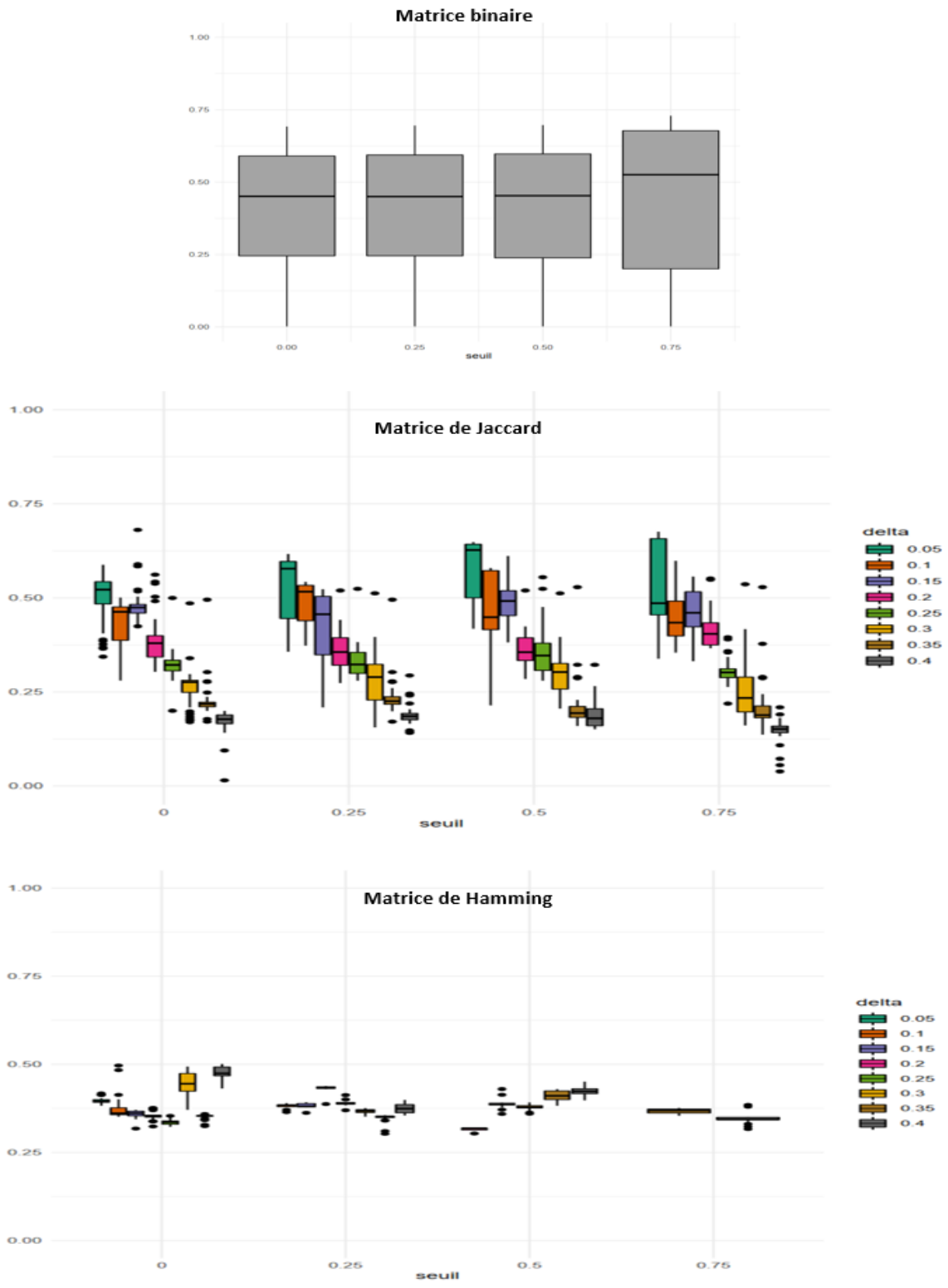


Figure 7.11 : Similarité inter-groupe en fonction des valeurs du *seuil* (en abscisse) et de *delta* (couleur de chaque boîte). De haut en bas : **A.** Classification à partir de la matrice binaire (ne prend pas en compte le *delta*), **B.** Classification à partir de de la matrice de Jaccard **C.** Classification à partir de la matrice de Hamming. Plus la similarité est faible, plus les profils phylogénétiques entre les groupes sont distincts.

Choix du *seuil*

Au vu des résultats obtenus avec les trois matrices et les différents paramètres testés, il ressort que nous n'avons pas obtenu des résultats satisfaisants avec la matrice de Hamming. Nous avons classé très peu d'activités enzymatiques par rapport aux autres méthodes et les similarités intra et inter-groupes ne sont guère meilleures. Ceci peut s'expliquer par le fait que la distance de Hamming va surtout mettre en valeur la différence entre les profils phylogénétiques. Cette distance va surtout éloigner les éléments à classer. Étant donné que l'algorithme de classification va regrouper les éléments par similarité, l'écartement des données rend la classification difficile.

Avec la matrice binaire, qui ne prend en compte que le *seuil* dans l'algorithme, il n'y a pas de réelles différences dans les résultats entre les différents *seuil*s utilisés. La plus grosse différence se situe au niveau du nombre de groupes obtenus par rapport aux deux autres matrices. Ce qui laisse à penser que les groupes obtenus sont de plus petites tailles. La Figure 7.12 nous montre que les 10 plus grands groupes obtenus sont constitués par des activités enzymatiques ayant une conservation supérieure à 85% et que nous avons considéré comme conservées chez toutes les espèces. Nous nous attendons à ce que ces activités enzymatiques très conservées forment un seul groupe. L'observation des groupes obtenus avec la matrice binaire nous montre que CLAG est trop sensible aux variations dans les profils phylogénétiques si la classification se fait directement à partir de la matrice de profil phylogénétique.

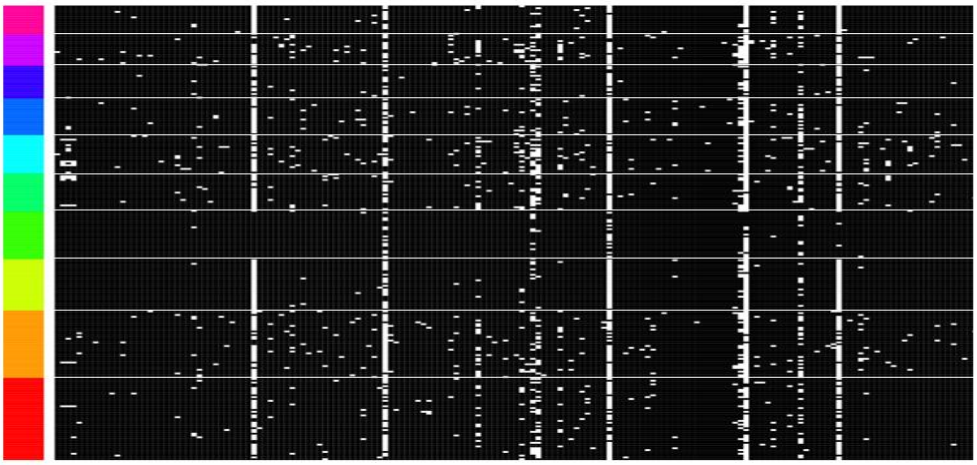


Figure 7.12 : Les 10 plus grands groupes obtenus avec la matrice binaire pour un *seuil* de 0.5. Chaque groupe est indiqué par la barre de couleur à gauche et séparé par une ligne blanche. Les lignes représentent les activités enzymatiques et les colonnes les espèces (ordonnés en fonction de l'arbre phylogénétique). Une cellule noire indique la présence de l'activité enzymatique, une cellule blanche indique son absence.

Avec la matrice de Jaccard, nous avons obtenu les résultats qui permettent de maximiser la similarité intra-groupe avec une valeur du ***seuil* de 0.75** et de minimiser la similarité inter-groupe avec la valeur de *delta* de 0.4. Or, pour maximiser la similarité intra-groupe avec un *seuil* de 0.75, il faut choisir les *delta*s 0.1, 0.15, 0.2 ou 0.25. Avec un *delta* de 0.4 nous minimisons la similarité inter-groupe mais nous ne maximisons pas la similarité intra-groupe (Figures 7.10 et 7.11). Par conséquent, il faut trouver un compromis entre les différentes valeurs de *delta*.

Choix du *delta* avec la matrice Jaccard avec un *seuil* de 0.75

Il est clair que pour maximiser la similarité intra-groupe, il faut utiliser un *seuil* de 0.75 (Figure 7.10 B). La difficulté se trouve au niveau du choix de *delta*. Nous observons qu'entre une valeur de *delta* de 0.1 ou de 0.4 il y a une très grande différence sur la similarité inter-groupe. C'est également vrai pour les valeurs de *delta* de 0.1, 0.15, 0.2 ou 0.25 pour lesquelles la similarité intra-groupe est très proche. En comparant les profils phylogénétiques des 10 premiers groupes de chaque valeur de *delta* avec un *seuil* de 0.75, nous pourrions mieux voir les groupes obtenus en fonction de la valeur de *delta*. Ceci nous permettra surtout de voir la différence entre le profil des groupes et l'impact sur la similarité inter-groupe.

Avec une valeur de *delta* de 0.05, l'algorithme est trop sensible à la moindre variation dans les profils phylogénétiques (Figure 7.13). Dans la Figure 7.13 nous pouvons observer que les groupes rouge, bleu, vert et jaune sont très similaires.

Nous avons analysé le profil global de chaque groupe pour les résultats de la classification avec des valeurs de *delta* de 0.1,0.15,0.2 et 0.25.
Avec une valeur de *delta* de 0.1 (Figure 7.14 haut gauche), nous remarquons une similitude de profil entre le profil du groupe bleu et du groupe rouge, à savoir un groupe d'activités enzymatiques très conservées. Ce groupe bleu est absent du résultat de la classification avec une valeur de *delta* de 0.15 (Figure 7.14 haute droite) car fusionné avec le groupe rouge.
Les profils globaux de chaque groupe avec une valeur de *delta* de 0.15 sont bien différents.
Pour une valeur de *delta* de 0.2 (Figure 7.14 bas gauche), le groupe bleu avec une valeur de *delta* de 0.15 (Figure 7.14 haute droite) a été fusionné avec le groupe rouge. Le fait d'augmenter le *delta* rend l'algorithme CLAG moins sensible.
Avec une valeur de *delta* de 0.25 (Figure 7.14 bas droite), nous observons que le groupe vert observé avec une valeur de *delta* de 0.2 a été clairement regroupé dans le groupe rouge.

Au vu de ses observations, nous avons choisi de travailler avec les résultats de la classification avec comme paramètre une valeur du *seuil* de 0.75 et une valeur de *delta* de 0.15. Ce couple de paramètres semble en effet offrir le meilleur compromis en termes de distance intra-clusters et inter-cluster. Ce couple de paramètre nous offre aussi la meilleure sensibilité pour obtenir des groupes de profils phylogénétiques bien distincts .

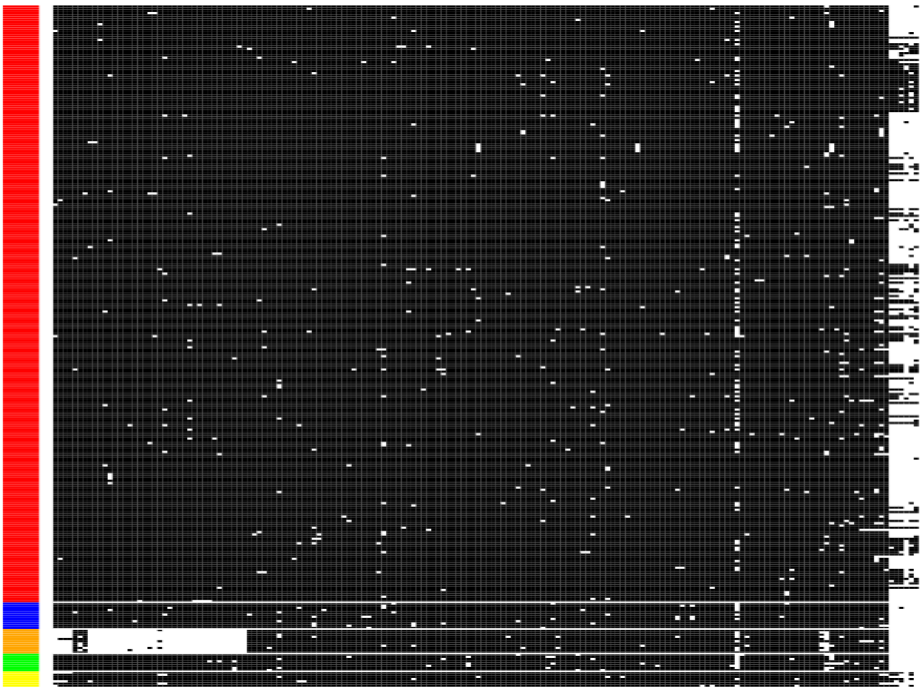


Figure 7.13 : Les 5 plus grands groupes obtenus avec une valeur du *seuil* de 0.75 et une valeur de *delta* de 0.05. Chaque groupe est indiqué par la barre de couleur à gauche et séparé par une ligne blanche. Les lignes représentent les activités enzymatiques et les colonnes les espèces (ordonnés en fonction de l'arbre phylogénétique). Une cellule noire indique la présence de l'activité enzymatique, une cellule blanche indique son absence.

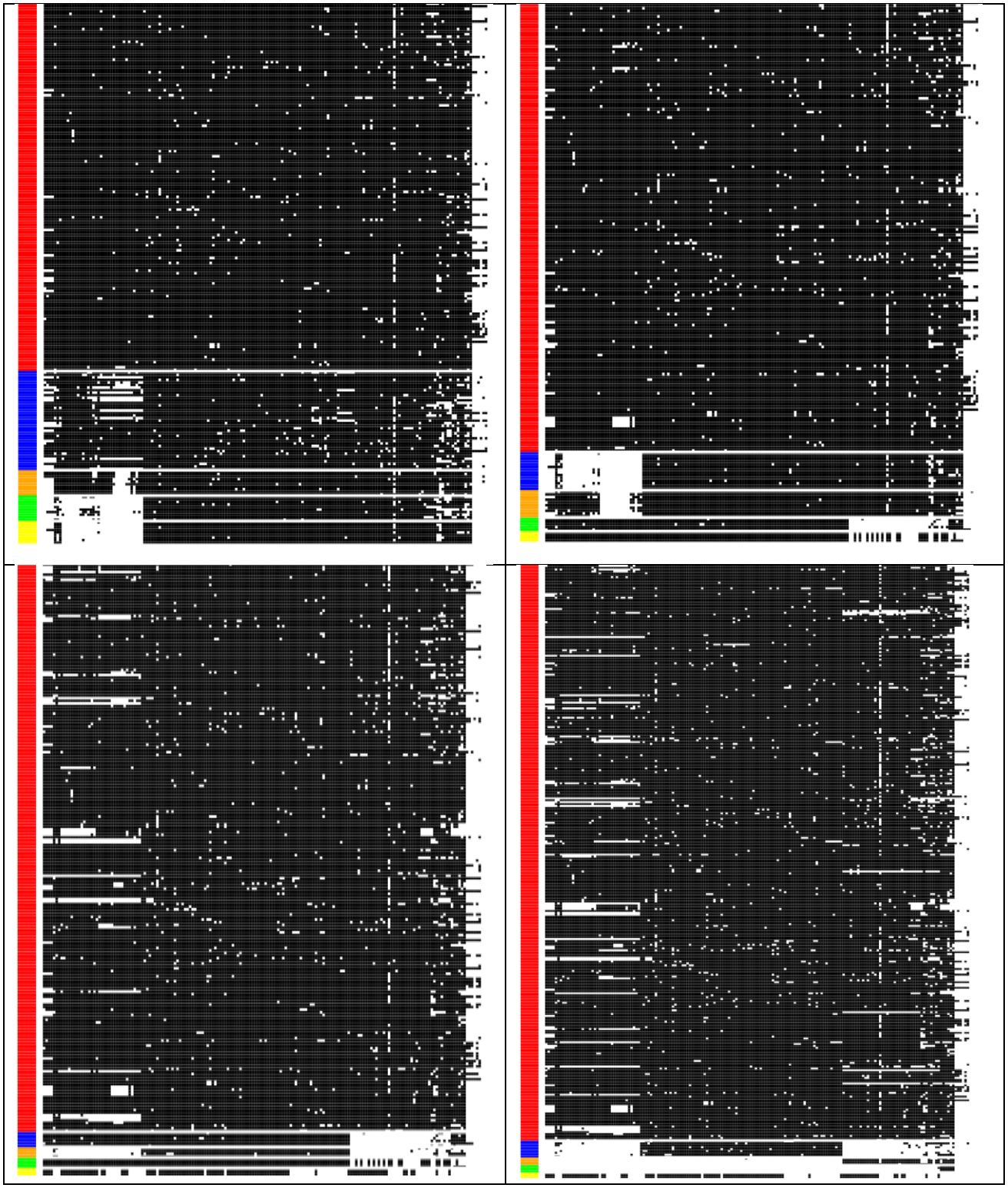


Figure 7.14 : Profil phylogénétique des 5 plus grands groupes obtenus pour un *seuil* de 0.75. De haut en bas et de gauche à droite chaque Figure représente le résultat de la classification pour une valeur de *delta* de : 0.1, 0.15, 0.2, 0.25. Chaque groupe est indiqué par la barre de couleur à gauche. Les lignes représentent les activités enzymatiques et les colonnes les espèces (ordonnés en fonction de l'arbre phylogénétique). Une cellule noire indique la présence de l'activité enzymatique, une cellule blanche indique son absence.

7.5.5 Description des résultats de la classification et visualisation des groupes obtenus

La classification obtenue avec CLAG -pour une valeur du *seuil* de 0.75 et une valeur de *delta* de 0.15 - a permis de classer 532 des 910 profils phylogénétiques. Ces 532 profils sont classés en 39 groupes distincts. Il est donc à noter que 378 activités enzymatiques n'ont pas été classées. Aucune similarité avec d'autres profils phylogénétiques n'a été détectée pour ces 378 activités enzymatiques.

Parmi tous les groupes obtenus, un groupe contient 420 activités enzymatiques. Ce groupe ne contient que des activités enzymatiques très conservées par toutes les espèces c'est-à-dire avec un niveau de conservation supérieur à 85%. En dehors de ce grand groupe, la taille des 38 autres groupes varie de 14 à 2. L'analyse des enzymes dans ces groupes nous a montré que dans 24 groupes, le groupe est composé de plusieurs activités enzymatiques qui sont portées par la même séquence protéique (la même enzyme). Cette multiple fonction peut être le fruit d'une protéine capable d'assumer plusieurs activités différentes (protéine multifonctionnelle). Par exemple les deux activités enzymatiques 2.7.4.7 (phosphooxymethylpyrimidine kinase) et 2.7.1.49 (hydroxymethylpyrimidine kinase) sont portées par une protéine bifonctionnelle (Mizote *et al.*, 1999). Cette multifonctionnalité peut être aussi due à un problème d'annotation des activités enzymatiques. Une activité enzymatique peut avoir deux EC-number différents car aucun consensus n'a encore été défini. Les deux annotations sont alors présentes dans les bases de données.

Nous avons décidé d'exclure les groupes dont toutes les activités enzymatiques ne sont portées que par une seule séquence protéique. Au final nous avons donc retenu 15 groupes de profils phylogénétiques similaires.

La visualisation des groupes de profils phylogénétiques similaires se fait sous forme de heatmap. Dans le heatmap, comme dans la matrice de profil phylogénétique, les colonnes représentent les espèces et les lignes représentent les activités enzymatiques. Les colonnes (les espèces) sont ordonnés en fonction de l'arbre phylogénétique des espèces. C'est la représentation que nous avons utilisée dans les parties précédentes pour différencier les groupes pour déterminer les paramètres adéquats pour la classification. Pour une meilleure visualisation des groupes, nous avons coloré les cellules où l'activité enzymatique est absente en fonction de la classe taxonomique des espèces. Ceci ayant pour objectif de mieux déterminer les limites taxonomiques afin de déterminer les groupes qui sont clades spécifiques (Figures 7.15 et 7.16).

Cas des *microsporidies*

La Figure 7.15 montre le plus grand groupe composé de 420 activités enzymatiques conservées dont la conservation est supérieure à 85%. Les cellules en noir indiquent la présence de l'activité enzymatique, en cas d'absence la cellule est colorée en fonction de la classe taxonomique de l'espèce (en rouge pour les microsporidies). La zone rouge sur la partie droite indique toutes les activités enzymatiques absentes chez les *microsporidies* et témoigne de la réduction maximale de son métabolisme.

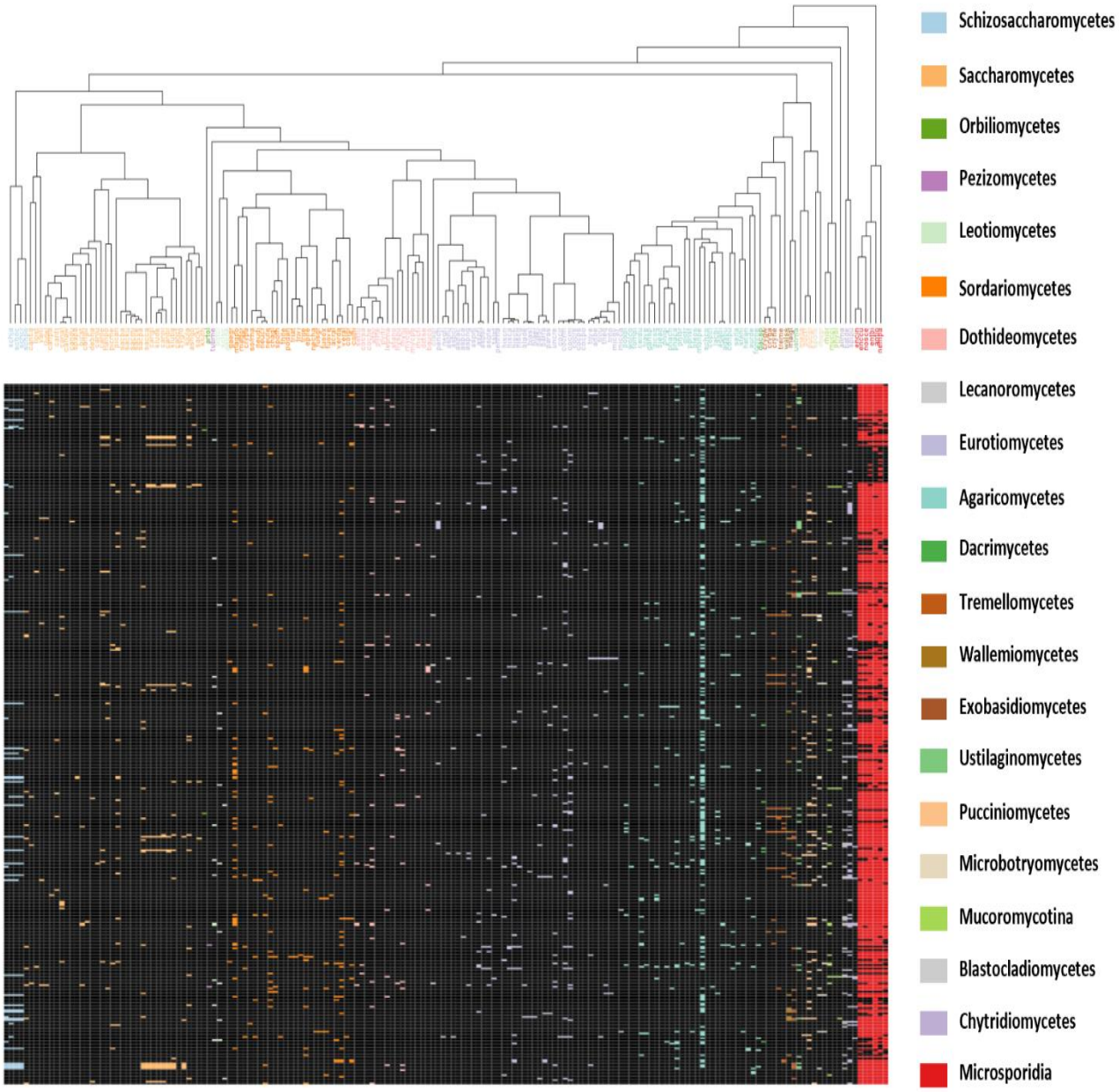


Figure 7.15: **Visualisation des profils phylogénétiques des 420 activités enzymatiques appartenant au plus grand groupe** identifié avec un *seuil* de 0.75 et un *delta* de 0.15. Chaque ligne représente une activité enzymatique, et chaque colonne représente une espèce. Les lignes ont été regroupées en fonction de la similarité des profils phylogénétiques. Les colonnes sont classées selon l'arbre phylogénétique affiché au-dessus de la matrice. Les noms d'espèces dans l'arbre sont codés en couleur selon la classe taxonomique. Dans la matrice, les cellules sont colorées en noir si l'activité enzymatique est présente, ou en fonction de la classe taxonomique si l'activité enzymatique est absente.

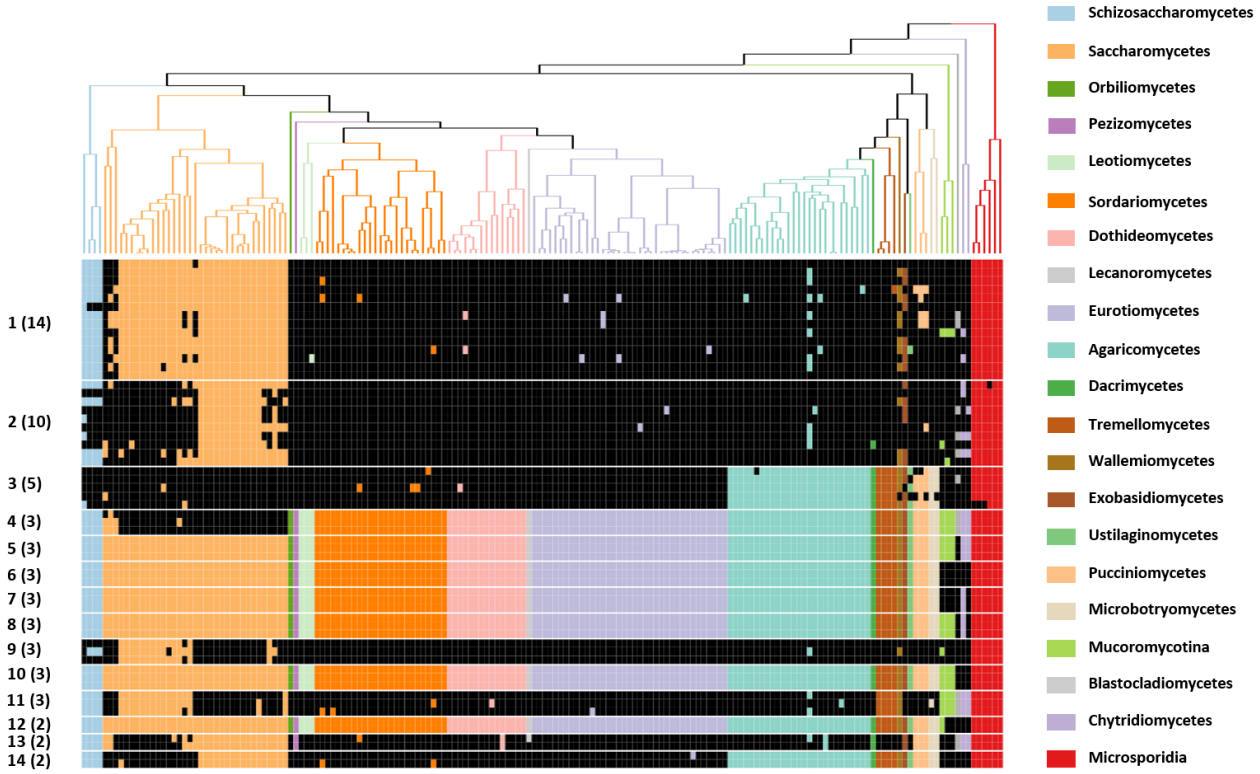


Figure 7.16 : Visualisation des profils phylogénétiques regroupés par profils similaires. Le groupe contenant 420 activités enzymatiques très conservées n'est pas montré ici. Chaque ligne représente une activité enzymatique, et chaque colonne représente une espèce. Les lignes ont été regroupées en fonction de la similarité des profils phylogénétiques. Chaque groupe de profils phylogénétiques similaire est séparé par une ligne blanche. Le numéro du groupe ainsi que la taille du groupe est indiqué à gauche. Les colonnes sont classées selon l'arbre phylogénétique affiché au-dessus de la matrice. Les branches dans l'arbre sont codées en couleur selon la classe taxonomique de l'espèce. Dans la matrice, les cellules sont colorées en noir si l'activité enzymatique est présente, ou en fonction de la classe taxonomique si l'activité enzymatique est absente. La liste des activités enzymatiques est disponible dans la table 7.2.

Analyse des autres groupes obtenus par CLAG

La liste des activités enzymatiques par groupe de profils similaires est disponible dans la table 7.2. L'ordre des groupes ainsi que l'ordre des activités enzymatiques est identique à l'ordre des activités enzymatiques dans le heatmap de la Figure 7.16. Cette table indique aussi pour chaque groupe le nombre d'activités enzymatiques présentes dans une même voie métabolique et le nom de la voie.

Le premier groupe de la Figure 7.16 (1) est un ensemble d'activités enzymatiques absentes chez presque toutes les espèces de *Saccharomycetes* et de *Schizosaccharomycetes*.
7 sur 14 des activités enzymatiques de ce groupe appartiennent à la voie du métabolisme de la valine, du leucine et de l'isoleucine (Figure 7.17 et Table 7.2). Ces activités enzymatiques forment un module évolutif. L'absence de toutes ces activités enzymatiques signifie que cette voie est très probablement absente chez les *saccharomycetes et les schizosaccharomycetes*. Plusieurs auteurs ont montré que *S.cerevisiae* ne peut pas utiliser la valine, leucine et l'isoleucine comme source de carbone pour sa croissance (Cooper, 1982; Large, 1986). En regardant le profil phylogénétique des activités enzymatiques de ce groupe, les espèces de cette classe partagent probablement toutes ce même phénotype.

4 activités enzymatiques appartenant à la voie du métabolisme du Butanoate sont présentes dans ce premier groupe composé des activités enzymatiques : 4.2.1.17, 1.1.1.157, 6.2.1.16 et 4.3.1.4. Aucune information concernant l'absence de ces activités enzymatiques chez les *Saccharomycetes* n'a été trouvée dans la littérature car les activités enzymatiques 4.1.3.4 et 4.2.1.17 sont des activités enzymatiques décrites comme présentes chez les *Saccharomycetes* (Moskowitz and Merrick, 1969; Iwahori *et al.*, 2000). La non identification de ces 2 activités enzymatiques chez cette classe est probablement du au fait que la séquence protéique chez les *Saccharomycetes* a beaucoup divergé des autres classes taxonomiques ce qui n'a pas permis son annotation.

Le deuxième groupe contient 10 activités enzymatiques, ces activités enzymatiques sont dispersées dans 10 voies différentes (Table 7.2).

Les 5 activités enzymatiques du troisième groupe de la Figure 7.14 et de la table 7.1 (2.7.1.36, 4.2.1.51, 2.5.1.16, 2.7.1.105 et 3.1.7.6) n'ont pas été détectées chez les Basidiomycètes. L'enzyme pour activité 2.7.1.36 est vraiment absente chez les *Basidiomycota* (Wilson *et al.*, 2012). Récemment, Lopez-Nieves et ses collaborateurs (Lopez-Nieves *et al.*, 2019) ont montré que l'enzyme ayant pour fonction 4.2.1.51 chez les *Basidiomycota* ne partage que 50% de similarités avec la séquence avec la même fonction présente chez les *Saccharomycetes*. Yang et ses collaborateurs (Yang *et al.*, 2023) ont montré que le gène codant l'enzyme ayant pour fonction 2.5.1.16 forme une séquence chimérique avec une autre gène chez les *Saccharomycetes*. Ces différences entre les séquences n'ont vraisemblablement pas permis d'annoter ces activités enzymatiques chez les Basidiomycètes et ont fait que des activités enzymatiques qui normalement n'ont pas co-

évolué ensemble le sont dans nos profils.
Aucune relation entre les profils et un phénotype n'a pu être déterminée à partir de la littérature avec les autres activités enzymatiques du groupe.

Dans le 6ème groupe, 1.5.1.34 et 1.14.16.1 participent à la voie du métabolisme du folate et ne sont retrouvées que chez quelques champignons inférieurs sauf chez les *microsporidie*s dans nos profils. Wang et ses collaborateurs (Wang *et al.*, 2013) ont montré que ces deux activités enzymatiques sont essentielles dans la dégradation de la phénylalanine et dans le métabolisme des lipides chez le champignon oléagineux *Mortierella alpina*, qui appartiennent aussi aux champignons inférieurs.
Deux activités enzymatiques 2.7.1.159 et 2.7.1.134 du 7ème groupe ne sont aussi retrouvées que chez les champignons inférieurs dans nos profils sauf chez les microsporidies et *Batrachochytrium dendrobatidis*. Ces deux activités enzymatiques sont très conservées chez les animaux et les plantes et ont été seulement décrites chez les champignons chez un champignon inférieurs *Funneliformis mosseae* (Campo and San Segundo, 2020).
Ces activités enzymatiques sont des activités enzymatiques très conservées chez les animaux et que les champignons inférieurs semblent avoir conservées.

Pour les autres groupes, nous n'avons pas pu faire le lien entre un phénotype et la co-conservation des activités enzymatiques dans les mêmes espèces à partir de la littérature.

La classification des activités enzymatiques par similarité de profil a permis de détecter des activités qui sont conservées ensemble et dont certaines opèrent sous la même voie. L'analyse des activités enzymatiques dans les groupes a aussi permis de détecter certains biais où des enzymes dont les séquences ont beaucoup divergé par rapport aux autres enzymes n'ont pas pu être annotées. Ceci a eu pour conséquence l'absence de l'activité enzymatique dans certaines espèces et dont le profil a été considéré comme similaire avec un autre profil même si les activités enzymatiques n'ont pas du tout évolué ensemble.

Liste des activités enzymatiques au profil similaire	**Nombre d'activités enzymatiques présentes dans une même voie**
4.2.1.17*** ; 1.1.1.157** ; 1.3.8.6 ; 6.2.1.16*** ; 4.1.3.4*** ; 2.3.3.8 ; 2.7.7.75 ; 2.10.1.1 ; 3.1.3.8 ; 1.2.4.4* ; 2.3.1.168* ; 1.8.3.6 ; 6.4.1.4* ; 1.3.8.4*	7 (Valine, leucine and isoleucine metabolism*, Figure 7.15) 4 (Butanoate metabolism**)
3.2.1.52 ; 1.4.3.3 ; 1.3.8.1 ; 1.7.3.3 ; 1.14.17.4 ; 3.1.2.22 ; 1.1.1.289 ; 2.7.1.83 ; 3.2.1.39 ; 1.1.1.31	0
2.5.1.16 ; 2.7.1.105 ; 4.2.1.51 ; 3.1.7.6* ; 2.7.1.36*	2 (Terpenoid backbone biosynthesis*)
1.11.1.9 ; 2.1.1.59 ; 1.3.1.94	0
3.5.2.7* ; 4.2.1.49* ; 1.17.4.2	2 (Histidine metabolism*)
1.5.1.34* ; 1.14.16.1* ; 1.14.19.3	2 (Folate metabolism*)
2.1.1.13 ; 2.7.1.159* ; 2.7.1.134*	2 (Inositiol phosphate metabolism*)
5.1.99.1* ; 5.4.99.2* ;1.11.1.7*	2 (Carbon fixation in prokaryotes, Glyoxylate and dicarboxylate metabolism, Propanoate metabolism, Valine, leucine and isoleucine degradation) *
5.1.1.13 ; 3.1.3.25 ; 5.1.1.11	0
3.5.1.14 ; 4.6.1.2 ; 1.1.1.88	0
4.2.1.79* ; 2.3.3.5* ; 4.1.3.30*	3 (Propanoate metabolism*)
4.1.1.9 ; 1.1.1.146	0
3.5.99.6* ; 3.5.1.25*	2 (Amino sugar and nucleotide sugar metabolism*)
1.13.11.20 ; 1.13.11.27	0

Table 7.2: Liste des activités enzymatiques par groupe de similarité de profil La première colonne indique la liste des activités enzymatiques par groupe de profils similaires. L'ordre des groupes et des activités enzymatiques correspondent à l'ordre des lignes dans le heatmap de la Figure 7.16. La deuxième colonne indique le nombre d'activités enzymatiques présentes dans une même voie ainsi que le nom de la voie. Pour chaque ligne du tableau, les étoiles indiquent quelles activités enzymatiques sont présentes dans la voie métabolique. Par exemple, les activités enzymatiques de la première ligne précédées d'une seule étoile (*) sont présentes dans la voie du métabolisme de la Valine, leucine et l'isoleucine, celles précédées de deux étoiles (**) sont présentes dans la voie du métabolisme du butanoate et celles précédées de trois étoiles (***) sont présentes dans les deux voies.

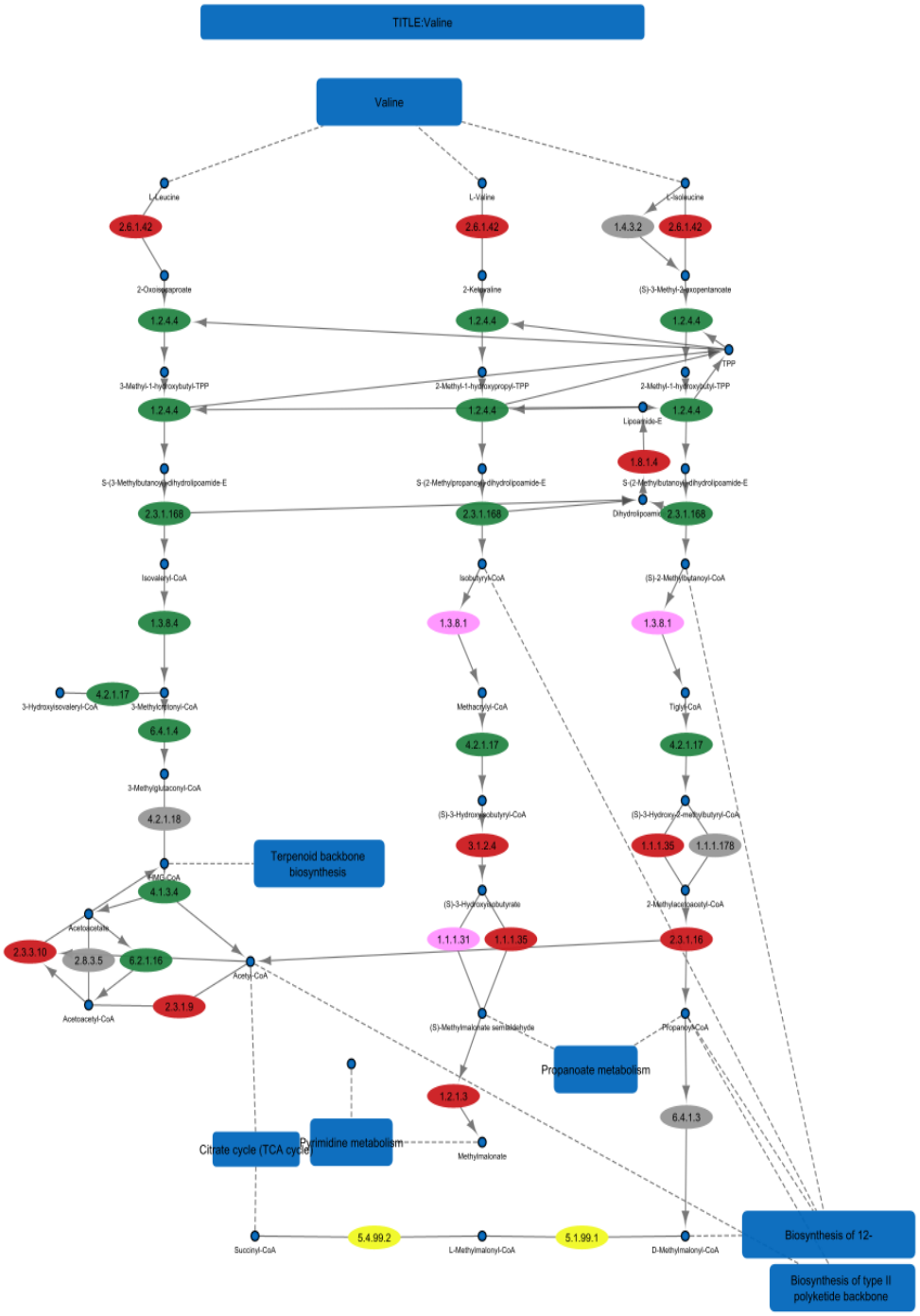


Figure 7.17 : Voie de dégradation de la valine, leucine et de l'isoleucine. Les ovales représentent les activités enzymatiques et sont colorés en fonction de leur groupe de similarité. En rouge les activités enzymatiques communes à toutes les espèces et en vert les activités enzymatiques absentes chez les *Saccharomycetes* et qui correspondent au premier groupe de la Figure 7.16.

Comment expliquer l'origine évolutive de l'absence des activités enzymatiques du premier groupe de la Figure 7.16 chez les *saccharomycetes* ?

Plus probablement, cet ensemble d'activité enzymatique a été perdu par cette classe durant l'évolution. En plus, il a été démontré qu'une perte massive de gène est survenue chez les *saccharomycetes* à la suite d'une duplication du génome suivi d'une perte massive durant l'évolution (Cliften *et al.*, 2006). L'autre hypothèse est que ces activités enzymatiques sont apparues chez les autres branches des champignons, mais cela signifierait que plusieurs évènements de gains se sont produits à différents nœuds de l'arbre phylogénétique au cours de l'évolution. Par parcimonie, la perte de ces activités enzymatiques est l'hypothèse privilégiée.

Comment expliquer l'origine évolutive d'une activité enzymatique si la présence de l'activité enzymatique est éparpillée dans l'arbre comme par exemple le dernier groupe de la Figure 7.16 ?

Dans ce groupe, les activités enzymatiques sont présentes chez une partie des *Saccharomycetes* et une partie des champignons inférieurs.

Dans ce cas, les deux hypothèses (gains et de pertes) impliqueraient plusieurs événements de pertes ou de gains et trancher sur l'origine évolutive de l'activité enzymatique ne sera pas trivial.

Chapitre :

8 Origines évolutives des activités enzymatiques

Les activités enzymatiques spécifiques présentes chez certaines espèces ou groupes d'espèces témoignent de la diversité métabolique chez les champignons. Comprendre la dynamique évolutive de ces activités enzymatiques spécifiques permettra de comprendre comment les pertes et les gains en activités enzymatiques ont façonné la diversité métabolique. Des méthodes et approches computationnelles peuvent être utilisées pour déduire les gains et les pertes des gènes associés aux activités enzymatiques.

Les modèles de Markov caché (HMM) et les modèles probabilistes peuvent être utilisés pour déduire des événements de gains et de pertes de gènes en fonction de la présence ou l'absence de gènes dans le génome. CAFE (Computational Analysis of gene Family Evolution) est un exemple de logiciel qui fonctionne avec une approche probabiliste (De Bie et al., 2006.

La phylostratigraphie (Domazet-Lošo *et al.*, 2007) est une méthode utilisée en génomique comparative pour étudier l'histoire évolutive des gènes et des familles de gènes. Elle permet de classer les gènes en différents « phylostrata » en fonction de l'âge présumé de leur origine, qui est inféré à partir de leur présence et de leur absence dans les groupes taxonomiques étudiés. C'est la méthode que nous avons décidé d'utiliser pour inférer l'histoire évolutive des enzymes associées aux activités enzymatiques.
Nous avons choisi cette méthode par rapport à la méthode probabiliste parce qu'elle se repose sur un principe très simple pour inférer l'origine évolutive d'un gène et est facile à interpréter. Le choix de cette méthode a été aussi motivé par le fait qu'un outil pour l'analyse phylostratigraphique a été développé au sein de l'équipe par Paul Roginski. Nous avons utilisé cette approche pour déterminer l'origine évolutive de chaque activité enzymatique des champignons.

8.1 La phylostratigraphie

La phylostratigraphie permet de dater l'âge de chacun des gènes présents dans le génome. L'âge d'un gène est défini en utilisant la longueur des branches dans l'arbre où l'ancêtre du gène est apparu. Ainsi, l'âge d'un gène qui est déterminé par l'approche phylostratigraphique n'est pas l'âge du dernier événement spéciation qui a fait émerger la fonction de ce gène mais plutôt la date de l'apparition de l'ancêtre de la famille de ce gène. (Domazet-Lošo *et al.*, 2007).

Pour mettre en œuvre une approche phylostratigraphique, une première étape de détection des homologues est nécessaire. Les homologues sont des gènes dont leur séquence présente une similitude. Et étant donné la dimension de l'espace des séquences possibles, une ressemblance importante entre deux gènes est interprétée comme une origine évolutive commune et non pas comme une évolution convergente.. Ce processus permet d'identifier les mêmes gènes dans les différentes espèces.
Une fois les homologues identifiés, la présence ou l'absence de chaque gène dans les

différents groupes taxonomiques peut être déduite. Sur la base de ces informations et de l'arbre phylogénétique des espèces, les gènes sont classés en différentes phylostratas. La classification la plus courante est fondée sur les principaux niveaux taxonomiques, comme les gènes propres à une espèce, les gènes présents dans un clade particulier (ex : saccharomycetes), ou un gène conservé dans un groupe plus vaste (ex : Ascomycota). Le phylostrata est défini comme le nœud de l'arbre taxonomique qui couvre toutes les espèces où un homologue du gène est retrouvé. Ce nœud est considéré comme l'origine de l'ancêtre du gène (Figure 8.1). La détermination du phylostrata permet donc de dater le moment où il y a eu l'émergence d'un nouveau gène.
A partir de la distribution des gènes dans le phylostrata, nous pouvons comprendre la dynamique évolutive de la famille du gène, en particulier les gains et les pertes au cours de l'évolution.

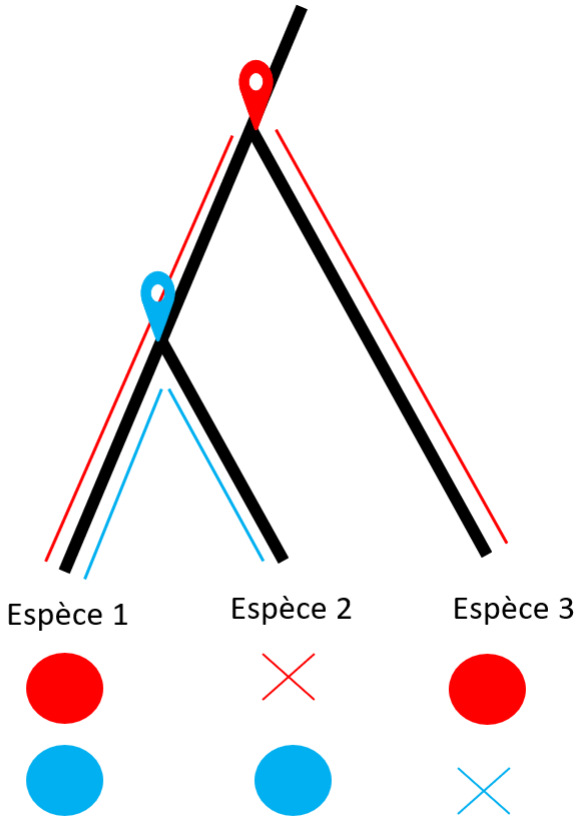


Figure 8.1 : Illustration du fonctionnement de la méthode par philostratigraphie. Le gène rouge est présent chez l'espèce 1 et l'espèce 3 mais est absent de l'espèce 2. Le nœud commun entre l'espèce 1 et 3 est le nœud rouge. C'est le nœud d'apparition du gène rouge. Par parcimonie, nous pouvons supposer que ce gène a été perdu chez l'espèce 2. Le gène bleu est présent chez l'espèce 1 et l'espèce 2 mais absent chez l'espèce 3. Le nœud commun entre l'espèce 1 et 2 est le nœud bleu. Le gène bleu est apparu au niveau de ce nœud.

8.2 Méthodes pour la phylostratigraphie

Pour déterminer l'histoire évolutive de chaque activité enzymatique, nous avons cherché à identifier les activités enzymatiques présentes chez les champignons et qui sont retrouvées en dehors des champignons (par exemple les plantes, animaux, bactéries, insectes..).

Notre idée étant que si la même activité enzymatique est retrouvée en dehors des champignons, la présence de cette activité enzymatique chez les champignons et en dehors des champignons est probablement due au fait que cette activité enzymatique a été héritée d'un ancêtre commun antérieur à l'apparition des champignons (anciennes).
Si l'activité enzymatique est retrouvée uniquement chez les champignons, cette activité enzymatique est donc une activité enzymatique qui est apparue pendant l'évolution des champignons et est spécifique des champignons. Cette activité enzymatique peut être commune à tous les champignons ou spécifique de certains clades ou d'un ensemble d'espèces de champignons.

Dans cette approche, nous ne nous intéressons pas aux activités enzymatiques qui peuvent avoir comme origine un transfert horizontal. Les transferts horizontaux ont été très bien étudiés et font partie de l'un des mécanismes à l'origine de la diversité génétique chez les bactéries. Chez les eucaryotes, il a été constaté que les transferts horizontaux sont un phénomène plus rare. Et plus particulièrement chez les champignons, il a été montré qu'au plus 2.8% des activités enzymatiques sont issues de transfert horizontal qui sont majoritairement des transferts intra-champignons (Wisecaver *et al.*, 2014).
Les cas les plus connus concernent le transfert de gènes codant pour des enzymes impliquées dans le catabolisme des xénobiotiques (Tiburcio *et al.*, 2010), la production de toxine issue d'un champignon pathogène vers une autre (Friesen *et al.*, 2006) ou un gène impliqué dans la fermentation du vin chez *Saccharomyces cerevisiae* (Novo *et al.*, 2009).

8.2.1 Détection des homologues en dehors des champignons

Pour chaque activité enzymatique, nous avons cherché les homologues de la séquence protéique qui a permis d'annoter chaque groupe d'orthologue par transfert d'annotation des EC-number à partir de la base de données NR (non-redundant database) du NCBI (téléchargé en mars 2022). Notre objectif étant de rechercher des séquences protéiques homologues en dehors des champignons, la base de données a été préalablement filtrée en enlevant toutes les séquences protéiques issues des champignons.

Nous avons 910 activités enzymatiques associées au métabolisme des champignons représentées par 3146 groupes d'orthologues. En effet, comme annoncée dans le chapitre 5, plusieurs groupes d'orthologues peuvent être associés à une même activité enzymatique (Figure 8.2). L'analyse de la distribution du nombre de groupes d'orthologues par EC-number montre que 305 activités enzymatiques sont représentées par un seul groupe

d'orthologues et 605 activités enzymatiques sont représentées par plus d'un groupe d'orthologues (Figure 8.2).
Parmi les 305 activités enzymatiques qui sont représentées par un seul groupe d'orthologues, 87 sont conservées (conservation supérieure à 85%) et 218 sont conservées que par certaines espèces (conservation inférieure à 85%).
Dans les activités enzymatiques représentées par plus d'un groupe d'orthologues, 369 sont fortement conservées et 236 sont espèces spécifiques.

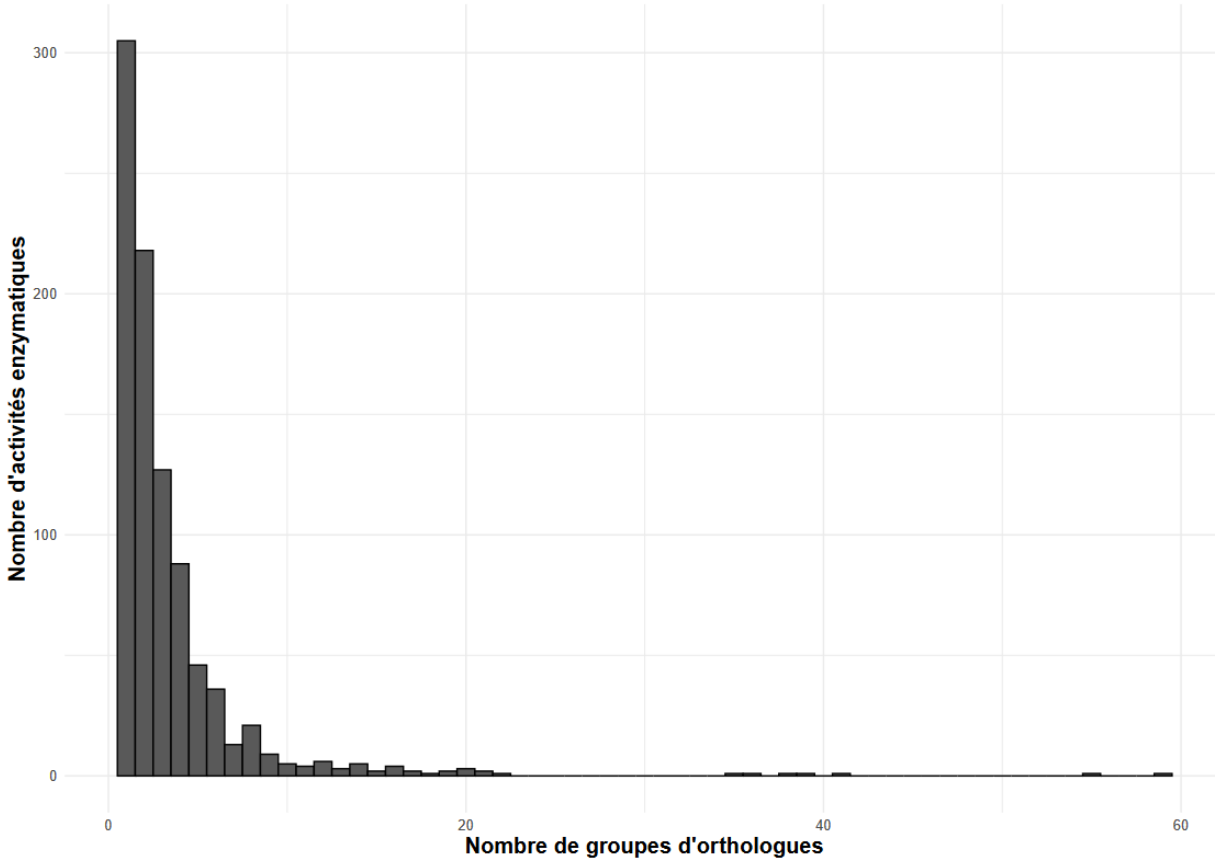


Figure 8.2 : Occurrence des activités enzymatiques dans les groupes d'orthologues. Chaque barre correspond à une intervalle de 1 et correspond au nombre d'activités enzymatiques associés au nombre de groupes d'orthologues auquel elles appartiennent.

Pour la recherche d'homologue en dehors des champignons nous avons utilisé le logiciel Diamond (Buchfink *et al.*, 2015). Comme BLAST (Altschul *et al.*, 1990) Diamond permet de comparer une séquence contre une base de données. L'avantage de Diamond par rapport à BLAST, réside dans sa rapidité quand il faut traiter d'énormes jeux de données. Malgré sa rapidité sur de gros jeux de données, Diamond ne perd pas en sensibilité (Buchfink *et al.*, 2015).
Dans la détection d'homologues, nous avons fixé comme critères une couverture d'au moins 70% sur la taille de la protéine requête et une e-value inférieure à 10-3. Ces valeurs ont été choisies pour pouvoir récupérer le plus de séquences homologues possibles et

étant donné l'échelle d'étude entre les champignons et les autres règnes du vivant, les séquences peuvent avoir divergé tout en gardant une certaine similarité pour assurer les mêmes fonctions.

Sur les 910 activités enzymatiques, nous trouvons au moins un homologue en dehors des champignons pour 907 activités enzymatiques.

Aucun homologue n'a été retrouvé pour les séquences protéiques correspondant aux activités enzymatiques : 2.1.1.59, 2.5.1.34, and 4.1.1.36.

En effet, 2.5.1.34 participe à la biosynthèse des ergots alkaloid et est seulement retrouvé chez certains champignons (Lee *et al.*, 1976). 4.1.1.36 est une enzyme ubiquitaire qui a été identifiée à la fois chez les eucaryotes et les procaryotes mais dont la séquence primaire chez les champignons est très différente de la séquence retrouvée par exemple chez les plantes. La taille de la protéine fait au moins 500 acides aminés chez les champignons alors que chez les plantes l'enzyme ne fait que 200 acides aminés. Ces 200 acides aminés constituent le site actif de l'enzyme qui est très conservé entre les champignons et les plantes (Petrényi *et al.*, 2016). 2.1.1.59 est seulement présent chez certains eucaryotes (plantes et les champignons) mais absent chez les animaux supérieurs (Polevoda *et al.*, 2000). Pourtant aucune séquence n'a été trouvée en dehors de champignons, ce qui semble suggérer que les séquences entre les plantes et les champignons ne présentent aucune homologie de séquence.

8.2.2 Annotation des homologues en EC-number

L'un des mécanismes principaux pour faire émerger une nouveauté fonctionnelle est la duplication divergence d'un gène suivie d'une divergence de séquences entre les deux copies du gène dupliqué. Des activités enzymatiques similaires peuvent donc partager une similarité de séquence ou des caractéristiques structurelles parce qu'elles ont évolué à partir d'un ancêtre commun ayant une activité catalytique similaire.
Par exemple, les enzymes dont l'EC-number commence par 2.7.1 sont des kinases qui catalysent le transfert d'un groupe phosphate de l'ATP vers un substrat. De nombreuses activités enzymatiques dans cette classe partagent un domaine conservé appelé domaine kinase, qui contient le site actif. Ce domaine permet d'identifier et de classer les kinases en fonction de leur séquence et de leur similarité structurelle (Hanks and Hunter, 1995).

Par conséquent, un homologue détecté en dehors des champignons peut être une séquence protéique avec une activité enzymatique différente. Afin de se prémunir de ces similarités de séquence entre les activités enzymatiques proches et d'identifier plus précisément l'origine de l'activité enzymatique et non l'origine de la séquence protéique, nous avons transféré un EC-number aux 100 premiers homologues trouvés en dehors des champignons (avec les 100 meilleurs scores) pour chacun des groupes d'orthologues.
Pour ce faire, nous avons utilisé la base de données Uniprot qui contient une base de

données de toutes les séquences protéiques ayant une activité enzymatique (avec l'EC-number associé). Pour les 100 premiers homologues trouvés en dehors des champignons, nous avons identifié la séquence la plus proche dans la base Unitprot en respectant une couverture de 70% avec la séquence requête et un e-value inférieur à 10-3, et nous avons transféré son EC-number à la séquence homologue. La séquence la plus proche est la séquence avec le meilleur score (Figure 8.3).

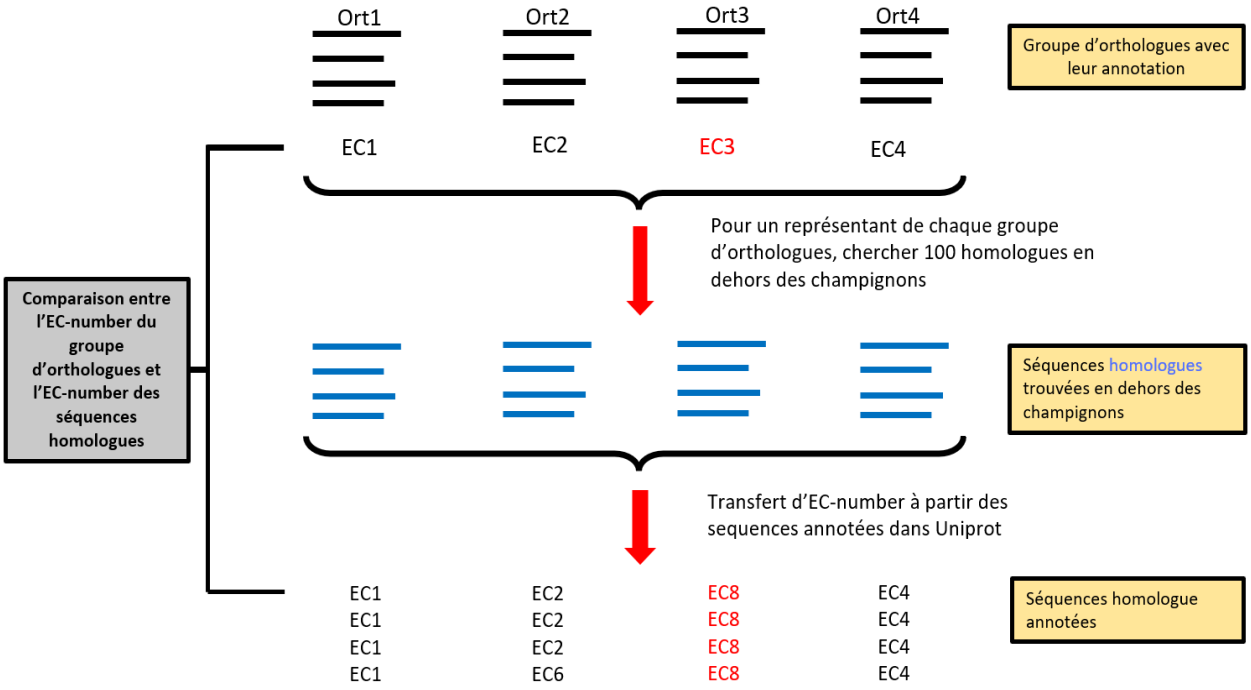


Figure 8.3 : Méthode de datation des activités enzymatiques. Avec un élément représentatif de chaque groupe, 100 séquences homologues sont recherchées en dehors des champignons. Pour chaque séquence homologue, l'EC-number des séquences qui sont annotées dans Uniprot est transféré si les critères d'homologie sont respectés (% de couverture > 70% et e-value < 1e-3). Les EC-number entre les homologues trouvés en dehors des champignons et celui du groupe d'orthologues de champignons sont alors comparés. Si aucun homologue ne possède le même EC-number que le groupe d'orthologue (groupe ort3 par exemple), l'activité enzymatique associée à ce groupe est considérée comme une activité spécifique des champignons.

