THÈSE DE DOCTORAT

NNT : 2025UPASL019

Distribution des structures de domaines protéiques dans les génomes et les communautés planctoniques marines.

Distribution of protein folds in genomes and marine planktonic communities.

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 577, Structure et Dynamique des Systèmes Vivants (SDSV)
Spécialité de doctorat : Écologie
Graduate School : Life Science and Health. Référent : Université d'Évry Val d'Essonne

Thèse préparée dans l'unité de recherche **Génomique Métabolique** (Université Paris-Saclay, Université d'Évry, CNRS, CEA), sous la direction d'**Olivier JAILLON**, chercheur et la co-direction de **Youri TIMSIT**, directeur de recherche

Thèse soutenue à Évry-Courcouronnes, le 11 avril 2025, par

Lucas PAVLOVIC

Composition du Jury

Membres du jury avec voix délibérative

Angela FALCIATORE Directrice de recherche, Sorbonne Université	Présidente
Mathilde CARPENTIER Maîtresse de conférences, HDR Sorbonne Université	Rapportrice & Examinatrice
Chris BOWLER Directeur de recherche, École Normale Supérieure Ulm	Examinateur
Guillaume POSTIC Maître de conférences, Université Paris-Saclay	Examinateur

Titre : Distribution des structures de domaines protéiques dans les génomes et les communautés planctoniques marines.

Mots clés : structures de domaines protéiques ; distribution ; plancton ; génomes ; océans

Résumé : Le plancton marin joue un rôle central dans les cycles biogéochimiques globaux et les réseaux trophiques. Il est composé d'une grande diversité d'espèces Eucaryotes ainsi que des Procaryotes et des Virus. Étant advecté par les courants, il est confronté à des variations de conditions environnementales sur de vastes échelles géographiques. La connaissance des dynamiques organisationnelles de ce microbiome est encore incomplète, alors que leur compréhension représente un enjeu central dans un contexte de changement climatique. Les travaux de recherche que j'ai menés ici visent à mieux comprendre l'organisation des communautés planctoniques marines à l'échelle des structures de domaines protéiques (folds). Des résultats précédents décrivent certaines conséquences de la pression de sélection abiotique sur les structures de protéines mais pas sur celles des domaines protéiques, qui sont pourtant l'unité évolutive et fonctionnelle de base des protéines. Leurs structures, à l'interface entre phénotype et génotype représentent une échelle intéressante pour étudier cette organisation. L'annotation structurale des protéomes a été réalisée avec CATH[1,2] sur les protéomes de différents génomes environnementaux (MAGs)[3,4] comprenant 700 Eucaryotes, 1900 Procaryotes, 31000 *Nucleocytoviricota*, et des protéomes de référence de 990 Eucaryotes [5]. Au total, 14.5 millions de gènes (9 millions pour les protéomes de référence et 5.5 millions pour les MAGs) ont été annotés. Dans le premier chapitre, j'ai validé l'utilisation des MAGs pour étudier la distribution des folds dans une grande diversité d'espèces.

Dans une deuxième partie, j'ai exploité les abondances relatives des MAGs dans les métagénomes pour étudier la distribution des folds dans les communautés planctoniques. Celle-ci est caractérisée par une loi puissance particulière, la loi de Pareto type II. La transition entre ces deux lois pourrait être liée au fait qu'évolution moléculaire et écologie combinées façonnent la composition des communautés à différents niveaux organisationnels, en mettant en jeu des processus à la fois génomiques et émergents résultant d'interactions écologiques. Enfin, j'ai pu identifier trois classes d'abondance de folds grâce aux propriétés de la loi de Pareto type II et observer que la distribution de ceux de certaines classes d'abondance dans certains phyla est structurée par la géographie, la température et la concentration en fer.

Dans l'ensemble, cette thèse a montré que les folds représentent une échelle biologique informative pour mieux comprendre l'organisation des communautés planctoniques marines et ses déterminants.

1 : Ian Sillitoe *et al.*, CATH: increased structural coverage of functional space, Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D266–D273, https://doi.org/10.1093/nar/gkaa1079.

2: CA Orengo *et al.*, CATH – a hierarchic classification of protein domain structures, Structure, Volume 5, Issue 8, 1997, Pages 1093-1109, ISSN 0969-2126, https://doi.org/10.1016/S0969-2126(97)00260-8.

3: Delmont *et al.*, Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean, Cell Genomics, Volume 2, Issue 5, 100123, doi: 10.1016/j.xgen.2022.100123.

4: S.Kijima, H.Hikida, T.O.Delmont, M.Gaïa and H.Ogata. "Complex Genomes of Early Nucleocytoviruses Revealed by Ancient Origins of Viral Aminoacyl-tRNA Synthetases". *Mol. Bio. Evol.*, vol.41, no.8. Aug. 2024, doi: 10.1093/molbev/msae149.

5: D. J. Richter *et al.*, EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes, Peer Community J., vol. 2, 2022, doi: 10.24072/pcjournal.173.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Title : Distribution of protein folds in genomes and marine planktonic communities.

Keywords : protein folds ; distribution ; plancton ; genomes : oceans

Abstract : Marine plankton plays a central role in biogeochemical cycles and food webs. It is made up of a wide variety of eukaryotic species as well as prokaryotes and viruses. As it is advected by currents, it is subject to variations in environmental conditions over vast geographical scales. Knowledge of the organisational dynamics of this microbiome is still incomplete, even though understanding it represent a central issue in the context of climate change.

My research aimed at gaining a better understanding of the organisation of marine plankton communities at the level of protein domain structures (folds). Previous results describe certain consequences of abiotic selection pressure on protein structures but not on those of protein domains, which are the basic evolutionary and functional unit of proteins. Their structures, at the interface between phenotype and genotype, represent an interesting scale for studying this organisation.

The structural annotation of proteomes was carried out using CATH [1,2] on the proteomes of 700 eukaryotes, 1,900 prokaryotes, 31,000 *Nucleocytoviricota* environmental genomes (MAGs) [3,4], and reference proteomes from 990 eukaryotes [5]. In total, 14.5 million genes (9 million for the reference proteomes and 5.5 million for the MAGs) were annotated. In the first chapter, I validated the use of MAGs to study the distribution of folds in a wide range of species.

In the second part, I used the relative abundances of MAGs in the metagenomes to study the distribution of folds in planktonic communities. This distribution is characterised by a particular power law, the Pareto type II law. The transition between power low in the genomes and Pareto type II law in the communities could be linked to the fact that molecular evolution and ecology combined shape the composition of communities at different organisational levels, by bringing into play both genomic and emergent processes resulting from ecological interactions. Finally, I was able to identify three fold abundance classes using the properties of the Pareto type II law and observe that the distribution of those in certain abundance classes in certain phyla is structured by geography, temperature and iron concentration.

Overall, this thesis has shown that folds represent an informative biological scale for better understanding the organisation of marine plankton communities and its determinants.

1 : Ian Sillitoe *et al.*, CATH: increased structural coverage of functional space, Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D266–D273, https://doi.org/10.1093/nar/gkaa1079.

2: CA Orengo *et al.*, CATH – a hierarchic classification of protein domain structures, Structure, Volume 5, Issue 8, 1997, Pages 1093-1109, ISSN 0969-2126, https://doi.org/10.1016/S0969-2126(97)00260-8.

3: Delmont *et al.*, Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean, Cell Genomics, Volume 2, Issue 5, 100123, https://doi.org/10.1016/j.xgen.2022.100123.

4: S.Kijima, H.Hikida, T.O.Delmont, M.Gaïa and H.Ogata. "Complex Genomes of Early Nucleocytoviruses Revealed by Ancient Origins of Viral Aminoacyl-tRNA Synthetases". *Mol. Bio. Evol.*, vol.41, no.8. Aug. 2024, doi: 10.1093/molbev/msae149.

5: D. J. Richter *et al.*, EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes, Peer Community J., vol. 2, 2022, doi: 10.24072/pcjournal.173.

Remerciements

Je souhaiterais d'abord remercier Angela Falciatore, Mathilde Carpentier, Chris Bowler et Guillaume Postic d'avoir accepté de faire partie de mon jury, d'avoir pris le temps de relire ce manuscrit et pour toutes les questions et discussions que nous avons pu avoir au cours de ma soutenance de thèse. Je remercie Angela Falciatore et Mathilde Carpentier pour leurs rapports, et Mathilde Carpentier pour ses suggestions de modifications et corrections.

Je remercie mon directeur de thèse, Olivier Jaillon, pour cette opportunité et pour avoir fait en sorte que je puisse bénéficier de six mois de prolongations, sans lesquels je n'aurais jamais pu finaliser cette thèse. Je le remercie également pour le temps qu'il a consacré à la relecture de ce manuscrit et à la préparation de ma soutenance.

Je remercie Youri Timsit et Magali Lescot pour leur implication tout au long de ma thèse. Je remercie particulièrement Youri Timsit pour le temps qu'il a consacré à la relecture de ce manuscrit ainsi qu'à la préparation de ma soutenance.

Je remercie les membres de mon comité de suivi individuel, Lucie Bittner et Gabriel Reygondeau, d'avoir pris le temps de faire ces trois réunions au cours de ma thèse. Vos conseils et encouragements m'ont beaucoup aidé.

Je remercie à nouveau Magali Lescot, ainsi que Caroline Vernette, d'avoir généré toutes les données nécessaires à la réalisation de ma thèse ainsi que les premières analyses. Je remercie Janaina Rigonato d'avoir participé aux discussions et à la réalisation des analyses préliminaires. Je remercie Daniele Iudicone et Emanuele Pigani pour les différentes discussions très enrichissantes que nous avons eues, et pour leur aide précieuse sur les lois de Pareto. Je remercie Téo Lémane et Paul Frémont pour les discussions sur mes travaux ainsi que pour le développement de Climap et le temps qu'ils ont consacré à nous former, Margaux, Manon et moi à son utilisation. Enfin et bien sûr, toutes les équipes impliquées dans la production des données *Tara* Oceans, de l'échantillonnage jusqu'au séquençage, grâce à qui nous disposons de ces données pour travailler, ainsi que les équipes informatiques du Genoscope, qui facilitent considérablement la réalisation de nos travaux et se rendent toujours disponibles pour résoudre nos problèmes informatiques.

Je voudrais aussi remercier Patrick Wincker de m'avoir permis de réaliser mon doctorat au sein du LAGE. Je remercie tous les permanents du LAGE pour leur soutien. Je remercie également l'ED SDSV pour le financement. Je remercie Sylvie Bobelet, Catherine Contrepois, Nancy Delpeche et Catherine Sarlande ainsi que les personnels de l'UEVE et de l'ED avec qui j'ai pu être en contact pour leur aide et leur bienveillance sur tous les aspects administratifs de la thèse.

Je remercie Valérie Chaudru et Élisabeth Petit-Teixeira de m'avoir donné l'opportunité de participer à différents enseignements à l'UEVE dans les meilleures conditions possibles.

Je remercie le consortium *Tara* Oceans, entre autres pour l'organisation des retraites *Tara* qui sont de très bons souvenirs, et plus particulièrement les membres du comité d'organisation du *Tara* Oceans Seminar Series avec qui j'ai eu le plaisir de gérer ces séminaires.

Je souhaiterais remercier profondément les trois stagiaires que j'ai co-encadré au cours de ma thèse, Louis, Manon, Alexandre. Votre implication dans mes travaux m'a énormément aidé et apporté beaucoup de motivation. Vous avez tous les trois étés des super stagiaires, motivés, volontaires et plein d'idées, merci infiniment à vous.

Un énorme merci à tous les non permanents du labo. Il y a eu de nombreuses journées où vous étiez ma principale source de motivation, et j'ai eu beaucoup de chance de faire votre connaissance à tous (bien que cela ne se soit pas reflété dans mon engagement pour les Genopub ou les raids Picard mais le cœur y était vraiment). Merci pour tous les moments passés ensembles, les discussions et les encouragements. J'ai une pensée en particulier pour les doctorants: Margaux et Chloé (je vous remercie également infiniment toutes les deux pour tout le temps que vous avez consacré à l'organisation de mon pot de soutenance et tout ce qui va avec), Lucie, Thibault, Clément, Nicolas. Merci également à tous les autres : Barbara, Ulysse, Benjamin, Manon, Emilie. Enfin, je remercie également tous ceux que j'ai eu la chance de connaître avant la fin de leurs contrats : Romuald, Julie, Paul, Marie, Janaina, Nina, Lise, Téo, Mathieu, Achraf. La solidarité très forte que j'ai toujours ressenti entre nous tous a été un aspect vraiment important de la façon dont j'ai vécu cette thèse et en a fait une vraie aventure humaine, en plus de tous les aspects professionnels.

Je remercie les encadrants de mes précédents stages, qui m'ont également soutenu tout au long de cette thèse : Dominique, Sébastien, Aude, Florence et Line. J'ai eu de la chance que mes premières expériences professionnelles se soient déroulées sous votre encadrement exigeant mais toujours bienveillant. Merci également à Stefano pour le temps que tu as consacré à la mise au point d'un sujet de thèse il y a trois ans et demi, et merci à toi ainsi qu'à Claudie pour le temps consacré à la rédaction du projet d'ATER.

Enfin, je remercie tous mes proches de m'avoir soutenu tout au long de mon parcours académique, de la prépa à la thèse. Merci à vous Benjamin, Antoine, Lou, Hadrien, Sariel, Elyès, Jad, Wassila, Fahd. Les derniers remerciements sont évidemment adressés à ma sœur et mes parents. Je n'ai pas de mots assez fort pour exprimer ma gratitude envers vous pour votre soutien inconditionnel. Merci, je n'en serais évidemment pas là aujourd'hui sans vous.

PARTIE I.
INTRODUCTION

I. Lexique

AFDB = AlphaFold Protein Structure Database

ASV = Amplicon Sequence Variant [1]

AV = Valeur d'abondance

BDIM = Birth Death Innovation Model

CATH = Class Architecture Topology Homology [2], [3]

- Classe (**C**): proportions relatives d'hélices α et de feuillets β (Class 1 = mainly Alpha; Class 2 = mainly Beta; Class 3 = Alpha Beta; Class 4 = few secondary structures; Class 6 = special)
- Architecture (C**A**) : caractéristiques générales de la forme de la structure (exemple : α-β sandwich)
- Topologie (CA**T**): niveau correspondant au fold = repliement = structure de domaine protéique. Nombre, arrangement et connectivité des structures secondaires entre elles
- Homologie (CAT**H**): fonction et origine évolutive
- exemple: 3.40.50.300
  → Classe 3: Alpha Beta
  → Architecture 3.40: 3-Layer(aba) Sandwich
  → Topologie 3.40.50: Rossmann Fold
  → Homologie 3.40.50.300: P-loop containing nucleotide triphosphate hydrolases

Chl *a* = Chlorophylle A

CO = Contact Order

DCM = Deep Chlorophyl Maximum

EC = Enzyme Commission number

Foldome = ensemble des folds d'un protéome donné

KS = Kolmogorov-Smirnov [4]

MAG = Génome Assemblé à partir d'un Métagénome [5]

Mpb = Mega paires de bases (10^6 paires de bases)

NCLDV = *Nucleocytoviricota* ou NucleoCytoplasmic Large Dna Virus

OMZ = Zone de Minimum d'Oxygène

OTU = Operational Taxonomic Unit [6]

OV = Valeur d'Occurrence

PII = Pareto type II

PFAM = Protein Families [7]

PDF = Densité

RAD = Distribution des Rangs des Abondances

RMSE = Racine carrée de l'Erreur Quadratique Moyenne

RP = Protéome de Référence (d'EukProt [8])

SAD = Distribution d'Abondance des Espèces

SSS = Super Secondary Structures

SST = Température de Surface de la Mer

Structurome = ensemble des structures de protéines d'un protéome

TO = *Tara* Oceans [9]

WOA = World Ocean Atlas [10]

Pour fluidifier l'écriture, les Virus seront inclus dans les domaines du vivant pour le reste de la thèse (bien que leur nature en tant qu'être vivants ou non soit débattue).

II. Préambule

Le plancton désigne l'ensemble des êtres vivants dont les déplacements sont principalement passifs et causés par les courants. Il inclut le phytoplancton (organismes autotrophes photosynthétiques) et le zooplancton (organismes hétérotrophes), et est composé d'espèces dont la taille varie du mètre au nanomètre (Figure 1).

Figure 1. Mandala du plancton. De haut en bas, mega-, macro-, meso-, micro- et nanoplancton. Le picoplancton n'est pas représenté. Photo de © Christian & Noé Sardet, extrait de Plancton Aux origines du vivant [11].

La part microscopique du plancton, correspondant au microbiome marin, recèle une diversité d'espèces très élevée, particulièrement pour les Eucaryotes aussi bien unicellulaires que pluricellulaires. Une de ses spécificités par rapport aux autres microbiomes du vivant est la capacité de dispersion très élevée de ses espèces, grâce à l'advection par les courants marins.

Le plancton a joué un rôle majeur dans l'évolution de la vie. Les premières cellules et êtres vivants étaient du plancton, et la diversité actuelle du plancton englobe tous les principaux groupes

taxonomiques du vivant. Au cours du temps, le plancton a également participé à façonner notre planète, autant du point de vue géologique, climatique que chimique. Au début du Protérozoïque (il y a environ 2500 millions d'années), les cyanobactéries ont été responsables du Grand Évènement d'Oxydation, qui a bouleversé les dynamiques climatiques et biologiques de l'époque. À l'heure actuelle, près de 50% du dioxygène dans l'atmosphère serait produit par le plancton. Il est également l'acteur principal de la composante biologique de la pompe à carbone marine qui stocke du carbone atmosphérique dans les sédiments marins. Le dioxyde de carbone atmosphérique se dissout en effet dans l'eau de mer, et plus sa concentration dans l'eau est faible et plus ce flux est important. Cet effet est amplifié dans les milieux froids. Le dioxyde de carbone dissout est consommé par le phytoplancton pour produire de la biomasse. Cette biomasse est soit intégrée dans les réseaux trophiques marins via les consommateurs primaires (faisant du plancton la base de tous les réseaux trophiques marins), soit re-transformée en matière minérale dissoute par le biais du court-circuit viral, soit sédimentée par le biais de la navette virale. Dans ce dernier cas, le carbone qui était à l'origine atmosphérique se retrouve piégé dans les couches sédimentaires profondes. Ce processus participe donc au piégeage à long terme du carbone. L'implication directe du plancton dans les cycles biogéochimiques le rend vulnérable au changement climatique en cours. Selon certains modèles, il serait en effet au cœur de boucles de rétroactions qui seraient négatives et précipiteraient l'intégralité des écosystèmes marins vers des conditions d'instabilité. L'acidification des océans, causée par l'augmentation de leur concentration en dioxyde de carbone dissout, est un premier exemple de phénomène à l'origine d'une telle boucle de rétroaction ayant un impact sur le plancton [12]. La stratification des masses d'eau océanique est un autre phénomène causé par le réchauffement des océans qui pourrait être impliqué dans une boucle de rétroaction négative [13].

Dans ce contexte, il est donc crucial de comprendre l'écologie et l'organisation du plancton à l'échelle globale.

III. Contexte, organisation et objectifs

La compréhension de l'écologie et de l'organisation du plancton à l'échelle globale ne peut passer que par la prise en compte des interactions entre plusieurs échelles biologiques et le milieu océanique. Ce cadre est nommé « Seascape » par analogie au paysage terrestre qui est composé de la réunion d'un biome et d'une biocénose [14]. Les échelles biologiques étudiées ont jusqu'ici principalement été la communauté, l'espèce, les fonctions et les propriétés génomiques (Figure 2).

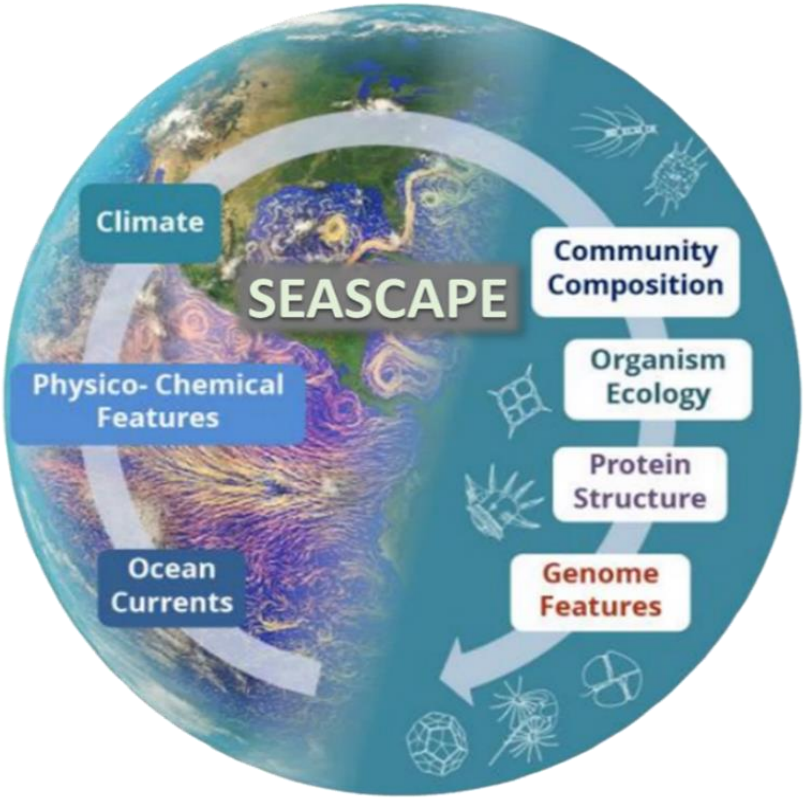


Figure 2. Les différentes échelles en interaction dans le seascape. Figure de Lescot, Jaillon, Timsit et Le Bescot.

L'un des enjeux de ces études a été de définir des biogéographies, c'est-à-dire la distribution géographique des espèces ou des communautés à l'échelle des bassins ou des biomes. L'un des concepts fondateurs de l'étude de la biogéographie des microbiomes marins a été développé au XX^{ème} siècle et a été résumé comme suit: « everything is everywhere », complété plus tard par « but the environment selects » [15], [16], [17], [18]. Elle traduit le fait que l'advection par les courants marins abolit les barrières majeures à la dispersion qui empêcheraient une espèce planctonique d'être présente dans tous les océans, et que ce n'est que la pression de sélection générée par le contexte environnemental qui modifie son abondance dans le milieu. Les observations à l'échelle des communautés et des espèces ont montré que les microbiomes marins étaient organisés en communautés advectées par les courants océaniques sur des échelles spatiales de l'ordre du bassin et temporelles de l'ordre de l'année. Les distinctions entre communautés planctoniques permettent de séparer des provinces génomiques, dont la similarité décroît en fonction du temps d'advection par les courants nécessaires pour passer de l'une à l'autre, et qui sont dans une certaine mesure cohérentes avec les provinces de Longhurst [19]. Les communautés les plus dissimilaires sont les communautés polaires et les communautés des biomes mésophiles (tempérés et tropicaux). Elles sont principalement distinguées par la différence de composition de leur phytoplancton, le phytoplancton polaire étant

composé essentiellement d'Eucaryote et le mésophile de Bactéries. La distribution des communautés planctoniques est principalement impactée par les paramètres environnementaux du milieu, qu'ils soient physiques (température, ensoleillement) ou chimiques (concentrations en azote, phosphate, fer, etc.).

Les cellules planctoniques, comme toutes celles du vivant, sont composées de macromolécules biologiques (protéines, acides nucléiques, sucres et lipides) qui jouent un rôle déterminant dans leurs adaptations aux conditions environnementales. Pour bien comprendre ces phénomènes, il est donc indispensable d'explorer comment les espèces planctoniques modulent l'usage et la composition de leur protéome en fonction des écosystèmes. Une manière de caractériser les protéomes est de se pencher sur son usage des folds, ou son « foldome », c'est-à-dire l'ensemble des folds de son protéome. Les folds constituent la topologie des structures de domaines protéiques. Ils représentent le chemin tridimensionnel que suit le squelette du domaine protéique, indépendamment de sa séquence et des chaînes latérales de ses acides aminés. Il en existe plusieurs milliers, pouvant être classées en fonction de leur structure tridimensionnelle. Les différentes catégories de folds se distinguent selon leurs compositions en hélices α, feuillets β et boucles. Elles peuvent être plus ou moins globulaires et compactes, certaines étant même irrégulières et désorganisées. Elles ont chacune des propriétés biochimiques et des cinétiques de repliement unique, et réalisent une ou plusieurs fonctions de manière autonome, comme dans les enzymes monodomaines, ou en combinaisons, chaque fold participant à la réalisation de la fonction globale. La même fonction peut aussi être réalisée par des folds différents.

De manière intéressante, l'usage des folds varie fortement au sein d'un protéome et entre protéomes. Certains sont adoptés par la majorités des domaines protéiques, alors que d'autres ne sont adoptés que par un seul domaine. Ces différences s'expliquent à la fois par les propriétés structurales et fonctionnelles des folds, ainsi que leur histoire évolutive. Les folds adoptés par un grand nombre de domaines sont appelés superfolds. Le Rossmann fold, les TIM-barrel ou les Ig-like en sont des exemples. Les autres folds, appelés unifolds ou mésofolds, ont une diversité beaucoup plus importante que les superfolds mais ne sont adoptés que par quelques ou un seul domaine dans un protéome. L'usage des folds varie également selon les espèces, certains folds n'étant présents que dans certains génomes, d'autres étant présents de façon universelle. Certains folds sont mêmes des synapomorphies qui permettent de distinguer les phyla entre eux [20]. Les folds sont un niveau biologique sous pression de sélection, autant pour la fonction qu'ils réalisent que pour leurs propriétés biochimiques et biophysiques, et sont donc d'une certaine manière en interaction avec leur milieu. Par exemple, les Diatomées vivant dans des milieux riches en zinc ont un usage accrus de folds dont la structure implique des doigts de zinc [21]. Des levures exposées à un stress thermique prolongé ont un usage accru de folds ne formant pas d'agrégats en conditions thermophiles, qui les empêche de réaliser correctement leurs fonctions [22].

Dans ce contexte, cette thèse se propose d'étudier les caractéristiques du plancton au niveau des structures de domaines protéique dans des protéomes de référence et des métagénomes. Cela n'avait jamais été fait auparavant à cette échelle. Ces structures se trouvent à un niveau biologique intermédiaire entre génotype et phénotype, particulièrement intéressant pour apporter des éléments de compréhension dans des approches holistiques pouvant permettre d'étudier le plancton.

Les deux problématiques principales de cette thèse seront donc :

Quelles sont les caractéristiques de la distribution des structures de domaines protéiques dans certaines espèces abondantes du plancton marin ?

La distribution des structures de domaines protéiques dans les communautés planctoniques marines est-elle structurée biogéographiquement, et quels sont les facteurs responsables de cette structuration ?

Pour répondre à ces questions, l'introduction qui suit a d'abord pour but de décrire les deux niveaux biologiques qui seront mis en relation dans cette thèse. D'abord, les propriétés des communautés planctoniques marines : composition, diversité taxonomique, répartition biogéographique à plusieurs niveaux organisationnels et principaux facteurs influant sur ces répartitions. Le deuxième niveau est celui des structures de domaines protéiques. Leur diversité, classification et propriétés structurales seront décrites, ainsi que leur évolution et leur réponse face aux pressions de sélections environnementales.

La seconde partie de cette thèse détaillera les méthodes et outils utilisés pour produire les résultats qui seront présentés dans une troisième partie.

La troisième partie est articulée en trois chapitres de résultats. Dans le premier, je décrirai les foldomes des espèces planctoniques dans différentes bases de données à différents niveaux taxonomiques et les comparerai aux foldomes d'espèces de référence. Le second sera consacré à la modélisation de la distribution des folds à deux échelles (foldomes planctoniques et communautés planctoniques) avec deux modèles mathématiques différents. Enfin, j'étudierai dans le dernier chapitre les propriétés de la distribution biogéographique des folds dans les stations TO et dans l'Océan global.

La dernière partie correspond à la conclusion et aux perspectives de ces travaux.

CHAPITRE 1. DIVERSITÉ ET STRUCTURATION BIOGÉOGRAPHIQUE DE LA DISTRIBUTION DU PLANCTON DANS LES OCÉANS

Le premier chapitre de cette introduction a pour but de décrire plusieurs propriétés des communautés planctoniques dans les océans : leur composition taxonomique et leur biomasse, leur diversité ainsi que leur distribution biogéographique à différents niveaux organisationnels. Elle abordera également les facteurs à l'origine de la variabilité de ces propriétés, et évoquera certains modèles utilisés pour décrire ces communautés.

I. description et classification des espèces planctoniques

I.1. méthodes de description du plancton

Plusieurs méthodes existent pour décrire la composition du plancton [23]. Historiquement, la première est l'observation au microscope. Actuellement, les observations utilisent plutôt des technologies comme le FlowCam [24] ou le ZooScan [25] pour le zooplancton, qui permettent non seulement de décrire les individus d'un échantillon d'eau de mer dans un environnement donné à l'aide de critères morphologique, souvent à l'aide d'intelligence artificielle, mais aussi d'évaluer des abondances. Des facteurs de conversions entre volume et masse sont souvent nécessaires pour analyser les résultats, et sont accessibles via les méthodes d'imagerie de type ZooScan. Des méthodes moins technologiques ont aussi fait leurs preuves comme le pesage de masse humide ou sèche après des traits de filets. Les limitations principales de ces méthodes sont la résolution taxonomique, liée à l'utilisation de critères morphologiques. L'évaluation de la concentration en pigments dans un échantillon est une autre méthode pour obtenir des informations sur composition du plancton. Cette mesure peut être faite sur des échantillons prélevés dans le milieu par HPLC (High-Performance Liquid Chromatography) ou à beaucoup plus grande échelle, par détection passive et à distance de la couleur des océans à l'aide de satellites [23]. Ils sont en effet capables, en analysant les longueurs d'ondes réfléchis par la surface de l'eau, d'estimer la concentration en chlorophylle *a*. Cette méthode est particulièrement efficace pour suivre à l'échelle de bassins entiers et pendant des années les dynamiques du phytoplancton en surface, et créer des modèles de sa dynamique au cours du temps [26]. Il est également possible de mesurer concentration particulaire en surface (donc de tout le plancton à cette profondeur dans la colonne d'eau) par LIDAR (Light Detection and Ranging) [27]. Ces méthodes ne permettent cependant pas d'avoir accès à des informations au rang des espèces. Elles ne permettent pas non plus d'effectuer des analyses plus profondément dans la colonne d'eau. Enfin, l'abondance et la diversité des espèces planctoniques peuvent être estimées à l'aide de la génomique via le métabarcoding ou la métagénomique. Le métabarcoding consiste à ne cibler que des régions précises dans tout l'ADN d'un échantillon, comme la V9 de l'ADNr 16S pour les Procaryotes ou celle du 18S pour les Eucaryotes. Ces régions sont isolées grâce à des amorces conçues pour les cibler spécifiquement, puis amplifiées par PCR. Chaque séquence unique de barcode est ensuite associée à une taxonomie à un niveau plus ou moins élevé, et permet de définir des Operational Taxonomic Units (OTUs) ou des Amplicon Sequence Variants (ASVs).

La métagénomique caractérise quant à elle des séquences nucléotidiques présentes dans un échantillon au sein d'une communauté. Elles peuvent être soit alignées sur des génomes utilisés comme référence, soit utilisées pour en reconstruire quand de tels génomes ne sont pas disponibles. Ces méthodes permettent généralement d'obtenir une résolution taxonomique relativement fine pouvant aller jusqu'au niveau de l'espèce et même du génotype dans le cas d'assemblages de génomes. Par l'approche de barcodes, il est cependant parfois compliqué d'atteindre une telle résolution car les séquences utilisées pour les identifications sont parfois complètement identiques à l'échelle d'un genre entier (cas de la région V9 de l'ADNr 18S des genres *Bathycoccus* ou *Pelagomonas* par exemple). Les méthodes génomiques permettent aussi d'estimer des abondances relatives en comparant le nombre de lectures générées par séquençage et associées à chaque espèce au nombre

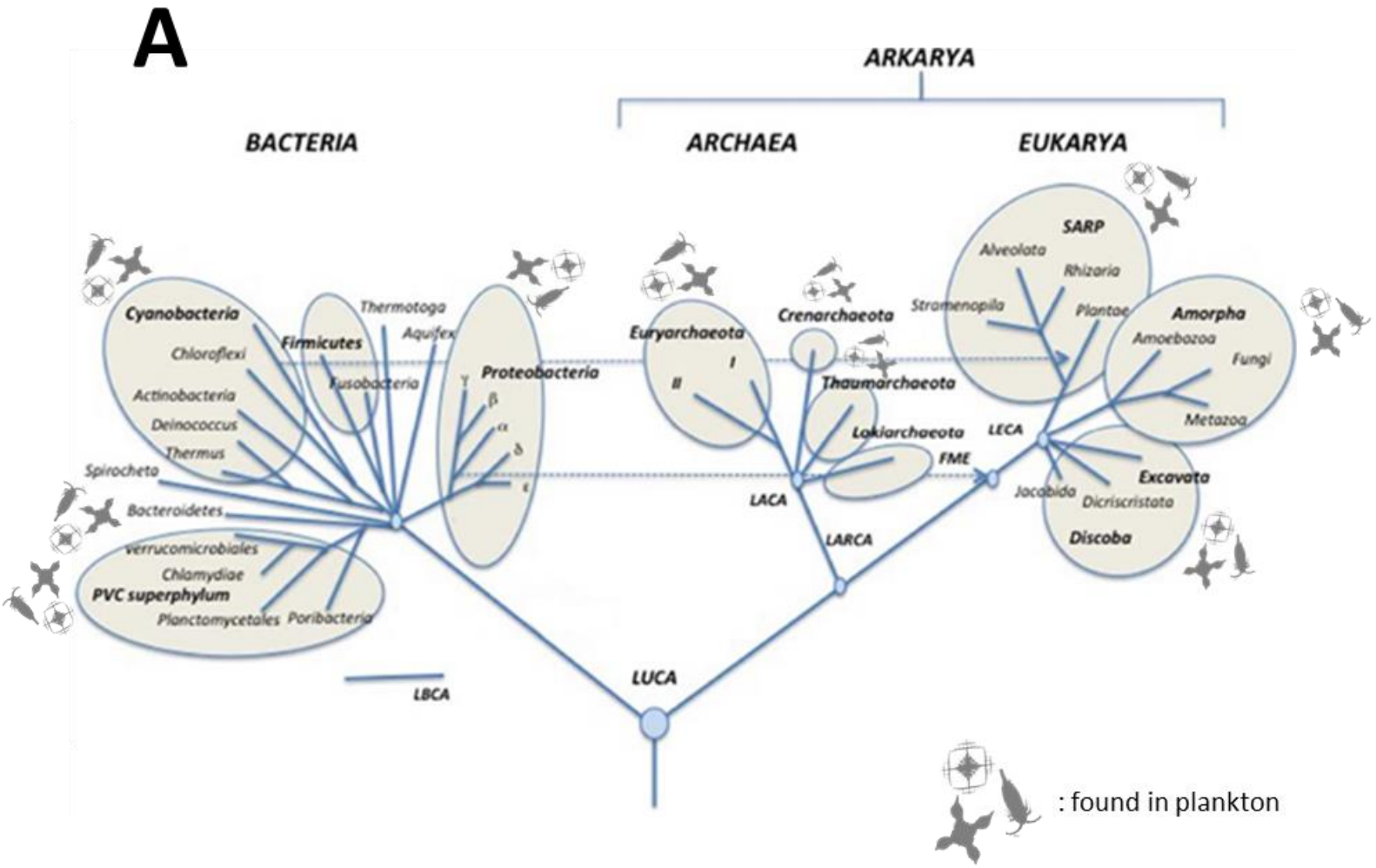
total de lectures produites dans l'échantillon. Ces abondances relatives comportent cependant plusieurs biais [28]. L'un d'entre eux est dit compositionnel : les échantillons à partir desquels ils sont calculés ne représentent en effet qu'une fraction de l'ensemble du plancton qui a été aléatoirement prélevée lors de l'échantillonnage. Deux échantillons prélevés au même endroit peuvent donc en théorie aboutir à des abondances relatives différentes pour les mêmes espèces, et, de la même façon, deux échantillons prélevés à des endroits différents ne sont pas directement comparables car ils ne représentent potentiellement pas la même part de l'ensemble du plancton. Le biais compositionnel dépend cependant beaucoup du volume de séquençage et affecte peu les espèces les plus abondantes dans la pratique. Des améliorations des méthodes de génomique sont en cours et à venir. Par exemple, la méthode de spike-in, dans laquelle la quantité connue d'un fragment d'ADN ajoutée à un échantillon avant séquençage est utilisée comme étalon pour transformer les abondances relatives en quantités [29].

I.2. classification des différentes lignées planctoniques

Les espèces planctoniques appartiennent à toutes les lignées principales du vivant, chez les Procaryotes comme chez les Eucaryotes (Figure 3 A).

