Analyse comparative de réseaux métaboliques à l'échelle du génome chez les algues brunes

Pauline Hamon-Giraud

Stage Master 2 mention Bio-informatique parcours IBI
Université de Rennes 1

HAMON-GIRAUD Pauline
2021 - 2022

Analyse comparative de réseaux métaboliques à l'échelle du génome chez les algues brunes.

Équipe Dyliss : IRISA Rennes
Encadrants : Anne Siegel (IRISA), Jeanne Got (IRISA) et Gabriel Markov (Station Biologique de Roscoff)

Master bioinformatique

UNIVERSITÉ DE RENNES 1

ENGAGEMENT DE NON PLAGIAT

Je, soussigné(e) ..HAMON-GIRAUD Pauline.....
étudiant(e) en..M2 Bioinformatique....parcours....IBI.....
déclare être pleinement informé que le plagiat de documents ou d'une partie de document publiés sur toute forme de support, y comprit l'internet, constitue une violation des droits d'auteur ainsi qu'une fraude caractérisée.

En conséquence, je m'engage à citer toutes les sources que j'ai utilisées pour la rédaction de ce document.

Date : 08 juin 2022

Signature :

Document à compléter de manière manuscrite et à insérer obligatoirement en première page du rapport de stage.

Remerciements

Dans un premier temps, je remercie mes encadrantes Anne Siegel et Jeanne Got de m'avoir accueillie au sein de l'équipe Dyliss à l'IRISA et mon encadrant Gabriel Markov de m'avoir accueillie à la station biologique de Roscoff. Ils m'ont tous les trois accompagnée et aiguillée régulièrement tout au long de mon stage. Jeanne Got a particulièrement pris du temps pour m'aider à installer, utiliser et comprendre les nombreux outils informatiques que je ne connaissais pas. Gabriel Markov a pris du temps pour m'expliquer et répondre à mes questions qui portaient sur le domaine de biologie spécifique des algues. Anne Siegel a su trouver de son temps pour m'aiguiller sur les problématiques à explorer lors de ce stage malgré son emploi du temps chargé.

Je remercie aussi Arnaud Belcour qui a toujours su être pédagogue et m'expliquer clairement certains concepts techniques et qui m'a aussi aidée sur des questions de programmation. Je tiens aussi à remercier l'équipe qui a participé au développement de l'outil AuCoMe et à la rédaction de son article associé, de m'avoir permis d'assister aux réunions concernant cette rédaction. Je remercie l'équipe Dyliss et Genscale pour leur accueil à l'Irisa et leur accompagnement donné aux stagiaires.

Je remercie aussi l'équipe de la Station Biologique de Roscoff qui m'a accueillie dans leurs locaux. Je tiens à remercier Olivier Godfroy, Mark Cock et Erwan Corre pour m'avoir donné des renseignements sur le projet Phaeoexplorer, expliqué les caractéristiques des données utilisées, aidé à trouver des pistes de solutions aux problèmes liés aux données rencontrés ainsi que de m'avoir fourni l'arbre phylogénétique de référence que j'ai utilisé pour l'analyse de résultats.

Pour finir, je remercie l'équipe Genouest de permettre l'accès à leur plateforme d'outils utilisés pour ce stage et de la disponibilité de leur support technique. Je remercie également la plateforme ABiMS de m'avoir permis d'accéder aux données génomiques du projet Phaeoexplorer.

Description des abréviations

— **GSMN** = Réseaux métaboliques à l'échelle du génome (Genome-Scale Metabolic Networks)

— **association GPR** = association Gène-Protéine-Réaction

— **LR** = espèces séquencées en longues lectures (Long Read) avec Nanopore et Illumina

— **SR** = espèces séquencées en lectures courtes (Short Read) avec Illumina

— **SBML** = (Systems Biology Markup Language) Langage de balisage pour la biologie des systèmes

— cor_{coph} = coefficient de corrélation cophénétique

— cor_{gbak} = coefficient de corrélation du Gamma de Baker

Sommaire

1 Introduction

L'accès aux données génomiques permet d'approfondir des connaissances biologiques sur les espèces du vivant. La génération de ces données génomiques, engendrant l'enrichissement des bases de données, permet de se concentrer sur des espèces de plus en plus spécifiques. Il sera dans cette étude question de se concentrer sur la classe précise des algues brunes. Les connaissances biologiques apportées à ces espèces peuvent se porter sur leur fonctionnement métabolique, comme il en sera l'objet de cette étude par la reconstruction de réseaux métaboliques.

Il sera premièrement introduit les spécificités des algues brunes dont une partie se concentrant sur une espèce d'intérêt. Dans un second temps, seront introduits les réseaux métaboliques (définitions et intérêts). Pour finir seront énoncés les objectifs du stage.

1.1 Les algues brunes

Les Phaeophyceae, communément dénommées algues brunes, sont des eucaryotes (Eucaryota) du règne des straménopiles (ou hétérocontes) et de l'embranchement des Ochrophyta. Les straménopiles font partie des plus grands groupes d'eucaryotes. Les nombreuses espèces de ce groupe comprennent, par exemple, des osmotrophes (qui se nourrissent à partir de substances dissoutes) telles que les mildious, mais aussi des phototrophes (qui tirent leur énergie à partir de la lumière). Ce sont ces espèces phototrophes qui constituent l'embranchement des Ochrophyta. Les Ochrophyta comprennent plusieurs classes dont les Phaeophyceae, mais aussi les Bacillariophyceae (ou diatomées) et les Eustigmatophyceae. Ces deux dernières classes se positionnent à la base de la classe des Phaeophyceae selon leurs relations phylogénétiques [20]. Les algues vertes (issues des Chlorobiontes) et les algues rouges (Rhodophytes), contrairement aux algues brunes, n'appartiennent pas au règne des straménopiles.

Les algues brunes comptent approximativement 2000 espèces qui se répartissent en 19 ordres dont les Ectocarpales, les Laminariales et les Fucales qui en rassemblent une partie importante comme illustré sur la Figure 1. Elles ont la particularité d'avoir évolué vers une multicellularité complexe. [8] Ce sont des organismes jouant des rôles essentiels dans l'écosystème côtier. Par exemple, les laminaires (Laminariales) forment des forêts sous-marines servant d'habitat à d'autres organismes ou encore constituent des sources d'énergie pour les herbivores. Les algues brunes montrent aussi certains intérêts sur le plan climatique, étant capables de séquestration du carbone. Mais aussi sur le plan économique, étant commercialisées à des fins alimentaires et aussi médicales ou cosmétiques, cela dû à leurs propriétés bioactives [8].

1.1.1 Espèce d'intérêt

Laminarionema elsbetiae est une espèce d'intérêt pour cette étude. *Laminarionema elsbetiae* est une algue de petite taille (environ 50 µm de long) à la structure filamenteuse et ramifiée appartenant à l'ordre des Ectocarpales.

L'objectif de son étude est de comprendre l'impact de son mode de vie endophytique sur ses fonctions métaboliques. Un endophyte se définit comme étant un organisme dont le cycle de vie s'opère au sein d'un hôte végétal. Dans le cas de *Laminarionema elsbetiae*, elle a d'abord

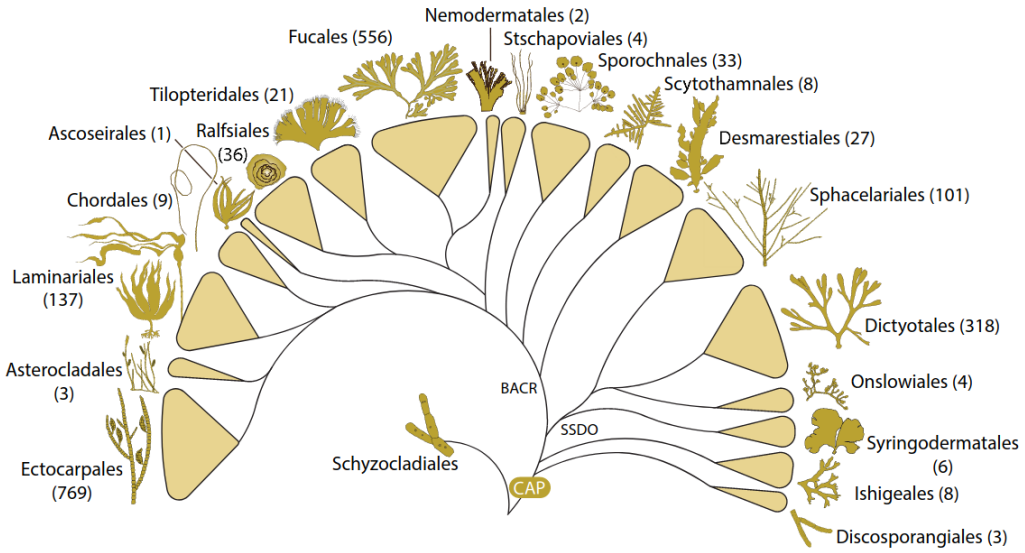

Figure 1. Phylogénie des ordres d'algues brunes. Entre parenthèses, le nombre d'espèces dans chaque ordre selon AlgaeBase. Figure extraite de *Bringloe et al. 2020* [8]

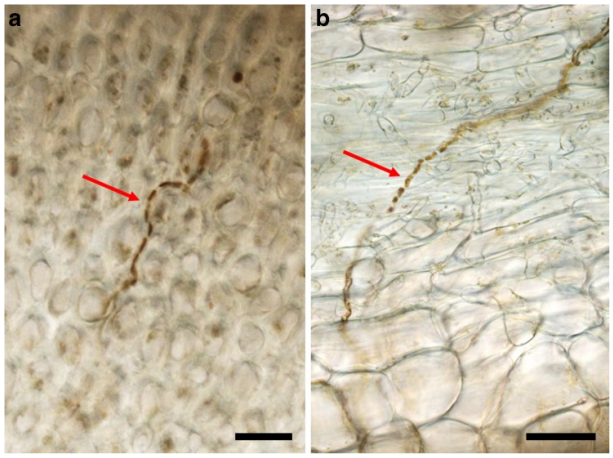

Figure 2. Algues de Bretagne Nord. **a.** Section microscopique du stipe de *Saccharina latissima*. **b.** Section microscopique de l'extrémité du thalle de *Saccharina latissima*. Les flèches rouges indiquent les filaments endophytes. La barre d'échelle représente 25 µm. [6]

été décrite comme infectant l'hôte *Saccharina japonica* sur l'île d'Hokkaido (Japon) par Kawai et Tokuyama en 1995 [15]. Elle a ensuite été décrite comme étant l'endophyte principal de *Saccharina latissima* (précédemment dénommée *Laminaria saccharina*) dans une étude de Elin Ellertsdottir et Akira F. Peters en 1997 sur l'archipel d'Heligoland en Allemagne [9]. Cette même étude montre qu'elle infecte aussi plus occasionnellement l'hôte *Laminaria digitata*. Ces hôtes sont toutes des algues de la famille des Laminariales, algues ayant un rôle important dans l'écosystème côtier et l'économie. Par ailleurs, une étude de Miriam S. Bernard et al. publiée en 2019 décrit que la diversité des hôtes des espèces d'endophytes diffère selon les localités. Elle montre en particulier, pour le cas de la Bretagne, une spécificité d'association hôte-endophyte claire. Ici, le génome de l'endophyte *Laminarionema elsbetiae* n'a été retrouvé que dans l'hôte *Saccharina latissima*[7]. Des filaments de l'endophyte *Laminarionema elsbetiae* présents dans son hôte *Saccharina latissima* sont observables Figure 2.

En plus des bactéries, champignons et de la pollution, les algues au mode de vie endophytique font partie de causes de maladies observées chez les algues hôtes. Ces maladies, constituant une perturbation de la structure ou de la fonction normale de l'hôte, vont avoir un impact sur leurs fonctions écologiques et leur valeur économique [9]. La réponse de l'hôte à l'infection de l'endophyte va causer des symptômes observables tels que des stipes (s'apparente à la tige de l'algue) tordus, des thalles (corps végétatif de l'algue) estropiés ou une réduction de la croissance.
Une étude de Xing et al. de 2021, montre cependant qu'en fonction de l'endophyte et de l'hôte infecté, la réponse à l'infection n'est pas similaire. Plus précisément, l'étude relève que, durant les premières heures après l'infection par *Laminarionema elsbetiae*, l'hôte occasionnel *Laminaria digitata* montrait des réactions de défense fortes, ce qui n'était pas le cas chez l'hôte habituel *Saccharina latissima* [21].

1.2 Les réseaux métaboliques

1.2.1 Définitions

Le **métabolisme** constitue l'ensemble des réactions biochimiques se produisant à l'intérieur des cellules d'un organisme.
Un **réseau métabolique** regroupe l'ensemble de ces réactions, renseignant, pour chaque réaction : les métabolites d'entrée appelés substrats, l'enzyme qui catalyse la réaction et les métabolites de sortie nommés produits.
Est définie comme **voie métabolique**, une succession de réactions catalysées par une série d'enzymes dans un ordre précis : chaque produit devient le substrat de la réaction suivante.

La construction de réseaux métaboliques à l'échelle du génome (GSMN) et un des moyens principaux servant à l'étude du métabolisme d'organismes. Ces GSMN permettent de renseigner sur les associations Gène-Protéine-Réaction (GPR) présentes dans un organisme. Avec l'amélioration de la qualité des GSMN et le développement de leurs champs d'applications, une meilleure compréhension du métabolisme de divers organismes a pu être apportée [13]. Cependant, la reconstruction de GSMN n'est pas triviale et de nombreux organismes ont un métabolisme encore mal connu.

L'utilisation de GSMN pour étudier le métabolisme a été bénéfique dans diverses applications. Ces applications sont par exemple le ciblage de médicaments contre des agents pathogènes, des prédictions de fonctions enzymatiques, l'analyse du pan-réactome, l'étude des interactions métaboliques entre un hôte et un pathogène ou la modélisation des interactions entre plusieurs cellules ou organismes [13].

Une étude de Schulz et Almaas en 2020 a aussi permis de montrer que les données métaboliques offrent une base pour la génération d'arbres phylogénétique fondés sur les fonctions métaboliques. Ces arbres phylogénétiques sont plus classiquement générés à partir des gènes hautement conservés. À partir de données de 975 espèces réparties dans les trois domaines du vivant : Eucaryotes, Archées et Bactéries, ils ont reconstruit un arbre métabolique basé sur la présence / absence des réactions chez chaque espèce. Leurs résultats montrent une forte similarité entre leur arbre métabolique et l'arbre de vie [17].

1.3 Objectifs de l'étude

L'objectif de l'étude présentée est de reconstruire des réseaux métaboliques à partir de génomes de différentes espèces d'algues brunes. Une fois ces réseaux reconstruits, ils serviront de base de résultats pour diverses analyses. Le stage comprend pour objectifs :

- Installer, déployer et utiliser l'outil de reconstruction de GSMN AuCoMe.
- Reconstruire les réseaux métaboliques de 35 espèces dont 29 algues brunes.
- Observer si le regroupement hiérarchique des espèces, à partir des réseaux métaboliques, se rapproche ou non de leur organisation phylogénique.
- Analyser si le mode vie endophytique de *Laminarionema elsbetiae* se retranscrit dans son métabolisme.

2 Matériel et méthodes

2.1 Les génomes étudiés

Les génomes étudiés sont essentiellement des génomes issus du projet Phaeoexplorer [1]. Ce projet vise à générer des données transcriptomiques et des assemblages de génomes annotés pour un large ensemble d'espèces d'algues brunes. Ces données sont générées afin de pouvoir élargir les connaissances biologiques propres à la classe des Phaeophyceae. Au total, 35 espèces ont été utilisées afin de reconstruire leurs réseaux métaboliques.
Parmi cet ensemble, 29 espèces sont des algues brunes, 23 proviennent des données séquencées au Genoscope pour le projet Phaeoexplorer et 6 sont des génomes publics. Pour certaines espèces, le choix du sexe (Mâle ou Femelle) était possible, le sexe n'ayant pas d'importance pour cette étude, il a été choisi le sexe lié à l'assemblage de meilleure qualité.