Pour chaque groupe d'orthologue après l'étape d'annotation, au moins un homologue en dehors des champignons a été associé avec un EC-number.
Si l'EC-number associé au groupe d'orthologues est identique à l'EC-number de l'un des 100 homologues, l'activité enzymatique est considérée comme **ancestrale**. Car la présence de l'activité enzymatique chez les champignons et en dehors des champignons signifie que l'activité enzymatique a été héritée d'un ancêtre commun.
Sinon l'activité enzymatique est considérée comme une activité enzymatique **spécifique** et

qui est apparue seulement chez les champignons au cours de l'évolution.

8.3 Résultat de la phylostratigraphie

Comme indiqué ci-dessus, si l'EC-number associé à un groupe d'orthologues de champignons est retrouvé à l'identique chez l'un des homologues obtenus en dehors des champignons, cela signifie que la même activité enzymatique est retrouvée en dehors des champignons. D'un point de vue évolutif, l'activité enzymatique était forcément présente au niveau de l'ancêtre commun (ancienne), et l'absence dans certaines lignées indique alors une perte de ces activités enzymatiques (ancienne spécifique).
Si l'EC-number associé au groupe d'orthologues des champignons n'est retrouvé pour aucun des homologues identifiés en dehors des champignons, alors l'activité enzymatique chez les champignons sera considérée comme une nouvelle activité enzymatique qui est apparue durant l'évolution des champignons et est donc spécifique des champignons (nouvelle spécifique).

La phylostratigraphie des activités enzymatiques nous a permis d'inférer que 860 activités enzymatiques sont d'origine ancestrale (pré-date l'apparition des champignons), dont 445 sont très conservées chez toutes les espèces (anciennes conservées) et 415 ont été perdues par certains clades ou espèces durant l'évolution (anciennes spécifiques).
Nous avons aussi identifié 50 activités enzymatiques dont les homologues en dehors de champignons ont des activités enzymatiques différentes. Ces activités enzymatiques sont considérées comme des activités enzymatiques spécifiques des champignons qui sont apparues (nouvelles spécifiques) au cours de l'évolution des champignons (Table 8.1).
D'après ce résultat, l'évolution du réseau métabolique chez les champignons a été majoritairement impactée par des pertes d'activités enzymatiques anciennes au cours de l'évolution.

1.1.1.39 ; 1.14.19.3 ; 1.2.1.75 ; 1.3.1.10 ; 1.4.1.13 ; 1.6.1.2 ; 2.1.1.1 ; 2.1.1.261 ; 2.5.1.31 ; 2.5.1.83 ; 2.7.1.91 ; 2.7.4.12 ; 2.7.4.22 ; 2.8.1.3 ; 3.5.4.26 ; 3.6.1.15 ; 3.6.1.17 ; 4.1.1.36 ; 4.1.1.53 ; 4.2.1.55 ; 6.3.5.6 ; 1.1.1.10 ; 1.1.1.188 ; 1.1.1.21 ; 1.1.1.267 ; 1.1.1.51 ; 1.1.1.81 ; 1.1.3.5 ; 1.1.5.5 ; 1.11.1.9 ; 1.14.13.178 ; 1.14.13.82 ; 1.3.99.5 ; 2.1.1.121 ; 2.1.1.59 ; 2.3.1.85 ; 2.5.1.34 ; 3.5.99.2 ; 4.2.1.130 ; 4.2.1.142 ; 4.2.1.66 ; 4.2.1.92 ; 4.2.2.3 ; 4.2.3.23 ; 5.4.99.32 ; 1.1.3.13 ; 1.1.5.9 ; 4.1.1.46 ; 4.2.1.143 ; 4.2.3.43

Table 8.1 : Liste des activités enzymatiques inférées comme spécifiques et nouvelles des champignons par homologie.

8.4 Limite de l'approche phylostratigraphique

Pour déterminer les activités enzymatiques présentes en dehors des champignons, notre méthode s'appuie principalement sur la détection de séquence homologue en dehors des champignons puis l'annotation de celle-ci.
Cette méthode présente cependant quelques limites. Si la séquence chez les champignons est très divergente, aucun homologue ne sera détecté ou l'homologue le plus proche de la séquence chez les champignons est une séquence avec une activité différente. Par conséquent l'enzyme sera annotée comme une nouvelle activité enzymatique spécifique des champignons.
Les problèmes d'annotations peuvent aussi être source d'erreurs. Les groupes d'orthologues Identifiés avec MARIO ont été annotés avec la version 2012 de Uniprot. Alors que les homologues en dehors des champignons ont été annotés avec la version de 2022. Des nouvelles séquences ont été ajoutées ainsi que l'activité enzymatique (EC-number) associée à une séquence a pu être modifiée. Si les annotations ont changé entre-temps ou l'enzyme en dehors des champignons est annotée différemment, l'activité enzymatique sera inférée comme une activité spécifique des champignons.
Des activités enzymatiques identifiées comme ancestrales peuvent ne pas l'être parce que l'homologue trouvé en dehors des champignons présente une ressemblance du fait de l'e-value utilisée qui est assez permissive. En outre si la séquence homologue en dehors des champignons n'a jamais été annotée, elle sera annotée avec la séquence la plus proche qui est la séquence chez les champignons. Par conséquent, la séquence chez les champignons et en dehors des champignons aura la même annotation.

L'analyse manuelle des activités enzymatiques inférées comme nouvelles et spécifiques des champignons a confirmé nos craintes sur la limite de notre méthode. Sur les 50 activités enzymatiques, en vérifiant la littérature, 42 activités enzymatiques que nous avons inférées comme spécifiques des champignons sont mentionnées dans les autres règnes du vivant. Et 8 activités enzymatiques ont été rapportées par la littérature comme des activités enzymatiques vraiment spécifiques des champignons (table 8.2).
Comme indiqué précédemment, la non détection d'une séquence homologue en dehors des champignons avec la même activité enzymatique peut être due à une erreur d'annotation ou une grande divergence des séquences homologues. Cela peut être aussi causé par une convergence fonctionnelle de deux séquences différentes où deux séquences indépendantes ont convergé vers une même fonction.
Par exemple l'Enoyl-ACP Reductases (1.3.1.10) est une activité enzymatique essentielle et ubiquitaire. Elle est retrouvée dans les 3 domaines du vivant. Il a été montré que l'organisation et la structure de l'enzyme impliquée diffèrent entre les mammifères, les plantes, les champignons et les bactéries (Massengo-Tiassé and Cronan, 2009). Cette diversité est probablement le fruit d'une pression évolutive différente dans les différents organismes qui a provoqué une grande divergence entre les différentes séquences pour s'adapter à l'environnement de chaque groupe.
Parmi les 42 activités enzymatiques dont l'homologue n'a pas la même activité enzymatique

mais dont l'activité enzymatique a été rapportée par la littérature comme présente chez les champignons, nous n'avons aucun moyen de différencier quelles sont les séquences issues d'une divergence, convergence et les erreurs d'annotations.
Par conséquent, nous avons considéré que l'origine évolutive de ces 42 activités enzymatiques n'est pas résolue.
Les **8 activités enzymatiques** qui sont nouvelles et spécifiques chez les champignons d'après la littérature sont considérées comme des activités enzymatiques qui sont apparues au cours de l'évolution chez les champignons. Ces activités enzymatiques sont impliquées soit dans la synthèse de métabolites secondaires par exemple les activités enzymatiques 4.2.1.143 et 4.2.1.142 qui sont impliquées dans la synthèse de l'aflatoxine soit dans la dégradation d'un substrat impliqué dans le métabolisme secondaire, par exemple l'activité 4.2.1.66 qui est impliquée dans la dégradation de la cyanide (Table 8.2).

EC-number	**Mention dans la littérature**	**Voie impliquée**
5.4.99.32	(Kimura *et al.*, 2010)	Antibiotic production
4.2.1.142	(Sakuno *et al.*, 2005)	Aflatoxine production
4.2.1.143	(Ren *et al.*, 2017)	Aflatoxine production
4.2.1.66	(Martínková *et al.*, 2015)	Cyanide degradation
2.1.1.261	(Wallwey *et al.*, 2012)	Ergot alkaloid biosynthesis
2.5.1.34	(Ding *et al.*, 2008)	Ergot alkaloid biosynthesis
4.2.3.43	(Toyomasu *et al.*, 2007)	Fusicoccin A production
1.1.3.13	(Westrick *et al.*, 2022)	Methane métabolisme

Table 8.2 : Liste des activités enzymatiques nouvelles et spécifiques des champignons qui sont confirmées par la littérature.

III. Construction du réseau métabolique

À ce stade nous avons les informations évolutives de chaque activité enzymatique. Cependant une activité enzymatique ne confère pas un avantage physiologique à elle seule . Chaque activité enzymatique est le maillon d'une chaîne qui assure la formation ou la dégradation de métabolites spécifiques. Elle fait partie d'une voie métabolique ou un réseau plus large pour travailler de manière coordonnée avec d'autres activités enzymatiques. Avec les profils phylogénétiques, nous n'avons pas la relation entre les activités enzymatiques entre elles.

Nous avons pu observer la conservation des activités enzymatiques à travers les 174 espèces de champignons ainsi que leur histoire évolutive. L'analyse par génomique comparée a permis de comprendre comment l'environnement a façonné la quantité d'activités enzymatiques (par exemple chez les microsporidies et les *eurotiales*) mais aussi comment les mécanismes à l'origine de la dynamique des génomes ont façonné le répertoire enzymatique de certaines espèces (cas des *Saccharomycetes*).
Pour obtenir la relation entre les activités enzymatiques, nous allons cartographier l'information évolutive sur les nœuds du réseau métabolique afin de comprendre si la topologie du réseau métabolique joue un rôle dans l'évolution des sommets (les activités enzymatiques).
Pour ce faire, nous avons d'abord construit un réseau métabolique global. Mais avant de construire ces réseaux, nous avons inféré les voies métaboliques présentes chez les champignons.
Une fois le réseau métabolique construit, les informations évolutives ont été cartographiées sur ce réseau afin de comprendre les contraintes exercées par ce réseau sur l'évolution des sommets.

La construction du réseau métabolique est un domaine encore récent. L'analyse du réseau métabolique est en constante évolution ces 10 dernières années du fait du nombre de génomes séquencés et annotés en constante augmentation grâce aux techniques de séquençage nouvelle génération. La construction du réseau métabolique et des voies métaboliques ont été surtout effectuées manuellement à partir des connaissances issues de la littérature et principalement à partir des organismes modèles (par exemple les voies dans KEGG et MetaCyc). Ces connaissances ont été par la suite stockées dans des bases de données (https://www.kegg.jp/ et https://metacyc.org/) pour être mises à la disposition de la communauté.
Depuis 20 ans, des outils qui permettent la construction semi-automatique du réseau métabolique d'un organisme ont été développés (Table 9.1). En conséquence, des centaines de voies métaboliques ont été reconstruites sous forme de modèle de prédiction pour les organismes dont le génome a été séquencé (Orth *et al.*, 2011; Thiele *et al.*, 2005). Ces reconstructions ont pour but de comprendre la structure des différents processus cellulaires et leurs fonctions (Radrich *et al.*, 2010).

CARMEN (Schneider *et al.*, 2010)	Comparative analysis and reconstruction of metabolic networks	http://carmen.cebitec.uni-bielefeld.de/cgi-bin/index.cgi
Pathway Tools (Karp *et al.*, 2002)	Genome annotation for an organism and infers probable metabolic reactions and pathways	http://bioinformatics.ai.sri.com/ptools/
ERGO (Overbeek *et al.*, 2003)	Genome annotation and metabolic analysis	https://www.igenbio.com/ergo
CoReCo (Pitkänen *et al.*, 2014)	Metabolic networks construction of multiple species from protein sequence and phylogenetic data	https://github.com/esaskar/CoReCo

Table 9.1 : Exemple d'outils qui permettent de reconstruire le réseau métabolique d'un génome annoté.

Ces différents outils sont basés sur le même principe pour la reconstruction du réseau métabolique : (1) Annotation du génome, (2) comparaison des réactions et des enzymes avec les bases de données de référence telles que MetaCyc et KEGG, (3) construction des voies et du réseau en se basant sur les représentations qui ont déjà été faites manuellement (Figure 9.1).

Chez les 174 espèces étudiées ici, l'étape d'annotation et d'identification des enzymes dans chaque espèce a déjà été effectuée à partir de FungiPath. La matrice de profil phylogénétique indique déjà quelles sont les activités enzymatiques présentes dans nos espèces étudiées. Pour construire le réseau métabolique global chez les champignons ou représenter une voie, il faut donc positionner ces activités enzymatiques sur un réseau métabolique déjà construit.

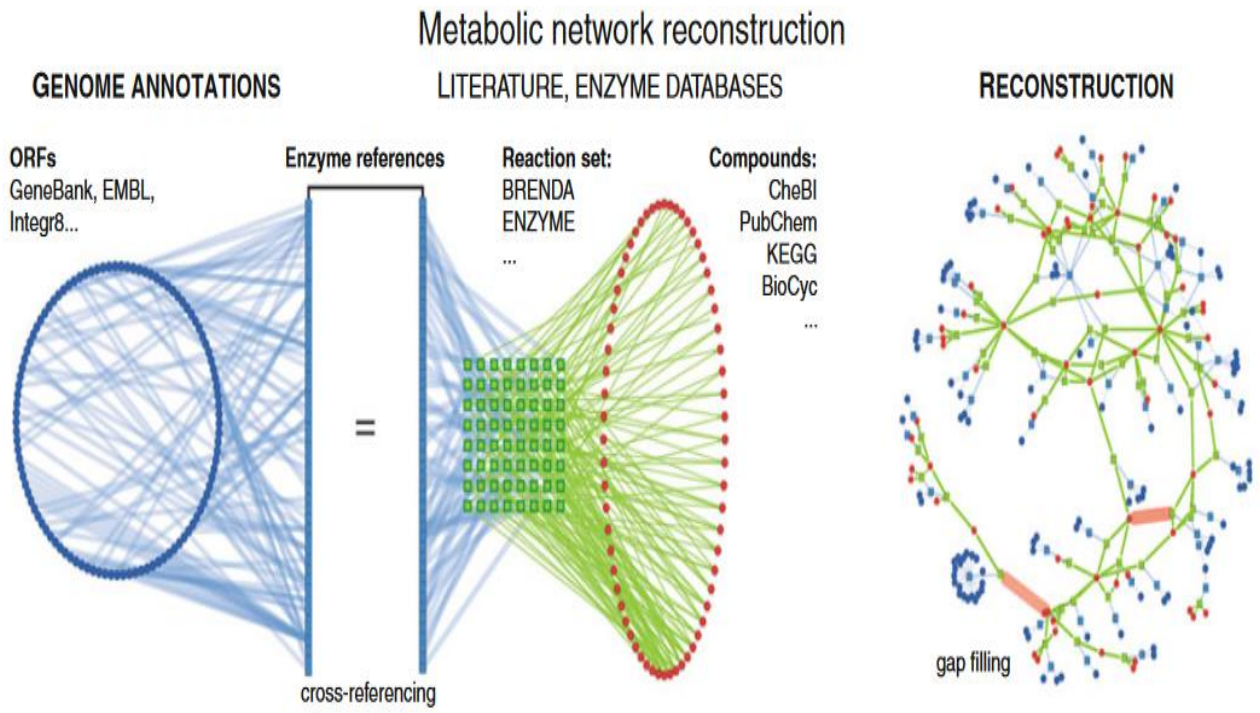


Figure 9.1 : Méthode de reconstruction du réseau métabolique. Les éléments annotés du génome sont comparés avec les données dans les bases de données publiques. Le réseau métabolique est construit en fonction des éléments en commun entre le génome et la base de données. Figure tirée de (Chalancon *et al.*, 2013).

La base de données KEGG met à disposition de ses utilisateurs un aperçu global du réseau métabolique. Dans ce réseau KEGG, les sommets représentent les métabolites et les liens entre les sommets représentent les réactions qui transforment un métabolite en un autre métabolite. Dans ce projet, nous voulons que les activités enzymatiques représentent les sommets et que les métabolites en commun entre les réactions enzymatiques représentent les arêtes car nous nous intéressons principalement à l'évolution des activités enzymatiques dans le réseau métabolique. Une première solution pour obtenir notre réseau c'est d'intervertir les sommets et les arêtes du réseau KEGG. Or le réseau global disponible dans KEGG (Figure 9.2) est une image où chaque élément du réseau (les sommets et les arêtes) est affiché en fonction de leurs coordonnées dans un fichier, par conséquent non exploitable pour une analyse computationnelle.

Pour construire le réseau métabolique global multi-espèces afin d'y extraire le réseau métabolique global des champignons, il faut commencer par connecter entre elles les voies métaboliques qui sont disponibles dans les bases de données comme KEGG et MetaCyc. Ces voies métaboliques sont exploitables de manière computationnelle. Leur analyse et leur manipulation peuvent être effectuées automatiquement. Pour construire le réseau métabolique global des champignons, il faut donc commencer par sélectionner les voies métaboliques présentes chez les champignons. Et pour construire le réseau métabolique

global des champignons, il faut commencer par prédire les voies présentes chez les champignons.

Le réseau métabolique global obtenu va nous permettre d'étudier la position et la relation des activités enzymatiques entre elles. Afin de construire ce réseau métabolique, j'ai développé un script python qui permet de connecter les voies métaboliques entre elles. Ce script permet de construire n'importe quel réseau métabolique en fonction des voies qui le constituent.

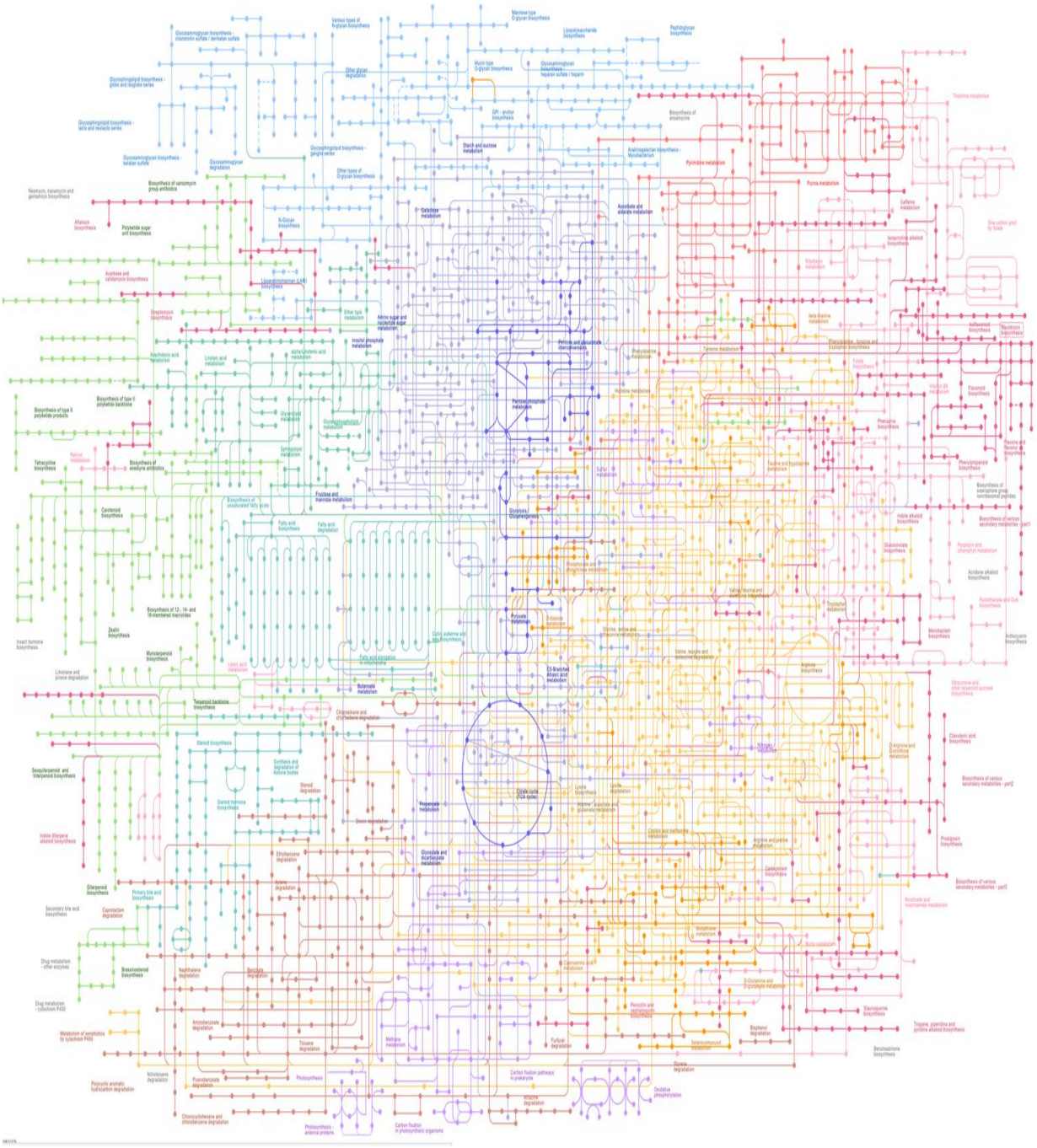


Figure 9.2 Représentation du réseau métabolique dans KEGG. Les sommets représentent les métabolites. Les liens entre les métabolites représentent la réaction qui permet de convertir un métabolite en un autre métabolite. Les voies métaboliques sont colorées en fonction de la nature du type de métabolites impliqués dans la voie.

Chapitre :

9 Comparaison entre KEGG pathways et MetaCyc d'un point de vue computationnel

Avant de construire un réseau basé sur l'agglomération de voies métaboliques, il faut se poser la question de la délimitation d'une voie métabolique. Par exemple, quel est le début et quelle est la fin d'une voie métabolique ? Ce que nous observons dans la pratique, c'est que chaque base de données définit les voies métaboliques en fonction de l'objectif de la base de données, par exemple le processus métabolique que la base de données veut illustrer.

Les deux bases de données de voies métaboliques les plus importantes sont KEGG et BioCyc. KEGG et BioCyc sont des collections de bases de données qui sont connectées entre elles. Dans KEGG, les informations sur les voies métaboliques sont stockées et organisées dans la base de données « Pathways » et dans « MetaCyc » pour BioCyc. Une comparaison sur le contenu de ces deux bases de données a été faite dans la partie introductive de ce manuscrit.

Les voies métaboliques de KEGG Pathways sont surtout destinées à faciliter la manipulation et la visualisation des voies alors que MetaCyc vise à centraliser et à organiser toutes les informations concernant les différents éléments d'une voie. Par exemple la voie de la Glycolyse ou *voie* d'Embden-Meyerhof-Parnas, dont le principal objectif est de générer des précurseurs métabolites et de l'énergie par assimilation du glucose n'est pas délimitée de la même manière dans KEGG et MetaCyc (Figure 9.3 et Figure 9.4). En effet, plusieurs voies de dégradations convergent à différents endroits vers la glycolyse, et il y a une différence dans les réactions utilisées entre différents organismes. Dans KEGG, cette voie est représentée par une seule entrée dans la base de données qui représente la voie globale issue de la collection des voies de plusieurs espèces. Dans MetaCyc, il y a 6 niveaux en fonction des points d'entrées. Ces différents niveaux sont des ensembles qui sont présents dans plusieurs espèces et ont été vérifiés expérimentalement. Plusieurs représentations globales appelées « superpathways » sont également mises à disposition dans MetaCyc. Ces « superpathways » permettent seulement une vision globale des différents niveaux et ne représentant pas la voie dans une espèce. KEGG recense 157 voies reliées au métabolisme contre 246 dans MetaCyc.

Dans MetaCyc, les voies sont téléchargeables au format BioPax. Biopax est un format standard RDF/OWL (Ressource Description Framework/ Web ontologie Language). Biopax est une combinaison de deux manières de représentation des connaissances qui sont utilisées pour décrire et modéliser des données sémantiques sur le web et même plus (par exemple ici pour décrire des voies métaboliques). RDF permet de modéliser des graphes des données sémantiques de manière structurée avec un ensemble de règles permettant des informations descriptives simples où les ressources sont reliées par des relations pour représenter des informations complexes et interconnectées. OWL est un langage de représentation des ontologies qui décrivent les concepts et les relations entre les concepts construits sur le modèle de données RDF. OWL permet de définir des ontologies qui précisent la structure des données (relations entre les éléments) pour ajouter plus de vocabulaire dans la description des données de façon formelle. Ensemble, les deux formats

permettent de créer des modèles de connaissances riches qui permettent l'intégration, l'échange, la visualisation et l'analyse des données (https://www.w3.org/).

KEGG utilise un format dérivé du XML pour modéliser les voies métaboliques : le format KEGG Markup Langage (KGML). C'est un format à balise destiné à faciliter la manipulation computationnelle des voies. Dans ce format, le format XML explique comment les boîtes bleues (enzymes) sont liées par des « relations » et comment les cercles (composés chimiques) sont liés par des « réactions ». Cette relation entre les enzymes facilite grandement la représentation et l'analyse de la voie sous forme de graphe (Figure 9.3).

Du fait de la simplicité et du côté pratique du format KGML pour manipuler et d'analyser de manière computationnelle la représentation de la voie, la base de données KEGG est plus pratique pour analyser et manipuler le réseau métabolique d'un point de vue topologique. De plus, le format KGML contient toutes les informations nécessaires pour représenter la relation entre les activités enzymatiques.

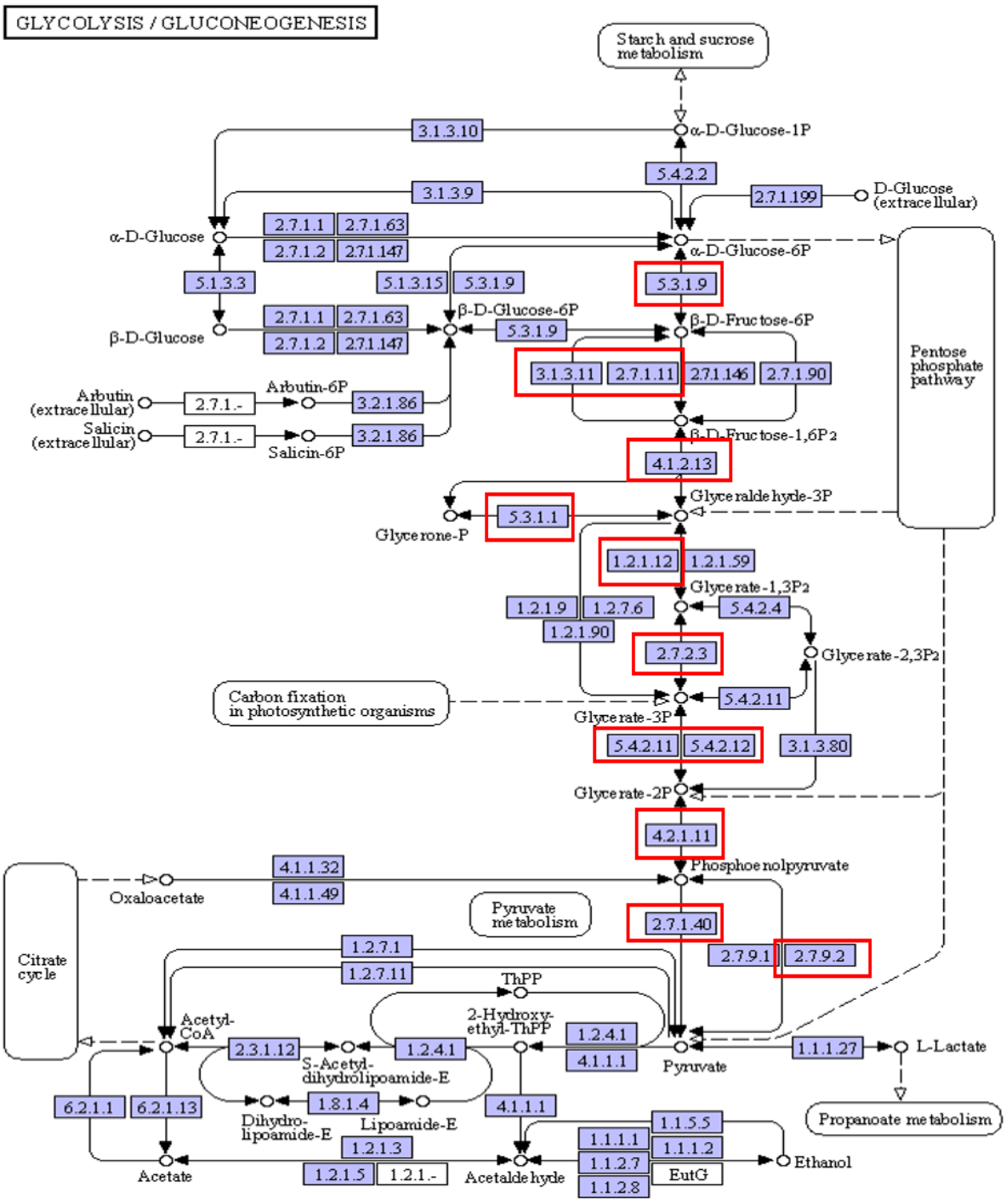


Figure 9.3 : Représentation de la voie de la Glycolyse dans KEGG. Les rectangles bleus (réactions) correspondent à l'activité enzymatique qui catalyse la réaction, les cercles blancs correspondent aux composés chimiques. Les flèches discontinues indiquent les composés en commun entre deux voies. Les activités enzymatiques encadrées en rouge sont les activités enzymatiques présentes dans la représentation de la voie de la glycolyse (from fructose 6-phosphate) de MetaCyc (Figure 9.4).

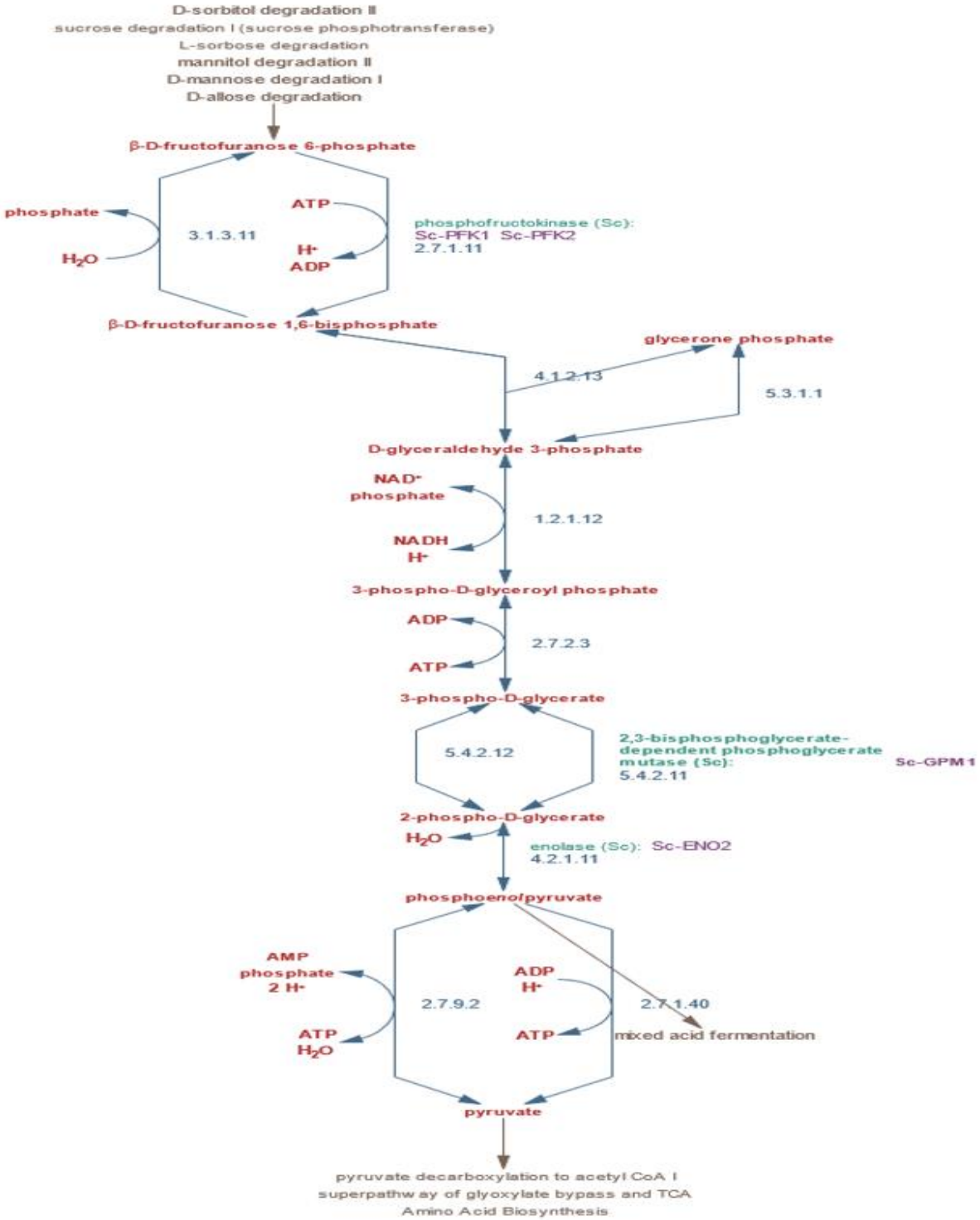


Figure 9.4 : Représentation de la voie de la Glycolyse I (from fructose 6-phosphate) dans MetaCyc. Les composés sont indiqués en rouge et les activités enzymatiques en bleu. En marron, les autres voies de MetaCyc qui sont connectées avec cette partie de la voie de la glycolyse. Les activités enzymatiques retrouvées dans la représentation de la voie de la glycolyse de KEGG sont encadrées en rouge dans la Figure 9.3.

Le format KGML se décrit comme suit (Figure 9.5) :

- Les sommets du graphe sont balisés « entry ». Il y a 3 types (attribut *type*) de sommets : *enzyme, compound* et *map* balisé « entry ». Chaque balise « entry » est identifiée par un identifiant (*id*) unique. Pour un « entry » de type *enzyme*, on a un attribut *name* et *reaction* qui donne le EC number ainsi que l'identifiant de la réaction dans KEGG catalysé par cette enzyme. Dans le cas d'un « entry » de type *compound*, l'attribut *name* contient l'identifiant du composé dans KEGG. Un sommet de type *map* indique une voie métabolique reliée à la voie. Il contient le nom de la voie métabolique ainsi que son identifiant dans KEGG.
- Les balises « reaction » donnent plus de détails sur les réactions catalysées par chaque enzyme. Plus précisément elles contiennent l'identifiant des composés de la réaction (le substrat et le produit) ainsi que le sens de la réaction si connu (réversible ou irréversible). Elles ont le même identifiant unique (*id*) que l'enzyme qui catalyse la réaction.
- Les balises « relation » indiquent la relation entre les enzymes. Le produit d'une réaction peut être le substrat d'une autre réaction. Ainsi chaque relation indique deux réactions qui partagent un composé en commun et le composé en question. Cette balise permet aussi d'indiquer la connexion avec d'autres voies. Ceci se matérialise à partir d'un composé en commun entre les deux voies.

Figure 9.5 : Exemple de fichier KGML avec deux enzymes qui catalysent deux réactions successives. Les deux activités enzymatiques 4.1.2.13 et 1.2.1.12 sont dans les balises « entry » de type « enzyme » ayant comme « id » 44 et 45. Les 3 composés de ces deux réactions sont contenus dans les balises « entry » de type compound. La balise « reaction » avec l'id 44 décrit la réaction catalysée par l'activité enzymatique 4.1.2.13.
La balise relation indique les activités enzymatiques adjacentes, c'est-à-dire qui partagent un composé en commun dans leur réaction. 4.1.2.13 (id=44) et 1.2.1.13 (id=45) sont en relation par le composé id=48 qui correspond au composé C00118.

Chapitre :

10 Identification des voies présentes chez les champignons

10.1 Filtre topologique

Dans la base de données KEGG, à partir de la base KEGG PATHWAYS, plusieurs voies et réseaux sont mis à disposition. Ces représentations sont divisées en 7 catégories :

- Metabolism
- Genetic information processing
- Environnemental information processing
- Cellular processes
- Organismal systems
- Human diseases
- Drug development

La partie « metabolism » contient les voies associées au métabolisme mais aussi des cartes globales pour un aperçu d'un réseau métabolique particulier, par exemple le réseau métabolique des acides gras. À partir de ces voies métaboliques on peut construire des réseaux métaboliques à la carte : réseau d'une espèce ou réseau à partir de voies d'intérêts.
Le nombre de voies métaboliques récupérées dans KEGG est de 157 (dernier téléchargement le 5 janvier 2021). Les cartes globales ont été exclues. Les voies dans KEGG ont chacun un identifiant à 6 chiffres (exemple : 00010 pour la glycolyse/Glycogenesis). Les cartes globales ont un identifiant qui commence par 01 ce qui permet très facilement de les identifier.

Dans le but de construire le réseau métabolique global chez les champignons, il faut déterminer quelles sont ces voies présentes chez les champignons. L'objectif n'est pas d'identifier si la voie présente est vraiment fonctionnelle (validation expérimentale) mais d'établir une règle formelle pour dire qu'avec les éléments qui constituent la voie on peut l'ajouter au réseau métabolique. Pour rappel, les voies dans KEGG sont des représentations globales qui regroupent sous une même représentation la collection de voies de plusieurs espèces.

Les voies métaboliques dans KEGG sont dessinées manuellement à partir des connaissances issues de la littérature. Certaines voies ont été largement étudiées chez plusieurs espèces et d'autres n'ont été étudiées que chez des espèces modèles ou des espèces d'intérêts. Il est important de noter que certaines voies métaboliques sont bien caractérisées, tandis que d'autres peuvent être moins connues ou en cours de découverte. Certaines ne sont pas encore complètement résolues, et la plupart des réactions enzymatiques de la voie ne sont pas encore caractérisées ou l'enzyme qui catalyse la réaction n'est pas encore correctement définie (activité enzymatique avec une nomenclature incomplète par exemple 2.1.1.-). Par conséquent, il y a des informations manquantes dans les voies. Afin de combler ces lacunes dans les représentations, dans certains cas, les réactions non élucidées sont remplacées par des balises sans aucune information ou incomplètes et le format KGML n'est pas respecté (composé de la réaction non identifiée et aucune relation avec les autres réactions). Sur quelques voies, des illustrations schématiques (images figées) ont été ajoutées par les

auteurs pour avoir une visualisation de la voie. Ces voies qui ne respectent pas le format KGML ont été exclues car non manipulables computationnellement pour des analyses de graphes. Au total 24 voies sur 157 ne respectaient pas le format défini par KEGG.
Malheureusement, l'exclusion de ces voies va exclure les activités enzymatiques qui composent ces voies du réseau métabolique global.

L'un des premiers objectifs avant de construire le réseau métabolique global est d'identifier les voies ou la partie de la voie qui sont présentes chez les champignons. Nous allons sélectionner ces voies sur des critères topologiques.

Nous avons retranscrit les informations contenues dans les balises « relation » du format KGML en format SIF (Simple Interaction File) pour permettre une analyse de graphe et décrire topologiquement chaque voie. Le format SIF est un fichier texte qui énumère toutes les activités enzymatiques (sommets) qui sont reliées par une relation dans le fichier KGML (partage un composé en commun).
Par exemple au format SIF, la série de réaction de la Figure 9.5 s'écrit :

44ec4.1.2.13 ECrel 54ec1.2.1.12

Le chiffre avant ec indique l'identifiant unique (id) contenu dans la balise « entry » du sommet dans le format KGML. Cet identifiant est essentiel car il permet de différencier les EC-numbers identiques mais qui sont présents à différents endroits de la voie. Une ligne dans le format SIF décrit deux sommets qui sont reliés dans le graphe.

Le format SIF peut être ouvert à partir de n'importe quel logiciel de manipulation et de visualisation de graphe par exemple Cytoscape (Shannon *et al.*, 2003) ou encore Gephi (Bastian *et al.*, 2009). Le format SIF offre la possibilité de le manipuler sur R ou Python avec des librairies d'analyses de graphes comme igraph pour R et NetworkX pour python.

Comme nous l'avons déjà dit, les voies dans KEGG représentent une synthèse de la voie pour toutes les espèces. Nous avons donc filtré les activités enzymatiques présentes chez les champignons dans chaque voie pour obtenir la voie spécifique des champignons (Figure 10.1).
Par définition, **une voie est une série de réactions enzymatiques**. Pour faire une série de réactions, il faut au minimum deux réactions successives. D'un point de vue de la théorie des graphes, cela se traduit par au moins deux sommets qui sont reliés par une arête et un diamètre supérieur ou égale à 1.
Nous avons donc décidé qu'une voie est probablement présente chez les champignons si au moins deux sommets sont reliés dans le graphe. Sinon la voie est considérée comme absente.

9 voies métaboliques de KEGG ne contiennent aucune activité enzymatique présente chez les 174 espèces de champignons. Dans les voies où au moins une activité enzymatique présente dans les profils phylogénétiques a été détectée, le diamètre de la voie a été calculé.

Si le diamètre est supérieur ou égale à 1, cela signifie qu'au moins deux activités enzymatiques sont connectées.
En plus des 9 voies où aucune activité enzymatique chez les champignons n'a été détectée, 17 autres voies ont un diamètre de 0. Dans ces 17 voies, il n'y a aucune succession de réactions enzymatiques. Les activités enzymatiques sont toutes isolées. Ces activités enzymatiques isolées peuvent être le fruit d'une voie dont la majorité des activités enzymatiques ne sont pas encore annotées (EC-number incomplet : 4.2.1.-), par conséquent ces informations manquantes vont créer des trous dans les voies (Figure 10.2). Ces activités enzymatiques peuvent être tout simplement des activités enzymatiques qui sont réutilisées dans plusieurs voies. L'activité enzymatique est un élément essentiel dans une autre voie et peut être utilisée dans d'autres voies chez des espèces en dehors des champignons pour constituer la voie. Mais la présence d'une activité enzymatique chez les champignons peut aussi être due à une erreur d'annotation d'une enzyme chez les champignons.

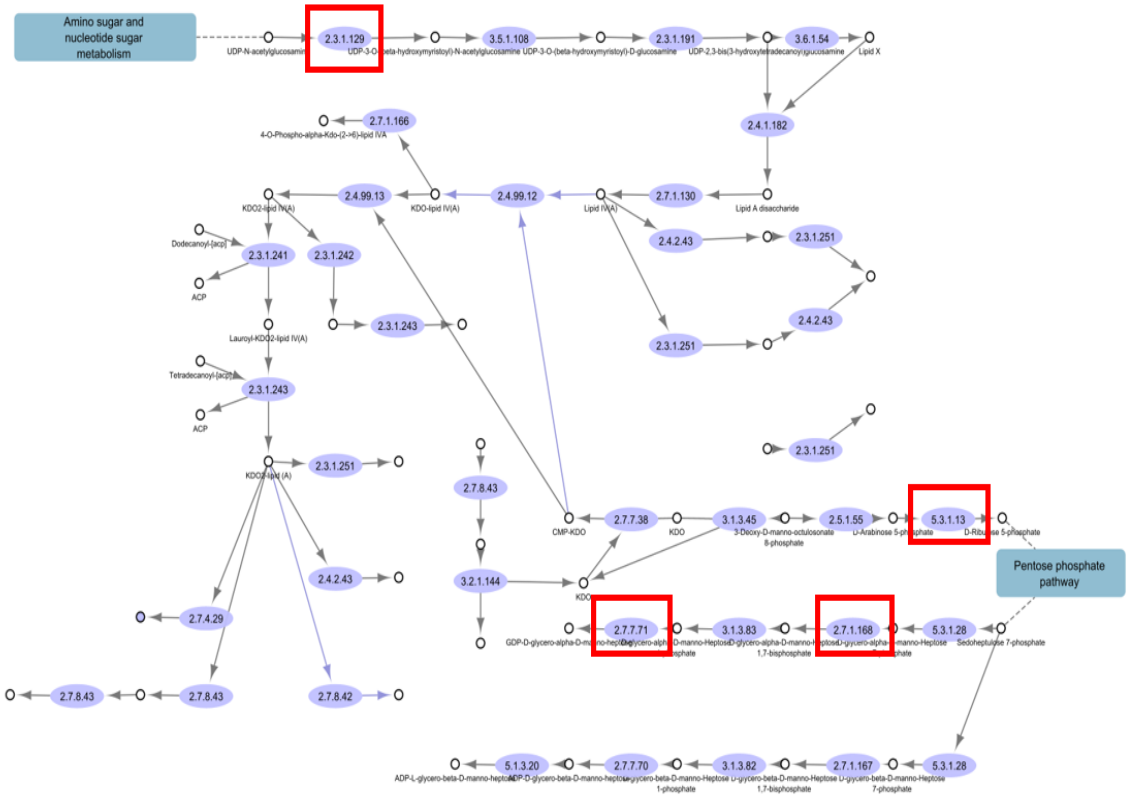


Figure 10.1 : Voie de la biosynthèse de la lipopolysaccharide. Les activités enzymatiques encadrées en rouge sont les seules activités enzymatiques identifiées chez les champignons et elles sont spécifiques de la voie. Elles ne partagent pas de composé en commun et par conséquent elles sont isolées. De ce fait, cette voie est considérée comme absente chez les champignons. La lipopolysaccharide est une composante spécifique de la membrane externe des bactéries gram-négatives (Silhavy *et al.*, 2010). La présence de ces 4 activités enzymatiques chez les champignons est probablement due à des erreurs d'annotations des enzymes chez les champignons ou ces activités enzymatiques assurent d'autres fonctions

qui ne sont pas encore connues dans d'autres voies.

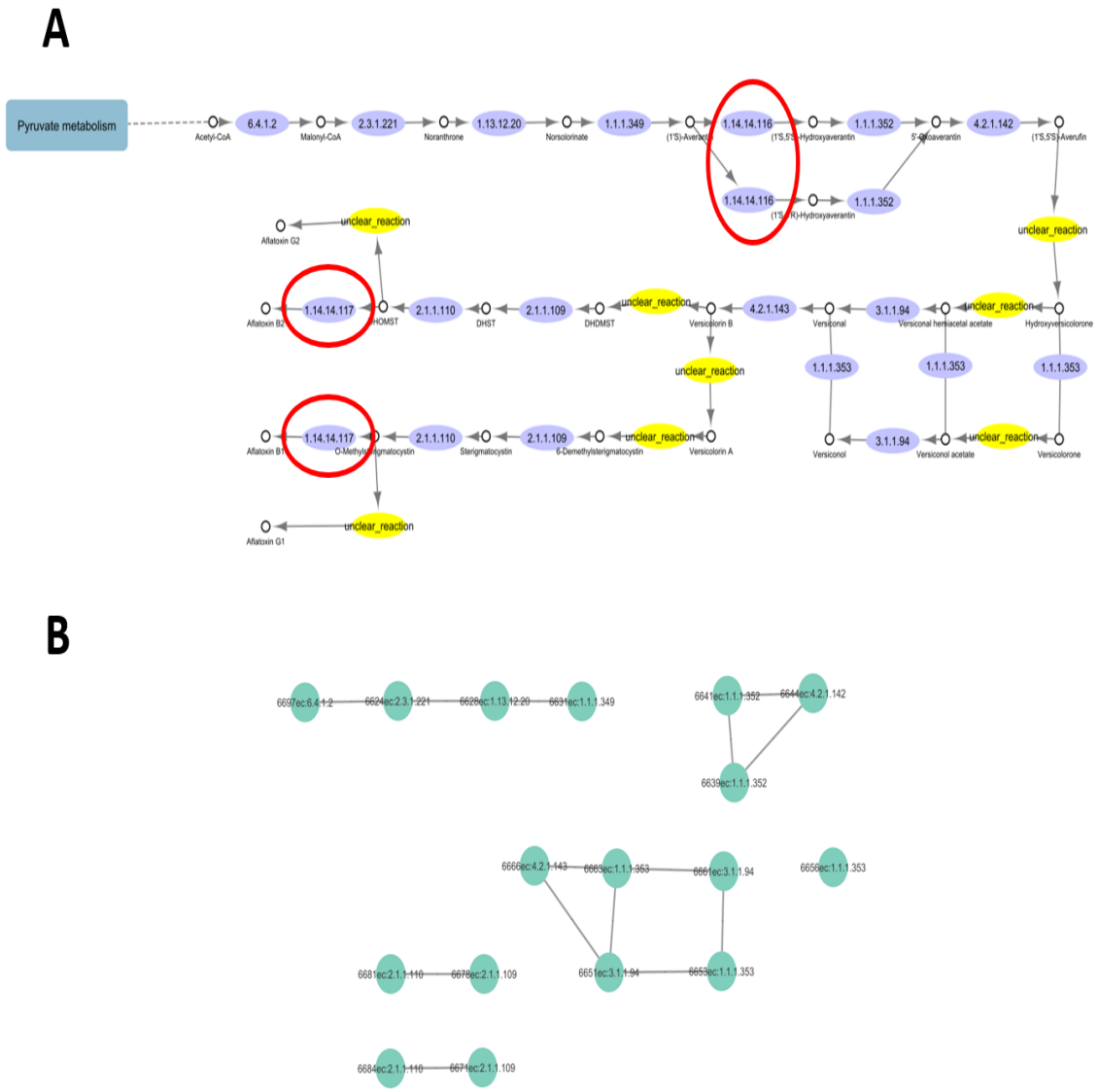


Figure 10.2 : Voie de synthèse de l'Aflatoxine (A) Représentation de KEGG réalisée sous Cytoscape. Les cercles violets représentent les activités enzymatiques. Les activités enzymatiques encerclées en rouge sont des EC-number qui ont changé d'EC-number entre l'annotation des groupes d'orthologues (Pereira *et al.*, 2014) et le téléchargement des voies KEGG (2021). Ces activités enzymatiques étant absentes des profils sont considérées comme absentes chez les champignons. Les cercles en jaune sont des réactions pas encore caractérisées et aucune activité enzymatique ne leur a été attribuée. **(B)** Représentation au format SIF (graphe des activités enzymatiques) de la voie avec seulement les activités enzymatiques identifiées chez les champignons dans les profils phylogénétiques. Les réactions sans activités enzymatiques ou avec EC-number incomplet ne sont pas incluses dans le graphe et créent des trous. L'activité enzymatique 1.1.1.363 apparaît comme complètement isolée à cause des réactions autour qui ne sont pas caractérisées.

Après cette étape de filtre, 107 voies métaboliques sont susceptibles d'être présentes chez les champignons.
Afin de déterminer quelles sont les voies qui sont les plus susceptibles d'être présentes chez les champignons et les voies qui ne contiennent que des modules qui sont réutilisés dans plusieurs voies, nous allons classer les voies en fonction du nombre d'activités enzymatiques et de leurs successions dans les voies. En effet, plus le nombre d'activités enzymatiques chez les champignons est élevé, plus il y a de grande chance que la voie soit présente chez les champignons.

10.2 Classification topologique des voies

L'objectif dans cette section est de classer les voies en fonction de leurs propriétés topologiques afin de sélectionner les voies qui sont très probablement présentes chez les champignons et qui vont faire partie du réseau. Cette classification pourra aussi servir afin de comprendre comment une voie métabolique évolue, car en fonction de leur topologie, on peut supposer que les voies peuvent évoluer différemment.

Pour chaque voie, nous avons défini plusieurs descripteurs topologiques : le nombre de composantes connexes, la taille de la plus grande composante connexe, le nombre d'activités enzymatiques de la voie, le nombre de sommets, le diamètre et la distance moyenne entre les sommets du graphe (la définition de tous ces termes est disponible dans la partie I.2.4.1).

Une analyse en composantes principales (ACP) a été effectuée sur ces descripteurs afin d'extraire et de visualiser les informations importantes (les composantes principales) qui permettent de classer les voies. L'ACP a été effectué sous R avec les librairies FactoMineR et factoxtra. Dans l'ACP les variables ont été normalisées afin qu'elles soient comparables.
Pour déterminer le nombre d'axes (composantes principales) de l'ACP, on a utilisé un critère « absolu », c'est-à-dire que l'on ne retient que les axes avec une valeur propre supérieure à 1. Dans notre analyse les 2 premiers axes (Dim.1 et Dim.2) expliquent 86% des variations (Table 10.1). C'est un pourcentage acceptable car ils représentent 86% de l'information contenue dans les jeux de données.

	Valeurs propres	Pourcentage de variance	Pourcentage de variance cumulé
Dim.1	4.05	67.55	67.55
Dim.2	1.12	18.81	86.37
Dim.3	0.52	8.67	95.04
Dim.4	0.26	4.44	99.48
Dim.5	0.021	0.35	99.84
Dim.6	0.009	0.15	100

Table 10.1 : Tableau des valeurs propre, du pourcentage de variance et le cumulé de chaque composante principale.

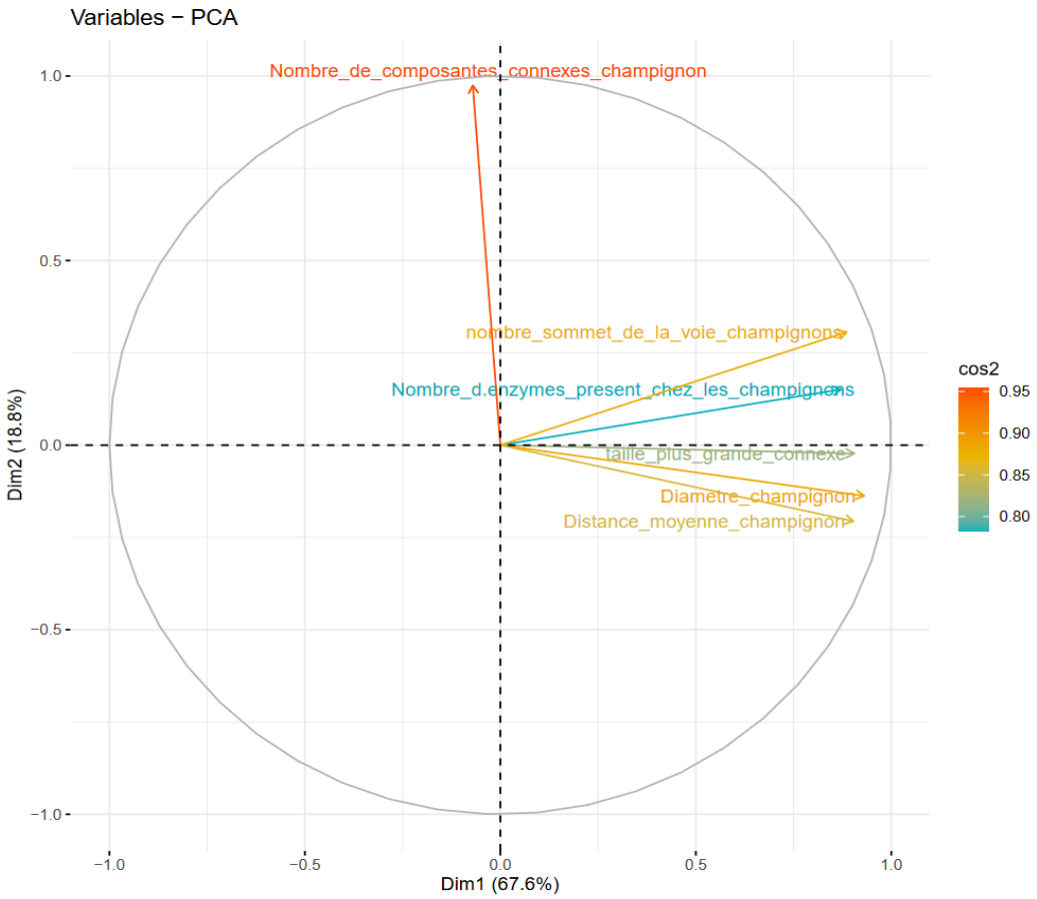


Figure 10.3 : Cercles de corrélation des variables. Sur l'axe des abscisses nous avons la dimension 1 et la dimension 2 sur l'axe des ordonnées. Les variables sont colorées en fonction de leur qualité de représentation (cos2) en suivant un gradient de couleur de rouge à bleu, rouge pour une qualité de représentation égal à 1 et bleu si égal à 0. Les variables corrélées sont regroupées.

Dans la Figure 10.3, les points qui nous intéressent sont les points qui sont les plus proches des axes (corrélés avec l'axe) et assez loin des origines afin d'avoir une idée de la qualité de la représentation. Dans cette représentation les variables corrélées sont regroupées. Le nombre de sommets, le nombre d'activités enzymatiques, la taille de la plus grande composante connexe, le diamètre et la distance moyenne sont corrélés avec l'axe horizontal. Le nombre de composantes connexes est corrélé avec l'axe vertical. L'angle quasi perpendiculaire entre le nombre de composantes connexes et les variables corrélées avec l'axe horizontal signifie que ces variables sont indépendantes. Le nombre de composantes connexes et une des variables corrélée avec l'axe horizontal vont permettre de bien différencier nos voies.
D'un point de vue biologique, le diamètre de la voie représente la plus longue chaîne de réaction de la voie et le nombre de composantes connexes le nombre de modules indépendants de la voie. Mais un nombre de composantes connexes élevé peut aussi signifier qu'un grand nombre de réactions de la voie est associé à une activité enzymatique dont l'EC-number n'est pas encore défini et va entraîner de ce fait des trous dans la représentation de la voie (Figure 10.2).
Nous avons classé les 107 voies en 3 catégories en fonction du nombre de composantes connexes et du diamètre : les voies avec de grands diamètres qui ont un diamètre supérieur à 3, les voies modulaires qui représentent un nombre de composantes connexes supérieur à 7 mais un diamètre inférieur à 3 et les « petites voies » qui ont un diamètre inférieur à 3 et un nombre de composantes connexe inférieur à 7 (Figure 10.4).

D'après cette classification, nous recensons 54 voies à grands diamètres et 8 voies modulaires. L'une des caractéristiques de ces voies est le nombre d'activités enzymatiques présentes dans la voie qui est élevé par rapport au petites voies. Nous recensons 45 petites voies (Figure 10.4).
Les voies à grands diamètres et modulaires sont très probablement des voies présentes chez les champignons car le nombre d'activités enzymatiques chez les champignons est élevé.

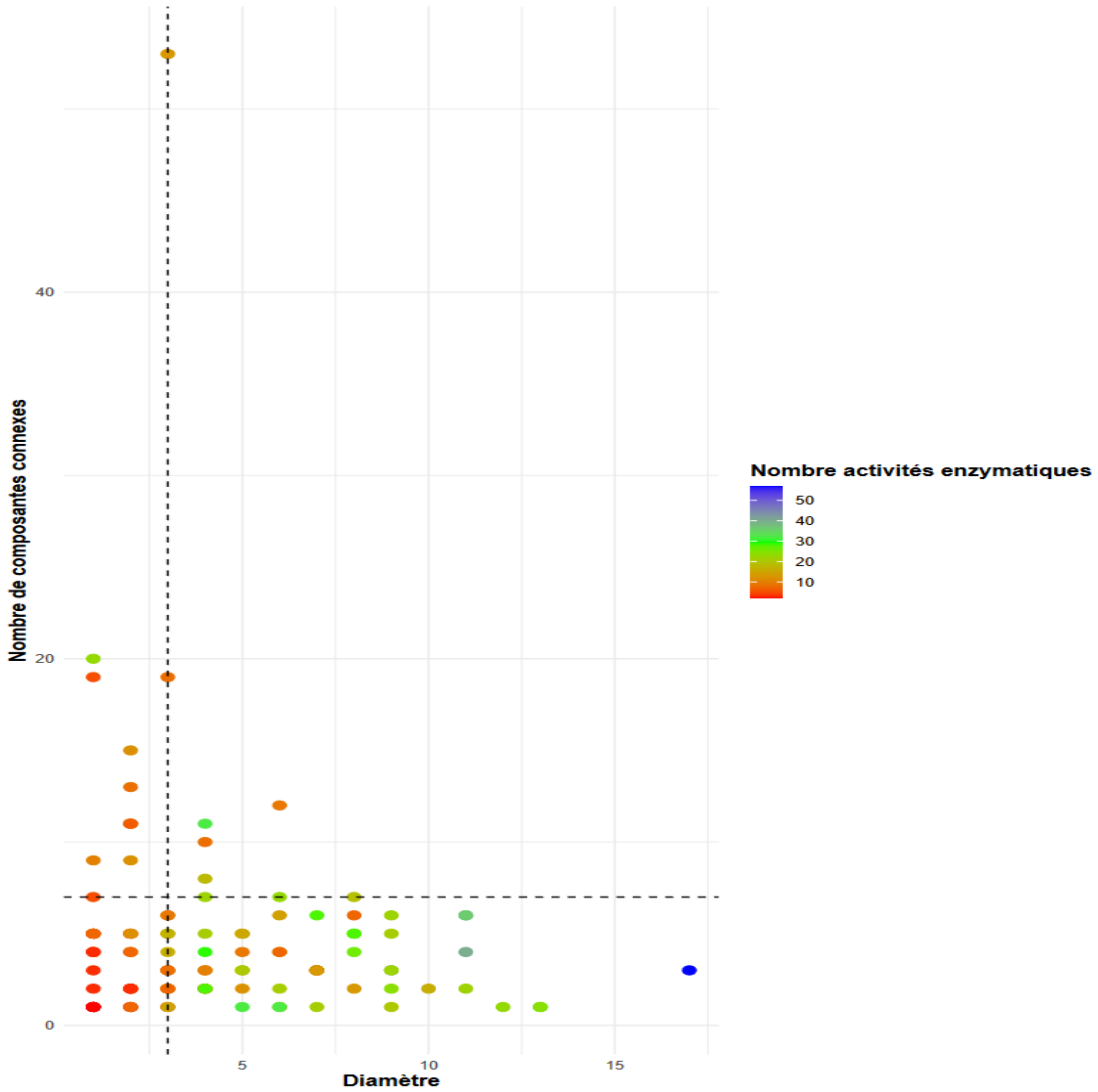


Figure 10.4 : Nombre de composantes connexes en fonction du diamètre de la voie. Le diamètre est indiqué sur l'axe des abscisses et le nombre de composantes connexes sur l'axe des ordonnées. Les voies sont colorées en fonction du nombre d'activités enzymatiques dans la voie. La ligne verticale discontinue indique un diamètre de 3 et la ligne horizontale indique un nombre de composantes connexes égal à 7.