La diversité des Procaryotes y avoisine les 37500 OTUs différents [30]. Les Bactéries planctoniques sont communément appelées « Bactérioplancton ». Il regroupe entre autres des Pseudomonadota (Alphaproteobactéries, Gammaprotéobactéries), des représentants du groupe PVC (notamment des Planctomycetota) ainsi que des Cyanobactéries. Le plancton héberge également une diversité élevée de Virus, pouvant être appelé « Virioplancton » avec de prophages, des virus à ADN simple brin, double brin ou à ARN ainsi que des *Nucleocytoviricota* (NCLDV, qui contient les virus géants dont la taille peut excéder le micron) et des Mirusvirus [31], [32], [33], [34], [35]. les *Nucleocytoviricota* ont une diversité de près de 6800 phylotypes dans le plancton dont 5000 appartenant aux Mimiviridae [34]. NCLDV et Mirusvirus sont caractérisés par une spécialisation pour les Eucaryotes, et ont une histoire évolutive particulière (Figure 3 C).

Toutes les lignées Eucaryotes connues sont représentées dans le microbiome marin, avec près de 250000 OTUs et 242500 ASVs [36], [37]. Il pourrait y avoir au total près de 16.5 millions d'espèces différentes [38]. Une grande partie des Unicontes ou Amorphea sont présents dans le plancton, dont des Amebozoaires et la majorité des Opisthocontes comprenant entre autres des Fungi et des Choanozoaires (Figure 3 B ; Figure 4). Au sein des Choanozoaires, les Choanoflagellés (Figure 4 F) ainsi qu'une partie des Métazoaires (Figure 4 E) sont représentés. Les Vertébrés, à l'exception de certains Actinoptérygiens (gamètes et stades larvaires), en sont en absents (ce qui est attendu puisque la majorité des Vertébrés ont une taille macroscopique tout au long de leur vie, excepté leurs gamètes). Les pluricellulaires au sein du plancton sont surtout des Arthropodes, des Cnidaires et les larves d'autres espèces de Métazoaires. La diversité des Bicontes est particulièrement importante et tous les principaux embranchements sont représentés, à savoir certains Archaeplastides (Figure 4 G), des Rhizaires (Figure 4 H,I,J), des Straménopiles (Figure 4 N,O,P), des Alvéolés (Figure 4 K,L,M) et d'autres comme les Haptophytes, les Cryptophytes et les Centroheliozoaires (Figure 4 Q,R). Les Excavata, dont la position est incertaine, y sont aussi bien représentés (Figure 4 B,C,D). Certaines lignées planctoniques Eucaryotes sont hyperdiversifiés, comme les Diatomées (Straménopiles ; Figure 4 N) et les Dinoflagellés (Alveolata ; Figure 4 K) avec plus de 1000 OTUs chacunes [38].


A
ARKARYA
BACTERIA
ARCHAEA
EUKARYA
Cyanobacteria
Chloroflexi
Firmicutes
Thermotoga
Aquifex
Fusobacteria
Actinobacteria
Deinococcus
Thermus
Spirocheta
Bacteroidetes
Verrucomicrobiales
Chlamydiae
PVC superphylum
Planctomycetales
Poribacteria
Proteobacteria
γ
β
α
δ
ε
Euryarchaeota
I
II
Crenarchaeota
Thaumarchaeota
Lokiarchaeota
FME
LACA
LARCA
LECA
SARP
Alveolata
Rhizaria
Stramenopila
Plantae
Amorpha
Amoebozoa
Fungi
Metazoa
Excavata
Jacobida
Dicriscristata
Discoba
LUCA
LBCA
: found in plankton
B
Archaeplastida
Rhodophyta
Rhodelphis
Chloroplastida
Glaucophyta
Ancoracysta
Picozoa
Amorphea
Apusomonada
Breviates
Obazoa
Opisthokonta
Amoebozoa
Cryptista
Palpitomonas
Katablepharida
Cryptophyta
Diphylleida
CRuMs
Rigifilida
Mantamonas
Haptista
Centrohelida
Haptophyta
Rappemonads
Discoba
'Excavates'
Metamonada
Malawimonadida
Ancyromonadida
Telonemia
TSAR
Rhizaria
Sar
Alveolata
Stramenopila
Hemimastigophora
Original 'Supergroup'
No molecular data in 2004

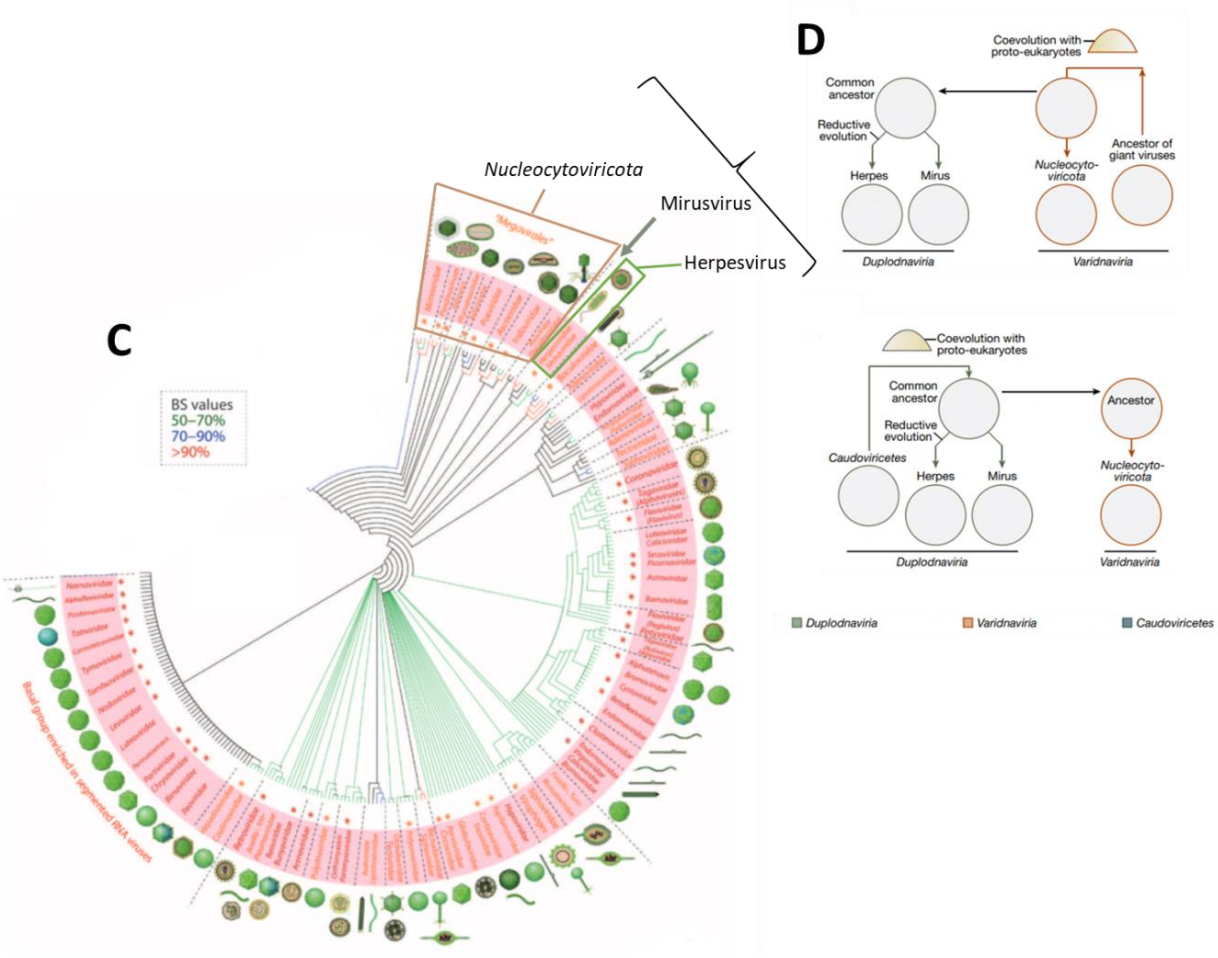


Figure 3. Arbres du vivant et des Virus. (A) Arbre Universel schématique basé sur les informations de l'ADNr et d'arbres de protéines, mis à jour de Woese *et al*. [39]. Les flèches en pointillés bleus représentent les transferts horizontaux d'ADN. LBCA : Dernier Ancêtre Commun des Bactéries ; LACA : Dernier Ancêtre Commun des Archées ; LARCA : Dernier Ancêtre Commun des Arkarya (Eucaryotes et Archées) ; FME : Premier Eucaryote avec des Mitochondries ; LECA : Dernier Ancêtre Commun des Eucaryotes ; SARP : Stramenopiles Alveolés Rhizaires Plantes. Adapté de Forterre *et al.* [40]. **(B)** Arbre des Eucaryotes basé sur un consensus d'études phylogénomiques antérieures à 2020. Ces études sont systématiquement basées sur l'utilisation de plusieurs gènes simultanément, dans la majorité des cas plus de 100. Les groupes colorés sont des « supergroupes ». Les lignes en pointillés indiquent les branches avec peu d'incertitudes sur la monophylie. Les étoiles indiquent les groupes anciennement considérés comme des supergroupes ; les cercles, les lignées majeures pour lesquelles il n'existait pas de données génomiques disponibles lors de l'établissement du modèle de supergroupe. Figure extraite de Burki *et al*. [41]. **(A-B)** L'absence de mention « found in plankton » sur un groupe n'exclue pas qu'une de ses espèces ne soit identifiée dans le plancton dans de futures études ; elle a simplement pour vocation de montrer les groupes principaux qui y sont trouvés. **(C)** Arbre des Virus et positionnement des Mirusvirus, basé sur les abondances de 442 folds universels (présents dans tous les domaines du vivant et les Virus) dans 368 protéomes de Virus.La couleur des branches représente la valeur de bootstrap associée. L'ordre des Megavirales, qui correspond globalement aux *Nucleocytoviricota*, n'est pas encore validé et est indiqué entre guillements. Les groupes monophylétiques sont représentés avec un astérisque. Quand cela était possible, une photo du morphotype provenant de ViralZone Web [42] a été ajoutée. Adapté de Nasir & Caetano-Anollés [43]. **(D)** Scenarii d'évolution possibles pour les Herpesvirus, Mirusvirus et des *Nucleocytoviricota*, basé sur la présence ou l'absence des gènes du module informationnel et du virion. Le module informationnel des Virus infectant les Eucaryotes des règnes *Duplodnaviria* et *Varidnaviria* serait apparu soit dans l'ancêtre des *Nucleocytoviricota* (en haut) soit des Mirusvirus (en bas). Adapté de Gaïa *et al*. [31].

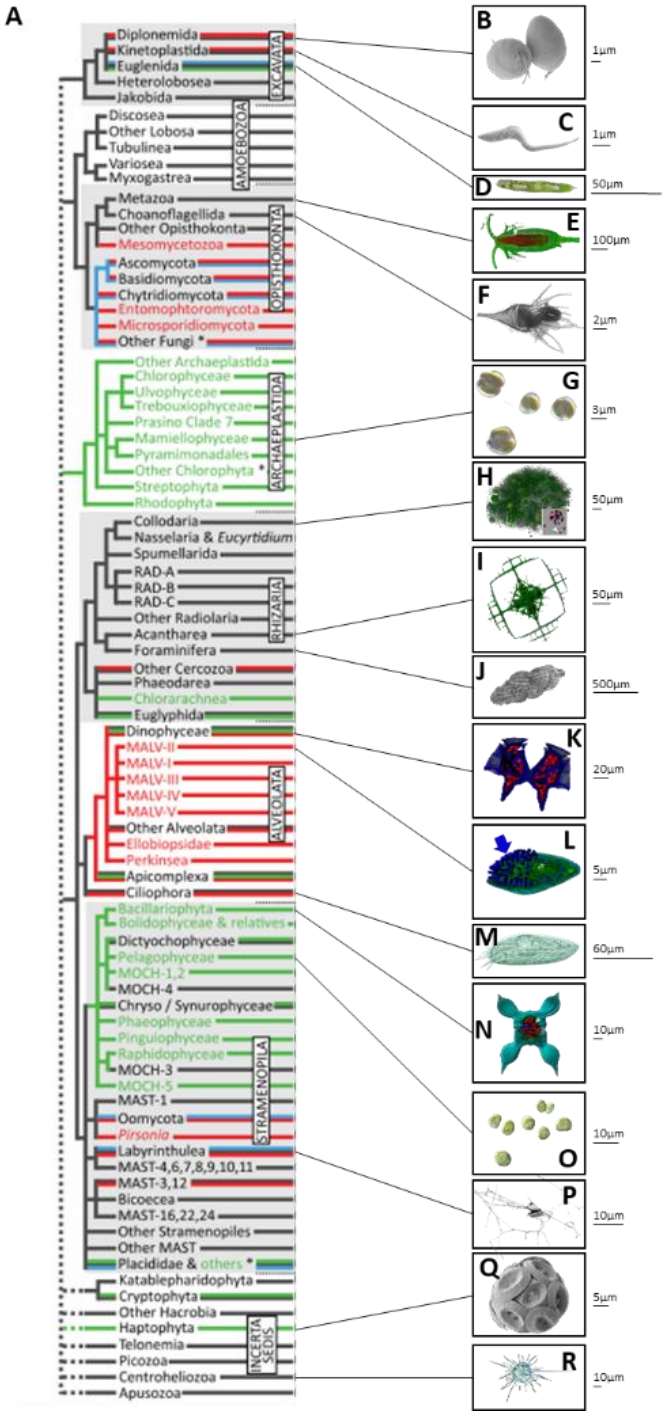


Figure 4. Classification du plancton Eucaryote. (A) Phylogénie schématique de 85 lignées Eucaryotes du plancton, basée sur la région V9 de l'ADNr 18S. Les couleurs indiquent les types trophiques connus : rouge = parasite, vert = photoautotrophe, bleu = osmo- ou saprotrophe ; noir = principalement phagotrophe (aussi la couleur des lignées identifiées uniquement à partir de séquences environnementales). Les trois branches avec un astérisque regroupent artificiellement plusieurs lignées distinctes. Extrait de De Vargas *et al*. [38]. **(B)** *Diplonema papillatum*. Extrait de Faktorová *et al*. [44]. **(C)** *Trypanosoma brucei*. Extrait de Faktorová *et al*. [44]. **(D)** *Euglena* sp.. Photo de Piotr Rotkewicz [45]. **(E)** Copépode. Extrait de De Vargas *et al.* [38]. **(F)** *Acanthoeca spectabilis*. Extrait de Leadbeater *et al*. [46]. **(G)** *Micromonas pusilla*. Extrait de la base de donnée de la Roscoff Culture Collection [47]. **(H)** Collodaire. Extrait de De Vargas *et al*. [38]. **(I)** *Lithoptera* sp.. Extrait de De Vargas *et al*. [38]. **(J)** *Uvigerina peregrina*. Photo de Claire McKay [48]. **(K)** *Dinophysis caudata*. Extrait de De Vargas *et al*. [38]. **(L)** *Amoebophrya* sp.. indiqué par la flèche bleue. Extrait de De Vargas *et al*. [38]. **(M)** *Spirotrichea* sp.. Photo de Y. Tsukii **[49]**. **(N)** *Chaetoceros bulbosus*. Extrait de De Vargas *et al*. [38]. **(O)** Pelagophyte. Extrait de Mitani *et al*. [50]. **(P)** *Aplanochytrium* sp.. Photo de Celeste Leander [51]. **(Q)** *Coccolithus pelagicus*. Extrait de Šupraha [52]. **(R)** *Raineriophrys echinata*. Extrait de Tice *et al*. [53].

I.3. groupement par taille des différentes espèces planctoniques

Le plancton rassemble des espèces dont la taille des individus peut varier du mètre à la dizaine de nanomètre. Ce gradient peut être découpé en plusieurs groupes de tailles, ou fractions, pour des raisons opérationnelles à des fins d'échantillonnage [54] (Figure 5). Elles sont imposées par les variations de densité en organismes dans l'eau de mer en fonction de leur taille. Il faut en effet prélever des volumes d'eau beaucoup plus importants pour échantillonner les petits Métazoaires, comme les Copépodes, que les Bactéries par exemple. Le picoplancton rassemble toutes les espèces dont la taille est inférieure à 2µm, donc la majorité des Procaryotes et des Virus, ainsi que certaines espèces phytoplanctoniques notamment au sein des Chlorophytes, des Pelagophytes et des Haptophytes. La fraction de taille entre 2 et 20µm est appelée nanoplancton et rassemble la majorité du phytoplancton (Diatomées, Dinoflagellés, certains Haptophytes, Pelagophytes et Chlorophytes). Il est également possible d'y retrouver des virus géants (*Nucleocytoviricota*), qui infectent les Eucaryotes de cette gamme de taille. Le microplancton rassemble toutes les espèces dont la taille est comprise entre 20 et 200µm, c'est-à-dire une partie des Diatomées et Dinoflagellés et la majorité des hétérotrophes unicellulaires. Enfin le mesoplancton est composé des espèces dont la taille excède les 200µm, à savoir certaines Diatomées géantes et Dinoflagellés ainsi que certains Radiolaires mixotrophes mais surtout des hétérotrophes pluricellulaires (Arthropodes, Cnidaires). Pour le reste de cette thèse, il ne sera question que de la partie microscopique du plancton, qui sera appelée soit « plancton » soit « microbiome planctonique ».

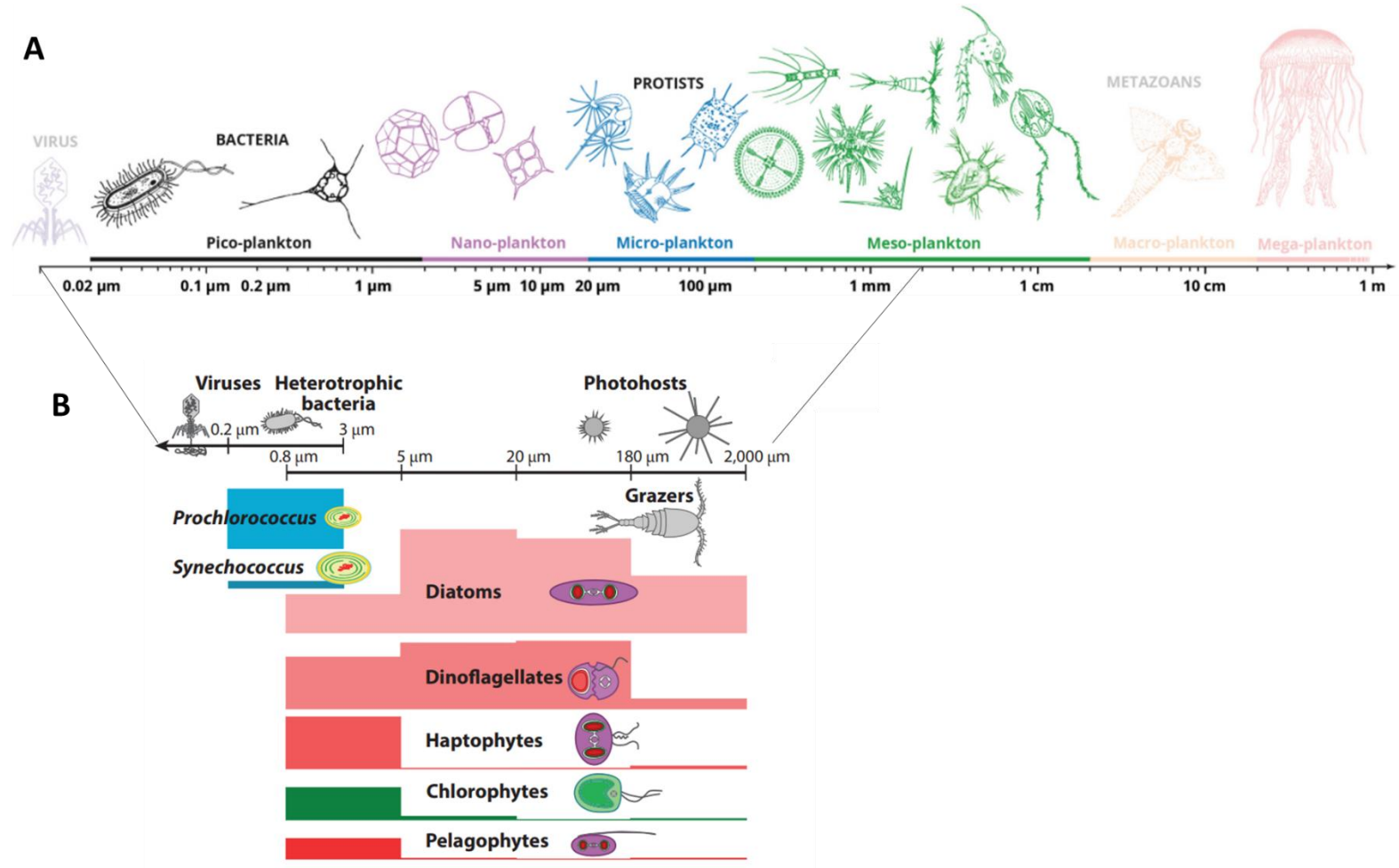


Figure 5. Le fractionnement par taille du plancton. (A) Extrait de Delmont *et al*. [5]. **(B)** Extrait de Pierella Karlusich *et al*. [55].

II. répartition biogéographique du plancton à plusieurs niveaux organisationnels : des communautés aux espèces

Ce sous-chapitre a pour but de décrire la répartition biogéographique du plancton à deux principaux niveaux organisationnels : les communautés et les espèces. Les communautés décrivent l'ensemble des espèces qui coexistent et interagissent dans un milieu donné avec ses paramètres environnementaux qui fluctuent avec le temps et la profondeur dans la colonne d'eau. Leurs caractéristiques peuvent être décrites de deux manières, qui seront abordées chacune dans les deux premières sections de cette partie: la biomasse et la productivité, qui donnent une indication sur l'importance des différents types trophiques au sein des communautés planctoniques, et la composition et la diversité, qui renseignent sur les dynamiques écologiques à une échelle plus fine. Ici la vision est plutôt écologie centrée. La troisième section de cette partie s'attachera à décrire la répartition de différentes lignées planctoniques au sein des Procaryotes, Virus et Eucaryotes indépendamment des communautés auxquelles elles appartiennent. Ici, les lignées planctoniques décrivent l'intégralité des espèces proches à un rang taxonomique donné, donc dans une approche plus taxonomie-centrée. Enfin la quatrième section visera à présenter plusieurs aspects rassemblant ces différents niveaux, à savoir l'intéraction entre diversité et productivité des communautés et le lien entre observations de diversité à l'échelle des communautés et caractéristiques des répartitions de différentes lignées.

II.1. description quantitatives des communautés planctoniques : biomasse et productivité

Les communautés peuvent être décrites par leur biomasse et leur productivité qui sont mesurées dans le milieu ou des échantillons par satellites ou méthodes de comptage, comme présenté dans I.1. Les unités utilisées sont variables, par exemple des densités en organismes par volume d'eau ou l'équivalent masse de Carbone (C). Dans cette section, je présenterai dans la première sous-section les caractéristiques de la biomasse planctonique, puis le mécanisme qui en est à l'origine, à savoir la production primaire, dans une seconde sous-partie. Enfin, je décrirai dans une troisième sous-partie le phénomène de bloom, qui est un exemple d'évènement impactant la production primaire et donc la biomasse.

II.1.a. variabilité géographique de la biomasse planctonique

La biomasse totale du plancton est estimée à environ 6 Gt_C. Cette mesure peut être découpée en fonction des types trophiques, ce qui donne une première idée des dynamiques écologiques au sein de cette biocénose. La biomasse des producteurs planctoniques est estimée à environ 0.5Gt de carbone (Gt_C) pour les Eucaryotes et 0.1Gt_C pour les Cyanobactéries [56]. Ces dernières peuvent représenter jusqu'à 15% de la biomasse bactérienne totale [56]. Pour les consommateurs, les animaux planctoniques représentent près de 2Gt_C, les Eucaryotes unicellulaires et Bactéries (dont environ 10% sont des SAR11) 1.5Gt_C chacun et les Archaea et Fungi 0.3Gt_C chacun [56]. Les virus ne contribuent qu'à hauteur de 0.03Gt_C à la biomasse totale, mais dominent néanmoins largement en terme de nombre d'unités (un ordre de grandeur de plus que les Bactéries et les Archées ensemble) [56], [57]. La relation entre biomasse du phytoplancton et biomasse du zooplancton est une fonction croissante positive, atteignant un plateau pour des valeurs de biomasse phytoplanctonique supérieure à 100$mg_C.m^{-3}$ [58].

La distribution de la biomasse dans les océans n'est pas homogène et est en partie structurée par la géographie. Globalement, elle a tendance à être plus élevée dans les gyres subpolaires et les pôles,

les régions tempérées, les côtes et les upwellings et moins élevée dans les gyres subtropicaux [59], [60], [61], [62]. Pour les Bactéries, elle n'est pas structurée par les provinces de Longhurst [63]. La concentration maximale de cellules de Picoeucaryotes et de Bactéries en surface à l'échelle globale est atteinte à des latitudes comprises entre 0 et -40° et est d'environ 10^6 cellules.mL^{-1} [64]. La variabilité de la concentration en cellules des Bactéries à l'échelle globale est élevée, avec une moyenne à $3.3.10^5 \pm 0.4.10^5$ cellules.mL^{-1} [65], [66]. Les Cyanobactéries atteignent leur concentration maximale entre 20 et -20° de latitude, à près de 2.10^5 cellules.mL^{-1}. *Prochlorococcus* est le genre dominant au sein des Cyanobactéries, pouvant représenter jusqu'à de ⅓ de toutes les cellules de Bactéries sous ces latitudes. Sa concentration peut y atteindre jusqu'à $1.5.10^5$ cellules.mL^{-1} [66], [67], [68], et reste de l'ordre de 10^5 cellules.mL^{-1} jusqu'à près de 150m de profondeur [68]. Le deuxième genre le plus important en terme de biomasse au sein des Cyanobactéries est *Synechococcus*. Sa concentration est beaucoup plus stable à l'échelle globale, d'environ 5.10^4 cellules.mL^{-1} sauf aux pôles [67]. Les Cyanobactéries sont généralement légèrement moins concentrées à la DCM qu'en surface, alors que les bactéries hétérotrophes y sont beaucoup plus nombreuses [64]. La concentration des Eucaryotes est en général plus faible, en moyenne de $1.7.10^3 \pm 1.5.10^3$ cellules.mL^{-1}, dont la majorité sont des Picoeucaryotes. La concentration des Coccolithophores, Diatomées et Dinoflagellés est de l'ordre de quelques cellules par mL [65], [66]. Le picophytoplancton et le nanophytoplancton représentent près de 80% de la biomasse phytoplanctonique dans les zones oligotrophes [69]. Le picophytoplancton représente plus de 50% de la biomasse quand la biomasse totale du phytoplancton est inférieure à 10mg$_{Chl}$.m^{-3} [66]. Les tendances à la DCM sont globalement les mêmes qu'en surface mais les concentrations y sont supérieures, avec des maxima atteints entre 0 et 20° N et S aux alentours de $2.5.10^6$ cellules.mL^{-1} (Figure 6 B).

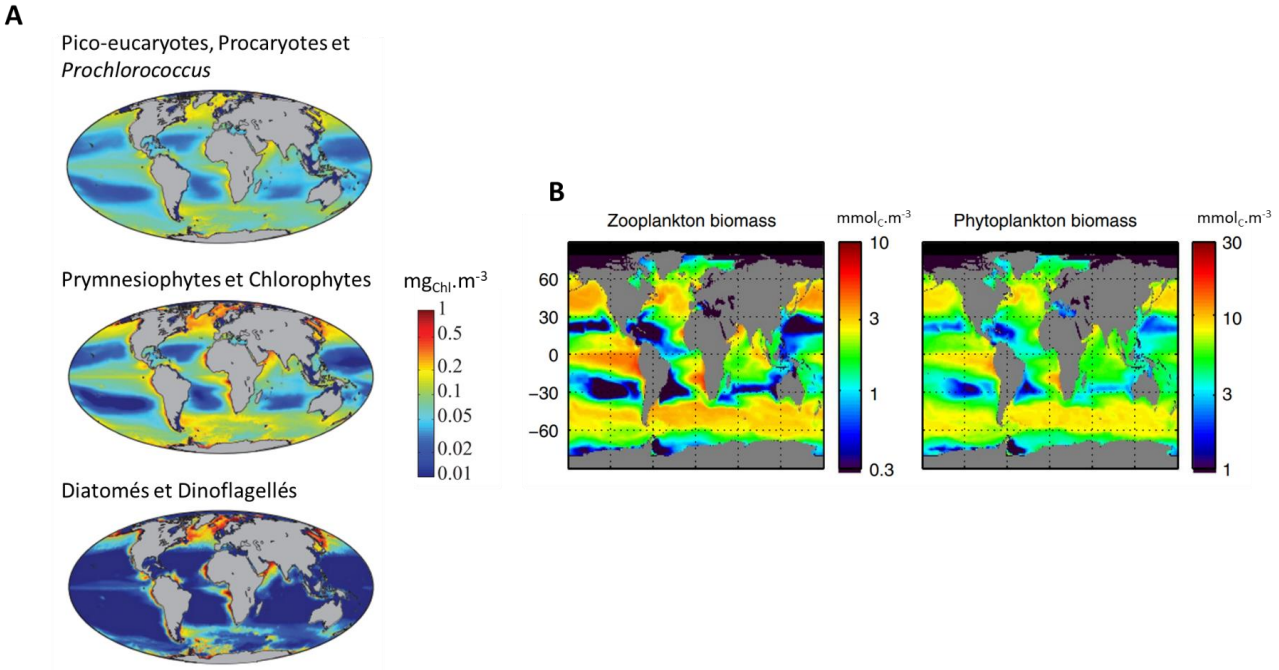


Figure 6. Biomasse et productivité de différents groupes planctoniques. (A) Productivité en concentration de chlorophylle (mg$_{Chl}$.m^{-3}) mesurée par satellite dans différentes lignées du phytoplancton. Adapté de Ward *et al*. [70] avec les données de Hirata *et al*. [71]. **(B)** Biomasse moyenne annuelle du zooplancton et du phytoplancton en concentration de carbone (mmol$_C$.m^{-3}), estimée par un modèle incluant deux fractions de taille de phytoplancton (16 *Prochlorococcus* et 16 *Synechococcus* dans la petite fraction de taille et 16 flagellés et et 16 diatomées dans la grande fraction) et deux fractions de taille de zooplancton (micro- et mesozooplancton). Extrait de Vallina *et al.* [61].

II.1.b. production primaire planctonique

La biomasse des communautés planctoniques résulte directement de leur productivité. La production primaire planctonique est à la base de tous les réseaux trophiques marins. Cette production est découpée en deux grandeurs : la production primaire brute, qui décrit l'intégralité de l'énergie assimilée par les producteurs primaires (en surface essentiellement le phytoplancton), et la production primaire nette, qui décrit la quantité de biomasse effectivement intégrée dans l'écosystème (énergie assimilée par la production primaire brute moins énergie consommée par la respiration). Dans les deux cas, l'unité est une masse par unité de surface et par unité de temps ($g_C.m^{-2}an^{-1}$ par exemple). Du point de vue fonctionnel, les producteurs primaires des communautés planctoniques de surface sont des photoautotrophes Eucaryotes comme les Prymnesiophytes, les Diatomées, les Chlorophytes et certaines Dinoflagellées, et des photoautotrophes Procaryotes qui sont principalement des Cyanobactéries. C'est de la production primaire nette (qui sera simplement appelée « production primaire » ici) dont il est question dans cette partie.

Comme décrit dans la sous-section précédente, la faible quantité de biomasse des producteurs planctoniques (environ 0.6 Gt_C) comparée à celle de tous les consommateurs (5.5 Gt_C) est compensée par le remplacement très rapide des producteurs primaires, qui est de l'ordre de la semaine [56], [60]. Cela donne une indication numérique de l'importance de la production primaire planctonique. Elle peut être approximée via la mesure de la concentration en chlorophylle dans le milieu. En réunissant toutes les fractions de tailles et types de chlorophylle, elles peuvent varier de $5mg_{Chl}.m^{-3}$ [60], [72] jusqu'à $25mg_{Chl}.m^{-3}$ [73]. Chaque fraction de taille contribue à la production primaire de façon différente. Dans le picophytoplancton, les concentrations en chlorophylle A varient entre 0.05 et $2mg_{ChlA}.m^{-3}$, sans structuration géographique nette à l'exception des maxima qui sont atteints près des côtes, surtout en Arctique. C'est cette fraction de taille qui contribue le plus à la production primaire aux latitudes absolues inférieures à 40°, réalisant plus de 50% de la photosynthèse dans les régions où la biomasse est inférieure à $100mg_Cm^{-3}d^{-1}$ [66], [74], [75]. Pour le nanophytoplancton, la concentration de chlorophylle A est maximale aux pôles aux alentours de $0.5mg_{ChlA}.m^{-3}$. Elle décroît ensuite pour atteindre un minimum entre 40 et 20°N et S (entre 0.001 et $0.005mg_{ChlA}.m^{-3}$) et remonte à près de $0.05mg_{ChlA}.m^{-3}$ à l'équateur [74], [75]. Dans le microphytoplancton, les concentrations en chlorophylle A ont globalement la même distribution biogéographique que dans les deux fractions de taille évoquées précédemment, mais avec des maxima et minima plus extrêmes : jusqu'à $2mg_{ChlA}.m^{-3}$ dans l'Océan Austral et moins de $0.003mg_{ChlA}.m^{-3}$ à environ 30°S [74], [75]. Le microphytoplancton est plus productif dans les milieux riches en biomasse [73]. La productivité des Diatomées est par exemple particulièrement importante aux hautes latitudes [76] (Fig.6 A). À noter que toutes ces mesures sont des moyennes et en général soumises avec une intensité plus ou moins forte à la saisonnalité [75]. Cet effet est particulièrement important aux pôles et dans les régions tempérées où les maxima de productivité ne sont atteints qu'à certaines périodes (entre juillet et décembre pour l'Arctique par exemple). Elle l'est un peu moins dans l'Océan Austral où les zones de fortes productivité sont plus ponctuelles et localisées. Elle n'a pas d'impact sur les côtes ni sous les tropiques, où la productivité est respectivement toujours forte et toujours faible.

À noter que la production primaire, dont le produit est du carbone organique, ne peut fonctionner de manière complètement autonome. La photosynthèse et plus généralement la survie des producteurs primaires requiert en effet que différents nutriments soient à disponibilité librement dans le milieu, l'un des plus importants étant l'azote. Cet élément est rendu biodisponible dans le milieu par des espèces dites diazotrophes, dont le rôle est crucial dans le cycle de l'azote. Les espèces de ce type trophique sont essentiellement des cyanobactéries (notamment *Trichodesmium*) ainsi que d'autres lignées bactériennes et des Archées [77], [78], [79], [80], [81].

II.1.c. un évènement majeur modifiant régulièrement la biomasse et la production primaire planctonique : description du phénomène de bloom

Les blooms ont aussi des impacts fort sur la production primaire et la biomasse, surtout pendant de courtes périodes (quelques semaines) au printemps et en automne dans les régions tempérés, au cours desquelles un nombre important d'espèces de tous les domaines du vivant et des Virus se succèdent [61], [82], [83]. Globalement, seules les espèces à croissance rapide, relativement grandes et avec des défenses contre la prédation arrivent à dépasser des concentrations de plus de 100mg.m^{-3}. Ce sont principalement des Dinoflagellés, des Diatomés, des Coccolihtophoridés ou des *Phaeocystis* coloniaux [58]. Les blooms sont généralement causés par des changements importants de concentrations en nutriments ayant lieu à de grandes échelles spatiales, ces concentrations étant souvent le facteur limitant la productivité des photoautotrophes [84]. Les variations de ces concentrations peuvent avoir plusieurs origines. Dans les milieux tempérés, c'est en général la fonte des glaciers d'hiver des régions polaires (surtout au Nord), causée par l'augmentation de l'irradiance et de la température. Dans les milieux tropicaux, une cause peut être l'«Island Mass Effect», qui consiste en un apport de nutriment par les îles tropicales (particulièrement celles avec un volcanisme actif) dans le milieu marin. Il peut en résulter plusieurs blooms successifs, les premiers étant côtiers et les suivants en pleine mer, par exemple des milieux pauvres en nutriments comme les gyres oligotrophiques [85], [86].

II.2. composition taxonomique et variabilité des communautés planctoniques

La première section de ce sous-chapitre visait à décrire quantitativement les communautés planctoniques. Dans cette section, je rentrerai plus en détail dans la composition des communautés planctoniques et comment celles-ci varient dans les océans. Si le premier sous-chapitre de ce chapitre (« description et classification des espèces planctoniques », p.15) avait pour but de fournir un inventaire le plus complet possible de toute la diversité des espèces planctoniques, le point de cette sous-section est plutôt d'inscrire ces espèces dans un contexte écologique et de décrire leurs différentes combinaisons et dynamiques au sein de leurs différentes communautés dans les océans. Les communautés peuvent être composées de différentes sous-communautés, assemblages ou modules au sein desquels les espèces interagissent plus entre elles qu'avec celles d'autres sous-communautés ou modules [36]. La modularité décrit le nombre de modules au sein d'une même communauté [87]. Les interactions entre espèces au sein d'un module peuvent être évaluées par des mesures de co-occurrence, qui en sont dans une certaine mesure un indicateur (malgré de possible biais, notamment lorsque des espèces partagent une niche très proche sans que leur niveau de compétition ne soit trop fort). La modularité des communautés ainsi que leur distribution peuvent varier en fonction des données et méthodes utilisées pour les définir.