Parmi les génomes publics, *Ectocarpus species7* est la souche de référence, car est la seule algue brune pour laquelle il y ait eu une curation experte (structurale et fonctionnelle) de l'annotation automatique. *Ectocarpus species7* était anciennement dénommée *Ectocarpus siliculosus* : espèce présente dans le jeu de données Phaeoexplorer (Figure 3). Les espèces, anciennement assignées à cette même espèce *Ectocarpus siliculosus*, ont, en 2017, été déclinées en 15 espèces distinctes suite à une étude de Montecinos & al. [16]. De ce fait, *Ectocarpus species7* et *Ectocarpus siliculosus* seront attendues comme espèces biologiquement très proches dans les résultats.

Schizocladia ischiensis est l'algue de la lignée apparentée aux algues brunes (Figure 1) [8]. Elle est, de plus, une algue dont le génome a été séquencé et annoté au sein du projet Phaeoexplorer. Cependant, elle n'appartient pas à la classe des Phaeophyceae et fait partie de l'extra-groupe. Cet ensemble extra-groupe se constitue, en plus de *Schizocladia ischiensis*, de 4 diatomées et d'une microalgue eustigmatophyceae. Cet extra-groupe vient constituer un socle d'ancêtres proches de la lignée des algues brunes, *Schizocladia ischiensis* étant la plus proche. L'extra-groupe ainsi que les génomes publics sont ici ajoutés au jeu de données afin de garantir au maximum la propagation de gènes par orthologie chez les espèces d'intérêt afin d'obtenir des réseaux finaux les plus complets possibles.

1. `https://phaeoexplorer.sb-roscoff.fr/home/`

Organisme	Ordre	Taille (bp)	# Contigs	# Gènes	Contigs avec gènes (%)
Chordaria linearis	Ectocarpale	214 613 037	217	17 198	87.1
Desmarestia herbacea	Desmarestiale	430 876 013	1 483	16 271	54.9
Dictyota dichotoma	Dictyotale	851 153 257	6 019	20 583	73.3
Ectocarpus crouaniorum	Ectocarpale	218 468 182	275	17 770	68.0
Ectocarpus fasciculatus	Ectocarpale	227 561 674	781	19 173	91.1
Ectocarpus siliculosus	Ectocarpale	200 168 972	1 633	17 801	41.0
Fucus serratus	Fucale	1 234 412 134	8 882	21 263	49.8
Pleurocladia lacustris	Ectocarpale	220 436 342	2 680	16 268	76.1
Porterinema fluviatile	Ralfsiale	167 191 340	110	15 519	99.1
Saccharina latissima	Laminariale	531 271 127	4 592	17 672	59.6
Schizocladia ischiensis	Schizocladiale	194 512 753	130	21 187	99.2
Scytosiphon promiscuus	Ectocarpale	193 199 992	111	19 218	100.0

Figure 3. Espèces séquencées en long read avec Illumina + Nanopore. Algues faisant toutes partie de la classe des Phaeophyceae (algues brunes) à l'exception de *Schizocladia ischiensis*.

Un aspect important à relever sur le jeu de données d'intérêt Phaeoexplorer, est que toutes les espèces n'ont pas été séquencées uniformément. Une partie des espèces a été séquencée en lectures longues (long reads) Nanopore et complétée par courtes lectures (short reads) Illumina (groupe LR). Tandis que l'autre partie ne l'a été uniquement en lectures courtes Illumina (groupe SR). Cela influe sur la qualité des génomes, en particulier sur les génomes SR qui sont beaucoup plus fragmentés que les génomes LR. Cette fragmentation se justifie par un nombre élevé de contigs chez les espèces SR (117 932 en moyenne et valeur médiane de 111 425) observable en détail dans le tableau Figure 4 par rapport aux espèces LR (2 242 en moyenne et valeur médiane de 1 132) observable en détail dans le tableau Figure 3. La différence marquante de ce nombre de contigs entre les SR et les LR se remarque figure 5 (**a.**), on y observe de plus la forte variabilité de ces valeurs chez les SR allant de 18 254 à 325 257 avec un écart type de 87 979 par rapport à celles des LR allant de 110 à 8 882 avec un écart type de 2 718. Cette fragmentation des SR implique un nombre élevé de petits contigs et donc une plus faible proportion de contigs où des structures de gènes sont prédites : entre 4.9% et 38.4% des contigs chez les SR contre 41% à 100% des contigs chez les LR (observable Figures 3, 4, 5 (**.c**).

2.2 Génération des fichiers d'entrée

Les fichiers d'entrée à utiliser afin de reconstruire les réseaux métaboliques avec l'outil AuCoMe (décrit partie suivante : 2.3) doivent être soumis au format GenBank. Les fichiers GenBank ont été générés à partir du package python *emapper2gbk* [5]. Ce package python a été développé afin de générer automatiquement des fichiers GenBank à partir de fichiers d'annotation Eggnog-mapper. Pour cette étude, un fichier GenBank a été généré pour chaque organisme étudié, à partir de son génome, donc ses séquences nucléotidiques, de son protéome, d'un fichier au format d'élément général (general feature format : GFF) et du fichier d'annotations Eggnog-mapper.
Pour chaque organisme, le fichier GenBank ainsi créé, renseignera pour chaque contig :

- sa séquence nucléotidique (avec nombre de bases)
- les séquences codantes présentes détaillant :

Organisme	Ordre	Taille (bp)	# Contigs	# Gènes	Contigs avec gènes (%)
Desmarestia dudresnayi	Desmarestiale	441 530 708	90 763	28 809	19.7
Feldmannia mitchelliae	Ectocarpale	205 927 934	27 497	17 647	30.8
Fucus disticus	Fucale	724 726 007	228 768	19 428	7.4
Hapterophycus canaliculatus	Ectocarpale	192 781 701	26 261	20 036	38.4
Heribaudiella fluviatilis	Sphacelariale	395 970 979	124 447	41 302	23.2
Himanthalia elongata	Fucale	785 461 444	325 257	17 308	4.9
Laminarionema elsbetiae	Ectocarpale	237 791 199	37 108	18 917	25.3
Macrocystis pyrifera	Laminariale	459 008 100	128 700	22 998	15.0
Myriotrichia clavaeformis	Ectocarpale	144 826 515	18 254	21 281	35.3
Pelvetia canaliculata	Fucale	570 570 704	171 596	24 108	8.8
Saccorhiza dermatodea	Tilopteridale	352 008 127	98 404	17 913	14.1
Saccorhiza polyschides	Tilopteridale	557 450 462	138 132	19 747	9.8

Figure 4. Espèces séquencées en short read avec Illumina. Algues faisant toutes partie de la classe des Phaeophyceae (algues brunes).

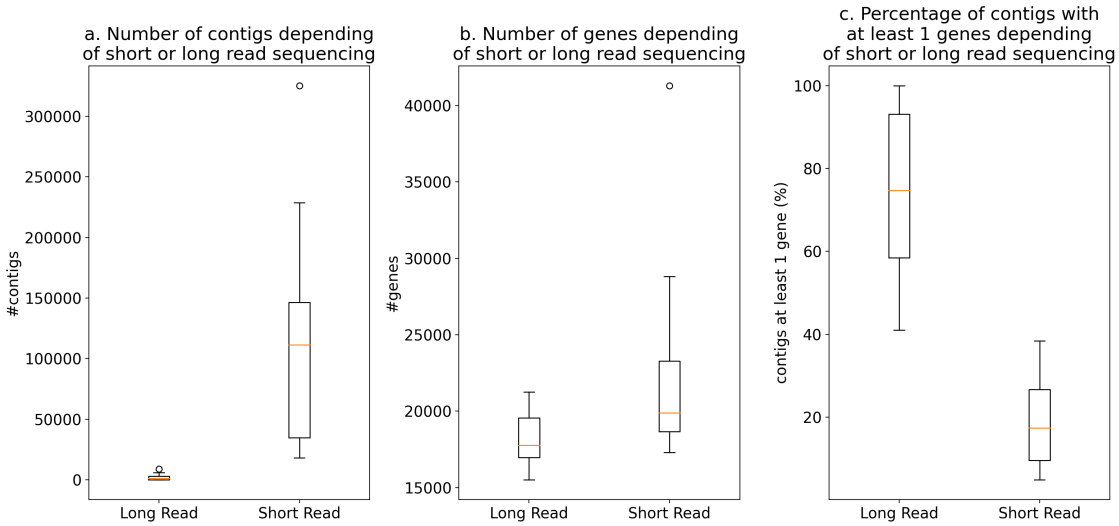

Figure 5. Comparaison avec boxplots des génomes assemblés entre ceux des espèces séquencées en courtes lectures et ceux des espèces séquencées en longues lectures. Comparaison : **a.** du nombre de contigs, **b.** du nombre de gènes et **c.** du pourcentage de contigs contenant au moins 1 gène.

Organisme	Ordre	Taille (bp)	# Contigs	# Gènes	Contigs avec gènes (%)
Cladosiphon okamuranus	Ectocarpale	129 920 920	541	15 166	94.82
Ectocarpus species7	Ectocarpale	196 804 589	30	18 462	100
Ectocarpus subulatus	Ectocarpale	239 578 281	4 677	25 846	100
Nemacystus decipiens	Ectocarpale	153 973 811	685	15 156	68.61
Saccharina japonica	Laminariale	518 091 773	12 993	17 898	22.48
Undaria pinnatifida	Laminariale	634 545 109	33	20 716	93.94

Figure 6. Espèces publiques d'algues brunes intégrées au jeu de données.

Organisme	Classe	Taille (bp)	# Contigs	# Gènes	Contigs avec gènes (%)
Fistulifera solaris	Diatomée	49 736 633	295	20 455	96.95
Fragilariopsis cylindrus	Diatomée	80 540 407	271	18 246	74.91
Phaeodactylum tricornutum	Diatomée	27 450 724	88	10 398	96.59
Thalassiosira pseudonana	Diatomée	32 437 365	64	11 771	96.88
Nannochloropsis gaditana	Eustigmatophyceae	27 589 342	684	10 705	50.44

Figure 7. Extra-groupe.

— leurs annotations structurelles (position)
— leurs annotations fonctionnelles avec les termes GO (Gene Ontology) et les nomenclatures EC (Enzyme Commission numbers)
— leur séquence protéique associée

Les informations taxonomiques de l'organisme en question y sont aussi indiquées.

2.3 Outil AuCoMe

AuCoMe est un package python dont l'objectif est de reconstruire simultanément plusieurs dizaines de réseaux métaboliques à l'échelle du génome (GSMN)[4]. Ce processus de reconstruction s'opère à partir de fichiers de génomes annotés au format GenBank. Pour chaque génome annoté, un GSMN associé sera construit. Un pan-métabolisme est par la suite créé en fusionnant les informations de tous les GSMN. Les annotations des génomes sont souvent hétérogènes en termes de qualité et d'exhaustivité entre les différents organismes. Afin de pouvoir comparer les GSMN créés en évitant ce biais, l'outil vise en particulier à homogénéiser les annotations structurelles et fonctionnelles lors de la construction des GSMN. L'outil reste un moyen de reconstruction automatique de réseaux métaboliques, une analyse experte reste donc nécessaire afin d'affiner les résultats obtenus à l'aide de cet outil.
La reconstruction de ces GSMN homogénéisés se déroule selon un pipeline divisé en 4 étapes principales détaillées ci suit.

2.3.1 Étape de reconstruction des GSMN préliminaires

Après une étape d'initialisation consistant à générer l'arborescence des dossiers utilisés par le programme et une étape de vérification des fichiers GenBank donnés en entrée, cette première étape principale va générer les GSMN préliminaires. Pour chaque organisme, les associations Gène-Protéine-Réaction (GPR) retrouvées à partir des annotations fonctionnelles associées au génome sont ajoutées à son GSMN (Figure en Annexe 17A).

Cette reconstruction est assurée grâce l'algorithme PathoLogic de PathwayTools qui permet de prédire les voies métaboliques à partir de génomes séquencés et annotés [14]. Afin de retrouver les réactions catalysées par les enzymes produites par l'organisme puis les voies métaboliques dans lesquelles les réactions sont présentes, l'algorithme va se référer à la base de données MetaCyc des réactions et des voies métaboliques [14]. La version de l'outil PathwayTools utilisée lors de cette étude est la 25.0. Cet algorithme est parallélisé grâce au package python *mpwt* (pour Multiprocessing PathwayTools) développé dans cet objectif [5]. Pour finir, les réactions

pouvant être associées à des gènes de l'organisme, sont conservées et ajoutées au GSMN sous forme d'association GPR.

2.3.2 Étape de propagation par orthologie des associations GPR

Cette seconde étape du pipeline a pour objectif de compléter, par orthologie, les GSMN préliminaires générés à l'étape précédente. Une protéine d'un certain organisme pourra être prédite comme étant orthologue à une autre protéine issue d'une association GPR d'un autre organisme retrouvée à l'étape précédente. Si cette orthologie est robuste, l'association GPR associée à cette protéine viendra compléter le GSMN concerné (Figure en Annexe 17B).

Cette étape est assurée par l'algorithme d'OrthoFinder qui va permettre de déterminer des groupes d'orthologues [11], [10]. Une fois les groupes d'orthologues constitués, une analyse est faite sur les paires de gènes orthologues de deux organismes distincts. Si parmi les deux gènes de cette paire, l'un est inclus dans une association GPR retrouvée à l'étape précédente, le second sera sélectionné comme association GPR potentielle du GSMN de l'organisme concerné. Pour finir, si l'option de filtre est sélectionnée, toutes les associations GPR potentielles sont filtrées. Seules les associations GPR validées par un critère de robustesse sont propagées dans les GSMN.

2.3.3 Étape de vérification des annotations structurelles

Cette troisième étape du pipeline vient compléter de nouveau les GSMN générés en sortie de l'étape précédente d'orthologie. Son objectif est de retrouver d'éventuelles annotations structurelles non renseignées dans les fichiers d'entrée. Cette recherche se fait par alignement des différents génomes entre eux (Figure en Annexe 17C).

Pour cette étape, chaque génome est comparé par paire avec tous les autres génomes du jeu de données. Le nombre total de comparaisons deux à deux est donc de $\binom{2}{n} = \frac{n!}{2(n-2)!}$ fois, n étant le nombre d'organismes du jeu de données, ce qui fait de cette étape une des plus chronophages si le nombre d'espèces n est élevé. Pour chaque paire et pour chaque séquence protéique appartenant à une association GPR d'un GSMN, un alignement est effectué contre le second génome grâce au package *Biopython*. Si une correspondance est trouvée et validée, alors l'association GPR associée à la séquence protéique est propagée dans le GSMN de l'espèce où elle manquait.

2.3.4 Étape fusion des GSMN et de complétion par réactions spontanées

Cette dernière étape consiste à compléter voies métaboliques des GSMN reconstruits à l'étape précédente par des réactions spontanées. Ces réactions spontanées sont des réactions qui ne sont pas associées à des gènes. Pour pouvoir ajouter ces réactions spontanées aux voies métaboliques, la recherche s'effectue à partir de la base de données MetaCyc pour chaque voie métabolique (Figure en Annexe 17D).

2.4 Outils d'analyse des résultats

2.4.1 Réseaux métaboliques et formats de fichiers

Les réseaux métaboliques sont stockés dans deux formats de fichiers distincts. Le premier format est le format Systems Biology Markup Language (SBML), ce format est utilisé, car libre

et est pris en charge par de nombreux logiciels.

Cependant, ce format classique n'est pas optimal pour stocker certaines métadonnées. Ces métadonnées peuvent, en effet, se révéler importantes par soucis de transparence lors de leur partage, mais aussi lors des analyses réalisées. Elles vont, par exemple, permettre d'indiquer comment les données ont été générées. Par exemple, pour AuCoMe, il va pouvoir être indiqué à quelle étape du pipeline (annotation préliminaire, propagation d'orthologie, vérification structurale ou complétion de réactions spontanées) une réaction a été ajoutée à un GSMN. L'outil *padmet* a donc été développé pour proposer un format de données pouvant stocker ces différentes métadonnées de manière plus structurée que SBML [1]. L'avantage de ce format par rapport au format SBML est qu'il peut contenir plus d'informations que SBML mais aussi qu'il est moins lourd que ce dernier.