.

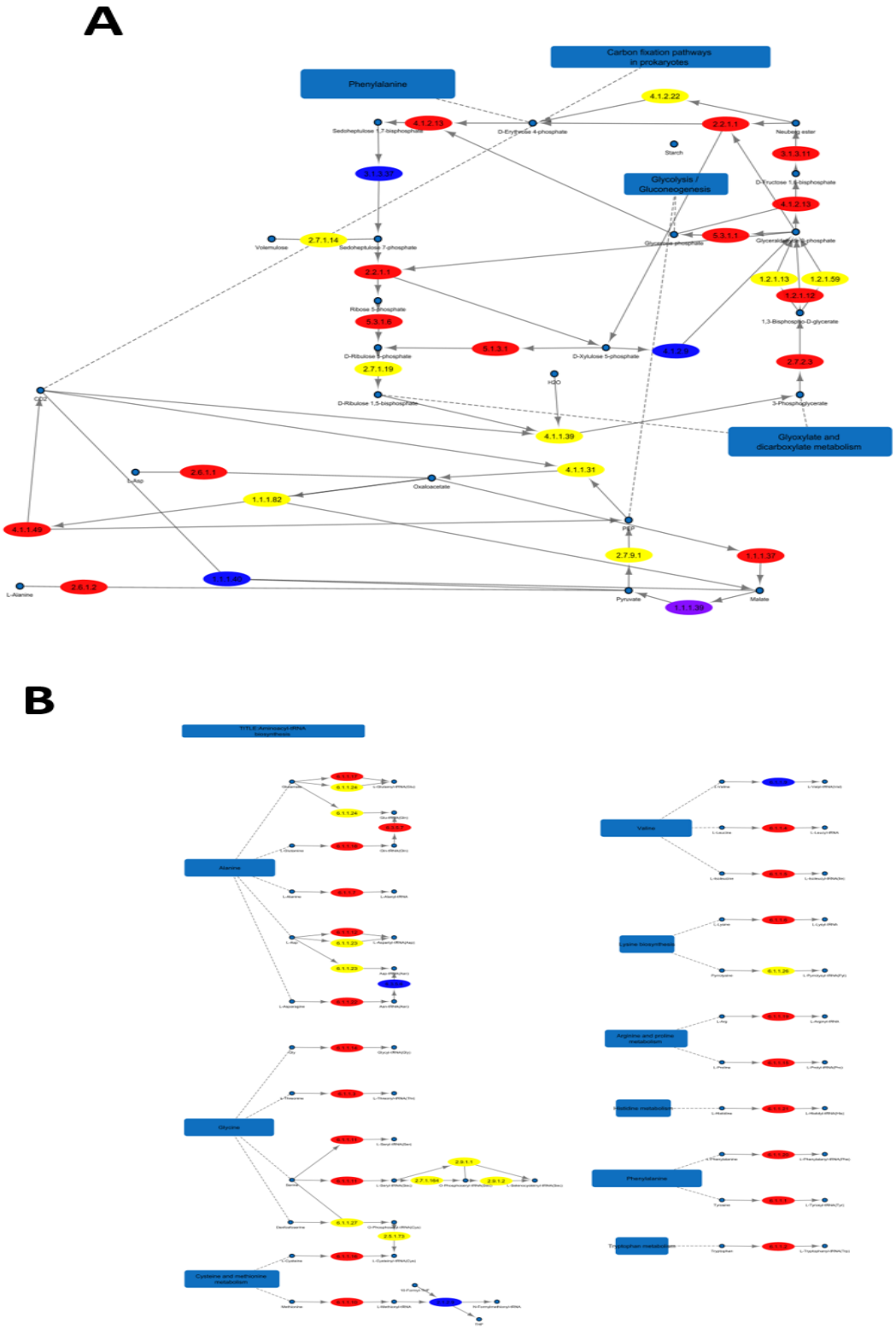


Figure 10.5 : A. Une voie métabolique avec un grand diamètre : le Carbon fixation in photosynthétic organism. **B.** Une voie métabolique modulaire : l'Aminoacyl-trna biosynthesis pathways. Les activités enzymatiques conservées par toutes les espèces sont colorées en rouge (conservation supérieure à 85%), les activités enzymatiques spécifiques de certaines espèces sont colorées en bleu et les activités enzymatiques absentes des champignons sont colorées en jaune.

Parmi les voies modulaires, on peut citer par exemple l'Aminoacyl-tRNA biosynthesis (Figure 10.5 B) . Les voies modulaires sont constituées de plusieurs modules indépendants.
Des voies censées ne pas être présentes chez les champignons figurent parmi ces deux catégories comme le « carbon fixation in photosynthetic organism ». Les champignons ne sont pas capables de pratiquer la photosynthèse. Les activités enzymatiques identifiées chez les champignons montrent que 16 des 25 activités enzymatiques de cette voie sont présentes chez les champignons (Figure 10.5 A), mais 15 d'entre elles également présentes dans d'autres voies.

La réutilisation d'une même activité enzymatique dans plusieurs voies est assez classique. Sur les 910 activités enzymatiques, 286 activités enzymatiques sont présentes dans au moins deux voies métaboliques (Figure 10.6). Les activités enzymatiques les plus redondantes, c'est-à-dire impliquées dans le plus grand nombre de voies métaboliques, sont : 4.2.1.17, 1.2.1.3, 2.3.1.9 et 1.1.1.35. Paradoxalement, 4.2.1.17 n'est cependant pas fortement conservé, elle fait partie d'un groupe d'activités enzymatiques qui sont absentes chez les *Saccharomycètes* (Figure 7.16 et table 7.1).

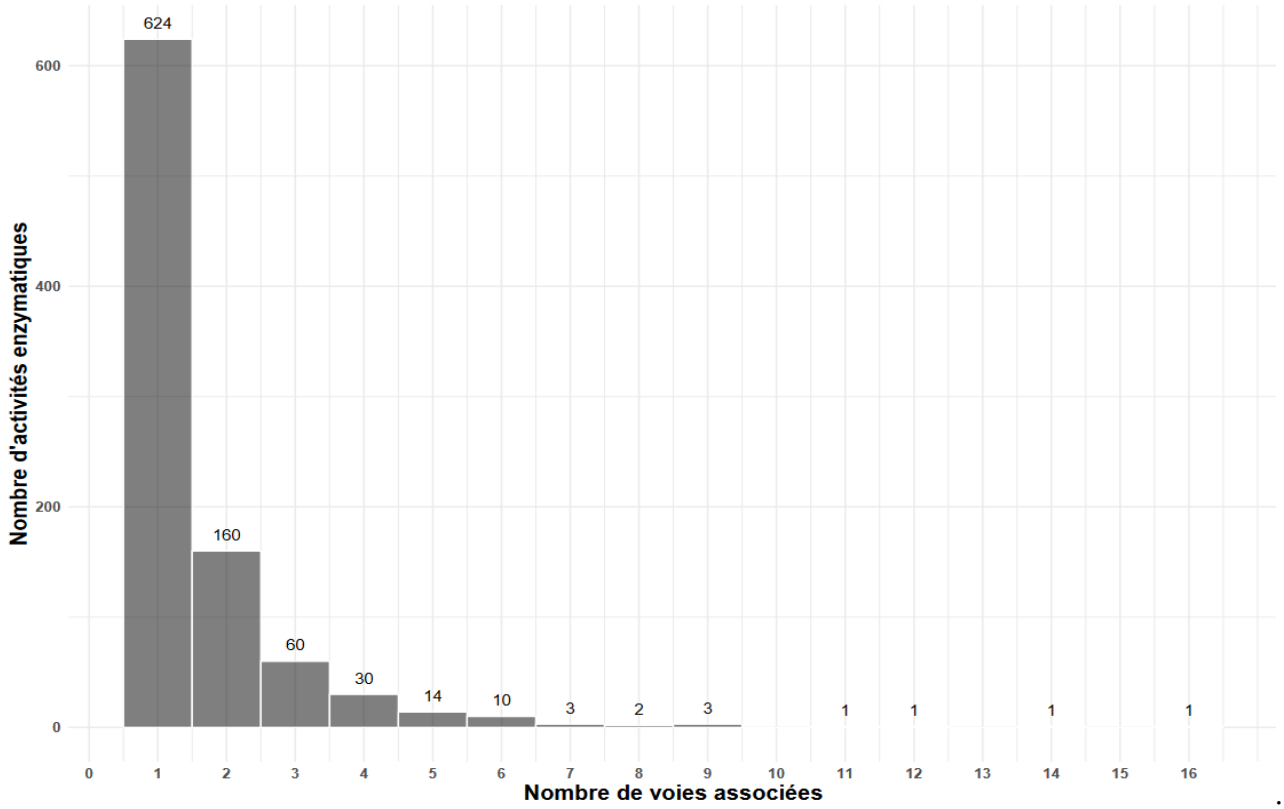


Figure 10.6 : Histogramme du nombre de voies associées par activités enzymatiques. Chaque barre équivaut à un intervalle de 1. Le nombre au-dessus de chaque barre indique le nombre d'activités enzymatiques.

La redondance des activités enzymatiques dans le réseau métabolique témoigne d'une réutilisation de la même activité enzymatique dans deux voies différentes. Ces activités enzymatiques redondantes indiquent des modules qui sont réutilisés dans d'autres voies. De telles activités enzymatiques soutiennent l'idée que le recrutement d'activités enzymatiques joue un rôle important dans l'évolution où de nouvelles voies peuvent émerger par le recrutement d'activités enzymatiques déjà présentes.

Comme nous l'avons vu précédemment, une activité enzymatique peut être présente dans deux voies différentes (réutilisation de la même activité enzymatique). Sur les 910 activités enzymatiques, 614 activités enzymatiques n'opèrent que dans une seule voie. Nous avons vérifié si une forte conservation de l'activité enzymatique est associée au fait que l'activité enzymatique est essentielle dans plusieurs voies. La Figure 10.7 montre la différence de conservation entre les activités enzymatiques présentes dans une seule voie et les activités enzymatiques présentent dans au moins 2 voies. Cette Figure montre qu'il y a une différence significative (la p-value associée à un test de Man-Whitney-Wilcoxon est de 10-4) entre les d'activités enzymatiques présentes uniquement dans une seule voie et les activités enzymatiques redondantes. Les activités enzymatiques présentes dans une seule voie sont moins conservées que les activités enzymatiques présentes dans plusieurs voies.

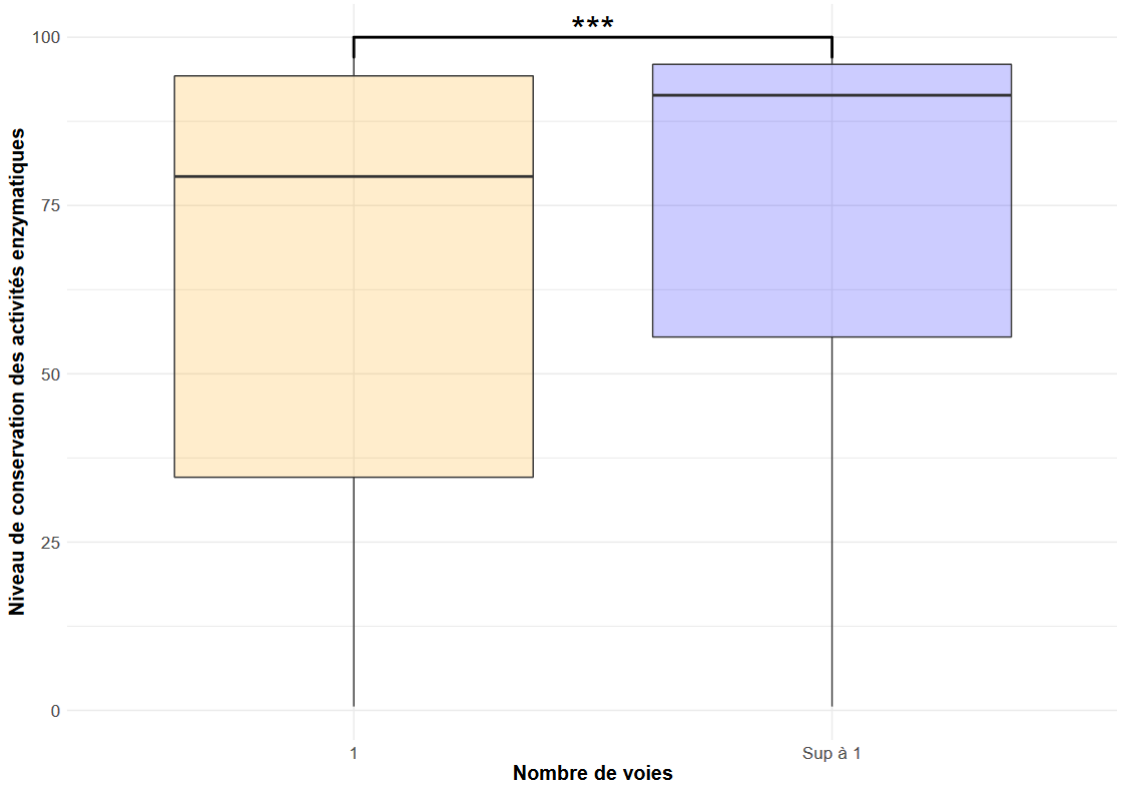


Figure 10.7: Comparaison de la conservation entre les activités enzymatiques présentes dans une seule voie (boîte jaune) et les activités enzymatiques présentes dans au moins 2 voies (boîte violette). ******* Signifie une p-value inférieure à 10-3 entre les deux distributions.

Dans le cas de la synthèse de l'aflatoxine, cette toxine est principalement produite par la section *Flavi* du genre des *Aspergillus* (Frisvad *et al.*, 2018). Il est à noter que les activités enzymatiques de cette voie sont spécifiques de la voie et les gènes associés sont organisés en cluster dans le génome (Caceres *et al.*, 2020). Or dans les profils phylogénétiques des activités enzymatiques de cette voie, nous observons que les activités enzymatiques associées sont réparties dans plusieurs espèces de champignons (Figure 10.8). L'identification de ces activités enzymatiques en dehors des *Aspergillus* peut correspondre à la présence du cluster de gènes ou une partie dans le génome d'autres espèces, mais qui ne sont plus exprimées ou incomplètes rendant la voie non fonctionnelle. Mais Frisvad et ses collaborateurs (Frisvad *et al.*, 2018) supposent qu'outre la production de l'aflatoxine, ces activités enzymatiques et les gènes associés semblent jouer un rôle dans d'autres processus biologiques.

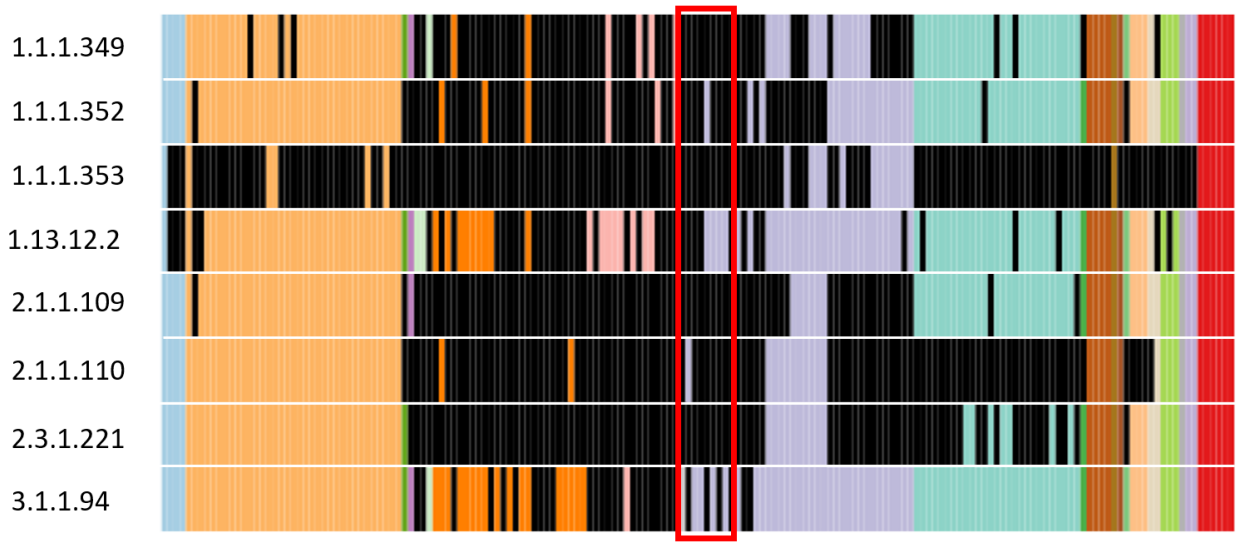


Figure 10.8 : Profils phylogénétiques des activités enzymatiques de la voie de biosynthèse de l'aflatoxine chez les champignons. Les lignes représentent les activités enzymatiques et les colonnes les espèces. Une cellule colorée en noire indique la présence de l'activité enzymatique, sinon la cellule est colorée en fonction de la classe taxonomique de l'espèce. Les espèces appartenant au genre *Aspergillus* sont encadrées en rouge.

L'une des principales difficultés dans notre prédiction topologique, surtout pour les petites voies constituées d'un petit nombre d'activités enzymatiques, est de déterminer si la présence des activités enzymatiques est seulement due au fait de leur réutilisation et la voie n'est pas fonctionnelle, comme par exemple dans le cas de novobiocine (Figure 10.9). Ou dans le cas d'activité enzymatique spécifique de la voie, quelles sont les activités enzymatiques essentielles pour que la voie soit fonctionnelle.

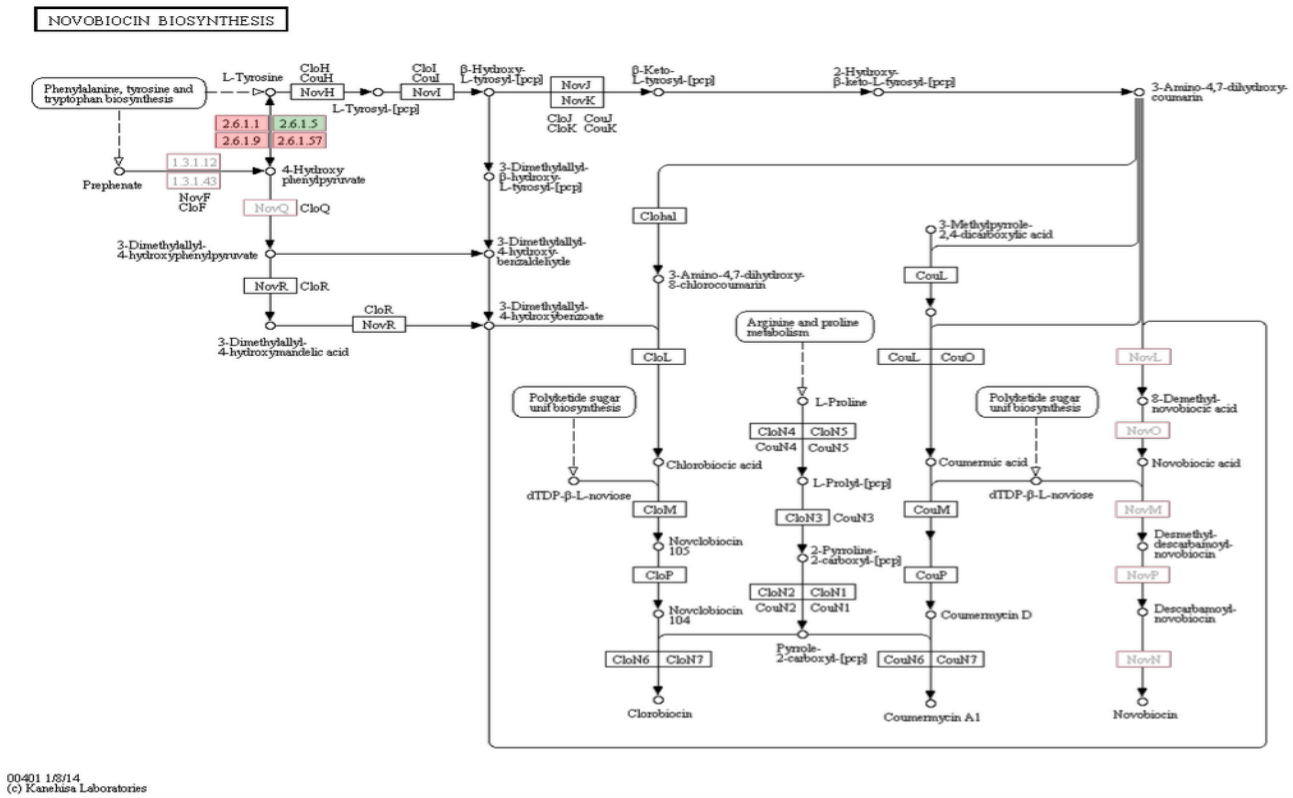


Figure 10.9 : Biosynthèse de la Novobiocine. La novobiocine est un antibiotique seulement produite que par quelques espèces de bactéries appartenant au *Streptomyces* (Steffensky *et al.*, 2000). Les activités enzymatiques surlignées en rouge et en vert sont les seules activités enzymatiques présentes chez les champignons. Les activités enzymatiques surlignées en rouge sont fortement conservées et en vert une activité enzymatique spécifique de certaines espèces. Les 4 activités enzymatiques sont présentes ensemble dans la voie du métabolisme de la Tyrosine. La voie de la Novobiocine est donc absente chez les champignons. Figure tirée de FungiPath (Grossetête *et al.*, 2010).

Si une activité enzymatique est spécifique d'une voie, nous avons considéré que cette voie est présente chez les champignons. Avec cette approche, nous avons retenu 33 voies parmi les 45 « petites voies ».

Pour les 12 voies restantes, nous avons regardé dans la littérature s'il y avait une évidence de la présence de la voie chez les champignons. 7 voies sur 12 ont été mentionnées dans la littérature comme présente chez les champignons. Probablement, une grande partie des activités enzymatiques de la voie n'ont pas été annotées chez les champignons ou la succession des réactions de la voie est totalement différente chez les champignons. Celles sur lesquelles rien n'a été trouvé dans la littérature n'ont pas été retenues pour la construction du réseau métabolique global chez les champignons.

Toutes ces étapes de sélection des voies sont résumées dans la Figure 10.10. Cette classification topologique a permis d'identifier les voies probablement présentes chez les champignons mais aussi va être utilisée pour comprendre comment les différentes voies ont évolué en fonction de leurs caractéristiques topologiques.

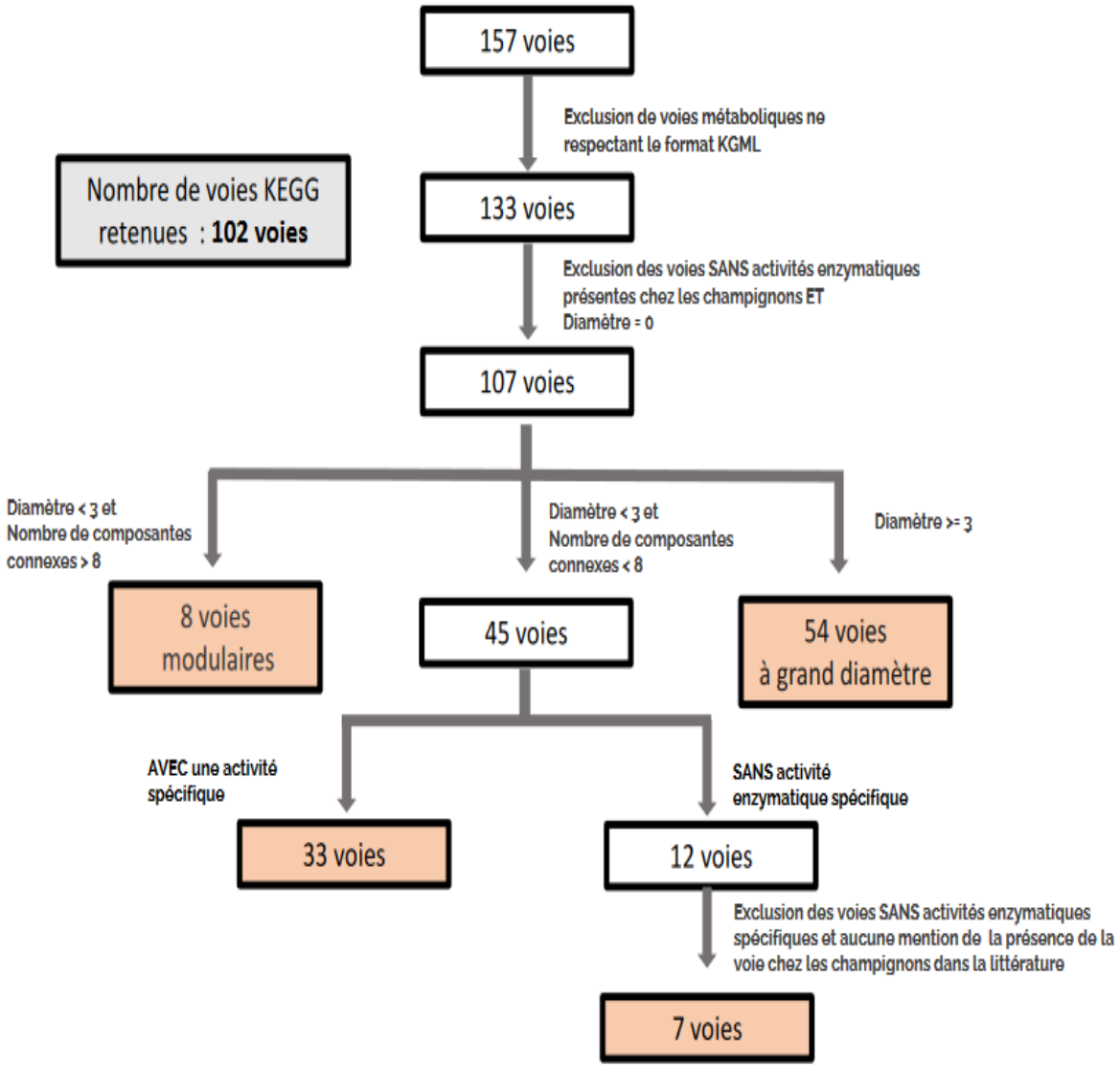


Figure 10.10 : Diagramme du protocole de sélection des voies métaboliques de KEGG. Les rectangles oranges indiquent les voies qui ont été retenues après l'application des critères de sélection.

Chapitre :

11 Construction du réseau métabolique

11.1 Connexion des voies de KEGG

La construction du réseau métabolique se fait en connectant les voies métaboliques entre elles. Ainsi pour construire le réseau métabolique global, il suffit de relier toutes les voies à partir des composés partagés. Dans la représentation KEGG, les composés en commun entre les voies sont indiqués dans le fichier KGML dans les balises « relations ».

En plus des composés en commun entre les voies, les mêmes réactions peuvent être retrouvées autour de ces points de connexions dans les voies à connecter (Figure 11.1). La duplication de ces réactions est surtout due au fait que ces réactions se trouvent à la frontière entre deux voies. Pour éviter ces duplications dans le réseau métabolique au niveau des points de connexion lors de la connexion des voies, les réactions identiques autour des points de connexion (même EC-number et même composés) sont fusionnées pour ne former qu'une seule réaction.

Connecter les voies entre elles va aussi créer de nouveaux liens entre les activités enzymatiques. Deux activités enzymatiques issues des deux voies vont partager un composé en commun. Pour garder la règle du format KGML, où deux activités enzymatiques qui partagent le même composé doivent être mises en relation, à chaque fois qu'une connexion est faite, il faut s'assurer que les réactions enzymatiques qui utilisent ce composé soient mises en relation.

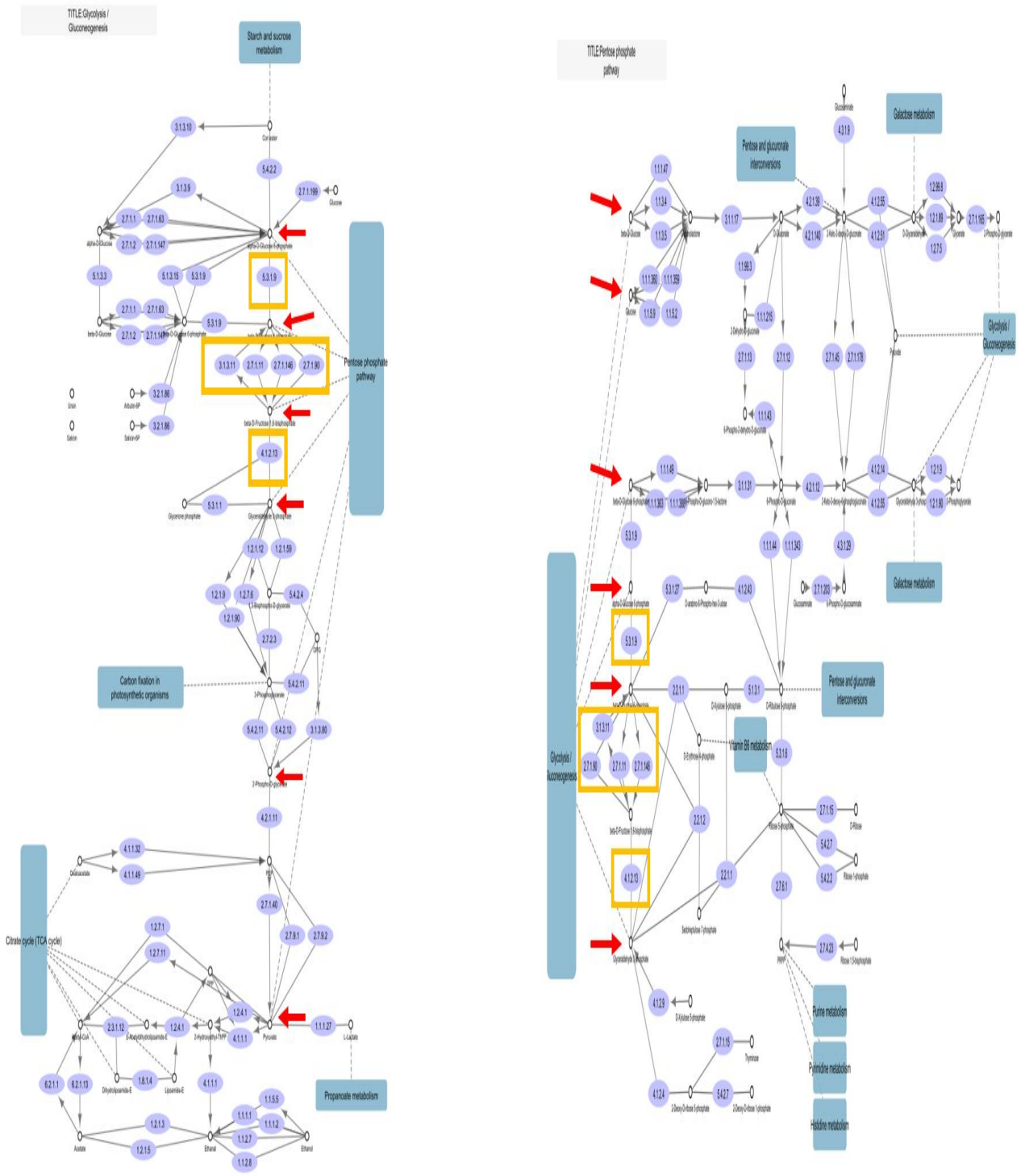


Figure 11.1 : Connexion entre la voie de la glycolyse (à gauche) et la voie du pentose phosphate (à droite). Tous les composés (cercles blancs) en commun entre les deux voies sont indiqués en rouge. Les réactions (cercles bleus) en commun entre les deux voies autour de ces points de connexions sont encadrées en jaune.

La connexion des voies entre elles est effectuée automatiquement avec un script Python (version 3.6.0) que j'ai développé. Le script prend en entrée la liste des voies KEGG à fusionner, et produit un réseau au format KGML mais aussi au format SIF.
La manipulation et la modification du fichier KGML qui est un format XML se font avec le module XML de python.
Une fois le réseau métabolique construit, le format KGML est converti en format SIF pour faciliter l'analyse topologique de la voie. Le format SIF peut être filtré pour ne garder que les activités enzymatiques qui sont présentes chez les champignons.

L'ensemble des scripts qui permettent de construire le réseau métabolique à partir des voies KEGG sont disponibles sur Github à l'adresse : https://github.com/Herinjiva/build_network_KEGG.

11.2 Caractéristique du réseau métabolique global

Le réseau métabolique global construit est composé de 2035 sommets qui représentent 910 activités enzymatiques (Figure 11.2 et Table 11.1). Ce nombre de sommets deux fois plus élevé que le nombre d'activités enzymatiques atteste que plusieurs activités enzymatiques sont présentes dans plusieurs voies métaboliques et sont donc présentes à plusieurs endroits du réseau. On dénombre 515 activités enzymatiques présentes qu'une seule fois dans le réseau (Figure 11.3).

Les activités enzymatiques 1.14.14.1 (55 fois), 2.3.1.85 (32 fois), 4.2.1.17 (31 fois), 2.3.1.86 (31 fois), 2.3.1.16 (21 fois), 1.1.1.35 (20 fois) sont les 6 activités enzymatiques les plus redondantes dans le réseau. Ces 6 activités enzymatiques ont comme point commun d'être impliquées dans le métabolisme des acides gras (fatty acid elongation et fatty acid degradation).

Figure 11.2 : Le réseau métabolique chez les champignons. Les sommets représentent les activités enzymatiques et deux activités enzymatiques sont reliées si elles partagent un composé en commun. La visualisation du réseau est effectuée en utilisant Cytoscape.

Le réseau métabolique en chiffres	
Densité	0.003
Diamètre	28
Distance moyenne entre les sommets	8.33
Nombre d'activités enzymatiques	910
Nombre de sommets	2035
Nombre de liens	7428
Nombre de composantes connexes	284

Table 11.1 : Le réseau métabolique des champignons en chiffres.

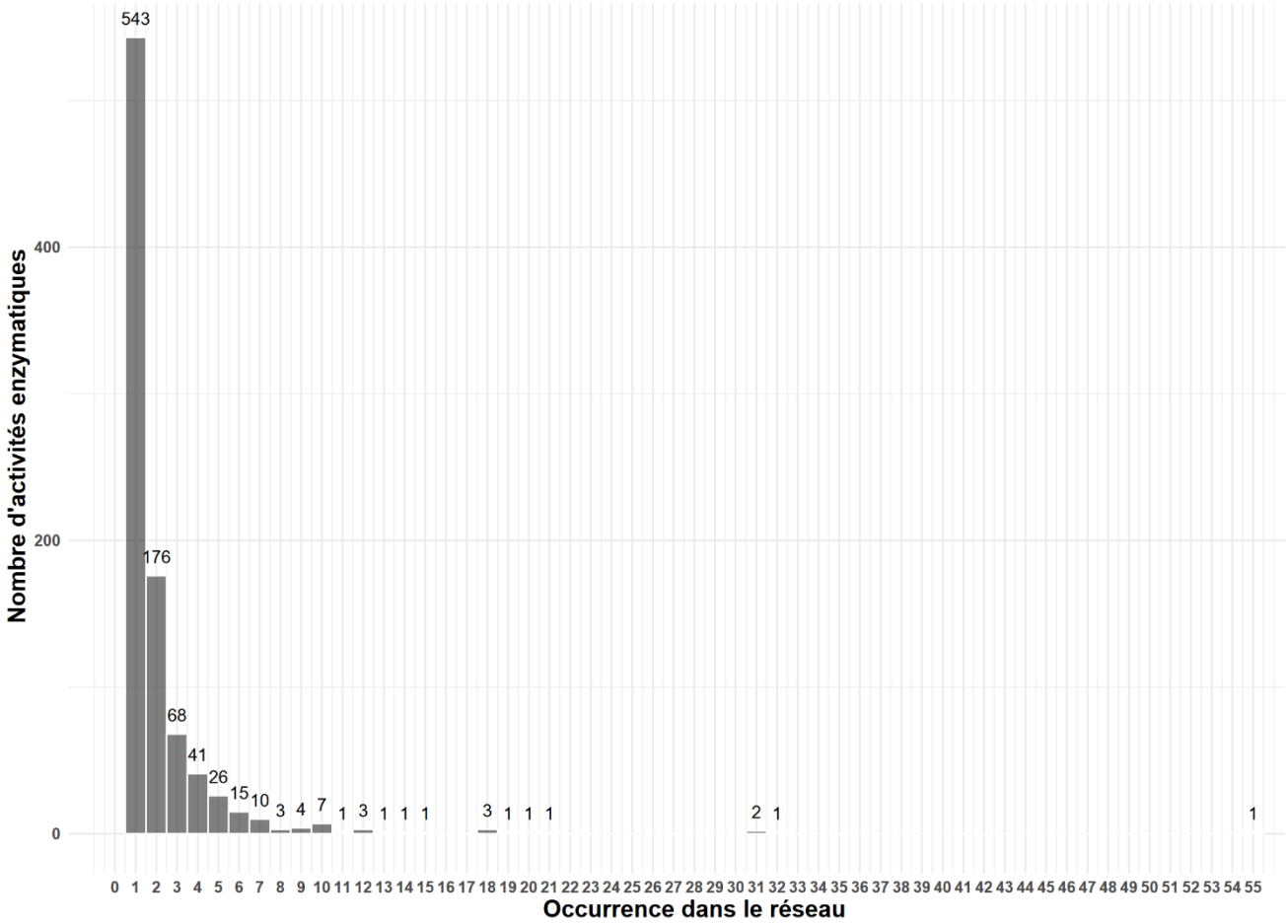


Figure 11.3 : Occurrence des activités enzymatiques dans le réseau métabolique. En abscisses, le nombre d'occurrence des activités enzymatiques dans le réseau. Chaque barre indique le nombre d'activités enzymatiques qui présentent un niveau d'occurrence donné.

Le nombre de liens dans le réseau est de 7428. Ce nombre correspond au nombre de paires de réactions enzymatiques qui partagent un composé en commun. En calculant la densité du réseau, qui est le rapport entre le nombre d'arêtes (ou d'arcs) observés divisé par le nombre d'arêtes (ou d'arcs) possibles, on obtient une densité de 0.003. Dans un graphe complet ou une clique, la densité est égale à 1. Il a été montré que les réseaux biologiques sont généralement peu connectés (Leclerc, 2008). Une densité faible du réseau métabolique indique que le nombre d'interactions entre les réactions enzymatiques est faible. Cette faible densité indique que l'interconversion des composés dans le réseau est principalement effectuée par des successions bien définies de réactions enzymatiques et que le nombre d'interactions entre les différentes réactions enzymatiques est très limité.
D'après les études de Fell and Wagner (Fell and Wagner, 2000), le réseau métabolique a une propriété de « small world » avec une distance moyenne entre les sommets de 3. Or ici le réseau métabolique obtenu a un diamètre de 28 et une distance moyenne entre les nœuds de 8. Ce qui montre que le réseau métabolique obtenu n'est pas si petit que ça. Ce résultat est cohérent avec la revue de Lima-Mendez and Helden, dans laquelle les auteurs expliquent que la propriété « small-world » est la conséquence de la présence des métabolites ubiquitaire dans la plupart des réactions (H2O, CO2, ATP…) dans le réseau (Lima-Mendez and Helden, 2009). La présence de ces métabolites va en effet créer des connexions qui ne devraient pas se faire entre différentes réactions. Dans les représentations KEGG, ces métabolites sont absents des réactions. Seuls les métabolites essentiels pour les réactions sont présents.
Le réseau métabolique est constitué de 284 composantes connexes. Avec une composante principale qui contient 1415 sommets. Ces composantes connexes peuvent représenter une voie totalement isolée du réseau métabolique ou des composantes artificielles. Ces composantes artificielles peuvent être dues à des activités enzymatiques manquantes dans les représentations de KEGG ou des erreurs d'annotations des enzymes chez les champignons.

IV. Exploration de l’évolution du réseau métabolique

Chapitre :

12 Localisation des activités enzymatiques dans le réseau en fonction de leur histoire évolutive

12.1 En fonction de la conservation

Le réseau métabolique représente la relation entre les différentes activités enzymatiques. Pour comprendre si le réseau métabolique exerce une pression sur l'évolution de ces nœuds (les activités enzymatiques), les informations évolutives sur les activités enzymatiques inférées dans les parties précédentes ont été cartographiées sur les nœuds du réseau métabolique (Figure 12.1).

Nous avons analysé les sommets du réseau ainsi que les liens entre ces sommets en utilisant deux métriques de la théorie des graphes: le degré et la centralité (I.2.4.1). Ces deux métriques ont été calculées à partir de Cytoscape (Doncheva *et al.*, 2012)
D'un point de vue enzymatique, le degré représente le nombre d'activités enzymatiques qui partagent le même métabolite avec le sommet étudié. Un sommet avec un degré élevé indique qu'il s'agit d'un hub où plusieurs séries de réactions transitent.
La centralité comme son nom l'indique, précise quelle est la position d'un sommet dans le réseau. Une valeur proche de 1 indique un nœud au centre du réseau et proche des autres sommets.

Nous avons comparé la distribution des degrés des activités enzymatiques très conservées, des activités enzymatiques ancestrales (qui prédatent l'apparition des champignons) mais perdues au cours de l'évolution chez certaines espèces et celles des activités qui sont apparues spécifiquement au cours de l'évolution chez les champignons (Figure 12.2).

Teste statistique pour la comparaison des distributions
Le test de Mann-Whitney-Wilcoxon a été utilisé pour comparer la distribution entre les échantillons par rapport à la médiane. Afin d'éviter le problème de la p-value inhérent aux grands échantillons (Lin *et al.*, 2013), lorsque la taille de l'échantillon fait plus de 500, nous avons réalisé 1000 tirages aléatoires d'échantillons de taille de 500 à partir de l'échantillon initial. La moyenne des p-values obtenues pour chaque échantillon a été ensuite calculée.

La comparaison des degrés montre que les activités enzymatiques anciennes conservées sont plus connectées que les activités enzymatiques anciennes spécifiques (perdues au cours de l'évolution) (la p-value du test de Mann-Whitney-Wilcoxon est de 10^{-04}). Ce résultat peut s'expliquer par le fait que la perte d'une activité enzymatique moins connectée aura moins d'impact que celle d'un sommet très connecté. La perte d'un sommet très connecté risque en effet de priver beaucoup de réactions d'un métabolite essentiel. Les activités enzymatiques les plus connectées étant celles dans la perte semblent la plus préjudiciable, il semble donc logique qu'elles soient plus conservées.
Ces résultats sont cohérents avec les résultats de Peregrin-Alvarez et ses collaborateurs, où les activités enzymatiques conservées chez les 3 domaine du vivant (Archée, Bactérie et Eukaryotes) sont plus connectées dans le réseau (Peregrín-Alvarez *et al.*, 2009). Ce résultat

confirme que le réseau métabolique est invariant d'échelle car une propriété du réseau métabolique plus globale est conservée chez le réseau métabolique des champignons.

En plus de leur forte connectivité, les activités enzymatiques anciennes conservées sont localisées au centre du réseau métabolique. Elles sont plus centrales que les activités enzymatiques anciennes spécifiques (Figure 12.2 : p-value 10^{-06}). Les activités enzymatiques anciennes spécifiques sont quant à elles plutôt localisées en périphérie du réseau.
Les 8 activités enzymatiques nouvelles spécifiques des champignons sont très peu connectées et sont localisées en périphérie du réseau métabolique par rapport aux activités enzymatiques anciennes (les p-value du test de Mann-Whitney-Wilcoxon entre la distribution des degrés et la centralité sont disponibles dans les annexes 3 et 4).

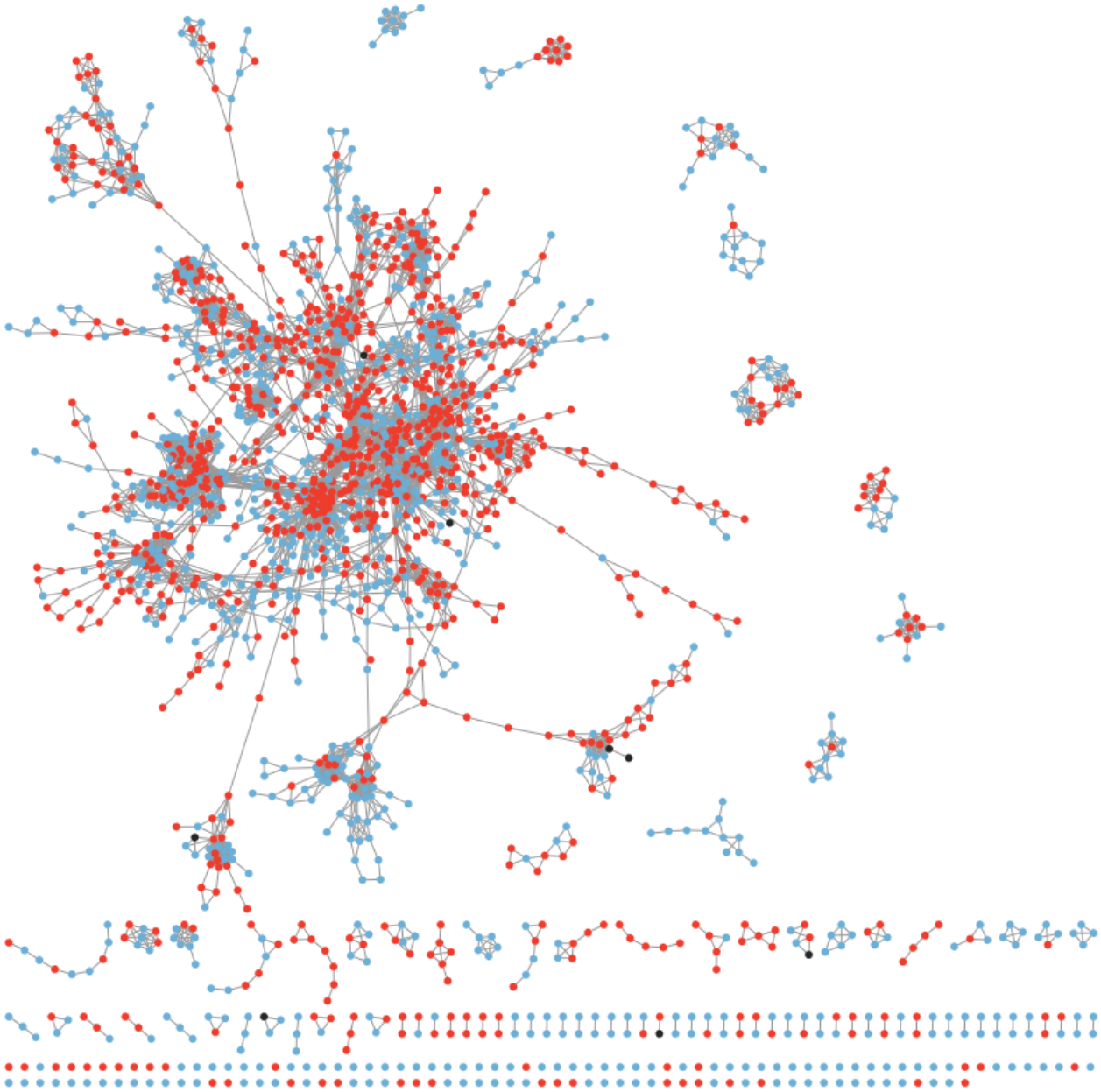


Figure 12.1 : Le réseau métabolique avec les informations évolutives des activités enzymatiques. Les sommets du réseau représentent les activités enzymatiques. Deux sommets sont connectés si les deux activités enzymatiques partagent un composé en commun. Les activités enzymatiques très conservées (présentes au moins chez 85% des 174 espèces de champignons) sont colorées en rouge, en bleu les activités enzymatiques ancestrales mais perdues au cours de l'évolution chez certaines espèces de champignons (anciennes spécifiques) et en noir les activités enzymatiques qui sont apparues spécifiquement chez les champignons (nouvelles spécifiques).

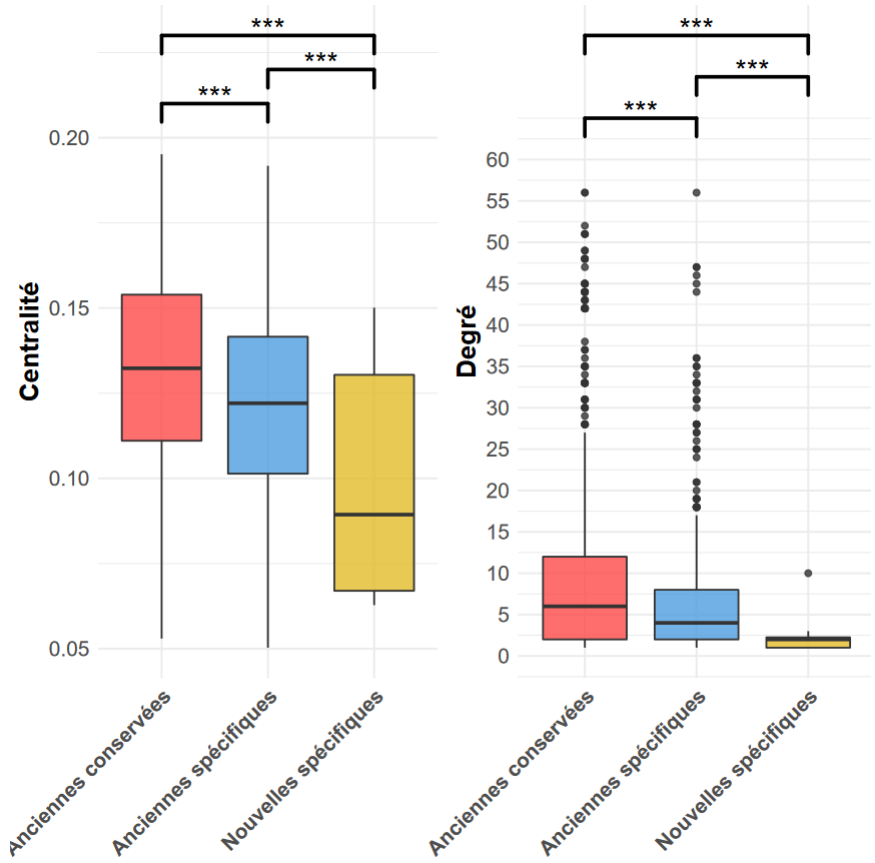


Figure 12.2 : A gauche La distribution de la centralité et **à droite** la distribution du degré des sommets en fonction de l'histoire évolutive des activités enzymatiques. Les boîtes rouges indiquent les activités enzymatiques conservées à plus de 85% et anciennes dans les espèces étudiées (anciennes conservées), en bleu les activités enzymatiques anciennes mais perdues chez certaines espèces au cours de l'évolution (anciennes spécifiques) et en jaune les activités enzymatiques nouvelles et spécifiques des champignons (spécifiques nouvelles). ******* signifie une p-value inférieure à 10-3 entre les distributions. Les différentes valeurs de p-value sont disponibles dans les annexes 3 et 4.

12.2 En fonction de la similarité de profil

Nous avons déterminé dans la partie II.7.5.5 que plusieurs activités enzymatiques avaient des profils similaires. Nous avons déterminé 15 groupes d'activités enzymatiques aux profils similaires. C'est-à-dire, les activités enzymatiques dans un même groupe sont présentes et absentes dans les mêmes groupes d'espèces. Nous avons aussi déterminé que certains de ces modules évolutifs forment des modules fonctionnels et se situent dans la même voie mais ce n'est pas le cas de toutes les activités enzymatiques aux profils similaires. Pour déterminer la localisation des activités enzymatiques avec des profils similaires dans le

réseau, et ainsi déterminer si ces activités enzymatiques sont colocalisées ensemble ou réparties aléatoirement dans le réseau, nous avons calculé la distance entre paire d'activités enzymatiques de profils similaires dans le réseau et comparé cette distance avec des sommets tirés aléatoirement dans le réseau (Figure 12.4). Pour rappel, la distance entre deux paires de sommets est le plus court chemin entre ces deux sommets. Une étude presque similaire a été faite par Yamada et ses collaborateurs (Yamada *et al.*, 2006). En effet, ces auteurs ont regardé la corrélation entre la distance par paires entre tous les sommets et la similarité par paires entre tous les profils d'un réseau métabolique global issu de plusieurs espèces qui couvrent les 3 domaines du vivant. Leur étude a montré qu'il y a une corrélation entre la similarité de profil et la distance dans le réseau : plus les profils sont similaires, plus ils sont proches dans le réseau.

Pour notre analyse, nous avons calculé la distance par paires entre profils similaires (dans le même groupe). Un des groupes que nous avons identifiés est constitué de 461 activités enzymatiques très fortement conservées. 71 activités enzymatiques sont réparties dans 15 groupes (espèces spécifiques) et 378 profils n'ont aucune similarité avec d'autres profils. Dans la Figure 12.3, nous avons comparé la distribution par rapport à la médiane de la distance entre les sommets tirés aléatoirement dans le réseau, le groupe de sommet avec des profils similaires mais fortement conservés et les autres groupes avec des sommets au profil similaire. Le groupe de sommets conservés a été séparé des autres groupes similaires car il représente 461/532 des activités enzymatiques qui ont une similarité de profil avec une autre activité enzymatique. Leur nombre risque de masquer l'information contenue dans les groupes de profils similaires qui sont spécifiques de certaines espèces.

Ainsi nous avons démontré que les activités enzymatiques très conservées sont plus proches entre elles que les activités enzymatiques non conservées tirées aléatoirement du réseau (la p-value du test de Mann-Whitney-Wilcoxon est de 10^{-5}). Ce résultat confirme l'impression visuelle que ces activités enzymatiques forment un module très conservé au centre du réseau. Par rapport aux activités enzymatiques très conservées et spécifiques tirées aléatoirement , les activités enzymatiques au profil similaire sont très proches entre elles dans le réseau métabolique (les p-value du test de Mann-Whitney-Wilcoxon sont de 10^{-16} et 10^{-42} respectivement). Même si les activités enzymatiques au profil similaire ne sont pas toutes localisées dans une même voie, ces modules évolutifs ne sont pas éloignés dans le réseau et sont probablement localisés dans des voies métaboliques reliées.

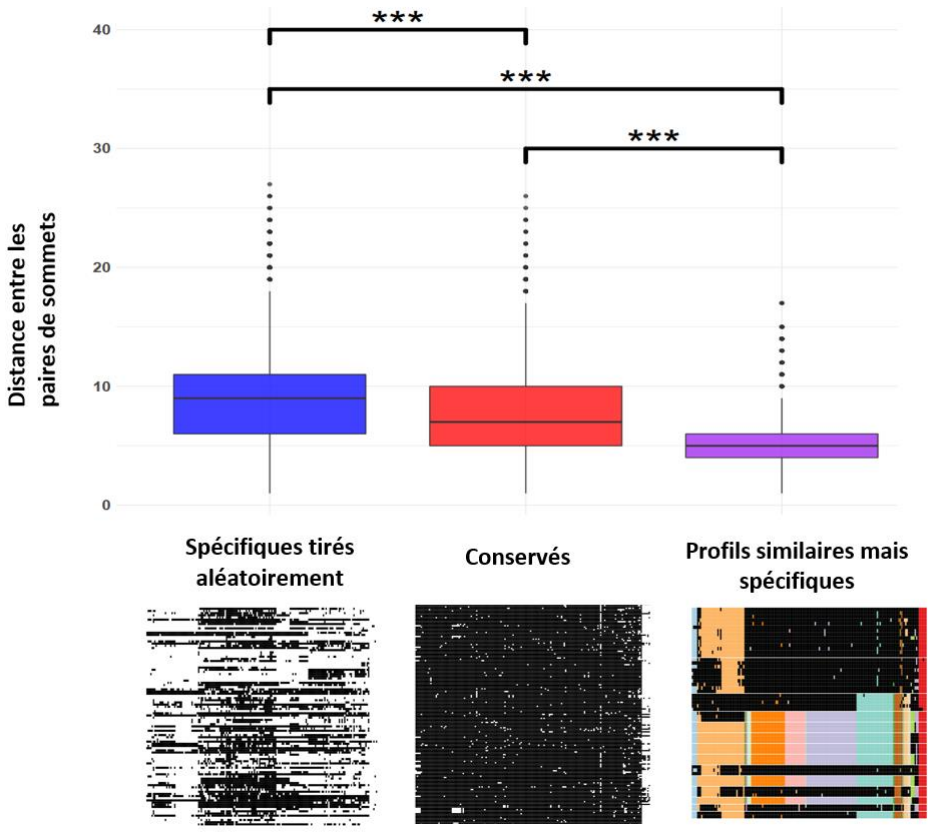


Figure 12.3 : Comparaison de la distribution entre les paires de distance dans le réseau entre les sommets: spécifiques de certaines espèces mais sélectionnés aléatoirement dans le réseau (en bleu), profils similaires mais conservés (en rouge) et entre profils similaires mais spécifiques de certaines espèces (en violet). Le profil des activités enzymatiques de chaque boîte est affiché au dessous du graphe. Une cellule noire indique la présence de l'activité enzymatique et une cellule blanche son absence. *** signifie une p-value inférieure à 10-3 entre les distributions.

Chapitre :

13 Caractéristiques évolutives des voies métaboliques

13.1 Les voies métaboliques essentielles et accessoires

Avec les informations évolutives des activités enzymatiques, nous avons inféré de manière formelle quelles sont les voies métaboliques essentielles (communes à toutes les espèces) et les voies métaboliques accessoires (spécifiques de certaines espèces).

Dans ce manuscrit, nous avons défini qu'une voie métabolique est une succession de réactions enzymatiques. D'après cette définition, pour faire une voie il faut minimum deux activités enzymatiques qui s'opèrent successivement. Ainsi, avec les informations évolutives des activités enzymatiques, nous avons décidé qu'une voie métabolique est conservée par toutes les espèces si au moins deux activités enzymatiques conservées (chez 85% des espèces étudiées) sont connectées dans la voie (Figure 13.1).

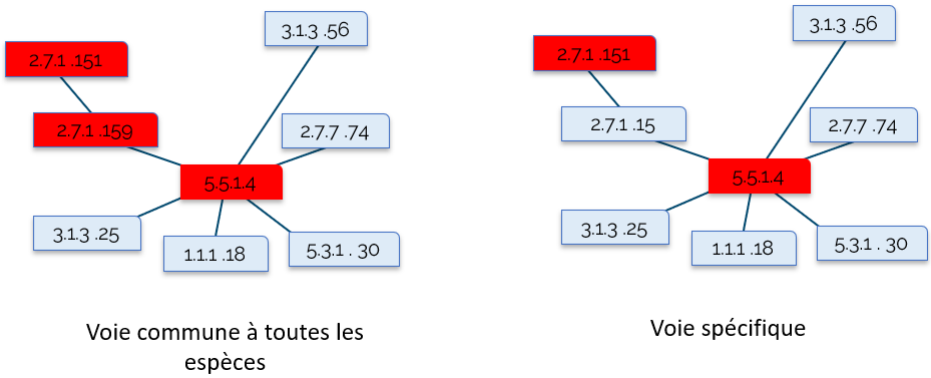


Figure 13.1 : Différence entre une voie commune à toutes les espèces et une voie spécifique. Les rectangles rouges indiquent des activités enzymatiques très conservées et les rectangles bleus des activités enzymatiques spécifiques de certaines espèces. Deux activités enzymatiques sont connectées si elles partagent un composé en commun dans leur réaction. **A gauche**: une voie commune à toutes les espèces avec au moins deux activités enzymatiques conservées sont connectées. **A droite** : aucune des activités enzymatiques conservées ne sont connectées entre elles.