Pour décrire la composition et la variabilité des communautés planctoniques, je commencerai par décrire la diversité des communautés planctoniques et comment celle-ci varie avec la latitude en surface, mais également quels sont les biais potentiels qui lui sont associés. Dans une deuxième sous-section, je présenterai les différences de répartition des communautés en fonction de la taille des espèces considérées. Enfin, la troisième sous-section sera consacrée à une description plus détaillée des différentes communautés planctoniques ainsi que leur variation avec la latitude, la profondeur dans la colonne d'eau et la saisonnalité.

II.2.a. variabilité géographique de la diversité du plancton

Le terme « diversité » peut se référerer à deux grandeurs : la macrodiversité, diversité spécifique, ou richesse, qui dénombre les espèces différentes ; et l'α-diversité, souvent mesurée avec l'indice de Shannon, qui prend en compte à la fois le nombre d'espèces mais aussi leurs abondances [88].

En surface, l'α-diversité du plancton varie avec la latitude de manière symétrique par rapport à l'équateur, avec une tendance globale à l'augmentation plus la latitude diminue chez les Eucaryotes et les Procaryotes [30], [37], [64], [87], [89], [90], [91], [92], [93], [94], [95]. Les maxima sont atteints à des latitudes différentes selon les lignées (Figure 7 A). Ils se situent vers 40° N et S chez les Procaryotes [30], les Archées et les Eucaryotes non phototrophes [63], [64], [96]. L'α-diversité décroit plus fortement avec la distance à l'équateur dans l'hémisphère sud pour les Bactéries et les Archées, chez qui elle est divisée par deux entre 43°S et 66°S [64], [94], [96]. Pour le phytoplancton (Procaryote et Eucaryote), le maximum de diversité est atteint à l'équateur [64], [95]. Il n'existe pas de relation entre diversité du phytoplancton et du zooplancton [58]. Globalement, le minima de diversité de tous les domaines du vivant est atteint aux pôles [30], [37], [64], [89], [90], [91], [92], [93], [94], [95]. Il n'y pas de gradient latitudinal clair de diversité dans certains groupes de Virus [64], [97]. À une échelle géographique plus fine, l'α-diversité des communautés Arctiques est moins élevée que celle des communautés de l'Atlantique Nord, malgré leur connectivité par les courants [91], [98]. Les courants de frontière d'ouest (Gulf Stream, Kuroshio) ainsi que les upwellings costaux sont généralement des régions de haute diversité pour les Diatomées, de Coccolithophores et de Dinoflagellés [76]. Au contraire, leurs minimas de diversité sont globalement localisés dans les gyres oligotrophiques des subtropiques [76], [99] (Figure 7 B). Certains systèmes océaniques sont aussi caractérisés par des variations importantes de diversité spécifique sur des petites échelles géographiques. Dans le courant des Aiguilles par exemple, la diversité est différente de part et d'autre des zones d'étranglement autant pour les Eucaryotes que les Procaryotes [100]. À une échelle encore plus petite, le mixage turbulent a un effet sur la diversité du phytoplancton, qui est en général plus élevée aux frontières des eddies que dans leur centre [62]. L'α-diversité peut aussi varier à une localisation donnée en fonction de la profondeur dans la colonne d'eau, chaque zone influant sur la diversité des autres [30], [37]. L'α-diversité est en général maximale juste en dessous de la Mixed Layer Depth (MLD) et dans le benthos dans les trois domaines du vivant [37], [91], [96], [101]. Chez les Eucaryotes, la diversité spécifique des espèces exclusivement trouvées à la DCM peut représenter jusqu'à un tiers de la diversité spécifique totale [91]. La diversité spécifique des NCLDV est plutôt maximale à la surface et décroît avec la profondeur [34], [97]. Dans les communautés profondes, la diversité globale des Procaryotes est relativement faible [102]. À noter que le gradient latitudinal de diversité observé à la surface reste globalement le même jusqu'à la DCM, puis a tendance à s'estomper à de plus grandes profondeurs [64]. La diversité peut également varier en fonction des saisons. Le gradient latitudinal de macrodiversité d'une partie du Bactérioplancton est par exemple différent en hiver et en été. La diversité est maximale en Arctique en hiver et décroît jusqu'en Antarctique ; en été, elle est maximale dans le milieu tempéré de l'hémisphère sud et décroit jusqu'en Arctique [89].

Les mesures de diversité dans les microbiomes marins sont soumises à des biais importants, notamment causés par l'effort d'échantillonnage et de séquençage, qui ne permet en général pas de détecter les espèces rares qui représentent la majorité de la macrodiversité totale de la communauté, et dont l'importance écologique et biogéochimique est probablement sous-évaluée [103], [104], [105]. L'absence de certaines espèces dans un milieu donné traduit beaucoup plus souvent un biais d'échantillonnage qu'une réalité biologique [106]. Ce constat s'inscrit dans la continuité de l'hypothèse « everything is everywhere but the environment selects » selon laquelle toutes les espèces du microbiome marin sont présentes partout, et seules leurs abondances varient [15], [18]. La difficulté

vient essentiellement du fait que la probabilité d'échantillonner une espèce très rare est basse, et que la génomique ne permet pas d'étudier des espèces dont les abondances sont trop faibles. Elle ne permet pas non plus de séparer les espèces rares et actives des espèces abondantes ou rares mais inactives [107], [108]. En intégrant les effectifs estimés des espèces rares, les gradient latitudinaux de diversité disparaissent pratiquement pour le nano- et le microphytoplancton [105]. Les autres biais sont d'ordre temporels. Par exemple, une partie des études se concentrant la diversité du Bactérioplancton des hautes latitudes utilise des données produites lors d'expéditions qui se sont déroulées en été, causant une surestimation de la diversité tropicale [89], [93], [109].

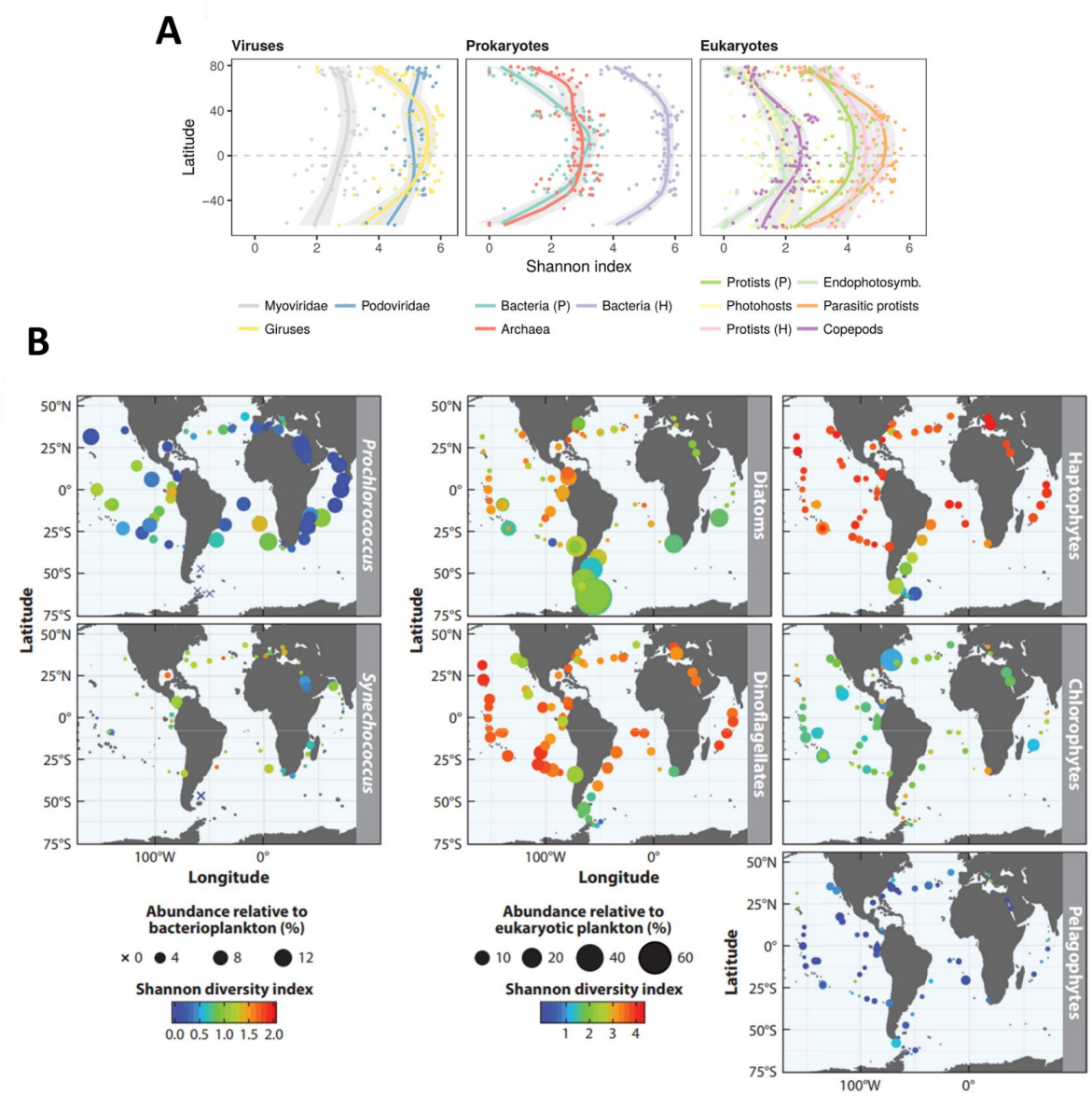


Figure 7. Diversité de différentes lignées planctoniques. (A) Gradient Latitudinal de Diversité (LDG) de différents groupes planctoniques dans les stations TO, calculé avec l'indice de Shannon. Extrait de Ibarbalz *et al.* [64]. **(B)** Abondance et α-diversité de différentes lignées phytoplanctoniques. À gauche, pour les principales picocyanobactéries à partir des données de Farrant *et al*. (fragments du gène *petB* provenant de métagénomes de la fraction de taille 0.22-3µm après séquençage Illumina)[110]; à droite, pour les principaux Eucaryotes à partir des données de De Vargas *et al*. (metabarcoding de la région V9 du gène de l'ARNr 18S, dans la fraction de taille 0.8-2000µm)[38]. La taille de chaque cercle correspond à l'abondance dans chaque station, et la couleur à la valeur de l'indice de Shannon. Extrait de Pierella Karlusich *et al.* [55].

II.2.b. les communautés planctoniques sont structurées différemment en fonction de la taille des organismes

Le nombre et la distribution des communautés dépend en partie de la taille des espèces considérées. À l'échelle planétaire, il existe plus de communautés d'espèces de petits organismes que de grands. L'aire de répartition des communautés de grandes espèces est généralement plus vaste que celui des communautés des petites espèces [38], [98] (Figure 8).

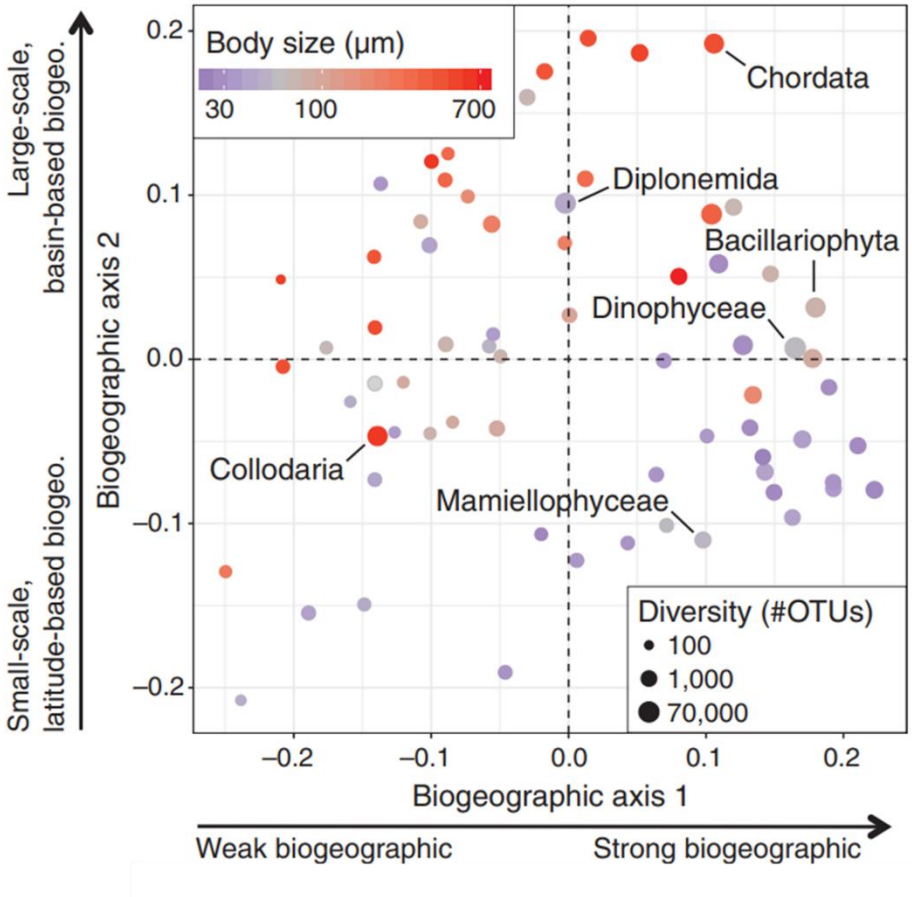


Figure 8. Effet de la taille des organismes sur l'intensité et le type de structuration biogéographique des lignées planctoniques. Analyse en Coordonées Principales des dissimilarités biogéographiques entre 70 groupes planctoniques couvrant le spectre complet des répartitions, d'une faible structuration spatiale (à gauche) à une structuration forte, à l'échelle des bassins (en haut à droite) ou à plus petite échelle et selon la latitude (en bas à droite). Chaque point correspond à la projection du groupe dans les deux premiers axes de variation; la taille reflète le log de la diversité, et la couleur le log de la taille moyenne des individus. Extrait de Sommeria-Klein *et al*. [98].

Pour les espèces dont la taille des organismes est comprise entre 20 et 2000µm, la structuration se fait principalement par biomes et bassins avec trois communautés très distinctes : polaire, tempérée et tropicale [111]. L'impact de la géographie sur la structuration est très fort [38]. Les communautés sont également distinctes entre le Pacifique et l'Atlantique, les continents les séparant représentant des barrières importantes à leur diffusion [98]. En revanche, des observations à des échelles plus réduites (de l'ordre d'une centaine de kilomètres) montrent une structuration plus importante que pour les communautés des plus petites fractions de taille [112]. Le nanoplancton (5-20µm) est séparé en trois communautés mais qui correspondent plutôt aux zones eutrophes, oligotrophes et tempérées et polaires, en cohérence avec leur mode trophique [111]. Ces fractions de tailles sont affectées par le

mixage turbulent qui génère une structuration régionale importante et maintient la diversité, tout en conservant une divergence entre sous-communauté plus faibles que les niveaux observés pour deux communautés provenant de bassins différents [62]. Les organismes de moins de 5µm ont une distribution beaucoup plus fractionnée et peu structurée par les biomes et les bassins chez Mamiellophyceae par exemple [98], avec jusqu'à sept communautés distinctes [111] (Figure 9).

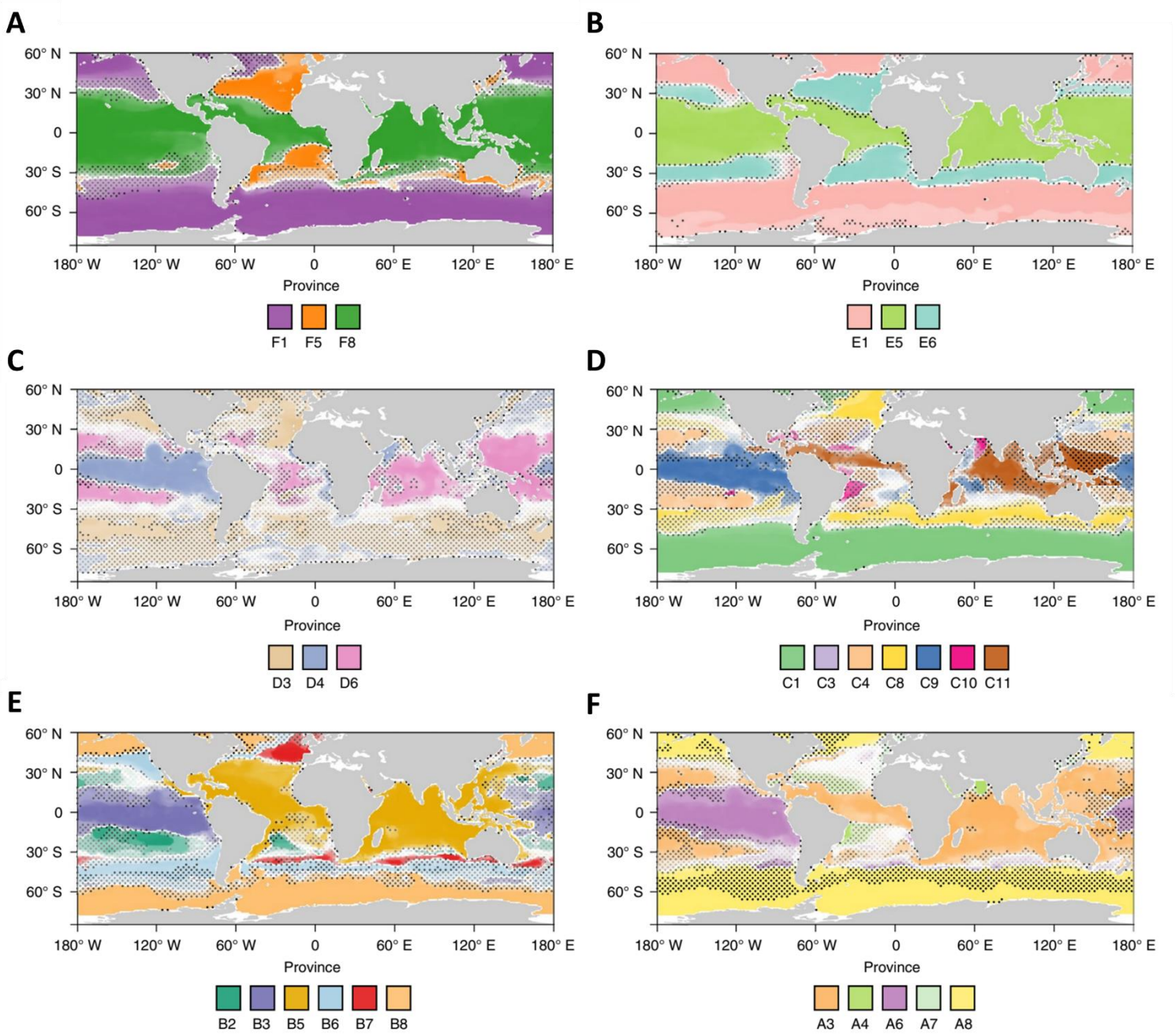


Figure 9. Biogéographie globale des provinces planctoniques en fonction de la taille des organismes. Pour les fractions : **(A)** enrichie en Metazoaires (180-2000µm); **(B)** enrichie en petits Metazoaires (20-180µm); **(C)** enrichie en protistes (5-20µm); **(D)** enrichie en protistes (0.8-5µm); **(E)** enrichie en Bactéries (0.22-3µm); **(F)** enrichie en virus (0-0.2µm). À chaque point de la grille sur la carte, l'intensité de la couleur de la province dominante représente sa probabilité de présence. Les points indiquent les zones d'incertitude (où le delta de probabilité de présence entre la province dominante et les autre est inferieur à 0.5). Extrait de Frémont *et al*. [111].

II.2.c. composition et variabilité des communautés planctoniques

- **variabilité de la composition des communautés de surface avec la géographie**

Plusieurs communautés planctoniques peuvent être distinguées en fonction des méthodes utilisées pour les définir. Il en existe seize [98] ou dix en ne considérant que les Procaryotes [87], six [113] ou cinq si on considère les Procaryotes et les Eucaryotes [74], [92]. Les assemblages sont souvent structurés par latitude et par bassin, avec une symétrie de part et d'autre de l'équateur et une cohérence globale avec les provinces de Longhurst [96], [111] [19]. Cela indique un couplage important avec les paramètres environnementaux et des capacités de dispersion réduites entre bassins. Le changement le plus important de communauté dominante s'observe pour tous les groupes au passage du front polaire, c'est-à-dire au-delà de 60° de latitude nord et sud [98]. Une partie des changements sont observés de manière symétrique dans les deux hémisphères. L'Arctique présente néanmoins des spécificités : contrairement à l'océan Austral, qui est ouvert sur plusieurs autres océans à la fois et a une connectivité globale élevée, l'accès à l'Arctique se fait en grande partie par l'Atlantique via la dérive nord Atlantique.

Les communautés de Procaryotes planctoniques étudiées via barcode sont composées en moyenne d'environ 90% d'hétérotrophes et 0 à 10% de photoautotrophes, dont les proportions sont maximales entre 0 et -20° de latitude (avec des maxima locaux pouvant aller au-delà de 10%) (Figure 10 A) [64]. D'un point de vue taxonomique, elles sont composées d'au moins 50% de Pseudomonadota [30]. Plus précisément, les Alphaproteobactéries représentent en moyenne entre 40 et 60% des communautés, avec une forte proportion de SAR11, et les Gammaprotéobactéries entre 10 et 20% [30], [63], [114], [115], [116], [117], [118]. Elles sont bien représentées aux pôles (particulièrement les Alteromonadales et Oceanospirillales) ainsi que dans les milieux oligotrophes occupent une plus petite part dans le biome tempéré [115], [116], [117]. En dehors des pôles, 10 à 20% des communautés sont composées de Deltaproteobacteria, Defferibacteres, Actinobacteria (qui sont présentes en Arctique), Verrucomicrobia, Chloroflexia et Planctomycetes [30], [115], [116]. Ces dernières peuvent à elles seules représenter plus de 20% des communautés bactériennes dans les zones côtières de l'Atlantique Nord [63], [119]. Dans le FCB group, les Bacteroidetes (particulièrement les Flavobacteriia) peuvent représenter jusqu'à 25% des communautés de l'Océan Austral [30], [100], [116], [117], [118]. Les communautés de cet Océan rassemblent également une proportion plus élevée de Betaprotéobactéries et Acidobactéries qu'ailleurs [116]. Enfin, les communautés sont aussi composées de Cyanobactéries, dans des proportions variables qui peuvent aller jusqu'à 35% dans les biomes tempérés et tropicaux, en particulier dans les upwellings, alors qu'elles sont pratiquement absentes des communautés polaires [30], [64], [70], [74], [100], [114]. Les Archées marines de surface sont essentiellement des Thaumarchaeota et Euryarchaeota, qui ne représentent jamais plus de 10% des communautés Procaryotes, en particulier aux pôles où elles sont quasiment absentes (Fig.7 A) [30], [64]. Ces compositions globales regroupent plusieurs modules distincts de Procaryotes. Celles du biome tropical sont globalement riches en Cyanobactéries et en d'Alphaproteobactéries (SAR11, Puneispirillales, Rhodospirillales) et contiennent également des Gammaproteobactéries et certains groupes d'Archées (Marine Group II des Euryarchaeota). Le biome polaire rassemble trois modules distincts constitués de Flavobacteriales, SAR11, 86 et 406, Puneispirillales, Rhodobacterales asinsi que des Archées du Marine Group II des Euryarchaeota [87]. Les communautés de l'Arctique et de l'Antarctique partagent un degré de similarité plus élevées entre elles qu'avec les communautés des régions plus proches géographiquement (Atlantique Nord et Pacifique Sud, respectivement), mais sont également bien distinctes entre elles : 78% des OTUs de l'Océan Austral ne sont pas retrouvés en Arctique et 70% des OTUs Arctiques ne sont pas retrouvés dans l'Océan Austral. Les deux régions ne partagent que 15% de leurs OTUs respectifs, qui sont principalement des SAR86 ou 92 et des

Polaribacter dont les abondances peuvent atteindre jusqu'à 0.1% des lectures Procaryotes [63], [90], [116], [120]. Les milieux épipélagiques profonds rassemblent trois modules contenant des lignées de Gammaproteobacteria, Nitrososphaeria et Thermoplasmata [87].

De façon générale, les communautés Eucaryotes de surface sont composées d'entre 10 et 25% de phototrophes (les maxima étant atteints aux pôles), le reste étant des hétérotrophes (Fig.7 A) [64], [70], [105]. À une échelle plus fine, elles rassemblent plusieurs modules distincts (16 par métabarcoding)[36]. Aux pôles, les communautés rassemblent plusieurs modules souvent exclusifs, ou présents partout mais avec une abondance plus élevée dans ce biome (Fig.7 B) [98], [113]. Ils sont globalement constitués de Dinoflagellés, d'Athropodes et du plusieurs lignées phytoplanctoniques (Cryptomonadales, Prymnesiophyceae, Mamiellophyceae, Dictyochophyceae) [113]. De la même façon que pour les Procaryotes, ces modules partagent plus de similarités entre eux qu'avec d'autres assemblages plus proches géographiquement, avec néanmoins des différences entre Arctique et Antarctique [74], [91], [92], [98], [113], [116]. Dans l'Arctique, deux modules composés d'espèces trouvées également dans d'autres océans sont majoritaires [74], [91]. Cinq modules sont présents en Antarctique, l'un d'entre eux étant le module majoritaire à l'échelle globale et un autre étant dominé par des Bacillariophytes [74], [113]. Le module majoritaire à l'échelle globale (sauf dans les zones côtières et les zones oligotrophes) est constitué essentiellement d'hétérotrophes et de mixotrophes (Dinophyceae, MALV I et II, Copepoda et Collodaria). Les Prymnesiophytes y représentent l'autotrophe le plus important mais leur abondance relative est faible en comparaison des autres groupes. [74], [92], [98], [113]. Malgré le fait que la majorité des assemblages soient dominés par le même module, leur composition est dans une certaine mesure structurée par latitude et par bassin [98]. Les assemblages du Pacifique tropical contiennent des modules non dominants plus riche en Metazoaires que ceux des océans Atlantique et Indien. Les modules non dominants des milieux tempérés et non tempérés se distinguent essentiellement par leurs différences de ratio hétérotrophes/phototrophes, l'importance des métazoaires et les espèces phototrophes présentes[98], [113].

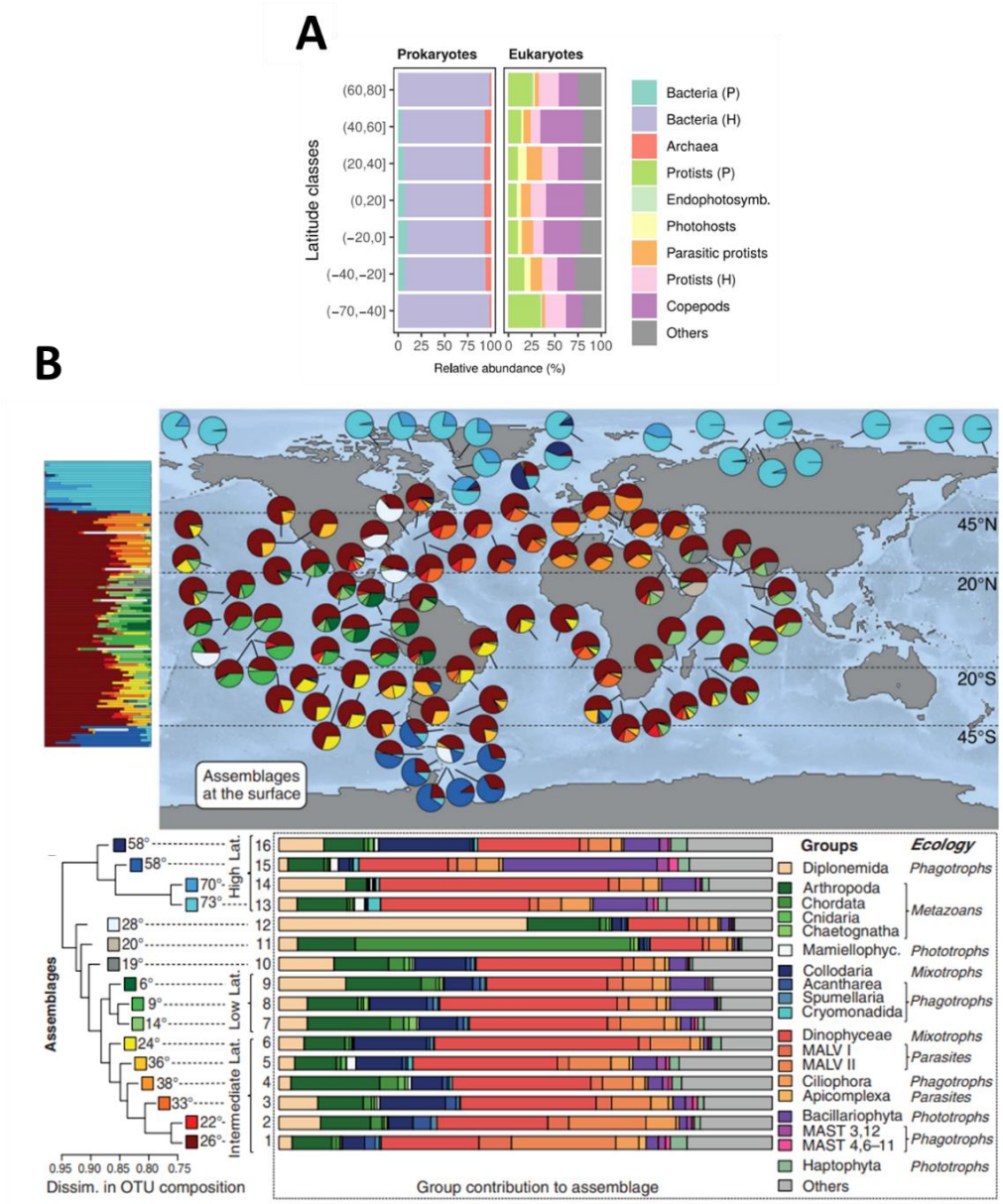


Figure 10. Exemples d'assemblages planctoniques et de leur distribution biogéographique. (A) Abondance relative moyenne des différents groupes planctoniques en fonction de la latitude. Procaryotes : gène ARNr 16S, fraction de taille 0.22-3µm ; Eucaryotes : gène ARNr 18S, 0.8-2000µm. Les groupes viraux ne sont pas représentés, leurs données d'abondance n'étant pas comparables aux autres. Extrait de Ibarbalz *et al*. [64]. **(B)** Répartition et composition des 16 assembalges d'OTUs co-fréquents dans le plancton Eucaryote. Chaque assemblage est associé à une couleur et à un nombre. En haut, contribution des assemblages aux communautés planctoniques de surface. En dessous de la carte, à gauche, le dendrogramme indique la dissimilarité taxonomique des assemblages, calculée avec l'indice de Simpson. La latitude moyenne absolue de chaque assemblage est indiquée. Le barplot à droite du dendrogramme montre la proportion de chacun des 19 principaux groupes Eucaryotes rassemblant plus de 1000 OTUs, groupés par proximité phylogénétique, dans les assemblages. Leurs types trophiques principaux sont indiqués en italique, à droite. Extrait de Sommeria-Klein *et al*. [36]. À noter que les observations des communautés via satellite ou métagénomiques par exemple, peuvent donner un aperçu sensiblement différent [74], [113], [121].

- **variabilité des communautés planctoniques avec la profondeur dans la colonne d'eau**

À une localisation donnée, les communautés sont différentes en fonction de la profondeur dans la colonne d'eau, mais restent plus similaires entre elles qu'avec des communautés éloignées géographiquement ou à la même localisation mais à une saison différente, en particulier pour les Procaryotes [64], [116]. La stratification verticale des communautés s'observe entre la surface, la DCM (Profondeur de concentration Maximale en Chlorophylle, n'existe pas toujours et est souvent à la même profondeur que la nutricline), la zone photique (entre la DCM et 200m) la zone mésopélagique (entre 200 et 1000m) et la zone benthique (entre 1000 et 4000m) [30], [37], [87], [90], [101], [117], [118], [122], [123] (Fig.10). La connectivité entre ces zones se fait principalement de la surface vers les profondeurs, par le phénomène de sédimentation au cours duquel des particules mortes de plancton coulent vers les abysses, entraînant avec elles un consortium de parasites, commensaux, Bactéries et Virus, expliquant la présence d'espèces phytoplanctoniques dans le plancton en cours de sédimentation et dans les sédiments benthiques (Fig.10). L'intensité des transferts dans ce sens est important : la proportion d'ADN pélagique dans le benthos est d'en moyenne 21%, et peut atteindre jusqu'à 50% sous les hautes latitudes [37]. En outre, des remontées sont aussi possibles notamment grâce aux courants ascendants ainsi que le transport par la mégafaune. La structuration des communautés benthiques suit globalement les provinces biogéographiques abyssales et la proportion d'espèces partagées entre régions profondes diminue avec la distance géographique comme en surface, avec une variabilité pouvant être importante à des échelles locales [37].

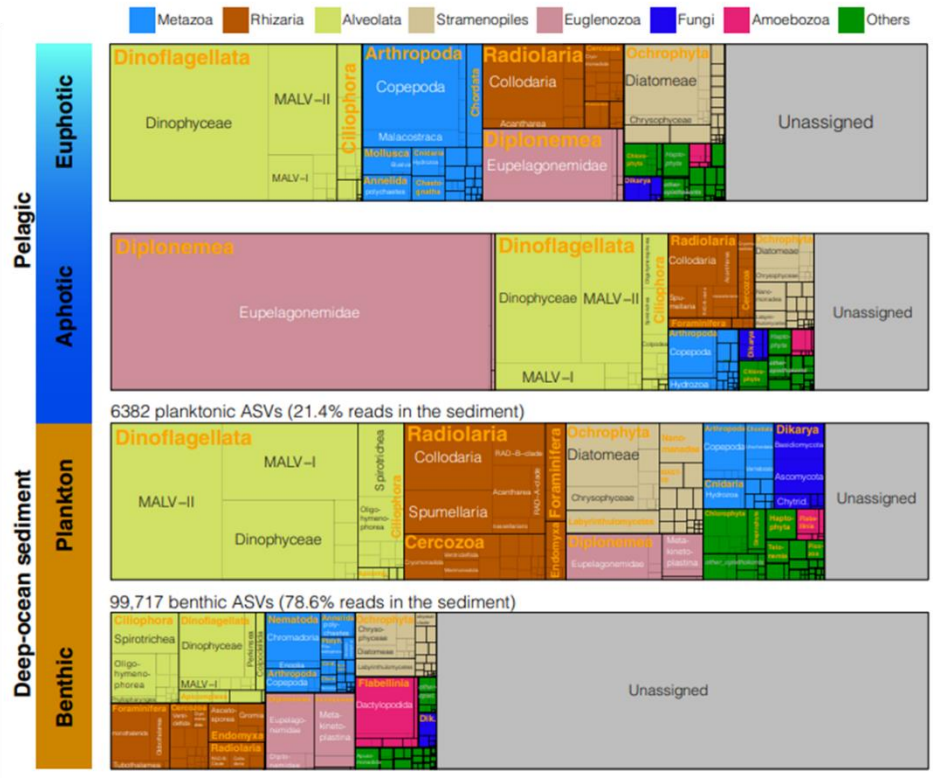


Figure 11. Variabilité de la composition des communautés d'Eucaryotes en fonction de la profondeur dans la colonne d'eau. La taille de chaque rectangle représente une proportion de macrodiversité. La catégorie « Deep-ocean sediment » « plankton » correspond à la communauté planctonique en cours de sédimentation, avant qu'elle n'atteigne le benthos. Extrait de Cordier *et al.* [37].

En considérant les communautés bathypélagiques complètes (tous les domaines du vivant et les NCLDV), 36 modules différents peuvent être distingués dans le milieu mésopélagique. Trois sont composés exclusivement d'espèces ne vivant qu'à cette profondeur [123]. Chaque domaine du vivant possède néanmoins des communautés spécifiques aux différentes profondeurs.

Chez les Eucaryote, la zone photique est dominée par les Dinophycées, les Copépodes et les Radiolaires Collodaires pour les hétérotrophes et les Diatomées pour le phytoplancton (Fig.10), à l'exception de la DCM où les Pelagophycées et Mamiellophycées peuvent ponctuellement dominer le phytoplancton et où les Radiolaires Polycystines représentent une plus grande part des hétérotrophes [37], [124]. De plus, certaines espèces ne sont présentes à la DCM que sous certaines latitudes, comme c'est le cas du petit zooplancton omnivore n'y est présent qu'entre 0 et 20° de latitude absolue [64]. Les communautés plus profondes ont également des spécificités, à commencer par une disparition du phytoplancton dès que la lumière ne pénètre plus. Dans le milieu bathypélagique, les communautés sont par exemple composées d'une plus grande fraction de Basidiomycètes, contrairement aux Dinoflagellés, MAST-1 MALV-I, MAST-4 dont l'abondance décroit avec la profondeur [124].