Le format PADMET permet aussi de pouvoir stocker les informations localement dans des interfaces de Wikis. Ces Wikis permettent de structurer l'ensemble des données générées tout en les reliant. Les wikis rendent aisément possible l'exploitation et la visualisation des informations des GSMN sans être informaticien grâce à leur interface utilisateur d'une meilleure clarté qu'avec les formats padmet. Ces données sont par exemple les méthodes utilisées dans les pipelines, les réactions, les métabolites, les voies métaboliques ou les gènes. Ces Wikis intègrent aussi des fonctionnalités de recherche sémantique. Il est aussi possible de mettre à jour les GSMN par l'utilisation de formulaires de curation manuelle assistée [1]. Pour pouvoir générer ces wikis, il est possible d'utiliser directement le package *padmet* ou alors d'installer l'environnement *AuReMe*[1].

2.4.2 Génération de dendrogrammes et tableaux

La génération des dendrogrammes par regroupement hiérarchique est réalisée à l'aide du package R *pvclust*[19]. Le regroupement hiérarchique est une méthode statistique qui vise à classer plusieurs éléments dans certains groupes en fonction des similitudes entre eux. Ce package R a pour intérêt de pouvoir effectuer des analyses bootstrap de regroupement hiérarchique tout en informant sur l'incertitude des groupes formés. Le bootstrap consiste à échantillonner aléatoirement des éléments des données et de leur appliquer l'analyse de regroupement hiérarchique un nombre n_{boot} de fois. Un nombre n_{boot} de réplicats de dendrogrammes vont être ainsi générés. Cela va permettre de calculer des indicateurs renseignant sur l'incertitude de chaque groupe sur le dendrogramme final. Le package *pvclust* renseigne sur deux types d'indicateur d'incertitude. Les premiers sont les BP values (Bootstrap Probability values) qui vont renseigner sur la fréquence d'apparition d'un groupe parmi les réplicats de dendrogrammes. Les seconds sont les AU p-value (Approximately Unbiased probability values) qui renseignent sur les valeurs de probabilité (p-values) approximativement sans biais. Les auteurs conseillent d'utiliser une valeur de 10 000 pour le paramètre n_{boot} afin de limiter les erreurs. [19]

La commande d'analyse de l'outil AuCoMe propose donc la création d'un dendrogramme métabolique à l'aide du package *pvclust*. Ce dendrogramme prendra en entrée un tableau de données binaire renseignant pour chaque espèce si chaque réaction est présente ou non dans son GSMN. La distance définie entre deux espèces distinctes sera donc basée sur la présence mutuelle d'une réaction chez ces deux espèces. Cette distance utilisée est la distance de Jaccard

et ne prend donc en considération comme similarité qu'une présence mutuelle des réactions et non l'absence mutuelle. Le nombre n_{boot} de réplicats est fixé à 10 000 comme conseillé par les auteurs.

L'étape d'analyse de l'outil AuCoMe ne permet pas uniquement l'exécution d'un regroupement hiérarchique. Le package *padmet* [1] va aussi permettre d'extraire des fichiers padmet, différents tableaux résumant certaines informations sur les GSMN. Il s'agit de quatre fichiers TSV (Tab Separated Values) : reactions.tsv, pathways.tsv, metabolites.tsv et genes.tsv.

- Le fichier reactions.tsv est celui utilisé pour générer le dendrogramme et renseigne donc sur la présence / absence (1 ou 0) de chaque réaction pour chaque espèce. Il précise, de plus, quels sont les gènes associés à chaque réaction pour chaque espèce.
- Le fichier pathways.tsv renseigne sur la complétion de chaque voie métabolique (sous la forme fractionnaire : *nombre de réactions de la voie métabolique présentes chez l'espèce* / *nombre de réactions total de la voie métabolique*). De plus, pour chaque voie métabolique, la liste des réactions associées est précisée.
- Le fichier metabolites.tsv renseigne, pour chaque métabolite : les réactions qui le consomment ainsi que les réactions qui le produisent (sous la forme d'une liste de réactions), pour chaque espèce.
- Le fichier genes.tsv va renseigner, pour chaque gène, si il est présent ou absent (1 ou 0), pour chaque espèce. De plus, pour chaque gène, les réactions associées sont indiquées.

3 Résultats

3.1 Génération des fichiers GenBank

L'étape préliminaire du projet est celle consistant à générer les fichiers au format d'entrée géré par l'outil AuCoMe, en l'occurrence le format GenBank. Pour ce faire, j'ai eu à utiliser le package python créé à cet effet. Par ailleurs, le format des données génomiques utilisées a permis de révéler que le package ne considérait pas ce format. Cela a permis une évolution du code du package de manière à ce qu'il puisse prendre en compte ce format de données.

Les fichiers GenBank ont été générés à partir des protéomes, des génomes, des annotations et des GFF de chaque organisme à l'aide du package python *emapper2gbk* [5]. Cependant, le format des identifiants renseignés dans ces fichiers n'étaient pas pris en compte par le package. Il a donc fallu adapter le code du package python *emapper2gbk* pour y ajouter une option permettant de considérer ce format. L'élément empêchant ce format d'être pris en compte était la présence de préfixes "mRNA" ou "prot" avant les identifiants des contigs ce qui empêchait leur stricte égalité. Le préfixe "prot" était utilisé dans les fichiers de protéome et dans les fichiers d'annotations. Le préfixe "mRNA" était lui utilisé dans le GFF. Le fichier de génome ne comprenait pas de préfixe. Le format en question est le format Gmove [2] qui permet de prédire des gènes et de générer les fichiers GFF.

2. `https://www.genoscope.cns.fr/gmove/`

3.2 GSMN reconstruits

Une fois les fichiers d'entrée à utiliser générés, le travail primordial de ce projet a donc été la reconstruction des réseaux métaboliques pour l'ensemble des 35 espèces. Pour ce travail, j'ai dû m'approprier l'outil AuCoMe sur des petits jeux de données de test. Par la suite, j'ai aussi appris à utiliser les différents outils permettant le passage à plus grande échelle (35 espèces). En plus de ces aspects de formation techniques, j'ai résolu des problèmes liés aux fichiers d'entrée. Il a donc été nécessaire que j'applique des modifications sur certains fichiers du jeu de données pour assurer la reconstruction des réseaux métaboliques de l'ensemble des espèces.

3.2.1 Déploiement et Installation de l'outil AuCoMe

Pour pouvoir utiliser l'outil AuCoMe afin de reconstruire les GSMN, plusieurs aspects techniques ont dû être appropriés. L'outil a d'abord été installé et testé avec des petits jeux de données sur une machine virtuelle distante. Cette machine virtuelle a été créée à l'aide de l'outil Genostack [3] proposé par la plateforme Genouest, cette machine virtuelle est exécutée à partir du cluster de calcul de Genouest. Ces premiers tests ont permis de rectifier certains problèmes survenus lors de l'installation de l'outil dus à des soucis de compatibilité de versions de package et de version de python.

L'installation de l'outil AuCoMe requiert l'installation de l'outil PathwayTools dont il dépend pour la première étape de reconstruction des GSMN préliminaires. Ici, la version installée a été la plus récente disponible au début du stage, à savoir la version 25.0. Sachant que la dernière version utilisée pour faire fonctionner l'outil avant cette étude était la 23.5, il a donc fallu s'assurer que la version 25.0 permettait de reconstruire les GSMN préliminaire de tous les génomes à étudier. Pour ce faire, tous les fichiers GenBank de chaque organisme ont été exécutés un par un dans la première étape d'AuCoMe utilisant PathwayTools. Ce test a mis en évidence plusieurs génomes qui ne parvenaient pas à passer cette première étape.

3.2.2 Réduction des génomes SR

Le test de PathwayTools a été effectué en analysant d'abord les génomes LR : étant ceux de meilleure qualité, ils ont été jugés prioritaire pour la reconstruction de leur GSMN. Sur ces génomes, tous sont passés dans l'outil PathwayTools. Cependant, concernant les génomes SR, 8 sur 12, soient les deux tiers, ne sont pas passés. L'analyse du rapport d'erreur indiquait la survenue d'une erreur de type "StackOverflowError" ce qui suggérait l'exécution d'une boucle trop longue. Suite à cette erreur a donc été calculé le nombre de contigs de chaque espèce. Cette analyse a révélé que les espèces ayant au moins 90 000 contigs étaient ceux faisant survenir cette erreur.
Ces génomes très fragmentés, de plus de 90 000 contigs, n'ont pas plus de 23% de contigs contenant au moins 1 gène prédit comme observable Figure 3. De ce constat, le choix de ne conserver que les contigs contenant au moins 1 gène prédit dans le fichier GenBank a été appliqué pour ces espèces. Le résultat de cette réduction du nombre de contigs est visible Figure 8 **(1)**. Il peut y être observé une réduction du nombre de contigs initiaux d'en moyenne 86.54%.

3. `https://www.genouest.org/outils/genostack/main-concept.html`

Organisme	(1) # Contigs initiaux	(1) # Contigs après sélection	(1) Réduction du nombre de contigs (%)	(2) Taille (bp) initiale	(2) Taille (bp) après sélection	(2) Réduction de la taille du génome (%)
Desmarestia dudresnayi	90 763	18 579	79.53	441 530 708	191 557 520	56.62
Fucus disticus	228 768	17 109	92.52	724 726 007	117 617 601	83.77
Heribaudiella fluviatilis	124 447	30 340	75.62	395 970 979	193 241 498	51.20
Himanthalia elongata	325 257	15 803	95.14	785 461 444	80 634 764	89.73
Macrocystis pyrifera	128 700	20 748	83.88	459 008 100	117 091 100	74.49
Pelvetia canaliculata	171 596	17 433	89.84	570 570 704	145 475 021	74.50
Saccorhiza dermatodea	98 404	13 947	85.83	352 008 127	119 475 084	66.06
Saccorhiza polyschides	138 132	13 857	89.97	557 450 462	183 925 850	67.01

Figure 8. Tableau indiquant, pour les 8 espèces ayant un nombre de contigs supérieur à 90 000 : **(1)** leur nombre de contigs avant et après sélection basée sur la présence de gène prédits ainsi que le pourcentage de réduction de ce nombre, **(2)** leur taille de génome avant et après la sélection des contigs ainsi que le pourcentage de la réduction de ce nombre.

Cette opération de réduction du nombre de contigs a permis de faire en sorte que tous les génomes SR puissent passer l'étape de reconstruction de GSMN par PathwayTools. Il est cependant important, pour la suite de l'analyse, de considérer que ces 8 espèces concernées ne présentent pas des GSMN basés sur l'entièreté de leur génome. En effet, la réduction du nombre de contigs implique une réduction d'en moyenne 70.42% du génome Figure 8 **(2)**. Il est cependant raisonnable d'émettre l'hypothèse que cette fraction du génome supprimée constitue principalement les régions non codantes ou répétées des organismes.

3.2.3 Passage à plus large échelle pour l'exécution d'AuCoMe

Par soucis de volume disponible et de performance (nombre de CPU et RAM), les exécutions de l'outil AuCoMe comprenant un nombre important d'espèces n'ont pas été effectuées sur la machine virtuelle de Genostack. Ces exécutions ont nécessité d'installer l'outil sur un conteneur docker crée à l'aide de l'outil GoDocker [4] aussi proposé par la plateforme Genouest. Ce conteneur a été créé avec 40 CPU et 180 Go de RAM et propose un espace de stockage de plusieurs To de données. Par comparaison, la machine virtuelle permettait au maximum 8 CPU, 32 Go de RAM et 20 Go d'espace de stockage (cet espace ne permettant pas de stocker tous les fichiers générés lors de l'exécution de l'outil). Avoir un espace de stockage élevé (environ 42.2 Go de données générées pour l'exécution comprenant les 35 espèces) est donc primordial pour pouvoir générer tous les fichiers de l'outil. De plus, un nombre de CPU élevé est important pour augmenter la vitesse d'exécution des différentes étapes du pipeline, car va permettre de paralléliser ces étapes en fonction du nombre d'espèces en accordant un CPU par espèce. Le conteneur GoDocker n'a pas impliqué l'installation d'une image docker pour l'outil AuCoMe, ce service n'a été utilisé ici uniquement pour ses performances et son volume de stockage proposés.

Au total, 35 organismes sont passés dans les étapes du pipeline de l'outil AuCoMe. Sur ces 35 organismes, 29 sont des algues brunes (les 6 autres sont des extra-groupes) et 24 sont des organismes séquencés pour le projet Phaeoexplorer.
De ces GSMN générés, ont été extraites les tables renseignant sur les réactions présentes, la

4. `http://www.genouest.org/godocker/`

complétion des voies métaboliques, les gènes présents ainsi que les métabolites consommés et produits par les réactions pour chaque organisme étudié. En plus de ces tables, les dendrogrammes basés sur la présence des réactions ont été générés ainsi que le pan-métabolisme.

3.2.4 Stockage des informations des GSMN dans des Wikis

Les réseaux métaboliques bien reconstruits pour l'ensemble des espèces, il était important de pouvoir diffuser leurs informations à l'ensemble des chercheurs susceptibles de les utiliser pour leurs études. Pour répondre à ce besoin, j'ai créé des Wikis à l'aide de l'outil AuReMe. L'utilisation de cet outil m'a permis en particulier d'apprendre à installer et à utiliser un conteneur Docker.

Contrairement à l'outil AuCoMe, l'installation de l'outil AuReMe proposant la fonction de création de Wikis à partir des GSMN stockés dans les fichiers padmet a nécessité l'installation d'une image docker.
Des Wikis ont ici été créés pour chaque organisme dont le GSMN a été reconstruit. Cette création de Wikis a eu pour objectif d'être mis à disposition des biologistes travaillant autour du projet Phaeoexplorer étant non familiarisés avec des formats de stockage de données tels que padmet ou SBML.

3.3 Nouveaux outils d'analyse des GSMN

Une fois les réseaux métaboliques générés et diffusés, la suite de l'étude consiste à exploiter les résultats afin de pouvoir en tirer des interprétations biologiques. L'étape d'analyse de l'outil AuCoMe permet d'obtenir une bonne base d'informations, mais ne propose pas de fonctions pour effectuer des analyses dans le détail. Afin d'enrichir et de faciliter les analyses des réseaux métaboliques produits par l'outil, j'ai pris l'initiative de développer des programmes permettant de répondre, pour le moment, en particulier aux besoins d'analyses rencontrés lors du stage. J'ai cependant programmé les différents outils de manière que leurs fonctionnalités puissent s'adapter à n'importe quels résultats d'exécutions d'AuCoMe. J'ai, de plus, tout de même intégré quelques fonctions qui ne m'ont pas été utiles, mais que j'ai jugé pouvant s'avérer pertinentes. Ces outils comprennent un script R consistant à enrichir la création de dendrogrammes en les mettant en forme et en permettant leur comparaison avec un dendrogramme phylogénétique et aussi en un package python proposant une bibliothèque de fonctions permettant diverses analyses et extractions d'informations.

3.3.1 Enrichissement de l'analyse par dendrogrammes

L'enrichissement de l'analyse par dendrogramme permet d'automatiser leur mise en forme pour les rendre visuellement plus lisibles et de faire ressortir des groupes d'espèces choisis. Il permet aussi la comparaison avec la phylogénie de référence. Pour réaliser ce script, je me suis donc documenté sur des packages qui proposaient les fonctionnalités recherchées. J'ai aussi recherché des moyens de quantifier rigoureusement les similitudes des deux dendrogrammes comparés pour ne pas se limiter à une interprétation visuelle subjective. J'ai donc sélectionné deux coefficients de corrélation agissant sur deux caractéristiques différentes des dendrogrammes.