En appliquant cette règle, nous avons identifié 76 voies métaboliques conservées par toutes les espèces et 28 voies métaboliques spécifiques de certaines espèces. Une voie métabolique conservée est forcément essentielle pour maintenir les organismes en vie.
Les voies métaboliques dans KEGG sont regroupées par super groupes en fonction des processus biologiques dans lesquels la voie est impliquée. Par exemple, KEGG regroupe 13 voies dans le super groupe du métabolisme des acides aminés (Figure 13.2).
La conservation des voies par super groupe est indiquée dans la Figure 13.2. La liste complète des voies par supergroupe ainsi que leur information évolutive est disponible dans l'annexe 3.
Dans cette analyse, nous constatons que toutes les voies associées aux métabolismes des acides aminés, de l'énergie, des nucléotides et la transcription sont conservées par presque

tous les champignons.
Les voies associées aux acides aminés incluent les voies de dégradations et de synthèses. Or nous avons démontré avec les méthodes de profilages phylogénétiques que la voie de la dégradation de la valine, leucine et isoleucine est absente des Saccharomycetes et Schizosaccharomycetes (Figure 7.15). Cette voie est inférée comme présente chez tous les champignons car quelques activités enzymatiques très conservées qui partagent un composé en commun sont présentes dans la voie. Mais la particularité de ces activités enzymatiques est qu'elles sont partagées entre plusieurs voies, par exemple 1.1.1.35 et 2.3.1.16 sont présentes ensemble dans d'autres voies qui sont fatty acid elongation, fatty acid degradation et benzoate degradation. Leur forte conservation est probablement due à leur activité essentielle dans plusieurs voies.
Autre cas étonnant, dans les voies associées au métabolisme de l'énergie, une grande majorité des activités enzymatiques associées à la fixation du carbone chez les organismes photosynthétiques sont présentes chez les champignons (Figure 10.4 A), raison pour laquelle la voie est inférée comme présente. Sur les 16 activités enzymatiques de la voie, 15 sont présentes dans au moins une autre voie métabolique.
La réutilisation des activités enzymatiques et leur forte conservation font que certaines voies normalement absentes chez les champignons sont identifiées à tort chez les champignons sur la base de la présence de certaines activités enzymatiques. La voie peut même être considérée comme commune à toutes les espèces car des successions de réactions enzymatiques très conservées ont été identifiées dans la voie. Ces successions constituent en effet des modules qui sont réutilisés dans d'autres voies.

Les voies spécifiques de certaines espèces sont majoritairement des voies métaboliques associées à la synthèse des métabolites secondaires. Il est à noter que certaines voies de synthèse de métabolites secondaires sont classées dans un autre super groupe à cause du bloc de construction initial du métabolite mais aussi de l'intersection avec les voies métaboliques essentielles. Par exemple la voie C5 Branched dibasic acid qui est associée à la synthèse de plusieurs métabolites secondaires comme l'acetoine (Bae *et al.*, 2016) et l'itaconate (Wierckx *et al.*, 2020) est classée dans le supergroupe des métabolismes des sucres parce que le bloc initial de construction de ces deux produits est le pyruvate et l'Acetyl-Coa.
La majorité des autres voies spécifiques sont associées aux dégradations des autres acides aminés et au métabolisme et dégradation des xénobiotiques. Ce sont les voies associées au métabolisme des substances étrangères présentes dans le milieu et étrangères pour l'organisme.

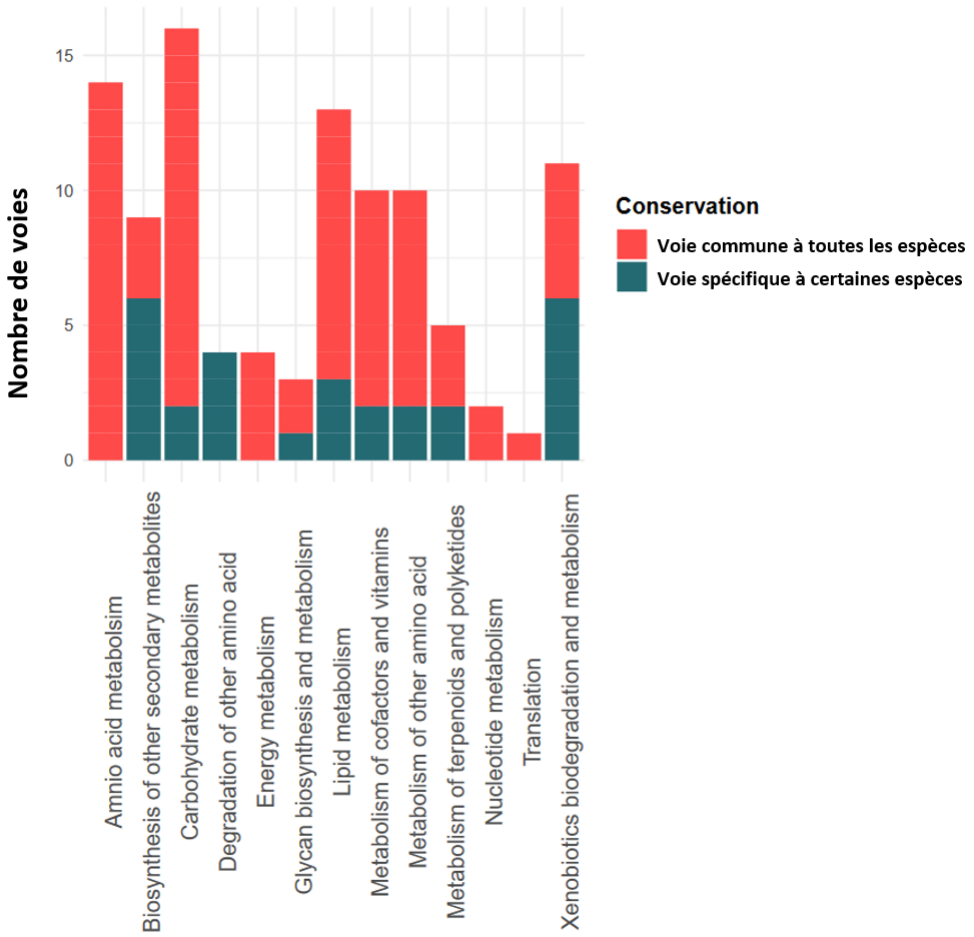


Figure 13.2 : Conservation des voies par supergroupe. En rouge les voies communes à toutes les espèces et en bleu les voies spécifiques de certaines espèces. Les voies sont regroupées par super groupe selon la classification de KEGG.

Les voies métaboliques essentielles sont communément acceptées comme conservées chez toutes les espèces alors que les voies métaboliques accessoires ne sont présentes que chez quelques espèces. Avec la théorie des graphes, nous avons caractérisé d'un point de vue topologique les voies communes à toutes les espèces et les voies spécifiques. En comparant le diamètre de ces voies, nous constatons que les voies communes ont un diamètre plus élevé que les voies spécifiques (la p-value du test de Mann-Whitney-Wilcoxon est de 10^{-57}). Les voies à grand diamètre sont donc plus conservées que les voies avec de petits diamètres (Figure 13.3).
D'un point de vue biochimique, les voies communes contiennent de longues séquences de réactions pour transformer un substrat en produit alors que les voies spécifiques ne nécessitent que quelques réactions enzymatiques pour transformer un substrat en métabolite secondaire (Figure 13.4).

Cette différence peut s'expliquer par l'importance des voies communes. Les voies communes assurent un processus biologique essentiel pour la survie et sont le fruit de

plusieurs milliards d'années d'évolution et d'adaptation à différents environnements. L'essentialité de la voie est une pression de sélection très forte ce qui a nécessité plusieurs étapes d'adaptation pour assurer la fonction principale de la voie et assurer la survie probablement par l'ajout de nouvelles réactions. Cette évolution a permis un meilleur contrôle des métabolites intermédiaires afin de prémunir de l'accumulation des intermédiaires indésirables pour mieux réguler le métabolisme mais aussi pour pourvoir la voie en précurseur.
Les voies accessoires sont majoritairement associées à l'adaptation dans un environnement particulier ou pour éliminer un excédent un composé potentiellement toxique. Ce sont probablement des voies spécifiques qui n'ont pas pour vocation à s'adapter à plusieurs environnements, par conséquent la pression évolutive autour de ces voies est moins forte et la succession des activités enzymatiques a peu évolué.

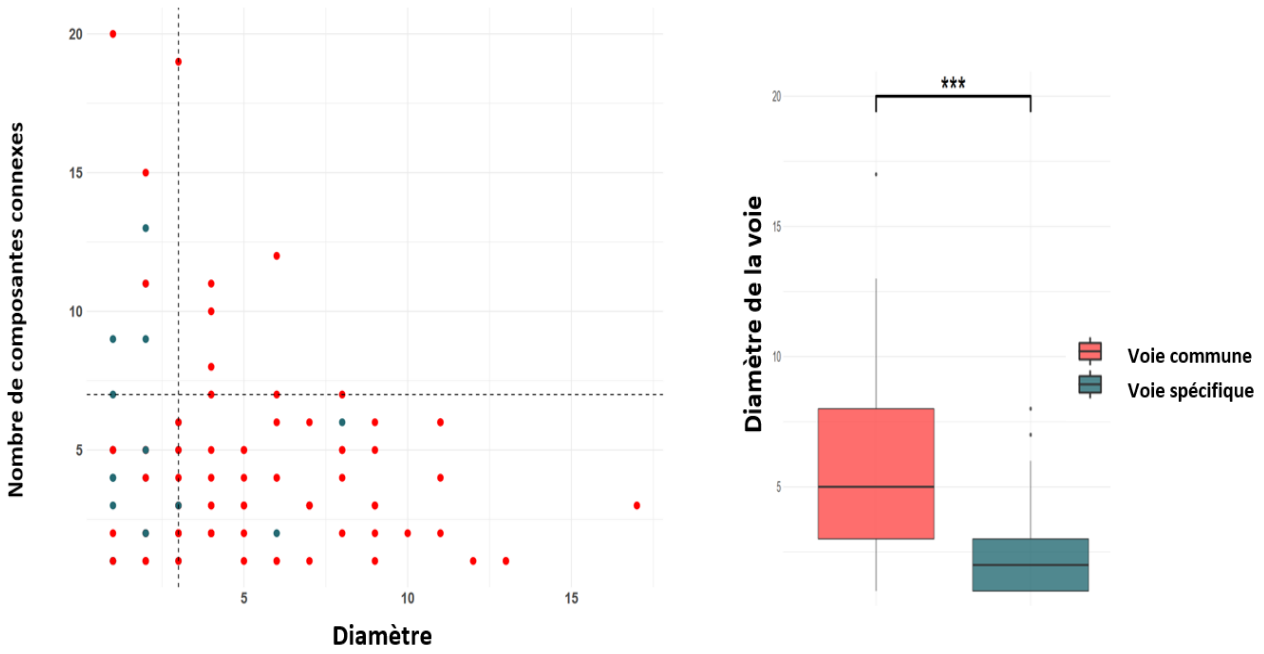


Figure 13.3: Comparaison topologique entre les voies spécifiques (en bleu) et les voies communes (en rouge). **À gauche** : le diamètre de chaque voie en fonction du nombre de composantes connexes. La ligne discontinue verticale indique un diamètre de 3. Les voies qui ont un diamètre supérieur ou égale à 3 sont des voies à grands diamètres. Les voies à grands diamètres sont majoritairement très conservées. **À droite** : comparaison des diamètres de chaque voie en fonction de la conservation. ******* signifie une p-value inférieure à 10-3 entre les distributions.

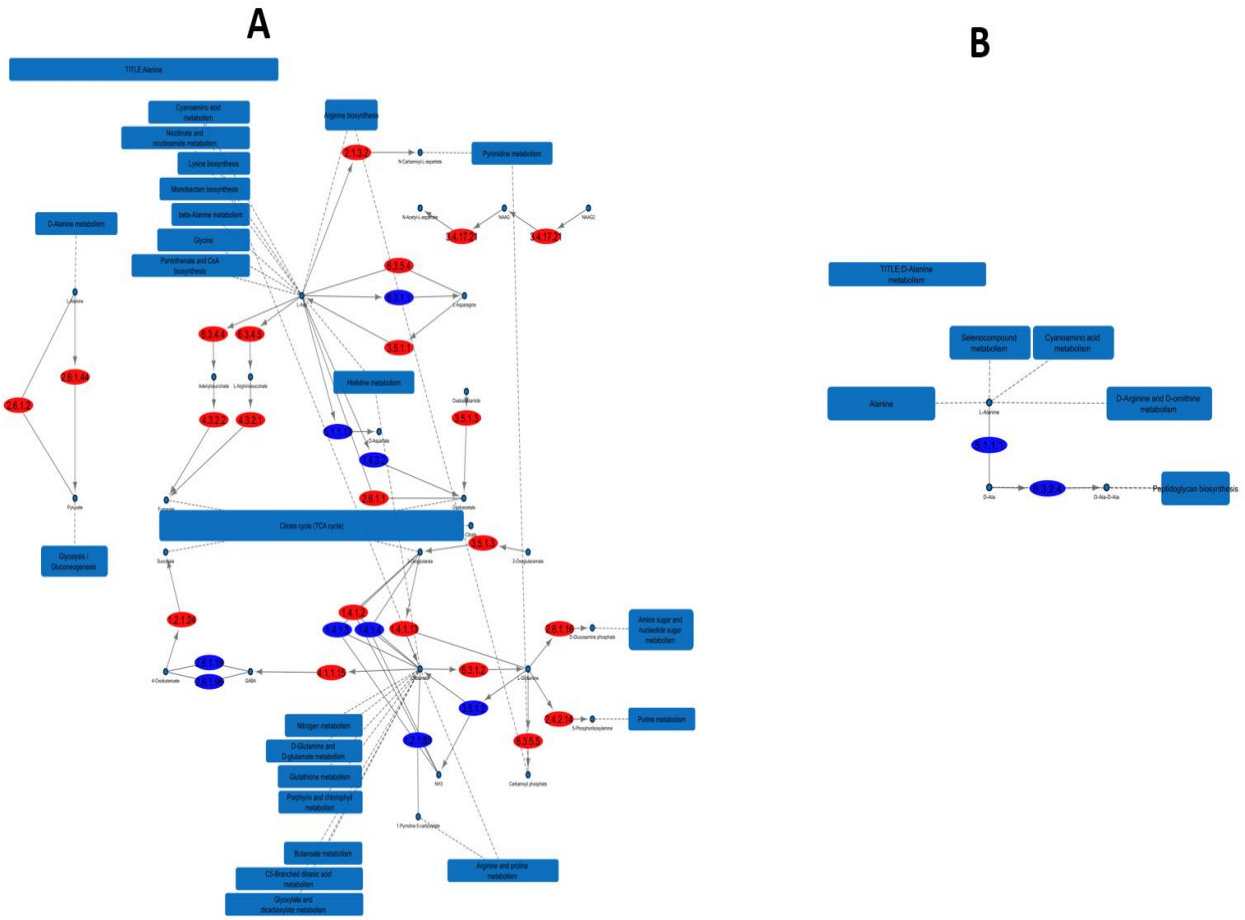


Figure 13.4 : Exemple d'une voie commune à toutes les espèces et d'une voie spécifique de certaines espèces. Les activités enzymatiques très conservées sont en rouge et les activités enzymatiques spécifiques en bleu **(A)** Alanine, aspartate and glutamate metabolism. La voie est composée de 29 activités enzymatiques chez les champignons et possède un diamètre de 4. C'est une voie commune à toutes les espèces **(B)** D-alanine metabolism. La voie est composée de 2 activités enzymatiques et possède un diamètre de 1. C'est une voie spécifique de quelques espèces.

13.2 Différence d'un point de vue topologique entre voies communes et accessoires

13.2.1 Position des voies dans le réseau en fonction de la conservation

Le réseau métabolique a été construit en connectant les voies métaboliques entre elles. Deux voies métaboliques sont connectées si elles partagent un composé en commun (comme indiqué dans le fichier KGML) et il faut que les activités enzymatiques qui utilisent ce métabolite dans les deux voies soient présentes chez les champignons. En construisant le réseau métabolique, nous avons aussi construit un réseau des voies métaboliques pour observer la localisation des voies métaboliques dans le réseau.

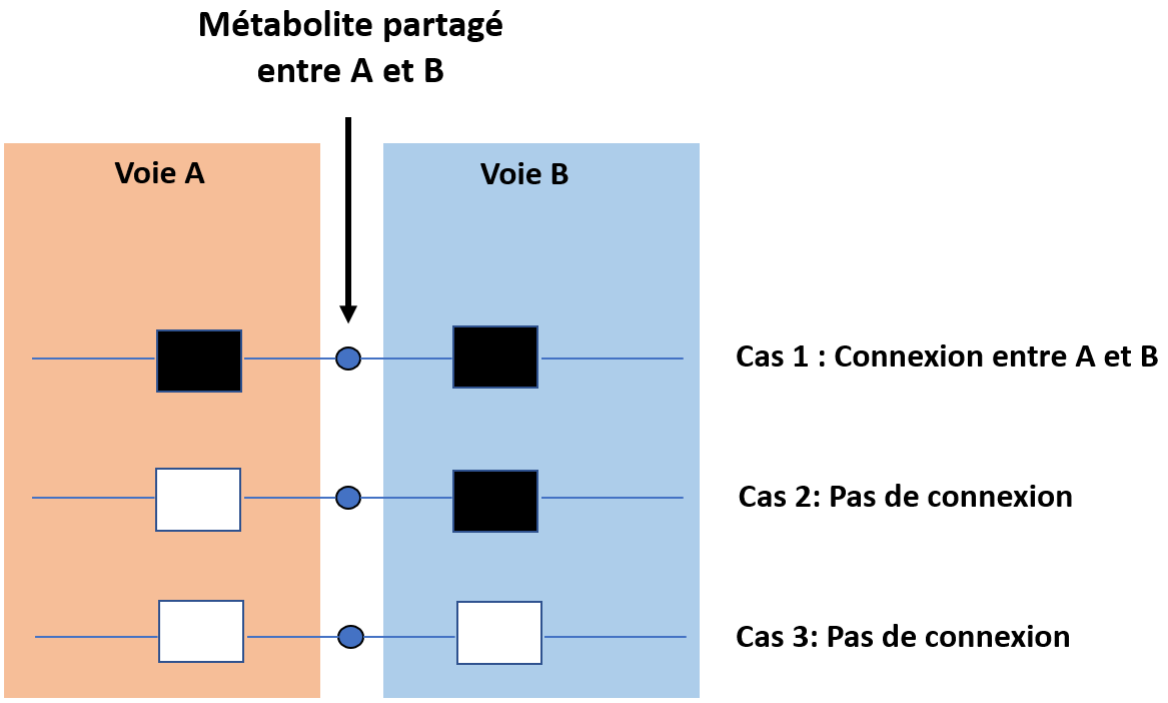


Figure 13.5 : Définition des liens entre les voies métaboliques. Les rectangles noirs indiquent des activités enzymatiques présentes chez les champignons et en blanc des activités enzymatique absentes chez les champignons. Dans le cas 1, une activité enzymatique (en noir) qui traite le composé en commun est présente dans les deux voies A et B, une connexion est établie entre la voie A et la voie B. Dans le cas 2, il n'y a pas d'activité enzymatique qui traite le composé dans la voie A, il n'y a pas de connexion entre les deux voies. Dans le cas 3, aucune activité enzymatique qui traite le métabolite en commun n'est présente dans les deux voies chez les champignons, il n'y a pas de connexion entre les deux voies.

Les informations évolutives sur chacune des voies inférées dans la partie précédente vont permettre de comprendre comment les voies communes à toutes les espèces et les voies spécifiques de certaines espèces sont localisées dans le réseau (Figure 13.6).
Le réseau des voies métaboliques construit est constitué de 17 composantes connexes. La plus grande composante connexe contient 85 voies métaboliques. Les 16 autres composantes sont de taille 1. Ce qui signifie que ce sont des voies probablement isolées du réseau. Ces voies isolées peuvent être artificielles dues aux informations incomplètes dans les voies métaboliques, certaines activités enzymatiques de la voie n'ont pas été détectées chez les champignons ou certaines activités enzymatiques n'ont pas encore été caractérisées. Ces absences font qu'il est impossible de connecter certaines voies entre elles.

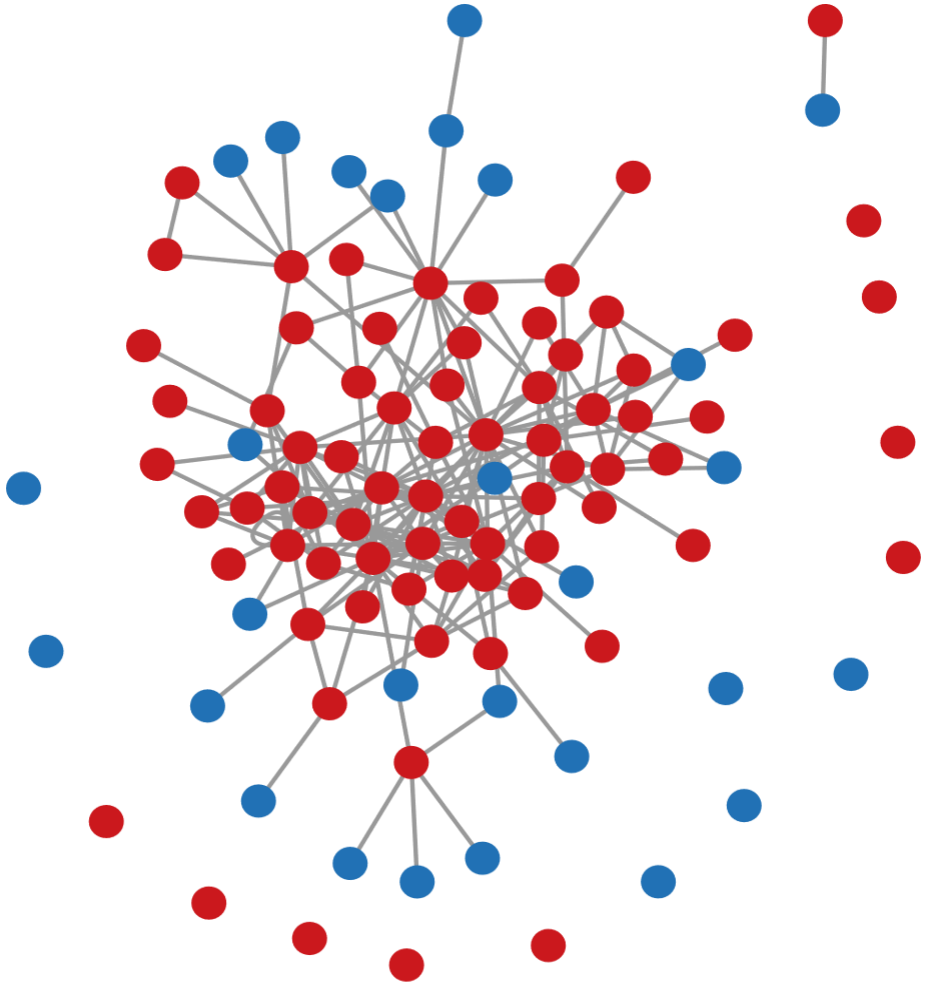


Figure 13.6 : Réseau des voies métaboliques en fonction de la conservation des voies métaboliques. Les sommets représentent les voies métaboliques et deux sommets sont connectés s'il y a un composé en commun entre les voies et les activités enzymatiques autour de ce composé dans les deux voies sont présentes chez les champignons. En rouge les voies communes à toutes les espèces et en bleu les voies spécifiques de certaines espèces.

Nous avons comparé le degré et la centralité des voies métaboliques communes à tous les champignons et les voies spécifiques à certaines espèces en comparant la distribution des données par rapport à la médiane (Figure 13.7).
Dans ce réseau des voies métaboliques, les voies communes à toutes les espèces sont plus connectées et en position centrale dans le réseau. Les voies spécifiques de certaines espèces sont localisées en périphérie du réseau (la p-value du test de Man-Whitney-Wilcoxon est de 10^{-33}) et de plus sont de très faible degré (la p-value test de Man-Whitney-Wilcoxon est 10^{-44}). Il est à noter que la majorité de ces voies spécifiques sont de degré 1. Un degré de 1 signifie que la voie ne partage un composé en commun qu'avec une seule autre voie qui lui sert probablement de précurseur. Les voies isolées du réseau métabolique sont soit artificielles, soit des réactions enzymatiques n'ont aucun lien avec le réseau métabolique principal. Par exemple, la voie de dégradation de l'Atrazine n'a aucun lien avec les autres voies métaboliques. Le processus de dégradation est totalement indépendant des autres voies (Figure 13.8)

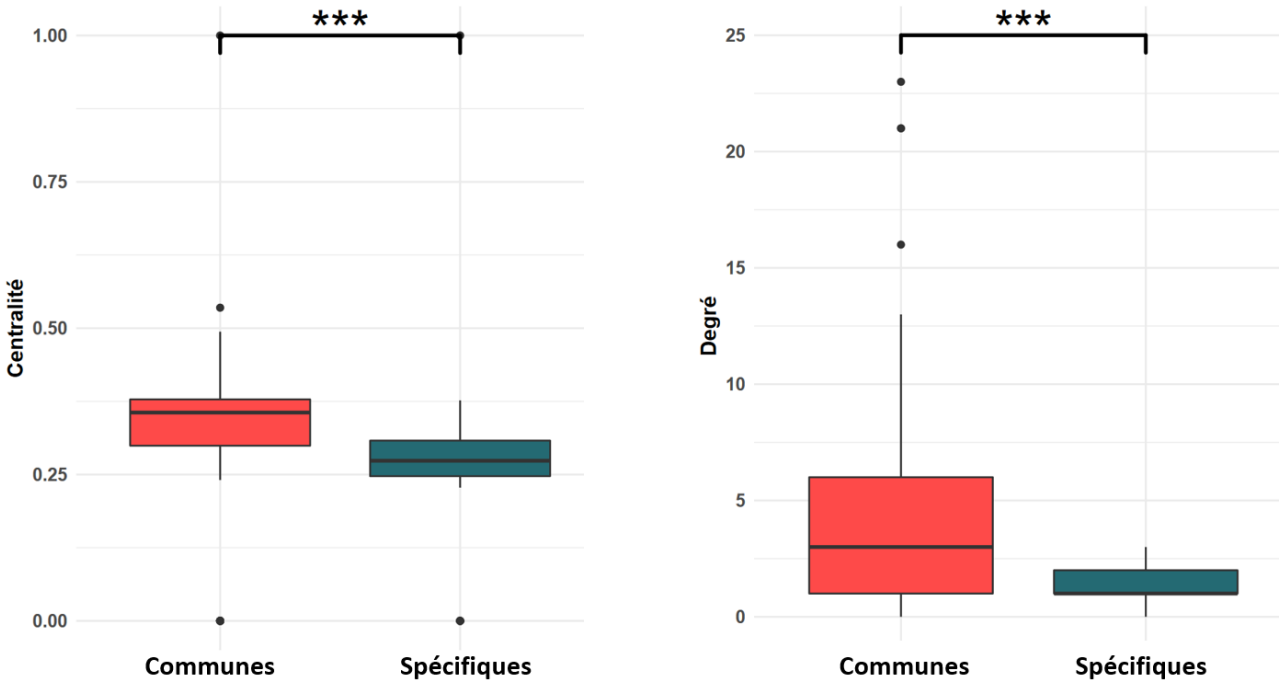


Figure 13.7. Comparaison de la centralité (à gauche) et des degrés (à droite) dans le réseau des voies métaboliques entre les voies communes à toutes les espèces et les voies spécifiques à certaines espèces. Les boîtes rouges représentent les voies communes et les boîtes bleues les voies spécifiques. *** signifie une p-value inférieure à 10-3 entre les distributions.

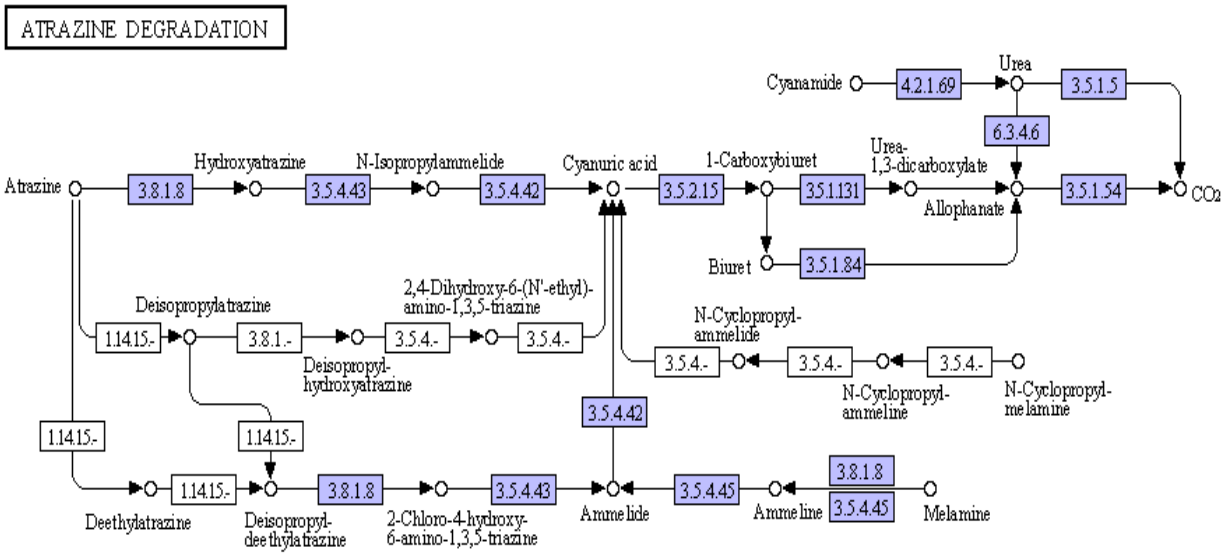


Figure 13.8: La voie de dégradation de l'atrazine. Cette voie ne partage aucun composé en commun avec les autres voies car aucun lien vers une autre voie n'est indiqué dans KEGG.

13.2.2 Localisation des voies dans le réseau en fonction des super groupes

Les supers groupes dans KEGG sont définis en fonction de leur processus global par rapport à un type de métabolite (acide aminé, nucléotide, sucre).
L'agencement des voies métaboliques par super groupe (Figure 13.9) montre que les voies d'un super groupe sont colocalisées dans le réseau (Figure 13.10). Pour démontrer cela, nous avons comparé la distribution par rapport à la médiane de la distance par paires entre les voies d'un même super groupe et des voies choisies aléatoirement dans le réseau (Figure 13.10) (la p-value du test de Mann-Whitney-Wilcoxon est 10^{-13}).
C'est un résultat attendu. Cette colocalisation dans le réseau s'explique par un ensemble d'activités enzymatiques partagées par les voies d'un même supergroupe. Ces activités enzymatiques partagées sont probablement des processus identiques dans le métabolisme d'un même type de métabolite (Peregrín-Alvarez *et al.*, 2009).

La distribution des degrés montre que les voies associées au métabolisme des sucres et des acides aminés sont les plus connectées (Figures 13.9 et 13.10). Toutes les autres voies s'articulent autour, ce qui témoigne de l'importance capitale de ces voies dans le métabolisme.
Nous observons aussi que les voies associées au métabolisme des xénobiotiques sont majoritairement isolées et en périphérie. Ces voies traitent des molécules étrangères et dangereuses pour le métabolisme ce qui nécessite probablement un isolement de ces processus biologiques.
Dans ce réseau, on observe aussi que la biosynthèse des métabolites secondaires (Figure 13.9 en fushia) a comme principale précurseur les acides aminés (Figure 13.9 en orange).
Les voies associées au métabolisme des lipides sont séparées dans le réseau dont 6 voies ne

sont pas connectées au réseau principal. D'après les représentations KEGG, les voies associées au métabolisme de l'acide linoléique, de l'acide alpha linoléique et de l'acide arachidonique ne partagent aucun composé commun avec d'autres voies du réseau principal. Les trois autres voies semblent être déconnectées du réseau à cause des informations manquantes.

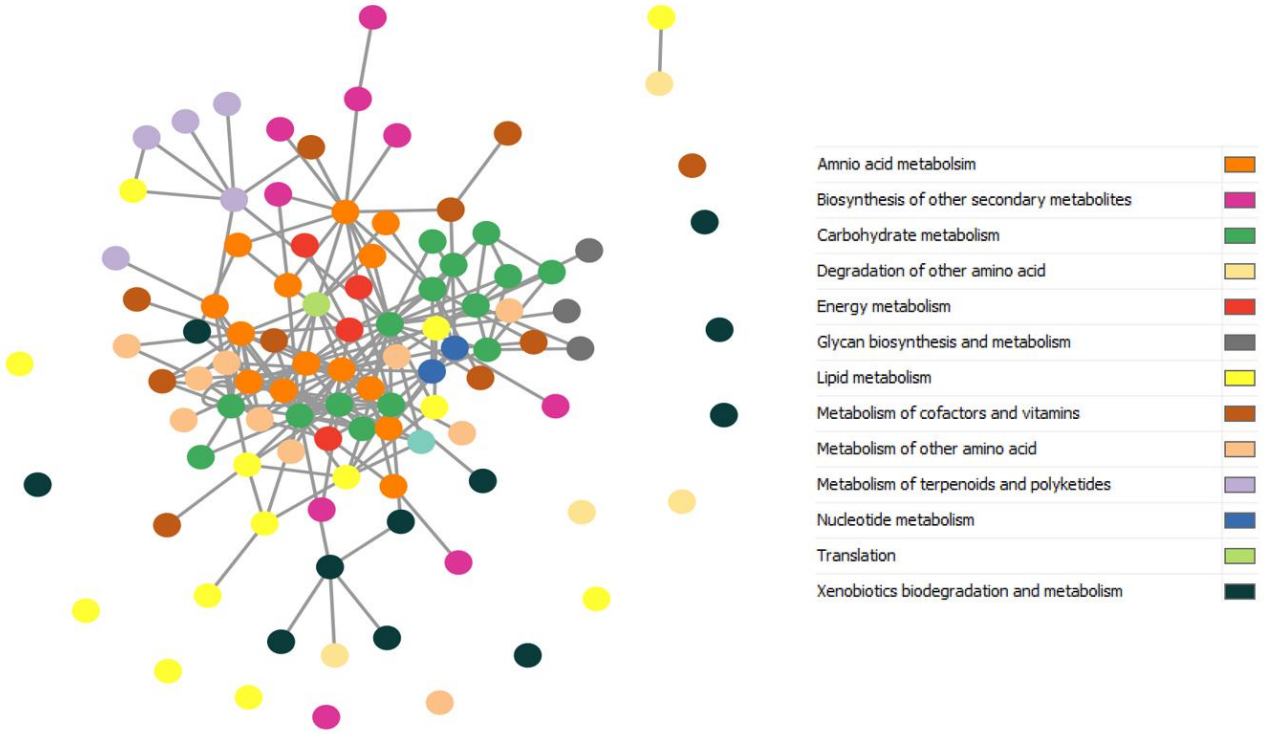


Figure 13.9 : Réseau des voies métaboliques. Les voies métaboliques sont colorées en fonction des supers groupes définis par KEGG. Deux voies métaboliques sont reliées entre elles si un lien (Figure 13.5) existe entre les deux voies.

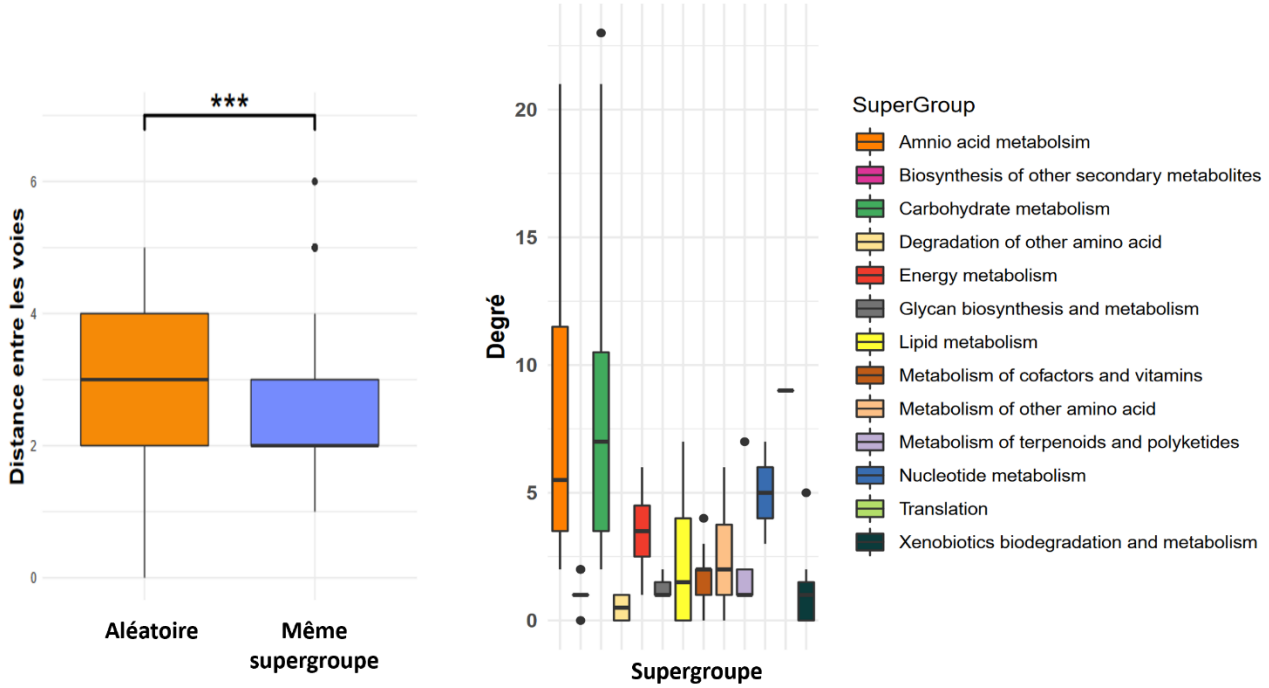


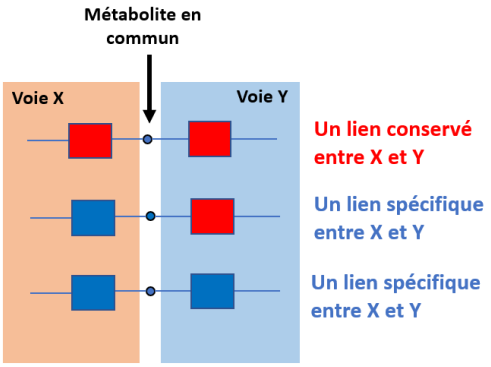
Figure 13.10 : A gauche : Comparaison des distances entre des voies aléatoires (boite orange) et des voies d'un même super groupe (boite bleue). **A droite :** Degré des voies dans le réseau par super groupe. Chaque couleur correspond à un super groupe.

13.2.3 Analyse des liens entre les voie métaboliques

Les liens entre les voies métaboliques sont déterminés à partir de la présence des activités enzymatiques à proximité des métabolites en commun entre les voies. Avec les informations évolutives des activités enzymatiques, nous avons analysé la conservation des liens entre les voies métaboliques en identifiant les liens qui sont conservés entre toutes les espèces et les liens qui sont spécifiques.

Un lien entre deux voies A et B est conservé si au moins une des activités enzymatiques dans A et au moins une des activités enzymatiques dans B qui se situent autour d'un composé en commun entre les deux voies sont communes à toutes les espèces (conservation supérieure 85%). Si aucune activité enzymatique n'est conservée par toutes les espèces , le lien entre les deux voies est considéré comme étant spécifique d'une espèce ou d'un groupe d'espèces (Figure 13.11).

Si plusieurs métabolites sont en commun entre les deux voies, (i) le lien global entre les deux voies est conservé si tous les liens avec les métabolites en commun sont conservés ou au moins un lien avec un métabolite est conservé. L'absence de lien spécifique dans certaines espèces ne va pas empêcher la connexion entre les deux voies du fait de la présence du lien conservé. (ii) Le lien global entre les deux voies est spécifique si tous les liens avec les métabolites sont spécifiques (Figure 13.11).



Lien avec un métabolite A	Lien avec un métabolite B	Lien global entre 2 voies en fonction des liens A et B
Conservé	Conservé	**Conservé**
Spécifique	Conservé	**Conservé**
Spécifique	Spécifique	**Spécifique**

Figure 13.11 : Définition de la conservation des liens entre les voies. À gauche, les activités enzymatiques sont représentées par les rectangles. En rouge les activités enzymatiques conservées par toutes les espèces et en bleu les activités enzymatiques spécifiques de certaines espèces. La conservation des liens dépend de la conservation des activités enzymatiques autour du métabolite en commun entre les voies X et Y. Si une des activités enzymatiques est spécifique, le lien entre les deux voies à partir du métabolite étudié est spécifique. **À droite** : cas où plusieurs métabolites sont en commun entre deux voies. Les deux premières colonnes définissent le type de lien avec le métabolite A et le métabolite B. La troisième colonne est le lien global défini en fonction des liens pour chaque métabolite. Par exemple, pour la deuxième ligne, le lien avec le métabolite A est spécifique et le lien avec le métabolite B est conservé. Par conséquent, le lien global entre les deux voies est conservé.

Nous avons identifié 40 liens strictement conservés entre les voies. C'est-à-dire qu'aucun lien à partir d'un métabolite commun entre les voies n'est spécifique. 64 liens ont un lien global conservé où certains points de liaisons entre les voies sont spécifiques mais un ou plusieurs liens sont communs entre toutes les espèces, et 92 liens spécifiques.
En analysant la connexion entre les voies métaboliques et en ne se focalisant que sur la grande composante connexe (composée de 66 voies communes à toutes les espèces et 21 voies qui sont spécifiques de certaines espèces), nous avons enlevé tous les liens spécifiques entre les voies et seulement gardé les liens conservés. Nous avons observé que 63 des voies communes à toutes les espèces continuent de former qu'une seule composante connexe et toutes les voies spécifiques se sont déconnectées de cette composante (Figure 13.12). Ceci met en évidence que la structure globale et l'agencement des voies communes à toutes les espèces sont très conservés chez tous les champignons et au cours de l'évolution des champignons. Cette structure très conservée assure probablement les fonctions essentielles pour la survie des organismes et leur connexion est essentielle.

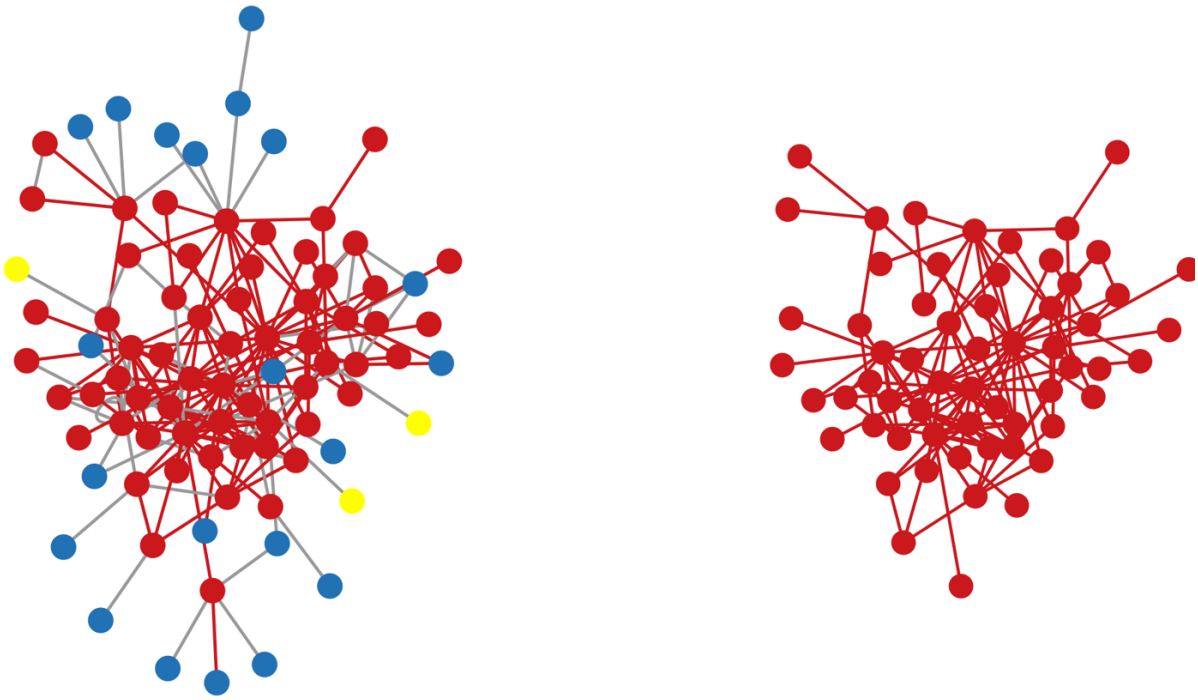


Figure 13.12 : Réseau des voies métaboliques. Les cercles rouges et jaunes indiquent les voies métaboliques communes à toutes les espèces, les cercles bleus les voies spécifiques de certaines espèces. Les liens en rouge sont des liens globaux conservés, en gris les liens globaux spécifiques. **À gauche.** La composante principale du réseau des voies métaboliques contenant toutes les voies et tous les liens. **À droite.** La composante principale du réseau si seul les liens conservés entre les voies sont gardés. Les voies colorées en jaune dans la Figure à gauche sont des voies communes absentes de cette composante principale composée seulement de lien global conservé.

Chapitre :

14 Rôles des activités spécifiques dans les voies.

Nous avons identifié une forte conservation des voies métaboliques et des activités enzymatiques chez les champignons. Dans les voies métaboliques conservées par toutes les espèces, un nombre d'activités enzymatiques spécifiques non négligeable est présent. 337 activités enzymatiques spécifiques de certaines espèces sont présentes dans les voies communes à toutes les espèces. Cela signifie qu'une partie de la voie est commune entre toutes les espèces (processus global et essentiel) mais que des parties spécifiques diffèrent entre différentes espèces et témoignent de la dynamique évolutive des voies métaboliques.

Dans ce chapitre, nous allons présenter l'étude de la position de ces activités enzymatiques spécifiques de certaines espèces dans les voies communes à toutes les espèces pour comprendre comment ces activités enzymatiques spécifiques ont façonné les voies aux cours de l'évolution des champignons.
Pour ce faire, nous avons défini un rôle pour chaque activité enzymatique spécifique en fonction de sa position dans les voies métaboliques par rapport aux activités enzymatiques conservées de la voie chez toutes les espèces.
Parmi les activités enzymatiques spécifiques dans les voies communes, 335 activités enzymatiques sont des activités qui ont été perdues par certaines espèces au cours de l'évolution et seulement 2 ont spécifiquement apparu chez les champignons. Ces 2 activités enzymatiques font partie des 8 activités enzymatiques qui sont nouvelles et spécifiques des champignons et pour lesquelles aucun homologue avec la même activité enzymatique n'a été détecté en dehors des champignons. Par ailleurs, comme nous l'avons vu précédemment, ces 8 activités enzymatiques ont été confirmées par la littérature comme étant spécifiques des champignons.

14.1 Définitions des différents rôles possibles

Les rôles des activités spécifiques de certaines espèces ont été déterminés à la fois par rapport à leur position dans les voies métaboliques communes mais également par rapport aux activités enzymatiques conservées par toutes les espèces de la voie. Il est à noter que la majorité de ces activités enzymatiques qui ne sont pas conservées sont en général perdues chez certaines espèces. Nous allons donc essayer de comprendre où ces pertes se sont principalement produites dans les voies et pourquoi elles sont tolérées.

Dans une voie on suppose que les activités enzymatiques conservées constituent la partie essentielle de la voie qui est donc commune à toutes les espèces. Les activités enzymatiques spécifiques participent quant à elles à la diversité de la voie entre les espèces. Dans les voies métaboliques, comme décrit dans les parties précédentes, certains sommets sont isolés et ne sont pas connectés avec un autre sommet dans la voie. Ces sommets (activités enzymatiques) ne seront pas traités.

Ainsi nous avons défini 6 rôles pour les activités enzymatiques spécifiques dans les voies communes. Ces rôles ont été choisis en raison de leur intérêt dans le fonctionnement et/ou la construction de la voie. Ces différents rôles ont été traduits dans le cadre de la théorie

des graphes pour permettre leur détection automatique par une analyse informatique.

Alternative à une/des activité(s) enzymatique(s) conservée(s) :

Un sommet associé à une activité enzymatique est considéré avoir un rôle alternatif s'il permet de substituer une activité enzymatique ou une série d'activités enzymatiques conservées. Les activités enzymatiques alternatives permettent de transformer un métabolite A en un métabolite B en parallèle avec d'autres activités(s) enzymatique(s) non conservée(s). La perte des activités alternatives ne va pas « casser » la voie métabolique.

La détection de ces sommets alternatifs est effectuée de manière automatique pour chaque voie à l'aide d'un script R et en utilisant la propriété topologique de ces sommets.
D'un point de vue de la théorie des graphes, une activité enzymatique est alternative si dans la représentation de la voie sous forme de graphe des activités enzymatiques elle respecte les propriétés suivantes :

(i) Si on retire l'activité enzymatique spécifique, l'absence de cette activité enzymatique ne créera pas deux composantes connexes (Figure 14 .1).
(ii) Si on part d'une voie constituée seulement des activités enzymatiques conservées, l'ajout de l'activité enzymatique ne va pas connecter deux nouvelles parties de la voie.

Les activités enzymatiques en alternatives peuvent être organisées en groupe (série) d'activités enzymatiques. Un groupe spécifique est seulement composé d'activités spécifiques.
Si le groupe est considéré comme un seul nœud dans le graphe, il a les caractéristiques d'un sommet alternatif à une autre activité enzymatique conservée.

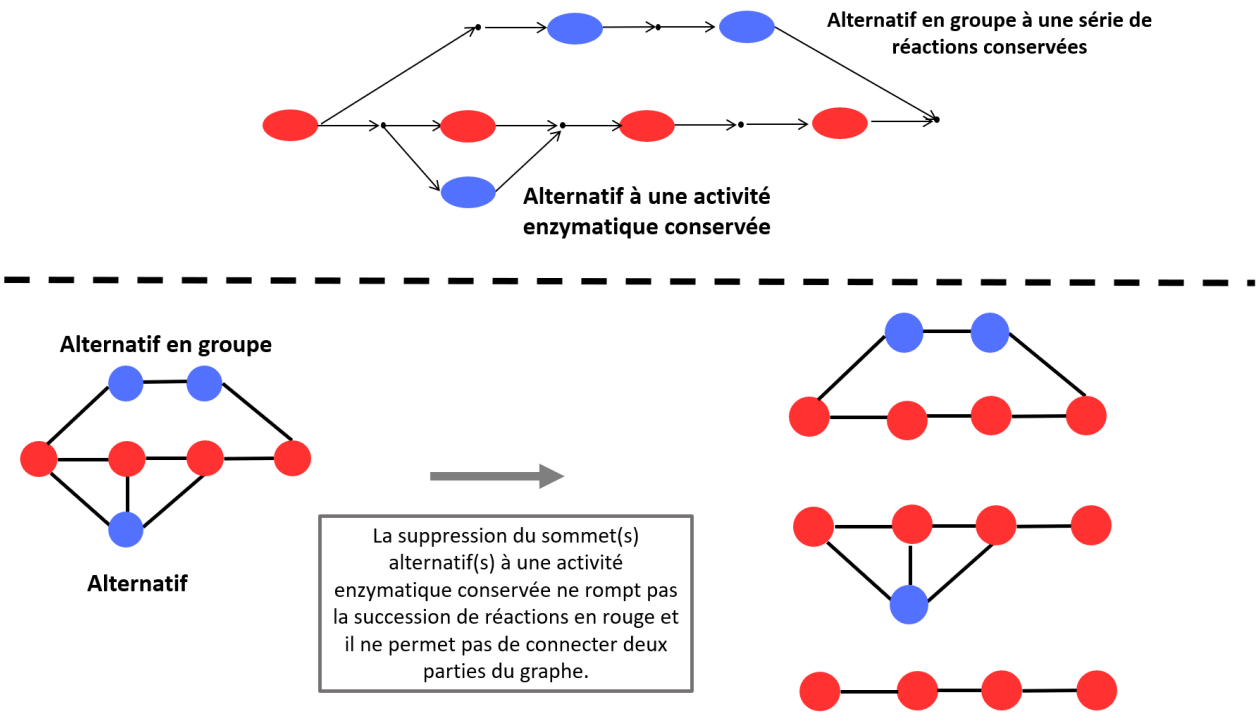


Figure 14.1 : Description des activités enzymatiques alternatives à une (des) activité(s) enzymatique(s) conservée(s). En rouge les activités enzymatiques conservées et en bleu les activités enzymatiques spécifiques. **En haut.** La représentation de la voie avec les activités enzymatiques et les métabolites. Les activités enzymatiques en bleu sont alternatives d'une activité enzymatique conservée ou de plusieurs activités enzymatiques conservées. **En bas.** La voie sous forme de graphe des activités enzymatiques. Deux activités enzymatiques sont connectées si elles partagent un métabolite en commun dans leur réaction. Dans une représentation sous forme de graphe, l'absence des sommets spécifiques en alternatifs ne casse pas la succession de réactions des sommets conservées chez toutes les espèces. À l'inverse, leur présence ne permet pas non plus de connecter de nouvelles parties du graphe.

Par exemple, dans la voie de la voie glycolyse, l'hexokinase : 2.7.1.1 (conservée) peut être substituée par le glucokinase : 2.7.1.2 qui est spécifique de certaines espèces (Encadrée en bleu dans la Figure 14.4). Les deux activités enzymatiques interviennent dans la première étape de la glycolyse pour convertir le glucose en glucose 6-phosphate. L'hexokinase est très peu spécifique mais a beaucoup d'affinité pour la phosphorylation d'hexoses comme le D-glucose, le D-mannose et le D-fructose. Le glucokinase, à l'inverse, est très spécifique au glucose mais présente une affinité faible aux autres hexoses (Panneman *et al.*, 1996). Il a été démontré que l'hexokinase est essentiel pour la croissance et le développement chez quelques champignons étudiés et l'absence du glucokinase retarde seulement la germination (Fleck and Brock, 2010; Laurian *et al.*, 2019).
Toujours dans la voie de la glycolyse, 1.2.1.3 (conservée) peut être substituée par 1.2.1.5 (spécifique). 1.2.1.3 a une plus grande spécificité de substrat que 1.2.1.5 (Datta *et al.*, 2017).
Dans la voie de la biosynthèse des acides gras, l'acide gras synthase 2.3.1.86 (conservée) peut être substituée par 2.3.1.85. L'activité 2.3.1.86 est très conservée chez les champignons

et 2.3.1.85 est très conservée chez les animaux (Encadré en bleu dans la Figure 14.2 B). Mais l'acide gras synthase présente chez les champignons est plus efficace que les autres acides synthase (Gajewski *et al.*, 2017).
Dans la voie du métabolisme de la tyrosine, il y a la présence d'une transaminase (2.6.1.5) qui permet de substituer 3 autres transaminases (2.6.1.1, 2.6.1.57, et 2.6.1.9 qui sont conservées) qui utilisent comme substrat la tyrosine (Encadré en bleu dans la Figure 14.2 A). La particularité de 2.6.1.5 est qu'elle a une forte spécificité pour la tyrosine alors que les autres ont une affinité pour plusieurs substrats.

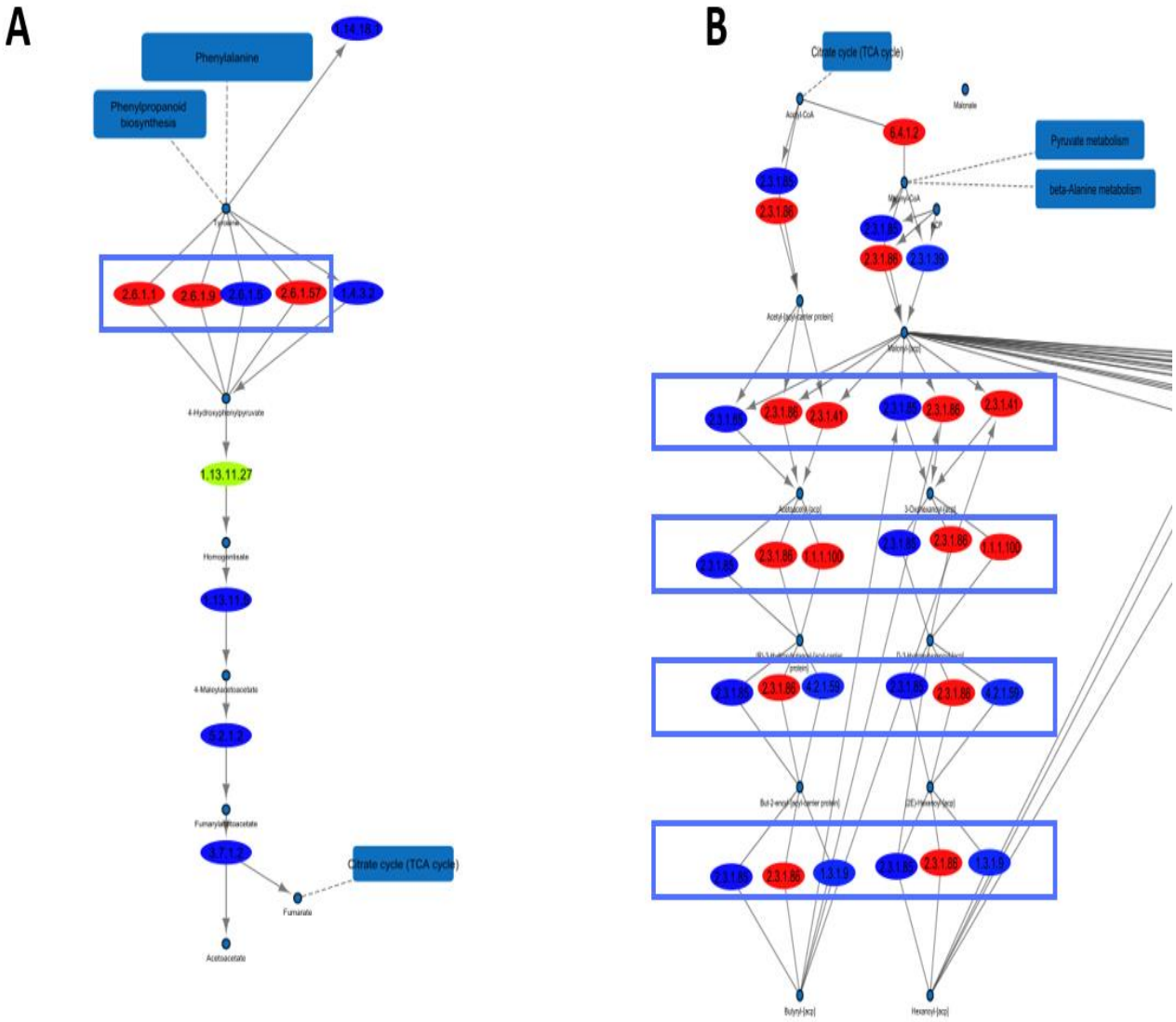


Figure 14.2 : Des exemple d'activités enzymatiques alternatives à une activité enzymatique conservée. Les activités enzymatiques conservées sont en rouge et en bleu les activités enzymatiques spécifiques. Encadré en bleu des activités enzymatiques alternatives à une activité enzymatique conservée. **A.** Une partie de la voie de la tyrosine **B.** Une partie de la voie de la biosynthèse des acides gras.

Les alternatives sont des activités enzymatiques qui permettent de substituer des activités enzymatiques conservées. En se basant sur ces quelques exemples, les alternatives perdues permettent de suppléer les activités enzymatiques conservées. Les activités enzymatiques conservées ont, soit une activité enzymatique ayant beaucoup d'affinité pour plusieurs substrats, soit présentent une activité enzymatique plus efficace que l'activité spécifique. Les activités enzymatiques plus conservées jouent probablement donc un rôle plus essentiel car elles sont capables de catalyser plusieurs substrats et sont plus efficaces. Leur capacité à catalyser plusieurs substrats semble jouer dans leur conservation dans toutes les espèces. Les alternatives étant plus spécifiques et moins efficaces sont soumises à une pression de sélection moins forte, mais leur présence dans certaines espèces indique qu'ils peuvent substituer l'activité enzymatique conservée probablement dans certaines conditions.

Alternative à une autre activité enzymatique spécifique :

Ce rôle définit les activités enzymatiques spécifiques de certaines espèces qui peuvent se substituer à une autre activité enzymatique aussi spécifique. Ce cas d'alternative à une activité enzymatique spécifique permet d'aller d'un métabolite A vers un métabolite B mais un autre chemin avec une ou des activités enzymatiques seulement spécifiques est possible.

Cette définition peut être traduite dans la théorie des graphes par un sommet qui vérifie les propriétés suivantes :

(i) Quand on retire l'activité enzymatique spécifique, son absence ne créera pas une nouvelle composante connexe. La voie reste ininterrompue car la liaison reste assurée par l'autre activité spécifique qu'elle substitue (Figure 14 .3).

(ii) Si on part d'une voie avec seulement les activités enzymatiques conservées, l'ajout de l'alternatif va connecter deux composantes connexes de la voie avec une activité enzymatique conservée dans les deux composantes (Figure 14.3).

L'alternatif peut être un **groupe (série) d'activités enzymatiques**. Un groupe spécifique est défini comme une activité enzymatique et toutes les activités enzymatiques qui peuvent être atteintes par un chemin seulement composé d'activité spécifique.
Si le groupe est considéré comme un seul nœud dans le graphe, il a les caractéristiques d'un sommet spécifique alternatif à une autre activité enzymatique spécifique.
Chaque activité enzymatique du groupe est considérée comme une alternative mais travaillant en groupe (Figure 14.3).