Dans les communautés bactériennes, le module dominant entre 20m et la DCM regroupe des Alphaproteobacteria (SAR11, Rickettsiales, Puniceispirillales, Rhodospirialles, Rhodobacterales, Pelagibacterales), des Cyanobactéries et des Flavobacteriales, avec des différences entre biome polaires et non polaires [30], [87], [115]. Les différences entre communautés sont en général plus marquées à la DCM qu'à la surface [87]. Dans la zone photique en dessous de la DCM, des différences sont observées entre Atlantique et Pacifique. Le module dominant en Atlantique est constitué de différents groupes de SAR, de Nitrosopumilales et de Synechococcales, et n'est pas dominant dans le Pacifique où le module principal est plutôt composée d'Alteromonades et de Sphingomonadales [87]. Les Cyanobactéries et Pelagibacter sont présentes dans toutes les communautés de cette profondeur et disparaissent en-dessous [30], [115]. À la couche mésopélagique, les communautés polaires ont plusieurs particularités. Près de 40% des OTUs de l'océan Austral et 60% des OTUs de l'Arctique y sont spécifiques, et environ 25% des communautés sont partagées entre ces deux bassins (principalement des Alphaproteobacteria, Gammaproteobacteria et Flavobacteria) et absentes ailleurs [116]. Dans le milieu tempéré, la divergence entre communautés est moins élevée qu'en surface et à la DCM [116]. Le milieu mesopélagique contient des zones où la concentration en dioxygène est particulièrement faible, appelées Zone de Minima d'Oxygène (OMZ) ; ces zones sont caractérisées par des communautés bactériennes différentes que celles vivant à la même profondeur mais avec plus de dioxygène [123]. Enfin, les communautés bathypélagiques en dehors des pôles sont caractérisées par une proportion élevée d'OTUs cosmopolites (jusqu'à 40%) [102].

Les Archées, et plus particulièrement les Euryarchaetoa et Thaumarchaeota, représentent en général une part plus importante des communautés dans le milieu mésopélagique [30].

Deux modules de Virus sont spécifiques des milieux profonds, un bathypélagique et un mésopélagique [97]. Les communautés mésopélagiques des espèces de NCLDV sont distinctes de celles à d'autres profondeurs dans la colonne d'eau [125].

• variabilité des communautés planctoniques avec la saisonnalité

À une localisation donnée, les communautés de surface changent plus avec la saisonnalité qu'avec la profondeur dans la colonne d'eau ; cet effet s'estompe néanmoins avec la profondeur, et ce dès la DCM chez les Procaryotes [64], [87], [116], [126], [127]. En surface, l'impact de la saisonnalité est variable selon le domaine du vivant considéré et le biome, et est globalement plus important dans les milieux polaires et tempérés que tropicaux [74], [84], [87], [95]. Elle n'est généralement pas responsable d'un changement de l'identité de la communauté dominante dans un milieu, mais entraîne plutôt des variations des proportions relatives des différentes communautés entre elles. Par exemple, la communauté Procaryote près des côtes de la Californie est en moyenne dominée par un seul module de l'hiver à la fin du printemps, qui reste dominant de l'été jusqu'au début de l'hiver malgré l'augmentation de l'importance de deux autres modules pratiquement absent au début de l'année [87]. Dans la mer Méditerranée, la communauté est dominée toute l'année par des Alphaproteobactéries, Flavobactéries et Cyanobactéries, mais la part des *Roseobacter* ainsi que les Thaumarchaeota et Euryarchaeota y augmente en hiver [84]. Dans les tropiques, les *Prochlorococcus* sont dominantes toute l'année mais avec différentes populations en fonction des saisons [84]. Le printemps dans les milieux polaires et tempéré est généralement marqué par des blooms qui changent la composition des communautés (voir *II.2.c* pour l'explication générale du phénomène) [84]. Pendant ces évènements, au moins une trentaine d'espèce de tous les domaine du vivant se succèdent, chacune dominant la communauté pendant quelques jours avant d'être exclue et remplacée par une autre [61], [82]. Jusqu'à 60% de la communauté phytoplanctonique (Eucaryotes et Procaryotes) change tout au long de l'année dans les zones de bloom [95]. Chez les Eucaryotes, les Diatomées sont la lignée comprenant le plus d'espèces responsables de ce phénomènes, à tel point que certaines ne sont détectable dans les communauté qu'à cette période ; d'autres espèces des lignées phytoplanctoniques y participent également [82], [99]

Aux pôles, les variations s'observent entre l'hiver, caractérisé par la nuit polaire, et l'été, où l'ensoleillement peut durer des journées entières (Fig.11). Les saisons sont décalées entre hémisphères, l'été austral se déroulant de décembre à février et le boréal de juin à septembre. En hiver, Les communautés d'hiver rassemblent entre autres des Dinophycés et des *Pelagibacter* [89], [113], [128], [129]. La fin de la nuit polaire et la fonte des glaces d'hiver entraîne des blooms à la fin du printemps qui sont généralement caractérisées des communautés dont la part phytoplanctonique est dominée par des Diatomés à la fin du printemps et au début de l'été, puis des Chlorophytes ensuite [74], [128]. Les communautés de fin d'été sont composées d'une plus grande part d'hétérotrophes et de mixotrophes, qui sont en Arctique essentiellement des Dinoflagellés et des Ciliophores [128].

Dans les milieux tempérés, le printemps est marqué par des blooms, qui changent de façon importante la composition des communautés (voir *II.2.c* pour l'explication générale du phénomène ; Fig.11) [84]. En été, l'augmentation des températures entraîne une augmentation de la stratification, rendant de vastes régions océaniques oligotrophes, et donc plus proche des conditions tropicales. Les communautés sont alors principalement composées de Cyanobactéries, des picophototrophes Eucaryotes ainsi que des Bactéries des groupes SAR11 et SAR86 [84]. De nouveaux blooms ont lieu quand les températures de surface diminuent suffisamment pour que le mixage de la colonne d'eau reprenne, en général au début de l'automne dans les milieux tempérés. Ils rassemblent généralement des Diatomées et des Dinoflagellés. Les Actinobacter, certaines Flavobacteria spécifiques de l'automne ainsi que d'autres espèces de SAR11 et des Planctomycetes augmentent [84].

Des phénomènes météorologiques plus rares comme El Niño peuvent aussi changer les concentrations en nutriments et provoquer des blooms à une fréquence inférieure à celle des saisons. Lorsque que cet évènement est particulièrement fort, les renversements de dominance écologique

s'observent à l'échelle des communautés. Par exemple, à la San Pedro Time Series en 2015 (année du plus fort El Niño depuis 2005), le module Procaryote habituellement dominant est renversé par un autre à la fin de l'été. Ce renversement disparait dès l'hiver suivant et la communauté renversée devient à nouveau dominante [87].

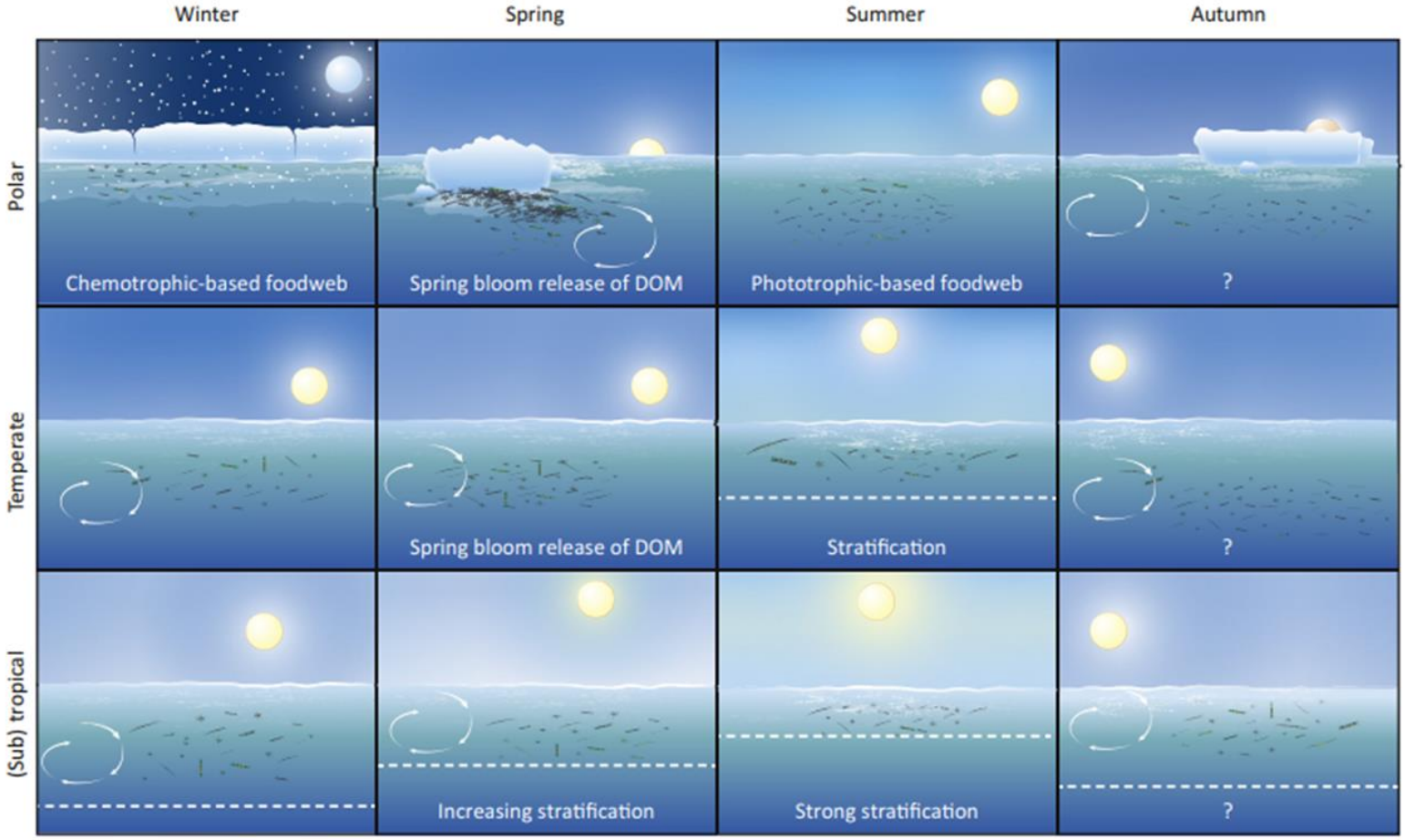


Figure 12. Impact de la saisonnalité sur le Bactérioplancton et le phytoplancton dans les trois biomes. Dans les régions polaires en hiver (en haut à gauche), l'obscurité et la glace limitent la production primaire de phytoplancton. Au printemps, le phytoplancton et les bactéries colonisent la glace, puis forment des blooms printaniers quand celle-ci fond et se brise. Dans les régions tempérées à la fin de l'automne et en hiver (au milieu à gauche), les remontées d'eau profondes combinées au effluents terrestre augmentent les concentrations en nutriments dans les eaux de surface, provoquant des blooms printaniers lorsque la lumière n'est plus limitante. La stratification augmente pendant l'été, entraînant des changements successifs de composition des communautés bactériennes. La diminution de la stratification au début de l'automne s'accompagne généralement de nouveaux blooms. Enfin, dans les régions tropicales (en bas), les masses d'eau au large sont mélangées par le vent, les tourbillons et les courants océaniques, tandis que les zones côtières sont souvent le lieu d'upwellings. En été, la stratification est particulièrement forte. Des blooms très épisodiques peuvent néanmoins avoir lieu en cas d'augmentations excepionnelles des concentrations en nutriments. En automne, les conditions estivales font place aux conditions hivernales; cependant, cette période de l'année est rarement étudiée et est donc représentée par « ? ». Extrait de Bunse et Pinhassi [84].

II.3. répartition biogéographique à une échelle taxonomique plus fine, celle des lignées : concept généraux et application aux lignées planctoniques

Après avoir détaillé les caractéristiques des communautés planctoniques de deux manières différentes dans les sections précédentes (quantitative et composition taxonomique et sa variabilité), cette section a pour but de décrire la répartition biogéographique du plancton non plus au niveau des communautés, qui sont des assemblages écologiques d'espèces dans un contexte environnemental donné, mais des lignées planctoniques à différents rangs (ordre, genre, OTU, espèce), qui ont donc une cohérence taxonomique mais pas forcément écologique. Décrire la distribution biogéographique d'une espèce revient en grande partie à décrire sa niche écologique ; la première sous-section de cette section sera donc consacrée à la définition de cette niche. Elle sera suivie par une sous-section décrivant certaines tendances générales des répartitions biogéographiques des lignées planctoniques, puis d'une sous-section finale qui s'attachera à décrire plus systématiquement la répartition biogégraphique des principales lignées planctoniques au sein des Procaryotes, Virus et Eucaryotes.

II.3.a. définition de la niche écologique

La distribution d'une espèce donnée s'explique en grande partie par sa niche écologique. Ce concept en englobe en fait deux : la niche écologique théorique et la niche écologique réalisée (Fig.13). La niche écologique théorique est souvent représentée comme un hypervolume formé par la combinaison de l'ensemble des gammes de paramètres environnementaux d'un milieu dans lequel un individu peut vivre. La niche écologique réalisée inclue dans cet hypervolume l'existence d'autres individus et donc de compétition pour les ressources. C'est cette dernière qui est à l'origine de la distribution biogéographique des espèces. C'est en grande partie la niche écologique qui définit la distribution biogéographique des espèces, indépendamment de leur histoire évolutive [130]. Il est connu que les microbiomes terrestres ont une distribution biogéographique structurée par la géographie, avec un effet net de la distance sur la similarité des communautés, ce qui a été démontré dans près de 68% des études traitant de ce sujet [126]. Pour les microbiomes du milieu marin, l'une des théories historiques principales sur la structuration est qu'elle n'est causée que par la sélection environnementale, indépendamment de la distance géographique (d'après Beijerinck puis Baas-Becking, comme présenté dans le préambule [15], [16]).

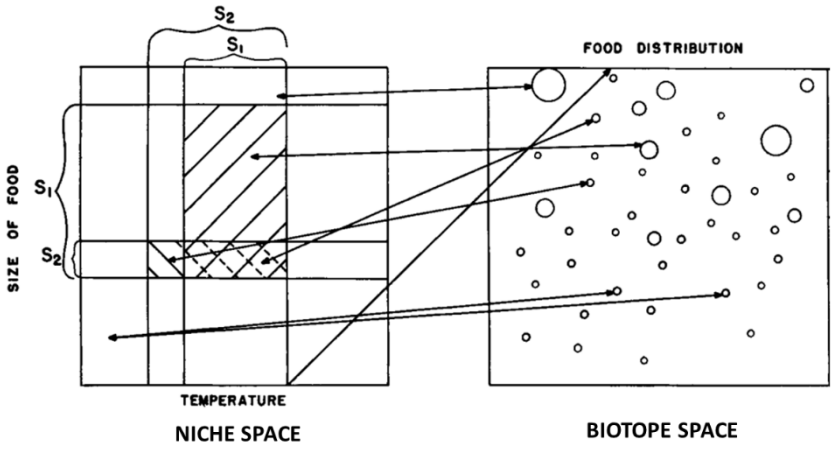


Figure 13. Niches écologiques théoriques et réalisées. À gauche**,** deux niches fondamentales définies par une paire de variable dans un espace de niche en deux dimensions. Les lignes indiquent les relations entre l'espace de niche à gauche, et l'espace du biotope, où la niche est réalisée. Figure issue de Hutchinson 1957 [130], Copyright © 2022 by Cold Spring Harbor Laboratory Press, reproduit avec la permission de Cold Spring Harbor Laboratory Press.

II.3.b. classes d'abondances des espèces planctoniques et notion de distribution antitropicale

Une première propriété fondamentale des communautés planctoniques est qu'elles rassemblent des espèces d'abondances très différentes, avec un gradient allant de l'espèce la plus abondante à celle étant la plus rare. Ce gradient peut être discrétisé en différentes catégories d'abondances. Il en existe plusieurs, essentiellement définies pour les écosystèmes terrestres [131]. Elles peuvent cependant être dans une certaine mesure transposées aux espèces planctoniques. Par exemple, la catégorie « rare et endémique » est appropriée pour décrire la distribution de certaines espèces de Bactérioplancton ou d'Eucaryotes qui sont des spécialistes côtiers capables d'exploiter efficacement les effluents terrestres [132]. Ce type de distribution est aussi caractéristique d'une proportion relativement élevé d'espèces dans les milieux profonds, par exemple à la DCM dans les milieux polaires où près de 30% des espèces sont endémiques mais globalement rares puisqu'elles n'y représentent au maximum que 2% des abondances relatives [64], [91].

La répartition de certaines lignées planctoniques est dite « bipolaires » ou « antitropicales », c'est-à-dire des distributions globalement symétriques par rapport à l'équateur [133], [134]. La zone entre 40° et 60° de latitude nord et sud représente l'une des barrières à la dispersion la plus importante à l'échelle globale et impacte fortement la distribution biogéographique des espèces. Elle est caractérisée par une variabilité saisonnière élevée et des transitions brutales avec les autres bassins océaniques (transition Atlantique Nord – Arctique par exemple). Les espèces planctoniques y subissent des remplacements et des variations significatives d'abondance [63], [90], [91], notamment pour les espèces abondantes et cosmopolites qui ont tendance à y devenir dominantes [132]. Dans milieux tempérés et polaires, où la variabilité environnementale est forte, seules quelques espèces dominent la communauté, la majorité étant rare. Certaines exceptions existent cependant. Par exemple, la taille de l'aire de répartition moyenne de certaines Bactéries et Archées a tendance à diminuer avec la latitude dans l'océan Pacifique Sud durant l'hiver polaire [96]. Malgré les difficultés et les potentiels biais que représentent leur identification, certaines espèces planctoniques ou genres planctoniques pourraient être véritablement cosmopolites c'est-à-dire présents avec une certaine abondance à la fois dans des environnements tempérés et tropicaux mais également polaires. Cela montre qu'une partie du plancton est capable non seulement de survivre à la traversée du front polaire mais aussi de continuer à y être actif voir compétitif. D'autres en revanche y entrent dans un état de pause physiologique en attendant le retour de condition plus favorables. Il est aussi probable que les populations de certaines espèces aient commencé à coloniser ces milieux seulement récemment à l'échelle des temps paléoclimatiques, et n'y sont pas encore complètement établies [91].

II.3.c. distribution biogéographique des principales lignées planctoniques

La distribution des lignées planctonique est souvent décrite en surface. C'est là que les observations sont les plus faciles mais c'est aussi la zone a la variabilité environnementale la plus élevée, puisqu'elle est beaucoup plus en interaction avec le milieu aérien, caractérisé par son instabilité, que les milieux profonds.

- **lignées Procaryotes**

Il n'existe pas de tendance générale concernant la distribution des différentes lignées du plancton Procaryote, mais une relation négative entre latitude et abondance est observée chez de nombreuses espèces rares [132]. Une autre particularité est que près de 30% des espèces présentes dans l'hémisphère sud le sont aussi dans l'hémisphère nord [132].

Grâce à l'advection par les courants marins, certaines lignées de Bactéries sont cosmopolites. C'est par exemple le cas des Bacteroidetes (particulièrement les Flavobacteriia), dont abondance varie avec des maxima dans l'Océan Austral (jusqu'à 25% contre 10% en moyenne ailleurs) [30], [100], [116], [117], [118]. D'autre ont une répartition structurées biogéographiquement [132]. Certaines espèces sont même endémiques, notamment au sein des SAR116, 92, 86 et 11 clades Ia, Ib et II, des Actinobactéries et des Verrucomicrobia [96], [120]. Les espèces endémiques sont plus fréquentes dans le milieu tropical, où certaines y sont même localement abondantes, c'est-à-dire présentes de façon ponctuelles avec des abondances relativement élevées [96], [120]. Dans ce biome, des espèces de Cyanobactéries des genre *Prochlorococcus* et dans une moindre mesure *Synechococcus* sont très abondantes, certaines étant dominantes dans les régions oligotrophiques et pas ailleurs [55], [63], [84], [89], [96], [135]. Dans le milieu tempéré, une part importante d'espèces a des aires de répartitions de l'échelle du bassin malgré des fluctuations saisonnières, résultant en un plus petit nombre d'espèces endémiques. Les espèces abondantes de ce biome (SAR86,11,116,406, Bacteroidetes, Verrucomicrobia, *Candidatus Actinomarina*) ont généralement des abondances plus faibles dans les autres [63], [115], [120]. Les Cyanobactéries y sont également abondantes mais moins que sous les tropiques, surtout le genre *Prochlorococcus* avec des renversements occasionnels de dominance par des *Synechoccocus* [55], [84], [96], [115]. Enfin, il n'y a pas d'espèces endémiques dans le biome polaire [120]. Certaines espèces de SAR11 des clades Ia et II, de SAR86, de Rickettsiales, de Rhodobacterales (notamment des Roseobacteraceae mais aussi des Paracoccaceae) et de Planktomarina sont particulièrement abondantes dans l'Océan Austral [96], [100], [117], [119]. Certaines Pelagibacterales sont cosmopolites à l'échelle de l'Atlantique [63]. Dans le milieu bathypélagique, les Alphaproteobacteria représentent en moyenne environ 20% des abondances et les Flavobacteria sont généralement moins abondantes qu'en surface, contrairement aux Deltaproteobacteria et Actinobacteria qui y sont plus abondantes qu'en surface, et ce sous toutes les latitudes [30], [115], [116], [123].

La distribution des Archées est essentiellement caractérisée par une augmentation des abondances relatives plus la profondeur augmente, pouvant passer pour les Thaumarchaeota en particulier de moins de 1% en surface à plus de 10% des barcodes en mésopélagique, dans pratiquement tous les océans [30]. Les Crenarchaeota sont plus abondantes dans l'Océan Austral [117].

- **lignées de Virus**

La distribution de beaucoup d'espèces du Virioplancton est cosmopolite. C'est le cas pour environ 93% des phages qui sont présents dans toutes les régions et à différentes profondeurs dans la colonne d'eau [123]. Certaines espèces d'autres groupes ont cependant des distributions plus réduites. Les coccolithovirus ont par exemple été identifiés dans l'Océan Austral mais pas dans l'Atlantique Sud ni l'Océan Indien [100]. Les communautés de DNA virus ont tendance à toutes regrouper la majorité de la diversité existante mais avec des différences d'abondances relative entre espèces [33]. La variabilité observée est principalement structurée par la nature des communautés bactériennes qui constituent la part la plus importante de leurs hôtes, résultant en une structuration par bassins et par biomes, avec une communauté dominante en Antarctique, une en Arctique et une partagée entre les biomes tempéré et tropical [97]. Chez les NCLDV, environ 29% des espèces sont cosmopolites et 70% des espèces sont trouvées à toutes les profondeurs dans la colonne d'eau, avec des abondances plus élevées en profondeur [34], [123]. Certaines espèces sont spécifiques du milieu mésopélagique [123].

- **lignées du phytoplancton Eucaryote**

Les lignées phytoplanctoniques présentent en général une distribution plutôt antitropicale, étant plus abondantes aux pôles que sous les tropiques [91]. Certaines espèces parmi les plus abondantes et ayant les aires de distribution les plus larges sont même indicatrices du milieu Arctique [91].

Les Dinoflagellés photosynthétiques sont le groupe le plus représenté dans les séquences de barcodes du nanophytoplancton Eucaryote dans les biomes tempéré et tropical, et y occupent également une part importante du microphytoplancton [96], [99], [136]. Elles peuvent en effet représenter jusqu'à 10% des lectures Eucaryotes et 20% des lectures de photoautotrophes Eucaryote en milieu tempéré, tropical et subtropical, et déclinent vers les pôles [55], [74], [99], [137]. Dans le biome polaire, leur abondance relative est plus faible que celle d'autres lignées du phytoplancton Eucaryote comme les Diatomées, les Chlorophytes et les Haptophytes et ne dépasse pas 10% des abondances relatives de photoautotrophes [74]. Certaines espèces comme *Ceratium hirundinella* sont cosmopolites ou quasi cosmopolites avec des préférences pour l'Arctique et peuvent former des blooms [55], [91], [138]. La région Indo-Ouest du Pacifique abrite des espèces endémiques (*Dinophysos miles*, *Ceratium dens*) [138].

Les Diatomées sont en moyenne la deuxième lignée la plus représentée derrière les Dinoflagellés au sein des communautés phototrophes (en moyenne 2.86% des lectures des barcodes Eucaryotes)[99]. Leur abondance relative varie fortement et est structurée par les biomes. Elles représentent une grande fraction de la communauté phototrophe dans les zones eutrophes comme les upwellings où les régions côtières et sont le groupe phototrophe le plus abondant dans les pôles [55], [84]. Elles sont bien connues pour former des blooms massifs au cours desquels leurs abondances relatives peuvent atteindre 40 à 80% des lectures Eucaryotes en barcode et métagénomique ainsi qu'en proportion de chlorophylle totale, et qui dépendent fortement des conditions du milieu. Leur abondance varie donc fortement en fonction des conditions environnementales [55], [70], [74], [98], [99], [100], [113], [139]. Du point de vue de la composition en Diatomées, la mer Méditerranée est la région la plus divergente, certaines espèces y étant particulièrement abondantes ; la composition intra région (à l'échelle de l'Océan Austral, du Pacifique Sud et Indien) est homogène [99], [139]. En outre, leur diversité est plus élevée entre 40 et -40°, bien que leur abondance globale y soit minimale [55], [139].

Les espèces de Pelagophyceae ont tendance à être indicatrices de milieux non polaires, avec des abondances relatives qui peuvent atteindre plus de 5% hors de ces latitudes et des zones oligotrophes [74], [139]. *Pelagomonas calceolata* est la Pelagophyceae la plus abondante en dehors des pôles et peut atteindre jusqu'à 4% des lectures Eucaryotes dans certaines communautés [140]. Une espèce de *Pelagomonas* ainsi que deux espèce du genre *Florenciella* (Dictyophyceae) sont signature du milieu polaire [91], [111]. Trois *Aureococcus* sont marqueurs de milieux tempérés, et cinq Bicosoecidae de milieux tropicaux [111]. Les différentes espèces de Chrysophyceae présentent des préférences biogéographiques distinctes, certaines étant cosmopolites, d'autre ayant une distribution bipolaire ou restreinte aux tropiques [138].

Les Chlorophytes sont très rares dans un grand nombre de milieux subtropicaux oligotrophes [70], [74]. Les espèces qui y sont présentes sont majoritairement spécialistes: deux espèces (genres *Bathycoccus* et *Micromonas*) sont endémiques de milieux tempérés, sept espèces le sont en milieu tropical (cinq *Chloropicon*, un *Chloroparvula*, un *Pycnococcus*) [111]. Leur abondance relative peut néanmoins dépasser les 10% dans certaines stations spécifiques des tropiques [55]. De façon général, elles sont plus abondantes aux latitudes élevées (au-delà de 40° de latitude absolue), en particulier les

pôles, ainsi que dans les upwellings [55], où elles peuvent représenter plus de 20% du nombre de cellules phototrophes localement [74], [98], [113], en particulier les *Micromonas*, *Ostreococcus* et *Bathycoccus* [63], [84], [135]. Il n'y a pas de tendance biogéographique claire concernant leur diversité, qui est globalement faible comparé aux Haptophytes et dans une moindre mesure aux Bacillariophytes [139]. Au sein du genre *Bathycoccus*, deux espèces très proches (jusqu'à récemment considérées comme deux écotypes), *B. prasinos* et *B. calidus*, ont ensemble une distribution quasi cosmopolite, la première étant présente dans tous les milieux tropicaux et la deuxième dans les milieux tempérés et côtiers [141], [142], [143].

La distribution des Haptophytes et plus particulièrement des Prymnesiophytes suit globalement celle des autres lignées du phytoplancton Eucaryote, à l'exception des Bacillariophytes. Leur abondance est faible dans les zones tropicales et subtropicales [55], [70]. Elles peuvent cependant être localement plus abondantes que les Bacillariophytes et les Chlorophytes et y dominer la communauté de nanophytoplancton [96], [99] en y représentant entre 5 et 10% en lectures de barcodes ainsi que proportion de Chl *a* [55], [74]. Elles sont plus abondantes en se rapprochant des pôles, en particulier dans l'Océan Austral où elles peuvent atteindre une abondance relative d'environ 15% en lectures barcode et métagénomique et en proportion de Chl *a* [55], [74], [98], [100], [113]. Neuf espèces de *Phaeocystis* sont signatures de milieux polaires, contre seulement deux New *Chrysochomulinaceae* en milieu tempéré et deux autres espèces (une *Chrysochromulinaceae* et une Haptophyte non classifiée) en tropical [111]. Grâce au riche registre fossile dans certains groupes de Prymnesiophyceae, il est possible d'observer que le niveau de provincialisme de ce groupe a varié au cours des périodes géologiques. Les périodes glaciaires ont notamment permis à certaines espèces subpolaires (notamment *Coccolithus pelagicus*) d'accroître leur aire de répartition, contrairement aux espèces tropicales [138].

La distribution des Cryptophyceae est globalement peu structurée biogégographiquement et ne suit pas les biomes. Elles sont cependant très rares dans les milieux tropicaux et peuvent attendre près de 10% du nombre de cellules phototrophes dans certaines zones tempérées [74]. Deux espèces du genre *Geminigera* sont même des marqueurs tempérés [111].

- **lignées d'hétérotrophes Eucaryotes**

La distribution des hétérotrophes Eucaryotes est en générale moins marquée par la symétrie par rapport à l'équateur que dans le phytoplancton Eucaryote, bien que des exceptions existent.

Dans les Alvéolés, la majorité des MALV I et II ainsi que les Syndiniales sont abondants et souvent cosmopolites (comme les Dinoflagellés phototrophes). Certaines espèces de Ciliophores sont indicatrices de milieux polaires et peuvent y représentent jusqu'à 15% de la communauté [91]. Certaines espèces ont été recensées dans les récifs coralliens (*Maristentor dinoferus*) et sur les côtes Antarctiques (*Heterostentor coeruleus*) [138].

Les MAST-4 peuvent représenter près de 0.5% des abondances relatives totales dans certaines stations. Certaines lignées ont une distribution cosmopolites, d'autres sont présentes partout à l'exception de l'océan Austral ou restreintes à l'Atlantique [122]. Les populations de MAST en général diffèrent entre milieu pélagique et côtier [122]. Au moins deux espèces de MAST-4 sont endémiques des milieux polaires [111].

Au sein des Rhizaires, les Cryomonadida sont, comme leur nom l'indique, souvent des marqueurs de milieux polaires et plus particulièrement Arctique [91]. La majorité des espèces de Foraminifera ont une distribution restreinte [138]. Chez les Radiolaires, les Acantharia et Collodaria sont abondants dans les larges fractions de tailles [38].

Des espèces de Picomonadida, dont la classification phylogénétique est incertaine mais qui seraient proches des Diplonemida, peuvent représenter une part importante de la communauté dans certains milieux (parfois plus de 15% en Arctique) [91]. Les Discoba sont représentés entre autres par des espèces parasitaires comme les Kinetoplastida dont certaines espèces sont des indicateurs de milieux non polaires avec une abondance relative pouvant atteindre jusqu'à 10% notamment dans le Pacifique Sud) [91].

Les Copépodes représentent en général une part importante des communautés planctoniques eucaryotes, avoisinant les 15% d'abondance relative métagénomique dans les communautés Eucaryotes [64]. Ils sont particulièrement abondants dans l'Océan Austral, le front subtropical, les régions oligotrophes des tropiques [96] où ils peuvent atteindre jusqu'à 35% des abondances relatives métagénomiques des Eucaryotes [64]. Vingt et une espèces connues sont marqueurs de milieux tempérés et trente et une de milieux tropicaux. Aucune espèce connue n'est marqueur de milieu polaire [111]. Les espèces observées à hautes latitudes y sont donc rarement endémiques contrairement aux espèces des milieux tempérés ou tropicaux. Les Calanoida, Oithonidae et Poecilostomatoida dominent les communautés zooplanctoniques dans l'Atlantique Sud et l'Océan Indien. Les Poecilostomatoida sont également présents dans l'Océan Austral. Ils peuvent être cosmopolites ou endémiques de certaines régions [100]. Ainsi dans la famille Oithonidae, *Oithona similis* a une distribution amphitropicale alors que *Oithona nana* est plutôt restreint aux tropiques et à la Méditerranée [144].

La même tendance générale, mais plus étendue, est observée chez les parasites qui représentent jusqu'à 10% des lectures métagénomiques des communautés entre 40 et -40° de latitude, et moins de 5% des lectures métagénomiques aux pôles [64].

II.4. lien entre les différents niveaux organisationnels de description des communautés planctoniques

Dans cette section finale du sous-chapitre dédié à la répartition biogéographique du plancton à différents niveaux organisationnels, j'évoquerai certaines des connexions existant entre eux, à savoir, le lien entre la biomasse d'une communauté et sa diversité dans la première sous-section, puis le lien grâce à la règle de Rapoport entre variabilité géographique de la diversité des communauté et répartition géographique des lignées planctoniques dans la deuxième et dernière sous-section.

II.4.a. lien entre biomasse et diversité

Biomasse et diversité permettent de décrire l'état écologique d'une communauté et peuvent être utilisées pour mieux comprendre les phénomènes qui en sont à l'origine. Dans cette perspective, il est intéressant de tenter de comprendre la relation liant ces deux grandeurs ; cependant, l'établissement d'une telle relation est rendue complexe par la question plus globale de la prise en compte des espèces rares dans le microbiome planctonique. En effet, leur prise en compte peut grandement changer les estimations de diversité, sans avoir d'impact important sur la biomasse dans une communauté donnée, et ce particulièrement dans les écosystèmes où la biomasse est faible[145].En plus de la prise en compte des espèces rares, les mesures de macrodiversité sont également impactées *in situ* par les échelles temporelles et spatiales des échantillonnages, et peuvent varier selon l'heure de la journée ou la saison à une localisation donnée. Les modèles *in silico* de diversité sont aussi impactés, en grande partie en conséquence des simplifications qu'ils produisent Ces différents biais sont à l'origine d'une incertitude concernant la possibilité même d'établir un lien entre biomasse et macrodiversité totale

[93], [105]. En cohérence avec ce constat, certaines analyses concentrées sur le bactérioplancton et le phytoplancton ont montré que la diversité n'était pas corrélée avec la productivité qu'elle soit représentée par la concentration en chlorophylle ou la production primaire [93], [105]. Une relation existe cependant si seules les espèces les plus abondantes et communes sont utilisées. Dans ce cas, la diversité et la biomasse/productivité sont liées par une relation unimodale, avec un maximum atteint pour des valeurs faibles de biomasse (en moyenne à partir de $100 mg_C.m^{-3}$) ou de productivité ($20 nmol.L^{-1}.h^{-1}$) dans tous les domaines du vivant et tous les bassins [58], [61]. Pour les Eucaryotes, la corrélation entre ces deux grandeurs est particulièrement forte pour les nanophytoplancton, et elle atteint son maximum aux plus hautes valeurs de productivité [61], [96]. D'autre part, certains modèles identifient des corrélations fortes entre diversité et biomasse/productivité : la diversité est plus élevée là où la productivité est plus élevée. D'un point de vu chronologique, la diversité atteindrait son maximum après celui de la diversité, particulièrement sous les hautes latitudes [61]. Ces résultats tendent à montrer que c'est la pression de sélection qui agit principalement sur la biomasse et la diversité des espèces les plus abondantes. et elle s'explique par la régulation par pression de sélection (biotique et abiotique) [61].

II.4.b. lien entre diversité des communautés et répartition des espèces : la règle de Rapoport

La diversité des communautés peut en partie s'expliquer par certaiens propriétés communes des répartitions biogéographiques des d'espèces planctoniques. Tout d'abord, le niveau de rareté des espèces est lié à la macrodiversité. En effet, la plus grande partie de la macrodiversité dans une communauté est généralement rare. Plus généralement, la modularité des communautés s'explique en partie par l'écologie des espèces, qui est l'un des principaux facteurs expliquant la variabilité de la diversité à l'échelle des communautés. Ainsi, les communautés composées d'espèces dont les niches écologiques se recouvrent en partie présentent généralement des gradients latitudinaux de diversité similaires. Ce n'est pas forcément le cas pour des communautés formées par des espèces apparentées [64]. Chez les Diatomées, la diversité en espèces des genres *Chaetoceros* et *Leptocylindrus*, appartenant tous les deux aux Chaetocerotophycideae, n'est pas du tout distribuée de la même manière dans les océans : le premier regroupe une diversité spécifique maximale aux pôles alors que le second y a une faible diversité, qui tend à être plus élevée dans la Méditerranée. *Chaetoceros* a cependant un gradient de diversité plutôt similaire à celui de *Fragilariopsis*, qui appartient aux Bacillariales [99].

Les liens entre modularité, diversité et écologie peuvent être partiellement expliqués dans le cas du plancton par la règle de Rapoport et l'hypothèse de tolérance physiologique, qui donnent une explication à la variabilité des niches écologiques des espèces planctoniques [146]. Les milieux très stables favorisent les espèces spécialistes, avec des niches écologiques étroites qui leur permettent d'éviter l'exclusion compétitive et d'exploiter le milieu au maximum. Les environnements à forte variabilité vont au contraire plutôt sélectionner des généralistes capables de dominer la compétition interspécifique dans un grand nombre de contextes environnementaux. Il en résulte que les espèces des pôles ont généralement de grandes aires de répartition, et la spécialisation locale y est moins forte. Le nombre d'espèce différentes s'y trouvant à un instant donné est donc moins élevé que dans des milieux de latitudes plus basses. La diversité pourrait même être prédite en estimant le nombre de niches différentes qu'un milieu peut supporter [147]. Cette théorie ne permet cependant pas d'expliquer toutes les observations de diversité [95]. La distribution d'un nombre important d'espèces planctonique est cohérente avec la règle de Rapoport, à savoir que l'endémicité aux pôles est globalement faible dans tous les groupes alors qu'elle est plus élevée dans les milieux tropicaux. Elle questionne également la pertinence de la théorie du « everything is everywhere but the environment

selects » [16], [18], puisque certaines espèces autant chez les Bactéries que chez les Eucaryotes et les Virus sont cosmopolites, mais un certain nombre ont une aire de répartition restreinte malgré la circulation océanique.