Le package AuCoMe propose par son étape d'analyse de générer un dendrogramme en fonction de la présence ou absence d'une réaction dans le GSMN. Dans le contexte de l'étude comparative de ce dendrogramme et de la phylogénie originale des espèces, ce dendrogramme était amené à être coloré selon les différents groupes de classification des espèces puis juxtaposé à l'arbre phylogénique des espèces correspondantes. Ces manipulations n'étant pas proposées par l'outil d'analyse d'AuCoMe, étaient réalisées manuellement a posteriori, ce qui rendait ces analyses chronophages.

Pour répondre à cette problématique, un script R a donc été réalisé de manière à automatiser la mise en forme et colorisation du dendrogramme ainsi que sa juxtaposition au dendrogramme représentant la phylogénie originale des espèces. De plus, deux indicateurs de corrélations entre les deux dendrogrammes sont calculés, l'un prenant en compte l'agencement des feuilles du dendrogramme et l'autre considérant la longueur des branches.

La génération du dendrogramme se fait toujours à l'aide du package *pvclust* [19]. Le dendrogramme est généré par bootstrap avec un nombre n_{boot} de réplicats par défaut de 10 000, mais qui peut être modifié par l'utilisateur. Il est possible de générer ce dendrogramme en fonction de l'absence/présence des réactions comme initialement dans l'outil AuCoMe mais aussi en fonction de la complétion/incomplétude de voies métaboliques selon un seuil choisi ou en fonction de la présence/absence de métabolites. La méthode de distance utilisée correspond à celle implémentée dans l'outil AuCoMe : il s'agit la méthode "binary" correspondant à la distance de Jaccard.

La mise en forme du dendrogramme est réalisée à l'aide du package R *dendextend*[12]. Pour assurer cette mise en forme, l'utilisateur aura à compléter un tableau TSV indiquant, pour les groupes d'espèces renseignés, de quelle couleur colorer la branche ou la feuille du dendrogramme. Ce tableau doit inscrire des noms de groupes correspondants à ceux renseignés dans un autre tableau TSV décrit partie suivante 3.3.2.
Si, de plus, l'utilisateur fournit un arbre phylogénique de référence au format NEXUS [5] indiquant la longueur des branches (raciné et binaire), il sera alors possible de le comparer au dendrogramme métabolique généré. Cette comparaison se fait visuellement à l'aide de la fonction *"tanglegram"* du package *dendextend* consistant à juxtaposer les dendrogrammes en miroir et de relier les identifiants similaires par des segments. Cette juxtaposition est optimisée par l'algorithme de la fonction *"untangle"* faisant pivoter les branches lorsque les identifiants ne sont pas alignés afin de trouver la meilleure disposition. En plus de ce résultat visuel, deux coefficients de corrélation entre les deux dendrogrammes sont calculés. Il s'agit du coefficient de corrélation cophénétique et du coefficient de corrélation du Gamma de Baker. Ces deux coefficients sont calculés à l'aide de fonctions disponibles dans le même package *dendextend* [12] :

— Le coefficient de corrélation cophénétique (cor_{coph}) prend en compte la longueur des branches pour calculer la similarité. Le principe est de calculer, pour chaque dendrogramme, une matrice de distance cophénétique. Pour ce faire, le dendrogramme est divisé en un certain de nombre de classes délimitées par des droites horizontales tracées sur celui-ci. La matrice de distance s'obtient en calculant pour chaque espèce, deux à deux,

5. Format utilisé en bio-informatique notamment pour le stockage d'informations sur les arbres

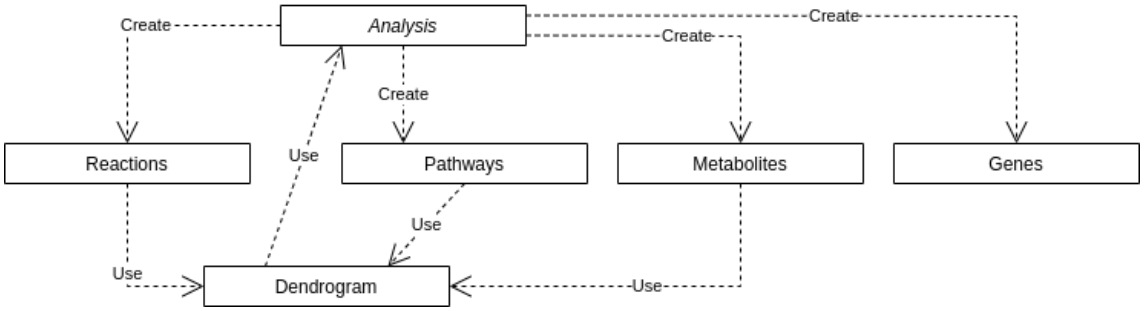

Figure 9. Diagramme des classes du package python *analysis_runs*

la classe dans laquelle leurs branches se rejoignent. Pour finir, le coefficient de corrélation est calculé en comparant les deux matrices obtenues pour chaque dendrogramme, ce coefficient correspond au coefficient de corrélation de Pearson. Il est compris entre -1 et 1, un coefficient proche de 0 indique une forte dissimilarité des dendrogrammes [18].

— Le coefficient de corrélation du Gamma de Baker (cor_{gbak}) (ou gamma de Goodman et Kruskal) va permettre de quantifier la concordance des deux dendrogrammes en fonction du positionnement relatif des feuilles. Pour cela, chaque dendrogramme va être divisé en un nombre k de groupes de clusters. Ensuite, pour chaque paire d'espèces, va être mesuré le plus haut niveau possible de k pour lequel les deux éléments appartiennent toujours au même arbre. Une fois ces mesures relevées pour chaque dendrogramme, elles sont comparées en calculant un coefficient de corrélation de Spearman. Ce coefficient est compris entre -1 et 1 un coefficient proche de 0 indique une forte dissimilarité des dendrogrammes [3].

3.3.2 Description d'un package python créé pour l'analyse des résultats générés par l'outil AuCoMe

En plus du script R enrichissant l'analyse par création de dendrogrammes, j'ai créé un package python mettant à disposition une bibliothèque de fonctions. J'ai donc programmé différents scripts python permettant d'extraire des informations sur des caractéristiques précises des réseaux métaboliques. J'ai, de plus, intégré à cette bibliothèque le script R de manière à centraliser les outils d'analyse dans ce package. Pour générer les différentes fonctions, j'ai dû me familiariser et me documenter sur plusieurs autres packages python dont certaines fonctions dépendent. J'ai aussi dû apprendre à lier plusieurs scripts dans un package.

Le package python *analysis_runs* [6] a donc été créé et permet d'extraire différentes informations des tables renseignant sur les réactions, voies métaboliques, métabolites et gènes retrouvés par l'exécution d'AuCoMe. Ce package se compose de six classes réparties dans six scripts codés dans le langage python. Ces classes et leurs dépendances sont illustrées sur le diagramme de classes Figure 9.

Afin de générer des analyses comparant plusieurs groupes d'espèces, l'utilisateur devra renseigner un tableau TSV indiquant, pour chaque espèce présente dans l'exécution d'AuCoMe, à

6. `https://github.com/PaulineGHG/analysis_runs.git`

quel groupe elle appartient. Il est possible de remplir plusieurs colonnes si plusieurs répartitions de groupes sont à étudier. Par exemple, pour cette étude, une première colonne renseigne sur la classe taxonomique de l'espèce (algue brune, diatomée ou Eustigmatophyceae pour *Nannochloropsis gaditana* et une deuxième colonne de son type de séquençage (LR ou SR pour les génomes Phaeoexplorer et "Public" pour les génomes publics).

La classe *Analysis* prendra en paramètres le fichier TSV ainsi créé, le chemin vers les résultats d'exécutions d'AuCoMe et un chemin dans lequel l'utilisateur souhaite stocker ses résultats d'analyse. Cette classe permet la création d'instances des classes *Reactions*, *Pathways*, *Metabolites* et *Genes* et aussi des analyses visant à comparer différents groupes choisis.

Les quatre classes *Reactions*, *Pathways*, *Metabolites* et *Genes*, de mon package *analysis_runs*, permettent d'extraire les informations des tables reactions.tsv, pathways.tsv, metabolites.tsv et genes.tsv fournies par AuCoMe. Il est possible d'effectuer ces analyses sur l'entièreté du tableau ou alors en filtrant les espèces et les reactions/pathways/metabolites/genes. Il est possible par exemple de ne conserver que les réactions présentes chez un nombre minimal défini d'espèces. Toutes les méthodes ne pourront pas être décrites en détail dans ce rapport, mais à titre d'exemple, certaines permettent d'effectuer des opérations telles que :

- Sélectionner les réactions absentes chez une espèce, mais présentes chez un certain nombre d'autres espèces.
- Sélectionner les voies métaboliques de complétion minimale chez une espèce par rapport aux autres espèces.
- Écrire dans des fichiers fasta, les séquences protéomiques associées à une liste de réactions.

De plus, les classes *Reactions*, *Pathways* et *Metabolites* utilisent la classe *Dendrogram* pour générer les dendrogrammes mis en forme et les comparaisons avec la phylogénie comme décrit partie précédente. La classe *Dendrogram* utilise le package python *rpy2* pour exécuter des lignes de scripts issues du script R sur le script python de la classe.

Ce package fournit donc une bibliothèque de fonctions permettant d'analyser les résultats des exécutions d'AuCoMe. Les fonctions implémentées sont majoritairement des fonctions qui ont été nécessaires aux analyses effectuées pour cette étude. Elles ont cependant été programmées de manière à ce qu'elles soient le plus généralisables possible pour des analyses de futures exécutions d'AuCoMe. L'ajout de nouvelles fonctions et l'adaptation de celles déjà fournies seront certainement nécessaires pour améliorer ce package.

3.4 Analyse comparative des dendrogrammes métaboliques des réactions et de la phylogénie des espèces

Les différents outils que j'ai programmés m'ont donc permis d'effectuer différentes analyses sur les réseaux métaboliques dans l'objectif de décrire des spécificités sur les algues brunes se retranscrivant par leur métabolisme. Ces analyses m'ont cependant surtout permis de révéler des résultats démontrant une répercussion de l'hétérogénéité des données utilisées sur les réseaux métaboliques finaux générés.

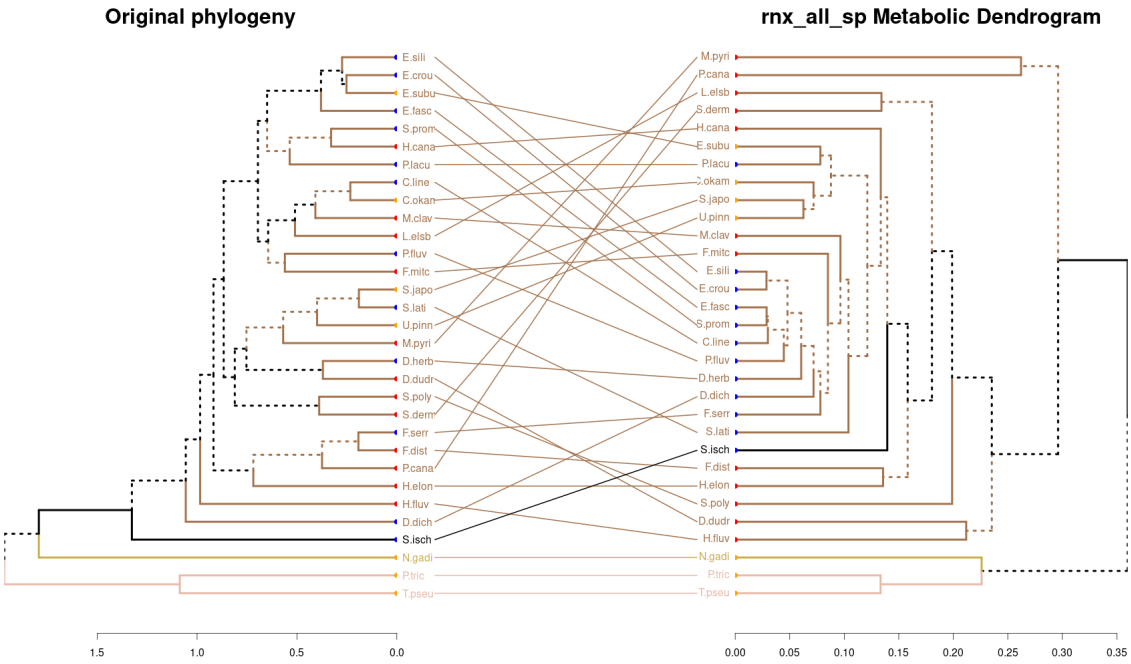

Figure 10. Tanglegramme sur les 33 espèces. À gauche le dendrogramme représentant la phylogénie originale des espèces et à droite le dendrogramme métabolique basé sur les réactions (représenté isolé avant l'application de l'algorithme d'optimisation de rotation des branches Figure 18 en Annexes). Pour les branches : en marron les algues brunes, en jaune les diatomées, en rose *N. gaditana* et en noir *S. ischiensis*. Pour les feuilles : en bleu les génomes LR, en rouge les génomes SR et en orange les génomes publics.

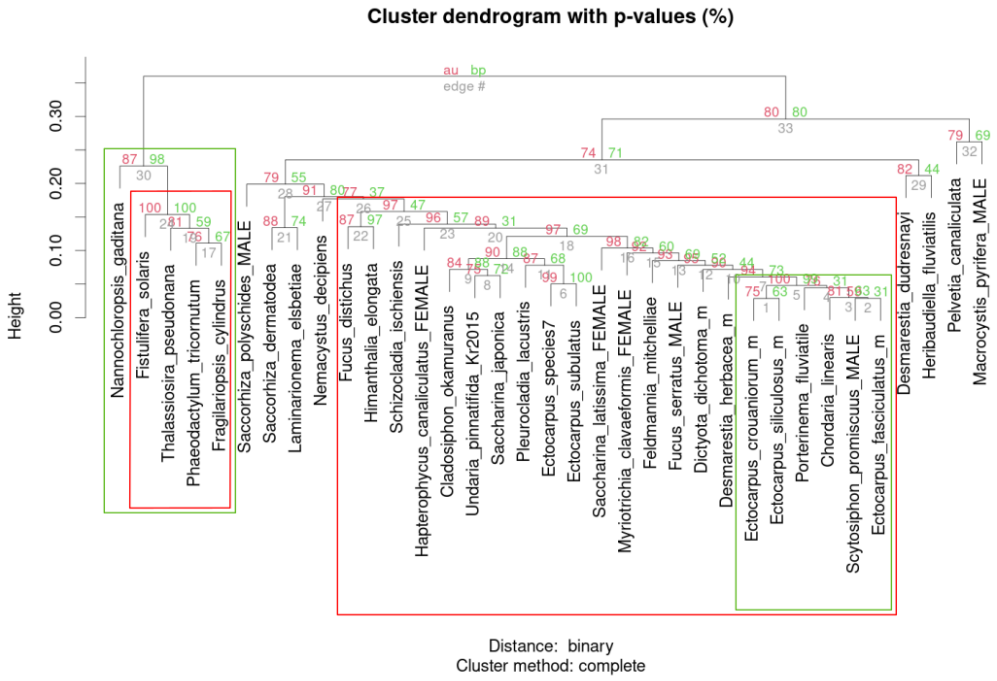

Figure 11. Dendrogramme sur les 35 espèces généré par *pvclust*. En rouge est indiqué la AU p-value (valeurs de probabilité approximativement sans biais) et en vert la BP value (fréquence d'apparition d'un groupe parmi les réplicats de dendrogrammes). Les rectangles verts désignent les grands groupes (plus de 3 espèces) ayant une BP value supérieure à 95%. Les rectangles rouges désignent les grands groupes (plus de 3 espèces) ayant une AU p-value supérieure à 95%

Le regroupement hiérarchique sur les données métaboliques a été généré avec un nombre total de réplicats de 10 000. Le dendrogramme de référence utilisé dans les comparaisons et considéré comme s'approchant au mieux de la réelle phylogénie des espèces correspond à un dendrogramme des protéomes prédits construit par analyse d'orthologie. Il comprend toutes les espèces de l'étude sauf deux espèces de l'extra-groupe : les diatomées *Fistulifera solaris* et *Fragilariopsis cylindrus*. Les comparaisons se feront donc sur 33 espèces sur 35.