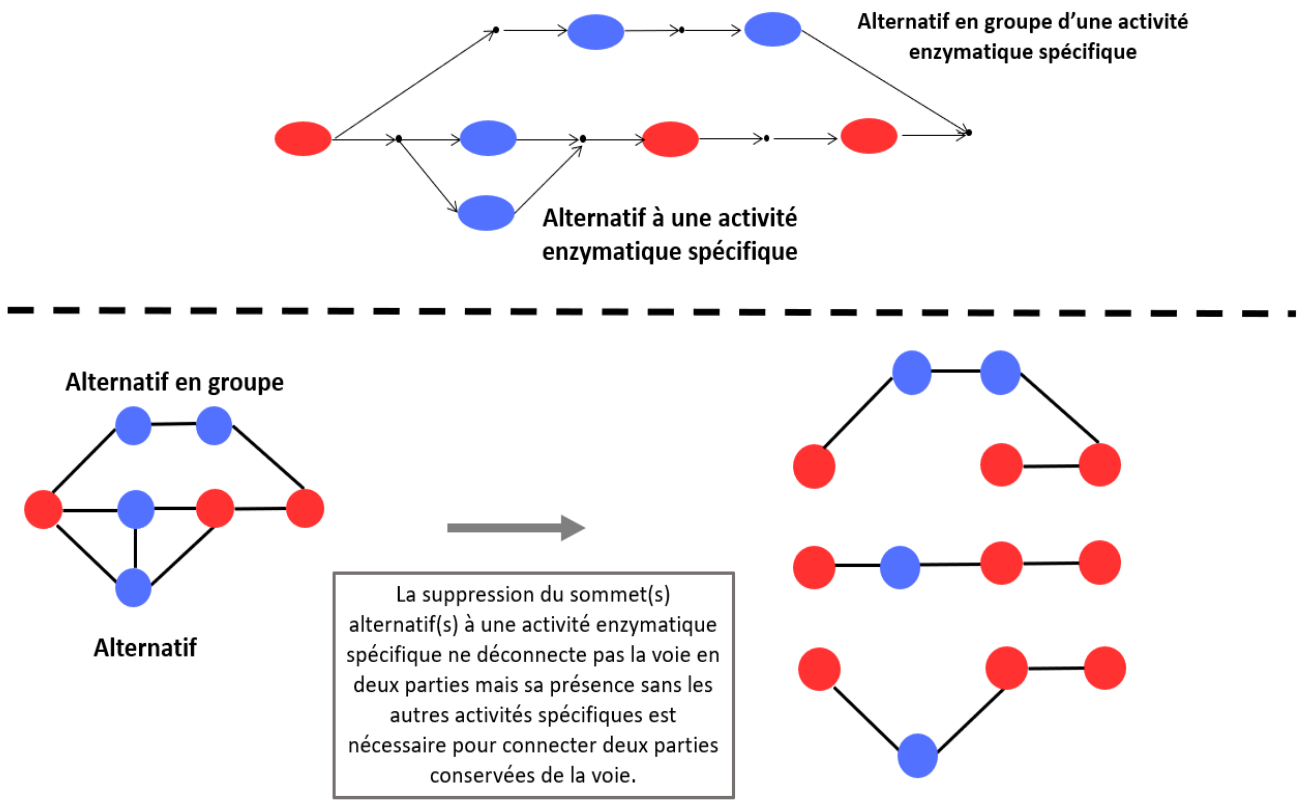


Figure 14.3 : Alternatif à une (des) activité(s) enzymatique(s) spécifique(s). En rouge les activités enzymatiques conservées et en bleu les activités enzymatiques spécifiques. **En haut,** la représentation de la voie avec les activités enzymatiques et les métabolites. Les activités enzymatiques en bleu sont en alternatives d'une activité enzymatique spécifique. **En bas,** la voie sous forme de graphe des activités enzymatiques. Deux activités enzymatiques sont connectées si elles partagent un métabolite en commun dans leur réaction. Un sommet spécifique est alternatif à un autre sommet alternatif, l'absence d'un seul sommet alternatif ne déconnecte pas le graphe en deux parties.

Nous avons observé des cas d'alternative à une activité enzymatique spécifique dans 14 voies.

Par exemple, dans la voie de la Glycolyse/Glycogenèse, 5.4.2.11 et 5.4.2.12 sont deux activités enzymatiques en alternatives (encadrées en noir dans la Figure 14.4 de gauche). L'absence simultanée de ces deux activités enzymatiques va couper la voie de la Glycolyse/Glycogenèse en deux parties. Mais en observant le profil phylogénétique de ces deux activités enzymatiques (Figure 14.4 droite), nous observons que les deux profils sont asymétriques. C'est-à-dire que la majorité des espèces ne possédant pas 5.4.2.11 ont conservé 5.4.2.12 et vice-versa.

Dans cette Figure 14.4, certaines espèces ne semblent néanmoins posséder ni l'une ni l'autre activité enzymatique. La non-détection des deux activités enzymatiques peut être due à une erreur d'annotation.

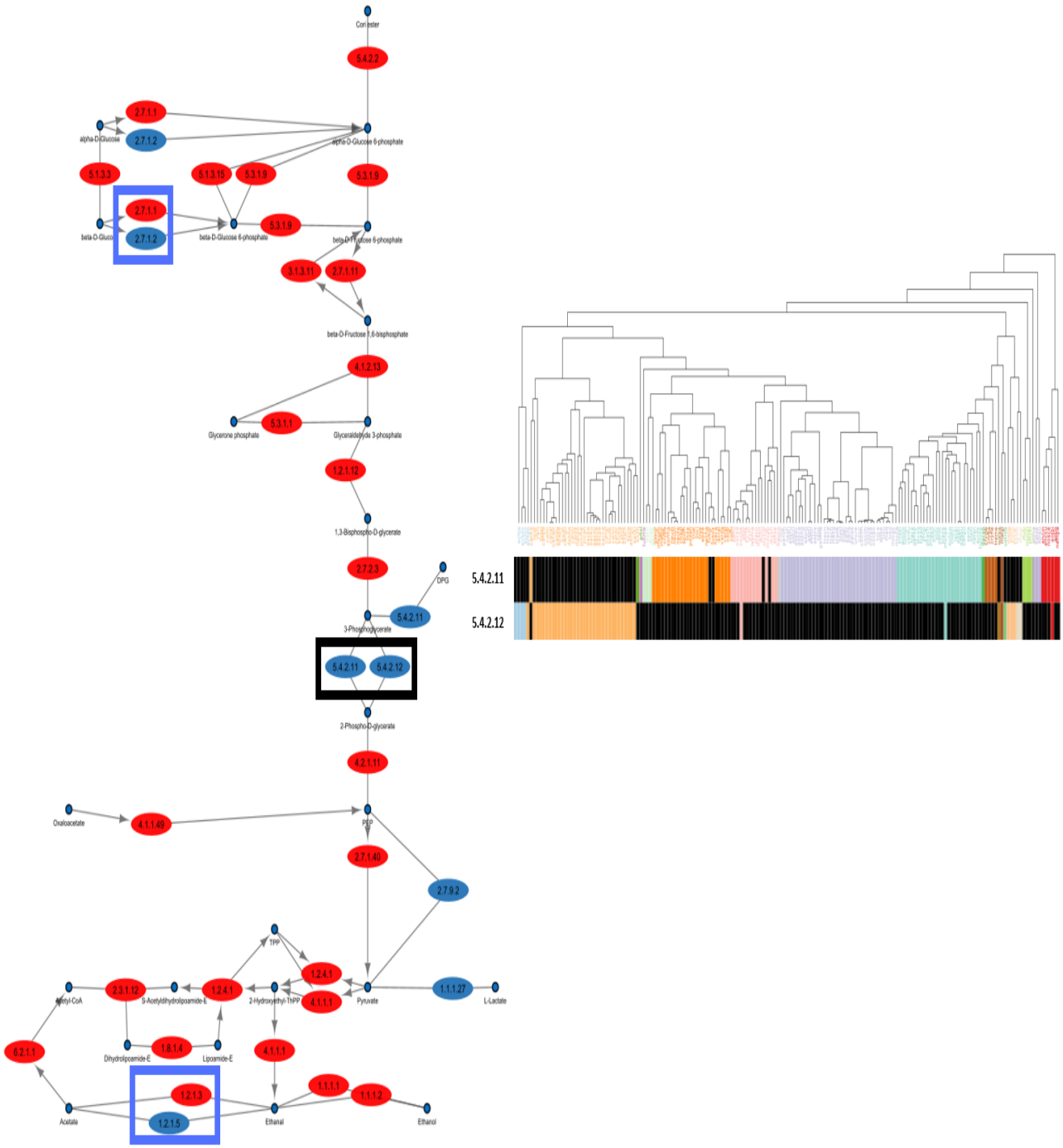


Figure 14.4 : A gauche : la voie de la Glycolyse/Glycogenèse. Les activités enzymatiques conservées sont en rouge et en bleu les activités enzymatiques spécifiques. Encadrées en bleu des activités enzymatiques alternatives à une activité enzymatique conservée (2.7.1.2 et 1.2.1.5). Encadrées en noir (5.4.2.11 et 5.4.2.12), des activités enzymatiques spécifiques ayant pour rôle une alternative à une autre activité enzymatique spécifique. **À droite** : le profil phylogénétique des deux activités enzymatiques spécifiques encadrées en noir. Les cellules en noires indiquent la présence de l'enzyme, si absente la cellule est colorée en fonction de la classe taxonomique. La liste des espèces et l'arbre phylogénétique sont indiqués au-dessus des profils. Les espèces dans l'arbre sont colorées en fonction de leur classe taxonomique.

Sur la base de cette observation, nous avons vérifié pour les activités enzymatiques spécifiques en alternatives à une activité enzymatique spécifique si une pression de sélection s'exerce pour conserver au moins une activité enzymatique afin de ne pas déconnecter la voie. Pour chaque couple d'alternatifs nous avons donc additionné les profils phylogénétiques et vérifié si la combinaison des deux profils couvre au moins 85% des espèces. Dans 8 voies sur 14, les profils phylogénétiques sont complémentaires c'est-à-dire qu'il y a eu une pression de sélection pour qu'une des activités enzymatiques soit présente dans le métabolisme comme dans le cas de la Glycolyse/Glycogenèse.

Quand les activités enzymatiques ne sont pas complémentaires, en vérifiant ces cas, nous avons deux activités enzymatiques avec deux spécificités différentes mais capables de catalyser certaines réactions communes. Par exemple, dans la voie de la biosynthèse des stéroïdes, la transformation de l'estrone en estradiol peut être effectuée par les activités enzymatiques 1.1.1.51 et 1.1.1.62 (Figure 14.5). Ces deux activités enzymatiques ont des spécificités communes pour l'estrone et l'estradiol mais reconnaissent aussi d'autres métabolites qui ne sont pas partagés par les deux activités enzymatiques (Marcus and Talalay, 1956). L'estrone chez les champignons est surtout associée à la synthèse de quelques mycotoxines (Fink-Gremmels and Malekinejad, 2007). Dans la voie de la biosynthèse de la phénylalanine, tyrosine et tryptophane, la transformation du L-arogenate en L-phénylalanine peut être effectuée par 4.2.1.51 et 4.2.1.91 (Figure 14.5). 4.2.1.91 est une activité enzymatique qui peut agir sur plusieurs substrats (Zamir *et al.*, 1988) alors que 4.2.1.51 (plus conservée) a plus d'affinité pour le L-arogenate. Ces deux activités enzymatiques offrent une alternative dans la synthèse du Phénylalanine.

Ces deux derniers exemples montrent, que quand les profils des activités enzymatiques spécifiques en alternatives ne sont pas complémentaires, la transformation qu'opèrent ces activités enzymatiques mène à une réaction enzymatique accessoire pour certaines espèces.

Nous avons retrouvé un groupe d'activités enzymatiques qui forme un alternatif à d'autres activités enzymatiques spécifiques de certaines espèces que dans une seule voie : la voie de fixation du carbone chez les procaryotes. C'est une voie que nous avons détectée comme présente chez les champignons à cause des nombreuses activités enzymatiques présentes chez les champignons.

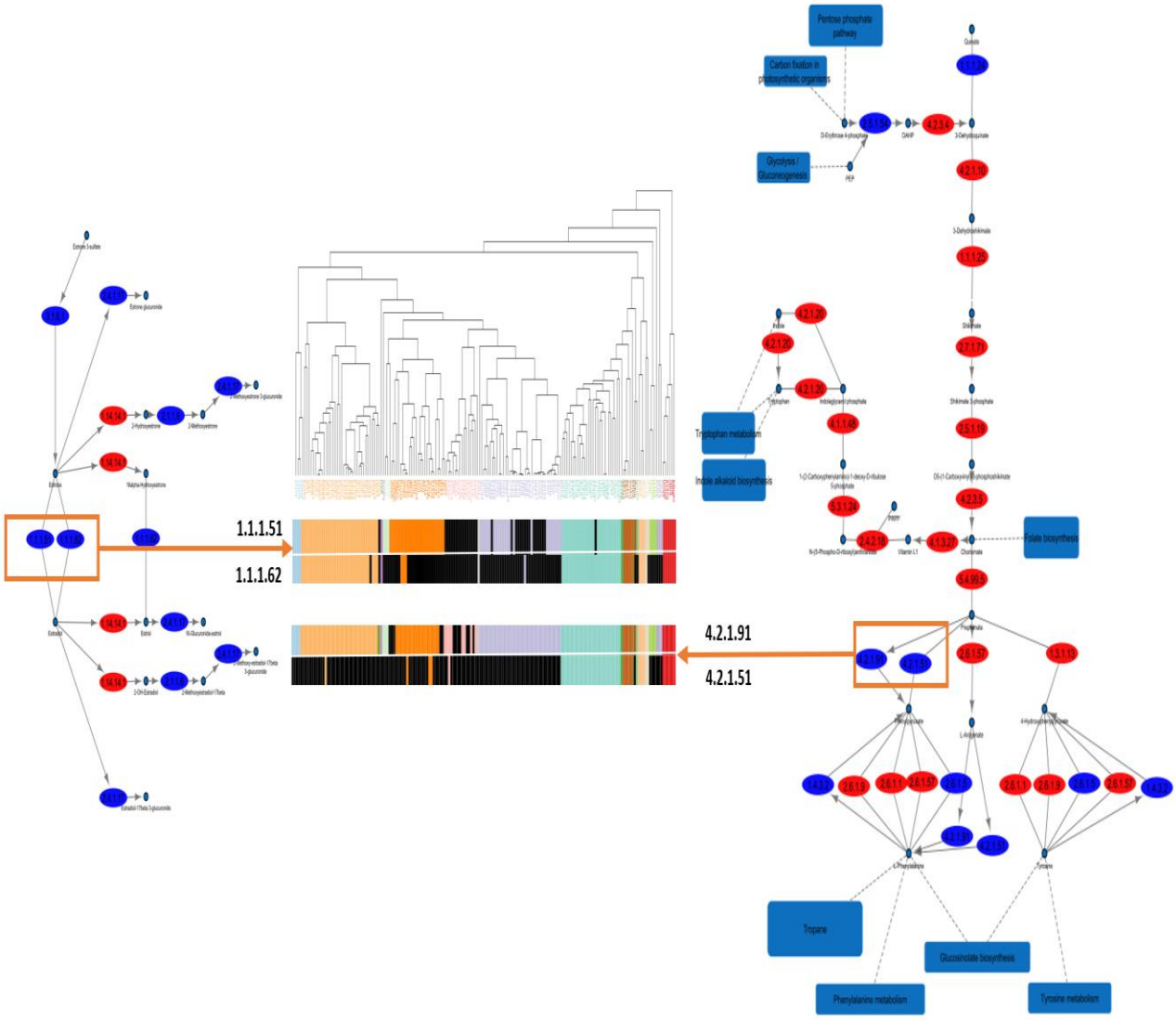


Figure 14.5 : A gauche, une partie de la voie de biosynthèse des stéroïdes. 1.1.1.51 et 1.1.1.62 sont encadrées en orange. **A droite,** la voie de la biosynthèse de la phénylalanine, de la tyrosine et du tryptophane. 4.2.1.91 et 4.2.1.51 sont encadrées en orange.
Les activités enzymatiques conservées sont en rouge et en bleu les activités enzymatiques spécifiques. Encadrées en orange des activités enzymatiques alternatives à une activité enzymatique spécifique. Le profil phylogénétique des activités enzymatiques spécifiques en alternatives est indiqué au milieu des deux voies. Les cellules noires indiquent la présence de l'activité enzymatique, si absente la cellule est colorée en fonction de la classe taxonomique. La liste des espèces et l'arbre phylogénétique sont indiqués au-dessus des profils. Les espèces dans l'arbre sont colorées en fonction de leur classe taxonomique.

Connecteur

Une activité enzymatique spécifique de certaines espèces a un rôle de connecteur si elle permet de connecter deux parties de la voie. Chaque partie contient une activité enzymatique conservée. Son absence va déconnecter deux parties de la voie. Théoriquement, c'est une activité enzymatique essentielle d'un point de vue topologique car elle permet de maintenir deux parties de la voie connectée.

Pour détecter ces sommets automatiquement, la définition précédente peut être traduite dans la théorie des graphes par un sommet qui possède les propriétés suivantes :

(i) L'absence de l'activité enzymatique sépare le graphe en deux parties (Figure 14 .6). Si on part d'une voie avec seulement les activités enzymatiques conservées, l'ajout de l'activité enzymatique connecte deux composantes qui étaient connexes (Figure 14.6).

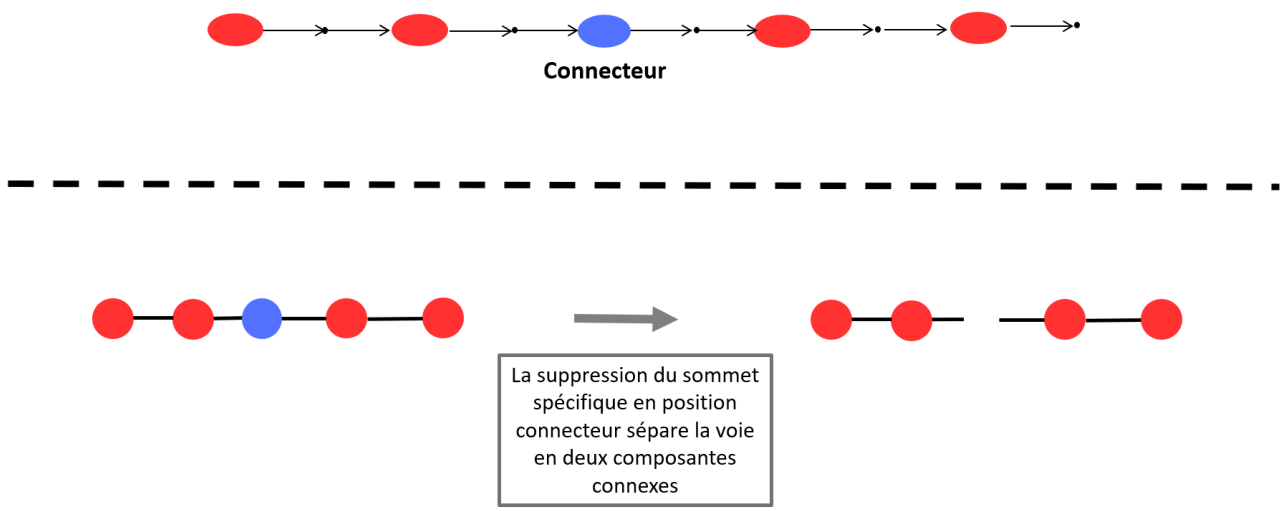


Figure 14.6 : Description d'un connecteur de voie. En rouge les activités enzymatiques conservées et en bleu les activités enzymatiques spécifiques. **En haut** la représentation de la voie avec les activités enzymatiques et les métabolites. **En bas,** la voie sous forme de graphe des activités enzymatiques. Deux activités enzymatiques sont connectées si elles partagent un métabolite en commun dans leur réaction. L'absence du connecteur va créer deux composantes connexes. L'ajout d'un connecteur dans la voie avec seulement la partie conservée va connecter deux parties du graphe. Les deux parties contiennent des activités enzymatiques conservées.

Nous avons recensé des connecteurs de voie répartis dans 22 voies métaboliques qui représentent 5% des pertes (Figure 14.15). Les connecteurs de voie permettent de comprendre comment des pertes d'activités enzymatiques qui connectent deux parties peuvent être tolérées dans certaines espèces.

En analysant la taille des deux composantes connexes connectées par le connecteur. Nous avons remarqué que la taille des deux composantes est en général très différente. Dans 15 voies sur 22, le connecteur relie une grande composante connexe contenant la majorité des activités enzymatiques conservées de la voie et une petite composante avec au moins une activité enzymatique conservée. La petite composante connexe est une courte série de réaction avec une activité enzymatique conservée dans la série.

Les métabolites et les activités enzymatiques associées à la petite composante connexe ont été analysés en regardant dans la littérature leur rôle potentiel chez les champignons ou chez d'autres espèces. Dans 11 voies sur 15, la petite composante connexe est une sous-voie accessoire. Par exemple, dans la voie du métabolisme de la cystéine et de la méthionine (Figure 14.7), une série de réactions composées de 1.13.11.20 (non conservée), 2.6.1.1 (conservée) puis une réaction spontanée permet la synthèse du sulfite et de la pyruvate à partir de la cystéine. Cette série de réaction permet d'éliminer une concentration élevée de cystéine qui est toxique à haute concentration (Hennicke *et al.*, 2013), mais aussi permet de synthétiser le sulfite qui a des propriétés antimicrobiennes et antioxidatives (Taylor *et al.*, 1986).

Dans la voie du métabolisme du folate (Figure 14.8), la petite composante connexe conduit à la biosynthèse de la Tetrahydrobiopterin. C'est un composé qui joue un rôle essentiel de régulation dans la lipogenèse chez les champignons oléagineux, c'est-à-dire les champignons capables d'accumuler une grande concentration de lipide comme *Mortierella alpina* (Wang *et al.*, 2020).

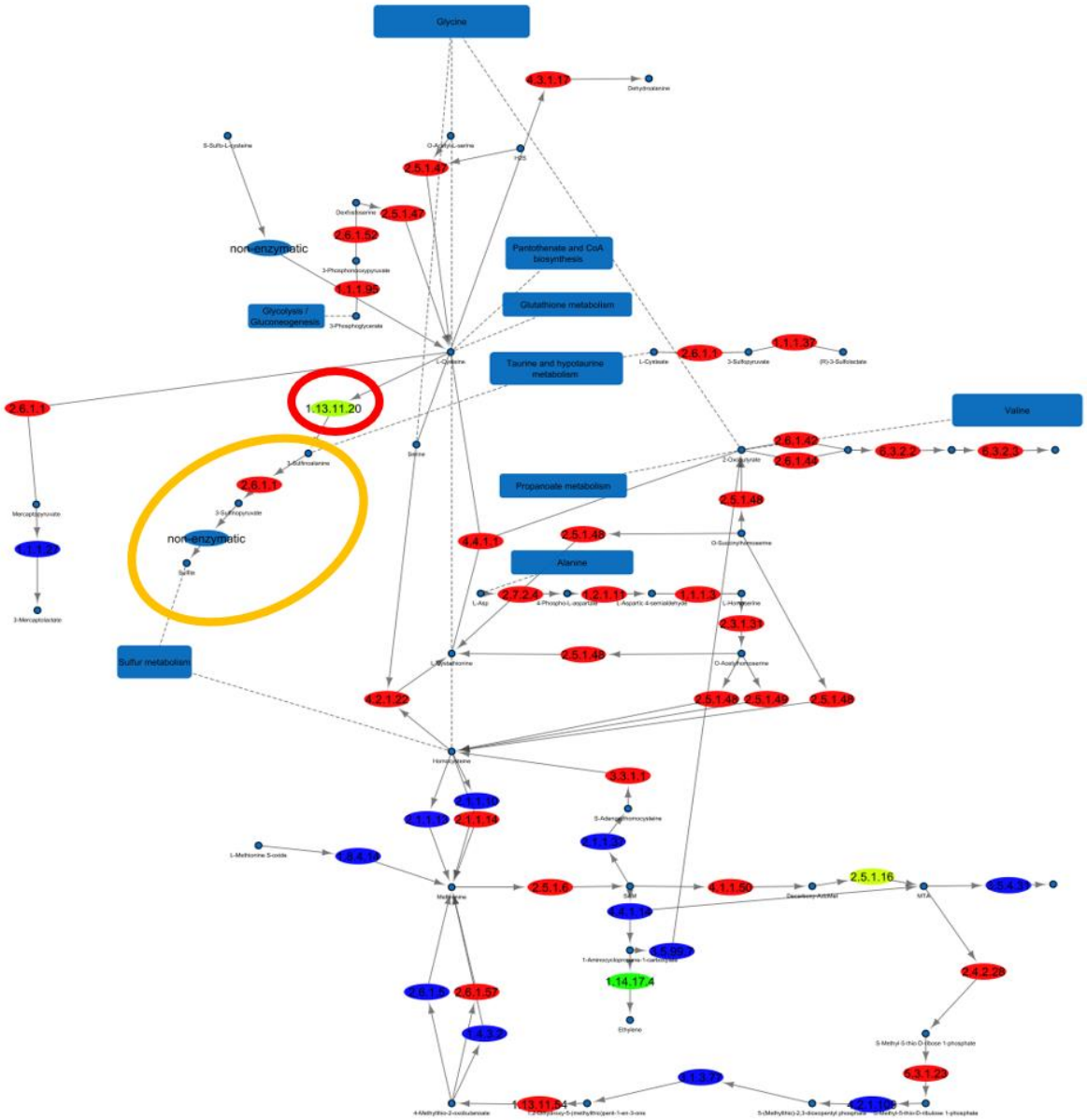


Figure 14.7 : Voie du métabolisme de la cystéine. Les activités enzymatiques en rouge sont conservées dans toutes les espèces, sinon elles sont colorées en fonction de leur similarité de profil : en bleu les activités enzymatiques avec des profils uniques. L'activité enzymatique encerclée en rouge (1.13.11.20) correspond à une activité enzymatique avec un rôle de connecteur. Ce connecteur connecte une petite composante connexe encerclée en jaune et la grande composante connexe qui correspond au reste de la voie. La partie de la voie encerclée en jaune est une série de réactions composées d'activité enzymatique conservée et non conservée qui permet de dégrader la cystéine et de synthétiser du sulfite. C'est une sous-voie accessoire.

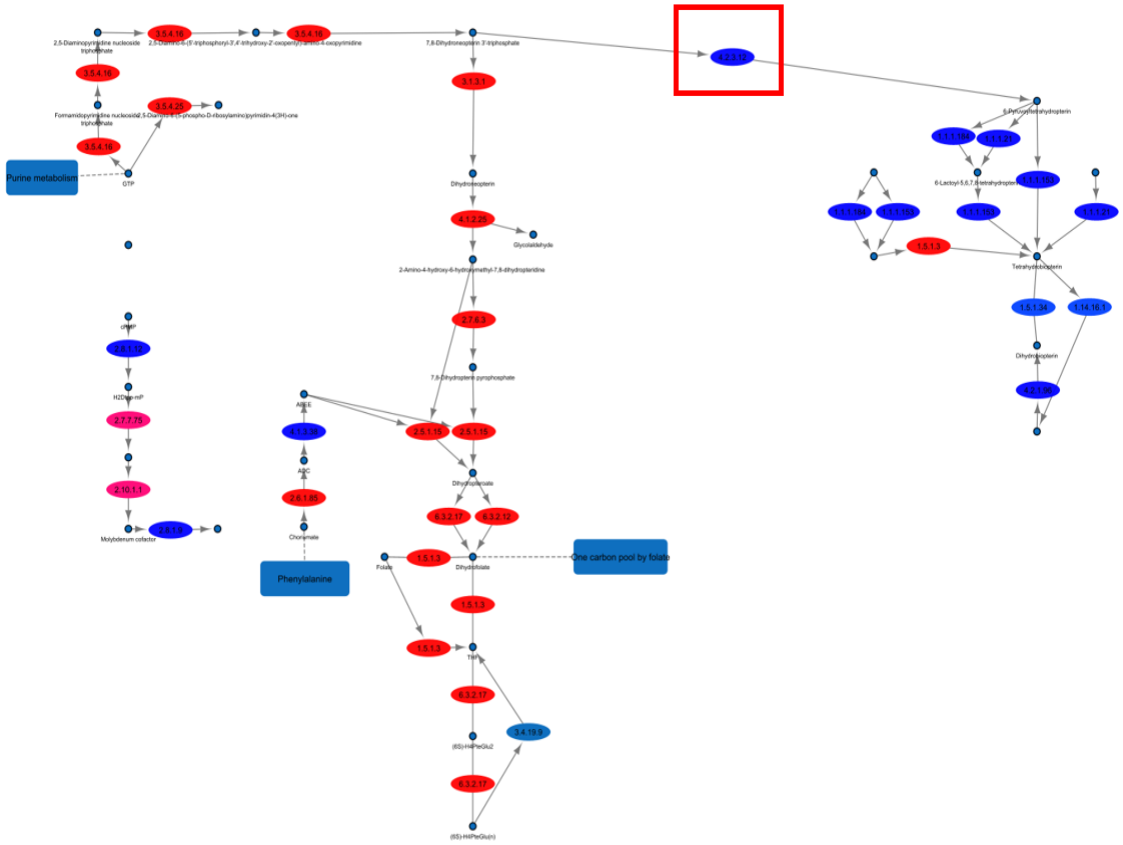


Figure 14.8 : Voie du métabolisme du folate. Les activités enzymatiques en rouge sont conservées dans toutes les espèces, sinon elles sont colorées en fonction de leur similarité de profil : en bleu les activités enzymatiques avec des profils uniques. L'activité enzymatique encerclée en rouge (1.5.1.3) correspond à une activité enzymatique avec un rôle de connecteur. Elle permet de connecter l'ensemble de la voie qui est très conservée et une sous-voie qui est accessoire contenant une activité enzymatique conservée.

La sous-voie contient aussi quelques activités enzymatiques conservées. Si la sous-voie est accessoire, nous avons essayé de comprendre pourquoi une des activités enzymatiques est conservée chez toutes les espèces. Notre hypothèse porte sur le fait que cette activité enzymatique joue un autre rôle probablement essentiel dans une autre voie métabolique.
La présence de l'activité enzymatique dans une partie accessoire témoigne d'une réutilisation des activités enzymatiques dans le réseau.
Par exemple, l'activité enzymatique conservée dans la sous-voie accessoire de la Figure 14.8 qui correspond à 1.5.1.3 est retrouvée dans le métabolisme du « One carbon pool by folate », qui est une voie essentielle chez tous les organismes vivant pour la synthèse des nucléotides et de certains acides aminés (Ducker and Rabinowitz, 2017).
En analysant toutes les activités enzymatiques conservées associées à la sous-voie accessoire, nous avons identifié que 16 activités enzymatiques conservées associées à une sous-voie accessoire 22 sont réutilisées dans d'autre voie. Leur conservation semble être due à leur rôle essentiel dans une autre voie.

Cependant dans certains cas, le connecteur connecte deux composantes connexes qui sont essentielles dans la biosynthèse d'un métabolite essentiel pour la survie de l'organisme. Par exemple, dans la voie du métabolisme de l'inositol (Figure 14.9), l'activité enzymatique 3.1.3.25 (inositol phosphatase) est absente chez une partie des *Saccharomycetes*, principalement du genre *candida*. L'inositol phosphatase permet la biosynthèse de l'inositol qui joue un rôle important dans la reproduction sexuée, la croissance mais aussi dans la virulence des organismes pathogènes (Chen *et al.*, 2007). Mais pour pallier à l'absence de cette activité enzymatique, *C.Albicans* a développé un mécanisme capable d'importer l'inositol présent dans son milieu (Chen *et al.*, 2008).

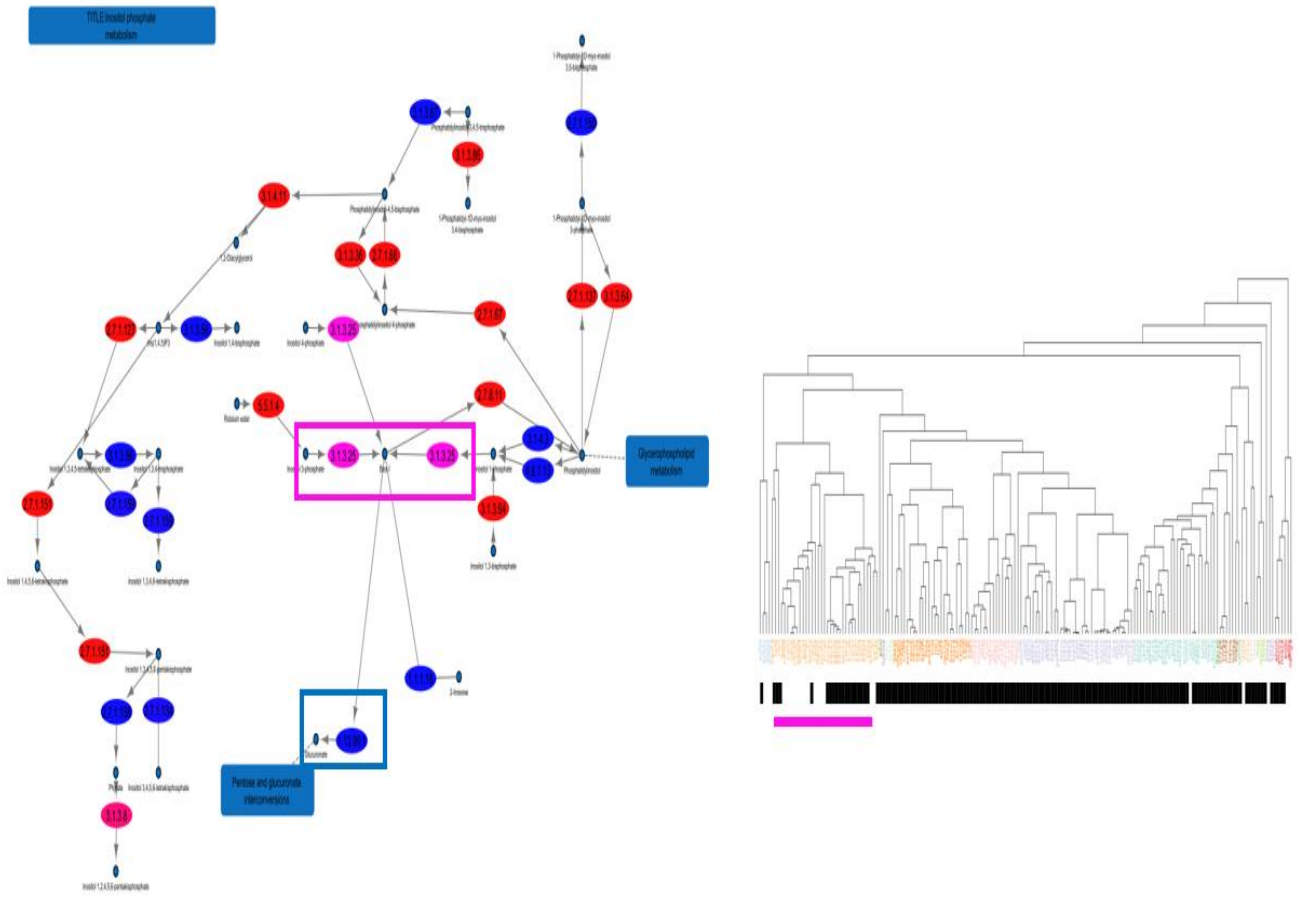


Figure 14.9 : A gauche : **la voie de l'inositol phosphate**. Les cercles rouges représentent les activités enzymatiques conservées. L'inositol phosphatase (3.1.3.25) est colorée en fuchsia et représente l'étape finale de la voie. L'activité enzymatique 1.13.99.1 qui a un rôle de développement est encadrée en bleu. **À droite** : le profil phylogénétique de l'inositol phosphatase. Les cellules noires indiquent la présence de l'activité enzymatique. La classe des *Saccharomycetes* est surlignée en fuchsia. La liste des espèces et l'arbre phylogénétique sont indiqués au-dessus des profils. Les espèces sont colorées en fonction de leur classe taxonomique.

Cet exemple montre que quand l'activité enzymatique joue un rôle dans la production d'un métabolite essentiel et indispensable, l'espèce a développé un autre mécanisme (non-métabolique) pour suppléer cette perte mais la majorité des connecteurs observés connectent une sous-voie qui est devenue accessoire pour certaines espèces.

Connecteur en groupe :

C'est un ensemble de sommets, qui permet de connecter deux parties de la voie. Chaque partie de la voie est composée au moins d'une activité enzymatique conservée (Figure 14.10).
Si on groupe sous un seul sommet toutes les activités enzymatiques qui peuvent être atteintes par un chemin seulement composé d'activité enzymatique spécifique, ce sommet a les mêmes caractéristiques qu'un connecteur. L'ensemble permet de connecter deux parties de la voie (Figure 14.10).
Chaque activité enzymatique du groupe est considérée comme un connecteur mais travaillant en groupe.

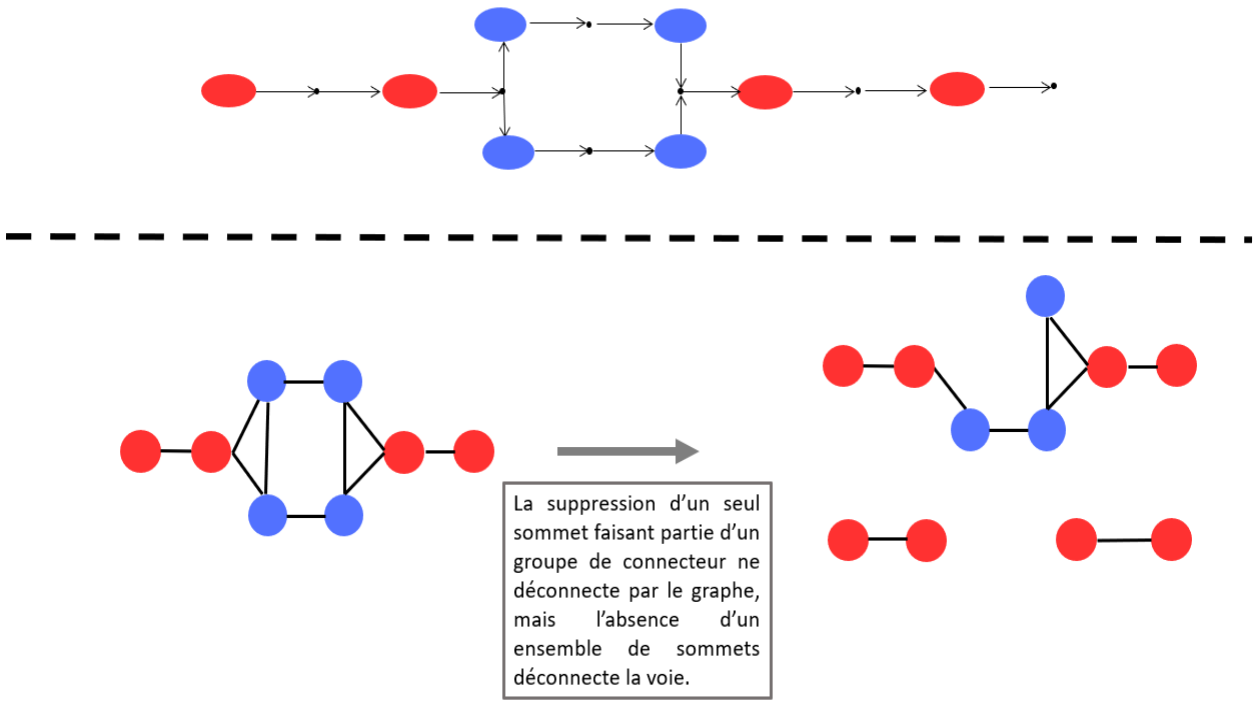


Figure 14.10 : Description des connecteurs en groupe. En rouge les activités enzymatiques conservées et en bleu les activités enzymatiques spécifiques. **En haut** la représentation de la voie avec les activités enzymatiques et les métabolites. **En bas,** la voie sous forme de graphe des activités enzymatiques. Deux activités enzymatiques sont connectées si elles partagent un métabolite en commun dans leur réaction. L'absence d'un seul sommet spécifique ne déconnecte pas la voie, mais l'absence de l'ensemble de sommets spécifiques adjacents déconnecte la voie en deux parties. Chaque partie contient au moins un sommet conservé.

Dans les quelques cas de connecteurs en groupe que nous avons observés, ces groupes permettent majoritairement de connecter une sous-voie composée majoritairement d'activité enzymatique spécifique avec quelques réactions catalysées par une activité enzymatique conservée et d'une partie principale composée majoritairement d'activités enzymatiques conservées. Les groupes de connecteurs permettent de relier l'activité enzymatique contenue dans la sous-voie au reste des activités enzymatiques de la voie. Des connecteurs en groupes ont été identifiés dans 14 voies.
Dans 6 voies, d'après la littérature, la sous-voie correspond à une sous-voie accessoire. Aucune information n'a été retrouvée dans les 8 autres voies. Par exemple, dans la voie du métabolisme du tryptophane (Figure 14.11), la sous-voie accessoire conduit à la biosynthèse de l'indole-2-acétate. 3 chemins conduisent à sa production constituée principalement d'activité enzymatique spécifique. La dernière étape de la réaction est effectuée par deux activités enzymatiques conservées. Ces deux activités enzymatiques sont présentes dans plusieurs voies. C'est un métabolite accessoire utilisé dans l'interaction entre les champignons et les plantes. C'est un métabolite qui n'a pas un effet immédiat sur la croissance des champignons mais affecte surtout la croissance des plantes (Fu *et al.*, 2015).

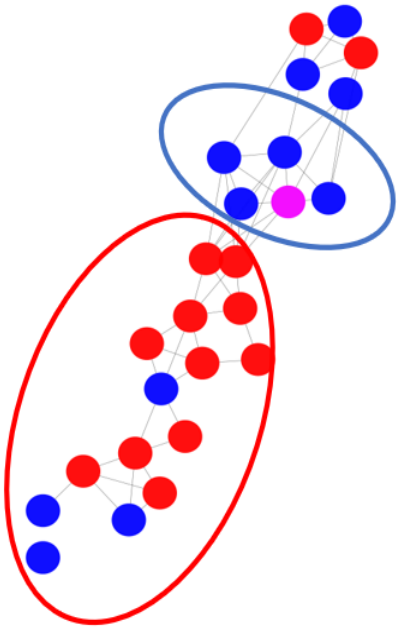


Figure 14.11 : Voie du métabolisme du tryptophane. Les sommets rouges indiquent des activités enzymatiques conservées. Deux activités enzymatiques sont reliées s'ils partagent un composé en commun dans leur réaction. La partie commune à toutes les espèces de la voie est encerclée en rouge et encerclée en bleue les sommets qui forment le groupe de connecteurs. Plusieurs chemins permettent de relier les activités enzymatiques conservées dans le module bleu avec le module rouge.

Développement ou extension :

Un sommet dans une voie a un rôle de développement si ce sommet se situe à l'extrémité de la voie. Les développements sont des formes d'extensions secondaires de la partie qui est commune et essentielle à toutes les espèces. Les développements permettent l'utilisation d'un substrat pour la voie (entrée) ou la synthèse d'un produit final (sortie) de la voie. L'utilisation d'un substrat ou la synthèse d'un produit peut être le fruit d'une série de réactions enzymatiques, ainsi les activités enzymatiques non conservées de cette série de réaction font partie du développement ou de l'extension.

Pour les détecter par analyse informatique, la définition précédente peut être traduite selon la théorie des graphes par un sommet qui possède les propriétés suivantes (Figure 14.12):

(i) Ce sont les sommets de degré un.

(ii) Ce sont des activités enzymatiques dont un de leurs métabolites n'est inclus que dans une seule activité enzymatique de la voie .

(iii) Ce sont des sommets dont l'absence va créer deux composantes connexes : une des composantes connexes ne sera constituée que d'activités enzymatiques spécifiques. Ce sommet fait partie d'une série de réactions pour la synthèse d'un produit ou la transformation d'un substrat.

Les développements signifient que la majorité des pertes sont des activités enzymatiques qui permettent l'utilisation d'un substrat particulier comme précurseur de la voie ou la synthèse d'un métabolite probablement accessoire. Ces sont probablement des extensions qui sont accessoires.

Un exemple de développement se trouve dans la voie de l'inositol phosphate (Figure 14.9), l'activité enzymatique 1.13.99.1 (encadrée en bleue dans la Figure 14.9) permet de synthétiser du glucuronate à partir de l'inositol. C'est une autre manière de synthétiser du glucuronate. Cette réaction enzymatique est accessoire car la synthèse du glucuronate est principalement synthétisée à partir du D-glucose (Mazerska *et al.*, 2016; Linster and Van Schaftingen, 2007). L'activité enzymatique 1.13.99.1 a été surtout perdue chez la majorité des *saccharomycetes* (Figure 14.13).

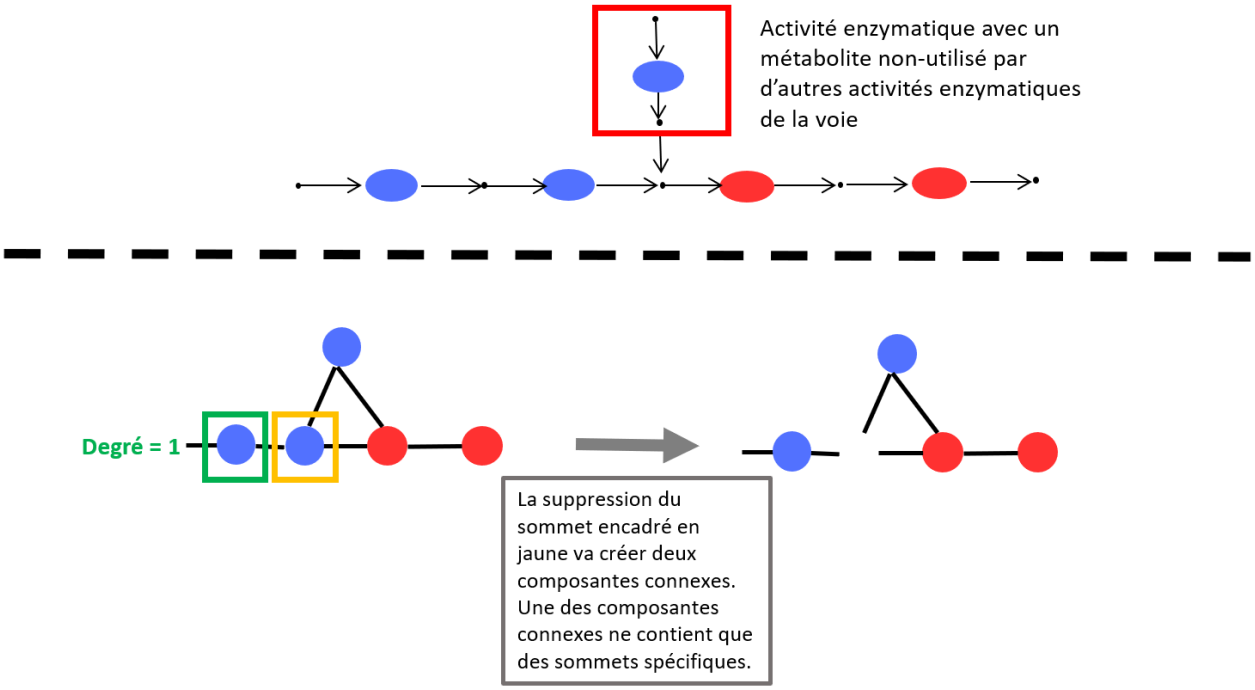


Figure 14.12 : Description des sommets en développement. En rouge les activités enzymatiques conservées et en bleu les activités enzymatiques spécifiques. **En haut** la représentation de la voie avec les activités enzymatiques et les métabolites. L'activité enzymatique encadrée en rouge possède un métabolite qui n'est traité que par cette activité enzymatique. **En bas :** la voie sous forme de graphe des activités enzymatiques. Deux activités enzymatiques sont connectées si elles partagent un métabolite en commun dans leur réaction. Le sommet encadré en vert est de degré 1 donc c'est un développement. Le sommet encadré en orange crée 2 composantes connexes si on le retire. Une des composantes est seulement constituée d'activité enzymatique spécifique. Ce sommet est donc lui aussi un développement et fait partie d'une succession d'activité enzymatique spécifique qui se situe soit en entrée soit en sortie de la voie.

Figure 14.13 : Profil phylogénétique de l'activité enzymatique 1.13.99.1 dans les 174 espèces de champignons. Les cellules noires indiquent la présence de l'activité enzymatique. La cellule est colorée en fonction de la classe taxonomique de l'espèce si absente. La classe des *Saccharomycetes* est surlignée.

Cas ambigu :
L'analyse sous forme de graphe des voies métaboliques peut créer des cas ambigus . Dans la Figure 14.14, si les activités enzymatiques spécifiques sont prises individuellement, elles sont considérées comme des développements qui permettent l'assimilation d'un substrat. Si elles sont prises deux par deux, les activités enzymatiques du haut forment un groupe d'alternatif à un autre groupe alternatif (celui du bas). Mais si les activités enzymatiques sont prises toutes ensemble, elles forment un groupe de connecteurs.
Pour pallier à ce problème, nous avons hiérarchisé les différents rôles. Un sommet ne peut pas avoir plusieurs rôles.
Tout d'abord nous avons cherché les activités enzymatiques qui permettent de connecter deux composantes connexes (les connecteurs). Ce sont des cas qui nous intéressent particulièrement pour comprendre comment une telle perte peut être tolérée dans certaines espèces ou comment deux parties de la voie ont été connectées ensemble.
Ensuite, nous avons déterminé les activités enzymatiques en alternatives à une autre activité enzymatique spécifique. Ces activités enzymatiques connectent deux parties de la voie car l'absence de tous les chemins parallèles va déconnecter la voie.
Nous avons ensuite cherché, les activités enzymatiques en développement et finalement les alternatives à une activité enzymatique conservée.

Ainsi, dans l'exemple de la Figure 14.14, les activités enzymatiques spécifiques sont considérées comme un groupe de connecteurs.

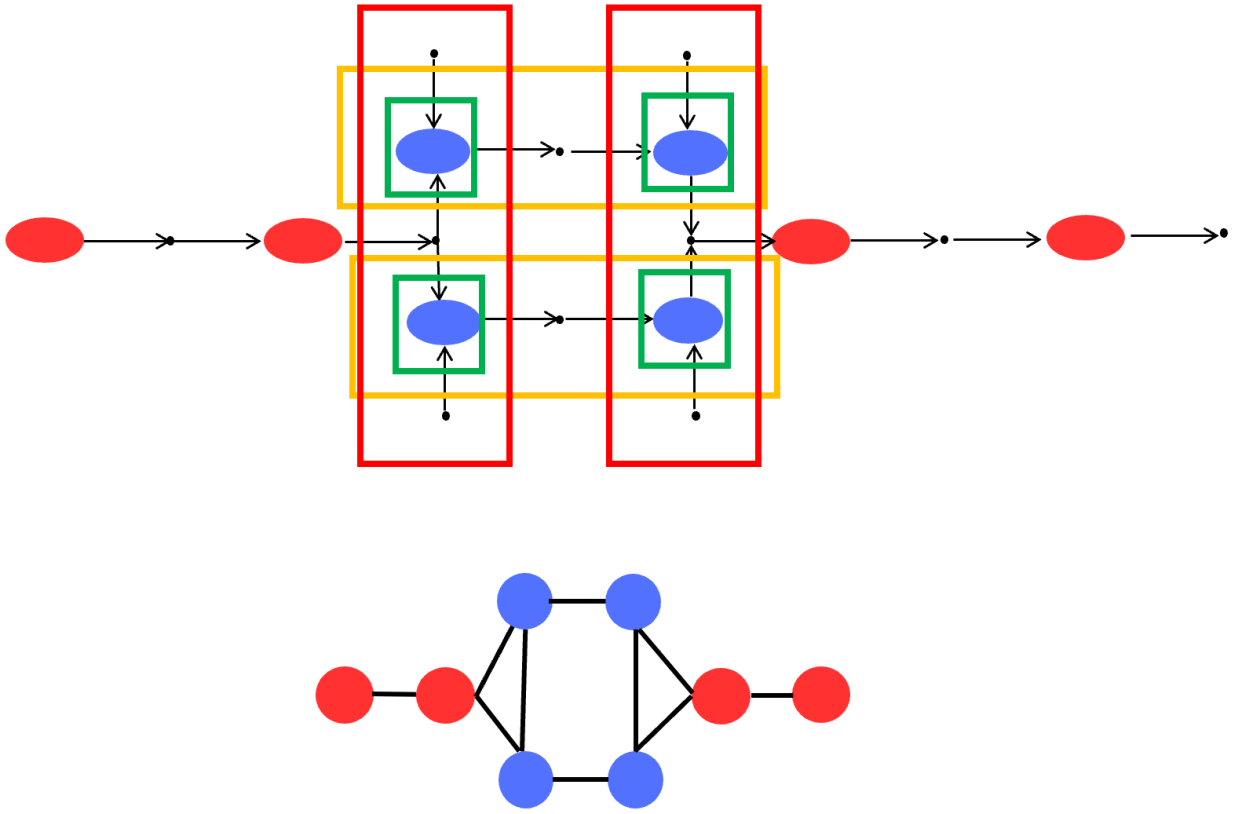


Figure 14.14 : Activités enzymatiques avec plusieurs rôles possibles. En haut, la représentation avec les activités enzymatiques et les métabolites. Les activités enzymatiques encadrées en vert peuvent être considérées comme des développements, en orange nous avons deux groupes spécifiques en alternatifs et en rouge des groupes de connecteurs. **En bas,** la représentation de la voie sous forme de graphes des activités enzymatiques.

14.2 Résultats de l'étude des rôles des sommets spécifiques

La localisation des activités enzymatiques perdues au cours de l'évolution dans les 63 voies communes montre que la majorité des cas de pertes sont des activités enzymatiques alternatives à une activités enzymatiques conservées ou des activités enzymatiques en développement. Ces deux cas de figure représentent 36% et 40% des pertes (Figure 14.15).

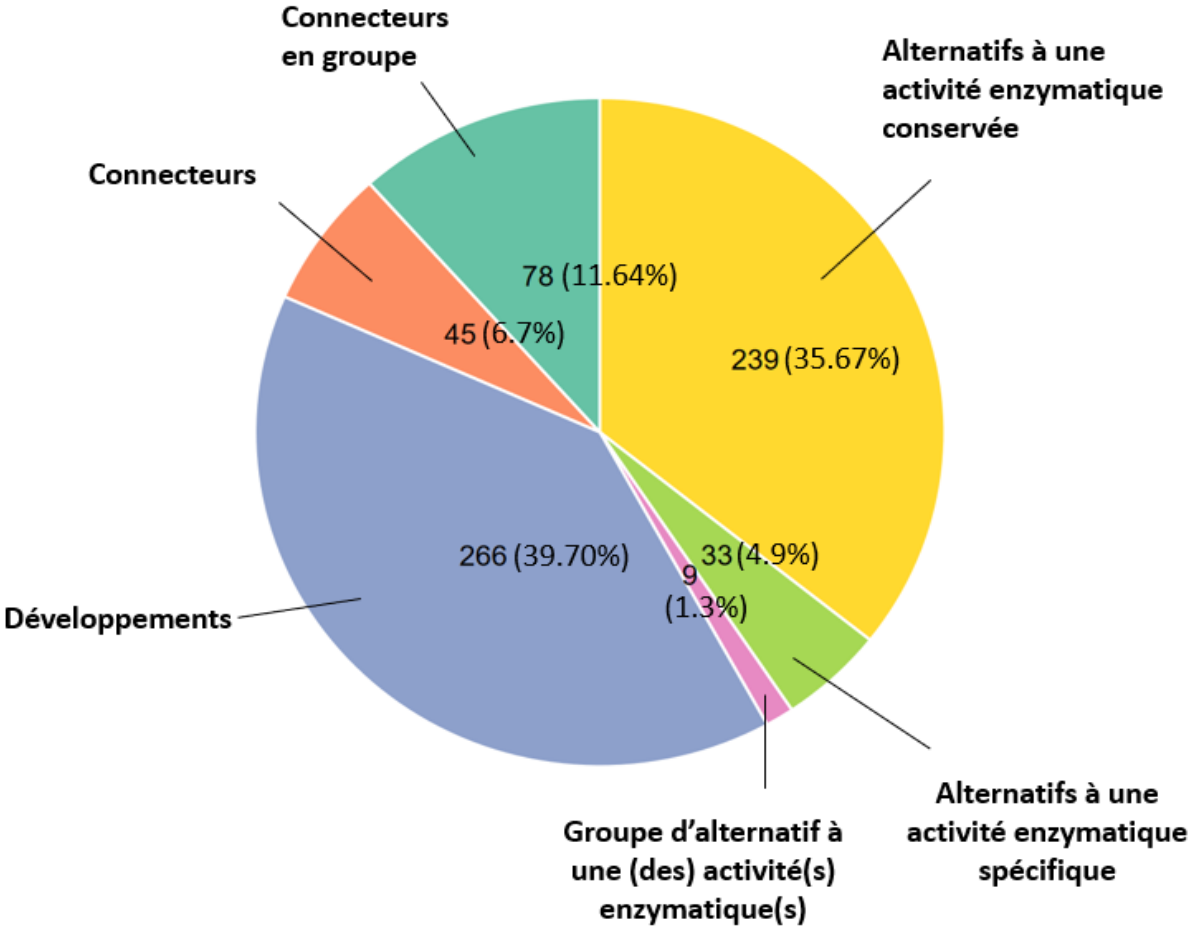


Figure 14.15 : Résultat de l'étude des rôles des activités enzymatiques perdues dans les voies communes à toutes les espèces. Chaque tranche représente un rôle et le nombre à l'intérieur de chaque tranche indique le nombre de sommets ayant le rôle indiqué par la tranche.

L'analyse des rôles des activités enzymatiques dans les voies a montré tout d'abord que 33% des sommets associés à des activités enzymatiques spécifiques sont en alternatifs d'une activité enzymatique conservée par toutes les espèces. La transformation d'un métabolite en un autre catalysée par une activité enzymatique conservée dans certaines espèces peut donc être effectuée par une autre activité enzymatique.

Dans la partie III.11.2, nous avons montré que le réseau métabolique à une densité très

faible. Ce qui signifie que le nombre d'activités enzymatiques reliées entre elles est très limité ce qui a pour conséquence que la succession de réactions pour convertir un métabolite en un autre est bien définie et très peu d'alternatives sont possibles.
L'analyse des rôles des activités enzymatiques spécifiques, nous a permis de détecter beaucoup d'alternatives à une activité enzymatique conservée. Ces alternatives correspondent surtout à des alternatives très localisées dans le réseau qui permettent de remplacer une activité enzymatique. Ces alternatives très localisées vont probablement permettre d'assurer que les séries de réactions bien définies ne soient pas rompues dans certaines espèces quand les activités enzymatiques conservées ne peuvent pas assurer leur fonction.
Dans les cas des activités enzymatiques spécifiques de certaines espèces alternatives à une autre activité spécifique, une pression de conservation sur les deux activités enzymatiques permet qu'au moins une des activités enzymatiques soit conservée dans chaque espèce pour éviter de rompre une voie essentielle.

La grande majorité des activités enzymatiques spécifiques sont aussi situées en périphérie des voies (développement). Ce sont des activités enzymatiques qui permettent l'utilisation d'un substrat particulier comme précurseur de la voie, vraisemblablement pour alimenter les successions de réactions enzymatiques communes à toutes les espèces à la synthèse d'un métabolite essentiel. Ce sont aussi des activités enzymatiques qui permettent de synthétiser un produit spécifique à certaines espèces.

L'environnement joue donc un rôle fondamental dans l'évolution du réseau métabolique, car en fonction de l'environnement certaines activités enzymatiques peuvent devenir accessoires. Même la perte des activités enzymatiques qui sont essentielles dans la topologie de la voie en connectant deux parties de la voie (connecteur) peut être tolérée si une des parties est accessoire dans certaines espèces. En plus de l'environnement, le réseau métabolique et les voies métaboliques exercent une pression pour conserver les séries de réactions enzymatiques essentielles (plus centrales et plus connectées) pour la survie de l'organisme. Et les activités enzymatiques responsables de la variété métabolique sont principalement situées en périphérie du réseau et des voies ou en alternatives d'une activité enzymatique conservée (les développements et les alternatifs). Les activités enzymatiques spécifiques de certaines espèces de champignons assurent aussi le maintien des séries de réactions enzymatiques conservées. Les activités enzymatiques apparues au cours de l'évolution des champignons que nous avons pu détecter sont majoritairement impliquées dans le métabolisme des métabolites secondaires.

V. Discussion

La construction du réseau métabolique chez les champignons

La construction du réseau métabolique chez les champignons a été effectuée en connectant entre elles les voies métaboliques décrites dans KEGG. Dans un second temps les activités enzymatiques du réseau ont été filtrées pour ne garder que les activités enzymatiques présentes chez les champignons. En effet, les voies métaboliques dans KEGG sont construites manuellement avec les connaissances décrites dans la littérature et représentent la voie globale de toutes les espèces où la voie a été étudiée. Les voies les plus connues telles que la glycolyse ou cycle de Krebs ont été largement étudiées dans les 3 règnes du vivant alors que certaines voies n'ont été étudiées que chez des organismes modèles ou quelques espèces d'intérêts.