III. impact des différentes forces évolutives sur la répartition du plancton

Dans ce sous-chapitre, je détaillerai les différentes forces évolutives à l'origine des répartitions biogéographiques présentées dans le sous-chapitre précédent. La première section sera consacrée à la migration et à la dérive écologique et la seconde à la sélection.

III.1. migration et dérive écologique

La migration et la dérive écologique dans les communautés planctoniques résultent principalement de l'advection par les courants marins, qui est la propriété la plus distinctive de ces communautés microbiennes.

L'advection par les courants est l'un des élement fondamental expliquant la structuration des communautés planctoniques [121]. La dissimilarité entre communautés augmente de façon exponentielle avec la distance et le temps d'advection par les courants [121], [138], [148]. Mesurée avec la distance de Bray-Curtis, elle passe de 0.25 pour des communautés séparées de 1000km à plus de 0.4 pour des communautés séparées de 5000km [30]. Deux communautés séparées par un temps de transport de quelques mois partagent en moyenne 10% de similarité métagénomique, contre en moyenne 5% pour un temps de transport de plus d'un an et demi [121] (Figure 14). À noter que ces pourcentages sont certainement des sous-estimations du niveau de similarité entre communautés, en raison de l'effet de sous-échantillonnage inhérent au séquençage métagénomique. Celui-ci n'invalide cependant pas les comparaisons présentées ici [121]. En ne considérant que la structuration géographique causée par le transport par les courant, il est possible d'expliquer entre 25% et 50% de la composition des communautés selon les différents groupes d'Eucaryotes [98]. Les courants augmentent la connectivité au sein d'un bassin océanique et limitent les transferts d'un bassin vers un autre [98]. Ils permettent la coexistence de différentes communautés et des remplacements de communautés dominante entre différents biomes [62], [92], [98]. À une échelle plus fine, ils sont le vecteur principal de la dispersion des organismes actifs, dormant ainsi que des gamètes. En les advectant en grand nombre d'un milieu à un autre, ils augmentent la capacité de dispersion de chaque espèce et sa capacité à coloniser d'autres milieux [117]. Les effets de l'advection ne sont pas les mêmes selon la fraction de taille considérée [100]. Pour un temps d'advection d'une année et demie, la similarité métagénomique entre deux communautés des fractions de tailles inférieure à 5µm est divisée par cinq contre deux pour les fractions de tailles supérieures à 20µm [121]. Une partie des espèces de cette fraction de taille, et en particulier les métazoaires qui sont les plus gros, peuvent en effet être transportés à l'échelle de bassins entiers [98], [121]. Leur capacité de dispersion est limitée par les continents et les grandes circulations océaniques [38]. Il en résulte, comme évoqué précédemment, une structuration par biomes et les bassins [38], [98], [111] Les espèces planctoniques de grande taille subissent de manière plus importante les effets du mixage turbulent [98], [112], [149]. Celui-ci génère une structuration régionale importante et maintient la diversité, tout en conservant une divergence entre sous-communauté plus faible que les niveaux observés pour deux communautés provenant de bassins différents [62]. La structuration des communautés de phytoplancton et leur diversité est surtout impacté par les effets « mesoscale » dont l'échelle spatiale est inférieure à la centaine de kilomètres et l'échelle temporelle est de l'ordre du mois [62], [105]. Par exemple, dans les régions hydrographiquement dynamiques, la diversité des diatomées est principalement contrôlée par

le mixage latéral [103]. Dans la fraction de taille 0.8-5µm, la similarité métagénomique entre deux environnements séparés par quelques mois d'advection est d'environ 9±1% ; elle diminue de façon exponentielle jusqu'à 3±1% pour des environnements séparés de 1.5 ans d'advection [121]. Le mixage turbulent n'a pas le même effet au sein du phytoplancton en fonction de la fraction de taille des cellules : il n'affecte globalement pas les plus petites mais tend à disperser les plus grosses [150]. La structuration observée à l'échelle globale résulte beaucoup de la grande capacité de diffusion de cette fraction de taille, et encore plus chez les Procaryotes [151]. Cette capacité de diffusion est la conséquence à la fois de la dispersion des gamètes mais aussi de l'existence de formes dormantes [105], [107].

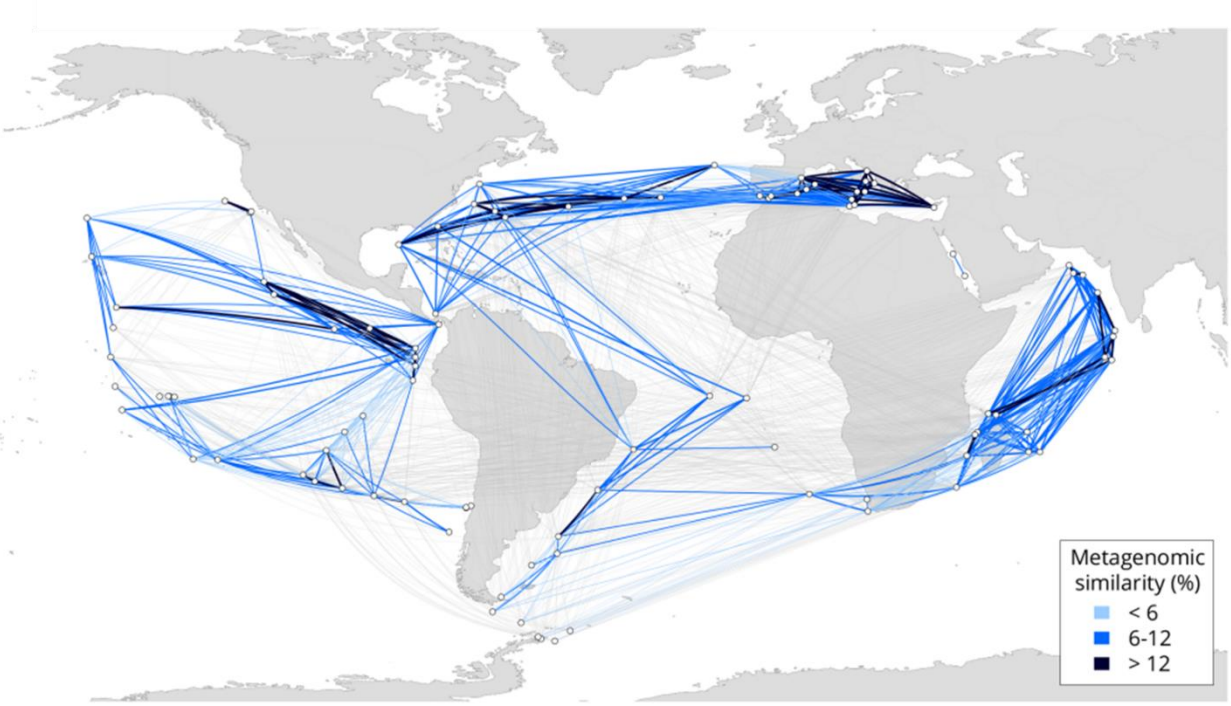


Figure 14. Dissimilarité métagénomique entre les stations Tara Oceans. Paires de stations connectées par un temps de transport inférieur à 1.5 années en bleu et supérieur à 1.5 années en gris. L'intensité du bleu reflète la similarité métagénomique pour la fraction de taille 0.8-5µm. Extrait de Richter *et al.* [121].

La dérive écologique désigne l'ensemble des évènements aléatoires comme les naissances et morts, qui dépendent en partie d'un grand nombre de facteur écologiques (prédation, parasitisme,…) et qui peuvent impacter les abondances relatives des espèces [152]. Son importance par rapport à la sélection est encore mal comprise et diffère selon les fractions de taille considérées. Elle a plus d'impact chez les Procaryotes que les Picoeucaryotes par exemple, en particulier en surface dans les tropiques et subtropiques [105], [153]. Il est difficile d'évaluer l'impact conjugué de l'advection par les courants et de la dérive écologique sur la structuration des communautés, en partie parce que ces phénomènes agissent à différentes échelles spatiales (du mètre au bassin) et temporelle (jusqu'à plusieurs dizaines d'années) qui sont à relier avec les échelles de temps évolutifs des espèces planctoniques. Leurs effets couplés sont néanmoins suffisant pour créer une structuration et de la diversité dans les communautés planctoniques à l'échelle des océans [154], [155]. Le milieu côtier constitue une exception dans la mesure où les courants y sont faibles ; les communautés y sont surtout structurées par la forte variabilité environnementale [148], [156].

III.2. sélection

La sélection joue un grand rôle dans la structuration des communautés planctoniques [98]. Les pressions de sélections peuvent être abiotiques ou biotiques (infections virales et bactériennes, prédation, exclusion compétitive), et la sélection en résultant peut être hétérogène ou homogène. Son importance, de la même façon que pour l'advection, n'est pas la même selon la fraction de taille considérée. Cependant, de façon générale, la sélection est plutôt hétérogénéisante dans les milieux où la variabilité saisonnière est importante, comme les pôles [92]. Elle participe, avec le flux important de diversité en provenance d'autres bassins, à y maintenir de la diversité à long terme [91].

Les courants marins et la sélection ne sont pas découplés, puisque les courants sont responsables du d'une homogénéisation des conditions environnementales sur de grandes échelles spatiales. Ainsi une pression de sélection relativement similaire s'exerce tout le long d'un courant donné, et les mêmes assemblages sélectionnés dans ce contexte peuvent rester dominant sur de très grandes échelles spatiales.

Pour les grandes fractions de tailles (jusqu'au microplancton), la sélection a tendance à être plutôt homogénéisante [87], [151]. Cette différence avec les plus petites fractions de taille est en partie expliquée par l'hypothèse taille-plasticité, qui postule que plus les organismes sont gros et plus leur plasticité métabolique est faible [151], [157], [158]. Les cycles de vie des Arthropodes planctoniques sont aussi généralement plus complexes et plus longs que ceux des plus petites espèces unicellulaires [38]. Leur reproduction est probablement obligatoirement sexuée, alors que les unicellulaires des petites fractions de tailles peuvent également recourir à la reproduction asexuée. Leur niche écologique est donc globalement plus étroite, ce qui n'exclue pas la possibilité d'une très vaste répartition biogéographique puisque les conditions environnementales peuvent être stables à de grandes échelles dans le milieu marin. À grande échelle, la structuration de leurs communautés est donc principalement contrôlée par la sélection abiotique, particulièrement la température chez les Copépodes et les Choanoflagellés [98]. Ce n'est cependant pas le cas à l'échelle locale ou c'est le mixage turbulent qui est responsable de la structuration, comme évoqué précédemment. Chez les phototrophes, les paramètres environnementaux ont un impact plus important sur la structuration que chez les métazoaires [98]. En outre, la plasticité métabolique de ce groupe leur permet en général de ne pas subir de pression de sélection abiotique trop forte lors de leur advection [151]. Le fait qu'ils puissent changer de mode de reproduction en fonction des conditions du milieu, et particulièrement recourir à de la reproduction asexuée sur de longues périodes, joue également un rôle clé dans leur capacité importante de dispersion [66], [69]. Les limites à la dispersion ont une importance environ cinq fois plus forte que la sélection dans la structuration des communautés de picoeucaryotes [153].

La sélection biotique est un facteur important dans la diversification des Eucaryotes, la compétition étant le type d'interaction le moins important dans ce phénomène [38], bien qu'elle contribue à la baisse de diversité du phytoplancton en milieu tempéré [95]. La prédation et le broutage ont un effet important sur la diversité et sa structuration (Figure 15 A) [61], [62], [76]. La pression de sélection biotique explique également en grande partie la faible abondance des espèces rares [32]. La mortalité élevée qui résulte de cette pression de sélection est compensée par l'advection par les courants qui est à l'origine d'apports permanents de nouveaux individus et d'un maintien de la diversité [159], [160]. Leur démographie est ainsi à l'équilibre [159].

Concernant la sélection abiotique, la température de surface (et plus particulièrement sa moyenne annuelle [93]) est l'un des paramètres physiques ayant le plus d'impact sur la distribution des espèces et la structuration des communautés dans tout l'arbre du vivant [30], [75], [90], [97], [113], [120], [122], [124], [148], [153]. Elle a un effet direct sur le métabolisme des organismes planctoniques

[93] ainsi que sur les tailles des populations et leur taux d'extinction [64] puisqu'elle affecte la dispersion des organismes en particulier vers les hautes latitudes où elle peut représenter une barrière forte à la dispersion [75], [154]. Elle peut expliquer jusqu'à deux tiers des fluctuations d'abondances du phytoplancton [95]. L'effet de la température n'est pas le même selon la fraction de taille dans le phytoplancton, et explique une plus grande part de la variation de concentration de picoplancton que dans les plus grandes fractions de taille [75]. La Photosynthetic Active Radiation (PAR), la longueur d'onde de la lumière et son intensité sont également des facteurs pouvant expliquer en partie la variabilité des communautés [113]. Ils sont à la base de la production primaire par la photosynthèse, mais peuvent aussi avoir un effet délétère via le mécanisme de photoinhibition qui peut en diminuer le rendement en surface [68]. Les paramètres chimiques, notamment les flux en nutriments et, dans une moindre mesure, leurs concentrations, participent également à la structuration des communautés [76], [114], [137]. Ils ont à la fois des impacts sur la diversité et la productivité. Ils constituent le facteur limitant pour la croissance du phytoplancton et impactent donc directement sa biomasse, en particulier au-delà de 40°N et S [75], [161], [162]. Le nitrate est la source d'azote principale pour le phytoplancton [100]. Dans l'Atlantique tropical et l'Océan Indien, le phosphate est particulièrement limitant, alors que c'est l'azote qui l'est dans le Pacifique Sud et l'Atlantique Nord [162]. Dans ces milieu, l'azote est majoritairement rendu biodisponible par les diazotrophes [163]. Les augmentations locales de concentration sont à l'origine de blooms phytoplanctoniques, notamment le nitrate [100], le phosphate et le fer [63], [84]. De façon plus générale, la variabilité des communautés est surtout expliquée par la combinaison de paramètres physiques et chimiques [64], [103], [119]. Dans l'océan Austral par exemple, les communautés sont structurées par la température, la disponibilité en différents nutriments (phosphate, silicate, nitrate) ainsi que le dioxygène, la salinité et la pression [117]. La composition des communautés Arctique est drivée par la proximité de la côte, la température ainsi que les nutriments [91]. La composition des communautés à la surface de l'Atlantique est drivée par la température et la moyenne annuelle de la concentration en nitrate [63]. La diversité du phytoplancton Eucaryote est contrôlée principalement par le taux d'approvisionnement en nutriment limitant, la température et l'intensité du mixage entre masses d'eaux de températures différentes [76] (Figure 15 B).

Enfin, la distribution actuelle des communautés planctoniques résulte forcément de l'histoire évolutive de chaque espèce et du contexte géologique et climatique dans lesquelles elles se sont déroulées. Les communautés planctoniques de l'Arctique ont par exemple été impactées par les différents processus tectoniques et les successions de périodes glaciaires qui ont contribué à isoler ce basin de la plupart des autres, hormis l'Atlantique et, dans une moindre mesure, le Pacifique Nord [91]. Les registres fossiles montrent la diminution croissante de la diversité du zooplancton en direction des pôles ainsi que leur dépendance à la température. Ces caractéristiques qui sont restées stables à l'échelle des temps géologiques [64], [91]. Ils montrent également que la colonisation de ce bassin par certaines populations planctoniques est encore en cours [91]. De façon plus globale, l'âge des espèces contribue à expliquer leur aire de répartition : plus une espèce est ancienne plus elle a eu du temps pour se disperser et pour augmenter l'efficacité de sa dispersion, causant une diminution des taux de spéciation allopatrique et *in fine*, une diminution du taux d'extinction [138]. Des évènements extrêmes récents comme El Niño peuvent également impacter la structuration des communautés sur des échelles de temps allant de l'année à la dizaine d'année [87].

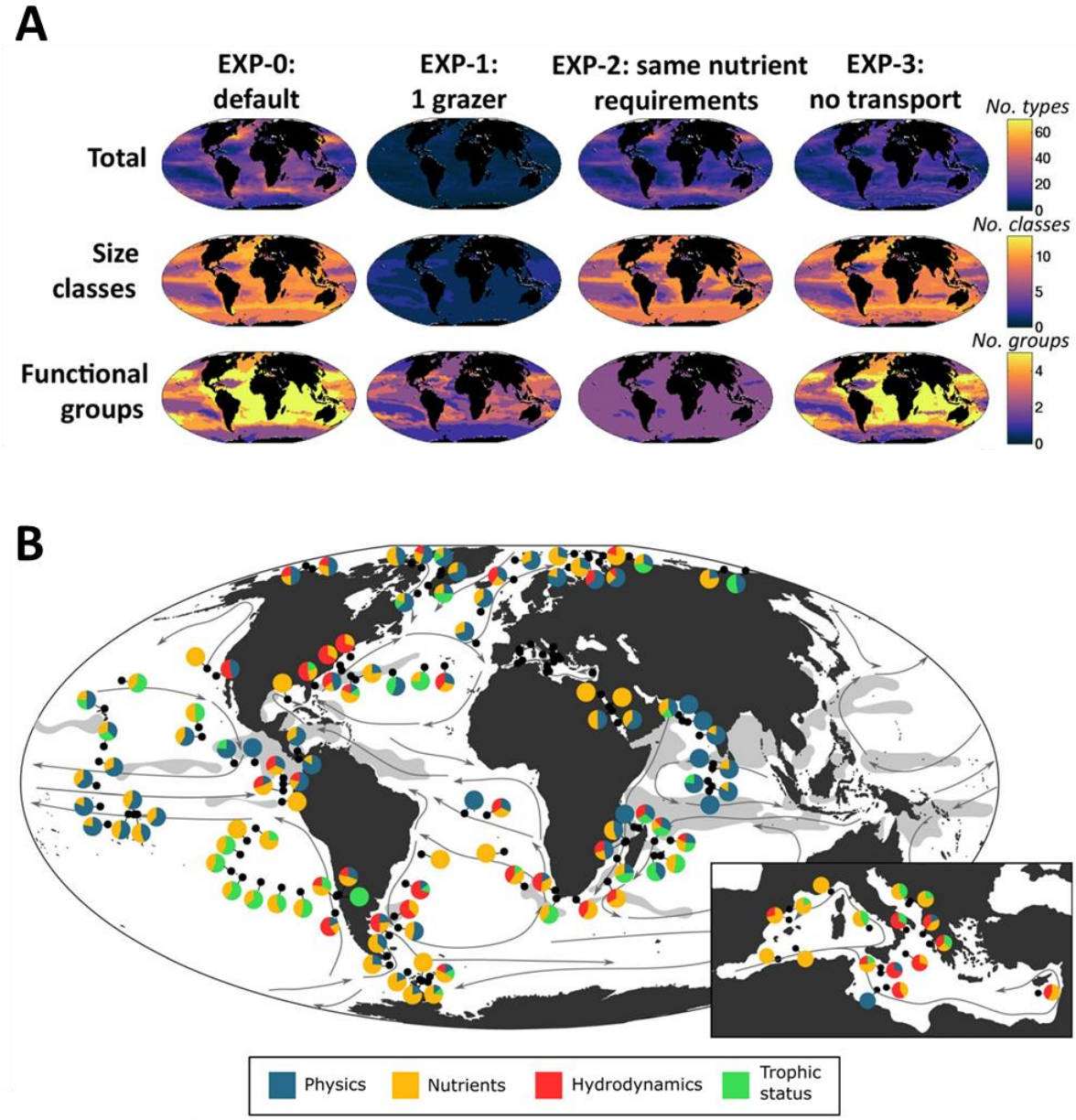


Figure 15. Exemples de résultats de modèles visant à évaluer l'importance de différents paramètres sur la diversité. (A) Simulations de sensibilité : richesse moyenne annuelle du modèle. EXP-1: un seul type d'hétérotrophe broutant les organismes de toutes les tailles. EXP-2: tous les groupes fonctionnels ont les mêmes besoins en nutriments. EXP-3: pas d'advection du plancton par les courants (mais advection des nutriments). Première ligne: nombre de types planctoniques; deuxième ligne: nombre de gammes de tailles différentes; troisième ligne: nombre de groupes fonctionnels; dernière ligne: nombre de niches thermiques différentes. Extrait de Dutkiewicz *et al*. [76]. **(B)** Contribution relative de différentes variables sur la prédiction de la richesse en Diatomées des stations *Tara* Oceans. Seules les contributions supérieures à 20% sont prises en compte et les résultats des trois modèles employés (arbre de régression boosté (BRT), réseau neuronal (NN) et random forest (RF)) sont agrégés. La carte utilise la projection de Mollweide. En arrière-plan, les principaux courants océaniques sont indiqués par des flèches et les zones de forte diffusivité latérale (selon Abernathey & Marshall [164]) sont représentées par des zones gris clair. Extrait de Busseni *et al.* [103].

L'estimation de l'importance de la sélection abiotique dans la distribution des espèces et la structuration des communautés se fait généralement via l'analyse de l'impact de chaque paramètre environnemental. Les province océanique et/ou latitude sont souvent considérés comme tels dans ces analyses, et présentent dans certains cas une bonne corrélation avec la distribution des espèces planctoniques [63], [120], [122]. Ils regroupent en fait toutes les conditions abiotiques et, indirectement, biotique locales (concentrations en nutriment, température, salinité, pression, luminosité, proportion d'hétérotrophes, de phototrophes, de parasites etc), et donnent d'une certaine façon une idée du niveau de sélection à un endroit donné. Ils n'aident pas à la compréhension fine des mécanismes responsables d'une observation à un instant précis, mais permettent de comprendre de manière globale l'échelle géographique pertinente de la structuration de l'espèce/la communauté/la fraction de taille étudiée. Le fait d'étudier chacune des variables citées précédemment individuellement pose des complications techniques, notamment parce que certaines sont fortement corrélées. La température est par exemple fortement corrélée à la concentration en dioxygène. Sa valeur à un instant précis en surface est en partie la conséquence de l'intensité lumineuse et du mixage [75]. La concentration en chlorophylle, qui reflète le niveau d'activité photosynthétique de la communauté est parfois utilisée comme un potentiel paramètre explicatif. C'est en fait un témoin des effets combinés de la disponibilité en nutriments ainsi que de l'intensité lumineuse (bien qu'elle ne soit pas systématiquement corrélée positivement à la concentration en chlorophylle), qui reflètent également à quel point la production primaire structure toute la communauté qui en dépend. Les changements de températures modifient les communautés de phytoplancton de surface, ce qui modifie ensuite l'efficacité de recyclage des nutriments disponibles, changeant leurs concentrations [75]. Une analyse comparative de différentes communautés pourrait laisser croire que température et nutriments sont responsables de l'originalité d'une communauté par rapport à une autre, alors que l'effet ne provient en réalité que de la température. Les analyses visant à expliquer les proportions de variances observées à différentes échelles peuvent ainsi être biaisées par les corrélations entre variables explicatives utilisées, et il est dans certains cas difficile d'établir des liens de causalité entre paramètres et observations biologiques. De plus, les modèles ne parviennent néanmoins jamais à expliquer 100% de la composition des communautés observées. Cela est impossible, mais un certain nombre de facteurs potentiels n'ont jamais été intégrés, comme les différences de capacités entre les espèces à stocker des nutriments ou changer les quotas cellulaires, la variabilité du spectre d'absorption entre les différentes espèces de phytoplancton, la morphologie, la motilité, la propension à former des colonies ou entretenir des symbioses ainsi que la régulation des capacités de flottaison par exemple [76].

IV. modèles de distribution des abondances des espèces planctoniques

Dans ce sous-chapitre final, je présenterai différents modèles utilisés pour décrire la distribution des abondances des espèces au sein des communautés planctoniques, et la façon dont ils peuvent être utilisés pour étudier les dynamiques écologiques des espèces au sein de ces communautés.

Les caractéristiques d'abondance, de diversité et de productivité du plancton peuvent être analysées au moyen de modèles. L'étude de la distribution des abondances (SAD) ou du rang (RAD) des espèces est une façon de visualiser les dynamiques écologiques dans une communauté (Figure 16).

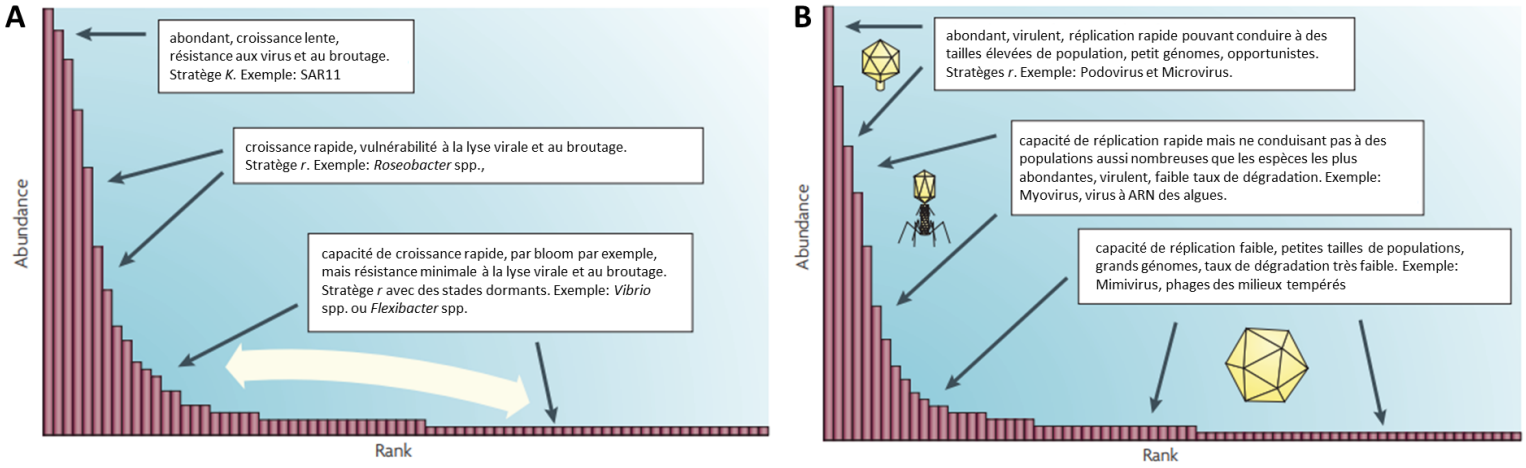


Figure 16. Écologie des espèces planctoniques en fonction de leur positionnement sur les distributions d'abondances. (A) RAD du plancton Procaryote et Eucaryote. Les espèces les plus abondants dans l'océan, tels que SAR11, sont probablement des organismes *K*-sélectionnés (taux de croissance maximum lent, résistance à la lyse virale et au broutage). En revanche, les espèces moins abondantes, tels que *Roseobacter* spp. et *Vibrio* spp. sont davantage *r*-sélectionnées (croissance rapide, sensibilité élevée aux infections virales et au broutage). La double flèche blanche représente les espèces généralement présentes en faible abondance dans un milieu donné et qui sont la majorité du temps exclues compétitivement par les stratèges *K*, mais qui peuvent périodiquement avoir une croissance rapide quand les conditions leurs sont propices. Elles sont cependant vulnérables aux Virus, qui les font rapidement retourner à une abondance faible. **(B)** RAD du Virioplancton. Contrairement aux autres domaines du vivant, les Virus les plus abondants sont *r*-sélectionnés. Ils sont virulents, ont des génomes de petite taille et une durée de vie courte. La structure de leurs populations est probablement inégale, un grand nombre de virus à un moment donné étant issus d'un nombre limité d'événements lytiques. Les virus les plus rares et les plus *K*-sélectionnés ont des génomes plus grands, se décomposent lentement et peuvent former des associations stables avec leurs hôtes. Ce groupe comprend certains virus à ARN et à ADN ayant une longue durée de vie et une faible virulence, comme les *Nucleocytoviricota*. Adapté de Suttle [32].

Les SAD et RAD sont souvent modélisées par des lois de la famille des lois puissance [159], [160], [165]. Ces lois ont la propriété d'être à « queue longue », c'est-à-dire qu'elles sont fortement déviées par rapport à une loi normale. Il avait été indiqué dans II.2.b que les communautés planctoniques rassemblent des espèces appartenant à différentes classes d'abondances, certaines étant rattachées à des distributions biogéographiques particulières. La modélisation des distributions d'abondance

permet d'estimer la part d'espèces rares par rapport aux espèces abondantes dans une communauté. Le fait que ces modèles suivent une loi puissance indique, d'un point de vue écologique, que la grande majorité des espèces dans une communauté est rare, et que seules quelques espèces sont dominantes. Jusqu'à 93% des espèces dans un microbiome planctonique sont rares [160]. Leur diversité est élevée en comparaison de celle des espèces dominantes, et maintenue à l'échelle des océans [159], [160]. Les espèces rares et actives participent au fonctionnement de la communauté, contrairement aux espèces abondantes ou rares mais inactives (gamètes, formes dormantes enkystées,…), qui jouent plutôt un rôle de « banques de semences », attendant des conditions favorables pour changer d'état physiologique et participer activement à la dynamique de la communauté [107], [108]. La comparaison des modèles entre communautés permet aussi d'évaluer la variabilité des espèces dominantes entre elles. Étant en général les espèces avec la fitness la plus élevée dans un contexte biotique et abiotique donné, la variabilité de leurs abondances renseigne sur les différences de dynamiques écologiques entre communautés. Chez les Eucaryotes, une même espèce est généralement dominante seulement dans une fraction des communautés, et cette fraction varie en fonction de la taille de ses individus [159]. Cela est en partie la conséquence de la saisonnalité des dynamiques populationnelles du plancton, dans lesquelles des blooms sont responsables de renversements de dominances en un temps très court. La forme générale des SAD est préservée dans ce cas mais ses paramètres et la position des espèces les unes par rapport aux autres ont tendances à varier. Cela s'observe particulièrement chez les Virus (Figure 16) [97].

En théorie, l'émergence du modèle de loi puissance des SAD/RAD découle d'une dynamique telle que décrite par la théorie neutre, c'est-à-dire un modèle dans lequel seuls des processus démographiques stochastiques (mortalité, natalité, immigration) entrent en jeu [106], [166]. La théorie neutre est particulièrement adaptée pour expliquer le fonctionnement de communautés dont la richesse spécifique est élevée et où les niches des espèces ont tendances à se recouvrir en partie, car elles permettent de se rapprocher de son cadre de validité qui repose sur l'hypothèse d'équivalence compétitive (qui n'a cependant jamais été démontrée expérimentalement) [159], [167]. Cette théorie ne peut cependant pas expliquer les déviations observées, entres autres à cause du fait qu'aucune communauté ne valide complètement cette hypothèse [165]. Les principales déviations à la loi puissance, et donc à la théorie neutre observée, sont causées par des facteurs biotiques ou abiotiques [106], [160]. À petite échelle, l'advection chaotique est responsable d'une déviation de la loi puissance ; la probabilité d'observer une grande valeur d'abondance est plus faible dans ce genre de système que dans un système sans courant. Cette déviation s'observe par exemple moins dans le milieu lacustre [160]. L'advection chaotique représente une barrière à la dispersion à cette échelle qui limite l'abondance des espèces dominantes [160]. Les autres paramètres biotiques et abiotiques ont généralement un effet sur la structuration biogéographique de la répartition des communautés, qui se reflète sur les SAD/RAD [165]. Aucune corrélation claire avec des variables physico-chimiques comme la température ou la concentration en nitrate n'a cependant été identifié à ce jour, mais de telles corrélations existent potentiellement avec d'autres grandeurs résumant la variabilité environnementale. De plus les mêmes paramètres n'ont pas forcément le même effet sur les distributions selon la taille des organismes ou le groupe taxonomique considéré. Le modèle neutre est par exemple plus pertinent pour les petits organismes que les grands [137]. Cela s'explique en partie par le fait que la position des espèces dans le réseau trophique change en fonction de la taille de leurs individus [54], [159], en accord avec la théorie métabolique de l'écologie qui lie métabolisme, taille des organismes et température et prédit une augmentation de taille moyenne des individus de quatre ordres de magnitudes entre niveaux trophiques [168]. La pertinence du modèle neutre pour les petits organismes s'explique aussi par des effets indirects liés à leur macrodiversité [54], [159]. Chez les Diatomées, les modèles de loi puissance des communautés présentent des déviations significatives

en lien avec la géographie (en particulier dans les régions polaires et les tropiques) (Figure 17) [165]. L'importance des déviations est corrélée négativement avec l'hétérogénéité de la communauté, c'est-à-dire l'écart en terme de taille de population entre les espèces les plus rares et les plus abondantes (plus la déviation à la théorie neutre est importante et plus l'hétérogénéité est faible) [165].

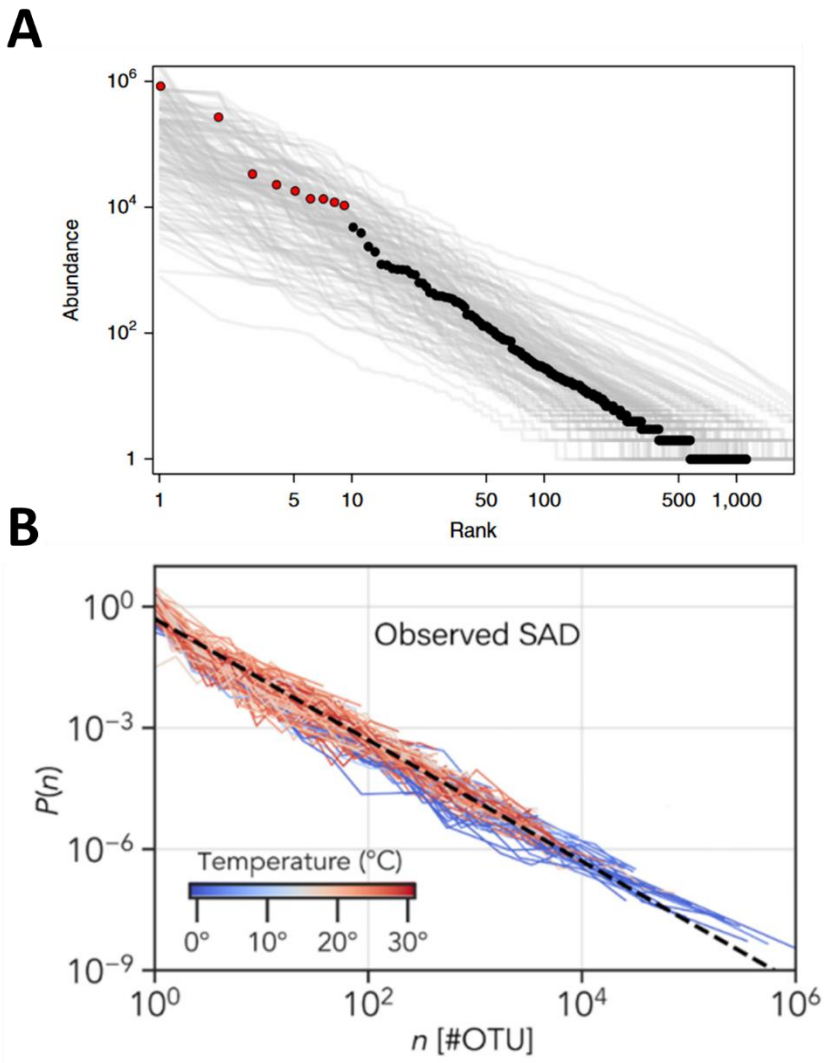


Figure 17. Exemples de distributions des abondances d'espèces planctoniques au sein d'une communauté. (A) RADs des espèces Eucaryotes dans les stations *Tara* Oceans (chaque ligne grise représente une station). Les cercles illustrent comment les espèces dominantes (en rouge) et non dominantes (en noir) sont partitionnées au sein d'une communauté. Extrait de Ser-Giacomi *et al.* [159]. **(B)** SADs basées sur les OTUs de Diatomées dans les stations *Tara* Oceans. Chaque ligne représente une station. La couleur correspond à la température moyenne en surface. Elles peuvent toutes être approximées par une loi puissance de paramètre λ=1.5, représentée par la ligne en tirets. Extrait de Pigani *et al.* [165].

CHAPITRE 2. DIVERSITÉ, USAGE ET ÉVOLUTION DES STRUCTURES DE DOMAINES PROTÉIQUES

Le premier chapitre de cette introduction visait à fournir une description du plancton aux plus hauts niveaux biologiques, à savoir ceux des communautés et des espèces. Cette thèse a pour but d'établir un lien entre ce niveau et un niveau d'un plus petit ordre de grandeur, celui des molécules, et plus particulièrement ici, celui des protéines qui participent activement à tous les aspects la vie cellulaire.

I. définition des folds et diversité

Dans ce premier sous-chapitre, je définirai les domaines protéiques et les folds dans une première section, puis présenterai certains de leurs systèmes de classification et les estimations de leur diversité dans la seconde. Je ferai un commentaire dans une troisième et dernière section sur certaines limites de ces définitions ainsi certaines propositions avancées par la littérature pour les contourner.