La génération du dendrogramme métabolique basé sur la présence / absence de réactions chez les 35 espèces étudiées permet de relever plusieurs points. Le regroupement hiérarchique permet bien de séparer l'extra-groupe de l'ensemble des algues brunes sauf pour *Schizocladia ischiensis* qui devrait normalement se positionner en position basale du cluster des algues brunes. Ces résultats peuvent s'observer à partir de la Figure 18 en Annexes qui correspond au dendrogramme de droite sur la Figure 10 avant l'exécution de l'algorithme *untangle* réagençant les branches. Pour la comparaison du dendrogramme avec la phylogénie (Figure 10) $cor_{coph} = 0.72$ et $cor_{gbak} = 0.59$, donc il y a une meilleure adéquation entre les dendrogrammes par rapport à la longueur des branches (donc la distance entre les espèces) que par rapport à l'agencement des feuilles (donc le positionnement hiérarchique relatif des espèces).

La figure du dendrogramme généré par *pvclust* sans mise en forme est représenté Figure 11. Cette figure permet d'afficher les mesures d'incertitudes calculées par le package. Ces indicateurs ne sont pas affichés sur le dendrogramme mis en forme consultable Figure 18 en Annexe par soucis de lisibilité. Ces valeurs renseignent sur le fait que, pour ce regroupement hiérarchique, les grands groupes ayant la valeur BP la plus élevée (supérieure à 95% soit moins de 5% d'incertitude) sont : celui des extra-groupes et celui de l'ordre des Ectocarpales du groupe LR (à l'exception de *Porterinema fluviatile*). Pour les groupes ayant la valeur de probabilité AU la plus élevée (supérieure à 95% soit moins de 5% d'incertitude), ce sont celui des Diatomées au sein de l'extra-groupe et un large groupe d'algues brunes comprenant toutes les espèces LR et les espèces publiques. Cependant, ce groupe ne comprend que cinq espèces SR et comprend aussi *Schizocladia ischiensis* qui n'est pas une algue brune. Les sept autres espèces SR exclues de ce groupe ne correspondent pas exactement aux deux tiers dont une partie du génome a été ablatée et comprend *Laminarionema elsbetiae*. Il est cependant à noter que, parmi les cinq espèces SR faisant partie du groupe, trois sont des Ectocarpales et sont à une position moins basale dans le groupe que les deux dernières. De plus, elles correspondent aux trois espèces SR ayant le plus faible nombre de contigs de base (voir Figure 4). Il peut être intéressant de noter pour la suite que la dernière et donc quatrième Ectocarpale du groupe SR est *Laminarionema elsbetiae*. En effet, il s'agit de la seule espèce de son ordre à se trouver aussi isolée des autres espèces de son ordre.

Pour finir, ce qui peut être intéressant de relever, au sein du groupe des algues brunes sur la Figure 18 en Annexes, est qu'elles semblent former des groupes distincts en fonction de leur moyen de séquençage (à quelques exceptions près) : LR ou SR.

3.5 Analyse des réseaux selon leur méthode de séquençage

À partir de l'observation de l'analyse précédente concernant le regroupement des données séquencées de manière similaire, j'ai effectué des analyses plus précises sur les deux

Figure 12. Diagramme de Venn des réactions parmi les 2 groupes LR et SR.

groupes LR et SR. Dans un premier temps, j'ai cherché à savoir si la génération de dendrogrammes ne comprenant que les espèces d'un seul groupe permettait d'améliorer ou de détériorer les coefficients de corrélations avec la phylogénie par rapport au dendrogramme comprenant toutes les espèces. Dans un second temps, j'ai comparé la distribution du nombre de réactions, gènes, voies métaboliques complétées entre les deux groupes à la recherche de différences significatives.

En observant le dendrogramme généré, la formation d'un groupe LR semble cependant plus robuste que celui d'un groupe SR dont les espèces sont plus dispersées en petits groupes. En effet, les espèces SR se rassemblent dans une position basale du dendrogramme, mais n'appartiennent pas à un groupe commun. À l'inverse, pour ce qui est des espèces LR ou des espèces publiques, il peut être observé le rassemblement d'une majorité de leurs espèces en un groupe commun. Il a donc été vérifié si cette génération de regroupement hiérarchique en ne sélectionnant que les génomes de séquençage similaire améliorait la concordance des dendrogrammes. En effet, pour les 12 espèces LR il est obtenu les coefficients $cor_{coph} = 0.73$ (+0.01) et $cor_{gbak} = 0.64$ (+0.05) ce qui ne suggère pas de changement significatif au niveau de la longueur des branches, mais on note une amélioration de l'agencement des feuilles. Les deux algues *Pleurocladia lacustris* et *Saccharina latissima* semblent être les deux se groupant le moins bien au sein de ce groupe. Cependant, l'analyse sur les 12 génomes SR donne des coefficients de $cor_{coph} = 0.24$ (-0.48) et $cor_{gbak} = 0.21$ (-0.38) ce qui montre que le dendrogramme métabolique au sein des SR ne concorde que peu avec le dendrogramme de référence, que ce soit au niveau de la longueur de branche ou que ce soit au niveau de l'agencement des feuilles.

Le dendrogramme étant basé sur la présence / absence des réactions dans les GSMN des espèces, il a été vérifié si une différence significative du nombre de réactions retrouvées en fonction des espèces expliquerait leur mauvais positionnement dans la hiérarchie du dendrogramme.
Entre les deux groupes LR et SR, il a pu être observé que l'union des réactions des espèces du groupe SR contenait plus de réactions que dans l'union des réactions des LR. En effet, sur 4825 réactions, 75.96% sont en commun entre les deux groupes. Les SR en comptent 99.32% et 24.04% n'appartiennent qu'à leur groupe. En moindre mesure, les LR en comptent 75.96% et seulement 0.68% n'appartiennent qu'à leur groupe (Figure 12).

Il a ensuite été regardé plus en détail la répartition du nombre de réactions chez chaque espèce de chaque groupe. Les résultats montrent une moyenne plus élevée de ce nombre chez les SR ($\mu = 3346$) que chez les LR ($\mu = 3279$) mais aussi que l'écart type s'y trouvait être sept fois supérieur ($\sigma = 263$ chez les SR et $\sigma = 38$ chez les LR). Cette différence peut être remarquée

clairement sur la Figure 13, le groupe SR contient des espèces avec beaucoup plus de gènes et de réactions par rapport au groupe LR mais aussi beaucoup moins. Par ailleurs, la complétion des voies métaboliques est en moyenne plus importante chez les LR que chez les SR, ce qui peut laisser à supposer que l'ensemble des réactions supplémentaires présentes uniquement chez les SR ne soient pas forcément très robustes et constitueraient des réactions isolées et réparties dans diverses voies métaboliques.

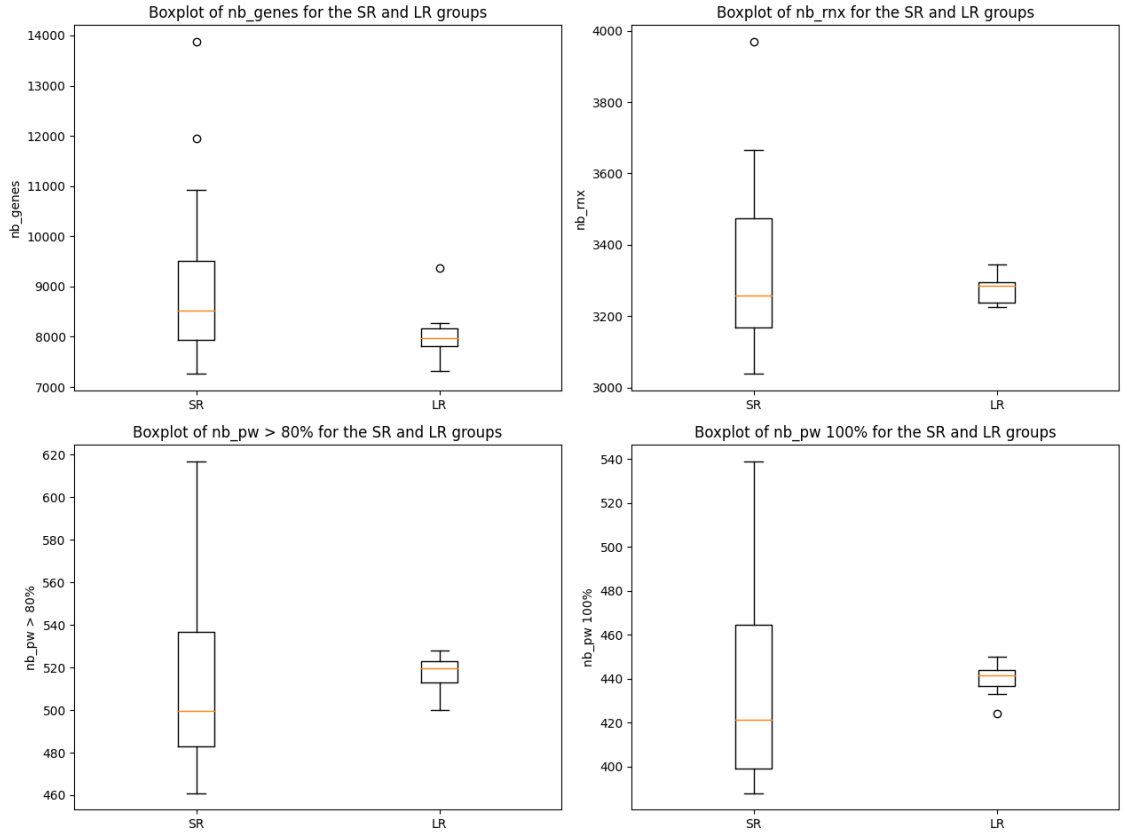

Figure 13. Boxplots comparant le nombre de gènes, le nombre de réactions, le nombre de voies métaboliques complétées à au moins 80% et le nombre de voies métaboliques complétées à 100% chez les LR et SR.

Pour étendre cette analyse en considérant cette hypothèse de réactions dont leur réelle présence dans l'organisme est contestable, une comparaison du nombre de réactions entre les deux groupes a, cette fois-ci, était réalisée sur la base d'un socle de réactions considérées comme très conservées au sein des algues brunes. Ce socle de réactions a été choisi comme l'ensemble des réactions présentes chez au moins 80% de l'ensemble des algues brunes.
Les résultats de cette analyse montrent que sur la base de cet ensemble de réactions, ce sont les LR qui en comptent le plus sur une majorité des espèces. En effet, comme observable Figure 14, la totalité des espèces LR comptent plus de réactions conservées que la moitié des espèces SR (valeur minimale du groupe LR exactement égale à la valeur médiane du groupe SR qui est de 2961 réactions).

Malgré l'hétérogénéité des données relevée, cela n'a pas invalidé la possibilité de leur exploitation pour dégager des pistes d'analyses biologiques. Des recherches ont donc été effectuées dans l'objectif de réussir à caractériser des particularités de l'endophyte *Laminarionema elsbetiae* supposées liées à son mode de vie.

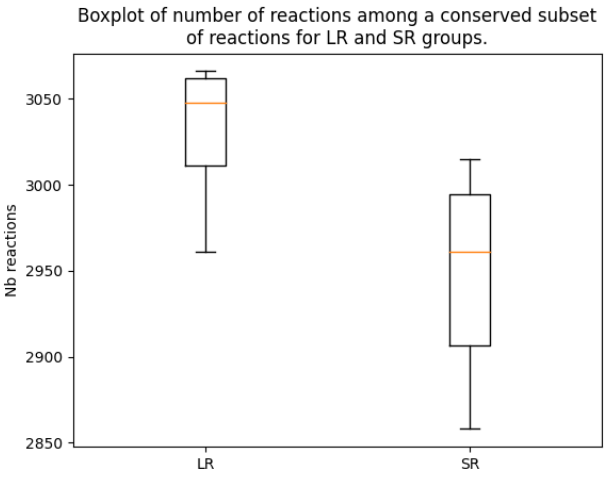

Figure 14. Boxplots comparant le nombre de réactions chez les LR et SR sur la base des réactions présentes chez au moins 80% des algues brunes.

3.6 Analyse du GSMN de *Laminarionema elsbetiae* pour comprendre l'impact de son mode vie endophytique sur son métabolisme

L'objectif de cette partie a donc été de réduire l'échelle de l'ensemble des algues brunes à une algue brune précise : *Laminarionema elsbetiae*. J'ai donc cherché à voir si des différences entre cette espèce et les autres algues brunes ressortaient des données métaboliques. Certaines différences significatives pourraient permettre d'ouvrir des pistes d'études permettant de mieux comprendre le fonctionnement biologique de cet endophyte.

L'étude du GSMN de *Laminarionema elsbetiae* comparé à ceux des autres algues brunes vise à conforter ou non l'hypothèse selon laquelle certaines réactions appartenant à un cœur métabolique défini des algues brunes seraient absentes chez *Laminarionema elsbetiae*. L'absence de ces réactions du cœur métabolique s'expliquerait par le fait que, étant donné son mode de vie endophytique, *Laminarionema elsbetiae* aurait perdue des réactions inutiles à son bon fonctionnement, car compensées par son hôte. Cette recherche s'est donc, dans un premier temps, réalisée à l'échelle des réactions isolées puis à l'échelle des voies métaboliques.

Le cœur métabolique de réactions a, pour cette étude, été défini par l'ensemble des réactions présentes chez toutes les algues brunes sauf une. Pour retrouver l'ensemble de ces réactions absentes chez *Laminarionema elsbetiae*, l'ensemble des réactions de ce cœur métabolique a dans un premier temps été extrait. À partir de cet ensemble, ont été sélectionnées les réactions absentes précisément chez *Laminarionema elsbetiae*. L'ensemble des réactions supposées perdues a permis de révéler l'absence de trois réactions appartenant à la voie des oxylipines chez *Laminarionema elsbetiae*. Cette absence s'est observée à partir d'une plus ancienne exécution d'AuCoMe que la finale (comprenant les espèces LR, les espèces publiques, l'extra-groupe et uniquement *Laminarionema elsbetiae* comme espèce SR). Des analyses visant à consolider l'hypothèse que ces réactions sont manquantes chez *Laminarionema elsbetiae* ont été effectuées par deux étudiants, Enora Corre et Garan Le Biwig, dont la présentation de leur travail est visible en Annexes A.3.

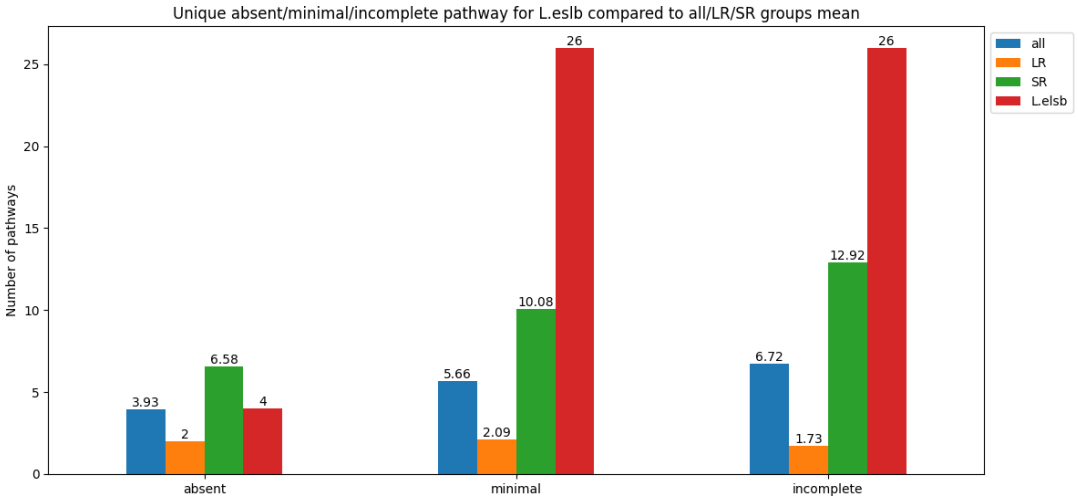

Figure 15. Barplot montrant le nombre de voies métaboliques absentes, de complétion minimale et incomplètes de manière unique chez chaque espèce. Sont comparés : ces nombres chez *Laminarionema elsbetiae* ainsi que leur moyenne entre toutes les espèces, les espèces du groupe SR et les espèces du groupe LR.