Du fait de ce déséquilibre, certaines voies peuvent refléter la voie globale chez tous les règnes du vivant alors que certaines voies représentent la voie de seulement quelques espèces. Par conséquent, si une voie n'a jamais été ou très peu décrite chez les champignons, les représentations KEGG peuvent ne pas du tout représenter la voie chez les champignons. Ceci aura comme conséquence la non-détection de la voie ou d'une partie de la voie chez les champignons si des activités enzymatiques diffèrent entre la représentation actuelle de la voie et la voie chez les champignons. Si la succession des réactions enzymatiques est différente chez les champignons, ou si les champignons utilisent d'autres activités enzymatiques qui n'ont pas été décrites dans d'autres espèces, alors la voie inférée chez les champignons à partir des voies KEGG ressemblera à un ensemble de petites composantes connexes.

Du fait de la construction manuelle des voies KEGG, certaines réactions enzymatiques spécifiques décrites dans certaines espèces spécifiques peuvent également ne pas être indiquées dans KEGG.

Ces connaissances manquantes dans les voies de KEGG, peuvent expliquer le nombre de composantes connexes élevées dans le réseau métabolique ainsi que l'exclusion de certaines voies de notre analyse même si la littérature confirme la présence de la voie chez les champignons, par exemple la dégradation du furfural (Wang *et al.*, 2012).

D'un autre côté, certaines activités enzymatiques identifiées dans les voies KEGG peuvent être non identifiées chez les champignons suite à une erreur d'annotation des enzymes correspondants.

L'annotation des activités enzymatiques dans les 174 génomes de champignons se fait par transfert d'annotation de la séquence dont la structure primaire la plus similaire est présente dans la base de données Uniprot. Les enzymes dont la fonction a été annotée chez les champignons et avec lesquelles nous avons travaillé ont été annotées avec la version de 2012 de Uniprot alors que nous avons travaillé avec la version 2022 de KEGG. Dans le tableau 1.2, nous pouvons observer que certains EC-numbers ont été transférés vers d'autres EC-number alors que certains ont été supprimés. Il se peut donc que ces EC-numbers aient été mis à jour dans KEGG et les annotations de certaines enzymes chez les champignons ne sont plus totalement à jour. Par conséquent, l'activité enzymatique dans KEGG n'est pas identifiée dans notre base de données chez les champignons et cela va créer une lacune dans la représentation de la voie chez les champignons.

La méthode de transfert d'annotation utilisée peut aussi conduire à une erreur d'annotation car deux séquences très similaires peuvent avoir deux activités différentes (mais probablement proches). Si l'activité enzymatique spécifique de l'enzyme n'est pas répertoriée dans Uniprot (en 2012), l'enzyme sera annotée avec la séquence la plus proche dans Uniprot qui peut avoir un EC-number différent de celui de l'activité enzymatique réelle de l'enzyme à annoter.
Ensemble, les erreurs d'annotations des enzymes mais aussi les biais dans les représentations de KEGG des voies peuvent donc introduire des biais dans l'analyse du réseau métabolique. L'amélioration des connaissances dans la représentation des voies, mais aussi dans les annotations en fournissant des bases de données plus fournies apportera plus de précisions dans l'analyse du réseau métabolique. Les biais évoqués ci-dessus ne remettent pas fondamentalement en cause nos résultats, mais constituent une source possible d'amélioration de la précision de nos résultats.

Détection des activités enzymatiques co-évoluantes

Les profils phylogénétiques des activités enzymatiques indiquent la présence et l'absence des activités enzymatiques chez les espèces de champignons étudiées. Afin de détecter des modules évolutifs, c'est-à-dire des activités enzymatiques qui sont à la fois présentes et absentes dans les mêmes espèces, nous avons classé les activités enzymatiques par similarité de profil.
Nous avons identifié 15 groupes d'activités enzymatiques aux profils similaires. Parmi ces 15 groupes, il y a un groupe qui contient 420 activités enzymatiques très conservées. Les 14 autres groupes témoignent d'une organisation modulaire du réseau métabolique. En plus de leur conservation, nous avons montré que dans le réseau métabolique, les activités enzymatiques co-évoluantes de ces 14 groupes sont proches les unes des autres. Cependant, ces 14 groupes ne contiennent que 39 activités enzymatiques. Cela signifie que 451 activités enzymatiques ont des profils uniques.
C'est un résultat surprenant, car avec la diversité des champignons (taxonomique, morphologique et environnementale) nous nous attendons à détecter plus de modules.
Une hypothèse qui permet d'expliquer ce résultat est la perte de modularité du réseau métabolique de l'ancêtre vers les espèces actuelles. En effet la comparaison du réseau métabolique chez les bactéries issues de différents environnements avec un réseau métabolique ancestral a montré que la spécialisation et les environnements extrêmes entraînent une baisse de la modularité (Parter *et al.*, 2007; Takemoto *et al.*, 2007; Kreimer *et al.*, 2008). C'est-à-dire que la spécialisation à un environnement particulier a entraîné la perte de plusieurs activités enzymatiques non essentielles pour cet environnement particulier.

Conservation et origine évolutive des activités enzymatiques

L'analyse de la conservation des activités enzymatiques chez 174 espèces de champignons nous a montré que la moitié des activités enzymatiques présentes chez les champignons est commune à toutes les espèces et que l'autre moitié participe à la diversité métabolique.
La répartition des activités enzymatiques par classes taxonomiques nous a montré que

certaines classes taxonomiques peuvent être définies en fonction de leur répertoire enzymatique. Les microsporidies ont subi une grosse réduction enzymatique par rapport aux autres classes vraisemblablement du fait de leur mode de vie parasitaire. Les *Saccharomycetes* se distinguent des autres classes appartenant au *Ascomycetes* par l'absence d'un ensemble d'activité enzymatique dont la moitié appartient à la voie de la dégradation de la lysine, leucine et de l'isoleucine. Dans la classe des *Eurotiomycetes*, les *Onygenales* et *Eurotiales* peuvent être clairement différenciés en fonction de leur contenu enzymatique. Le répertoire enzymatique permet de définir certains groupes taxonomiques reflétant l'évolution de ces groupes au cours de l'évolution.

Avec les approches de phylostratigraphie, nous avons inféré l'origine évolutive de chaque activité enzymatique. Nous avons inféré que 860 activités enzymatiques sont d'origines ancestrales car une séquence homologue en dehors des champignons avec la même activité enzymatique a été trouvée. Parmi les activités enzymatiques qui sont d'origine ancestrale, 445 sont fortement conservées et 415 semblent avoir été perdues au cours de l'évolution chez certaines espèces. D'après ces résultats, l'évolution du réseau métabolique chez les champignons a été surtout façonnée par de nombreuses pertes d'activités enzymatiques. La perte de ces activités enzymatiques témoigne probablement de l'adaptation et de la spécialisation à un environnement particulier. Du fait de cette spécialisation, certaines activités enzymatiques deviennent dispensables et la pression pour les conserver diminue car réguler une réaction enzymatique nécessite de l'énergie pour maintenir l'équilibre dans le réseau. Par exemple, le métabolisme des microsporidies est réduit au strict minimum car ils profitent du métabolisme de leur hôte (Corradi, 2015). Les petites composantes connexes accessoires connectées au reste de la voie par une seule activité enzymatique (connecteur) sont donc des parties de voie qui sont dispensables.
Dans cette analyse, nous avons seulement identifié 8 activités enzymatiques qui ne sont retrouvées que chez les champignons. C'est un chiffre qui nous semble très faible par rapport à la diversité et la capacité métabolique qu'offrent les champignons.

Contrainte du réseau métabolique sur l'évolution des activités enzymatiques.
L'analyse de la conservation et de l'origine évolutive des activités enzymatiques dans le réseau montre que le réseau métabolique exerce une pression sur l'évolution des activités enzymatiques. Le centre du réseau métabolique est principalement composé d'activités enzymatiques fortement conservées et les activités enzymatiques spécifiques de certaines espèces se situent principalement en périphérie du réseau. De plus, les activités enzymatiques fortement conservées sont plus fortement connectées que les activités enzymatiques spécifiques. Ces résultats sont cohérents avec les analyses faites par Peregrin-Alvarez et ses collaborateurs (Peregrín-Alvarez *et al.*, 2009) qui ont identifié les mêmes caractéristiques sur le réseau métabolique global partagé par tous les règnes du vivant.
Nous avons également montré que le centre de ce réseau métabolique est surtout constitué des voies métaboliques essentielles pour la survie et qui sont communes à toutes les espèces et que les voies métaboliques accessoires qui sont spécifiques de certaines espèces s'articulent autour. Cette organisation peut s'expliquer par le fait que le précurseur des

voies métaboliques accessoires est principalement les métabolites issus des voies métaboliques essentielles et communes à toutes les espèces (Keller, 2019).

En plus de leur position en périphérie du réseau et de leur faible connectivité, une grande majorité des activités enzymatiques spécifiques de certaines espèces sont en situation alternative à une autre activité enzymatique qui est généralement conservée. C'est-à-dire que la transformation d'un métabolite A en un métabolite B est effectuée par une activité enzymatique conservée mais peut être substituée par une activité enzymatique spécifique chez certaines espèces. Dans cette configuration, l'activité enzymatique qui est probablement capable de catalyser la réaction plus efficacement dans les conditions normales est conservée par toutes les espèces, et l'activité enzymatique spécifique assure la catalyse de la réaction dans les conditions plus spécifiques. Par conséquent, Il y a plus de pression pour conserver l'activité enzymatique le plus indispensable. Le réseau exerce aussi une pression pour qu'au moins une activité enzymatique soit conservée dans une espèce pour ne pas casser la voie métabolique.

On peut alors se poser la question de comment expliquer une grande conservation des activités enzymatiques entre les champignons et les autres règnes du vivant , et surtout de comment expliquer cette contrainte dans l'expansion du réseau métabolique.

Faire émerger une nouvelle activité enzymatique à partir d'une séquence aléatoire du génome (gène de novo) est un phénomène extrêmement rare (Jacob, 1977). François Jacob disait « L'évolution ne tire pas ses nouveautés du néant. Elle travaille sur ce qui existe déjà ». Faire émerger une nouvelle enzyme à partir d'une séquence aléatoire va aussi à l'encontre des principaux modèles sur l'évolution des voies métaboliques : le modèle d'évolution patchwork (Jensen, 1976) et le modèle d'évolution rétrograde (Horowitz, 1945) qui sont principalement basés sur la duplication-divergence. De plus, il a été montré que les familles d'activités enzymatiques (activités enzymatiques proches) partagent en général une similarité de séquence et un site actif très conservé (Cantarel *et al.*, 2009; Laurian *et al.*, 2019). Il est donc plus que probable que les activités enzymatiques d'une même famille dérivent d'un même ancêtre commun.
Pour faire émerger une nouvelle activité enzymatique, le chemin le plus court est de le faire émerger par duplication-divergence en mutant une copie du gène dupliqué d'une activité enzymatique existante et proche de la nouvelle activité enzymatique.
Si deux activités enzymatiques identiques émergent indépendamment dans deux espèces non reliées taxonomiquement, il semble cependant plus parcimonieux d'imaginer qu'il y a eu convergence à partir d'un élément qui existait déjà et qui était en commun entre les deux espèces : c'est une convergence fonctionnelle à partir d'un élément en commun (Figure 15.1).
Étant donné que le point de départ est commun (activité enzymatique commune entre les deux espèces), pour s'adapter à une même contrainte, la même activité enzymatique va probablement émerger. Ce qui permet d'expliquer pourquoi des activités enzymatiques

peuvent être retrouvées à la fois chez les champignons et les autres règnes du vivant.

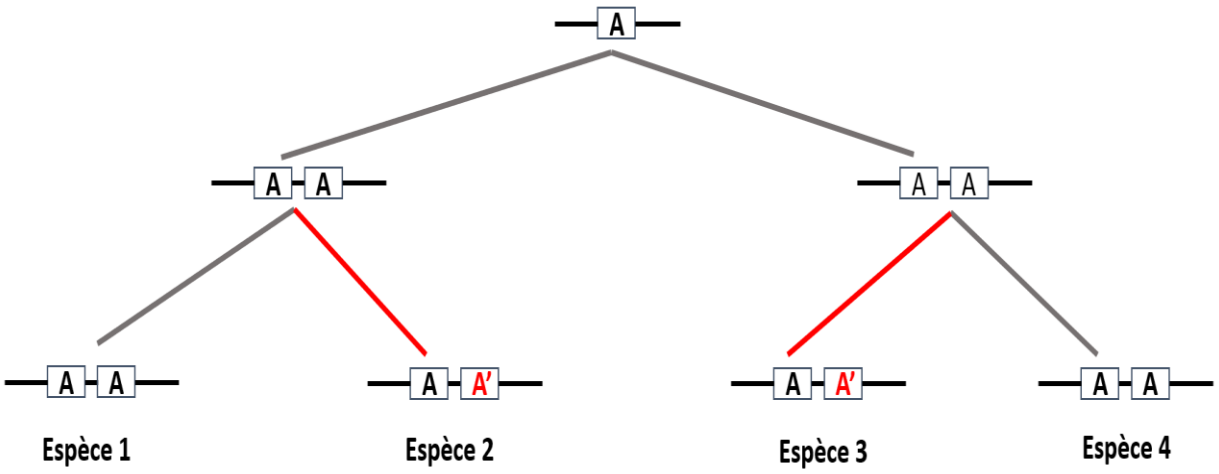


Figure 15.1 : Convergence fonctionnelle à partir d'une même enzyme de départ. L'activité enzymatique **A** est présente chez l'ancêtre des 4 espèces. Chez l'espèce 2 et 3 une activité enzymatique **A'** est apparue à partir d'une copie de A du fait de la même pression de sélection sur les deux espèces.

Si la duplication-divergence est le principal moteur de l'évolution du réseau métabolique, et que la convergence fonctionnelle a aussi joué un rôle dans la dynamique évolutive des activités enzymatiques en plus des évènement de pertes, les activités enzymatiques que nous avons identifiées en dehors des champignons peuvent donc avoir comme origine une apparition indépendante par convergence fonctionnelle. Ces activités enzymatiques étaient probablement nécessaires à la survie de l'espèce. Du fait de la même pression, l'innovation a débuté à partir d'un élément qui existait déjà et était probablement déjà commun. Des cas de convergences fonctionnelles ont été reportés chez les champignons, plus particulièrement pour activités enzymatiques associées à la dégradation de la lignine. Une enzyme ancestrale incapable de dégrader la lignine était présente chez l'ancêtre commun des champignons puis cette enzyme a évolué indépendamment dans différente espèce non relié taxonomiquement pour développer les activités enzymatiques actuelles capables de dégrader la lignine (Ayuso-Fernández *et al.*, 2018).
Différencier les activités enzymatiques présentes au niveau de l'ancêtre commun puis perdues au cours de l'évolution ainsi que les activités enzymatiques apparues indépendamment dans les différentes espèces par convergences fonctionnelles à partir d'un élément en commun est cependant loin d'être trivial si l'on dispose uniquement des profils phylogénétiques.

En outre, la conservation des activités enzymatiques peut aussi s'expliquer par la conservation des métabolites. Les activités enzymatiques spécifiques alternatives imposent une conservation des métabolites. Il y a donc une pression pour maintenir les métabolites déjà présents dans le réseau. En effet, depuis plusieurs milliards d'années d'évolution, et malgré la diversité enzymatique dans le réseau, les métabolites de bases que catalysent ces enzymes sont restés inchangés et sont les mêmes entre tous les organismes vivants. Les cofacteurs qui activent les enzymes sont aussi universels et communs à tous les règnes du vivant : les ions métalliques (fer, cuivre, zinc et magnésium), l'ATP, le coenzyme A, le FAD et le NAD+….. Les métabolites principaux sont les mêmes, par conséquent les activités enzymatiques les catalysant sont très conservées et les activités enzymatiques qui sont reliées à ces métabolites sont forcément partagées par plusieurs espèces.

La relation métabolite-enzyme joue un rôle dans la conservation des activités enzymatiques. Alam et ses collaborateurs ont montré qu'une enzyme a en moyenne 5 métabolites qui peuvent l'inhiber dans le réseau (Alam *et al.*, 2017). Ces métabolites inhibiteurs ont probablement des structures similaires à celles des substrats de l'enzyme et sont principalement des métabolites des enzymes adjacents dans le réseau. Afin d'échapper à ces inhibitions, les enzymes ont suffisamment évolué pour échapper à la majeure partie de ces inhibiteurs. Faire émerger une enzyme avec une nouvelle activité enzymatique à partir d'une séquence aléatoire (de novo) semble imposer plus de contrainte (développement d'un site actif, reconnaissance du substrat et échappement aux inhibiteurs) que faire émerger une activité enzymatique par duplication divergence à partir d'une enzyme déjà présente. Une enzyme déjà présente dans le réseau a en effet suffisamment évolué pour échapper à certains inhibiteurs, possède déjà un site actif capable de reconnaître un substrat et facilement mutable pour reconnaître un autre substrat de la même famille. Une enzyme déjà présente constitue donc une base plus fiable et plus rapide pour faire émerger une nouvelle activité enzymatique.

Si le mécanisme principal et le plus simple pour faire émerger une nouvelle activité enzymatique semble donc être la duplication-divergence. L'ensemble des contraintes et mécanismes et des contraintes évoqués précédemment font que probablement des activités enzymatiques sont partagées entre plusieurs espèces en plus des activités enzymatiques qui étaient déjà présentes chez l'ancêtre commun.

Le mécanisme utiliser pour faire émerger une nouvelle activité enzymatique fait que les mêmes activités enzymatiques peuvent donc émerger indépendamment dans plusieurs espèces. En outre, le réseau métabolique joue un rôle essentiel pour limiter son expansion. Nous avons en effet montré qu'une des caractéristique du réseau métabolique est sa faible densité. Une faible densité signifie une très forte réduction des interactions entre les différentes réactions enzymatiques. Par conséquent le réseau va limiter le nombre de réactions enzymatiques dans le réseau. Réduire la densité du réseau peut être une stratégie pour économiser de l'énergie. Les réactions métaboliques nécessitent en effet de l'énergie pour se produire, et un réseau plus restreint peut minimiser les coûts énergétiques

nécessaires à la synthèse et au maintien de nombreuses protéines enzymatiques. Cette densité faible contribue aussi à réduire le réseau métabolique à un niveau minimal qui est plus facile à maintenir et réguler afin de maximiser l'efficacité des réactions déjà présentes.
Pour maintenir l'intégrité de ce réseau sans affecter la densité, nous avons observé que les activités enzymatiques spécifiques sont en grande partie des formes d'alternatives à une activité enzymatique conservée. Cela permet de suppléer certaines activités enzymatiques sans créer de nouvelles connexions entre de nouvelles réactions métaboliques.
Certaines réactions métaboliques peuvent aussi produire des produits toxiques. Un réseau métabolique à densité faible minimise le risque de production excessive de tels composés indésirables. Maintenir cette densité faible a pour conséquence de limiter l'expansion du réseau métabolique et exerce une pression pour maintenir si possible les réactions déjà présentes et éliminer les réactions enzymatiques dispensables.

Origine de la diversité métabolique chez les champignons

Si les champignons partagent une grande majorité des activités enzymatiques avec les autres règnes du vivant, la question reste entière de savoir comment expliquer la diversité métabolique des champignons. Il reste également à comprendre comment les champignons vont utiliser cette diversité pour s'adapter à leur milieu.

Les champignons ont colonisé de nombreuses niches écologiques, et les pertes d'activités enzymatiques témoignent probablement de leur adaptation en éliminant les activités enzymatiques non essentielles et en perfectionnant les activités enzymatiques essentielles (Ayuso-Fernández *et al.*, 2018). Dans ce projet, nous avons identifié seulement 8 activités enzymatiques spécifiques des champignons. Ce constat permet de se demander comment les champignons ont innové d'un point de vue métabolique pour s'adapter à leur milieu.

La diversité métabolique peut s'expliquer au niveau des systèmes de régulation des voies métaboliques. Par exemple dans le domaine de l'industrie, la dégradation de la biomasse végétale a de nombreuses applications dans le domaine alimentaire, textile ou encore celui de la production de biocarburants. Seulement quelques espèces d'ascomycètes ont été choisies pour une utilisation industrielle (Pariza and Johnson, 2001). Ces espèces ont été spécifiquement choisies par la présence d'inducteurs et d'un système de régulation de gène qui permet d'augmenter leur productivité (Benocci *et al.*, 2017). Mais ces systèmes de régulations sont seulement limités à des sous-groupes d'espèces (Todd *et al.*, 2014). Malgré le nombre d'activités enzymatiques partagées entre les espèces éloignées ou même proches, la régulation des voies métaboliques et des réactions enzymatiques peuvent donc jouer un rôle essentiel dans l'adaptation à un environnement particulier et participe ainsi à la diversité métabolique. Par exemple l'inhibition ou la surproduction d'un métabolite à partir de sa réaction enzymatique par la régulation du gène codant l'enzyme ou par un autre mécanisme peut avoir une conséquence sur la capacité métabolique d'un organisme et définir son phénotype (Wang *et al.*, 2017).

La classification d'une enzyme est principalement basée sur le(s) substrats, le(s) produits et

la nature de la réaction enzymatique. Cette classification ne prend pas en compte le pouvoir catalytique de l'enzyme. Par exemple dans la production de la pénicilline, *Penicilium notatum* a été initialement utilisé pour sa production mais le criblage de plusieurs espèces de *Penicilium* a permis d'identifier une autre espèce pour une production à grande échelle qui est *Penicilium chrysogenum* (Ziemons *et al.*, 2017). De la même façon, la voie de production de l'aflatoxine est très bien étudiée ainsi que l'organisation des gènes impliqués qui sont organisés en cluster (Ingolia and Queener, 1989). Malgré cela, le mécanisme à l'origine des différentes capacités de synthèse entre les différentes espèces de champignons reste encore inexpliqué. Plusieurs hypothèses ont été proposées comme le nombre de copies de gènes dans le génome, différents mécanismes de régulation ainsi que le pouvoir catalytique des enzymes entre différentes espèces (Ziemons *et al.*, 2017).
Il est donc possible que la capacité catalytique des enzymes entre les différentes espèces soit une hypothèse pour rendre compte de la capacité métabolique des champignons. La différence de vitesse de réaction entre les enzymes des différentes espèces pourrait ainsi avoir un impact physiologique. Cette information n'est pas contenue dans la classification des enzymes (EC-number). La représentation des voies métaboliques indique l'enchaînement des activités enzymatiques et nous avons montré que ces activités enzymatiques sont conservées chez beaucoup d'espèces. La succession des réactions enzymatiques peut donc être beaucoup conservée du fait aussi de la conservation des métabolites, mais le pouvoir catalytique des enzymes responsable de ces réactions peut en revanche être totalement différent pour permettre à l'organisme de s'adapter à son environnement.

En plus de leur pouvoir catalytique, certaines enzymes sont capables de reconnaître plusieurs substrats.
La promiscuité enzymatique se définit comme la capacité d'une enzyme à catalyser une réaction autre que celle pour laquelle elle a été spécialisée. Les enzymes sont surtout annotées en fonction de leur activité principale (Danchin, 2009; Khersonsky and Tawfik, 2010) car la plupart du temps on ne trouve que ce que l'on cherche et que l'on connaît déjà. Dans les voies qui sont bien étudiées, par exemple la glycolyse, la comparaison des enzymes permettant la transformation du beta-D-glucose en beta-D-glucose-6P (glucokinase) a permis de démontrer qu'une des enzymes conservées par toutes les espèces possède une spécificité pour le D-mannose et le D-fructose (2.7.1.1) et l'autre enzyme spécifique de certaines espèces est spécifique du beta-D-glucose (2.7.1.2). 2.1.1.1 est donc une alternative à 2.7.1.2 qui est présent dans plusieurs réactions enzymatiques du réseau. Ces découvertes ont permis de corriger la classification (EC-number) de ces enzymes en apportant une plus grande résolution sur l'activité enzymatique de l'enzyme.
Si la plupart des enzymes sont annotées en fonction de leur activité enzymatique principale, les voies métaboliques aujourd'hui ne sont donc tout simplement que la représentation des successions des activités principales des enzymes. Ce qui est probablement le plus essentiel pour assurer la survie de l'organisme et par conséquent présentes chez beaucoup d'espèces même en dehors des champignons.
On peut supposer qu'en plus des activités principales des enzymes chez les champignons, la

plupart des enzymes sont capables d'assurer d'autres activités.
L'ensemble de ces réactions catalysées par des enzymes non spécifiques constituent l « underground métabolisme » (D'Ari and Casadesús, 1998). L'analyse de l'underground metabolism chez E.Coli a permis de démontrer que ces activités enzymatiques issues d'enzymes infidèles facilitent l'adaptation à un milieu mais aussi de prédire les milieux dans lequel l'espèce peut s'adapter (Notebaart *et al.*, 2014).
Ceci montre donc que l'activité secondaire d'une enzyme peut jouer un rôle essentiel dans la diversité métabolique et phénotypique. Le réseau métabolique que nous observons aujourd'hui n'est peut-être que la partie émergée de l'iceberg qui assure les fonctions et les activités enzymatiques essentielles.

VI. Projet d’article

Cette partie du manuscrit contient une version préliminaire de l’article qui reprend les informations contenues dans les parties II et IV. L'article sera déposé sur BioRxiv dans un premier temps, puis soumis dès que possible.

La première partie de l’article décrit la conservation des activités enzymatiques chez les champignons ainsi que la classification des activités enzymatiques à l’aide de profils phylogénétiques similaires."

Dans un second temps nous regarderons l’inférence des origines évolutives des activités enzymatiques par une approche phylostratigraphique

Enfin, les informations évolutives seront cartographiées sur le réseau métabolique global et sur les différentes voies métaboliques pour comprendre les contraintes exercées par le réseau métabolique sur l’évolution des activités enzymatiques.

1

Title

Exploring the evolution of metabolic network in fungi

Authors

Vahiniaina Herinjiva Andriamanga[1], Anne Lopes[1], Olivier Lespinet[1]

Affiliations

[1] Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvettes.

Corresponding author

Olivier Lespinet, Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Université Paris-Saclay, 91198, Gif-sur-Yvette, France

Email : olivier.lespinet@i2bc.paris-saclay.fr

Keywords

Fungi, Metabolic network evolution, Comparative genomics, Graph theory

Abstracts

The metabolic network represents the relationships between all biochemical reactions. It defines the metabolic capacity of the organism to use compounds available in the environment and to synthesize new products. Consequently, the environment plays a crucial role in constraining the evolution of metabolic networks. To unravel the evolution dynamics of the metabolic network, we investigated the evolution of 910 enzyme activities in 174 species of fungi using a unique combination of phylogenetic profiles and graph-based analysis techniques. The enzyme activities were categorized based on their conservation, with the initial 454 enzyme activities being universally present in all the species studied, whereas the remaining 456 were associated with particular clades or specific species. Using a phylostratigraphy approach, we reconstructed the evolutionary history, encompassing both the losses and acquisitions, of enzyme activities related to specific clades or species. Our study revealed that 406 of these enzyme activities were already present in fungal ancestors but subsequently lost during the course of evolution, while 8 were newly emerged fungal-specific enzyme activities. Regarding their location within the metabolic network, lineage-specific enzyme activities tend to occupy the periphery. They exhibit lower connectivity within the metabolic network compared to common enzyme activities and often serve as alternatives to the common ones. Furthermore, when we group the enzyme activities based on the similarity of their phylogenetic profiles , we observe that those with similar profiles tend to cluster together within the network. Additionally, our observations indicate that the loss of network-disrupting enzyme activities tolerated for two reasons: either the affected portion of the subnetwork is considered accessory, or there exists an alternative enzyme activity in other species.

Introduction

Fungi are renowned for their diverse metabolic capabilities (Wainwright, 1988). Fungi exhibit exceptional chemical aptitude, enabling them to synthesize a wide variety of metabolites. These include compounds of therapeutic interest (Dutta *et al.*, 2022), such as antibiotics like penicillin (Fleming, 1929), anticancer drugs (Kornienko *et al.*, 2015), and immunosuppressants (Wong *et al.*, 2017). Furthermore, fungi produce enzymes of industrial importance (El-Gendi *et al.*, 2021) as well as mycotoxins (Bennett and Klich, 2003). They can break down an extensive range of substrates (Wainwright, 1988), including cellulose and lignin, which are the two most abundant biopolymers on earth (Bouws *et al.*, 2008).

In addition to their remarkable biochemical repertoire, fungi exhibit an astonishing diversity of lifestyle (Naranjo-Ortiz and Gabaldón, 2019). They have successfully colonized every corner of the earth. Fungi can adopt various roles, such as saprobic, lichenized, pathogenic, endophytic, symbiotic, and commensal, while interacting with a wide spectrum of hosts.

Adapting to different lifestyle, metabolism serves as the forefront, continuously adjusting to maintain balance and ensure the organism's survival. Metabolism comprises a connected series of enzyme-catalyzed reactions that take place within the organism. These reactions within the metabolic network are facilitated by enzymes. Enzymes are proteins or RNA molecules (Cech *et al.*, 1981), serving as biological catalyst that accelerates biochemical reactions in living organisms at a pace compatible with life. Enzymes are classified based on the overall chemical transformation of substrates into products (Tipton and Boyce, 2000). Each enzyme is assigned a unique four-digit code, known as the Enzyme Commission or EC-number, which describes its specific enzymatic activity. The sequence of these reactions constitutes metabolic pathways, and their interconnections forms the complete

metabolic network of an organism. This metabolic network defined the organism's metabolic capacities, including its capacity to use chemical compounds present in the environment and to synthesize new products. The immense variety of living organisms reflects the extensive diversity of metabolic networks.

The diversity of species and lifestyles raises the intriguing question of how the metabolic network and metabolic pathways have evolved to support life in different environments. With the growing wealth of genomic data available, comparative studies that highlight differences and similarities between species serve as powerful tools for understanding the organization and the evolution of intracellular structure. Phylogenetic profiles elucidate the collective presence and absence of traits (such as protein, gene, phenotype) within a set of species. Phylogenetic profile was initially employed to assign potential function to uncharacterized proteins, based on the premise that a shared profile suggests a functional relationship between proteins (Pellegrini *et al.*, 1999). Phylogenetic profiles from species covering the three domains of life, have been used in the analysis of metabolic network element to identify evolutionary modules, which are sets of enzymes that have co-evolved and exhibit similar profiles, and they are specific to each domain of life (Yamada *et al.*, 2006; Peregrín-Alvarez *et al.*, 2009; Li *et al.*, 2016).

Several databases dedicated to metabolic pathways are now available (Labena *et al.*, 2018), and metabolic pathways are the most preferred representation of metabolism by regrouping enzymes based on their involvement in specific processes, as seen in resources like KEGG (Kanehisa and Goto, 2000) and MetaCyc (Karp *et al.*, 2000). However, various techniques can reconstruct an organism's metabolic networks by leveraging sequence similarity searches on known enzymes (Chalancon *et al.*, 2013).

By applying tools from graph theory, researchers have delved into the topological

characteristics and organizational structures of metabolic networks. Analyses of these metabolic networks have unveiled properties such as small-worldness (Fell and Wagner, 2000; Jeong *et al.*, 2000), scale-freeness (Barabási and Albert, 1999), and modularity (Hartwell *et al.*, 1999; Ravasz *et al.*, 2002). It's worth noting that while these features have been observed, some have also faced criticism (Lima-Mendez and Helden, 2009).

Network comparison was widely used for assessing similarities and differences between the organization of the metabolic network from different taxa (Jeong *et al.*, 2000; Ma and Zeng, 2003; Zhu and Qin, 2005; Banerjee, 2012), as well as for understanding how environmental factors influence their structure (Parter *et al.*, 2007; Takemoto *et al.*, 2007; Kreimer *et al.*, 2008).

Some studies started to shed light on the constraints imposed by the network on the evolution of its components (Peregrín-Alvarez *et al.*, 2009) demonstrating that highly conserved enzyme activities tend to be more connected. Additionally, it was revealed (Montanucci *et al.*, 2018) that the most conservative genes within the human metabolic network are located in the final stages of the metabolic pathways, while genes initiating these pathways are more susceptible to evolutionary changes. All these studies suggest that the structure of metabolic network and pathways exerts constraints on the evolution of their individual elements.

To date, the majority of analyses in this field have been conducted on prokaryotes or on broad taxonomic groups encompassing all three domains of life. However, during the early 2000s, substantial efforts were made to sequence fungal genomes (Cuomo and Birren, 2010; Grigoriev *et al.*, 2014). In parallel, a dedicated tool for the functional annotation of genes encoding enzymes has been developed (Grossetête *et al.*, 2010). This deluge of

data, along with the development of reliable methods and the wide diversity of fungal metabolic profiles, presents an exceptional opportunity to explore the evolution of fungal metabolic networks.

In this article, we present an integrative approach to assess the constraint exerted by the metabolic network on the evolution of fungal enzyme activity. Our goal is to gain insight into how the extensive diversity of fungal metabolism influences the evolution of the metabolic network. This integrative approach combines evolutionary information related to enzyme activity with tools from graph theory to analyze the structure the metabolic network.

Results

In order to investigate the evolution of enzyme activities (as represented by enzyme commission number, EC numbers) in fungi, enzyme activity identification was performed for 174 fungal genomes [see methods]. We identified a total of 968 enzyme activities. These enzyme activities were mapped on metabolic pathways available from KEGG, and out of the total, 910 enzyme activities correspond to biochemical reactions that operate in fungal metabolism.

The distribution of enzyme activities across 174 fungal species shows different amounts of enzyme activity per taxonomic class (Figure 1). The 174 fungal species are distributed in 21 classes (the number of species per taxonomic class is available in supp1). The distribution of enzymatic activities per species shows that microsporidia have significantly reduced metabolism compared to other classes. Microsporidia are obligate intracellular parasites and have lost many of their metabolic functions (Corradi, 2015). Most *Ascomycota* classes exhibit a higher diversity of enzyme activities compare to other fungi classes. However, it's worth noting that *saccharomycetes* and *schizosaccharomycetes* display more limited enzyme activities when compared to other classes within Ascomycota. Within *Eurotiomycetes*, two orders emerge: Eurotiales with 700 enzymatic activities on average and *Onygenales* with 600 enzymatic activities on average. *Eurotiales* are mainly associated with plants, while *Onygenales* are mainly human pathogens. Wang and collaborators showed that most gene reductions in *Onygenales* are genes associated with cellulose degradation and genes directly related with plant interactions (Wang *et al.*, 2022). The reduction in enzymatic diversity between the two orders reflects this difference in lifestyle.

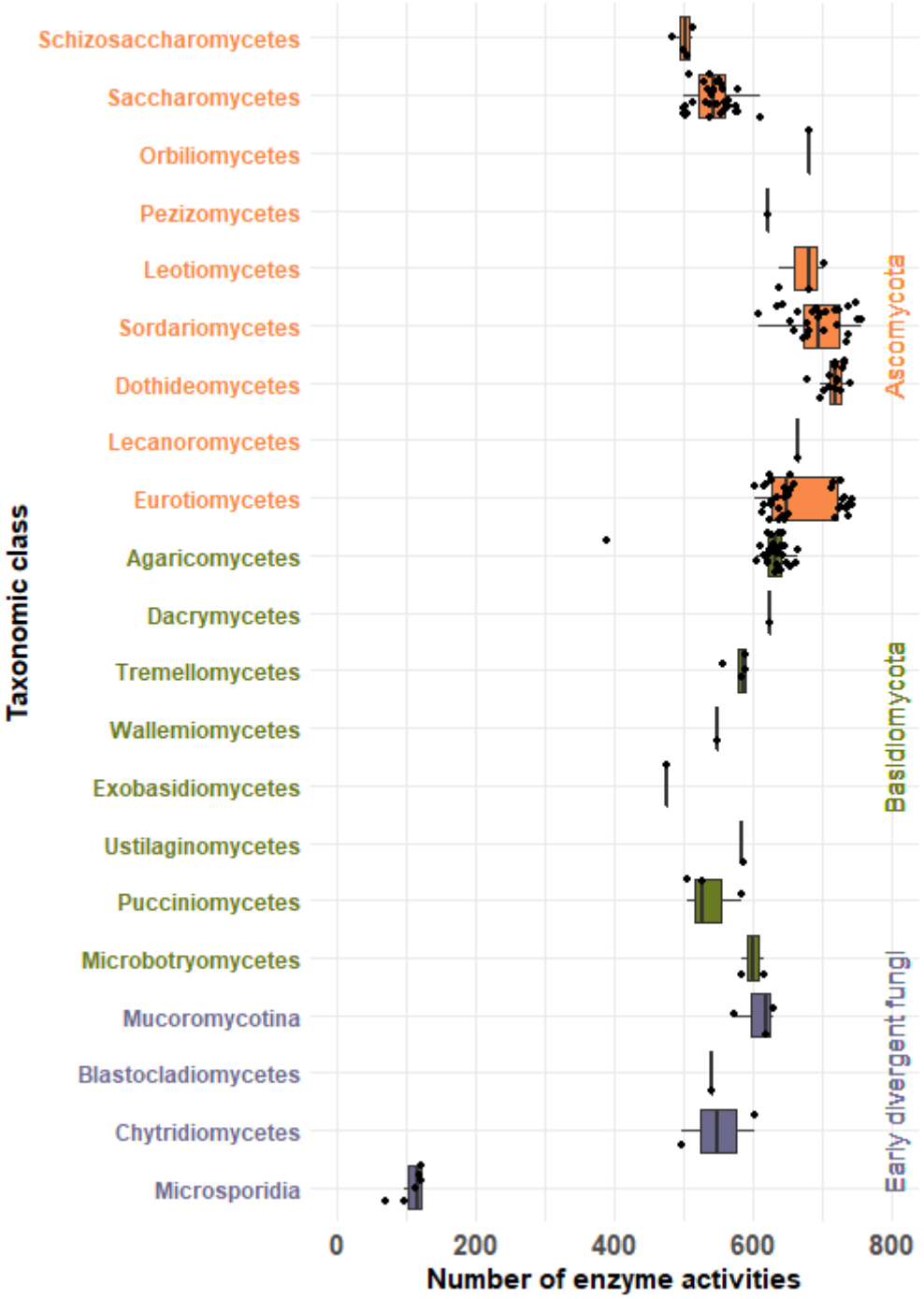


Figure 1: **Boxplot representing the number of enzymatic activities per species grouped by taxonomic class.** On the x-axis is the number of enzymatic activities per species , listed in order alongside the respective taxonomic class of each species. The coloration differentiates between various fungal groups, with Orange representing Ascomycota, green representing Basidiomycota, and purple indicating early divergent fungi.

Enzyme activities evolution in fungi

The enzyme activities were categorized based on their conservation across the studied species to determine which enzyme activities were present in all species and which were lineage-specific, meaning they are only found in specific clades or species. Conservation analyses revealed that half of the enzyme activities exhibited a conservation rate ranging from 85% to 100%. The remaining enzyme activities demonstrated a conservation lower than 85% with an average of 20 enzyme activities per bin (Figure 2).

As a result, enzyme activities that were found in at least 85% of species were considered shared by all species, indicating they were common to all (highly conserved). Out of the 910 enzyme activities studied, 454 were shared by all studied species, while 456 were lineage-specific enzyme activities. Highly conserved enzyme activities most likely represent reactions essential for life, whereas lineage-specific enzyme activities contribute to the origin diversity of metabolic profiles.

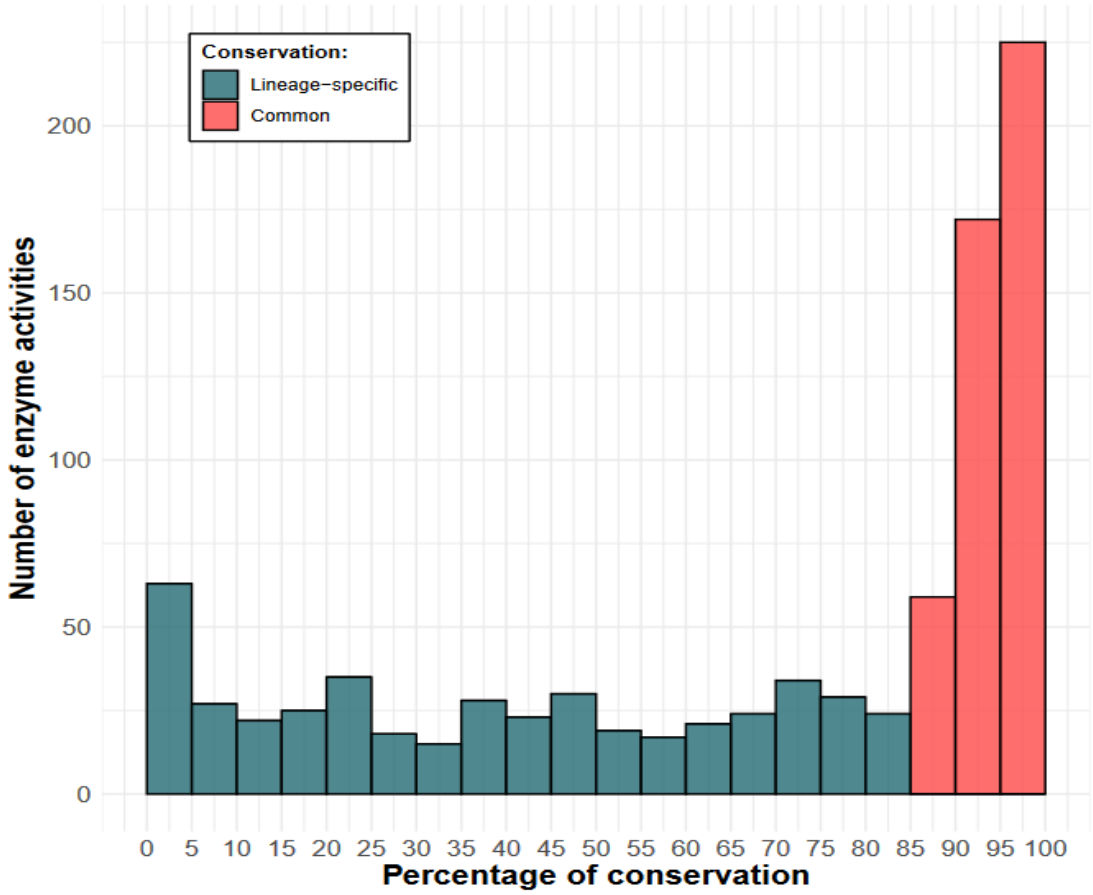


Figure 2. The number of enzyme activities based on their conservation level. Each bin represents a 5% conservation interval. Enzyme activities with at least 85% of conservation were considered shared by all studied species and are represented by the red bin, containing 454 enzyme activities. In contrast, lineage-specific enzyme activities with less than 85% conservation are exclusive to specific species and are depicted by the blue bins, which encompass 456 enzyme activities.

Among the lineage-specific enzyme activities, some are observed to be either present and absent within the same species or unique to specific clades. Therefore, these enzyme activities exhibit similar profiles. To identity enzyme activities with similar profiles, enzyme activities were clustered according to the similarity of their phylogenetic profiles using a

Clustering Aggregation method (Dib and Carbone, 2012). As a result, we identified 15 clusters of enzyme activities that share a similar evolutionary profile, each comprising distinct sequences bearing EC numbers that match the same profile.

It should be noted that 431 enzyme activities remained unclustered, meaning that their profiles did not exhibit similarity with any other profiles. Among the 15 clusters, CLAG identified a larger cluster corresponding to highly conserved enzyme activities, which consist of 420 enzyme activities. The size of the other 14 clusters ranges from 15 to 2 (Supp 2). These 14 clusters are mainly composed of lineage-specific enzyme activities. Representing these clusters on an organized heatmap where species are ordered according to the phylogenetic tree reveals a subset of enzyme activities that are exclusively present or absent within specific taxonomic classes (Figure 3), indicating co-evolution of enzyme activities within the same cluster.

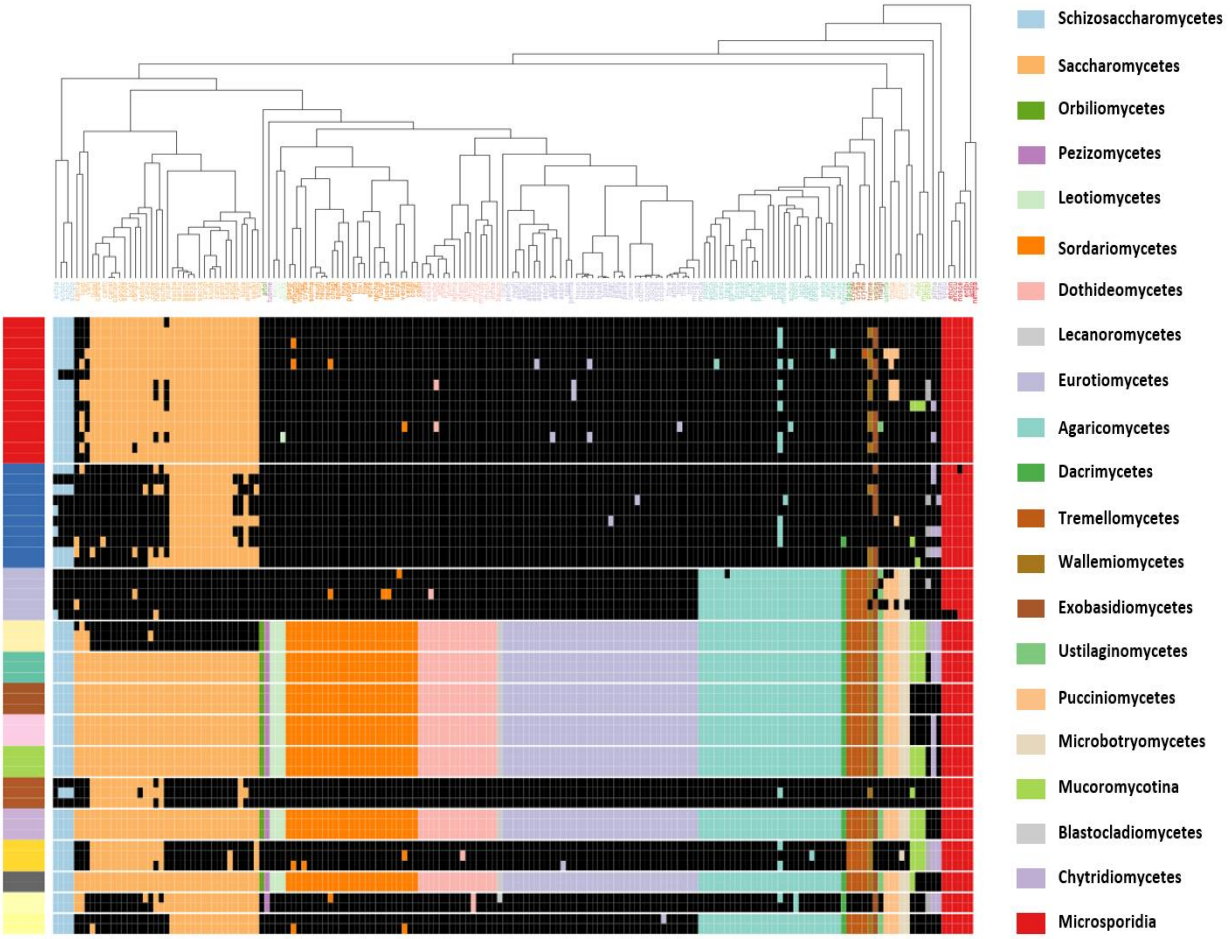


Figure 3: Heatmap displaying the conservation of enzyme activities sharing similar profiles. Matrix showing the presence or absence of enzyme activities across 174 fungal species. Each row represents an enzyme activity, and each column represents a species. Rows are clustered by similarity in enzyme activity profiles. The colored bar on the left indicates the phylogenetic cluster to which the enzyme activities belong to. Columns are ordered according to the displayed phylogenetic tree above the matrix. The species names in the tree are color-coded according to the taxonomic class. Cells in the matric are black when the enzyme activity is present, or colored according to the taxonomic class when absent. On the left, a colored bar groups enzyme activities based on their conservation among the studied species.

The first cluster in Figure 3 comprise enzyme activities that are largely absent in most *saccharomycetes* species. Out of 14 enzyme activities of this cluster, 7 of them are associated with the leucine, isoleucine, and valine metabolism (Figure 4). In this cluster, the absence of these enzyme activities in specific lineages can be attributed to evolutionary losses. These lineages likely underwent genetic changes or adaptations that led to the loss of specific enzyme activities over time. Indeed, the presence of enzyme activities in specific lineages may indicate gain events that occurred at different points along the evolutionary tree. This suggests that these lineages independently acquired these enzyme activities. Furthermore, clusters encompassing phylogenetically unrelated species suggest specific losses or independent gains of those enzyme activities within those lineages. These observations highlight the dynamic nature of enzyme activity evolution, including both gains and losses. These processes contribute to the diversity and complexity of metabolic profiles across different species.

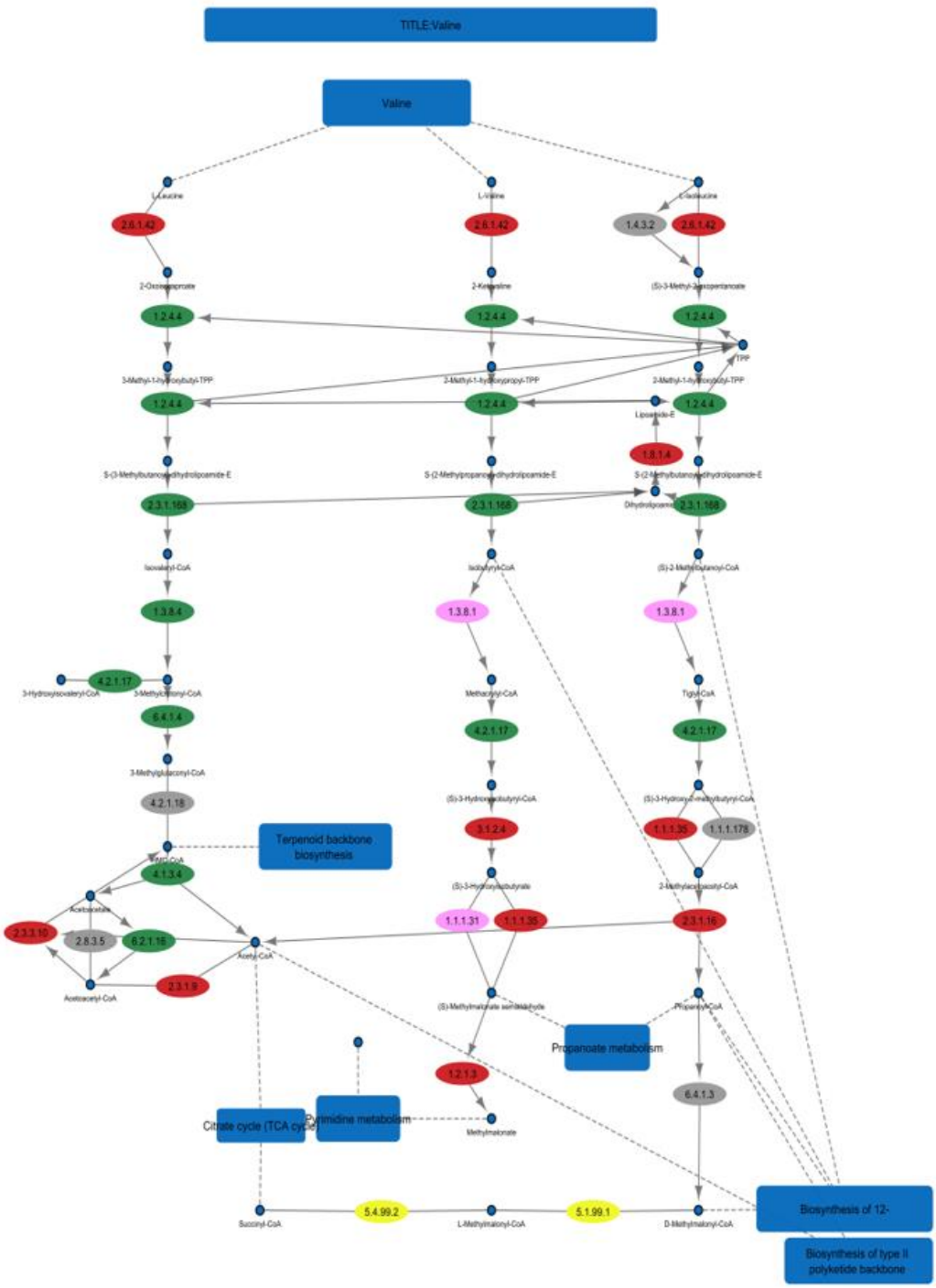


Figure 4: Valine, leucine, and isoleucine pathway. Enzyme activities are color-coded according to their profile similarity. Grey enzyme activities represent enzyme activities without similarity with other phylogenetic profiles. Enzyme activities colored in green signify enzymes that are absent in most saccharomycetes. Red enzyme activities denote highly conserved enzymes

Inferring the evolutionary origin of an enzyme activity becomes challenging when the presence of the enzyme activity is scattered across the phylogenetic tree. Tracing the specific evolutionary history of such an enzyme activity can be complex. The scattered distribution may indicate multiple independent acquisition, gene losses, or horizontal gene transfer events at various evolutionary points. To address this, we employed phylostratigraphy to identify enzymatic activities that have been lost within a specific lineage, as well as fungal-specific enzyme activities that emerge during fungal evolution (see methods). This approach helps in reconstructing the evolutionary history of enzyme activities through homology analysis by examining their distribution across different taxonomic groups. It allows us to infer the timing of their origin and potential losses. Enzyme activities present in ancient lineages but absent in more recent ones suggest ancestral activities that have been lost over time.

On the other hand, enzyme activities specific to fungal lineages but absent in other organisms indicate innovations that arose during fungal evolution. The presence of the EC number outside the fungi suggests that this enzyme activity was already present in their last common ancestor. We excluded horizontal gene transfer events (HGT), considering that horizontal EC number transfer is rare in fungi (Wisecaver *et al.*, 2014).

In our analysis, we identified a total of 860 enzymatic activities with ancestral origins Among these, 447 enzyme activities were conserved across all species, while 416 enzyme activities were lost at various points during evolution. Only 50 enzyme activities lack homologous sequences outside of the fungi. Based on the literature, 8 out of 50 enzyme activities are EC-numbers only found in fungi (table 1). Collectively, these analyses support the idea that the evolution of the metabolic network was predominantly shaped by the loss of various enzyme activities in different lineages.

EC-number	Metabolic pathway	References
2.1.1.261	Ergot alkaloid biosynthesis	Wallwey *et al.*, 2012
2.5.1.34	Ergot alkaloid biosynthesis	Ding *et al.*, 2008
4.2.1.142	Aflatoxin production	Sakuno *et al.*, 2005
4.2.1.66	Cyanide degradation	Martínková *et al.*, 2015
5.4.99.32	Sesquiterpenoid and triterpenoid biosynthesis (Protostadienol antibiotic production)	Kimura *et al.*, 2010
1.1.3.13	Methane metabolism	Westrick *et al.*, 2022
4.2.1.143	Aflatoxin production	Ren *et al.*, 2017
4.2.3.43	Diterpenoid biosynthesis (Fusicoccin A production)	Toyomasu *et al.*, 2007

Table 1 : List of fungal specific enzyme activities and the associated metabolic pathway inferred using phylostratigraphy techniques.

Conservation level with respect to their localization within the metabolic network

We subsequently tried to understand the relationship between the conservation, loss, gain of enzyme activities and topology of the metabolic network.

We investigated the conservation level of enzyme activities relative to their localization within the metabolic network to investigate whether the metabolic network imposes constraints on the evolution of nodes (enzyme activities)(Figure 5 A). The metabolic network was constructed by combining the metabolic pathways from KEGG. The metabolic network comprises a total of 102 metabolic pathways. There are a total of 2,035 nodes representing 910 enzyme activities and 7,428 edges. The number of nodes highlights the recurrent use of the same enzyme activities across various pathways. Out of these nodes,

1047 represent highly conserved enzyme activities, 894 denote lost enzyme activities, and 8 signify novel enzyme activities. The observation of this metabolic network reveals a central "core" composed of highly conserved enzymes. We showed that highly conserved enzyme activities tend to be located at the center of the network compared to lineage-specific enzyme activities (Figure 5 B). This is reflected in their significantly higher closeness centrality values (P-value against loss: 10e-06, against gain: 10e-20). Closeness centrality measures a node's centrality within the network; a value closer to one indicates a more central position in the network. In addition, lost enzyme activities exhibit lower connectivity compared to highly conserved ones (Figure 5 C, P-value against highly conserved: 10e-06), implying that they share their compounds with fewer enzyme activities than highly conserved and novel enzyme activities.

Specific enzyme activities are less connected and are situated towards the periphery of the metabolic network compared to highly conserved and lost enzyme activities.

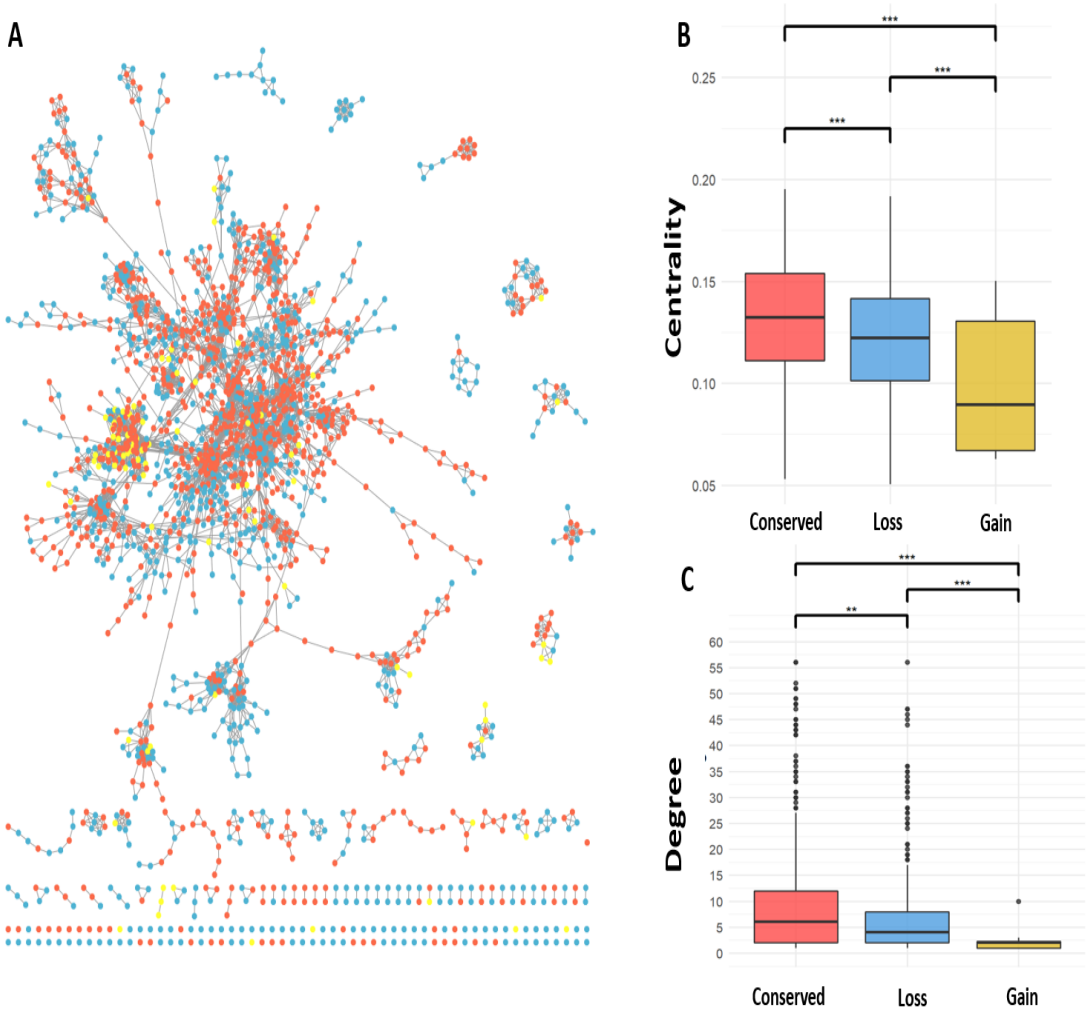


Figure 5: Enzyme activities topological properties. (A) The metabolic network: Nodes in the graph represent enzyme activities, connected if they share a common compound. Nodes are colored based on conservation and origin. Red nodes represent common enzyme activities, blue nodes denote ancestral enzyme activities lost by some species or clades, and green nodes signify novel enzyme activities. **(B) Closeness Centrality comparison**: A comparison of closeness centrality values among common, lost, and novel nodes. **(C) Connectivity comparison:** A comparison of connectivity levels among highly conserved, lost, and novel nodes.