I.1. définition d'un domaine protéique et d'un fold

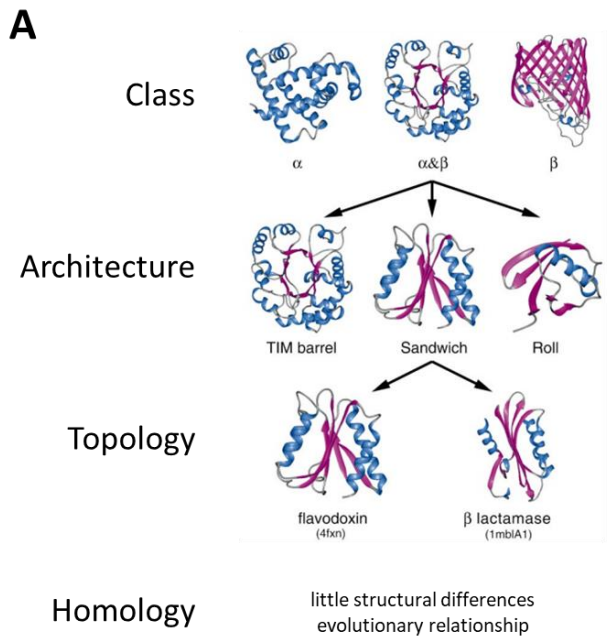
Un domaine protéique est une chaîne polypetidique suivant de façon autonome ou coopérative et assistée un chemin de repliement au terme duquel son squelette (succession des carbones α, sans considérér les chaînes latérales des acides aminés) adopte une structure tridimensionnelle, appelée **fold** [169], [170], [171], [172], [173]. Les domaines protéiques ont une longueur en général comprise entre 40 et 150 résidus [171]. Ils peuvent réaliser une ou plusieurs fonctions telles que la catalyse enzymatique, la régulation de l'expression génétique, la signalisation ou le contrôle du développement chez les organismes pluricellulaires [170], [171]. Les protéines peuvent être composées d'un ou plusieurs domaines, et sont qualifiées de « monodomaine » dans le premier cas et « multidomaine » dans le second. La majorité des protéines du vivant sont des protéines multidomaines (entre 2/3 et 4/5 de toutes les protéines) [174].

I.2. diversité et systèmes de classification

La diversité totale des domaines dont la structure est globulaire avec une proportion de régions désordonnées inférieure à un certain seuil (par exemple moins de 65% des résidus du repliement [175]) est probablement finie [176], [177], [178], mais en estimer le nombre total est délicat. Il pourrait y en avoir entre 1510 et 1970, voir entre de 3000 et 4000 [179], [180], [181]. Concrètement, il est difficile d'évaluer cette diversité de manière absolue car les folds sont des objets dynamiques dont la structure peut changer, et qui peuvent être définis de plusieurs manières. À noter que UniProtKB contient actuellement 227.10^6 séquences protéiques ; la base de donnée d'AlphaFold (AFDB) regroupe les structures prédites de 214 millions de séquences UniProt [182] et la Protein Data Bank (PDB) 215000 structures de protéines résolues expérimentalement [183]. Les 214 millions de structures protéiques de l'AFDB contiennent près de 365 millions de domaines protéiques [184].

Plusieurs méthodes ont été développées pour classifier la diversité structurale des folds. Leurs différences proviennent principalement de l'objectif dans lequel elles ont été créées et les définitions de domaine protéique et de repliement sur lesquelles elles s'appuient [172], [185]. CATH (Class Architecture Topologie Homologie ; Figure 18) [2], [3], SCOPe (Structural Classification Of Proteins (extended)) [186], [187], SCOP2 (Structural Classification Of Protein 2) [188] ont été développées pour recenser la diversité des folds qu'elles classifient selon leur topologie. CATH met l'accent plutôt sur la structure, alors que SCOPe/SCOP2 plutôt sur la fonction et l'évolution, résultant en deux bases de données globalement équivalentes mais utilisant un mode de classification différent. ECOD propose également une classification mais qui regroupe les folds en fonction de leurs relations évolutives et

leurs homologies plutôt que leurs topologies [189], [190], [191]. Enfin DALI a été développé pour permettre la comparaison rapide de structures entre elles [192].



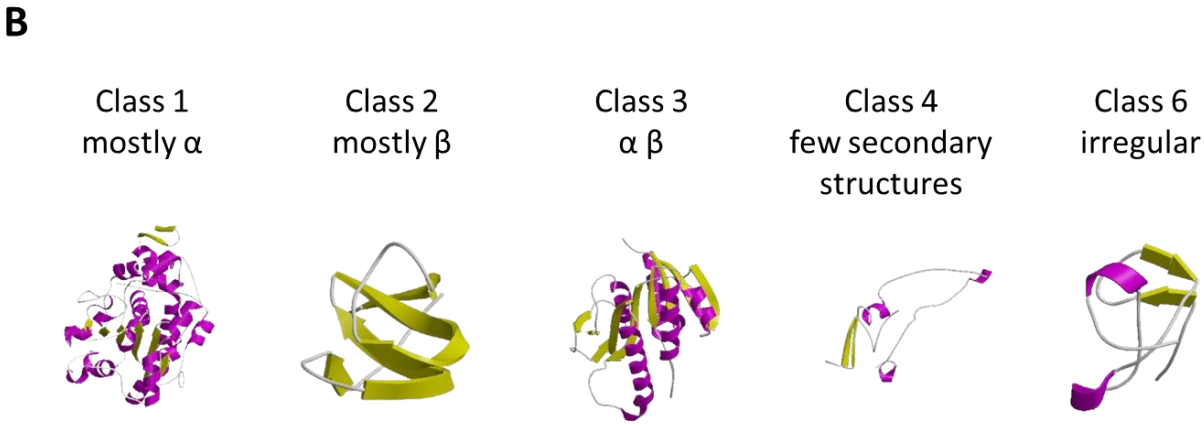


Figure 18. Classification CATH. (A) Les quatre niveaux hiérarchiques et le type de structures qui leurs sont associés. **(B)** Exemple de fold dans chacune des Classes. Extrait de Orengo *et al.* [3] et de https://www.cathdb.info/browse/tree.

Les domaines sont définis manuellement dans SCOP2 et SCOPe, et deux domaines sont considérés distincts s'ils n'ont pas d'homologie évidente de séquence. Ils sont définis au contraire dans CATH en prenant le consensus de trois méthodes automatisées, qui fonctionnent dans 53% des cas. Les 47% restant sont annotés manuellement. CATH propose une annotation structurale pour un nombre de domaines plus importants que les autres ressources (actuellement 1060659 par intégration des structures produites par AlphaFold2), entre autres grâce à son automatisation qui permet un traitement plus rapide des nouvelles séquences protéiques [193]. De plus, c'est l'annotation CATH qui a été sélectionnée pour classifier les domaines recensés dans l'Encyclopédie des Domaines (TED), qui est la ressource la plus vaste en termes de nombre de domaines à ce jour. Elle comprend 251 millions de domaines dont les folds ont pu être annotés avec CATH et au moins 33 millions de domaines non globulaires ou avec peu de structures secondaires [184]. Dans l'annotation CATH, l'homologie est en grande partie estimée sur la séquence bien que des cas particuliers existent. L'annotation structurale est ensuite faite par des algorithmes qui se basent sur des matrices de contact [185]. DALI et ECOD

définissent les domaines de manière entièrement automatique. DALI compare les structures directement en utilisant des matrices de distance qui permettent de clusteriser les protéines dans un espace de structures, les protéines proches dans cet espace partageant le même fold [185]. ECOD utilisant trois approches différentes (par séquence, profil et structure en utilisant DALI). Des méthodes ont été proposées pour définir les domaines de manière univoque, par exemple SWORD qui se base sur une mesure d'ambiguïté d'une protéine pour en proposer plusieurs décompositions alternatives en domaines afin de choisir la plus consensuelle [194]. Les méthodes les plus performantes actuellement sont Chainsaw [195] et UniDoc [196], qui sont complémentaires plutôt qu'antagonistes ; les deux sont utilisées dans le pipeline de classification de CATH [193].

Les classifications CATH, SCOP2 et SCOPe sont hiérarchiques (les niveaux dans SCOP2 et SCOPe étant les mêmes, bien que SCOP2 ai plutôt vocation à remplacer la classification hiérarchique par un réseau), ce qui permet de comparer les structures à plusieurs niveaux de finesse. Par ordre de finesse croissante ces niveaux sont dans CATH la Classe, l'Architecture, la Topologie (qui correspond au fold) et l'Homologie (qui correspond à la Superfamille)(Figure 18 A). L'annotation structurale se déroule en quatre étapes principales avec un pipeline semi-automatisé. La structure primaire de la protéine est d'abord clusterisée avec les structures primaires les plus proches dans la base de donnée de protéines CATH (niveau d'identité à 35%). Si des structures primaires proches sont trouvées, la deuxième étape consiste à évaluer si la protéine a un ou plusieurs domaines, de la façon décrite au paragraphe précédent. Si elle en a plusieurs, elle est découpée de sorte à obtenir une séquence pour chaque domaine. Les domaines sont à nouveau clusterisés avec la base de donnée de domaine CATH (niveau d'identité à 35%). Ces deux étapes de clustering se basent sur des Hidden Markov Models créés dans CATH qui sont représentatifs de la diversité des domaines protéiques de bases de données de protéines de référence Uniprot [197] et Ensembl [198]. Une fois les domaines clusterisés, leurs Classe, Topologie et Homologie sont déterminées automatiquement. Les Architectures sont déterminées manuellement. Cette assignation manuelle repose en grande partie sur les principes de classification des structures établis par Richardson [173]. La Classe décrit la proportion d'hélices α et feuillets β dans la structure ; l'Architecture l'arrangement global des structures secondaires entre elles; la Topologie l'arrangement plus fin des structures secondaires et la façon dont elles sont connectées entre elles. Deux domaines ayant la même Homologie sont proche du point de vue évolutif et réalisent des fonctions très similaires. Il y a quatre niveaux hiérarchiques dans SCOP2 et SCOPe, la Classe, le Fold, la Superfamille et la Famille. La Classe 1 dans SCOPe/SCOP2 et CATH rassemble les folds composés uniquement d'hélices α ; la 2 ceux composés uniquement de feuillets β. Pour les structures de domaines ayant à la fois des hélices α et des feuillets β, CATH ne propose qu'une Classe, la 3. SCOPe et SCOP2 font une distinction entre les protéines dans lesquelles les hélices et feuillets sont chacun regroupés entre eux (classe 4, α+β) et celles dans lesquelles les structures secondaires sont mélangées dans la protéine (classe 3, α/β). En plus des Classe 1,2 et 3, CATH a également une Classe 4 (domaines avec peu de structures secondaires) et une Classe 6 (domaines spéciaux) qui regroupe des Architecture non globulaires (Fig.18 B). SCOPe propose en plus des quatre catégories décrites ci-dessus huit autres catégories, en particulier une pour les protéines non naturelles. SCOP2 n'a qu'une autre classe, les petites protéines, mais permet aussi de parcourir la classification par type de protéine : globulaire, membranaire, fibreuse ou sans structure. En termes de diversité, CATH recense 42 Architectures (dont la moitié se trouvent dans la Classe 2 et environ un quart dans la Classe 3), 1472 Topologies (634 dans la Classe 3 et 404 dans la Classe 1) et 6631 Superfamilles (85% sont réparties entre les Classe 1, 2 et 3). Cette diversité sera probablement revue à la hausse après validation par curation manuelle de centaines de nouveaux folds potentiels identifiés à l'aide d'AlphaFold [184]. SCOP2 et SCOPe contiennent 1562 et 1257 Folds, 2816 et 2067 Superfamilles et 5936 et 5084 Familles, respectivement. 72544 domaines ont été annotés dans SCOP2 contre 348214 dans SCOPe. La classifications ECOD est

sensiblement différentes : elle distingue des architectures, des X-groupes, des H-groupes, des T-groupes et des F-groupes. Les architectures ECOD ont des similarités avec celles de CATH et décrivent grossièrement l'organisations des structures secondaires entre elles dans la structure, mais il n'y en a que vingt et une. Les X-groupes (2459) rassemblent tous les domaines (homologues ou non) ayant la même architecture, les H-groupes (3717) les domaines homologues de même architecture, les T-groups (3950) ceux qui sont homologues sans avoir la même topologie et les F-groupes (12948) ceux appartenant à la même famille (donc étant homologues avec la même topologie). Les X-groupes sont ce qui se rapproche le plus du fold de SCOP et de la Topologie de CATH, mais des différences conceptuelles existent entre les deux.

Dans l'ensemble, les domaines définis par CATH, SCOP, DALI et ECOD sont différents: seul 67559 (à 80% d'identité de séquence) sont partagés entre CATH, SCOP et ECOD ; DALI produit les résultats les plus divergents [185], [189], [199]. En revanche, CATH et SCOP partagent 70% de leurs domaines, ce qui en fait les bases de données avec la plus grande similarité [199]. La part des domaines annotés manuellement de CATH et tous les domaines de SCOP sont généralement plus en adéquation avec les résultats de cristallographie, notamment en comparaison de DALI [172]. La part des domaines définis automatiquement est en outre plus proche de ceux de DALI [172]. Concernant l'annotation structurale, les résultats de DALI sont plus en adéquation avec ceux de SCOP [185]. Les différences observées avec CATH proviennent en partie du fait que CATH a tendance à agréger les folds de domaines avec plus d'une topologie possible plutôt que d'en créer plusieurs [185]. Une trentaine de folds de SCOP sont agrégés dans un petit nombre de Topologies dans CATH (notamment le Rossmann fold de CATH qui rassemble plusieurs folds différents de SCOP) [199].

I.3. diversité au-delà des structures de domaines protéiques

L'existence de domaines délimités différemment selon les bases de données est en partie liée au fait que certains (possiblement jusqu'à 4% de la PDB) domaines peuvent se replier de plusieurs manières [200]. Dans CATH, environ 14% des folds peuvent être attribués de manière ambigüe à plusieurs Superfamilles au sein desquelles la divergence structurale entre domaines est très élevée [201]. Pour prendre en compte les différents repliements possibles de ces domaines, la notion de *metafold* a été proposée. Un *metafold* regroupe des folds proches topologiquement et avec des relations d'homologie [185], [202], [203]. Il peut être considéré comme un fold consensus pour des domaines avec plusieurs folds possibles, donc une structure moyenne, probable et dynamique. Il y en aurait environ 1130 différents [185]. Les *concepts* sont une autre définition proposée. Ils décrivent à une échelle plus fine que les folds la géométrie des assemblages de structures secondaires ainsi que leurs motifs de contact et se résument en 1493 structures différentes [204]. À une échelle encore plus fine, les TERtiary structural Motifs (TERMs) visent à décrire l'environnement structural de chaque acide aminé dans une protéine en prenant en compte les résidus proches (deux positions dans la structure primaire en amont et en aval) ainsi que ceux susceptibles d'être en interaction avec l'acide aminé d'intérêt. Environ 625 TERMs différents suffisent à décrire 50% des résidus dans la PDB [205].

II. phénomène de repliement et acquisition d'une structure à l'origine d'une fonction

Dans ce sous-chapitre, je décrirai les différents évènements du début du mécanisme de repliement de la chaîne polypeptidique pendant la traduction jusqu'à l'acquisition d'une structure complète (dite conformation native) réalisant une fonction. La première section détaillera donc les caractéristiques principales du processus de repliement, la seconde la diversité des structures protéiques acquises à l'issu de ce repliement et la troisième le lien entre ces structures et leurs fonctions.

II.1. processus de repliement

Le postulat d'Anfinsen, publié en 1961, stipule qu'une séquence donnée d'acides aminés donnée se replie spontanément en une structure 3D dans un environnement propice [206]. Ce repliement débute dans le ribosome dès le début de la traduction et s'achève généralement dans le cytoplasme, en particulier lorsqu'il fait entrer en jeu des chaperonnes. Le taux de traduction et la présence de certains cofacteurs a un impact sur le repliement des chaînes polypeptidiques en sortie de ribosome, leurs variations pouvant même dans certains contextes induire des repliements différents pour la même structure primaire [22], [207]. À l'échelle d'un fold, le temps nécessaire au repliement, aussi appelé taux de repliement, dépend entre autres de sa longueur en nombre de résidus. Les protéines avec des folds complexes ont généralement les taux de repliement les plus faibles [208].

Pour les folds globulaires, le processus de repliement démarre généralement dans une région précise, qui constituera à terme son noyau [209]. Ce dernier est souvent composé d'une proportion élevée d'acides aminés hydrophobes, fondamentaux pour le processus de repliement car les liaisons qu'ils établissent entre eux (liaisons hydrophobes) sont parmi les plus fortes des liaisons spontanées [210]. La position de ces résidus dans la chaîne polypeptidique est très conservée au cours de l'évolution. Elle est sous une pression de sélection plus forte que les acides aminés eux-mêmes, dont les substitutions ne sont pas contre sélectionnées tant que le résidu appartient à la même classe d'hydrophobicité [209]. Une fois que le noyau est replié, le reste du processus est différent selon le niveau de complexité de la protéine. Dans le cas des protéines monodomaines, et de certains domaines dans des protéines multidomaine, la cinétique est simple. Elle se déroule en deux étapes, sans intermédiaires observables [208], [211]. Dans les autres cas, le processus de repliement est plus complexe avec de nombreuses transitions entre différents états qui peuvent être observables en conditions physiologiques [208], [212]. Dans les protéines multidomaine, le processus de repliement est en partie lié à l'ordre des domaines entre eux par rapport aux extrémités N et C-terminale. En effet, les différents domaines ont tendance à se replier indépendamment si leurs interfaces avec les autres sont petites et flexibles ; ils sont plus dépendant les uns des autres lors de ce processus si leurs interfaces sont denses et hydrophobes. Certaines protéines rassemblent des folds dont les connexions se font via des interfaces réduites mais dont le repliement se fait de façon coopérative, comme les spectrines [213]. L'ordre des domaines entre eux influe donc sur leurs interfaces de contacts, et n'est engénéral pas aléatoire [174], [214], [215]. La très grande majorité des combinaisons de deux folds ne sont trouvées que dans un ordre et pas l'autre, à l'exception de certaines impliquant par exemple un Rossmann fold ou une EF-hand [214], [215]. Dans le cas du Rossmann fold, toutes les possibilités de connexion avec d'autres domaines existent : via l'extrémité C ou N-terminale du domaine catalytique, inséré dans le domaine catalytique ou inversement (domaine catalytique inséré dans le domaine adoptant le Rossmann fold) [215].

L'enfouissement des résidus hydrophobes au cœurs de la structure est un processus actif ou passif au terme duquel elle atteint un minimum (local ou absolu) d'énergie libre de Gibbs *G* (Figure 19). Cette grandeur contrôle le repliement d'un point de vu thermodynamique ainsi que l'état de la structure finale [216]. Une structure est considérée repliable si son énergie libre à l'état natif est inférieure à l'énergie libre minimale accessible à toutes ses conformations mal repliées [217]. En outre, l'environnement moléculaire dans lequel le repliement a lieu conditionne également l'état de la structure repliée. Celui-ci est surpeuplé et en perpétuelle agitation à cause de l'agitation moléculaire ; en fonction de sa configuration, le minimum d'énergie qu'il est possible d'atteindre pour une structure peut changer. Dans certaines cas, le delta d'énergie libre du repliement est même positif (donc énergie-dépendant) et ne peut être atteint que via l'intervention de chaperonnes [216] (Figure 19). Les protéines dont le paysage énergétique est formé de cette façon possèdent plusieurs conformères, chacun pouvant adopter un ou plusieurs folds spécifiques [212]. Deux stratégies sont observées : soit un grand nombre de conformères d'énergies proches, soit peu de conformères avec des barrières énergétiques fortes entre eux, pouvant nécessiter une chaperonne pour passer de l'un à l'autre [218]. Plus les repliements sont rigides et plus ils ont tendance à avoir un paysage énergétique correspondant à la deuxième stratégie, c'est-à-dire avec un seul ou un faible nombre de minima locaux et peu de conformères possibles [219]. Il existe plusieurs processus de repliements différents pour atteindre les différents conformères possibles. Dans certains, la région à changement de fold s'hétéro ou s'homo-oligomérise via une région hydrophobe ; dans d'autres, cette région s'homo-oligomérise via une région à liaison hyrdogènes [200]. Le changement de fold peut aussi se faire sans modification du niveau d'oligomérisation [200]. Dans tous les cas, les structures capables de changer de fold possèdent des régions qui restent structuralement stables quand d'autres subissent des phénomènes de repliement contexte-dépendant, avec des repliements indépendants les uns des autres [200].

Pour les folds non globulaires (non structurés ou spéciaux), le processus de repliement est particulier et n'est en général déclenché que lors de la liaison avec un substrat [220]. Certains folds restent même toujours désordonnées et réalisent leurs fonctions dans cet état. Leurs structures sont généralement des random coil ou des molten globule [220]. À l'échelle des protéines, la désorganisation concerne certaines régions, plus ou moins longues. Des différences sont observées entre domaines du vivant : il y a proportionnellement plus de protéines avec des régions désordonnées de plus de 50 résidus consécutifs chez les Eucaryotes (plus de 30%) que chez les Bactéries (environ 4%) et chez les Archées (environ 2%) [220], [221], [222]. À noter que la proportion de protéines avec des régions désordonnées longues n'est pas corrélée avec la taille du protéome [222].

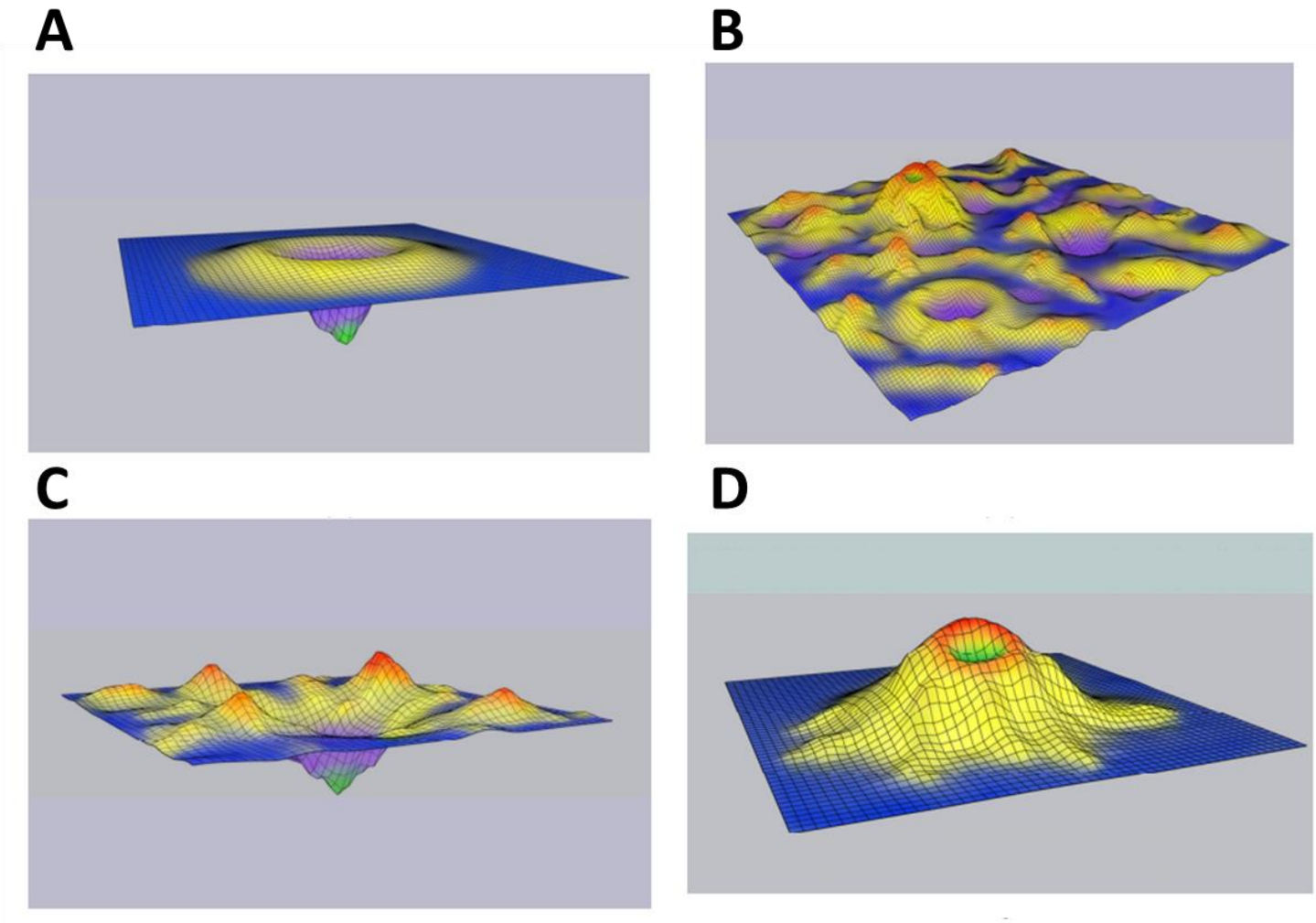


Figure 19. Possibles paysages énergétiques associés au repliement d'une protéine *in vitro* et *in vivo*. Zones bleues = conformations « parfaitement dépliées » sans interactions stables entre les résidus non contigus ; zones rouges et violettes = conformations plus compactes (l'énergie dans les zones rouges résulte des interactions entre la protéine et les composants cellulaires dans un environnement encombré) ; zones vertes = conformation native. **(A)** Paysage énergétique canonique en forme d'entonnoir qui ne s'applique probablement qu'au repliement de petites protéines thermodynamiquement stables, tel qu'il se produit spontanément *in vitro*. **(B)** *In vivo*, le paysage énergétique du repliement est mal connu. Il est très probablement complexe, dynamique et façonné par des interactions entre le polypeptide qui se replie et les molécules dans son environnement cellulaire proche. **(C)** Probable paysage énergétique du repliement de la même protéine qu'en **(A)** *in vivo*. **(D)** Les conformations natives de la plupart des protéines sont susceptibles d'occuper des minima thermodynamiques locaux avec une énergie libre de Gibbs plus élevée que leurs conformations dépliées (ΔG positif du repliement). Une telle conformation native ne peut résulter que d'un processus de repliement actif et donc consommateur d'énergie. Extrait de Sorokina *et al.* [216].

II.2. espace des structures tridimensionnelles

Le repliement d'une protéine abouti à une structure tridimensionnelle qui peut être caractérisée par ses structures secondaires. La majorité des folds globulaires sont construits autour d'hélices α et feuillets β liés entre eux par des liaisons hydrogènes, dont le nombre et la position définissent son identité (sa Topologie d'après la définition CATH) [223]. La diversité structurale en périphérie des hélices et des feuillets, notamment en termes de boucles et de filaments, est élevée [223], [224]. Le nombre de résidus dans la structure primaire a un impact sur les structures secondaires : plus il est élevé et plus la proportion de feuillets a tendance à augmenter, contrairement à celle des hélices [205]. Les hélices α sont parfois remplacées par des hélices β (feuillets β tronqués organisés en un squelette hélicoïdal) dans certains séquences courtes [225].

Plusieurs métriques existent pour mesurer les propriétés des structures secondaires et de la structure d'une protéine en général. Le Contact Order (CO) permet par exemple d'estimer la proportion d'interactions locales versus non locales et de comparer finement la topologie de deux repliements de longueurs différentes [208], [211]. Il augmente plus des résidus lointains établissent des liaisons entre eux. Les Super Secondary Structures (SSS) consistent en trois sous-structures (α -hairpin, β -hairpin et les βαβ-unit), qui sont retrouvées dans un grand nombre de folds différents et dont la quantification peut permettre de classifier les structures [226]. Les différences structurales entre enzymes monodomaines peuvent être mesurée à l'aide d'une grandeur appelée « polarité », qui décrit le degré de séparation entre les résidus du site actif et les autres [227], [228].

Les espaces de structures peuvent recenser par exemple la structure des chaînes principales (Figure 20 A) ou de la surface (Figure 20 B) des protéines. Dans l'ensemble, ces espaces sont continus, c'est-à-dire que toutes les structures sont connectées les unes aux autres par des changements structuraux mineurs [201], [229], [230]. Certaines régions sont cependant discontinues. De façon générale, plus le niveau de similarité entre structures considérées est faible et plus l'espace peut être discrétisé, et donc conduire à une classification hiérarchique [172], [231]. Dans l'espace structural des chaînes principales (Figure 20 A), les folds all α, all β et α/β sont clairement séparés [229], [232], [233]. Les folds α+β ont plutôt tendance à se trouver à l'intersections de ces trois catégories, résultant en un positionnement moins net [233]. Les folds avec peu de structures secondaires (type coil-like) se positionnent près de l'origine de l'espace des structures, à proximité des folds α+β mais plus proche encore du point de convergence entre les all α, all β et α/β [232]. La taille des familles associées à chaque fold a un impact sur la position de ces dans l'espace des structures: les folds ayant les plus petites familles ont tendances à être plus proche de l'origine [232], [233]. Les folds constitués d'un grand nombre de répétitions de SSS se recouvrent généralement dans cet espace ; cela ne concerne en outre qu'une minorité de folds [201]. L'espace des structures de surface (Figure 20 B) est différent de celui des chaînes principales dans la mesure où les folds all α, all β et α/β n'y sont pas clairement séparés. Les formes de surface ont globalement un degré de similarité plus élevé que celles des chaînes principales [229], [230]. Il y a par exemple un plus grand nombre de folds all α dont la forme de surface présente le degré de similarité maximal avec un fold all β (1291) qu'avec un fold all α (713) [230]. Les protéines à structure quaternaires peuvent atteindre des portions de l'espace des structures inaccessibles aux protéines monodomaines par combinaison d'éléments structuraux plus ou moins distants dans cet espace. Leur structure a alors tendance à être plus proches de celle de folds α/β ou α+β que de folds all α ou all β [233].

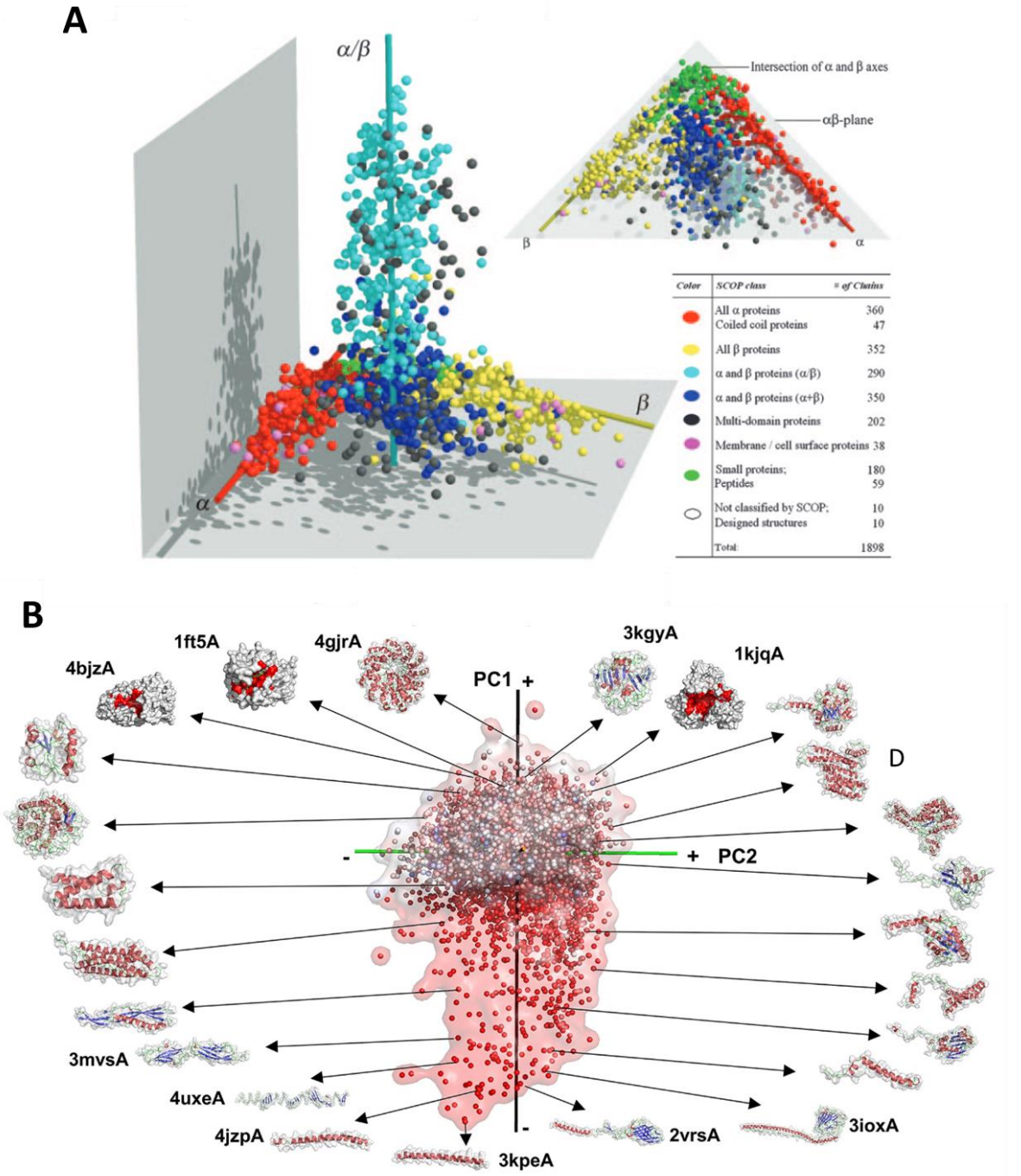


Figure 20. Exemples d'espaces de structures de protéines. (A) Deux visualisations de la carte de l'espace des structures de protéines. Chacune des 1898 protéines est représentée par une sphère, dont la position est déterminée par la composition en hélices α, feuillets β et à leur agencement. À gauche, visualisation en 3D avec chaque dimension correspondant à une Classe (α à gauche, β à droite, α/β en haut). L'intersection des classes α et β correspond à l'origine. Les couleurs et le nombre de protéines dans chacune des Classes sont listés en bas à droite. En haut à droite, l'espace des structures est représenté dans le plan α/β. Extrait de Hou *et al*. [233]. **(B)** Espace basé sur la forme générale du repliement (indépendamment de sa composition en structures secondaires). Chaque point correspond à une protéine. La distance entre points représente la similarité de leurs formes. La couleur indique l'excentricité (degré d'élongation d'une forme) avec un gradient allant du bleu (0.0, sphère) au rouge (1.0, forme allongée). La représentation est dans le plan PC1-PC2. Extrait de Han *et al*. [230].

II.3. lien entre structure des domaines protéiques et fonctions

Dans l'ensemble, une protéine ou un domaine ne peut réaliser sa fonction que s'il est replié. Dans les enzymes, la fonction catalytique est réalisée dans le site actif. la structure du site actif impacte de façon importante la fonction réalisée [234]. Leurs ligands peuvent dans certains cas participer au processus de repliement [235]. La fonction en tant que telle est maintenue par la préservation de certains résidus clés ainsi que leur position dans le site actif, deux propriétés qui peuvent être obtenues avec des folds différents [236]. Il est donc imaginable que des protéines puissent changer de folds tout en maintenant la même structure de site actif afin de réaliser la même fonction tout en s'adaptant à des contextes cellulaires différents. Certaines protéines, dites métamorphiques (comme la lymphotactine par exemple), peuvent adopter plusieurs structures, et donc réaliser plusieurs fonctions [237], [238], [239]. Elles changent de folds de façon réversible en fonction du contexte cellulaire ou environnemental, tout en conservant généralement un cœur structuralement stable [200], [237]. Les protéines avec de longues régions désordonnées constituent un autre cas particulier. Elles peuvent avoir des fonctions bien précises notamment de régulation et de signalisation. Leurs régions désordonnées leurs permettent justement d'avoir plus de partenaires et de s'impliquer dans de nombreux mécanismes cellulaires différents [220], [240].

Certains folds sont impliqués dans une diversité très importante de fonction. Par exemple, le Rossmann fold est impliqué dans près de 40% des réactions métaboliques du vivant [241], [242]. La diversité structrale des protéines qui en disposent est élevée [233]. Deux protéines constituées d'une combinaison de Rossmann fold avec le même domaine mais dans un ordre différent n'ont dans la majorité des cas pas la même fonction [215]. La structure du Rossmann fold est caractérisée par une vaste cavité qui peut abriter une diversité très élevée de ligands de tailles et de formes différentes, malgré une préférence pour les nucléotides [241]. Dans le cas du P-loop (3.40.50.300), c'est l'ATP. Les protéines monodomaine qui l'adoptent ont généralement des fonctions de kinases ou de transférases, et les multidomaines des fonctions de production d'énergie (via hydrolyse de l'ATP ou du GTP et oxidation/réduction de cofacteurs type NAD ou NADP) [214]. Certains autres folds comme le TIM barrel sont aussi adoptés par de nombreuses protéines avec des fonctions différentes (parfois plusieurs centaines) et sont associés à beaucoup de superfamilles très diversifiées du point de vue fonctionnel [227], [243], [244]. De la même façon que pour le Rossmann fold, ils accomplissent en général des fonctions basiques, à partir desquelles les autres domaines, ayant généralement une structure plus complexe, peuvent réaliser une fonction plus spécifique [174], [245], [246].