Il a ensuite été regardé (à partir de l'exécution finale des 35 espèces), si le nombre de réactions absentes du cœur métabolique était plus important chez *Laminarionema elsbetiae* que chez les autres algues brunes. Sur l'ensemble des réactions, 25 sont retrouvées absentes chez *Laminarionema elsbetiae* pour une moyenne de $\mu = 15,59$ chez l'ensemble des algues brunes et un écart type de $\sigma = 14.82$. Elle se retrouve à la 6$^{\text{ème}}$ position (sur 29) des algues brunes ayant le plus de réactions absentes. Cependant, comme montré précédemment Figure 14, les espèces SR possèdent moins de réactions très conservées, donc plus de réactions absentes du cœur métabolique. En effet, chez les LR la moyenne des réactions perdues est de $\mu = 6.09$ avec un écart type de $\sigma = 6.92$. Chez les SR la moyenne des réactions perdues est de $\mu = 25,67$ avec un écart type de $\sigma = 17.05$.

Pour ce qui est des voies métaboliques, il a été regardé combien de voies métaboliques étaient absentes, de complétion minimale et incomplètes chez *Laminarionema elsbetiae* comparé aux autres algues brunes. Pour ce qui est du nombre de voies métaboliques uniquement absentes chez *Laminarionema elsbetiae*, il y en a 4. Par comparaison aux autres algues, ce nombre est presque égal à sa moyenne pour toutes les espèces ($\mu_{all} = 3.93$) et inférieur à sa moyenne uniquement chez les espèces SR ($\mu_{sr} = 6.58$) (Figure 15). Cependant, pour ce qui est du nombre de voies métaboliques de complétion minimale uniquement chez *Laminarionema elsbetiae* et du nombre de voies métaboliques uniquement incomplètes chez *Laminarionema elsbetiae*, ce nombre et beaucoup plus important chez *Laminarionema elsbetiae* que chez les autres espèces. En effet, il est observable Figure 15, qu'il y a 26 voies métaboliques de complétion minimale uniquement chez *Laminarionema elsbetiae* contre une moyenne de $\mu_{all} = 5.66$ chez toutes les espèces et de $\mu_{sr} = 10.08$ chez les espèces SR. Dans le même ordre de grandeur, il est observé 26 voies métaboliques incomplètes uniquement chez *Laminarionema elsbetiae* contre une moyenne de $\mu_{all} = 6.72$ chez toutes les espèces et de $\mu_{sr} = 12.97$ chez les espèces SR.

3.7 Affinage des données pour générer les dendrogrammes métaboliques

Pour finir, considérant l'hétérogénéité des données, j'ai essayé de trouver un moyen de réduire les différences présentes entre les différents groupes. L'objectif a été de vérifier si l'application d'un filtre sur les réactions ou voies métaboliques permettait d'améliorer les résultats.

Afin de pallier l'hétérogénéité des génomes des espèces et de leurs GSMN associés, il a été essayé de régénérer les dendrogrammes en modifiant les données d'entrée pour voir si cela avait une influence sur sa qualité basée sur sa concordance avec la phylogénie de référence.
L'objectif est de trouver un moyen de filtrer les données pour enlever les informations qui semblent les plus improbables sans pour autant appliquer un filtre trop stringent. En effet, les données doivent rester suffisamment différenciables pour pouvoir former les clusters. Pour cela, plusieurs pistes ont été essayées comme :

- filtrer les réactions en supprimant celles présentes chez au maximum un certain pourcentage d'espèces des espèces. (Figure 16 **a.**)
- générer les dendrogrammes sur la base de différentes complétions des voies métaboliques. (Figure 16 **b.**)
- filtrer les voies métaboliques pour ne garder que celles possédant au moins un certain nombre de réactions au total. (Figure 16 **c.**)

Il peut être observé Figure 16 les différents coefficients de corrélation obtenus en appliquant les différentes approches de modification des données d'entrée pour le regroupement hiérarchique. Cependant, aucune de ces approches ne permet de conclure sur une amélioration significative de la concordance des dendrogrammes. Les coefficients cor_{coph} et cor_{gbak} ont tendance à se rapprocher encore plus de 0 dans les cas **b.** et **c.** agissant au niveau des voies métaboliques. Dans le cas **a.** où le filtre a été appliqué sur les réactions, il est observé une amélioration du coefficient de corrélation cophénétique mais une dégradation du coefficient de corrélation du Gamma de Baker. Ce résultat suppose que le filtre des réactions présentes chez peu d'espèces permet d'améliorer la concordance des longueurs de branches, mais cela détériore la qualité de leur agencement.

4 Discussion

4.1 Le mauvais placement de certaines espèces dans les dendrogrammes métaboliques

Il a pu être observé que les GSMN obtenus, suite à l'utilisation de l'outil de reconstruction AuCoMe, ne permettent pas un positionnement fidèle des espèces sur le dendrogramme métabolique comparé à leur positionnement phylogénétique.

4.1.1 L'impact de la qualité des données génomiques et de la composition du jeu de données

L'étude de Schulz et Almaas avait permis de conclure sur des phénomènes qui permettaient d'expliquer le mauvais positionnement de certains organismes. Ils ont observé que les espèces

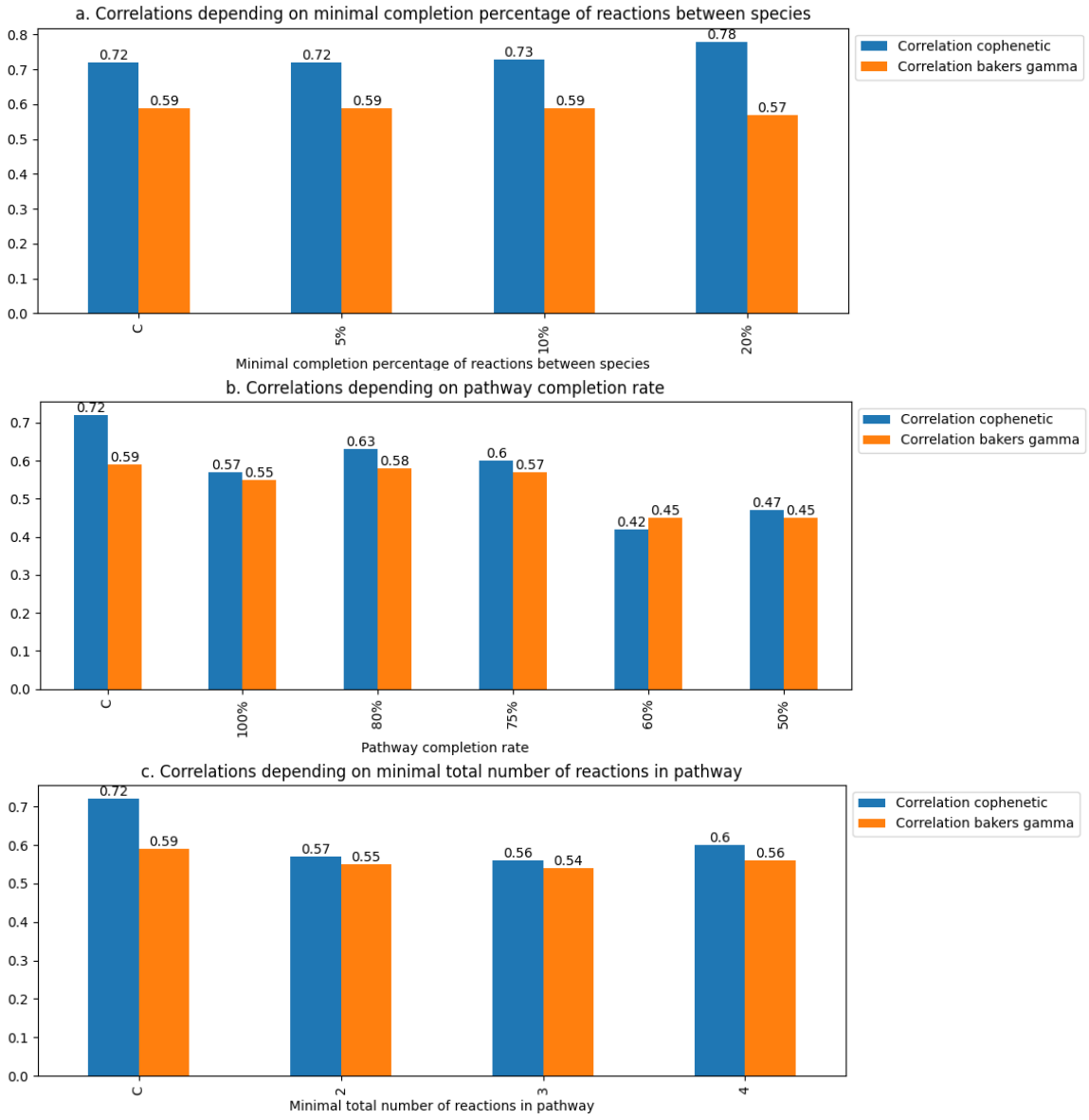

Figure 16. Coefficients de corrélations cophénétiques (bleu) et du Gamma de Baker (orange) dans les différents dendrogrammes générés avec les données suivantes : **a.** Filtrage des réactions présentes chez au minimum 5%, 10% et 20% des espèces, **b.** Base de complétions de voies métaboliques de 100%, 80%, 75%, 60% et 50%, **c.** Filtrage des voies métaboliques contenant au minimum 2, 3 ou 4 réactions au total sur la base de complétion de 100%. Le ”C” représente le contrôle : correspond aux valeurs obtenues sur le dendrogramme métabolique obtenu partie 3.4.

mal placées pouvaient s'expliquer par leurs annotations incomplètes ou significativement manquantes. Ils ont aussi observé que les organismes présentant trop de réactions spécialisées et trop peu de réactions plus communément partagées se retrouvées mal placées [17]. De plus, la réduction de leur jeu de données de 975 à 21 espèces montrait aussi une augmentation du nombre d'espèces mal placées.

Dans l'étude présentée ici, la forte hétérogénéité de la qualité des données est donc un facteur important dans le mauvais positionnement des espèces. De plus, la taille du jeu de données est assez faible (35 espèces) et se concentre sur des espèces très proches. À cette échelle, les éléments permettant de discriminer les espèces entre elles lors du regroupement hiérarchique sont plus subtiles qu'à la large échelle des trois domaines : eucaryotes, archées et bactéries. De ce fait, il n'est pas spécialement surprenant que des réseaux reconstruits automatiquement sans aucuns traitements a posteriori, de plus sur des données n'étant pas toutes de grande qualité, n'aboutissent pas à la génération d'un regroupement hiérarchique de qualité. Il peut ensuite être relevé que l'ordre des Ectocarpales semblait réussir à s'attirer au sein d'un groupe et compte la majorité des espèces (48%). Il peut donc être supposé qu'un jeu de données ajoutant beaucoup plus d'espèces des autres ordres de manière à équilibrer leur représentativité pourrait permettre d'au moins former des groupes par ordre lors de regroupement hiérarchique.

Il est cependant important de préciser que l'outil AuCoMe est un outil de reconstruction automatique. De ce fait, il permet d'obtenir une base de résultats qui nécessiterait une curation experte réseau par réseau pour être plus exhaustive. Il est aussi important d'ajouter que AuCoMe dépend de la richesse des annotations disponibles sur la base de donnée MetaCyc. De plus, l'ordre des algues brunes, comprend beaucoup d'espèces peu étudiées et dont l'annotation mériterait de gagner en expertise. Pour finir, il serait aussi intéressant d'avoir à disposition des données plus homogènes en qualité de séquençage sur la base de celle des LR. L'ensemble de ces aspects, amenés à évoluer dans le temps, pourrait permettre la génération de meilleurs résultats lors d'une étude similaire.

4.1.2 L'impact de la nature du jeu de données

Ici les données utilisées pour la génération des dendrogrammes ont été des données binaires renseignant sur la présence / absence de réactions. Elles ont aussi été des données binaires renseignant sur le dépassement / non-dépassement d'un seuil de complétion de voies métaboliques.

Les données se basant sur les voies métaboliques sont donc la traduction de proportions (appartenant à l'intervalle [0, 1]) en des données binaires (appartenant à l'ensemble $\{0, 1\}$). Cette traduction implique donc une perte de précision des informations. La génération de dendrogrammes à partir des données des proportions, aurait pu permettre de pouvoir plus finement discriminer les espèces entre elles lors du regroupement hiérarchique.

Pour finir, il aurait aussi pu être intéressant de comparer les deux approches précédentes (base de réactions et base de voies métaboliques) à une approche basée sur la présence / absence de métabolites.

4.1.3 L'impact du calcul de la matrice de distance du regroupement hiérarchique

En plus des données, la méthode de génération du dendrogramme va aussi influer sur les résultats obtenus. La distance utilisée pour dans cette étude pour générer le dendrogramme peut aussi être remise en question. La distance utilisée pour calculer la matrice de distance utile à la création du dendrogramme est la distance de Jaccard. Cette distance, ne prenant en compte comme similarité que la présence commune d'un élément, il a été discuté de la remplacer par une autre méthode de distance prenant en compte comme similarité toujours les présences communes, mais aussi les absences communes. Or, cette méthode de distance n'est pas disponible parmi les méthodes de distances implémentables dans l'outil de création de dendrogramme de *pvclust*. Les recherches pour implémenter cette distance sur une création de dendrogramme par bootstrap constituent une perspective d'amélioration des résultats obtenus.

4.2 L'hétérogénéité importante du nombre de gènes et de réactions

L'ensemble des espèces SR ayant un nombre de réactions inférieur à la moyenne pourrait s'expliquer par la fragmentation du génome et donc l'incapacité à pouvoir y prédire certains gènes qui seraient coupés et dispersés sur plusieurs contigs. Le sur-enrichissement en nombre de réactions est par contre plus énigmatique, une piste avancée serait la présence de contaminants dans le génome. Ces hypothèses restent à confirmer par des analyses plus fines.

4.3 Les pertes de gènes chez *Laminarionema elsbetiae*

Dans un premier temps, la définition du cœur métabolique des algues brunes peut être remise en cause. Le choix de sélectionner les réactions retrouvées chez toutes les algues brunes sauf une, permet en effet d'obtenir un ensemble de réactions sans doute très conservées au sein de la classe. Il est cependant important de noter que l'algue brune de référence étant *Ectocarpus species7*, les espèces proches de celle-ci risquent d'être plus facilement enrichies en réactions que les espèces plus éloignées. De plus, la définition choisie du cœur métabolique ne prend pas en considération la position phylogénétique des espèces. En effet, la position phylogénétique des espèces peut avoir une influence sur ce cœur métabolique, les algues aux positions les plus basales sont les plus susceptibles d'avoir subi des pertes ou gains de gènes en évoluant. Une autre définition plus affinée du cœur métabolique pourrait permettre de révéler des résultats plus pertinents sur les suppositions de pertes de gènes chez les espèces.