Metabolic networks evolution with metabolic pathways viewpoint

The highly conserved enzyme activities at the core of the metabolic network corresponds to metabolic pathways shared by all species. Accessory metabolic pathways, specific to particular lineage are primarily located at the network's periphery. In addition to these two features, conserved pathways display a longer diameter than lineage-specific ones (p-value : 10e-30). Unlike lineage-specific pathways, conserved pathways contain multiple enzymatic reaction chains or routes for converting compounds into usable material (Supp 3). Highly conserved pathways are associated with essential processes such as amino acid metabolism, carbohydrates metabolism, energy metabolism, and nucleotide metabolism. These supergroups carry metabolic pathways essential for life. In contrast, most lineage-specific pathways in fungi are associated with the biosynthesis of secondary metabolites and the biodegradation and metabolism of xenobiotics.

Enzyme activities co-evolution

Subsequently, we examined the spatial distribution of enzymes with similar phylogenetic profiles (those present and absent in the same species) as identified through the clustering method within the metabolic network. We computed the distances between such enzyme activities and compared these distances with those between randomly selected lineage-specific nodes (Figure 6). Notably, the largest cluster, comprising highly conserved enzymes, encompasses approximately half of the enzymatic activities and nodes of the metabolic networks and was therefore analyzed separately. Our findings revealed that enzymes activities with similar phylogenetic profiles tend to be in close proximity to each other (nodes with similar profiles showed p-value of 10-16 against highly conserved nodes and 10-42 against lineage-specific nodes). Moreover, they are likely to be part of the same metabolic pathways (Figure 4). This emphasized the strong connection between their

functions and underscores the relationship between genomic and topological aspects.

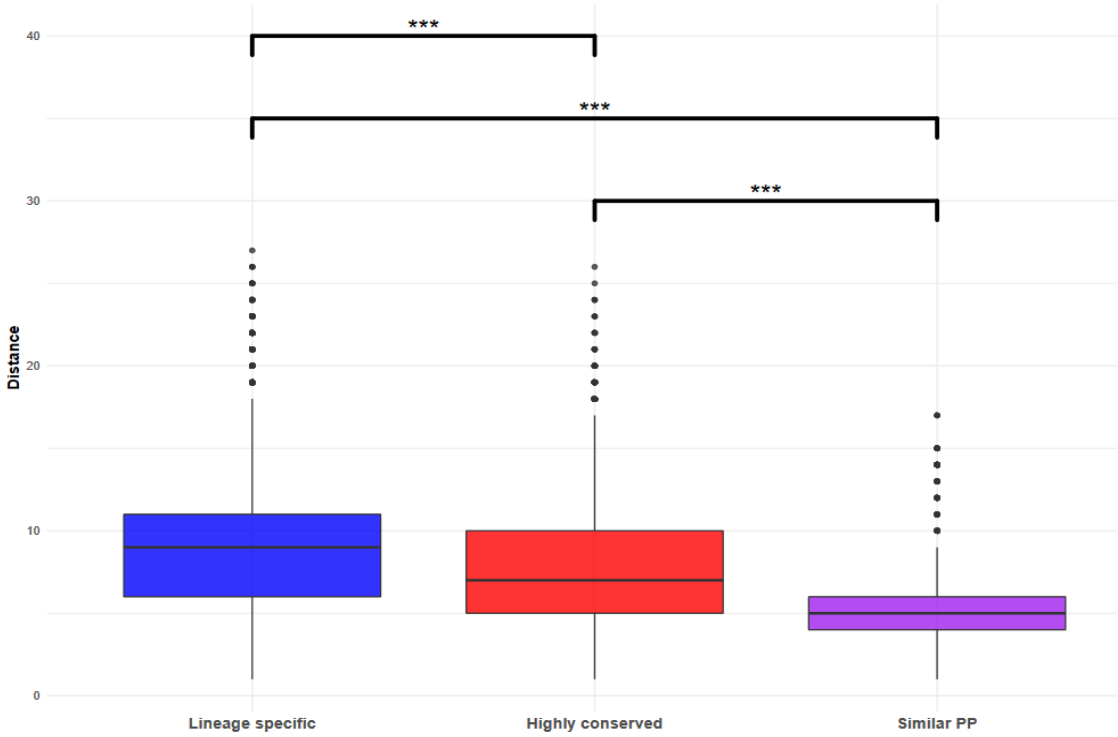


Figure 6: Enzymes with similar phylogenetic profiles (PP) and their promiscuity. Comparing pairwise distances between enzyme activities with similar profiles (in purple) and random enzyme activities (either from the highly conserved enzyme activities cluster in red or only lineage-specific enzyme activities in blue).

For example, the majority of enzymes within the first cluster in Figure 2 are associated with valine, leucine, and isoleucine metabolism (highlighted in green in Figure 6). This metabolic pathway catabolizes these three amino acids as carbon sources. Interestingly this set of enzymes is notably absent in the majority of the saccharomycetes species. When examining the evolutionary history of the enzyme that make up this pathway, it becomes evident that most species within the saccharomycetes class lack this pathway. Additionally, it has been demonstrated that *S. cerevisiae* cannot use valine, leucine and isoleucine as carbon sources (Cooper, 1982).

The placement of Lineage-specific node within Metabolic Pathway

To gain a more precise understanding of the specific locations within metabolic pathways where enzyme activity loss and gain predominantly occur, we have defined six distinct categories for lineage-specific enzyme activities within a metabolic pathway (Figure 7): (1) alternative: enzyme activity serve as alternatives to a reaction catalyzed by highly conserved enzyme activities, either for the transformation of metabolite A to B, or as an alternative to a chain of reactions. Alternative enzyme activity, therefore, uses or produces metabolites that have already been processed by highly conserved enzyme activities. (2) Development: Enzyme activity in this category are responsible for reactions that enable the synthesis of specific substrates or to synthesize specific metabolites. (3) Essential connector: these enzyme activities serve as connectors between two components (subpathways) with highly conserved enzyme activities on both sides. The loss or gain of this enzyme activity would disconnect or connect two subpathways. (4) A connector multiple: A set of interconnected enzyme activities that act as network-breaking elements by connecting two subnetworks. (5) essential alternative: these are enzyme activities that act as alternatives to lineage-specific enzyme activities. The absence of both enzymes

breaks the network. (6) a set of essential alternatives. This category includes interconnected lineage-specific enzyme activity that serve as alternative to either lineage-specific enzyme activity or another set of alternatives. The absence of both enzymes breaks the network.

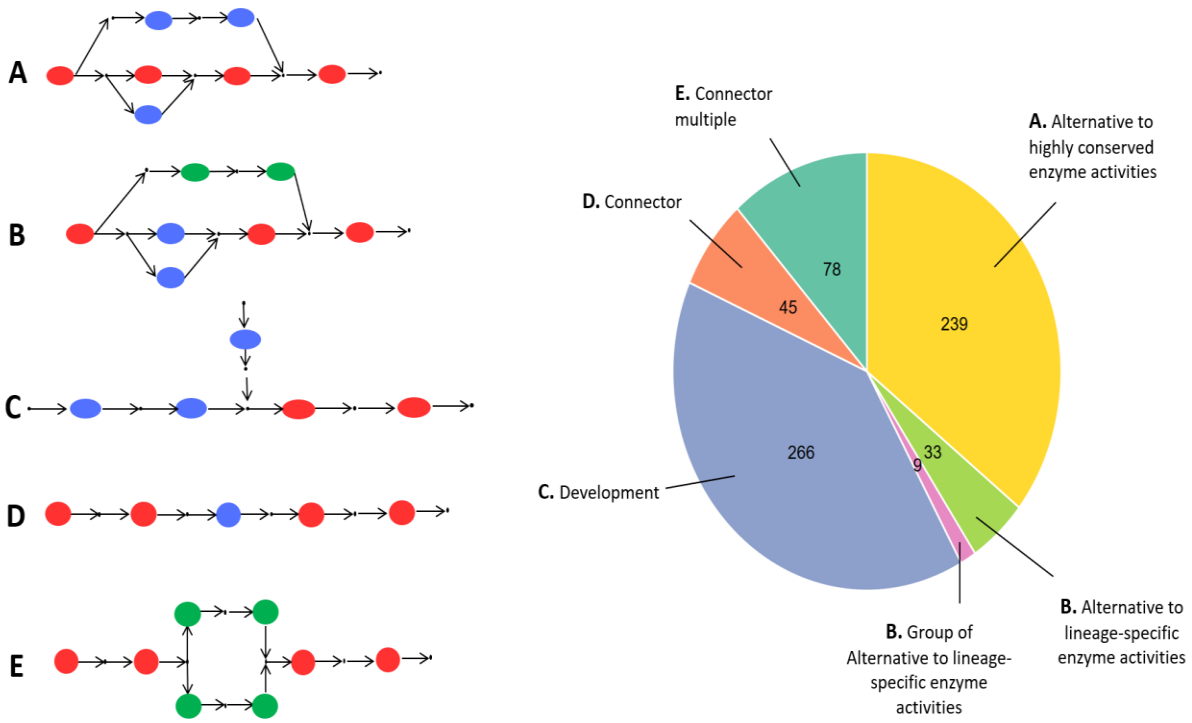


Figure 7: Specific node position. Definition of categories for each node within metabolic pathways. Red nodes indicate highly conserved nodes, blue and green nodes indicate lineage-specific nodes. Green nodes have to be considered togeether **A.** Alternative to highly conserved enzyme activities. **B.** Alternative to lineage-specific enzyme activities **C.** Development **D.** Connector **E.** Group of connector. The pie chart illustrates the distribution of lineage-specific nodes in each position.

While enzyme activities associated with metabolism are mostly of ancestral origin, there are numerous instances of losses within specific species or clades. In most cases, these losses occur when the reaction is already catalyzed by highly conserved enzyme activity, and the lineage-specific enzymes serve as alternatives, accounting for 35.7%. The second most prevalent category comprises enzyme activities involved in a development, representing 39.7%. This result is consistent with the observation that lineage-specific enzyme activities are usually occupy peripheral position of the metabolic network. Enzyme activities involved in development include enzymes that facilitate the use of alternative substrates or produce compounds that are accessory for a specific clade or species (Figure 7). The loss of these enzymes along with the compound linked to these reactions , exerts a limited impact on the overall metabolic network. This is due to the fact that only one or two adjacent enzymes rely on the compound associated with these reactions, thereby minimizing the risk of disrupting the metabolic pathway (the same holds true for alternative enzymes).

Conversely, losing enzyme activities that serve as essential connectors between two subnetworks, particularly when these connectors involved highly conserved enzyme activities in both subnetworks, appears implausible. This is because disrupting the connectivity between two crucial components could potentially be lethal for the organism. We observed that all essential connector enzyme activities are distributed among 22 pathways. Intriguingly, in 15 of these pathways, the connector node links a principal component housing the most highly conserved enzyme activities, referred to as essential routes. The small component, on the other hand, comprises one or two conserved alongside a multitude of lineage-specific enzyme activities. It’s worth noting that these lineage-specific enzyme activities exhibit no similarity in their phylogenetic profiles.

Assessing the function of this smaller subnetwork in the existing literature revealed that 11 out of 15 instances within this component can be characterized as accessory subnetwork. Moreover, the highly conserved enzyme activity within this subnetwork is frequently employed in different pathways, as evidence by 35 out of 46 common enzymes that are disconnected. This can elucidate its conservation in most of the studied species , as it serves as an essential component in these other pathways.

As an illustration, in tryptophan metabolism, such an extension generates a metabolite (the Indole-3-acetic acid) that is only essential for the competition between fungal species and fungal-plant interactions and mainly produced by fungi which interact with plant (Fu *et al.*, 2015).

If the subnetwork is essential for producing an essential metabolite, an alternative mechanism exists to supply for the absence of the enzyme activity. In the cases of biotin metabolism and inositol phosphate metabolism, the two components linked by lineage-specific enzyme activities represent essential pathways but include an essential connector (Figure 8B). These connectors enzyme activities are absent in Saccharomycetes. However, this clade can transport both metabolites externally (Chen *et al.*, 2008; Stolz *et al.*, 1999).

We observed the same features within the connector multiple. These nodes connect a main component housing the most highly conserved enzyme activities, often referred to as essential routes. Meanwhile, the small component encompasses one or two conserved and numerous lineage-specific enzyme activities, functioning as an accessory in 6 out of 14 pathways.

Alternative essential enzyme activities refer to a minimum of two lineage-specific enzyme

activities that can catalyze the conversion of metabolite A to B. The concern lies in the possibility that certain species may lack both of these enzyme activities, potentially disrupting the pathway. To address this, we conducted an analysis of the enzyme activity profiles by considering their combined presence. We observed that species lacking the first enzyme activities often possess the second one, as seen in the example of glycolysis metabolism involving 5.4.2.12 and 5.4.2.11 (Figure 8 A). In other cases, we encounter accessory sub-pathways featuring two enzymes with different specificities.

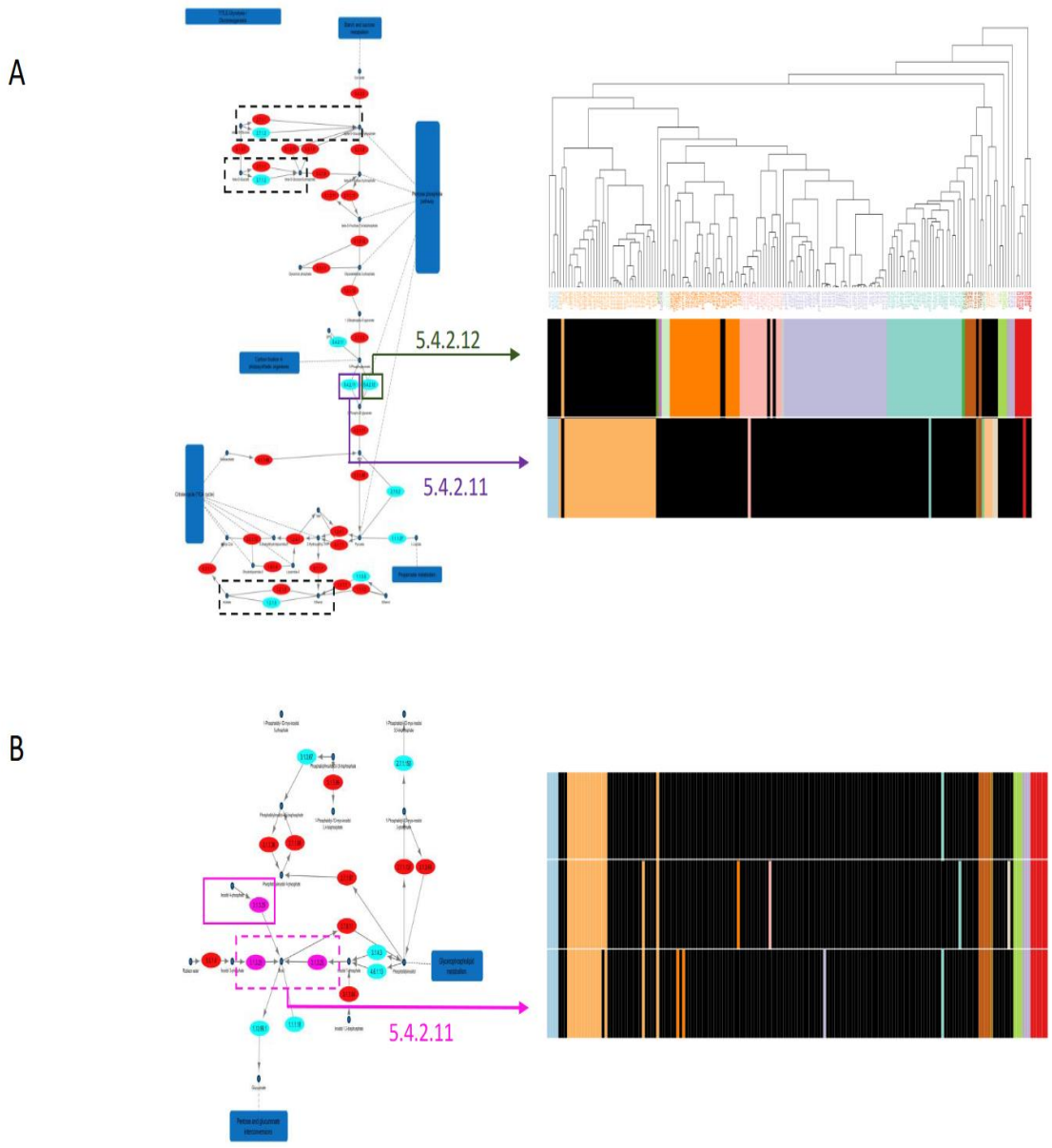


Figure 8: Representative example of alternative, essential connector, and essential alternative enzyme activities. In each pathway, the enzyme activities are color-coded for clarity: highly conserved enzyme activities are represented in red, while other enzyme activities are colored according to their profile similarity (excluding grey, which denotes enzyme activities without similarities). In the enzyme profile, black cells signify the presence of enzyme activity, while cells are otherwise colored according to the taxonomic class of the species.

A Glycolysis/Gluconeogenesis metabolism. Lineage-specific enzyme activity alternatives to a highly conserved enzyme activity are denoted by a black dashed frame. 5.4.2.11 and 5.4.2.12 serve as essential alternatives. Their respective phylogenetic profile are displayed on the right.

B Inositol phosphate metabolism. In inositol phosphate metabolism, the absence of 3.1.3.25, indicated by a magenta dashed frame results in the isolation of 5.5.1.4 (represented in grey) leading to the disruption of inositol production. The phylogenetic profile of 3.1.3.25 is displayed on the right.

Discussion

In this study, an analysis of the enzyme activity conservation in 174 fungal species revealed that half of these enzyme activities are universally shared among all fungal species, while the remaining half contributes to metabolic diversity. The distribution of enzyme activities across taxonomic classes in fungi allowed us to define these based on their enzyme activities repertoire. Consequently, The enzyme activity repertoire has played a pivotal role in delineating specific taxonomic groups during the course of evolution.

The phylogenetic profiles of enzyme activities were clustered based on their profile similarity, leading to the identification of 14 groups of enzymatic activities sharing similar profiles. However, within these 14 groups, only 39 enzyme activities are represented, indicating that a total of 451 enzyme activities possess distinctive profiles. This finding is surprising, given the diversity of fungi both taxonomically and environmentally. One might anticipate a higher number of modules, yet the presence of a relatively small number of distinct profiles suggests a unique aspect of fungal metabolic evolution. One hypothesis that could account for this result is the potential loss of modularity in the metabolic network from ancestor species to the current ones. Indeed, when comparing the metabolic network in bacteria across different environments to an ancestral metabolic network, previous studies have shown that specialization and adaptation to extreme environments can result in a reduction in modularity (Parter et al., 2007; Takemoto et al., 2007; Kreimer et al., 2008). In other words, the process of specialization for a particular environment has lead to loss of several non-essential enzymatic activities, allowing organisms to better adapt to that specific environment and enhance their environmental specificity.

Through the application of phylostratigraphy techniques, we inferred the evolutionary origin of each enzyme activity, and we found only 8 enzymatic activities specific to fungi. This is

a remarkably low number, considering the extensive diversity and metabolic potential that fungi are known to possess.

We identified 860 enzyme activities through homology outside of fungi, and these enzymatic activities are likely of ancestral origin. Notably, about half of these ancestral enzymes are specific to particular lineages , suggesting that other species may have lost these enzyme activities during the course of evolution. The loss of these enzyme activities likely reflects adaptation and specialization in response to a particular environment. As a result of this specialization, certain enzymatic activities can become dispensable, especially smaller related components that are connected to the rest of the pathway solely through a single enzyme activity (referred to as a connector). The pressure to retain such enzyme activity decreases because regulating an enzymatic reaction requires energy to maintain network equilibrium. The elimination of these enzymatic activities also serves as an essential mechanism for reducing the density of the network, that is, to limit the number of connections between enzymatic reactions. Reducing network density can be a strategy to save energy. Metabolic reactions require energy for their execution, and a more compact network can help minimize the energy costs associated with the synthesis and maintenance of numerous enzymatic proteins. Enzymatic activities occupying alternate positions also provide a way to maintain low network density while preserving network integrity. Enzymatic activities in alternative positions further signify a conservation of metabolites within the network. The introduction of a new metabolite can be dangerous by inhibiting other enzymes (Alam et al., 2017).

The two primary models for the evolution of metabolic pathways are the patchwork evolution model (Jensen, 1976) and the retrograde evolution model (Horowitz, 1945), both of which are mainly based on the concept of duplication-divergence. Consequently, it

appears more parsimonious to generate a new enzyme activity from an existing element. It is noteworthy that Identical enzyme activities from taxonomically unrelated species that are exposed to the same constraint can therefore independently arise from a shared common element. Hence, certain ancestral enzyme activities found exclusively in specific fungi might therefore be derived from an independent functional convergence. However, distinguishing between enzyme activities present in the common ancestor but lost during evolution and enzyme activities that appeared independently in the different species through functional convergences from a common element is a complex task from and not easily discernible solely through phylogenetic profiles.

The analysis of enzyme activities conservation and evolutionary origins within the network shows that the metabolic network exerts constraints on the evolution of enzyme activities. The center of the metabolic network is mainly composed of highly conserved enzyme activities, while the lineage-specific enzyme activities are predominantly located at the network's periphery. In addition, highly conserved enzyme activities exhibit greater connectivity compared to specific enzymatic activities. These findings are consistent with analyses conducted by others (Peregrín-Alvarez et al., 2009), who identified similar characteristics in the global metabolic network shared across all kingdoms of life. We have also demonstrated that the center of this metabolic network is mainly composed of the essential metabolic pathways shared by all species, while the accessory metabolic pathways specific to certain species surround them. This structure can be explained by the fact that the precursors of accessory metabolic pathways are primarily metabolites derived from essential metabolic pathways shared by all species (Keller, 2019).

If the metabolic network imposes constraints on the conservation of enzyme activities, and

many of these activities are shared between species, how can we explain metabolic diversity in fungi ?

First, regulatory systems can play an essential role on organism's phenotype (Wang et al., 2017). Enzymes may also have the same enzyme activities but different reaction rates. These differences in reaction rate between enzymes from different species can have significant physiological consequences. Furthermore, some enzymes are able to recognize multiple substrates. This capacity, known as enzymatic promiscuity, refers to an enzyme's ability to catalyze a reaction other than the one for which it was originally specialized. Enzymes were indeed mostly annotated based on their main activity (Danchin, 2009; Khersonsky and Tawfik, 2010). If most enzymes are annotated according to their primary enzymatic activity, it's likely that current representations of metabolic pathways may therefore simply be the representation of the successions of the primary enzymes activities. This is probably the most essential to ensure the organism's survival and are therefore conserved across many species, extending beyond fungi. All reactions catalyzed by non-specific enzymes constitute what is referred to as the underground metabolism (D'Ari and Casadesús, 1998). The analysis of underground metabolism in E.Coli has revealed that these enzymatic activities from versatile enzymes not only facilitate adaptation to a environment but also serve as predictors of the specific environments in which the species can adapt (Notebaart et al., 2014). This demonstrates that the secondary activity of an enzyme can indeed play an essential role in shaping metabolic and phenotypic diversity. The metabolic network we currently observe might represent only the surface layer, while beneath it lies a reservoir of additional enzyme functions and activities that contribute significantly to the diversity seen in various organisms.

Materials and Methods

Phylogenetic profile construction and visualization

We worked with 174 species of fungi that cover the entire fungal tree (Supp 1) whose genome were fully sequenced and annotated. These data were downloaded from various databases, mostly from the Joint Genome Institute and the Broad Institute. These 174 species of fungi encompassed all fungi with fully sequenced genome as of January 1, 2011.

For functional annotation of each genome's genes, we use MARIO (Pereira *et al.*, 2014) to identify biologically relevant orthologous groups.This approach combines multiple orthologous methods to identify the relevant orthologous group. Subsequently, EC number annotations from homologous sequences in the Uniprot database were transferred to each group of orthologs using the FungiPath pipeline method (Grossetête *et al.*, 2010). This approach enables the annotation of previously unannotated amino acid sequences. Using these annotations and focusing only on enzyme activities, we can construct phylogenetic profiles that indicate the presence or absence of specific EC numbers for each species.

Clustering of similar profiles

Enzyme activities with similar profiles were clustered using the unsupervised non-hierarchical clustering algorithm CLAG (Dib and Carbone, 2012). CLAG relies on two input parameters for the clustering: *threshold* and *delta*. The enzyme activity phylogenetic profile was transformed into a similarity distance matrix. To assess the similarity between two profiles, we use Jaccard index. We selected a *threshold* of 0.75, which minimizes intra-

cluster distance (supp 6). The optimal delta value was determined through visualization of clusters identified, with of 0.1, 0.15, 0.2, and 0.25. A delta of 0.15 yielded the most distinct cluster (Supp 7).

The phylogenetic profile ordered the species based on the inferred phylogenetic tree, which was obtained from the MycoCosm database (Grigoriev *et al.*, 2014). Since the tree only encompasses the primary fungal classes, we further refined the species tree within each class. For each class, we derived a core genome for the species in that class using the Bidirectional Best Hits method (Overbeek *et al.*, 1999). The gene sequences of each core genome were then subsequently aligned and utilized to construct a class-specific tree using IQTree (Nguyen et al., 2015). We use the library "heatmap.2" from the R (version 4.0.2) package to display the clustered phylogenetic profiles.

Phylostratigraphy

Enzyme activity dating is performed using the phylostratigraphy method. Doing so, for a representative of each ortholog group, we conducted a search for homologous sequences outside fungi, against the Non-Redundant database. This search was conducted with DIAMOND (Buchfink *et al.*, 2021), employing an E-values cutoff of 10-3 and a query coverage of 70%.

As a result, we found homologous sequences for all enzyme-associated sequences in fungi, except for enzyme activities: 2.1.1.59, 2.5.1.34, and 4.1.1.36. Indeed, 4.1.1.36 participates in ergot alkaloid biosynthesis and is only found in some fungi (Lee *et al.*, 1976). On the contrary, 2.1.1.59 is a ubiquitous enzymatic reaction identified both in prokaryotes and eukaryotes organisms, but their primary structure differs (Petrényi *et al.*, 2016). 2.1.1.59 is a Cytochrome C methyltransferase only found in certain eukaryotes, including plants and

fungi, but not higher animals (Polevoda *et al.*, 2000). However, no homologous sequences were found outside fungi (especially in the plant), which may suggest that the two sequences from the plant and fungi exhibit a difference in their primary structure.

Some enzyme activities with similar EC numbers (similar family) may share sequence or structural features because they have evolved from a common ancestor (duplication-divergence) with similar catalytic activity. For example, enzymes with the EC number 2.7.1 are kinases that catalyze the transfer of a phosphate group from ATP to a substrate. Many enzyme activities within this class share a conserved domain called the kinase domain, which contains the active site. This domain identifies and classifies kinases based on their sequence and structural similarity (Hanks and Hunter, 1995). Two enzymes with close enzyme activity may share similar sequences. Hence, the presence of homologous sequence outside fungi is not enough for inferring the presence of enzyme activity outside fungi. To determine the evolutionary origin of each enzyme activity, we annotated the 100 first homologous sequences from outside fungi with EC numbers.

Our goal was to determine if any outgroup sequence shares the same EC number as the fungal sequence. Fungal enzyme activities are considered ancestral if any of the homologous sequence enzyme activities match the fungal enzyme activity.

The hundred first hits were annotated by transferring the EC annotation from the best homologous sequences found in Uniprot, based on the highest score and meeting the criteria of an E-value smaller than 10-3 and query coverage of 70%. This allows for a comparison between the original fungal EC number and the hits EC numbers, enabling the inference of the enzyme activity's origin .

Metabolic network construction and graph analyses

The metabolic network is constructed by merging selected metabolic pathways from KEGG databases using a Python script. We select KEGG as our reference database because the KEGG markup language (KGML) provides computational analysis and modeling capabilities. We retrieved 157 metabolic pathways using the KEGG API(released: 5/2021), but selected only 102 metabolic pathways based on their reaction sequences in fungi. We retained metabolic pathways with at least two sequences of enzymatic reaction in fungi. Due to the reuse of enzymes in different pathways, metabolic pathways with at least three enzymatic reaction series are selected in fungi. Metabolic pathways with less than three series of reactions are selected if they contain specific enzymatic activities only in the pathway. Metabolic pathways without specific enzymatic activities are kept if there is evidence of the fungal pathways in literature. 102 metabolic pathways from the KEGG PATHWAY database were selected.

The KGML file is formatted in SIF format for graph analysis, where two nodes (enzymes) are connected if they share a common compound. The topological analysis of the metabolic network, including degree and closeness centrality of all nodes, was assessed using Cytoscape NetworkAnalyzer (version 3.9.1), while the diameter of metabolic pathways was evaluated using the "igraph" library in R.

Metabolic Pathways conservation

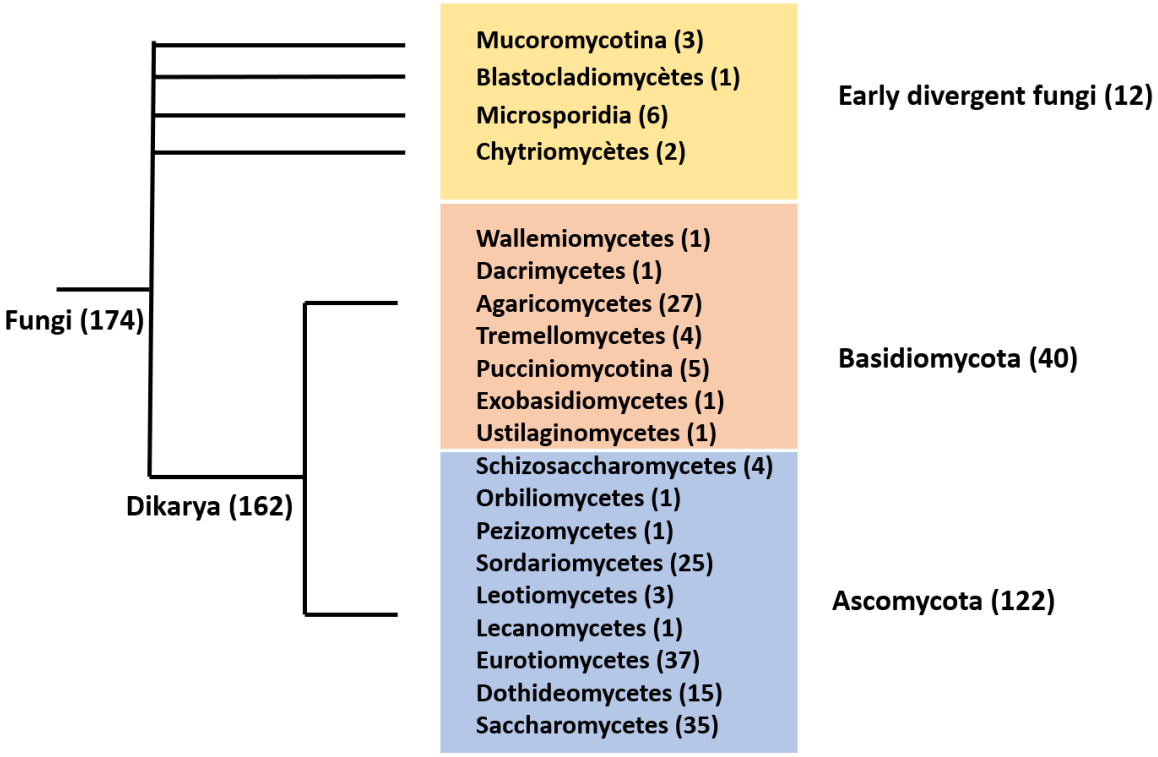
The metabolic network can be divided into metabolic pathways, which are series of enzymes that produce usable materials. With enzyme activities evolutionary information, we were able to distinguish pathways common to all fungal species and those specific to subsets of species. We categorized the 102 metabolic pathways from KEGG into two

categories based on the conservation of their nodes. First, we identified metabolic pathways common to all species when at least two highly conserved (conserver in at least 85% of studied species. Figure 2) enzymes in the pathway are connected (i.e., they share the same compound). These pathways are likely to be ancestral. Second, metabolic pathways not common to all species have all edges that involve at least one non-conserved enzyme, making them specific to a subset of species. We identified 76 metabolic pathways common to all species and 28 that are specific to a subset of species, corresponding to specialized pathways.

Statistical test

We conducted all statistical analyses that aimed to compare median using the Mann-Whitney-Wilcoxon test in R (version 4.0.2). To address the p-value problem inherent in large samples (Lin et al. 2013), we conducted tests iteratively 1000 times on samples of 500 individuals randomly selected from the initial sample when it exceeded 500 individuals. The averaged p-value over the 1000 iterations was subsequently computed.

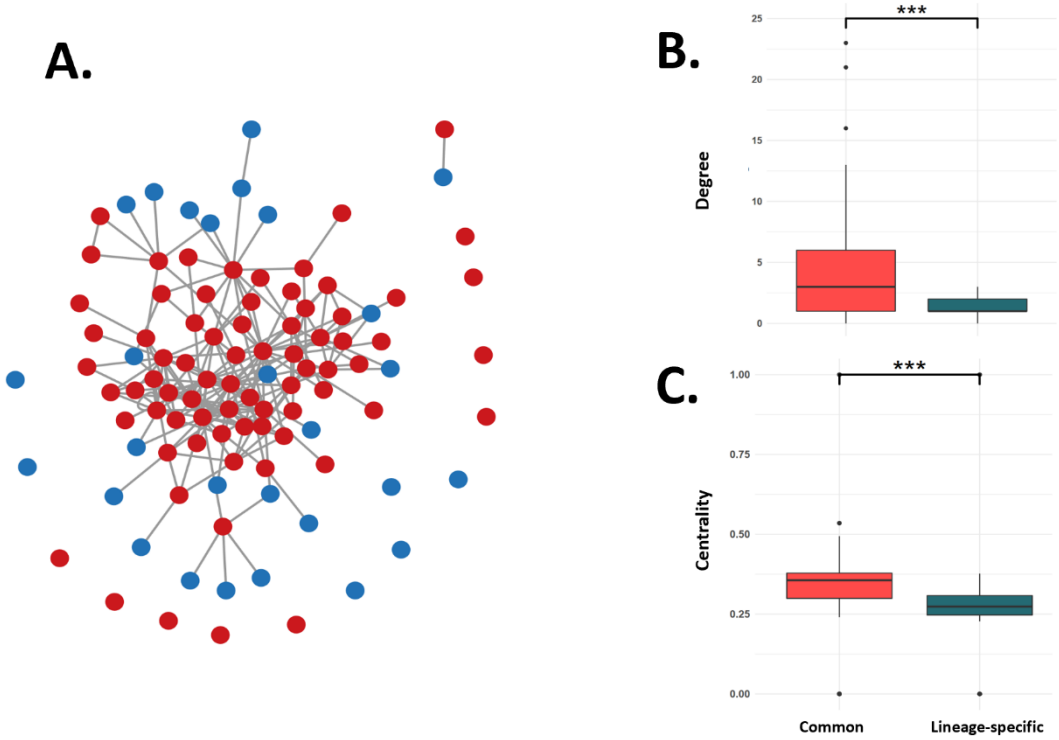
Supplemental Figures



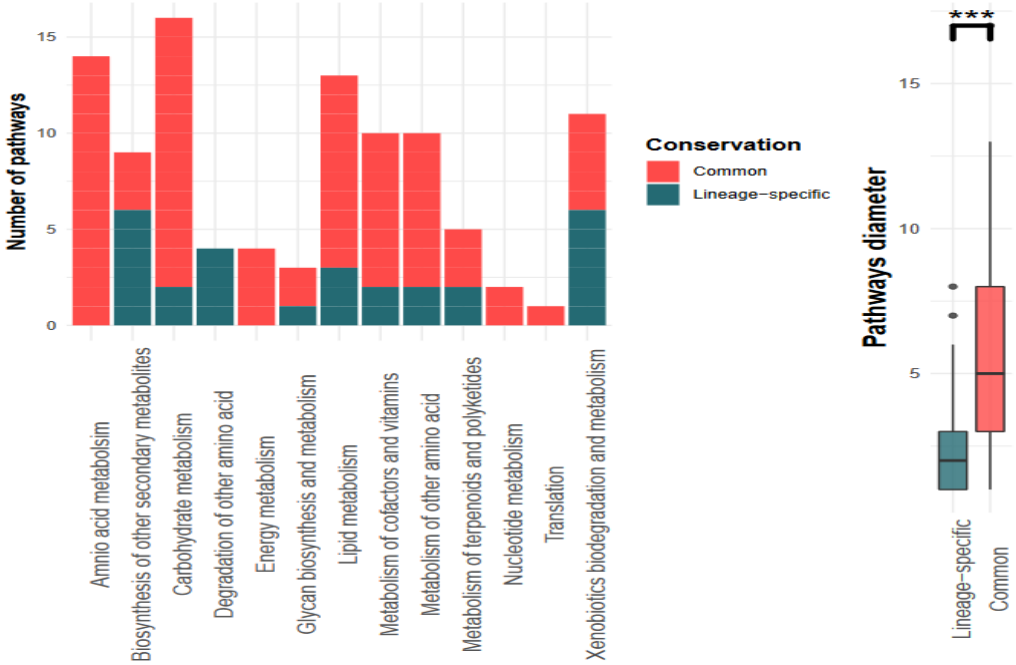
Supplemental Figure 1: **Number of species per taxonomic class**. Ascomycota groups are colored in blue, Basidiomycota in orange, and early divergent fungi are depicted in yellow.

4.2.1.17 ; 1.1.1.157 ; 1.3.8.6 ; 6.2.1.16 ; 4.1.3.4 ; 2.3.3.8 ; 2.7.7.75 ; 2.10.1.1 ; 3.1.3.8 ; 1.2.4.4 ; 2.3.1.168 ; 1.8.3.6 ; 6.4.1.4 ; 1.3.8.4
3.2.1.52 ; 1.4.3.3 ; 1.3.8.1 ; 1.7.3.3 ; 1.14.17.4 ; 3.1.2.22 ; 1.1.1.289 ; 2.7.1.83 ; 3.2.1.39 ; 1.1.1.31
2.5.1.16 ; 2.7.1.105 ; 4.2.1.51 ; 3.1.7.6 ; 2.7.1.36
1.11.1.9 ; 2.1.1.59 ; 1.3.1.94
3.5.2.7 ; 4.2.1.49 ; 1.17.4.2
1.5.1.34 ; 1.14.16.1 ; 1.14.19.3
2.1.1.13 ; 2.7.1.159 ; 2.7.1.134
5.1.99.1 ; 5.4.99.2 ;1.11.1.7
5.1.1.13 ; 3.1.3.25 ; 5.1.1.11
3.5.1.14 ; 4.6.1.2 ; 1.1.1.88
4.2.1.79 ; 2.3.3.5 ; 4.1.3.30
4.1.1.9 ; 1.1.1.146
3.5.99.6 ; 3.5.1.25
1.13.11.20 ; 1.13.11.27

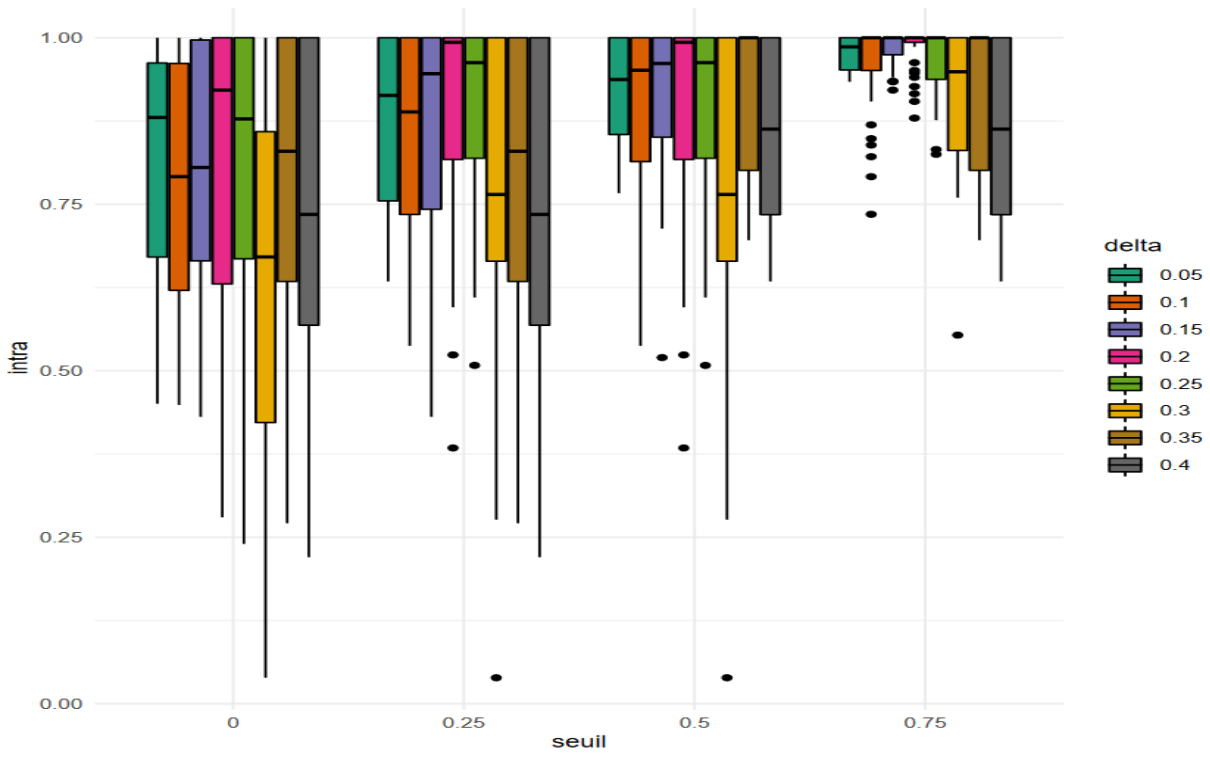
Supplemental Figure 2: List of enzyme activities grouped by similar profile. The order of each line are in the same order as the phylogenetic profile heatmap rows in the Figure 3.



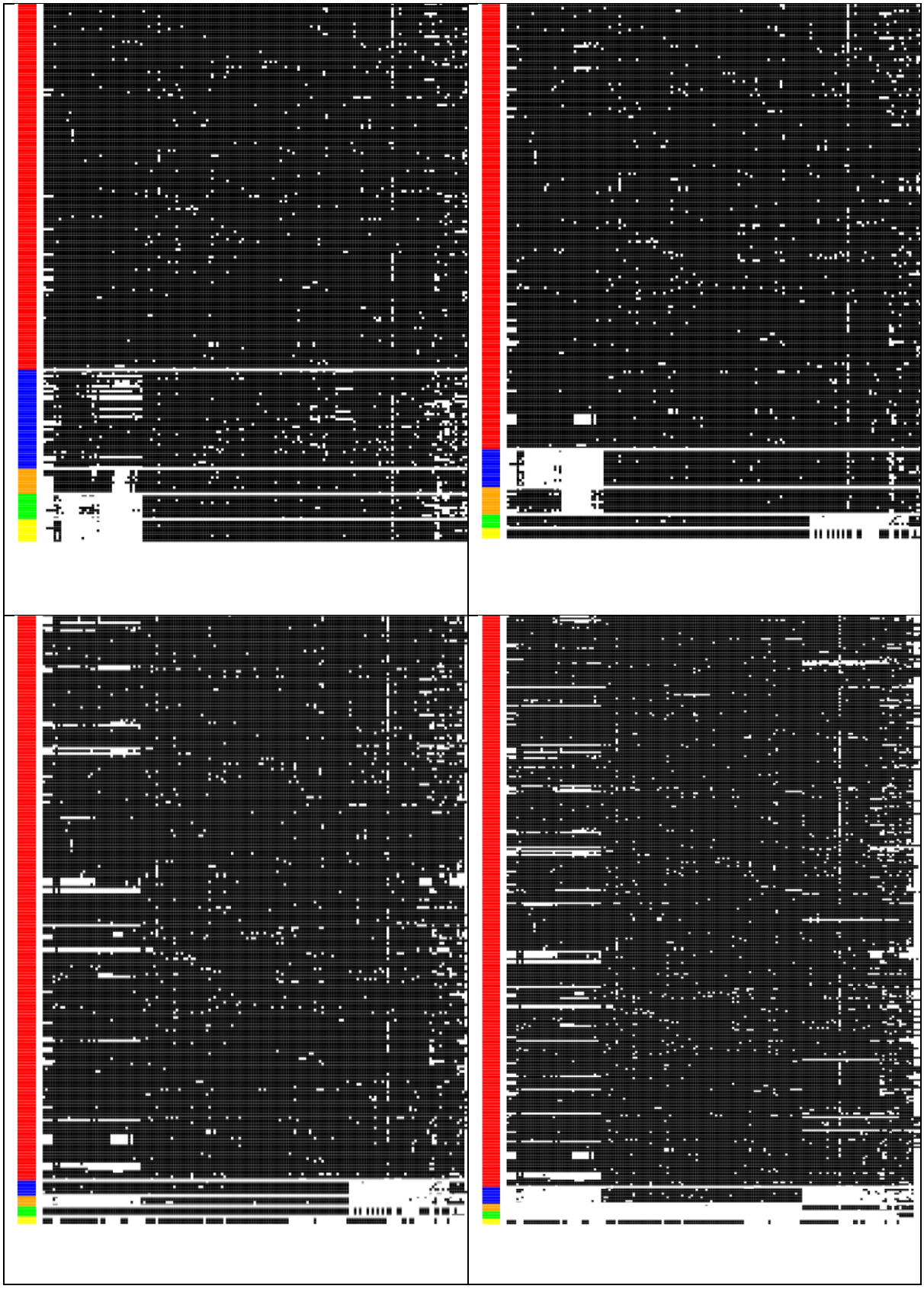
Supplemental Figure 3 : Metabolic pathways conservation within the network. (A) The Metabolic pathways network. Two nodes (pathways) are connected if a metabolite are shared between the two pathways. Node (pathway) color-coded according to their level of conservation**.** Common pathways are colored in red and lineage-specific pathways are colored in blue. **(B)** The closeness centrality boxplot as a function of metabolic pathways conservation**. (C)** The degree boxplot as a function of metabolic pathways conservation.



Supplemental Figure 4: Pathways supergroup conservation and properties. (A) KEGG categorizes pathways in supergroups based on their function. The number of pathways is represented with respect to conservation as a function of pathways supergroup**.** **(B)** Pathways diameter is illustrated in relation to metabolic pathways conservation.



Supplemental Figure 5: Intra-cluster similarities based on CLAG threshold and delta. Each boxplot represents the pairwise similarities of enzyme profiles within the same cluster. We use Jaccard coefficient as a similarity distance metric. The boxplot are color-coded according to the delta values. The x-axis corresponds to the threshold values.



Supplemental Figure 6 : Phylogenetic profile of the 5 largest groups obtained with a

threshold of 0.75 , as a function of delta. From top to bottom and left to right, we have dela values of 0.1, 0.15, 0.2, and 0.25. Each group is represented by the color bar on the left side of the visualization. Rows represent enzymatic activities, and columns represent species are ordered according to the phylogenetic tree. A black cell indicates the presence of enzymatic activity, while a white cell indicates its absence. The blue cluster with delta 0.15 has been merged with the red cluster in delta 0.2 and 0.25. The blue cluster at delta 0.1, without a distinct motif, has been partially merged with the red cluster at delta 0.15 and divided into the orange cluster.

References

Banerjee,A. (2012) Structural distance and evolutionary relationship of networks. *Biosystems*, **107**, 186–196.

Barabási,A.-L. and Albert,R. (1999) Emergence of Scaling in Random Networks. *Science*, **286**, 509–512.

Bennett,J.W. and Klich,M. (2003) Mycotoxins. *Clin Microbiol Rev*, **16**, 497–516.

Bouws,H. *et al.* (2008) Fungal secretomes--nature's toolbox for white biotechnology. *Appl Microbiol Biotechnol*, **80**, 381–388.

Cech,T.R. *et al.* (1981) In vitro splicing of the ribosomal RNA precursor of Tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell*, **27**, 487–496.

Chalancon,G. *et al.* (2013) Metabolic Networks, Reconstruction. In, Dubitzky,W. *et al.* (eds), *Encyclopedia of Systems Biology*. Springer New York, New York, NY, pp. 1259–1263.

Chen,Y.-L. *et al.* (2008) Candida albicans Uses Multiple Mechanisms To Acquire the Essential Metabolite Inositol during Infection. *Infect Immun*, **76**, 2793–2801.

Cooper,T.G. (1982) Nitrogen Metabolism in Saccharomyces cerevisiae. *Cold Spring Harbor Monograph Archive*, **11**, 39–99.

Corradi,N. (2015) Microsporidia: Eukaryotic Intracellular Parasites Shaped by Gene Loss and Horizontal Gene Transfers. *Annu Rev Microbiol*, **69**, 167–183.

Cuomo,C.A. and Birren,B.W. (2010) The fungal genome initiative and lessons learned from genome sequencing. *Methods Enzymol*, **470**, 833–855.

Dib,L. and Carbone,A. (2012) CLAG: an unsupervised non hierarchical clustering algorithm handling biological data. *BMC Bioinformatics*, **13**, 194.

Dutta,B. *et al.* (2022) Fungi in Pharmaceuticals and Production of Antibiotics. In, Shukla,A.C. (ed), *Applied Mycology: Entrepreneurship with Fungi*, Fungal Biology. Springer International Publishing, Cham, pp. 233–257.

El-Gendi,H. *et al.* (2021) A Comprehensive Insight into Fungal Enzymes: Structure, Classification, and Their Role in Mankind's Challenges. *J Fungi (Basel)*, **8**, 23.

Fell,D.A. and Wagner,A. (2000) The small world of metabolism. *Nat Biotechnol*, **18**, 1121–1122.

Fleming,A. (1929) On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to their Use in the Isolation of B. influenzæ. *Br J Exp Pathol*, **10**, 226–236.

Fu,S.-F. *et al.* (2015) Indole-3-acetic acid: A widespread physiological code in interactions of fungi with other organisms. *Plant Signal Behav*, **10**, e1048052.

Grigoriev,I.V. *et al.* (2014) MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Research*, **42**, D699–D704.

Grossetête,S. *et al.* (2010) FUNGIpath: a tool to assess fungal metabolic pathways predicted by orthology. *BMC Genomics*, **11**, 81.

Hartwell,L.H. *et al.* (1999) From molecular to modular cell biology. *Nature*, **402**, C47–C52.

Jeong,H. *et al.* (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, **28**, 27–30.

Karp,P.D. *et al.* (2000) The EcoCyc and MetaCyc databases. *Nucleic Acids Research*, **28**, 56–59.

Kornienko,A. *et al.* (2015) Towards a Cancer Drug of Fungal Origin. *Medicinal research reviews*, **35**, 937.

Kreimer,A. *et al.* (2008) The evolution of modularity in bacterial metabolic networks. *Proceedings of the National Academy of Sciences*, **105**, 6976–6981.

Labena,A.A. *et al.* (2018) Metabolic pathway databases and model repositories. *Quant Biol*, **6**, 30–39.

Li,Z.-W. *et al.* (2016) On the Origin of De Novo Genes in Arabidopsis thaliana Populations. *Genome Biol Evol*, **8**, 2190–2202.

Lima-Mendez,G. and Helden,J. van (2009) The powerful law of the power law and other myths in network biology. *Mol. BioSyst.*, **5**, 1482–1493.

Ma,H. and Zeng,A.-P. (2003) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, **19**, 270–277.

Montanucci,L. *et al.* (2018) Influence of pathway topology and functional class on the molecular evolution of human metabolic genes. *PLOS ONE*, **13**, e0208782.

Naranjo-Ortiz,M.A. and Gabaldón,T. (2019) Fungal evolution: diversity, taxonomy and phylogeny of the Fungi. *Biological Reviews*, **94**, 2101–2137.

Parter,M. *et al.* (2007) Environmental variability and modularity of bacterial metabolic networks. *BMC Evol Biol*, **7**, 169.

Pellegrini,M. *et al.* (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, **96**, 4285–4288.

Peregrín-Alvarez,J.M. *et al.* (2009) The conservation and evolutionary modularity of metabolism. *Genome Biology*, **10**, R63.

Ravasz,E. *et al.* (2002) Hierarchical Organization of Modularity in Metabolic Networks. *Science*, **297**, 1551–1555.

Stolz,J. *et al.* (1999) Identification of the Plasma Membrane H+-Biotin Symporter of Saccharomyces cerevisiae by Rescue of a Fatty Acid-auxotrophic Mutant*. *Journal of Biological Chemistry*, **274**, 18741–18746.

Takemoto,K. *et al.* (2007) Correlation between structure and temperature in prokaryotic metabolic networks. *BMC Bioinformatics*, **8**, 303.

Tipton,K. and Boyce,S. (2000) History of the enzyme nomenclature system. *Bioinformatics*, **16**, 34–40.

Wainwright,M. (1988) Metabolic diversity of fungi in relation to growth and mineral cycling in soil — A review. *Transactions of the British Mycological Society*, **90**, 159–170.

Wang,Y. *et al.* (2022) Comparative genome analysis of plant ascomycete fungal pathogens with different lifestyles reveals distinctive virulence strategies. *BMC Genomics*, **23**, 34.

Wisecaver,J.H. *et al.* (2014) The Evolution of Fungal Metabolic Pathways. *PLoS Genet*, **10**.

Wong,S.S.W. *et al.* (2017) Treatment of Cyclosporin A retains host defense against invasive pulmonary aspergillosis in a non-immunosuppressive murine model by preserving the myeloid cell population. *Virulence*, **8**, 1744–1752.

Yamada,T. *et al.* (2006) Extraction of phylogenetic network modules from the metabolic network. *BMC Bioinformatics*, **7**, 130.

Zhu,D. and Qin,Z.S. (2005) Structural comparison of metabolic networks in selected single cell organisms. *BMC Bioinformatics*, **6**, 8.