La diversité fonctionnelle n'est pas répartie de manière homogène dans l'espace des structures (Figure 21 A). La région de cet espace regroupant les folds α/β en concentre la plus grande diversité, et celle-ci diminue plus avec la distance (Figure 21 B). La diversité fonctionnelle élevée des folds α/β est liée entre autres à leur ancienneté et leurs multiples apparitions indépendantes, qui leurs ont permis d'explorer l'espace des fonctions de façon beaucoup plus conséquente que les folds des autres classes [229], [246]. La taille des familles de protéines est en outre mieux expliquée par le type de fonctions qu'elles réalisent plutôt que leur âge. Les familles adoptant le fold all-β immunoglobulin-like sont par exemple très grandes parce qu'elles sont impliquées dans des fonctions de signalisation et d'interaction cellule-cellule, qui sont particulièrement diversifiées [245], [247]. Les histones, dont la fonction est très conservée, sont associées à une famille de petite taille bien que très ancienne [247].

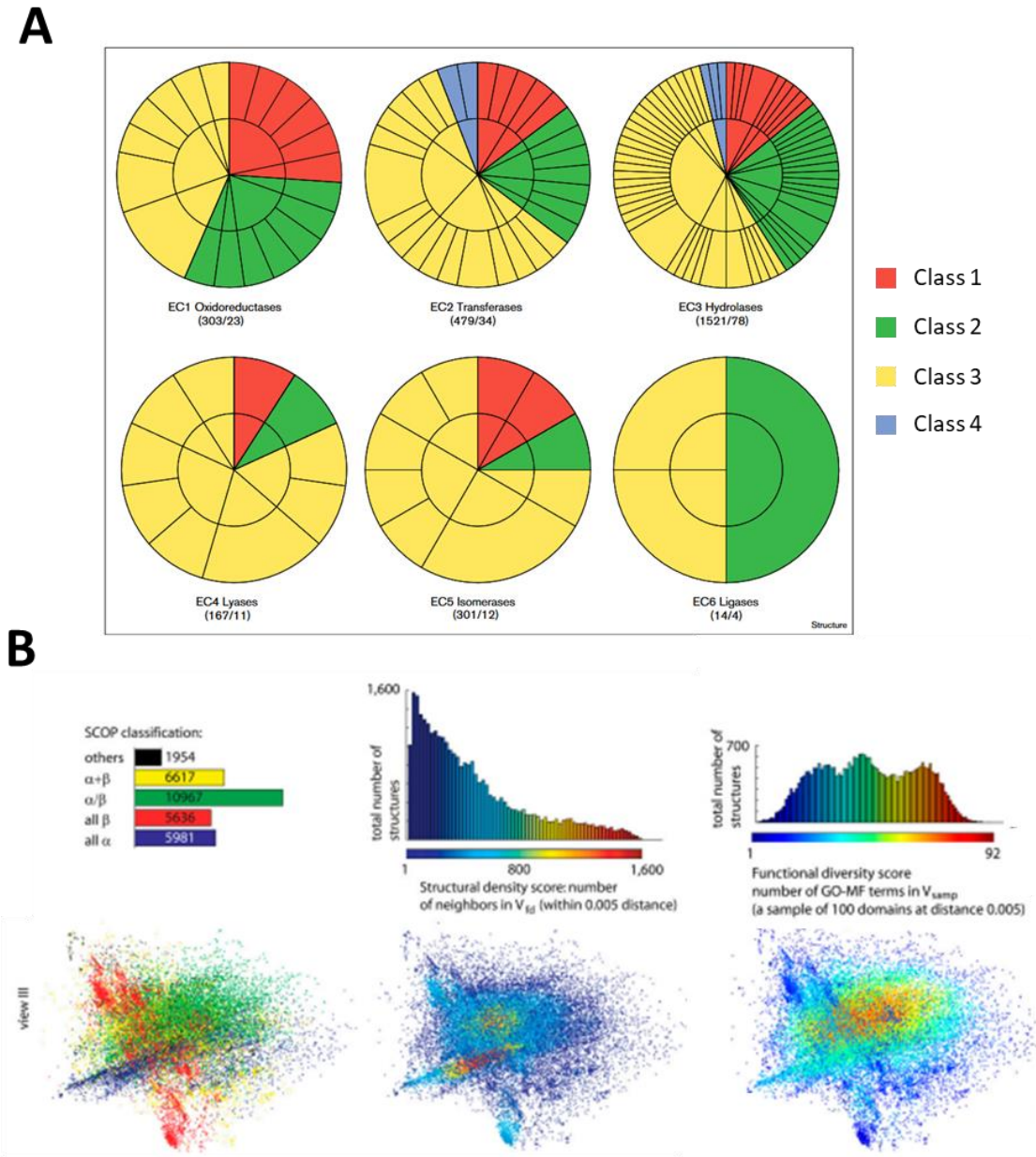


Figure 21. Diversité fonctionnelle des domaines protéiques. (A) Classification structurale de toutes les enzymes monodomaines de la PDB, groupées selon leur numéro EC. Le nombre d'exemples, suivi du nombre de familles non-homologues (définies par CATH) sont indiqués entre parenthèses sous chaque diagramme. Extrait de Martin *et al.* [235]. **(B)** Lien entre espace des structures et espace des fonctions. Chaque point représente un domaine SCOP. La distance entre les points reflète la dissimilarité structurale entre leurs domaines. À gauche, la couleur de chaque point indique sa Classe (la correspondance des couleurs est indiquée en haut à gauche). Au milieu, la couleur représente la densité en domaines dans les différentes régions de l'espaces. Les zones rouges sont celles qui concentrent la plus grande diversité de domaine. La légende indiquant les valeurs de densité est en haut au milieu, sous la forme d'une distribution des densités de domaines. À droite, la couleur de chaque point indique le degré de diversité fonctionnelle du domaine. La légende correspondante est en haut à droite, sous la forme d'une densité du score de diversité fonctionnelle. Extrait de Osadchy & Kolodny [229].

III. différences d'usage des folds entre protéomes et au sein d'un protéome

Après avoir décrit le processus de repliement, la diversité des structures et leurs fonctions, ce sous-chapitre sera consacré à une description plus statique de l'usage des folds dans les protéomes à deux niveaux : au sein d'un protéome et entre protéomes. La première section décrira les caractéristiques de l'usage des folds au sein d'un protéome, qui peuvent être étudiées via la distribution des occurrences des folds. La seconde section présentera comment certains usages de folds sont protéomes-spécifiques à l'échelle de l'arbre du vivant.

III.1. distribution des structures de domaines protéiques dans les génomes

La distribution des structures de domaines protéiques dans les génomes peut être observée à différents niveaux. La première sous-section décrira la variabilité de la taille des familles de domaines protéiques et comment cette variabilité est à l'origine de plusieurs catégories d'occurrence de folds. Dans la dernière sous section, je rentrerai plus en détail sur les propriétés de la distribution des folds dans les protéomes et les modèles mathématiques associés.

III.1.a. taille des familles de domaines et catégories de folds

Les domaines protéiques adoptant le même fold appartiennent à des familles de séquences. Chaque fold peut être considéré comme un réseau formé d'un ensemble de séquences (une Superfamille dans la définition CATH) séparées par une mutation ponctuelle, chacune correspondant à un domaine, adoptant la même structure et réalisant globalement la même fonction [212]. Certains réseaux peuvent aussi être connectés par certaines séquences qui, à un mutation près, vont adopter un fold plutôt qu'un autre [212]. Le regroupement de tous ces réseaux est appelé « super-network » [248]. Environ 4% des superfamilles dans CATH correspondent à des super-networks, et ces familles adoptent 25% de toutes les structures dans cette base de donnée [201]. Près de 40% des domaines dans les génomes appartiennent à ces superfamilles [201].

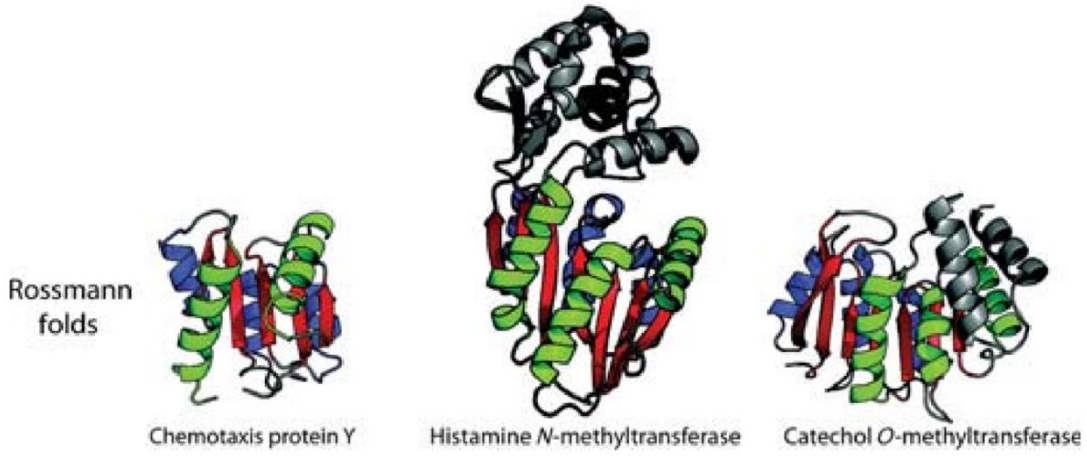


Figure 22. Exemples de protéines avec un domaine Rossmann. Extrait de Schaeffer & Dagget [172].

En général, les protéines provenant d'une même famille partagent au moins 30% d'identité de séquence, un pourcentage qui peut baisser à environ 20% pour deux protéines appartenant à des familles différentes mais apparentées [247]. La taille des familles peut donc varier de plusieurs centaines de domaines différents à quelques-uns, voir un seul, ce qui est le cas pour la majorité d'entre elles [249]. Les folds associés aux plus grandes familles agissent donc comme des « attracteurs » de

séquence et sont adoptés par une très grande diversité de domaines. Ils sont appelés superfolds [245], [248], [250], [250], [251]. Il en existe plusieurs, le plus important étant le Rossmann fold, ou three-layered α/β/α sandwich (3.40.50). Son Architecture, qui correspond à un α/β doubly wound (3.40), est observée dans plusieurs autres superfolds. Le Rossmann fold est le fold adopté par le plus grand nombre de domaines (38685 PDB structures, soit 20% de toutes les structures connues) [233], [242]. Ses spécificités structurales sont dans une certaine mesure partagées par tous les superfolds, qui ont en général une proportion plus élevée de SSS que les autres folds [226]. La connexion de leurs feuillets parallèles se fait à droite plutôt qu'à gauche [252]. Les autres superfolds sont la globin (1.10.490), le trefoil (2.80.10), l'α-up-down, la greek key-immunoglobulin (2.60.40), le split αβ sandwich (plusieurs fold dans l'Architecture 3.30), le jelly roll (2.60.120), le UB αβ Roll (3.10.20) et le TIM barrell (3.20.20) [252] [250] (Figure 23). Les superfolds ont aussi la particularité d'être souvent retrouvés dans des protéines multidomaines en combinaison avec d'autres domaines. Le Rossmann fold forme par exemple des combinaisons avec au moins sept familles différentes de domaines catalytiques [215] (Figure 22). Il existe deux autres catégories de folds, les mesofolds et les unifolds [251]. Les unifolds sont définis par le fait qu'ils ne sont associés qu'à une seule famille protéique. Tous les autres folds sont des mesofolds [251]. Les folds de ces catégories sont généralement trouvés en combinaison avec une faible diversité d'autres folds, voir aucun. Ainsi près d'un tiers des familles de domaines ne forme de combinaison avec aucune autre famille. Les domaines de ces familles ne sont donc trouvés que dans des protéines monodomaine ou en combinaison avec un domaine appartenant à la même famille, c'est-à-dire en « tandem » [174], [214].

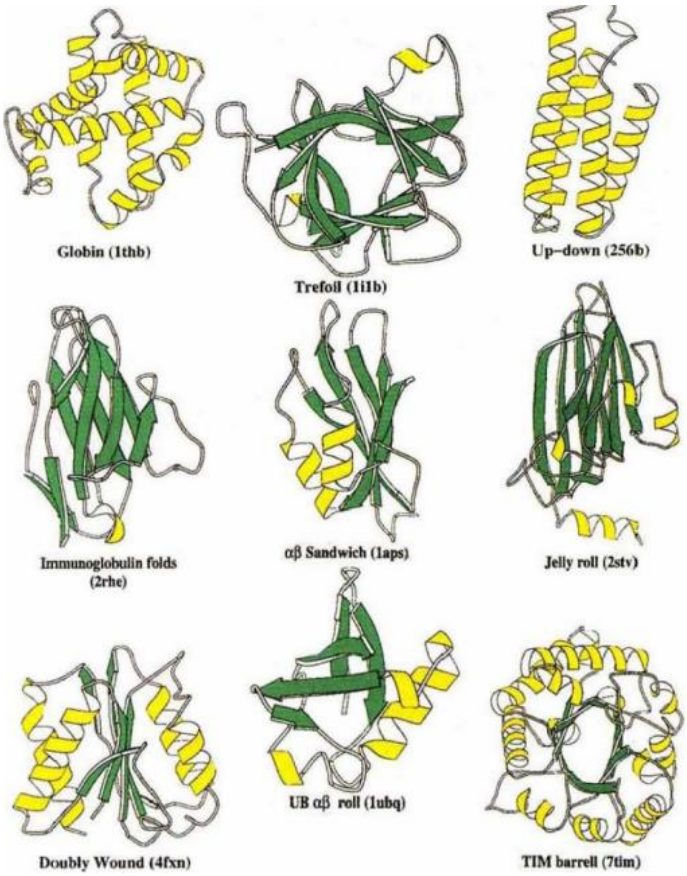


Figure 23. Structure de neuf superfolds. Extrait de Orengo *et al.* [250].

III.1.b. distribution des domaines dans les protéomes

La conséquence numérique des différences de taille de familles de domaines et de l'existence de trois catégories de folds est que la distribution de plusieurs grandeurs associées aux domaines dans les protéome peut être modélisée par une loi puissance (ou une Generalized Discrete Pareto, qui est une généralisation de la loi puissance, donc le même type de loi que celles utilisées pour modéliser la distribution des abondances des espèces dans les communautés planctoniques (*Chapitre 1.IV: lois mathématiques décrivant la composition des communautés*)) [253]. Ces grandeurs sont la distribution des occurrences des domaines [249], [253], [254], [255], [256], la diversité de leurs partenaires dans les protéines multidomaines [174], [214], [254], les occurrences ou nombre de copie de folds dans un génome [180], [223], [257], [258] (Figure 24 A) ainsi que le nombre de familles par folds ou la taille des familles de domaines protéiques [181], [247], [254], [255], [257] (Figure 24 B,C). Cette loi est dite invariante d'échelle, c'est-à-dire que, dans le cas de la taille des familles de protéines, un sous-ensemble de ces familles suivra également une loi puissance malgré une certaine variabilité dans ses paramètres. Cette variabilité est en partie liée au fait que les structuromes sont des systèmes finis (un protéome est constitué d'un nombre fini de protéines). Cet effet est particulièrement important pour les familles rassemblant un nombre très important de domaines. Dans un système infini, tous les sous ensemble suivraient exactement la même loi puissance [247].

Des différences de paramètres de loi puissance sont observées entre domaines du vivant. Par exemple, les lois puissances modélisant la distribution des folds dans les protéomes viraux ont des paramètres sensiblement différents de celles modélisant la distribution des folds dans les protéomes de Procaryotes et d'Eucaryotes (Figure 24 C). Le coefficient directeur de la droite de régression associée au modèle est plus grand en valeur absolue, indiquant un plus grand nombre de domaines de faible occurrence et une plus faible occurrence pour les domaines correspondants aux superfolds. Il n'existe en revanche pas de relation claire entre nombre moyen d'occurrences d'un fold (moyenne de la valeur d'occurrence dans tous les protéomes disponibles) et valeur d'occurrence du même fold dans un protéome donné. De plus, il n'y a pas de corrélation entre variation d'occurrence d'un fold dans différents protéomes et celle d'un autre, bien que certains folds soient en combinaisons dans des protéines multidomaines [259]. Enfin, il n'existe pas non plus de relation directe entre occurrence moyenne d'un fold et nombre d'Homologies associées à la Topologie de ce fold [257].

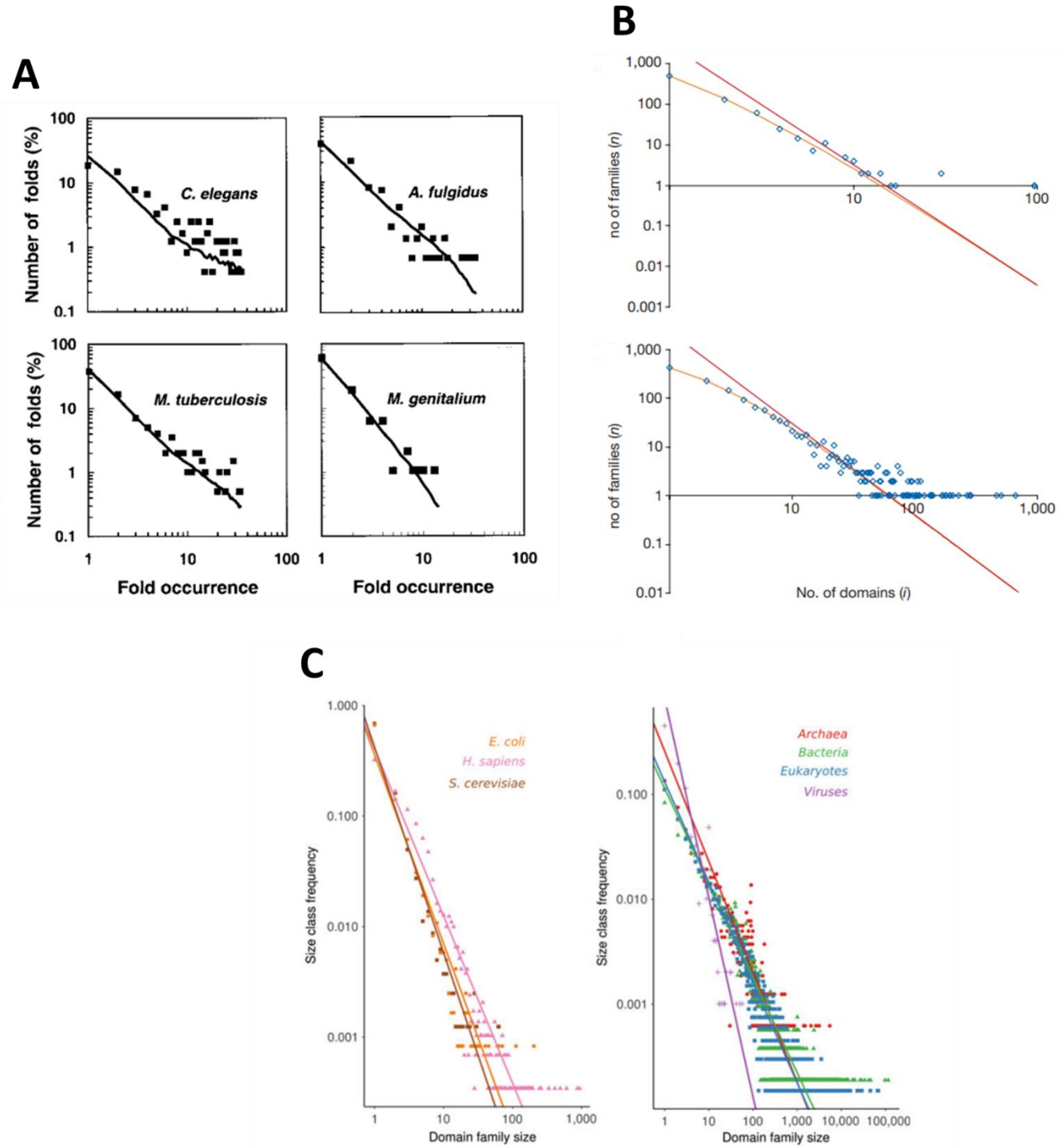


Figure 24. Exemples de modèles de loi puissance dans les protéomes. (A) Pour les occurrences des folds dans quatre génomes de référence Eucaryotes et Procaryotes. Extrait de Qian *et al.* [258]. **(B)** Pour la taille des familles de domaines protéiques dans un génome. En haut, *Thermotogata maritima*, $n=2.972(i+0.8)^{-3.0}$. En bas, *Caenorhabditis elegans*, $n=2.395(i+1.5)^{-1.9}$. La ligne rouge représente l'asymptote suivant une loi puissance. Extrait de Koonin *et al.* [255]. **(C)** Pour la taille des familles de domaines protéiques dans un génome et un domaine du vivant. La fréquence f d'une famille de domaines de taille X peut être calculée avec la relation $f = cX^{a}$. À gauche, pour *S.cerevisiae* a = -1.9, pour *E.coli* a = -1.7, pour *H.sapiens* a = -1.5. À droite, pour les Bactéries a = -0.9, pour les Archées a = -1.1, pour les Eucaryotes a = -0.8 et pour les Virus a = -1.9. Extrait de Forslund [254].

III.2. différences lignées-spécifiques d'usage des domaines et des folds entre protéomes

Cette section est dédiée à l'usage différentiel des domaines et des folds selon les protéomes à l'échelle de l'arbre du vivant, aux causes fonctionnelles de ces différences et comment celles-ci peuvent être utilisées pour classifier les espèces.

III.2.a. principales différence d'usage des domains, folds et protéines

Concernant les domaines protéiques, seusl 1.8% sont communs aux Procaryotes, Eucaryotes et Virus (Figure 25 A) [254]. Les domaines spécifiques aux Eucaryotes représentent 40% de la diversité totale (Figure 25 A) [254], [260]. Il n'est pas surprenant que près de la moitié des domaines protéiques du vivant leurs soient exclusifs : leur histoire évolutive est en effet caractérisée par un très grand nombre d'innovations fonctionnelles, notamment en lien avec l'acquisition de la pluricellularité [261]. Les Bactéries sont le deuxième groupe avec le plus grand nombre de domaines exclusifs, représentant 25% de la diversité totale (Figure 25 A) [254], [260]. Les Archées sont le domaine du vivant avec le moins de domaines spécifiques. Au moins 50 familles de domaines sont spécifiques aux Virus [260]. Les Eucaryotes partagent près de 11% de la diversité totale des domaines exclusivement avec les Bactéries, contre seulement 0.7% avec les Archées et 0.7% avec les Virus (Figure 25 A)[20], [254], [260]. À une échelle taxonomique plus fine chez les Eucaryotes, les vertébrés et les plantes terrestres ont en moyenne un nombre de domaines spécifiques et de combinaisons de domaines exclusives plus élevé que dans toutes les autres lignées [262]. Au sein des plantes, seuls 65% des domaines sont partagés entre toutes les lignées [263].

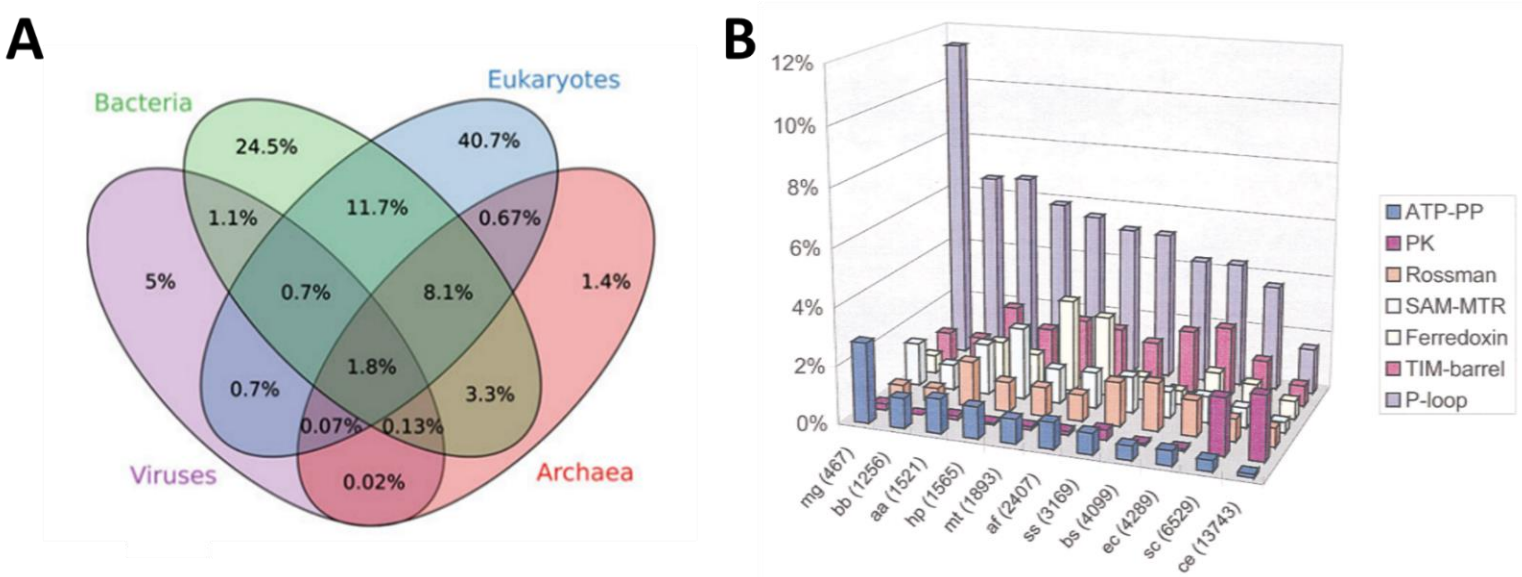


Figure 25. Différences d'usages de domaines et de folds entre domaines du vivant. (A) Distribution de 10330 domaines uniques définis avec PFAM [7] dans les quatres domaines du vivant. Extrait de Forslund [254]. **(B)** Distribution des sept folds les plus fréquents dans des protéomes de Bactéries (*Mycoplasma genitalium* = mg, *Borrelia burgdorferi* = bb, *Aquifex aeolicus* = aa, *Synechocystis* sp. = ss, *Bacillus subtilis* = bs, *E. coli* = ec), d'Archées (*Methanobacterium thermoautotrophicum* = mt, *Archaeoglobus fulgidus* = af) et d'Eucaryotes (*S. cerevisiae* = sc, *C. elegans* = ce). Extrait de Wolf *et al*. [266].

L'usage de tous les folds n'est pas le même dans toutes les lignées du vivant. Certains folds sont absents de certaines lignées, d'autres sont particulièrement fréquent dans une lignée donnée. Les Eucaryotes ont en moyenne la diversité de folds la plus élevée, suivi par les Bactéries puis les Archées dont la diversité en folds est comparable à celle Virus [254], [257], [264]. Les foldomes des Archées sont caractérisés par une proportion particulièrement élevée de superfolds [245], [254], [260]. De façon générale, les folds présents dans tous les génomes (universels) sont aussi ceux avec les plus grandes occurrences, particulièrement chez les Eucaryotes [257], [265]. Les folds α/β , bien qu'étant la catégorie concentrant le plus de folds universels et adoptés par la plus grande diversité de domaines, font exception puisqu'ils ont en moyenne des occurrences supérieures chez les Bactéries et sont moins diversifiés dans les grands génomes [229], [257]. Des tendances domaine-spécifiques sont observées à l'échelle de certains fold. Le fold de la serine–threonine protein kinase est par exemple beaucoup plus fréquent chez les Eucaryotes ; celui de la Ferredoxine l'est plus chez les Procaryotes [266] (Figure 25 B). La spécificité ou non d'un fold dans une lignée est souvent liée à sa fonction. Environ la moitié des quelques folds partagés exclusivement entre Archées et Eucaryotes sont adoptés par des protéines impliquées dans des systèmes de régulation, alors que ceux partagés entre Bactéries et Eucaryotes sont plus impliqués dans la traduction [264]. Les Opisthokontes ont des folds spécifiques impliqués dans des fonctions de motilités ainsi que dans le cycle cellulaire. Les folds spécifiques des Metazoaires sont adoptés par des protéines de signalisation et d'apoptose, ceux des Gnathostomes par des protéines du système immunitaire adaptatif, ceux des Tetrapodes par des protéines réalisant des fonctions associées à leur changement majeur de milieu de vie et ceux des Ecdyzosoaires par des neurohormones et l'hémocyanine [264].

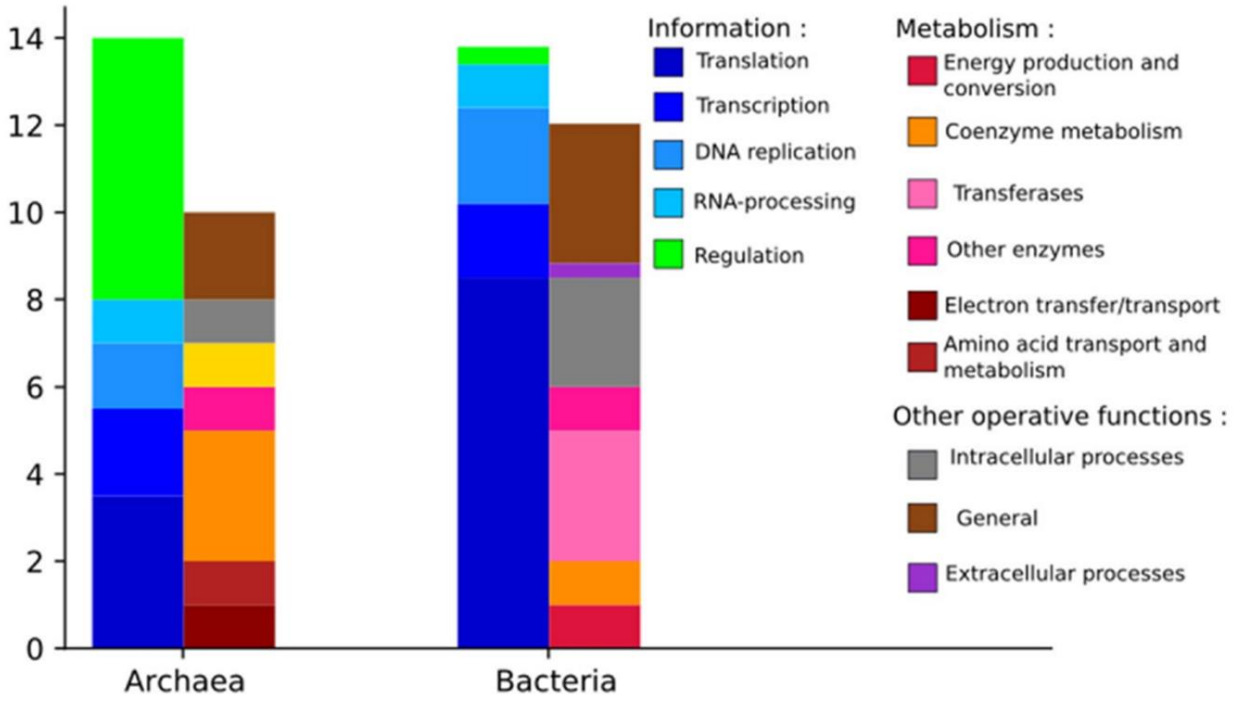


Figure 26. Principales fonctions des folds partagés entre Eucaryotes et Procaryotes. Nombre de folds partagés entre Eucaryotes et Archées à gauche et entre Eucaryotes et Bactéries à droite. Les folds liés à des fonctions d'information sont en bleu et vert, ceux liés à des fonctions opérationnelles sont en rouge, rose et marron. Extrait de Romei *et al*. [264].

III.2.b. classification du vivant basé sur l'usage des folds

L'usage des folds de chaque espèce peut être utilisé pour la classer dans l'arbre du vivant, au moins au rang du domaine ou du règne (Figure 27). Les folds sont stables à l'échelle des temps évolutifs et sont apparus de manière suffisamment fréquente, surtout lors d'évènements évolutifs majeurs ; ils peuvent donc être considérés comme des synapomorphies [20]. Les classifications d'espèces basées sur les Homologies se rapprochent plus de la phylogénie que celles basées sur les Topologies, ce qui est attendu puisqu'elles rassemblent des domaines qui sont homologues entre lignées [260]. Cet effet peut cependant être mitigé par l'effet des HGT, comme c'est le cas chez les Bactéries chez qui les protéomes sont moins stables à de grandes échelles temporelles [20]. Les folds représentent néanmoins une échelle intéressante pour classer les protéomes bactériens et peuvent permettre d'identifier des caractéristiques sur la composition des génomes et le phénotype qui ne pourraient pas l'être uniquement par des données génomique [20], [267]. À noter que les arbres d'espèces basés uniquement sur l'usage des folds all β sont généralement en désaccord avec la taxonomie [223], [260], [268]. Une grande partie des folds de cette classe sont des immunoglobulin-like, souvent impliqués dans des fonctions de signalisations propres aux pluricellulaires ; ils sont donc de mauvais marqueurs taxonomiques quand ils sont utilisés seuls [245]. Les arbres basés uniquement sur les résidus importants pour le repliement, les sites hydrophiles ou ceux codant pour des résidus chargés sont plus robustes que ceux construits uniquement avec les sites hydrophobes ou codant pour des résidus neutres [269]. En outre, l'information de présence absence des folds est plus informative que leurs occurrences, qui sont plus affectées par les duplications, HGT et pertes qui ne sont pas forcément taxon-spécifiques [260], [270].

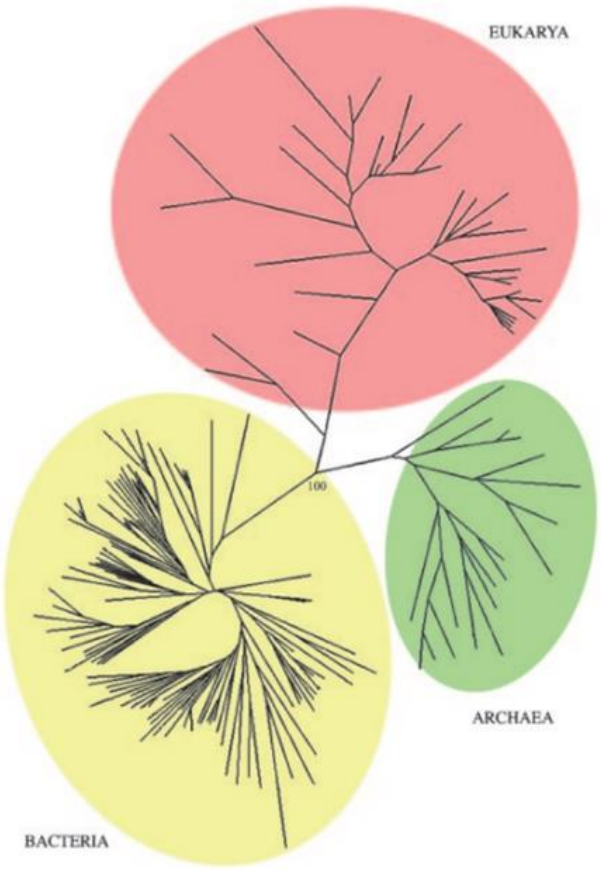


Figure 27. Classification du vivant basée sur les répertoires de folds. Phylogénie (Neighbour Joining) de 174 organismes de référence. Les valeurs de bootstrap ne sont indiquées que pour les branches majeures. Extrait de Yang *et al.* [260].

IV. évolution des structures de domaines protéiques

Ce sous-chapitre a pour but d'expliciter les mécanismes évolutifs responsables des différences d'usages de folds aux différents niveaux évoqués dans la partie précédente. La première section sera consacrée aux modes d'évolution des folds et la seconde à la description de leur histoire évolutive.

IV.1. principaux modes d'évolution des folds

Dans cette section, je décrirai dans une première sous-section que les folds évoluent principalement en conséquence d'évènements au niveau des génomes, arrivant entre autres par mutations et variants structuraux. La deuxième sous-section montrera l'importance de la pression de sélection sur les fonctions sur l'évolution des folds. Dans la troisième sous-section, je détaillerai comment mutation et pression de sélection ont façonné l'évolution des folds, en partie par convergence et par divergence. Enfin, j'expliquerai comment les modes d'évolutions des folds sont à l'origine de leurs propriété d'usage au sein des protéomes.

IV.1.a. évènements génomiques ayant un impact sur l'évolution des folds

La nature à la fois continue et discrète de l'espace des structures des folds, responsable des différences de classification détaillées plus haut, est une conséquence directe de la façon dont les folds évoluent. Un parallèle peut être fait du point de vue évolutif entre l'ontogénie et le processus de repliement des folds : de la même façon que la première récapitule la phylogénie lors du développement embryonnaire, les différentes étapes de repliements des folds complexes récapitulent dans une certaine mesure leur évolution [212].