Malgré le fait que *Laminarionema elsbetiae* ait un nombre de réactions absentes du cœur métabolique supérieur à la moyenne de ce nombre chez l'ensemble des algues brunes, cela ne permet pas de conclure sur une réelle absence biologique. En effet, en considérant séparément les SR et les LR son nombre de réactions absentes se trouve être égal à celui de la moyenne du groupe SR. De ce fait, il est donc difficile de conclure sur l'origine des réactions absentes chez *Laminarionema elsbetiae*. Elles peuvent être liées à une réelle absence biologique, mais sont peut-être juste causées par sa qualité de séquençage. Pour ce qui est de l'étude des pertes à l'échelle des voies métaboliques, elle ne laisse pas percevoir un nombre important de voies métaboliques absentes chez *Laminarionema elsbetiae* et présentes chez les autres algues brunes. Cependant, il est observé un nombre important de voies métaboliques de complétion minimale ou incomplètes particulièrement chez *Laminarionema elsbetiae*. Si ces voies incomplètes ne

sont pas causées par une absence de l'annotation, il pourrait s'agir de voies métaboliques au sein desquelles *Laminarionema elsbetiae* a perdu des réactions, n'ayant pas l'utilité des fonctions métaboliques qu'elles remplissent. Dans ce cas, ces résultats peuvent ouvrir des pistes de réactions potentiellement perdues à étudier plus précisément.

Ici encore, l'hétérogénéité de la qualité des données permet seulement de dégager des hypothèses, mais pas de les confirmer. Des analyses biologiques de génétique inverse (CRISPR-Cas9 [2] à partir de l'organisme modèle *Ectocarpus species7*) pourraient cependant permettre de valider ou non la réelle perte de réactions (et donc du gène associé) parmi celles sélectionnées par les analyses du réseau métabolique d'une espèce. L'amélioration de la qualité des données, permettrait ici de consolider les hypothèses de départ avant d'engager ces procédures de validations biologiques.

5 Conclusion et perspectives

Ce stage m'a permis d'enrichir mes connaissances dans le domaine précis de la biologie des algues. De plus, j'ai pu apprendre à utiliser des outils développés par l'équipe Dyliss et d'autres outils informatiques, comme Docker, que je n'avais jamais utilisés.

Pour ce stage, j'ai pu apporter plusieurs contributions. La première a été de générer les fichiers d'entrée à utiliser, ce qui a permis un enrichissement du code du package utilisé. Une fois ces fichiers générés, j'ai appris à utiliser divers outils informatiques et j'ai dû surmonter les problèmes techniques rencontrés, notamment en apportant des modifications aux fichiers d'entrée. Une fois ces outils maîtrisés, je suis parvenue à reconstruire les réseaux métaboliques de toutes les espèces du jeu de données. Ayant eu une certaine autonomie, j'ai pris l'initiative de développer différents scripts en R et en python pour analyser les réseaux créés. Pour finir, j'ai exploré plusieurs pistes d'interprétations à partir des résultats obtenus, dont principalement une analyse sur l'hétérogénéité des données et ses répercussions. J'ai aussi contribué à générer des données utilisées pour des études précises de *Laminarionema elsbetiae*.

Les résultats obtenus ont permis de dégager plusieurs pistes d'analyses, celles choisies ont donc été celles présentées dans ce rapport. En perspectives, il pourrait être intéressant d'analyser les aspects non étudiés, comme la caractérisation détaillée des fonctions métaboliques des algues brunes, cela en décrivant les rôles biologiques des voies métaboliques propres aux algues brunes. Ayant développé les outils d'analyse seule, une révision et amélioration du code par des personnes avec une meilleure expertise serait sûrement nécessaire.

Le sujet traité lors de ce stage est la première étude où sont exploités 29 GSMN d'algues brunes (+6 espèces constituant l'extra-groupe). J'ai évalué les limites d'une étude dans laquelle sont couplées différentes techniques de séquençage pour reconstruire automatiquement des GSMN. Cette étude est donc une première approche pour ensuite pouvoir étudier les GSMN d'algues à plus grande échelle. À partir de ce travail, il sera possible par la suite d'augmenter le nombre de génomes à étudier et d'améliorer la qualité des GSMN que j'ai obtenus.

Références

[1] Méziane Aite et al. « Traceability, reproducibility and wiki-exploration for "à-la-carte" reconstructions of genome-scale metabolic models ». en. In : *PLOS Computational Biology* 14.5 (mai 2018). Sous la dir. de Jens Nielsen, e1006146. doi : 10.1371/journal.pcbi.1006146. url : https://dx.plos.org/10.1371/journal.pcbi.1006146.

[2] Yacine Badis et al. « Targeted CRISPR□Cas9□based gene knockouts in the model brown alga Ectocarpus ». en. In : *New Phytologist* 231.5 (2021), p. 2077-2091. doi : 10.1111/nph.17525. url : https://onlinelibrary.wiley.com/doi/10.1111/nph.17525.

[3] Frank B. Baker. « Stability of Two Hierarchical Grouping Techniques Case 1 : Sensitivity to Data Errors ». In : *Journal of the American Statistical Association* 69.346 (1974), p. 440. doi : 10.2307/2285675. url : https://www.jstor.org/stable/2285675?origin=crossref.

[4] Arnaud Belcour et al. « AuCoMe : inferring and comparing metabolisms across heterogeneous sets of annotated genomes ». en. In : (2022 (en préparation)).

[5] Arnaud Belcour et al. « Metage2Metabo, microbiota-scale metabolic complementarity for the identification of key species ». In : *eLife* 9 (déc. 2020). Sous la dir. de María Mercedes Zambrano et al., e61968. doi : 10.7554/eLife.61968. url : https://doi.org/10.7554/eLife.61968.

[6] Miriam Bernard et al. « qPCR-based relative quantification of the brown algal endophyte Laminarionema elsbetiae in Saccharina latissima : variation and dynamics of host—endophyte interactions ». en. In : *Journal of Applied Phycology* 30.5 (2018), p. 2901-2911. doi : 10.1007/s10811-017-1367-0. url : http://link.springer.com/10.1007/s10811-017-1367-0.

[7] Miriam S. Bernard et al. « Diversity, biogeography and host specificity of kelp endophytes with a focus on the genera Laminarionema and Laminariocolax (Ectocarpales, Phaeophyceae) ». en. In : *European Journal of Phycology* 54.1 (jan. 2019), p. 39-51. doi : 10.1080/09670262.2018.1502816. url : https://www.tandfonline.com/doi/full/10.1080/09670262.2018.1502816.

[8] Trevor T. Bringloe et al. « Phylogeny and Evolution of the Brown Algae ». In : *Critical Reviews in Plant Sciences* 39.4 (3 juill. 2020), p. 281-321. issn : 0735-2689. doi : 10.1080/07352689.2020.1787679. url : https://doi.org/10.1080/07352689.2020.1787679 (visité le 24/01/2022).

[9] E. Ellertsdottir et A. F. Peters. « High prevalence of infection by endophytic brown algae in populations of Laminaria spp. (Phaeophyceae) ». In : *Oceanographic Literature Review* 7.44 (1997), p. 740. url : https://www.infona.pl//resource/bwmeta1.element.elsevier-ad8e9f4a-2b2f-30cf-87be-bd3c0caf437e.

[10] David M. Emms et Steven Kelly. « OrthoFinder : phylogenetic orthology inference for comparative genomics ». In : *Genome Biology* 20.1 (nov. 2019), p. 238. doi : 10.1186/s13059-019-1832-y. url : https://doi.org/10.1186/s13059-019-1832-y.

[11] David M. Emms et Steven Kelly. « OrthoFinder : solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy ». In : *Genome Biology* 16.1 (août 2015), p. 157. doi : 10.1186/s13059-015-0721-2. url : https://doi.org/10.1186/s13059-015-0721-2.

[12] Tal Galili. « dendextend : an R package for visualizing, adjusting and comparing trees of hierarchical clustering ». en. In : *Bioinformatics* 31.22 (nov. 2015), p. 3718-3720. doi : 10.1093/bioinformatics/btv428. url : https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv428.

[13] Changdai Gu et al. « Current status and applications of genome-scale metabolic models ». en. In : *Genome Biology* 20.1 (juin 2019), p. 121. doi : 10.1186/s13059-019-1730-3. url : https://doi.org/10.1186/s13059-019-1730-3.

[14] Peter D. Karpe, Mario Latendresse et Ron Caspi. « The Pathway Tools Pathway Prediction Algorithm ». en. In : *Standards in Genomic Sciences* 5.3 (nov. 2011), p. 424-429. doi : 10.4056/sigs.1794338. url : https://doi.org/10.4056/sigs.1794338.

[15] Hiroshi Kawai et Masashi Tokuyama. « Laminarionema elsbetiae gen. et sp. nov. (Ectocarpales, Phaeophyceae), a new endophyte in Laminaria sporophytes ». In : *Phycological Research* 43.4 (1995), p. 185-190. doi : 10.1111/j.1440-1835.1995.tb00024.x. url : https://onlinelibrary.wiley.com/doi/10.1111/j.1440-1835.1995.tb00024.x.

[16] Alejandro E. Montecinos et al. « Species delimitation and phylogeographic analyses in the Ectocarpus subgroup siliculosi (Ectocarpales, Phaeophyceae) ». en. In : *Journal of Phycology* 53.1 (2017). Sous la dir. de M. Cock, p. 17-31. doi : 10.1111/jpy.12452. url : https://onlinelibrary.wiley.com/doi/10.1111/jpy.12452.

[17] Christian Schulz et Eivind Almaas. « Genome-scale reconstructions to assess metabolic phylogeny and organism clustering ». en. In : *PLOS ONE* 15.12 (déc. 2020). Sous la dir. de Zhong-Hua Chen, e0240953. doi : 10.1371/journal.pone.0240953. url : https://dx.plos.org/10.1371/journal.pone.0240953.

[18] Robert R. Sokal et F. James Rohlf. « THE COMPARISON OF DENDROGRAMS BY OBJECTIVE METHODS ». en. In : *TAXON* 11.2 (1962), p. 33-40. doi : 10.2307/1217208. url : https://onlinelibrary.wiley.com/doi/abs/10.2307/1217208.

[19] R. Suzuki et H. Shimodaira. « Pvclust : an R package for assessing the uncertainty in hierarchical clustering ». en. In : *Bioinformatics* 22.12 (juin 2006), p. 1540-1542. doi : 10.1093/bioinformatics/btl117. url : https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btl117.

[20] Rabindra Thakur, Takashi Shiratori et Ken-ichiro Ishida. « Taxon-rich Multigene Phylogenetic Analyses Resolve the Phylogenetic Relationship Among Deep-branching Stramenopiles ». en. In : *Protist* 170.5 (2019), p. 125682. doi : 10.1016/j.protis.2019.125682. url : https://linkinghub.elsevier.com/retrieve/pii/S1434461018300865.

[21] Qikun Xing et al. « Different Early Responses of Laminariales to an Endophytic Infection Provide Insights About Kelp Host Specificity ». In : *Frontiers in Marine Science* 8 (2021). url : https://www.frontiersin.org/article/10.3389/fmars.2021.742469.

A Annexes

A.1 Étapes pipeline AuCoMe

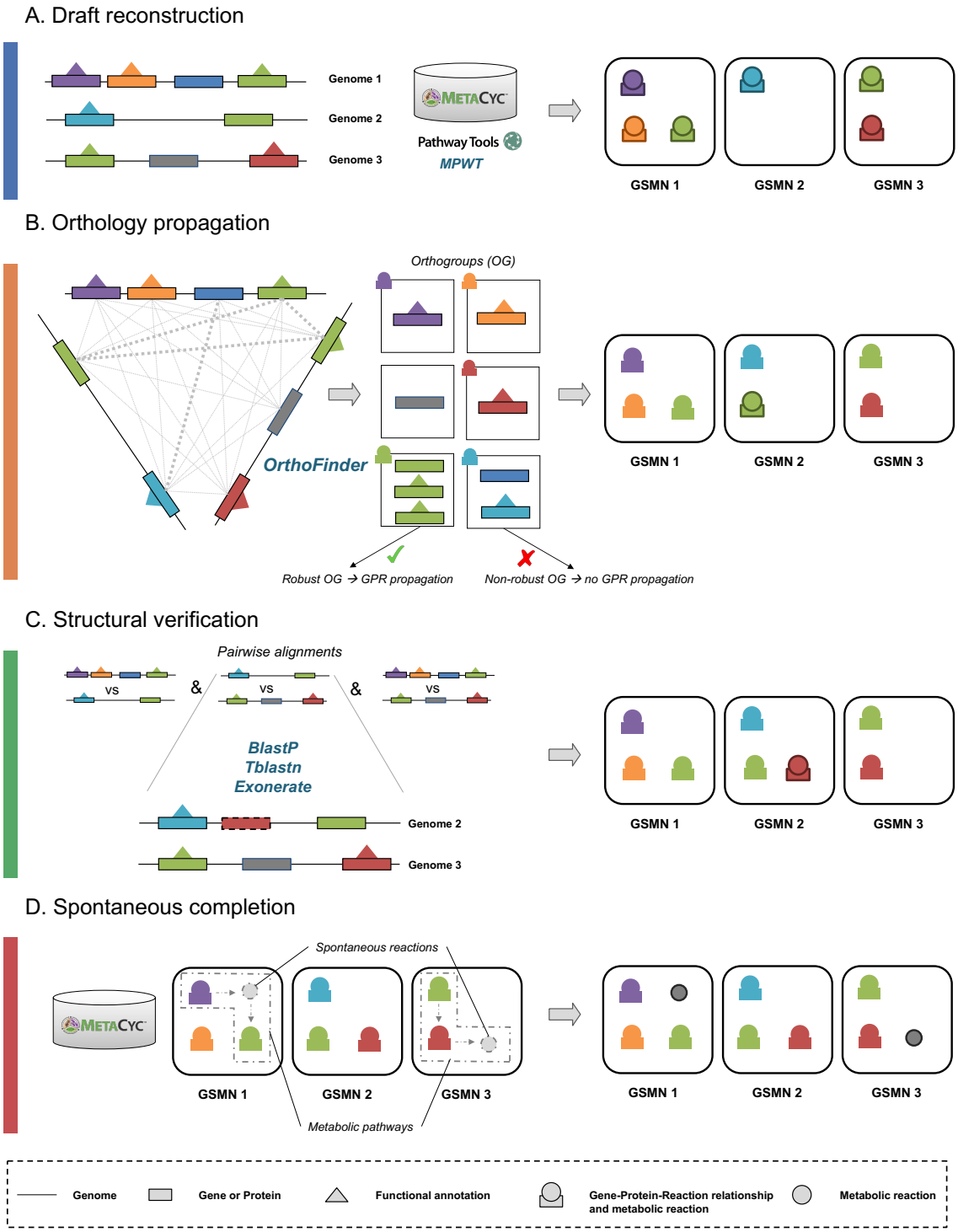

Figure 17. Étapes du pipeline de l'outil AuCoMe. Figure extraite de la publication des auteurs de l'outil [4].