VII. Annexes

Amnio acid metabolsim

Arginine biosynthesis*
Purine metabolism*
Alanine, aspartate and glutamate metabolism*
Glycine, serine and threonine metabolism*
Cysteine and methionine metabolism*
Valine, leucine and isoleucine degradation*
Valine, leucine and isoleucine biosynthesis*
Lysine biosynthesis*
Lysine degradation*
Arginine and proline metabolism*
Histidine metabolism*
Tyrosine metabolism*
Phenylalanine metabolism*
Tryptophan metabolism*
Phenylalanine, tyrosine and tryptophan biosynthesis*

Biosynthesis of other secondary metabolites

Caffeine metabolism*
Aflatoxin biosynthesis*
Penicillin and cephalosporin biosynthesis*
Neomycin, kanamycin and gentamicin biosynthesis*
Acarbose and validamycin biosynthesis*
Benzoxazinoid biosynthesis*
Staurosporine biosynthesis*
Phenazine biosynthesis*
Novobiocin biosynthesis*
Carbapenem biosynthesis*
Prodigiosin biosynthesis*
Streptomycin biosynthesis*
Clavulanic acid biosynthesis*
Glycine, serine and threonine metabolism*
Monobactam biosynthesis*
Tropane, piperidine and pyridine alkaloid biosynthesis*
Glucosinolate biosynthesis*
Biosynthesis of various secondary metabolites - part 3*
Biosynthesis of various secondary metabolites - part 2*
Biosynthesis of various secondary metabolites - part 1*
Indole alkaloid biosynthesis*
Phenylpropanoid biosynthesis*
Flavonoid biosynthesis*
Anthocyanin biosynthesis*
Isoflavonoid biosynthesis*
Flavone and flavonol biosynthesis*
Stilbenoid, diarylheptanoid and gingerol biosynthesis*
Isoquinoline alkaloid biosynthesis*
Tropane, piperidine and pyridine alkaloid biosynthesis*
Betalain biosynthesis*
Glucosinolate biosynthesis*

Carbohydrate metabolism

Glycolysis / Gluconeogenesis*
Citrate cycle (TCA cycle) *
Pentose phosphate pathway*
Fructose and mannose metabolism*
Galactose metabolism*
Ascorbate and aldarate metabolism
Synthesis and degradation of ketone bodies*
Starch and sucrose metabolism*
Amino sugar and nucleotide sugar metabolism*
Inositol phosphate metabolism*
Pyruvate metabolism*
Glyoxylate and dicarboxylate metabolism*
Propanoate metabolism*

Butanoate metabolism*
C5-Branched dibasic acid metabolism*

Degradation of other amino acid
Atrazine degradation*
Limonene and pinene degradation*
Caprolactam degradation*

Energy metabolism
Oxidative phosphorylation*
Photosynthesis*
Methane metabolism*
Carbon fixation in photosynthetic organisms*
Carbon fixation pathways in prokaryotes*
Nitrogen metabolism*
Sulfur metabolism*

Glycan biosynthesis and metabolism
N-Glycan biosynthesis*
Glycosylphosphatidylinositol (GPI)-anchor biosynthesis*
Mucin type O-glycan biosynthesis*
Lipopolysaccharide biosynthesis*
O-Antigen nucleotide sugar biosynthesis*
Lipoarabinomannan (LAM) biosynthesis*
Arabinogalactan biosynthesis - Mycobacterium*
Other glycan degradation*
Various type of N-Glycan biosynthesis*
Other types O-glycan biosynthesis*
Mannose type O-glycan biosynthesis*
Glycosaminoglycan degradation*
Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate*
Glycosaminoglycan biosynthesis - keratan sulfate*
Glycosaminoglycan biosynthesis - heparan sulfate / heparin*
Peptidoglycan biosynthesis*
Lipoarabinomannan (LAM) biosynthesis*
Arabinogalactan biosynthesis – Mycobacterium*
Glycosphingolipid biosynthesis - lacto and neolacto series*
Glycosphingolipid biosynthesis - globo and isoglobo series*
Glycosphingolipid biosynthesis - ganglio series*

Lipid metabolism
Fatty acid biosynthesis*
Fatty acid elongation*
Fatty acid degradation*
Cutin, suberine and wax biosynthesis
Steroid biosynthesis*
Primary bile acid biosynthesis*
Steroid hormone biosynthesis*
Glycerolipid metabolism*
Glycerophospholipid metabolism*
Ether lipid metabolism*
Arachidonic acid metabolism*
Linoleic acid metabolism*
alpha-Linolenic acid metabolism*
Secondary bile acid biosynthesis*

Metabolism of cofactors and vitamins
Ubiquinone and other terpenoid-quinone biosynthesis*
Retinol metabolism*
One carbon pool by folate*
Thiamine metabolism*
Riboflavin metabolism*
Vitamin B6 metabolism*
Nicotinate and nicotinamide metabolism*
Pantothenate and CoA biosynthesis*
Biotin metabolism*
Lipoic acid metabolism
Folate biosynthesis*

Metabolism of other amino acid
Taurine and hypotaurine metabolism*
beta-Alanine metabolism*
Phosphonate and phosphinate metabolism*
Selenocompound metabolism*
Cyanoamino acid metabolism*
D-Glutamine and D-glutamate metabolism
Sphingolipid metabolism*
D-Arginine and D-ornithine metabolism*
D-Alanine metabolism*
Glutathione metabolism*
Porphyrin and chlorophyll metabolism*

Metabolism of terpenoids and polyketides
Geraniol degradation*
Biosynthesis of 12-, 14- and 16-membered macrolides*
Polyketide sugar unit biosynthesis*
Tetracycline metabolism*
Terpenoid backbone biosynthesis*
Monoterpenoid biosynthesis*
Diterpenoid biosynthesis*
Brassinosteroid biosynthesis*
Carotenoid biosynthesis*
Zeatin biosynthesis*
Sesquiterpenoid and triterpenoid biosynthesis*
Insect hormone biosynthesis*

Nucleotide metabolism
Pyrimidine metabolism*
Purin metabolism*

Translation
Aminoacyl-tRNA biosynthesis*

Xenobiotics biodegradation and metabolism
Chlorocyclohexane and chlorobenzene degradation*
Benzoate degradation*
Bisphenol degradation*
Fluorobenzoate degradation
Furfural degradation*
Dioxin degradation*
Xylene degradation*
Toluene degradation
Polycyclic aromatic hydrocarbon degradation*
Chloroalkane and chloroalkene degradation*
Naphthalene degradation*
Aminobenzoate degradation*
Nitrotoluene degradation*
Ethylbenzene degradation*
Styrene degradation*
Metabolism of xenobiotics by cytochrome P450*
Steroid degradation*
Drug metabolism - cytochrome P450*
Drug metabolism - other enzymes*
Metabolism of xenobiotics by cytochrome P450*

Annexe 1 : Liste des voies de KEGG par super groupe (Version 2021). Les voies communes à toutes les espèces sont précédées d'un astérisque rouge *. Les voies spécifiques de certaines espèces sont précédées d'un astérisque bleu*. Les voies non exploitables de KEGG sont précédées d'un astérisque noir *. Et les voies inférées comme absente des champignons sont précédées d'un astérisque vert*

1.10.3.3; 1.1.1.1; 1.1.1.10; 1.1.1.100; 1.1.1.101; 1.1.1.102; 1.1.1.103; 1.1.1.105*; 1.1.1.12; 1.11.1.21; 1.1.1.136; 1.1.1.144; 1.1.1.146; 1.1.1.153; 1.1.1.156; 1.1.1.157; 1.11.1.6; 1.1.1.169; 1.1.1.17; 1.11.1.7; 1.1.1.170; 1.1.1.178; 1.1.1.179; 1.1.1.18; 1.1.1.184; 1.1.1.188; 1.1.1.189; 1.11.1.9; 1.1.1.2; 1.1.1.205; 1.1.1.208*; 1.1.1.21; 1.1.1.22; 1.1.1.23; 1.11.2.3; 1.1.1.24; 1.1.1.243; 1.1.1.25; 1.1.1.26; 1.1.1.267; 1.1.1.27; 1.1.1.270; 1.1.1.271; 1.1.1.283; 1.1.1.284; 1.1.1.289; 1.1.1.29; 1.1.1.3; 1.1.1.302; 1.1.1.305; 1.1.1.307; 1.1.1.31; 1.1.1.313; 1.1.1.318; 1.1.1.330; 1.1.1.34; 1.1.1.349; 1.1.1.35; 1.1.1.352; 1.1.1.353; 1.1.1.37; 1.1.1.38; 1.1.1.39; 1.1.1.40; 1.1.1.41; 1.1.1.42; 1.1.1.44; 1.1.1.49; 1.1.1.51; 1.1.1.6; 1.1.1.62; 1.1.1.65; 1.1.1.67; 1.1.1.8; 1.1.1.81; 1.1.1.83; 1.1.1.85; 1.1.1.86; 1.1.1.87; 1.1.1.88; 1.1.1.9; 1.1.1.93; 1.1.1.95; 1.1.2.3; 1.1.2.4; 1.13.11.1; 1.13.11.11; 1.13.11.20; 1.13.11.27; 1.13.11.29; 1.13.11.37; 1.13.11.39; 1.13.11.4; 1.13.11.40; 1.13.11.5; 1.13.11.51; 1.13.11.52; 1.13.11.54; 1.13.11.58; 1.13.11.6; 1.13.11.60; 1.13.11.62; 1.13.12.16; 1.13.12.20; 1.13.12.3; 1.13.12.4; 1.1.3.13; 1.1.3.15; 1.1.3.17; 1.1.3.38; 1.1.3.4; 1.1.3.5; 1.1.3.9; 1.13.99.1; 1.14.11.1; 1.14.11.17; 1.14.11.37; 1.14.11.38*; 1.14.11.8; 1.14.13.1*; 1.14.13.101*; 1.14.13.113; 1.14.13.114; 1.14.13.127; 1.14.13.131*; 1.14.13.148; 1.14.13.168; 1.14.13.171*; 1.14.13.178; 1.14.13.22; 1.14.13.50; 1.14.13.59*; 1.14.13.63; 1.14.13.7; 1.14.13.8; 1.14.13.82; 1.14.13.84; 1.14.13.9; 1.14.14.1; 1.14.14.5*; 1.14.15.7; 1.14.16.1; 1.14.17.4; 1.14.18.1; 1.14.19.1**; 1.14.19.3; 1.14.19.6**; 1.14.99.46; 1.1.5.3; 1.1.5.5; 1.1.5.9; 1.16.3.1; 1.17.1.4; 1.17.4.1; 1.17.4.2; 1.1.99.1; 1.1.99.2; 1.1.99.31; 1.2.1.11; 1.2.1.12; 1.2.1.24; 1.2.1.3; 1.2.1.31; 1.21.3.1; 1.21.3.7*; 1.2.1.38; 1.2.1.41; 1.2.1.44; 1.2.1.5; 1.2.1.75; 1.2.1.84; 1.2.1.88; 1.2.4.1; 1.2.4.2; 1.2.4.4; 1.3.1.10; 1.3.1.100; 1.3.1.13; 1.3.1.20; 1.3.1.21; 1.3.1.22; 1.3.1.3; 1.3.1.32; 1.3.1.38; 1.3.1.42; 1.3.1.6; 1.3.1.70; 1.3.1.71; 1.3.1.76; 1.3.1.9; 1.3.1.94; 1.3.1.98; 1.3.3.3; 1.3.3.4; 1.3.3.5; 1.3.3.6; 1.3.5.1; 1.3.5.2; 1.3.8.1; 1.3.8.4; 1.3.8.6; 1.3.8.8; 1.3.98.1; 1.3.99.23*; 1.3.99.30; 1.3.99.4; 1.3.99.5; 1.4.1.13; 1.4.1.2; 1.4.1.3; 1.4.1.4; 1.4.3.2; 1.4.3.21; 1.4.3.22; 1.4.3.3; 1.4.3.4; 1.4.3.5; 1.4.4.2; 1.5.1.10; 1.5.1.11; 1.5.1.15; 1.5.1.16; 1.5.1.2; 1.5.1.20; 1.5.1.3; 1.5.1.34; 1.5.1.38; 1.5.1.39; 1.5.1.44; 1.5.1.46; 1.5.1.5; 1.5.1.7; 1.5.1.8; 1.5.1.9; 1.5.3.1; 1.5.3.14; 1.5.3.16; 1.5.3.17; 1.5.3.19; 1.5.3.6; 1.5.3.7; 1.6.1.2; 1.6.2.2; 1.6.5.2; 1.6.5.4; 1.6.99.3**; 1.7.1.1**; 1.7.1.14; 1.7.1.3**; 1.7.1.4; 1.7.2.1; 1.7.3.1; 1.7.3.3; 1.8.1.2*; 1.8.1.4; 1.8.1.7; 1.8.1.9; 1.8.3.1*; 1.8.3.6; 1.8.4.14; 1.8.4.8*; 1.8.5.1; 2.10.1.1; 2.1.1.1; 2.1.1.10; 2.1.1.100; 2.1.1.107; 2.1.1.109; 2.1.1.110; 2.1.1.114; 2.1.1.121; 2.1.1.13; 2.1.1.14; 2.1.1.140; 2.1.1.17; 2.1.1.201; 2.1.1.222; 2.1.1.254*; 2.1.1.261; 2.1.1.37; 2.1.1.41; 2.1.1.45; 2.1.1.59; 2.1.1.6; 2.1.1.64; 2.1.1.71; 2.1.2.1; 2.1.2.10; 2.1.2.11; 2.1.2.13; 2.1.2.2; 2.1.2.3; 2.1.2.9; 2.1.3.2; 2.1.3.3; 2.1.4.1; 2.2.1.1; 2.2.1.2; 2.2.1.3; 2.3.1.1; 2.3.1.102*; 2.3.1.12; 2.3.1.129*; 2.3.1.133; 2.3.1.15; 2.3.1.158; 2.3.1.16; 2.3.1.161*; 2.3.1.164; 2.3.1.165**; 2.3.1.168; 2.3.1.176; 2.3.1.181; 2.3.1.199; 2.3.1.20; 2.3.1.205; 2.3.1.22; 2.3.1.221; 2.3.1.23; 2.3.1.24; 2.3.1.26; 2.3.1.31; 2.3.1.32; 2.3.1.35; 2.3.1.36; 2.3.1.37; 2.3.1.39; 2.3.1.4; 2.3.1.41; 2.3.1.42; 2.3.1.47; 2.3.1.5; 2.3.1.50; 2.3.1.51; 2.3.1.61; 2.3.1.74; 2.3.1.85; 2.3.1.86; 2.3.1.9; 2.3.1.94*; 2.3.2.2; 2.3.3.1; 2.3.3.10; 2.3.3.13; 2.3.3.14; 2.3.3.5; 2.3.3.8; 2.3.3.9; 2.4.1.1; 2.4.1.109**; 2.4.1.11; 2.4.1.117; 2.4.1.12; 2.4.1.123; 2.4.1.131; 2.4.1.132; 2.4.1.134**; 2.4.1.135**; 2.4.1.141; 2.4.1.142; 2.4.1.149**; 2.4.1.15; 2.4.1.16; 2.4.1.17; 2.4.1.18; 2.4.1.198; 2.4.1.214**; 2.4.1.217; 2.4.1.223**; 2.4.1.232**; 2.4.1.25; 2.4.1.255**; 2.4.1.256; 2.4.1.257; 2.4.1.258; 2.4.1.259; 2.4.1.260; 2.4.1.261; 2.4.1.265; 2.4.1.267; 2.4.1.280; 2.4.1.34; 2.4.1.80; 2.4.1.82; 2.4.1.83; 2.4.1.91*; 2.4.2.1; 2.4.2.10; 2.4.2.12; 2.4.2.14; 2.4.2.17; 2.4.2.18; 2.4.2.19; 2.4.2.28; 2.4.2.7; 2.4.2.8; 2.4.2.9; 2.4.99.18; 2.5.1.1; 2.5.1.10; 2.5.1.15; 2.5.1.16; 2.5.1.17;

2.5.1.18; 2.5.1.19; 2.5.1.21; 2.5.1.29; 2.5.1.3; 2.5.1.31; 2.5.1.32; 2.5.1.34; 2.5.1.39; 2.5.1.47; 2.5.1.48; 2.5.1.49; 2.5.1.54; 2.5.1.58; 2.5.1.6; 2.5.1.61; 2.5.1.7; 2.5.1.75*; 2.5.1.78; 2.5.1.83; 2.5.1.87; 2.5.1.9; 2.5.1.91; 2.5.1.93; 2.6.1.1; 2.6.1.11; 2.6.1.13; 2.6.1.16; 2.6.1.19; 2.6.1.2; 2.6.1.39; 2.6.1.42; 2.6.1.44; 2.6.1.48; 2.6.1.5; 2.6.1.52; 2.6.1.57; 2.6.1.62; 2.6.1.7; 2.6.1.85; 2.6.1.9; 2.6.1.96; 2.7.1.1; 2.7.1.105; 2.7.1.108; 2.7.1.11; 2.7.1.12; 2.7.1.127; 2.7.1.134; 2.7.1.137; 2.7.1.15; 2.7.1.150; 2.7.1.151; 2.7.1.158; 2.7.1.159; 2.7.1.16; 2.7.1.167*; 2.7.1.17; 2.7.1.173; 2.7.1.174; 2.7.1.2; 2.7.1.20; 2.7.1.21; 2.7.1.22; 2.7.1.23; 2.7.1.24; 2.7.1.25; 2.7.1.26; 2.7.1.28; 2.7.1.29; 2.7.1.3; 2.7.1.30; 2.7.1.31; 2.7.1.32; 2.7.1.33; 2.7.1.35; 2.7.1.36; 2.7.1.39; 2.7.1.40; 2.7.1.43; 2.7.1.46; 2.7.1.48; 2.7.1.49; 2.7.1.50; 2.7.1.59; 2.7.1.6; 2.7.1.67; 2.7.1.68; 2.7.1.71; 2.7.1.82; 2.7.1.83; 2.7.1.91; 2.7.2.1; 2.7.2.11; 2.7.2.3; 2.7.2.4; 2.7.2.8; 2.7.4.12; 2.7.4.14; 2.7.4.2; 2.7.4.22; 2.7.4.25; 2.7.4.26; 2.7.4.3; 2.7.4.6; 2.7.4.7; 2.7.4.8; 2.7.4.9; 2.7.6.1; 2.7.6.2; 2.7.6.3; 2.7.7.1; 2.7.7.12; 2.7.7.13; 2.7.7.14; 2.7.7.15; 2.7.7.2; 2.7.7.23; 2.7.7.3; 2.7.7.4; 2.7.7.41; 2.7.7.53; 2.7.7.70*; 2.7.7.75; 2.7.7.9; 2.7.8.1; 2.7.8.11; 2.7.8.15; 2.7.8.2; 2.7.8.23; 2.7.8.29; 2.7.8.5; 2.7.8.7; 2.7.8.8; 2.7.9.2; 2.8.1.1*; 2.8.1.12; 2.8.1.3; 2.8.1.6; 2.8.1.7; 2.8.1.8; 2.8.1.9; 2.8.3.5; 3.10.1.1**; 3.1.1.1; 3.1.1.11; 3.11.1.2; 3.1.1.13; 3.1.1.17; 3.1.1.2; 3.1.1.23; 3.1.1.24; 3.1.1.3; 3.1.1.31; 3.1.1.4; 3.1.1.45; 3.1.1.47; 3.1.1.5; 3.1.1.57; 3.1.1.65; 3.1.1.7; 3.1.1.83; 3.1.1.84*; 3.1.1.94; 3.1.2.1; 3.1.2.12; 3.1.2.14; 3.1.2.2; 3.1.2.22; 3.1.2.27; 3.1.2.4; 3.1.2.6; 3.1.3.1; 3.1.3.11; 3.1.3.12; 3.1.3.15; 3.1.3.18; 3.1.3.2; 3.1.3.21; 3.1.3.25; 3.1.3.27; 3.1.3.3; 3.1.3.36; 3.1.3.37; 3.1.3.4; 3.1.3.46; 3.1.3.5; 3.1.3.56; 3.1.3.6; 3.1.3.64; 3.1.3.67; 3.1.3.7*; 3.1.3.74; 3.1.3.77; 3.1.3.8; 3.1.3.81; 3.1.3.86; 3.1.4.11; 3.1.4.12; 3.1.4.16; 3.1.4.17; 3.1.4.2; 3.1.4.3; 3.1.4.35**; 3.1.4.4; 3.1.4.46; 3.1.4.50; 3.1.4.53; 3.1.6.1; 3.1.7.2; 3.1.7.6; 3.1.8.1; 3.2.1.1; 3.2.1.10; 3.2.1.106; 3.2.1.108; 3.2.1.11; 3.2.1.113; 3.2.1.132; 3.2.1.14; 3.2.1.15; 3.2.1.165; 3.2.1.18; 3.2.1.20; 3.2.1.21; 3.2.1.22; 3.2.1.23; 3.2.1.24**; 3.2.1.25**; 3.2.1.26; 3.2.1.28; 3.2.1.3; 3.2.1.31; 3.2.1.37; 3.2.1.39; 3.2.1.4; 3.2.1.50**; 3.2.1.51**; 3.2.1.52; 3.2.1.55; 3.2.1.58; 3.2.1.64; 3.2.1.67; 3.2.1.78; 3.2.1.80; 3.2.1.91; 3.2.1.96**; 3.2.2.3; 3.3.1.1; 3.3.2.10; 3.3.2.6; 3.4.11.1; 3.4.11.5; 3.4.13.18; 3.4.16.4**; 3.4.17.21; 3.4.19.13; 3.4.24.84; 3.5.1.1; 3.5.1.10; 3.5.1.107; 3.5.1.110; 3.5.1.14; 3.5.1.18; 3.5.1.19; 3.5.1.2; 3.5.1.23; 3.5.1.25; 3.5.1.3; 3.5.1.32; 3.5.1.4; 3.5.1.41; 3.5.1.46; 3.5.1.49; 3.5.1.5; 3.5.1.54; 3.5.1.56; 3.5.1.89; 3.5.1.9; 3.5.2.15; 3.5.2.17; 3.5.2.2; 3.5.2.3; 3.5.2.5; 3.5.2.6; 3.5.2.7; 3.5.2.9; 3.5.3.1; 3.5.3.11; 3.5.3.12; 3.5.3.4; 3.5.4.1; 3.5.4.10; 3.5.4.12; 3.5.4.16; 3.5.4.19; 3.5.4.2; 3.5.4.25; 3.5.4.26; 3.5.4.3; 3.5.4.31; 3.5.4.4; 3.5.4.5; 3.5.4.6; 3.5.4.9; 3.5.5.1; 3.5.5.4; 3.5.5.7; 3.5.99.2; 3.5.99.6; 3.5.99.7; 3.6.1.1**; 3.6.1.11; 3.6.1.13; 3.6.1.15; 3.6.1.17; 3.6.1.21; 3.6.1.22; 3.6.1.23; 3.6.1.29; 3.6.1.3**; 3.6.1.31; 3.6.1.43; 3.6.1.5; 3.6.1.9; 3.7.1.2; 3.7.1.3; 3.8.1.3; 3.8.1.5; 4.1.1.1; 4.1.1.15; 4.1.1.17; 4.1.1.2; 4.1.1.21; 4.1.1.23; 4.1.1.25; 4.1.1.28; 4.1.1.33; 4.1.1.36; 4.1.1.37; 4.1.1.45; 4.1.1.46; 4.1.1.48; 4.1.1.49; 4.1.1.5; 4.1.1.50; 4.1.1.53; 4.1.1.55*; 4.1.1.6; 4.1.1.61; 4.1.1.65; 4.1.1.68; 4.1.1.8; 4.1.1.82; 4.1.1.86; 4.1.1.9; 4.1.2.13; 4.1.2.25; 4.1.2.27; 4.1.2.4; 4.1.2.48; 4.1.2.49; 4.1.2.53; 4.1.2.9; 4.1.3.1; 4.1.3.16; 4.1.3.27; 4.1.3.30; 4.1.3.34; 4.1.3.38; 4.1.3.4; 4.1.99.1; 4.1.99.12; 4.1.99.16; 4.2.1.1; 4.2.1.10; 4.2.1.103; 4.2.1.104; 4.2.1.109; 4.2.1.11; 4.2.1.116; 4.2.1.118**; 4.2.1.119**; 4.2.1.126; 4.2.1.130; 4.2.1.132*; 4.2.1.134; 4.2.1.142; 4.2.1.143; 4.2.1.17; 4.2.1.18; 4.2.1.19; 4.2.1.2; 4.2.1.20; 4.2.1.22; 4.2.1.24; 4.2.1.25; 4.2.1.3; 4.2.1.33; 4.2.1.36; 4.2.1.40; 4.2.1.47; 4.2.1.49; 4.2.1.51; 4.2.1.55; 4.2.1.59; 4.2.1.6; 4.2.1.66; 4.2.1.68; 4.2.1.69; 4.2.1.75; 4.2.1.76; 4.2.1.77; 4.2.1.79; 4.2.1.9; 4.2.1.90; 4.2.1.91; 4.2.1.92; 4.2.1.96; 4.2.2.2; 4.2.2.3; 4.2.3.1; 4.2.3.12; 4.2.3.22; 4.2.3.23; 4.2.3.4; 4.2.3.43; 4.2.3.5; 4.2.3.6; 4.2.3.75; 4.2.3.9; 4.3.1.17;

4.3.1.18; 4.3.1.19; 4.3.1.2; 4.3.1.24; 4.3.2.1; 4.3.2.2; 4.3.2.3; 4.3.3.5*; 4.3.3.6; 4.4.1.1; 4.4.1.14; 4.4.1.17; 4.4.1.22; 4.4.1.3*; 4.4.1.5; 4.6.1.1; 4.6.1.13; 4.6.1.2; 4.99.1.1; 4.99.1.4; 4.99.1.7; 5.1.1.1; 5.1.1.11; 5.1.1.13; 5.1.1.17; 5.1.1.18; 5.1.1.3; 5.1.1.7; 5.1.3.1; 5.1.3.15; 5.1.3.18; 5.1.3.2; 5.1.3.3; 5.1.99.1; 5.2.1.2; 5.3.1.1; 5.3.1.13*; 5.3.1.16; 5.3.1.23; 5.3.1.24; 5.3.1.6; 5.3.1.8; 5.3.1.9; 5.3.3.10; 5.3.3.17*; 5.3.3.2; 5.3.3.5; 5.3.3.6; 5.3.3.8; 5.4.2.11; 5.4.2.12; 5.4.2.2; 5.4.2.3; 5.4.2.8; 5.4.2.9; 5.4.3.2; 5.4.3.8; 5.4.4.5; 5.4.99.17; 5.4.99.2; 5.4.99.32; 5.4.99.5; 5.4.99.7; 5.5.1.12; 5.5.1.19; 5.5.1.2; 5.5.1.4; 5.5.1.5; 5.5.1.9; 6.1.1.1; 6.1.1.10; 6.1.1.11; 6.1.1.12; 6.1.1.14; 6.1.1.15; 6.1.1.16; 6.1.1.17; 6.1.1.18; 6.1.1.19; 6.1.1.2; 6.1.1.20; 6.1.1.21; 6.1.1.22; 6.1.1.3; 6.1.1.4; 6.1.1.5; 6.1.1.6; 6.1.1.7; 6.1.1.9; 6.2.1.1; 6.2.1.12; 6.2.1.16; 6.2.1.2; 6.2.1.3; 6.2.1.4; 6.2.1.5; 6.3.1.1; 6.3.1.2; 6.3.2.1; 6.3.2.12; 6.3.2.17; 6.3.2.2; 6.3.2.26; 6.3.2.3; 6.3.2.4; 6.3.2.5; 6.3.2.6; 6.3.2.8; 6.3.3.1; 6.3.3.2; 6.3.3.3; 6.3.4.10; 6.3.4.11; 6.3.4.13; 6.3.4.15; 6.3.4.2; 6.3.4.21; 6.3.4.3; 6.3.4.4; 6.3.4.5; 6.3.4.6; 6.3.4.9; 6.3.5.1; 6.3.5.2; 6.3.5.3; 6.3.5.4; 6.3.5.5; 6.3.5.6; 6.3.5.7; 6.4.1.1; 6.4.1.2; 6.4.1.3; 6.4.1.4

Annexe 2 : Liste des activités enzymatiques identifiées chez les 174 espèces de champignons. Les EC-number suivit d'un seul * sont des activités enzymatiques contenus dans les voies non exploitables de KEGG. Les EC-numbers suivit d'un double ** sont contenus dans les voies qui ont été inférés comme absentes chez les champignons. La conservation de chaque activité enzymatique est disponible sur http://fungipath.i2bc.paris-saclay.fr/.

	Anciennes conservées	Anciennes spécifiques	Nouvelles spécifiques
Anciennes conservées		4e-4	5.6e-46
Anciennes spécifiques	4e-4		2.4e-21
Nouvelles spécifiques	5.6e-46	2.4e-21	

Annexe 3 : p-value du test de Mann-Whitney-Wilcoxon de la distribution des degrés des sommets du réseau métabolique en fonction de l'origine évolutive de la Figure 12.2.

	Anciennes conservées	Anciennes spécifiques	Nouvelles Spécifiques
Anciennes conservées		5.3e-5	6.3e-15
Anciennes spécifiques	5.3e-5		8.16e-29
Nouvelles spécifiques	6.3e-15	8.16e-29	

Annexe 4 : les p-value du test de Mann-Whitney-Wilcoxon de la distribution des valeurs de centralités des sommets du réseau métabolique en fonction de l'origine évolutive de la Figure 12.2.

VIII. Bibliographie

Alam,M.T. *et al.* (2017) The self-inhibitory nature of metabolic networks and its alleviation through compartmentalization. *Nat Commun*, **8**, 16018.

Albalat,R. and Cañestro,C. (2016) Evolution by gene loss. *Nat Rev Genet*, **17**, 379–391.

Alexander,W.G. *et al.* (2016) Horizontally acquired genes in early-diverging pathogenic fungi enable the use of host nucleosides and nucleotides. *Proceedings of the National Academy of Sciences*, **113**, 4116–4121.

Altman,T. *et al.* (2013) A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics*, **14**, 112.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Alves,R. *et al.* (2002) Evolution of enzymes in metabolism: a network perspective. *J Mol Biol*, **320**, 751–770.

Aravind,L. *et al.* (2000) Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc Natl Acad Sci U S A*, **97**, 11319–11324.

Arita,M. (2005) Scale-freeness and biological networks. *J Biochem*, **138**, 1–4.

Arita,M. (2004) The metabolic world of Escherichia coli is not small. *Proceedings of the National Academy of Sciences*, **101**, 1543–1547.

Ayuso-Fernández,I. *et al.* (2018) Evolutionary convergence in lignin-degrading enzymes. *Proc Natl Acad Sci U S A*, **115**, 6428–6433.

Bae,S.-J. *et al.* (2016) Efficient production of acetoin in Saccharomyces cerevisiae by disruption of 2,3-butanediol dehydrogenase and expression of NADH oxidase. *Sci Rep*, **6**, 27667.

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res*, **28**, 45–48.

Baldauf,S.L. (2003) The deep roots of eukaryotes. *Science*, **300**, 1703–1706.

Barabási,A.-L. and Albert,R. (1999) Emergence of Scaling in Random Networks. *Science*, **286**, 509–512.

Barabási,A.-L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet*, **5**, 101–113.

Bastian,M. *et al.* (2009) Gephi: An Open Source Software for Exploring and Manipulating Networks. *ICWSM*, **3**, 361–362.

Becker,J. *et al.* (2012) GKK1032A2, a secondary metabolite from Penicillium sp. IBWF-029-96, inhibits conidial germination in the rice blast fungus Magnaporthe oryzae. *J Antibiot*, **65**, 99–102.

Benocci,T. *et al.* (2017) Regulators of plant biomass degradation in ascomycetous fungi. *Biotechnology for Biofuels*, **10**, 152.

Berbee,M.L. *et al.* (2017) Early Diverging Fungi: Diversity and Impact at the Dawn of Terrestrial Life. *Annu Rev Microbiol*, **71**, 41–60.

Bhan,A. *et al.* (2002) A duplication growth model of gene expression networks. *Bioinformatics*, **18**, 1486–1493.

Bowers,P.M. *et al.* (2004) Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biology*, **5**, R35.

Brimacombe,R. and Stiege,W. (1985) Structure and function of ribosomal RNA. *Biochem J*, **229**, 1–17.

Buchfink,B. *et al.* (2015) Fast and sensitive protein alignment using DIAMOND. *Nat Methods*, **12**, 59–60.

Bullerwell,C.E. and Lang,B.F. (2005) Fungal evolution: the case of the vanishing mitochondrion. *Curr Opin Microbiol*, **8**, 362–369.

Caceres,I. *et al.* (2020) Aflatoxin Biosynthesis and Genetic Regulation: A Review. *Toxins (Basel)*, **12**, 150.

Campo,S. and San Segundo,B. (2020) Systemic induction of phosphatidylinositol-based signaling in leaves of arbuscular mycorrhizal rice plants. *Sci Rep*, **10**, 15896.

Cantarel,B.L. *et al.* (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res*, **37**, D233–D238.

Carvunis,A.-R. *et al.* (2012) Proto-genes and de novo gene birth. *Nature*, **487**, 370–374.

Cech,T.R. *et al.* (1981) In vitro splicing of the ribosomal RNA precursor of Tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell*, **27**, 487–496.

Chalancon,G. *et al.* (2013) Metabolic Networks, Reconstruction. In, Dubitzky,W. *et al.* (eds),

Encyclopedia of Systems Biology. Springer New York, New York, NY, pp. 1259–1263.

Chaudhary,B. *et al.* (2022) Fungal Dispersal Across Spatial Scales. *Annual Review of Ecology, Evolution, and Systematics*, **53**.

Chen,M. *et al.* (2007) Transcriptional regulation of yeast phospholipid biosynthetic genes. *Biochim Biophys Acta*, **1771**, 310–321.

Chen,Y.-L. *et al.* (2008) Candida albicans Uses Multiple Mechanisms To Acquire the Essential Metabolite Inositol during Infection. *Infect Immun*, **76**, 2793–2801.

Choi,B. *et al.* (2019) Identifying genetic markers for a range of phylogenetic utility–From species to family level. *PLoS One*, **14**, e0218995.

Choi,J. and Kim,S.-H. (2017) A genome Tree of Life for the Fungi kingdom. *Proc Natl Acad Sci U S A*, **114**, 9391–9396.

Chung,M. *et al.* (2018) Using Core Genome Alignments To Assign Bacterial Species. *mSystems*, **3**, e00236-18.

Cliften,P.F. *et al.* (2006) After the Duplication: Gene Loss and Adaptation in Saccharomyces Genomes. *Genetics*, **172**, 863–872.

Cooper,T.G. (1982) Nitrogen Metabolism in Saccharomyces cerevisiae. *Cold Spring Harbor Monograph Archive*, **11**, 39–99.

Copley,S.D. (2020) Evolution of new enzymes by gene duplication and divergence. *The FEBS Journal*, **287**, 1262–1283.

Corradi,N. (2015) Microsporidia: Eukaryotic Intracellular Parasites Shaped by Gene Loss and Horizontal Gene Transfers. *Annu Rev Microbiol*, **69**, 167–183.

Crouch,J.A. *et al.* (2009) What is the value of ITS sequence data in Colletotrichum systematics and species diagnosis? A case study using the falcate-spored graminicolous Colletotrichum group. *Mycologia*, **101**, 648–656.

Crow,K.D. and Wagner,G.P. (2006) What Is the Role of Genome Duplication in the Evolution of Complexity and Diversity? *Molecular Biology and Evolution*, **23**, 887–892.

Danchin,A. (2009) Myopic selection of novel information drives evolution. *Curr Opin Biotechnol*, **20**, 504–508.

D'Ari,R. and Casadesús,J. (1998) Underground metabolism. *BioEssays*, **20**, 181–186.

Datta,S. *et al.* (2017) Different specificities of two aldehyde dehydrogenases from Saccharomyces cerevisiae var. boulardii. *Biosci Rep*, **37**, BSR20160529.

Davison,E.M. (1998) Phytophthora Diseases Worldwide. *Plant Pathology*, **47**, 224–225.

Dean,P. *et al.* (2016) Microsporidia: Why Make Nucleotides if You Can Steal Them? *PLoS Pathog*, **12**, e1005870.

Dentinger,B.T.M. *et al.* (2011) Comparing COI and ITS as DNA barcode markers for mushrooms and allies (Agaricomycotina). *PLoS One*, **6**, e25081.

Dezső,Z. *et al.* (2003) Bioinformatics Analysis of Experimentally Determined Protein Complexes in the Yeast Saccharomyces cerevisiae. *Genome Res*, **13**, 2450–2454.

Díaz-Mejía,J.J. *et al.* (2007) A network perspective on the evolution of metabolism by gene duplication. *Genome Biology*, **8**, R26.

Dib,L. and Carbone,A. (2012) CLAG: an unsupervised non hierarchical clustering algorithm handling biological data. *BMC Bioinformatics*, **13**, 194.

Ding,Y. *et al.* (2008) Molecular analysis of a 4-dimethylallyltryptophan synthase from Malbranchea aurantiaca. *J Biol Chem*, **283**, 16068–16076.

Domazet-Lošo,T. *et al.* (2007) A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in Genetics*, **23**, 533–539.

Doncheva,N.T. *et al.* (2012) Topological analysis and interactive visualization of biological networks and protein structures. *Nat Protoc*, **7**, 670–685.

Douglas,L.M. and Konopka,J.B. (2014) Fungal membrane organization: the eisosome concept. *Annu Rev Microbiol*, **68**, 377–393.

Drott,M.T. *et al.* (2017) Balancing selection for aflatoxin in Aspergillus flavus is maintained through interference competition with, and fungivory by insects. *Proceedings of the Royal Society B: Biological Sciences*, **284**, 20172408.

Ducker,G.S. and Rabinowitz,J.D. (2017) One-Carbon Metabolism in Health and Disease. *Cell Metab*, **25**, 27–42.

Dupont,S. *et al.* (2012) Ergosterol Biosynthesis: A Fungal Pathway for Life on Land? *Evolution*, **66**, 2961–2968.

Dutta,B. *et al.* (2022) Fungi in Pharmaceuticals and Production of Antibiotics. In, Shukla,A.C. (ed), *Applied Mycology: Entrepreneurship with Fungi*, Fungal Biology. Springer International Publishing, Cham, pp. 233–257.

Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.

Eick,C.F. *et al.* (2004) Supervised clustering - algorithms and benefits. In, *16th IEEE*

International Conference on Tools with Artificial Intelligence., pp. 774–776.

Eisenberg,E. and Levanon,E.Y. (2003) Preferential Attachment in the Protein Network Evolution. *Phys. Rev. Lett.*, **91**, 138701.

Ejaz,U. *et al.* (2021) Cellulases: From Bioactivity to a Variety of Industrial Applications. *Biomimetics (Basel)*, **6**, 44.

Emamalipour,M. *et al.* (2020) Horizontal Gene Transfer: From Evolutionary Flexibility to Disease Progression. *Frontiers in Cell and Developmental Biology*, **8**.

Emilia Hannula,S. and Morriën,E. (2022) Will fungi solve the carbon dilemma? *Geoderma*, **413**, 115767.

Enzyme Nomenclature. Recommendations 1992 (1994) *European Journal of Biochemistry*, **223**, 1–5.

Fell,D.A. and Wagner,A. 12 Structural properties of metabolic.

Fell,D.A. and Wagner,A. (2000) The small world of metabolism. *Nat Biotechnol*, **18**, 1121–1122.

Felsenstein,J. (1978) Cases in which Parsimony or Compatibility Methods Will be Positively Misleading. *Systematic Zoology*, **27**, 401–410.

Fink-Gremmels,J. and Malekinejad,H. (2007) Clinical effects and biochemical mechanisms associated with exposure to the mycoestrogen zearalenone. *Animal Feed Science and Technology*, **137**, 326–341.

Finn,R.D. *et al.* (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*, **39**, W29–W37.

Fleck,C.B. and Brock,M. (2010) Aspergillus fumigatus Catalytic Glucokinase and Hexokinase: Expression Analysis and Importance for Germination, Growth, and Conidiation. *Eukaryotic Cell*, **9**, 1120.

Fondi,M. *et al.* (2009) Origin and evolution of operons and metabolic pathways. *Research in Microbiology*, **160**, 502–512.

Fox Keller,E. (2005) Revisiting "scale-free" networks. *BioEssays*, **27**, 1060–1068.

Fraley,C. and Raftery,A.E. (1998) How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *The Computer Journal*, **41**, 578–588.

Frank,D.N. and Pace,N.R. (1998) Ribonuclease P: unity and diversity in a tRNA processing ribozyme. *Annu Rev Biochem*, **67**, 153–180.

Frey,B.J. and Dueck,D. (2007) Clustering by Passing Messages Between Data Points. *Science*, **315**, 972–976.

Friesen,T.L. *et al.* (2006) Emergence of a new disease as a result of interspecific virulence gene transfer. *Nat Genet*, **38**, 953–956.

Frisvad,J.C. *et al.* (2018) Taxonomy of Aspergillus section Flavi and their production of aflatoxins, ochratoxins and other mycotoxins. *Stud Mycol*, **93**, 1–63.

Fu,S.-F. *et al.* (2015) Indole-3-acetic acid: A widespread physiological code in interactions of fungi with other organisms. *Plant Signal Behav*, **10**, e1048052.

Gabaldón,T. (2021) Origin and Early Evolution of the Eukaryotic Cell. *Annu Rev Microbiol*, **75**, 631–647.

Gajewski,J. *et al.* (2017) Engineering fungal de novo fatty acid synthesis for short chain fatty acid production. *Nat Commun*, **8**, 14650.

Garcia-Vallvé,S. *et al.* (2000) Horizontal Gene Transfer in Bacterial and Archaeal Complete Genomes. *Genome Res.*, **10**, 1719–1725.

Gascuel,O. (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol*, **14**, 685–695.

Gazis,R. *et al.* (2011) Species delimitation in fungal endophyte diversity studies and its implications in ecological and biogeographic inferences. *Mol Ecol*, **20**, 3001–3013.

Gilbert,W. (1986) Origin of life: The RNA world. *Nature*, **319**, 618–618.

Ginsburg,H. (2006) Progress in in silico functional genomics: the malaria Metabolic Pathways database. *Trends in Parasitology*, **22**, 238–240.

Goffeau,A. *et al.* (1996) Life with 6000 genes. *Science*, **274**, 546, 563–567.

Grigoriev,I.V. *et al.* (2014) MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Research*, **42**, D699–D704.

Grossetête,S. *et al.* (2010) FUNGIpath: a tool to assess fungal metabolic pathways predicted by orthology. *BMC Genomics*, **11**, 81.

Guindon,S. *et al.* (2010) New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, **59**, 307–321.

Haldane,J.B.S. (1933) The Part Played by Recurrent Mutation in Evolution. *The American Naturalist*, **67**, 5–19.

Hamming,R.W. (1950) Error detecting and error correcting codes. *The Bell System Technical*

Journal, **29**, 147–160.

Hanks,S.K. and Hunter,T. (1995) The Eukaryotic Protein Kinase Superfamily. In, *The Protein Kinase FactsBook*. Elsevier, pp. 7–47.

Hartl,L. *et al.* (2012) Fungal chitinases: diversity, mechanistic properties and biotechnological potential. *Appl Microbiol Biotechnol*, **93**, 533–543.

Hartwell,L.H. *et al.* (1999) From molecular to modular cell biology. *Nature*, **402**, C47–C52.

Heames,B. *et al.* (2020) A Continuum of Evolving De Novo Genes Drives Protein-Coding Novelty in Drosophila. *J Mol Evol*, **88**, 382–398.

Hebert,P.D.N. *et al.* (2003) Biological identifications through DNA barcodes. *Proc Biol Sci*, **270**, 313–321.

Hennicke,F. *et al.* (2013) Factors Supporting Cysteine Tolerance and Sulfite Production in Candida albicans. *Eukaryot Cell*, **12**, 604–613.

Hérivaux,A. *et al.* (2017) The Identification of Phytohormone Receptor Homologs in Early Diverging Fungi Suggests a Role for Plant Sensing in Land Colonization by Fungi. *mBio*, **8**, 10.1128/mbio.01739-16.

Horowitz,N.H. (1945) On the Evolution of Biochemical Syntheses. *Proc Natl Acad Sci U S A*, **31**, 153–157.

Houbraken,J. *et al.* (2020) Classification of Aspergillus, Penicillium, Talaromyces and related genera (Eurotiales): An overview of families, genera, subgenera, sections, series and species. *Stud Mycol*, **95**, 5–169.

How,C.W. *et al.* (2022) How far have we explored fungi to fight cancer? *Seminars in Cancer Biology*, **86**, 976–989.

Huey,C.J. *et al.* (2020) Mycorrhiza: a natural resource assists plant growth under varied soil conditions. *3 Biotech*, **10**, 204.

Huws,S.A. *et al.* (2018) Addressing Global Ruminant Agricultural Challenges Through Understanding the Rumen Microbiome: Past, Present, and Future. *Frontiers in Microbiology*, **9**.

Ingolia,T.D. and Queener,S.W. (1989) Beta-lactam biosynthetic genes. *Medicinal Research Reviews*, **9**, 245–264.

Iwahori,A. *et al.* (2000) cDNA-derived amino acid sequence of acetoacetyl-CoA synthetase from rat liver. *FEBS Lett*, **466**, 239–243.

Jacob,F. (1977) Evolution and Tinkering. *Science*, **196**, 1161–1166.

Jensen,R.A. (1976) Enzyme Recruitment in Evolution of New Function. *Annual Review of Microbiology*, **30**, 409–425.

Jeong,H. *et al.* (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Jeong,H. *et al.* (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.

Jing,L.S. *et al.* (2014) Database and tools for metabolic network analysis. *Biotechnol Bioproc E*, **19**, 568–585.

Kaessmann,H. (2010) Origins, evolution, and phenotypic impact of new genes. *Genome Res*, **20**, 1313–1326.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, **28**, 27–30.

Kanemaki,M.T. (2022) A rethink about enzymes that drive DNA replication. *Nature*, **605**, 228–229.

Karp,P.D. *et al.* (2000) The EcoCyc and MetaCyc databases. *Nucleic Acids Research*, **28**, 56–59.

Karp,P.D. *et al.* (2002) The Pathway Tools software. *Bioinformatics*, **18**, S225–S232.

Katoh,K. *et al.* (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, **30**, 3059–3066.

Keeling,P.J. and Palmer,J.D. (2008) Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet*, **9**, 605–618.

Keeling,P.J. and Slamovits,C.H. (2004) Simplicity and Complexity of Microsporidian Genomes. *Eukaryot Cell*, **3**, 1363–1369.

Keller,N.P. (2019) Fungal secondary metabolism: regulation, function and drug discovery. *Nat Rev Microbiol*, **17**, 167–180.

Khalid,S. *et al.* (2018) NRPS-Derived Isoquinolines and Lipopetides Mediate Antagonism between Plant Pathogenic Fungi and Bacteria. *ACS Chem Biol*, **13**, 171–179.

Khanin,R. and Wit,E. (2006) How Scale-Free Are Biological Networks. *Journal of Computational Biology*, **13**, 810–818.

Khersonsky,O. and Tawfik,D.S. (2010) Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu Rev Biochem*, **79**, 471–505.

Kim,H.S. *et al.* (2006) MANET: tracing evolution of protein architecture in metabolic networks. *BMC Bioinformatics*, **7**, 351.

Kimura,M. *et al.* (2010) Protostadienol synthase from Aspergillus fumigatus: functional conversion into lanosterol synthase. *Biochem Biophys Res Commun*, **391**, 899–902.

Kinene,T. *et al.* (2016) Rooting Trees, Methods for. *Encyclopedia of Evolutionary Biology*, 489–493.

Knowles,D.G. and McLysaght,A. (2009) Recent de novo origin of human protein-coding genes. *Genome Res*, **19**, 1752–1759.

Krantz,M. *et al.* (2006) Comparative genomics of the HOG-signalling system in fungi. *Curr Genet*, **49**, 137–151.

Kreimer,A. *et al.* (2008) The evolution of modularity in bacterial metabolic networks. *Proceedings of the National Academy of Sciences*, **105**, 6976–6981.

Kumar,A. *et al.* (2023) Industrial applications of fungal lipases: a review. *Front Microbiol*, **14**, 1142536.

Labena,A.A. *et al.* (2018) Metabolic pathway databases and model repositories. *Quant Biol*, **6**, 30–39.

Large,P.J. (1986) Degradation of organic nitrogen compounds by yeasts. *Yeast*, **2**, 1–34.

Laurian,R. *et al.* (2019) Hexokinase and Glucokinases Are Essential for Fitness and Virulence in the Pathogenic Yeast Candida albicans. *Frontiers in Microbiology*, **10**.

Lazcano,A. and Miller,S.L. (1999) On the origin of metabolic pathways. *J Mol Evol*, **49**, 424–431.

Leclerc,R.D. (2008) Survival of the sparsest: robust gene networks are parsimonious. *Mol Syst Biol*, **4**, 213.

Lees,J.A. *et al.* (2018) Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study. *Wellcome Open Res*, **3**, 33.

Li,Y. *et al.* (2014) Expansion of biological pathways based on evolutionary inference. *Cell*, **158**, 213–225.

Li,Z.-W. *et al.* (2016) On the Origin of De Novo Genes in Arabidopsis thaliana Populations. *Genome Biol Evol*, **8**, 2190–2202.

Lima-Mendez,G. and Helden,J. van (2009) The powerful law of the power law and other myths in network biology. *Mol. BioSyst.*, **5**, 1482–1493.

Lin,M. *et al.* (2013) Research Commentary—Too Big to Fail: Large Samples and the p-Value Problem. *Information Systems Research*, **24**, 906–917.

Lin,Y. *et al.* (2011) Comparative studies of de novo assembly tools for next-generation sequencing technologies. *Bioinformatics*, **27**, 2031–2037.

Linster,C.L. and Van Schaftingen,E. (2007) Vitamin C. Biosynthesis, recycling and degradation in mammals. *FEBS J*, **274**, 1–22.

Lopez-Nieves,S. *et al.* (2019) Biochemical characterization of TyrA dehydrogenases from Saccharomyces cerevisiae (Ascomycota) and Pleurotus ostreatus (Basidiomycota). *Archives of Biochemistry and Biophysics*, **665**, 12–19.

MacDonald,N.J. and Beiko,R.G. (2010) Efficient learning of microbial genotype-phenotype association rules. *Bioinformatics*, **26**, 1834–1840.

MacQueen,J. (1967) Some methods for classification and analysis of multivariate observations. In, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. University of California Press, pp. 281–298.

Malik,V.S. (1980) Microbial secondary metabolism. *Trends in Biochemical Sciences*, **5**, 68–72.

Marcus,P.I. and Talalay,P. (1956) Induction and purification of alpha- and beta-hydroxysteroid dehydrogenases. *J Biol Chem*, **218**, 661–674.

Margulis,L. (1971) The Origin of Plant and Animal Cells: The serial symbiosis view of the origin of higher cells suggests that the customary division of living things into two kingdoms should be reconsidered. *American Scientist*, **59**, 230–235.

Marques,C.A. and McCulloch,R. (2018) Conservation and Variation in Strategies for DNA Replication of Kinetoplastid Nuclear Genomes. *Curr Genomics*, **19**, 98–109.

Martínková,L. *et al.* (2015) Cyanide hydratases and cyanide dihydratases: emerging tools in the biodegradation and biodetection of cyanide. *Appl Microbiol Biotechnol*, **99**, 8875–8882.

Massengo-Tiassé,R.P. and Cronan,J.E. (2009) Diversity in Enoyl-Acyl Carrier Protein Reductases. *Cell Mol Life Sci*, **66**, 1507.

Mazerska,Z. *et al.* (2016) The role of glucuronidation in drug resistance. *Pharmacology & Therapeutics*, **159**, 35–55.

McDonald,A.G. *et al.* (2009) ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic Acids Research*, **37**, D593–D597.

Michal,G. (1998) On representation of metabolic pathways. *Biosystems*, **47**, 1–7.

Millán,J.L. (2006) Alkaline Phosphatases. *Purinergic Signal*, **2**, 335–341.

Mironov,A.S. *et al.* (2002) Sensing Small Molecules by Nascent RNA: A Mechanism to Control Transcription in Bacteria. *Cell*, **111**, 747–756.

Mizote,T. *et al.* (1999) Cloning and characterization of the thiD/J gene of Escherichia coli encoding a thiamin-synthesizing bifunctional enzyme, hydroxymethylpyrimidine kinase/phosphomethylpyrimidine kinase. *Microbiology*, **145**, 495–501.

Montanucci,L. *et al.* (2018) Influence of pathway topology and functional class on the molecular evolution of human metabolic genes. *PLOS ONE*, **13**, e0208782.

Moran,N.A. (2002) Microbial minimalism: genome reduction in bacterial pathogens. *Cell*, **108**, 583–586.

Moskowitz,G.J. and Merrick,J.M. (1969) Metabolism of poly-beta-hydroxybutyrate. II. Enzymatic synthesis of D-(-)-beta hydroxybutyryl coenzyme A by an enoyl hydrase from Rhodospirillum rubrum. *Biochemistry*, **8**, 2748–2755.

Müller,V. *et al.* (2008) Discovery of a Ferredoxin:NAD+-Oxidoreductase (Rnf) in Acetobacterium woodii. *Annals of the New York Academy of Sciences*, **1125**, 137–146.

Murphy,D.N. and McLysaght,A. (2012) De Novo Origin of Protein-Coding Genes in Murine Rodents. *PLoS One*, **7**, e48650.

Nakjang,S. *et al.* (2013) Reduction and Expansion in Microsporidian Genome Evolution: New Insights from Comparative Genomics. *Genome Biol Evol*, **5**, 2285–2303.

Naranjo-Ortiz,M.A. and Gabaldón,T. (2019) Fungal evolution: diversity, taxonomy and phylogeny of the Fungi. *Biological Reviews*, **94**, 2101–2137.

Neyman,J. (1971) MOLECULAR STUDIES OF EVOLUTION: A SOURCE OF NOVEL STATISTICAL PROBLEMS**This investigation was supported in part by research grant GM 10525-08 from the National Institutes of Health, Public Health Service. In, Gupta,S.S. and Yackel,J. (eds), *Statistical Decision Theory and Related Topics*. Academic Press, pp. 1–27.

Nguyen,L.-T. *et al.* (2015) IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, **32**, 268–274.

Noda-Garcia,L. *et al.* (2018) Metabolite–Enzyme Coevolution: From Single Enzymes to Metabolic Pathways and Networks. *Annu. Rev. Biochem.*, **87**, 187–216.

Notebaart,R.A. *et al.* (2014) Network-level architecture and the evolutionary potential of underground metabolism. *Proceedings of the National Academy of Sciences*, **111**, 11762–11767.

Novo,M. *et al.* (2009) Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast Saccharomyces cerevisiae EC1118. *Proceedings of the National Academy of Sciences*, **106**, 16333–16338.

Ohno,S. (2013) Evolution by Gene Duplication Springer Science & Business Media.

Olson,M.V. (1999) When less is more: gene loss as an engine of evolutionary change. *Am J Hum Genet*, **64**, 18–23.

Ooi,H.S. *et al.* (2010) Biomolecular pathway databases. *Methods Mol Biol*, **609**, 129–144.

Orth,J.D. *et al.* (2011) A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011. *Mol Syst Biol*, **7**, 535.

Overbeek,R. *et al.* (2003) The ERGOTM genome analysis and discovery system. *Nucleic Acids Res*, **31**, 164–171.

Panneman,H. *et al.* (1996) Cloning and biochemical characterisation of an Aspergillus niger glucokinase. Evidence for the presence of separate glucokinase and hexokinase enzymes. *Eur J Biochem*, **240**, 518–525.

Pariza,M.W. and Johnson,E.A. (2001) Evaluating the safety of microbial enzyme preparations used in food processing: update for a new century. *Regul Toxicol Pharmacol*, **33**, 173–186.

Parter,M. *et al.* (2007) Environmental variability and modularity of bacterial metabolic networks. *BMC Evol Biol*, **7**, 169.

Pavlicek,A. *et al.* (2006) Retroposition of processed pseudogenes: the impact of RNA stability and translational control. *Trends Genet*, **22**, 69–73.

Pellegrini,M. *et al.* (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, **96**, 4285–4288.

Peregrín-Alvarez,J.M. *et al.* (2009) The conservation and evolutionary modularity of metabolism. *Genome Biology*, **10**, R63.

Pereira,C. *et al.* (2014) A meta-approach for improving the prediction and the functional annotation of ortholog groups. *BMC Genomics*, **15**, S16.

Pereira,C. (2015) Nouvelles approches bioinformatiques pour l'étude à grande échelle de l'évolution des activités enzymatiques.

Pereira-Leal,J.B. and Teichmann,S.A. (2005) Novel specificities emerge by stepwise duplication of functional modules. *Genome Res*, **15**, 552–559.

Petrényi,K. *et al.* (2016) Analysis of Two Putative Candida albicans Phosphopantothenoylcysteine Decarboxylase / Protein Phosphatase Z Regulatory Subunits Reveals an Unexpected Distribution of Functional Roles. *PLoS One*, **11**, e0160965.

Pirovich,D.B. *et al.* (2021) Multifunctional Fructose 1,6-Bisphosphate Aldolase as a Therapeutic Target. *Front Mol Biosci*, **8**, 719678.

Pitkänen,E. *et al.* (2014) Comparative Genome-Scale Reconstruction of Gapless Metabolic

Networks for Present and Ancestral Species. *PLOS Computational Biology*, **10**, e1003465.

Radrich,K. *et al.* (2010) Integration of metabolic databases for the reconstruction of genome-scale metabolic networks. *BMC Systems Biology*, **4**, 114.

Ralser,M. *et al.* (2021) The evolution of the metabolic network over long timelines. *Current Opinion in Systems Biology*, **28**, 100402.

Rancurel,C. *et al.* (2017) Alienness: Rapid Detection of Candidate Horizontal Gene Transfers across the Tree of Life. *Genes*, **8**, 248.

Rannala,B. and Yang,Z. (1996) Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *J Mol Evol*, **43**, 304–311.

Ravasz,E. *et al.* (2002) Hierarchical Organization of Modularity in Metabolic Networks. *Science*, **297**, 1551–1555.

Ravasz,E. and Barabási,A.-L. (2003) Hierarchical organization in complex networks. *Phys. Rev. E*, **67**, 026112.

Ren,S. *et al.* (2017) Functional analyses of the versicolorin B synthase gene in Aspergillus flavus. *Microbiologyopen*, **6**, e00471.

Ribichich,K.F. *et al.* (2006) Comparative EST analysis provides insights into the basal aquatic fungus Blastocladiella emersonii. *BMC Genomics*, **7**, 177.

Rives,A.W. and Galitski,T. (2003) Modular organization of cellular networks. *Proc Natl Acad Sci U S A*, **100**, 1128–1133.

Romero,P. *et al.* (2004) Computational prediction of human metabolic pathways from the complete human genome. *Genome Biology*, **6**, R2.

Ross,M.G. *et al.* (2013) Characterizing and measuring bias in sequence data. *Genome Biology*, **14**, R51.

Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, **4**, 406–425.

Sakuno,E. *et al.* (2005) Aspergillus parasiticus cyclase catalyzes two dehydration steps in aflatoxin biosynthesis. *Appl Environ Microbiol*, **71**, 2999–3006.

Schindel,D.E. and Miller,S.E. (2005) DNA barcoding a useful tool for taxonomists. *Nature*, **435**, 17.

Schneider,J. *et al.* (2010) CARMEN - Comparative Analysis and in silico Reconstruction of organism-specific MEtabolic Networks. *Genet Mol Res*, **9**, 1660–1672.

Schoch,C.L. *et al.* (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences*, **109**, 6241–6246.

Scossa,F. and Fernie,A.R. (2020) The evolution of metabolism: How to test evolutionary hypotheses at the genomic level. *Computational and Structural Biotechnology Journal*, **18**, 482–500.

Segal,E. *et al.* (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*, **34**, 166–176.

Seifert,K.A. *et al.* (2007) Prospects for fungus identification using CO1 DNA barcodes, with Penicillium as a test case. *Proc Natl Acad Sci U S A*, **104**, 3901–3906.

Shannon,P. *et al.* (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.*, **13**, 2498–2504.

Sievers,F. and Higgins,D.G. (2018) Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci*, **27**, 135–145.

Silhavy,T.J. *et al.* (2010) The Bacterial Cell Envelope. *Cold Spring Harb Perspect Biol*, **2**, a000414.

Slot,J.C. and Rokas,A. (2011) Horizontal transfer of a large and highly toxic secondary metabolic gene cluster between fungi. *Curr Biol*, **21**, 134–139.

Spatafora,J.W. *et al.* (2017) The Fungal Tree of Life: from Molecular Systematics to Genome-Scale Phylogenies. *Microbiology Spectrum*, **5**, 10.1128/microbiolspec.funk-0053–2016.

Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Steffensky,M. *et al.* (2000) Identification of the Novobiocin Biosynthetic Gene Cluster of Streptomyces spheroides NCIB 11891. *Antimicrob Agents Chemother*, **44**, 1214–1222.

Stentiford,G.D. *et al.* (2016) Microsporidia – Emergent Pathogens in the Global Food Chain. *Trends in Parasitology*, **32**, 336–348.

Stobbe,M.D. *et al.* (2011) Critical assessment of human metabolic pathway databases: a stepping stone for future integration. *BMC Systems Biology*, **5**, 165.

Stobbe,M.D. *et al.* (2014) Knowledge representation in metabolic pathway databases. *Briefings in Bioinformatics*, **15**, 455–470.

Stumpf,M.P.H. *et al.* (2005) Subnets of scale-free networks are not scale-free: Sampling properties of networks. *Proceedings of the National Academy of Sciences*, **102**, 4221–4224.

Takemoto,K. *et al.* (2007) Correlation between structure and temperature in prokaryotic metabolic networks. *BMC Bioinformatics*, **8**, 303.

Tam,P. *et al.* (2015) Candida glabrata, Friend and Foe. *Journal of Fungi*, **1**, 277–292.

Tautz,D. and Domazet-Lošo,T. (2011) The evolutionary origin of orphan genes. *Nat Rev Genet*, **12**, 692–702.

Taylor,J.W. (1995) Molecular Phylogenetic classification of fungi. *Arch Med Res*, **26**, 307–314.

Taylor,S.L. *et al.* (1986) Sulfites in foods: uses, analytical methods, residues, fate, exposure assessment, metabolism, toxicity, and hypersensitivity. *Adv Food Res*, **30**, 1–76.

Thiele,I. *et al.* (2005) Expanded Metabolic Reconstruction of Helicobacter pylori (iIT341 GSM/GPR): an In Silico Genome-Scale Characterization of Single- and Double-Deletion Mutants. *J Bacteriol*, **187**, 5818–5830.

Tiburcio,R.A. *et al.* (2010) Genes acquired by horizontal transfer are potentially involved in the evolution of phytopathogenicity in Moniliophthora perniciosa and Moniliophthora roreri, two of the major pathogens of cacao. *J Mol Evol*, **70**, 85–97.

Tiwari,P. and Bae,H. (2020) Horizontal Gene Transfer and Endophytes: An Implication for the Acquisition of Novel Traits. *Plants (Basel)*, **9**, 305.

Todd,R.B. *et al.* (2014) Prevalence of transcription factors in ascomycete and basidiomycete fungi. *BMC Genomics*, **15**, 214.

Toyomasu,T. *et al.* (2007) Fusicoccins are biosynthesized by an unusual chimera diterpene synthase in fungi. *Proceedings of the National Academy of Sciences*, **104**, 3084–3088.

Vakirlis,N. *et al.* (2018) A Molecular Portrait of De Novo Genes in Yeasts. *Mol Biol Evol*, **35**, 631–645.

Van Dyke,M.C.C. *et al.* (2019) Fantastic yeasts and where to find them: the hidden diversity of dimorphic fungal pathogens. *Current Opinion in Microbiology*, **52**, 55–63.

Větrovský,T. *et al.* (2016) The rpb2 gene represents a viable alternative molecular marker for the analysis of environmental fungal communities. *Molecular Ecology Resources*, **16**, 388–401.

Vickers,T.J. and Beverley,S.M. (2011) Folate metabolic pathways in Leishmania. *Essays Biochem*, **51**, 63–80.

Vitkup,D. *et al.* (2006) Influence of metabolic network structure and function on enzyme evolution. *Genome Biol*, **7**, R39.

Vosseberg,J. *et al.* (2021) Timing the origin of eukaryotic cellular complexity with ancient duplications. *Nat Ecol Evol*, **5**, 92–100.

Wallwey,C. *et al.* (2012) Genome mining reveals the presence of a conserved gene cluster for the biosynthesis of ergot alkaloid precursors in the fungal family Arthrodermataceae. *Microbiology (Reading)*, **158**, 1634–1644.

Wang,H. *et al.* (2013) Role of the Phenylalanine-Hydroxylating System in Aromatic Substance Degradation and Lipid Metabolism in the Oleaginous Fungus Mortierella alpina. *Applied and Environmental Microbiology*, **79**, 3225–3233.

Wang,H. *et al.* (2020) Tetrahydrobiopterin Plays a Functionally Significant Role in Lipogenesis in the Oleaginous Fungus Mortierella alpina. *Front Microbiol*, **11**, 250.

Wang,X. *et al.* (2012) [Furfural degradation by filamentous fungus Amorphotheca resinae ZN1]. *Sheng Wu Gong Cheng Xue Bao*, **28**, 1070–1079.

Wang,Y. *et al.* (2022) Comparative genome analysis of plant ascomycete fungal pathogens with different lifestyles reveals distinctive virulence strategies. *BMC Genomics*, **23**, 34.

Wang,Z. *et al.* (2017) Combining inferred regulatory and reconstructed metabolic networks enhances phenotype prediction in yeast. *PLoS Comput Biol*, **13**, e1005489.

Westrick,N.M. *et al.* (2022) A broadly conserved fungal alcohol oxidase (AOX) facilitates fungal invasion of plants. *Mol Plant Pathol*, **24**, 28–43.

Wick,R.R. and Holt,K.E. (2019) Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Res*, **8**, 2138.

Wierckx,N. *et al.* (2020) Metabolic specialization in itaconic acid production: a tale of two fungi. *Current Opinion in Biotechnology*, **62**, 153–159.

Will,J.L. *et al.* (2010) Incipient Balancing Selection through Adaptive Loss of Aquaporins in Natural Saccharomyces cerevisiae Populations. *PLOS Genetics*, **6**, e1000893.

Wilson,R. and Turner,A.P.F. (1992) Glucose oxidase: an ideal enzyme. *Biosensors and Bioelectronics*, **7**, 165–185.

Wilson,R.A. *et al.* (2012) Towards Defining Nutrient Conditions Encountered by the Rice Blast Fungus during Host Infection. *PLOS ONE*, **7**, e47392.

Wisecaver,J.H. *et al.* (2014) The Evolution of Fungal Metabolic Pathways. *PLoS Genet*, **10**.

Wittig,U. and De Beuckelaer,A. (2001) Analysis and comparison of metabolic pathway databases. *Brief Bioinform*, **2**, 126–142.

Woese,C.R. and Fox,G.E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A*, **74**, 5088–5090.

Wong,S.S.W. *et al.* (2017) Treatment of Cyclosporin A retains host defense against invasive

pulmonary aspergillosis in a non-immunosuppressive murine model by preserving the myeloid cell population. *Virulence*, **8**, 1744–1752.

Woodhouse,M.R. *et al.* (2021) A pan-genomic approach to genome databases using maize as a model system. *BMC Plant Biology*, **21**, 385.

Wu,D.-D. *et al.* (2011) De Novo Origin of Human Protein-Coding Genes. *PLoS Genet*, **7**, e1002379.

Yamada,T. *et al.* (2006) Extraction of phylogenetic network modules from the metabolic network. *BMC Bioinformatics*, **7**, 130.

Yang,Y. *et al.* (2023) Spermidine Synthase and Saccharopine Reductase Have Co-Expression Patterns Both in Basidiomycetes with Fusion Form and Ascomycetes with Separate Form. *Journal of Fungi*, **9**, 352.

Yook,S.-H. *et al.* (2005) Self-similar scale-free networks and disassortativity. *Phys. Rev. E*, **72**, 045105.

Zamir,L.O. *et al.* (1988) Structure of D-prephenyllactate. A carboxycyclohexadienyl metabolite from Neurospora crassa. *J Biol Chem*, **263**, 17284–17290.

Zhang,L. *et al.* (2019) Rapid evolution of protein diversity by de novo origination in Oryza. *Nat Ecol Evol*, **3**, 679–690.

Zhao,J. *et al.* (2007) Modular co-evolution of metabolic networks. *BMC Bioinformatics*, **8**, 311.

Zhu,D. and Qin,Z.S. (2005) Structural comparison of metabolic networks in selected single cell organisms. *BMC Bioinformatics*, **6**, 8.

Ziemons,S. *et al.* (2017) Penicillin production in industrial strain Penicillium chrysogenum P2niaD18 is not dependent on the copy number of biosynthesis genes. *BMC Biotechnology*, **17**, 16.