A.2 Dendrogrammes

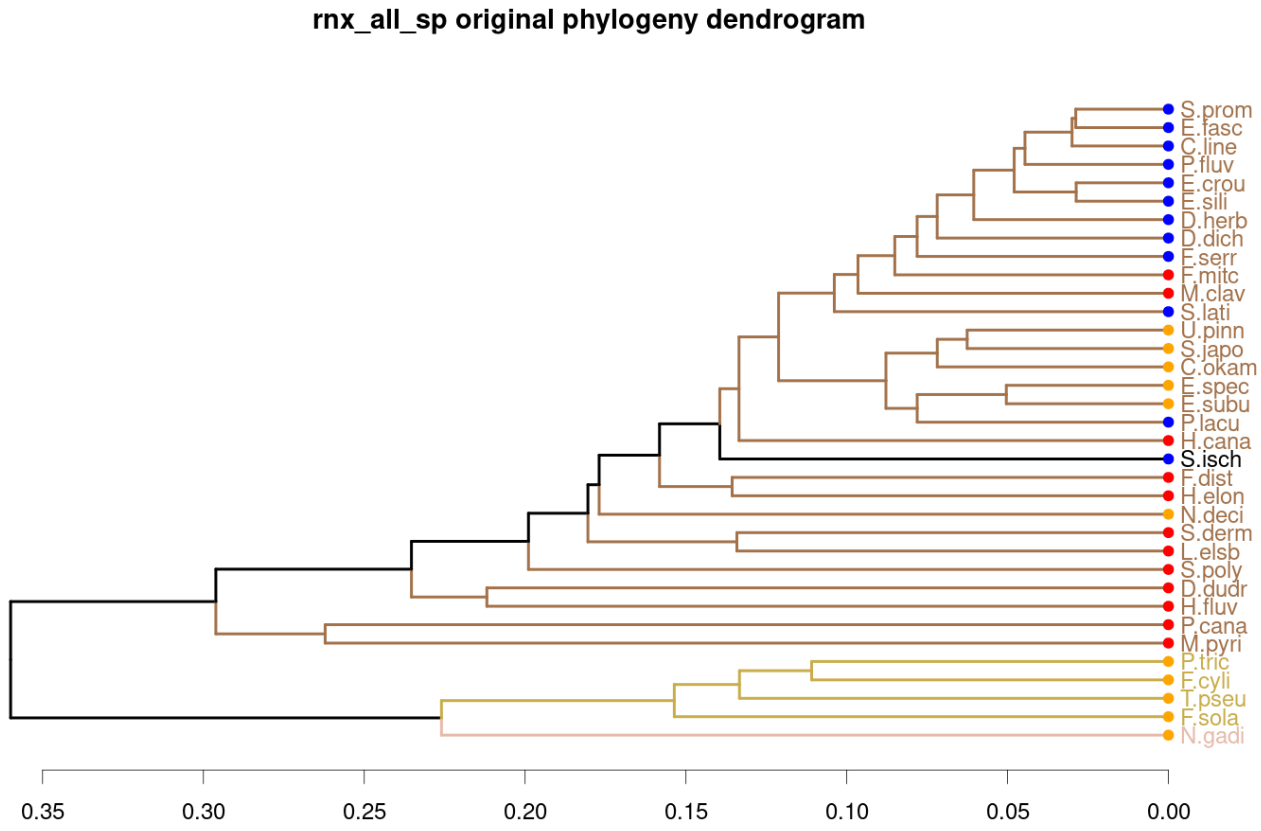

Figure 18. Dendrogramme des 35 espèces mis en forme par le package *dextend*. Pour les branches : en marron les algues brunes, en jaune les diatomées, en rose *N. gaditana* et en noir *S. ischiensis*. Pour les feuilles : en bleu les génomes LR, en rouge les génomes SR et en orange les génomes publics.

A.3 Poster étude des pertes de gènes chez *Laminarionema elsbetiae*

Document page suivante.

Analyse intégrative du réseau métabolique à l'échelle du génome de l'algue brune endophyte *Laminarionema elsbetiae*

CNRS • SORBONNE UNIVERSITÉ
Station Biologique de Roscoff

Enora Corre* (encorre@sb-roscoff.fr) **Garan Le Biwig*** (g.biwig@laposte.net) Encadrant : Gabriel Markov**
*Licence Biologie Modélisation et Analyse de Données, Sorbonne Université
**Algal Biology and Interactions with the environment UMR 8227, Station Biologique de Roscoff

SCIENCES SORBONNE UNIVERSITÉ

Introduction

Laminarionema elsbetiae est une **ectocarpale filamenteuse endophyte**, c'est-à-dire qu'elle vit à l'intérieur d'une autre plante. (Fig. 1)

Afin de reconstruire **l'histoire évolutive des variations des voies métaboliques** dues à ce mode de vie endophyte, des premières analyses comparatives ont permis de déterminer 98 réactions métaboliques susceptibles d'avoir été perdues.

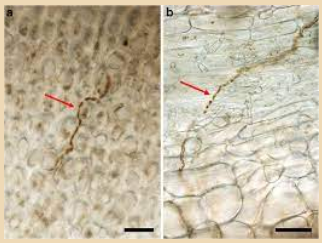

Fig. 1 : Section microscopique de *S. latissima* provenant de Bretagne nord. Les flèches rouges indiquent les filaments endophytes, et l'échelle présente 25 µm. [1]

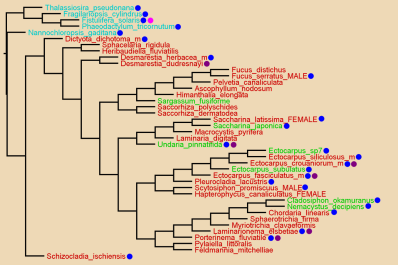

Fig. 2 : Arbre phylogénétique de référence. Provenance des génomes : Phaeoexplorer [2], publics, out-groups. Les points colorés correspondent à différents runs AuCoMe, cf Fig. 3 *Méthodes et analyses*.

Nous nous sommes donc intéressés aux potentielles pertes chez *L. elsbetiae*, de manière **globale** mais également plus **ciblée**, notamment sur la **voie des oxylipines**, impliquée dans les réactions de défense.

Dans le but de déterminer et valider ces pertes, nous avons réalisé des **analyses de comparaison de réseaux métaboliques à l'échelle du génome** provenant de 41 algues, présentées sur la Fig 2.

Méthodes et analyses

1 Nous avons récupéré les données sortantes du logiciel AuCoMe [3][4], outil permettant de construire des réseaux métaboliques à l'échelle du génome. Le logiciel a permis de faire des comparaisons entre les métabolismes des espèces présentées en Fig. 3, ainsi que tester la présence / absence d'un grand nombre de réactions chez *L. elsbetiae*.

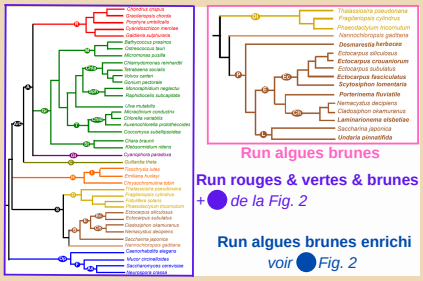

Fig. 3 : Présentation des espèces comprises dans les différents runs AuCoMe.

2 Afin de traiter les données de différents runs d'AuCoMe, nous avons élaboré un script Python [5] permettant d'obtenir des listes de réactions perdues à partir des données de présence absence. L'intersection de 3 runs (Fig. 4) a permit d'établir une liste de 27 réactions potentiellement perdues.

Fig. 4: Diagramme de Venn représentant l'intersection des 3 runs de la Fig. 3.

3 Les séquences des enzymes associées aux réactions potentiellement perdues ont permis de construire des arbres phylogénétiques avec le logiciel Seaview (méthode PhyML). Si l'arbre s'approche de la phylogénie de référence, l'hypothèse de la perte est consolidée.

Résultats

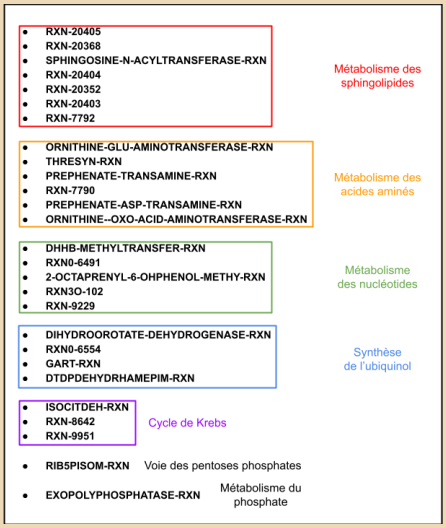

Fig. 5 : Répartition des 27 réactions provenant de l'intersection de la Fig. 4 dans les différentes voies métaboliques.

- Les 27 réactions nous donnent une vue d'ensemble sur les pertes potentielles de *L. elsbetiae*, mais ne montrent pas de tendance générale.
- Nous élargissons donc notre recherche aux unions entre les différents runs AuCoMe. (Fig 6)

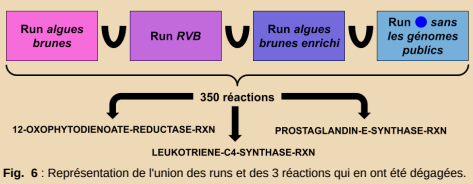

Fig. 6 : Représentation de l'union des runs et des 3 réactions qui en ont été dégagées.

- 3 réactions présentes dans l'union font parties du métabolisme des oxylipines.
- Les oxylipines sont des molécules oxygénées formées à partir d'acides gras polyinsaturés. Ces molécules sont impliquées dans des réactions de défenses chez les eucaryotes [6].

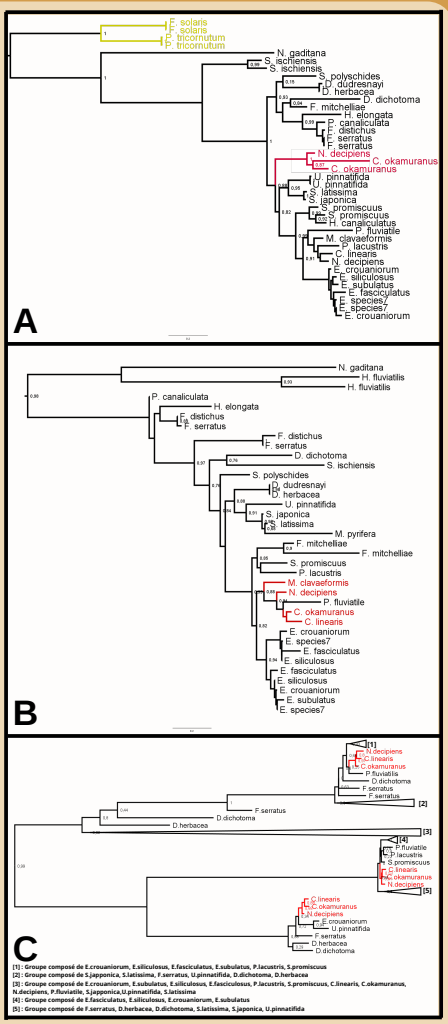

Fig. 7 : Arbres phylogénétiques construits par maximum de vraisemblance : A) 12-OXOPHYTODIENOATE-REDUCTASE-RXN, B) LEUKOTRIENE-C4-SYNTHASE-RXN, C) PROSTAGLANDIN-E-SYNTHASE-RXN

- 3 arbres correspondants aux enzymes étudiées afin d'appuyer la perte potentielle. En rouge est indiqué la famille des Chordariaceae, qui comprend le genre Laminarionema.

Discussion

Nous avons construit un modèle de voie métabolique contenant les 3 réactions concernées chez les Laminaires, en indiquant les pertes chez *L. elsbetiae*, afin de les replacer dans un contexte biologique.

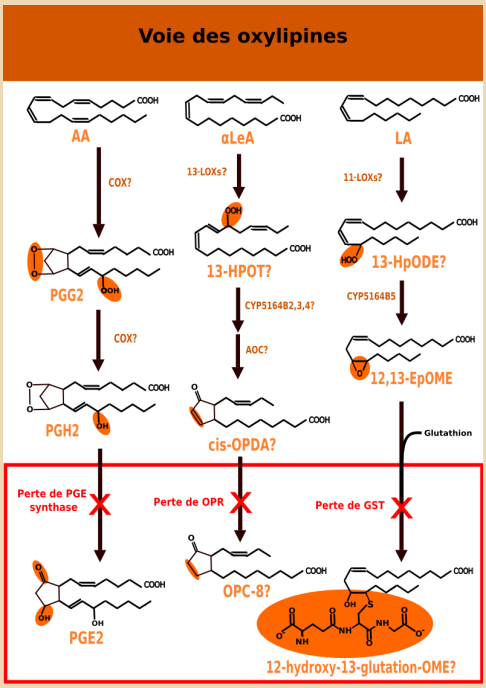

Fig. 8 : Modèle de la voie des oxylipines chez les Laminaires. Les enzymes encadrées en rouge indiquent l'absence de ces réactions chez *L. elsbetia*, due à une perte de gènes. Les molécules et enzymes marquées d'un point d'interrogation sont celles dont la présence n'a pas été prouvée chez les Laminaires.

Conclusion

→ *L. elsbetiae* a potentiellement perdu des **réactions impliquées dans son mécanisme de défense**, et ces pertes pourraient être **compensées** par son mode de vie endophyte.

→ Cependant, ces pertes sont hypothétiques, il faudrait faire des **tests biochimiques** [7] pour valider l'absence de réaction chez l'endophyte, et notre modèle métabolique pour les Laminaires.

→ En revanche, l'algue aurait possiblement conservé le **mécanisme de dégradation** de ces molécules afin d'**inactiver la défense de l'hôte**.

Références :
[1] : Bernard et al, « QPCR-Based Relative Quantification of the Brown Algal Endophyte Laminarionema Elsbetiae in Saccharina Latissima: Variation and Dynamics of Host—Endophyte Interactions »
[2] : https://phaeoexplorer.sb-roscoff.fr/
[3] : GitHub - AuReMe/aucome: Automatic Comparison of Metabolism
[4] : Arnaud Belcour et al. « AuCoMe: building panmetabolic maps for families of eukaryotic organisms » (en prep.)
[5] : https://github.com/PaulineGHG/Reactions_loss/blob/master/data/Proj%20tut%201.ipynb
[6] : « Oxylipin ». In Wikipedia, 12 septembre 2021 https://en.wikipedia.org/wiki/Oxylipin
[7] : Yacine Badis et al, « Targeted CRISPR-Cas9-based gene knockouts in the model brown alga Ectocarpus »

Master Bioinformatique : Université Rennes 1 (2021-2022) HAMON-GIRAUD Pauline.

Résumé

L'objectif du sujet du stage présenté dans ce rapport est de reconstruire les réseaux métaboliques d'algues brunes. À partir de ces réseaux seront analysés les résultats obtenus et sera plus particulièrement étudié les fonctions métaboliques de l'algue *Laminarionema elsbetiae*. Ce rapport détaille le processus de reconstruction des réseaux métaboliques en partant de la génération des fichiers d'entrée, du déploiement et de l'utilisation de l'outil de reconstruction des réseaux : AuCoMe, jusqu'à l'obtention des réseaux finaux. Les réseaux obtenus ont permis de mettre en avant plusieurs résultats. L'étude des résultats comprend l'analyse de dendrogrammes métaboliques reconstruits, la comparaison des réseaux des espèces selon la méthode de séquençage de leur génome et la recherche de réactions métaboliques supposées perdues chez l'algue *Laminarionema elsbetiae*. Les résultats décrivent aussi des scripts en langage R et python qui ont été créés pour ce projet. Il s'agit la première étude où sont exploités 29 réseaux métaboliques d'algues brunes comprenant des espèces encore assez méconnues. Cette étude soulève les limites du couplage de différentes techniques de séquençage lors de la reconstruction automatique des réseaux métaboliques. Les résultats obtenus constituent une première approche qui permettra ensuite d'améliorer la qualité des réseaux métaboliques générés ainsi que d'ajouter d'autres espèces d'algues brunes au jeu de données.

Mots clés : Réseaux métaboliques, Algues brunes, Phaeophyceae, Reconstruction, Analyse comparative

Abstract

The objective of the internship presented in this report is to reconstruct the metabolic networks of brown algae. From these networks will be analyzed the results obtained and will be more particularly studied the metabolic functions of the alga *Laminarionema elsbetiae*. This report details the process of reconstruction of metabolic networks starting from the generation of input files, the deployment and use of the network reconstruction tool : AuCoMe, until the final networks are obtained. The obtained networks allowed to highlight several results. The study of the results includes the analysis of reconstructed metabolic dendrograms, the comparison of the metabolic networks of the species according to the sequencing method of their genome and the search for supposedly lost metabolic reactions in the alga *Laminarionema elsbetiae*. The results also describe scripts in R and python language that were created for this project. This is the first study to exploit 29 metabolic networks of brown algae including species that are still relatively unknown. This study raises the limits of the coupling of different sequencing techniques when automatically reconstructing metabolic networks. The results obtained constitute a first approach that will allow to improve the quality of the metabolic networks generated and to add other species of brown algae to the dataset.

Key words : Metabolic networks, Brown algae, Phaeophyceae Reconstruction, Comparative analysis