L'évolution des folds se fait essentiellement via des évènements ayant lieu à l'échelle des génomes, par exemple les transferts horizontaux de gènes (HGT), les permutations circulaires, les expansions/réductions (notamment par duplications ou indels pouvant mener à des éliminations) ainsi que les phénomènes de fusion/fission [210], [215], [252], [256], [265], [271], [272]. Tous ces phénomènes peuvent être à l'origine de nouvelles combinaisons et/ou de réarrangement de domaines protéiques. Les permutations circulaires sont des phénomènes à l'issue desquels les anciennes extrémités N et C terminales d'une protéine se trouvent liées et d'autres apparaissent à une nouvelle position. Elles ne changent pas l'arrangement spatial des éléments secondaires mais la façon dont ils se connectent les uns aux autres, résultant en une modification du fold [272]. Elles peuvent se dérouler de deux façon différentes. La plus rare est mécanisme post-traductionnel au cours duquel les extrémités du polypeptide sont activement liées puis une boucle est activement clivée. Il n'a été observé que pour la concavaline A [273]. La plus fréquente est par « creative destruction », un mécanisme qui implique une succession de duplication, fusion puis d'adaptation du nouveau domaine [271] (Figure 28). Il a été très impliqué dans l'évolution des folds, en particulier ceux associés à des fonctions de régulation [177], [210], [214], [243]. En effet, deux domaines voisins dans une séquence (séparés par moins de trente résidus) proviennent généralement d'une recombinaison au cours de l'évolution [174], [214]. Après duplication-fusion ou recombinaison, la co-évolution des deux domaines se fait essentiellement via des mutations. Celles résultant en des substitutions non synonymes participent à augmenter la complexité de l'espace des structures, et peuvent conduire à une divergence suffisante pour que leurs folds ne soient plus les mêmes, ou même à un seul domaine codant pour une nouvelle protéine monomérique [210], [212], [254], [259], [271], [274].

De nombreuses protéines multidomaines impliquant un Rossmann fold sont probablement apparues de façon indépendante au cours de l'évolution. En effet, ce fold s'est combiné avec les domaines d'autres familles dans différents contextes. En outre, toutes les séquences protéiques impliquant un domaine Rossmann en combinaison avec un autre sont organisées dans le même ordre au sein d'une famille de protéines [215]. Dans le cas d'une duplication les deux domaines peuvent malgré leur co-évolution continuer d'appartenir à la même famille, résultant en un tandem. La recombinaison de deux domaines se trouvant appartenir à la même famille produit le même type de structure [214]. À noter que les recombinaisons ont lieu la majorité du temps entre des domaines universels et anciens plutôt qu'entre des domaines rares et présents uniquement dans certaines lignées (ou à minima entre un domaine universel et un domaine rare) [214], [259], [272]. La diversité augmente plus rapidement à l'échelle des combinaisons de domaines qu'à l'échelle des domaines eux-mêmes [214].

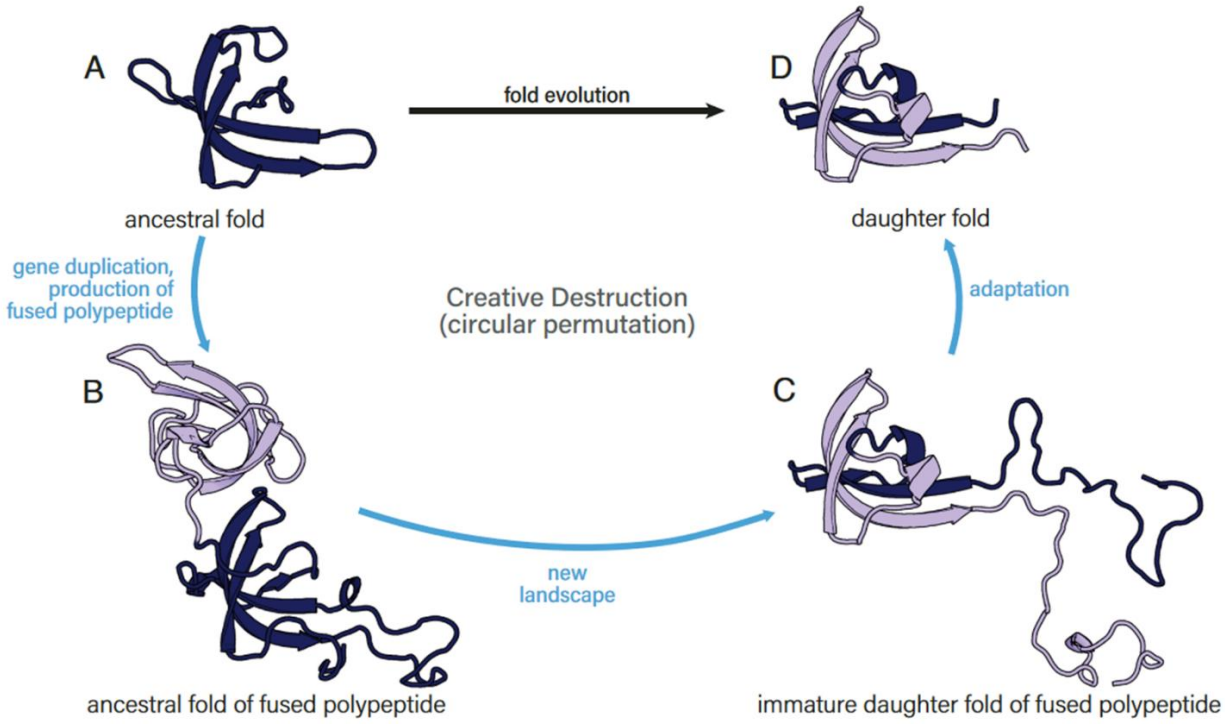


Figure 28. Exemple de mécanisme d'apparition d'un nouveau fold : permutation circulaire par « creative destruction ». Les gènes sont fusionnés, résultant en la disparition du fold ancestral et l'apparition d'un nouveau fold. Ici le fold ancestral est représenté en **(A)** (PDB: 5YYA). **(B)** Fold ancestral théorique du polypeptide fusionné après duplication du gène (PDB: 5YYA). **(C)** Fold « fils » immature. Des parties du fold ancestral ainsi que certaines de ses structures secondaires ont été détruites (PDB: 7D4A). **(D)** fold « fils » mature (PDB: 7D4A) qui a hérité d'une partie des éléments de son ancêtre. Extrait de Alvarez-Carreño *et al*. [271].

IV.1.b. importance des fonctions dans l'évolution des structures de domaines

L'évolution des folds est liée à l'évolution de leurs fonctions, qui subissent une pression de sélection au même titre que leurs propriétés structurales intrinsèques (cinétique de repliement, thermostabilité). Au sein d'un fold, un résidu peut généralement être substitué par un autre du moment qu'ils ont en commun les mêmes propriétés de charge, de taille et d'hydrophobicité [236], [275]. Il existe néanmoins des cas de convergence fonctionnelle à des pourcentages d'identités inférieurs à 40% et, inversement, des cas de domaines à plus de 40% d'identité de séquence ne

réalisant pas la même fonction [236], [247]. La divergence à long terme des orthologues est donc principalement limité par la pression sur la fonction plus que sur la structure [236].

Les évènements évolutifs à l'échelle des génomes (duplication/fusion, recombinaison, transferts par HGT, etc) peuvent être sources d'innovations fonctionnelles, en particulier dans le cas où les domaines acquièrent au terme de ces processus des repliements différents. Par exemple, un dimère rassemblant un fold avec une activité catalytique et un fold associé à la production d'énergie (de type Rossmann) peut résulter en une nouvelle protéine de type hélicase ou kinase [210]. L'évolution par créative destruction est à l'origine de nombreuses protéines de fonctions diverses, comme la Scr kinase family homology (SH3), l'oligonucleotide/oligosaccharide-binding (OB), et le cradle loop barrel (CLB) [271]. Le processus conduisant à la combinaison de plusieurs folds déjà existant mais résultant en une protéine capable de réaliser une nouvelle fonction s'appelle « exaptation » [264]. Ces exaptations permettent l'acquisition de nouvelles fonctions plus complexes, souvent impliquées dans la de signalisation et la transduction des signaux, la régulation et la réponse immunitaire, en particulier quand l'assemblage concerne plus de trois domaines [276]. Les folds sont donc bien l'unité de base du point de vue fonctionnel pour les protéines : chacun peut être considéré comme un module dont le potentiel fonctionnel peut s'exprimer seul ou s'accroître en se combinant avec d'autres. À ce titre, le Rossmann fold a eu une place particulière au cours de la diversification fonctionnelle des protéines. D'un point de vue évolutif et fonctionnel, la sélection de l'ATP comme monnaie énergétique principale et universelle des cellules a, d'une certaine façon, entraîné par coévolution, la sélection du p-loop basé sur le domaine Rossmann comme le domaine clé pour exploiter cette source d'énergie.

Le processus de fonctionnalisation des polypeptides ayant abouti aux folds est complexe à retracer. Se pose la question de qui du repliement ou de la fonction est apparu en premier, une sorte de problème de l'œuf et de la poule à l'échelle des domaines protéiques. Les protéines ne réalisent en effet leurs fonctions qu'une fois repliées, mais le repliement en lui-même ne procure pas d'avantages sélectifs et n'a donc pas pu être sélectionné [238]; or atteindre un niveau de repliement suffisant pour être en capacité de réaliser une fonction aurait nécessité des étapes intermédiaires qui, si elles n'ont pas pu être sélectionnées, ont très peu de chance d'avoir pu se succéder sans avoir disparu avant. Dans tous les cas, ces polypeptides ont commencé à évoluer à partir du moment où ils accomplissaient une fonction et l'information pour les coder pouvait être transmise d'une génération à une autre. Ils ont conduit dans un premier temps à des protéines enzymatiques dont les repliements étaient probablement en capacité de réaliser plusieurs fonctions [223]. Ensuite, les mutations ponctuelles conduisant à une augmentation de leur activité si elles avaient lieu dans le site catalytique, soit à une augmentation de la stabilité du repliement si elles avaient lieu en dehors du site catalytique, ont été sélectionnées [238]. Elles se seraient par ce processus spécialisées fonctionnellement, en lien avec l'émergence du design modulaire des protéines [223]. En outre, La capacité des protéines primordiales à se lier à des substrats et l'impact de cette liaison sur leur repliement pourrait également avoir joué un rôle majeur dans l'apparition de folds stables à l'échelle des temps évolutifs, mais également dans l'apparition de nouveaux folds dans le cas d'appariements accidentels aboutissant à une nouvelle structure stable [212].

IV.1.c. importance relative des phénomènes de convergence et de divergence

L'évolution des folds s'est faite à la fois par des phénomène de divergence et de convergence, dont il est complexe d'évaluer l'importance respective [255], [277]. Si les repliements n'avaient évolués que par divergence, ils auraient tous une origine commune et auraient subi une forte pression de sélection diversifiante [172]. Celle-ci aurait conduit à une accumulation de mutations dans certains domaines au point de causer des erreurs de repliements et *in fine*, à leur élimination [247]. Les

fonctions perdues par ces éliminations auraient été retrouvées par convergence évolutive à partir de structures simples et symétriques [255]. Les patterns de similarité structurale observés dans les protéomes s'expliquent mieux par des modèles qui placent la divergence comme type d'évolution le plus fréquent [278]. Dans les modèles plaçant la convergence comme processus majeur, un certain nombre de folds seraient apparus plusieurs fois au cours de l'évolution. Ils auraient ensuite divergés et se seraient diversifiés, explorant ainsi plus rapidement l'espace des fonctions que dans le cas d'une origine commune [172]. L'existence de « super-barrières » dans l'espace des structures semble en accord avec ce scénario : il existe des folds associés à des réseaux de séquences discontinus, fragmentés par les super-barrières. Les séquences de part et d'autre de ces barrières ne peuvent pas être rejointes par des mutations ponctuelles, témoignant de leur polyphylie [202], [248]. Cependant, seulement 10% des réactions biochimiques impliquent des enzymes réalisant des fonctions similaires sans origine évolutive commune (enzymes analogues) ; dans la très grande majorité des cas, deux enzymes réalisant la même fonction avec le même fold sont homologues, même si leur similarité de séquence est faible [235], [279], [280], [281]. Les phénomènes de convergence sembleraient donc avoir eu une importance limitée dans leur évolution. Ils ont également concerné certains motifs structuraux plus que d'autres. C'est par exemple le cas des motifs Rossmann-like (composants essentiels du Rossmann fold) ou β-hairpin-like, mais aussi des TIM-barrel, des α/β hydrolases ou des HINT folds qui sont observés dans des enzymes non homologues mais catalysant les mêmes réactions avec les mêmes ligands [241], [242], [280], [282], [283].

IV.1.d. lien entre modèle de loi puissance de la distribution des folds dans les protéomes et modes d'évolution des folds

Le modèle de loi puissance de la distribution des folds dans les protéomes résulte directement de leur mode d'évolution.

Les systèmes naissance/mort/innovation (BDIM) peuvent aboutir à une distribution suivant la loi puissance lorsqu'ils évoluent de façon neutre, c'est-à-dire indépendamment des propriétés fonctionnelles et structurales des folds [251], [255], [256], [284]. Le principe est simplement que, pour un ensemble de domaine protéiques ayant une probabilité égale de se dupliquer, les duplications vont avoir une probabilité accrue d'avoir lieu sur des domaines qui ont déjà été dupliqués. Cette propriété est nommée « attachement préférentiel ». Elle implique que les plus grandes familles de domaines sont dans l'ensemble anciennes, à quelques exceptions près (notamment les protéines ribosomales) [256], [259]. Les domaines de ces familles adoptent généralement des superfolds, qui sont les folds les plus anciens et les plus universels [251], [254]. Les domaines qui au contraire ont eu de faibles taux de duplications ou ne ce sont pas (encore) dupliqués adoptent plutôt des mesofolds, qui correspondraient à un âge intermédiaire, ou des unifolds, qui sont les plus récents et n'ont pas eu le temps de se diversifier dans l'espace des séquences [251].

L'évolution des folds n'a cependant pas pu reposer uniquement sur de la dérive génétique et a impliqué de la sélection. Il est d'ailleurs probable que ce soit avant tout des contraintes fonctionnelles, plus que stochastiques, qui aient été à l'origine de la loi puissance [251]. La fonction qu'ils réalisent est en effet sous pression de sélection. Dans certains cas, les domaines associés à des fonctions qui subissent une pression de sélection positive ont une probabilité plus faible de disparaître et plus forte de se dupliquer [251], [255]. Du point de vue fonctionnel, les unifolds pourraient s'être maintenus au cours de l'évolution car ils répondaient à des besoins précis, remplissant une fonction isolée dans l'espace des fonctions, sans possibilité d'en réaliser d'autres par des changements mineurs de séquence. Les mesofolds auraient au contraire subi des changements de fonctions via divergence au cours de l'évolution, par des modifications mineures de structure primaire [251]. L'expansion de certaines familles de domaines adoptant des superfolds peut aussi être, par des effets indirects, à

l'origine de l'accès à de nouvelles niches fonctionnelles, et donc être sélectionnée positivement en tant que phénomène favorisant la diversification fonctionnelle [243], [247]. Plus généralement, les superfolds « tolèrent » une variabilité de séquence beaucoup plus élevée que les autres folds ; cette tolérance leur permet d'accéder à une vaste gamme de fonction, en particulier quand certaines boucles se trouvent dans les régions à fortes variabilité de séquence. Les domaines associés à des fonctions dont la réalisation implique le maintien d'une identité de séquence forte sont au contraire rapidement contre sélectionnés au-delà d'un faible seuil de divergence, empêchant la diversification de leurs familles. Les contraintes fonctionnelles ont donc joué une grande importance dans la taille des familles de protéines observées aujourd'hui [247]. Cependant, la pression de sélection sur la fonction n'est probablement pas la seule qui se soit exercée sur les folds [255]. Les propriétés thermodynamiques et la cinétique du repliement entrent également probablement en jeu [249], [251], [255], [259]. En effet, les folds les plus fréquents ont souvent des structures particulièrement stables du point de vue thermodynamique, alors que les folds moins stables ne peuvent être adopté que par un nombre limité de séquences [249], [285]. C'est donc probablement l'interaction de ces processus (duplication plus ou moins neutre, sélection sur la combinaison fonction-structure) qui explique la permanence et l'universalité des lois puissances observées. C'est aussi cela qui explique la variabilité taxonomique de cette loi : chaque lignée a son propre BDIM et ses propres évènements aléatoires d'apparition, duplication et disparition ; de même, les pressions de sélections sur les fonctions sont différentes d'une lignée à une autre, favorisant certains folds par rapport à d'autres. Bien que la stabilité du fold ainsi que sa cinétique de repliement soient des propriétés qui lui sont intrinsèques, l'environnement cellulaire, qui peut participer à le stabiliser ou le déstabiliser, peut varier entre espèces et donc moduler la pression de sélection sur la structure elle-même.

Des analogies ont été proposées entre langage ou idéogrammes et protéines [276]. L'une des observations à leur origine est le fait que les mots dans le langage suivent la loi de Zipf, qui est une catégorie de loi puissance qui peut également s'appliquer aux folds [223]. Les protéines multidomaines, mais également le structurome d'un individu donné, peuvent être considérés dans une certaine mesure comme une phrase, dans laquelle chaque mot a un rôle précis. Les folds les plus fréquents sont analogues en terme d'utilisation à certains mots courts comme les déterminants ou les conjonctions de coordination, ils sont indispensable à la construction de la phrase mais ne lui donnent pas de spécificité (= de fonction). Les autres folds communs correspondent au vocabulaire utilisé pour la vie courante (= impliqués dans des fonctions de ménage) Enfin les folds rares, dont il existe un très grand nombre, peuvent être comparés au vocabulaire spécialisé : extrêmement riche mais dont seulement une fraction est utilisée dans un domaine précis (assimilable aux fonctions propres à certains taxons ou des espèces non apparentés mais occupant des niches écologiques similaires). Les limites de cette analogie sont le fait que tous les protéomes sont riches en protéines n'ayant qu'un domaine, alors qu'une phrase est par définition constituée de plusieurs mots. De manière plus fondamentale, l'évolution des protéines s'est déroulée sur un temps infiniment plus long que celui des langages, et a été soumise à des effets stochastiques beaucoup plus forts [276]. Il est également difficile de catégoriser les fonctions réalisées par chaque fold, contrairement à la fonction de chaque mot dans une phrase.

IV.2. histoire évolutive des folds

Cette section sera dédiée à l'histoire évolutive des folds, et comment celle-ci a façonné les différences d'usage des folds entre protéomes observés dans la partie précédente.

L'évolution des folds s'est déroulée conjointement à l'évolution du vivant et est caractérisée par plusieurs évènements majeurs. Elle a démarée par l'apparition des domaines protéiques en tant qu'unités modulaires, qui proviennent probablement à l'origine d'un vaste répertoire primordial de polypeptides courts et diversifiés. Leurs structures avaient probablement toutes des d'énergies libres similaires, leur permettant de coexister sans compétition des uns par rapport aux autres et de changer principalement par des effets stochastiques. Cette théorie est la « random origin hypothesis » [223], [286], [287], [288], [289], et elle correspond à la première phase de l'évolution des folds, qui est appelée « communal world ». Les plus anciens folds, au nombre d'une vingtaine, seraient apparus au cours de cette période. Parmis eux se trouvaient probablement des folds comme l'ABC transporter ATPase domain like (1.10.8.280) ainsi que plusieurs α/β/α three-layered sandwiches (structure Rossmann like) et des barrels ayant évolué à partir de domaines primordiaux capables de se lier avec des nucléotides [223], [242], [243], [265], [290], [291], [292]. Les Classes mostly Alpha, mostly Beta et Alpha Beta (contenant les folds α/β et α+β; dans les folds α+β, les hélices α sont regroupées dans une région de la structure et les feuillets β dans une autre, contrairement aux folds α/β) sont donc probablement toutes apparues lors du « communal world », mais pas en même temps. Les folds α/β sont probablement les plus anciens, en partie parce qu'ils sont actuellement majoritaires dans les génomes [223], [229], [257], [265], [293]. Une partie de ces structures auraient subi des réorganisations menant à une ségrégation des hélices et feuillets et conduisant à des structure α+β, suivie par une spécialisation par séparation pour les mostly α et mostly β [223], [265]. Les β-barrels et helical bundles sont probablement les premières structures mostly α et mostly β apparues au cours de ce processus. L'apparition des domaines p-loop et winged constituent des vagues majeures d'innovation [292]. Par la suite, une tendance à la transition de structures α vers β est observée (par exemple la conversion d'une hélice en un 3-stranded-meander) [223]. Malgré cela, le positionnement des folds α/β dans l'espace des structures et leur placement phylogénétique est parfois moins profond que celui des autres folds [232]. Cela résulte probablement du fait qu'ils soient apparus plus de fois de manière parallèle au cours de l'évolution que les autres classes [232]. Dans l'ensemble, il est probable que LUCA disposait déjà d'un protéome rassemblant une part importante d'architecture existant encore aujourd'hui [265].

Après cette première phase, la divergence des principaux domaines du vivant a été marquée dans un premier temps par une perte de folds dans chaque domaines, suivi d'une phase d'innovation domaine-spécifiques. Ce processus est appelé la « looser trend » [265]. En divergeant, les Virus auraient perdu des folds comme l'Acetyl-CoA synthetase like et le Thiolase related [291]. En-dehors des Virus, les folds apparus au cours de cette période sont globalement impliqués dans des core metabolic processes [291]. Plusieurs scenarii sont ensuite possibles, en ne se basant que sur les informations fournies par les foldomes: le premier avance une origine mixte de l'origine des Eucaryotes provenant à la fois des Archées et des Bactéries, supportée par le fait que les folds Eucaryotes proviennent en proportions égales de folds d'Archées et de Bactéries [20]. Le second se déroule en plusieurs phases, avec une divergence d'abord des Archées, suivi par les Bactéries et enfin, les Eucaryotes. La divergence des Archées aurait eu lieu il y a 2.9 milliards d'années [290]. Enfin, une divergence primordiale entre Eucaryotes et Procaryote est également possible [223] (Figure 29 A,B). Dans l'hypothèse d'une divergence plus récente des Eucaryotes et des Archées par rapport aux Bactéries, le pourcentage plus élevé de folds partagés exclusivement entre Eucaryotes et Bactéries qu'entre Eucaryotes et Archées pourrait s'expliquer par le degré de spécialisation très élevé des Archées qui ont probablement perdu

une grande partie des gènes et des folds qu'elles avaient en commun avec les Eucaryotes. La fonction des folds partagés entre ces deux domaines du vivant va également dans le sens d'un dernier ancêtre commun plus récent entre Eucaryotes et Archées qu'entre Eucaryotes et Bactéries. En effet, les folds partagés exclusivement entre Eucaryotes et Bactéries ont des fonctions d'information liées à la traduction, donc très anciennes et primordiales (Figure 26). Il est possible que les Bactéries aient ensuite acquis leurs propres folds pour les fonctions d'informations liées à la régulation des différents processus biologiques, de la même façon que l'ancêtre commun des Eucaryotes et des Archées, encore existant au moment de la divergence des Bactéries, a acquis les siens. Les Eucaryotes et Archées auraient ensuite divergé mais auraient gardé une partie de ces folds en commun, ce qui est observé actuellement [20]. Les génomes des Archées auraient dans tous les cas subi une forte évolution réductrice, leur faisant perdre de nombreux folds (par exemple le Sigma2 domain of RNA polymerase sigma factor), suivie d'une phase de spécialisation [265], [291]. À la même période, au moins 300 nouveaux folds impliqués dans le métabolisme serait apparus dans les autres groupes.

La phase suivante correspondrait à la diversification des Bactéries, avec innovations de nombreux folds spécifiques à ce domaines, à commencer par le TilS substrate-binding domain, et dont certains sont associés à une pathogénicité [265], [290], [291]. Il est probable que la coévolution des structures entre hôtes et pathogènes ait joué un rôle non négligeable dans l'évolution des folds [294]. Le taux élevé d'innovation en fold dans chez les Bactéries à cette période pourrait être la conséquence d'une compétition intense avec les Eucaryotes, les deux ayant à l'époque des écologies proches [265].

La cinquième phase est celle de la diversification des Eucaryotes, Virus et Archées [265], [290], [291]. La sixième phase correspond à la poursuite de l'innovation chez les Eucaryotes, en deux vagues. La première correspond à l'apparition de la pluricellularité, avec des folds impliqués dans la régulation, la signalisation et les processus intracellulaires) ; en parallèle, les Eucaryotes parasites perdent un grand nombre de folds [265]. La deuxième vague correspond à la divergence des Métazoaires, associées à des vagues de duplications massives et des nouveaux repliements impliqués dans l'adhésion cellulaire et la signalisation [264], [265], [290], [291]. Les évènements de recombinaisons de domaines ont été tellement importants dans l'histoire évolutive de ce groupe que plus de 80% de leurs protéines en résultent, contre seulement 2/3 chez les unicellulaires [214]. Les Streptophytes pluricellulaires ont également subi de nombreux évènements d'apparitions et de pertes d'architectures spécifiques liés à leurs changements de ploïdie nucléaires [263]. Ces évènements sont cependant relativement minoritaires, les taux de duplications étant jusqu'à 20 fois supérieurs aux taux d'innovations chez les Streptophytes pluricellulaires et les animaux, alors qu'ils sont quasiment égaux chez les Procaryotes [256].

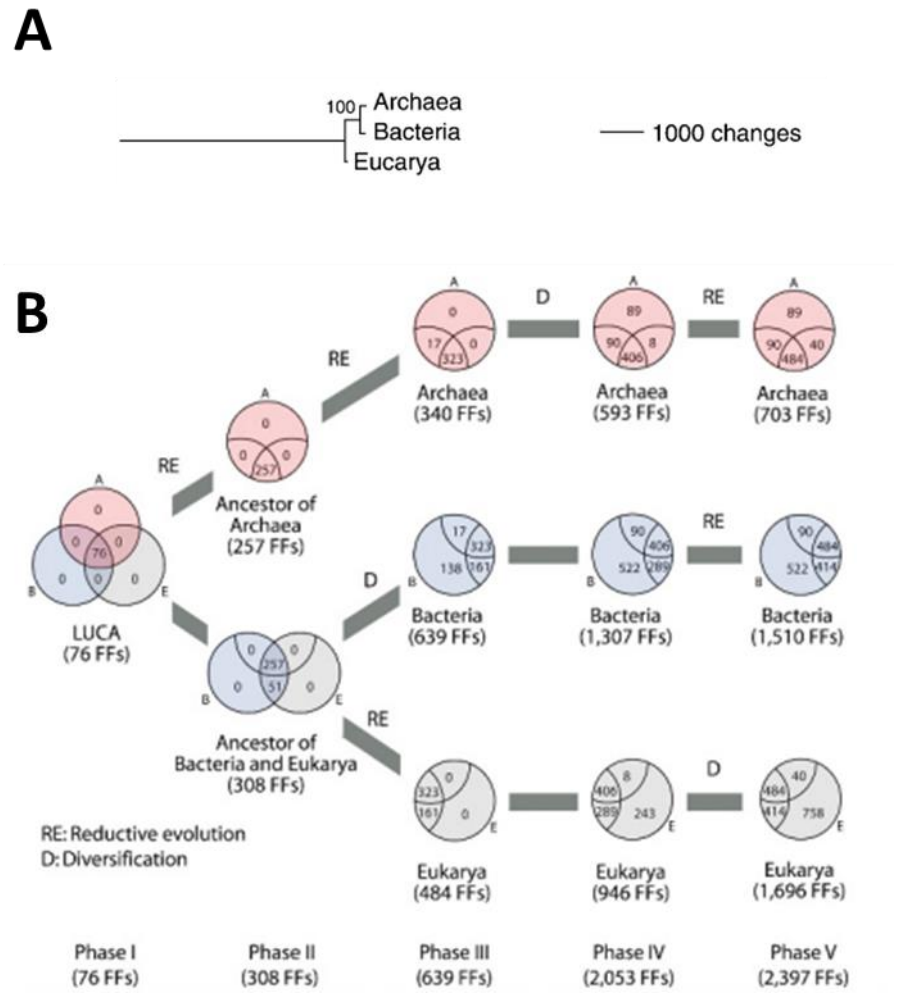


Figure 29. Scenarii possibles d'émergence des trois domaines du vivant basés sur les folds. (A) Arbre reconstruit à partir des occurrences de folds, avec 300 caractères informatifs et 20 états possibles. Extrait de Caetano-Anollés & Caetano-Anollés [223]. **(B)** Scénario évolutif des répertoires de fold dans les trois domaines du vivant à partir de LUCA. Les répertoires sont représentés par les diagrammes de Venn. Les cinq phases principales de leur histoire évolutive sont indiquées en bas. Les phases d'Évolution Réductrice (RE) ou de Diversification (D) sont indiquées au-dessus des flèches. Le modèle a été assemblé à partir de données phylogénomiques. Extrait de Kim & Caetano-Anollés [290].

L'histoire évolutive des folds a aussi été impactée par le milieu dans lequel elle s'est déroulée, plus spécifiquement par la variation de la disponibilité en cofacteur métalliques, les plus importants étant le fer (Fe), le cuivre (Cu), le manganèse (Mn) et le zinc (Zn). Les folds associés au Fe et au Mn sont ainsi apparus plus tôt que ceux associés aux deux autres métaux, le Fe et le Mn ayant été biodisponibles plus tôt au cours des temps géologiques. Plus précisément, le fold associé au Mn est le nucleotide-diphospho-sugar transferase, celui associé à un hème Fe globin-like/ α -helical ferredoxin, celui associé au Zn le zincin like et celui associé au Cu le cupredoxin like [295]. Le Rossmann fold, qui utilise des clusters souffre – Fe, a été adopté par des protéines réalisant du transport d'électron alors que l'atmosphère était encore anoxique [296]. À noter que les virus ont eu un rôle très important dans l'évolution des folds au cours des temps évolutifs, en particulier via les HGT et dans les protéomes des Procaryotes [256], [291].

V. variabilité des séquences de domaines protéiques en fonction des conditions environnementales

Dans ce dernier sous-chapitre, je décrirai les conséquences des pressions de sélection abiotiques sur les structures de domaines protéiques. La première section visera à préciser les échelles auxquelles les principaux processus adaptatifs des structures ont lieu, à savoir les structures primaires et secondaires ; la seconde et la troisième section seront consacrées aux réponses face aux stress générés par les augmentations de viscosité cellulaire et les milieux psychrophiles (qui ont plutôt tendance à « rigidifier » le milieu cellulaire, par opposition à une augmentation de viscosité), respectivement ; enfin, la dernière section aura pour objectifs de présenter certaines observations dans les structures de protéines planctoniques et de montrer à quel point celles-ci sont exposées aux pressions de sélection abiotiques.

V.1. échelle de la variabilité et tendances générales en réponse à un stress environnemental

Les folds subissent une pression de sélection sur la réalisation de leur fonction et leurs propriétés thermodynamiques. Les mutations sur la séquence génomique résultant en des substitutions non synonymes peuvent être contre-sélectionnées si elles ont lieu à des positions clés du repliement ou de la réalisation de la fonction. En particulier, la flexibilité à ces positions doit rester élevée mais peut être atteinte avec différents types d'acides aminés en fonction du contexte moléculaire et environnemental [172]. Il en résulte qu'une partie de la variabilité interspécifique des fréquences en acides aminés dans les protéomes peut être imputée à la niche écologique des espèces considérées, en particulier chez les Procaryotes chez qui les substitutions et les fréquences en acides aminés sont liées à la température optimale de croissance de l'organisme [297], [298], [299]. Ce n'est pas le cas chez les Eucaryotes chez qui ces fréquences sont impactées par la dérive plus que par la sélection par l'environnement [300]. À l'échelle des espèces, des biais de résidus espèce-dépendants existent dans tout l'arbre du vivant [297].

Les modifications de pression, d'acidité, de salinité et de température ont des impacts variables sur les protéomes en fonction de la tolérance de la cellule au stress abiotique. Elles peuvent changer la viscosité du milieux intracellulaire et donc avoir un impact sur le repliement des domaines [301]. Les protéines globulaires des espèces vivant dans des milieux extrêmes conservent généralement le même squelette structural que leurs homologues mésophiles, mais les propriétés cinétiques et thermodynamiques de leur repliement, et plus généralement, leur flexibilité à une température donnée sont différentes [302], [303]. Suite à une exposition prolongée sur plusieurs génération à une nouvelle température générant un stress thermique, les mutations non synonymes ayant pour conséquence un déplacement de la température optimale des repliement (via un changement de leur cinétique et de la flexibilité de l'état replié) sont souvent sélectionnées si elles aboutissent *in fine* à une température optimale proche de celle responsable du stress thermique [304], [305], [306]. Les nouvelles structures atteignent alors à cette température des flexibilités comparable à celles atteintes à l'ancienne température optimale, de sorte à ce que les taux catalytiques et de liaison au ligand, qui sont sous forte pression de sélection, soient les mêmes [305]. Ce processus adaptatif est appelé «hypothèse des corresponding states » [307]. Le taux d'adaptation à la température est cependant variable à l'échelle d'un protéome car les toutes les fonctions et tous les repliements ne sont pas impactés de la même manière par le stress thermique [301]. Certaines fonctions ne s'expriment en effet que dans lorsque ce type de stress est détecté, et sont parfois réalisées par des protéines qui ont une fonction différente dans d'autres contextes. Il s'agit d'un phénomène de plasticité protéomique,

par lequel certaines protéines sont capables de changer de folds pour réaliser de nouvelles fonctions [22], [194]. Dans les protéines enzymatiques, la pression de la sélection sur la réalisation de la fonction conduit à une conservation stricte du site actif même dans une situation d'adaptation à la température. Des modifications globales risqueraient d'entraîner une perte de la fonction et donc une contre-sélection. Ce sont les autres régions du repliement qui subissent des modifications, qui suffisent à changer les propriétés cinétiques de la structure globale [304]. En outre, la flexibilité d'un site actif peut également être modifiée sans avoir d'impact sur la stabilité de la structure globale [308]. Les enzymes subissent en général une pression particulièrement forte, étant les protéines ayant le plus besoin de flexibilité lors de l'accomplissement de leur fonction, et encore plus pour celles impliquées dans l'apoptose, la défense antioxydante et certaines voies métaboliques [309]. L'augmentation de la température du milieu peut être responsable d'une accélération du métabolisme et donc de la production de radicaux libres nocifs, deux facteurs pouvant entraîner une perte de contrôle des cellules résultant en l'activation de mécanismes d'apoptose [310]. Les protéines intervenant dans la transcription et la traduction ne sont au contraire que peu impactées par le stress thermique du point de vue structural [301]. Deux hypothèses (ne s'excluant pas l'une et l'autres) sont avancées pour justifier ces distinctions : la présence de chaperonnes qui pourraient limiter le stress thermique de façon particulièrement importante pour cette fraction des protéomes, ou un lien avec le métabolome [301]. Ce dernier peut en effet jouer un rôle dans la modulation du stress thermique, et les enzymes métaboliques sont celles qui présentent le plus d'adaptation à la température [301], [309]. Plus globalement, les processus d'adaptation des protéines à la température sont complexes. Il n'existe probablement pas un processus unique et universel, ou qui réponde à des règles simples ; les effets de la sélection conduisent plutôt à des réponses anecdotiques et contingentes aux circonstances.

V.2. effets des milieux augmentant la viscosité intracellulaire sur les repliements

Dans le cas d'un milieu où pression, acidité, salinité et/ou température sont élevé(s), la viscosité du milieu intracellulaire augmente. Cette augmentation a tendance à stabiliser les repliements, en diminuant l'intensité de l'agitation moléculaire ; dans ces conditions, les repliements trop rigides ont de plus fortes chances d'être contre sélectionnés [301], [311], [312]. Les protéines se repliant dans ces conditions sont donc généralement moins flexible (ou plus rigides) que les autres [22], [301], [311], [312], [313]. Des modifications sont observées dans leurs structures secondaires et primaires. Concernant les premières, le contenu et la longueur des hélice des protéines thermophiles a tendance à augmenter, contrairement aux feuillets et boucles [314]. Pour les structures primaires, les réponses sont surtout de type repliement-spécifiques et peuvent conduire à des observations opposées. Sur les sites exposés du repliement, les résidus chargés (autant positifs dans certains repliements que négatifs dans d'autres) ainsi que les résidus acides ont tendance à augmenter en fréquences et les résidus polaires (particulièrement la sérine) à être remplacés par des résidus rigides et non polaires [309], [311], [314], [315], [316]. Dans d'autres cas, la fréquence des résidus polaires (glutamine et thréonine en particulier) augmente , ainsi que celle de certains résidus neutres comme la glycine [297], [309], [316]. L'asparagine, polaire mais dotée d'un groupement amide pouvant être impliqué dans la formation de liaisons hydrogènes et de pont salins, augmente généralement en fréquence dans les repliements riches en ce type de liaison mais diminue dans les autres [309], [314], [316]. Une augmentation de la fréquence de résidus aromatiques volumineux est parfois observée [314], sans être une tendance générale [309]. Dans les sites enfouis, des substitutions de résidus hydrophobes sont observées, et peuvent résulter en une diminution de l'entropie conformationelle des chaînes latérales [314].

V.3. effets des milieux froids sur les repliements

Les adaptations au froid ont été beaucoup plus étudiées que celles aux chaleurs extrêmes. La pression de sélection imposée par le froid sur les repliements survient en effet à un delta de température par rapport à celle des milieux tempérés beaucoup plus faible que la chaleur. Beaucoup d'espèces vivent donc dans des milieux où le froid exerce une pression forte sur les repliements, alors que la chaleur n'exerce une telle pression quasiment que chez les Archées extrêmophiles [312]. Le froid, de la même façon que les températures élevées, peut provoquer une dénaturation des protéines appelée « cold denaturation » [317]. Certaines protéines constituent des cibles de choix dans l'adaptation au froid : c'est par exemple le cas de la chaperonine cytosolique (CCT), un complexe protéique présentant des traces d'adaptations structurales dans ce milieu et impliqué dans le repliement d'environ 10% des protéines ainsi que dans la réponse au stress causé par le froid [312]. Le fait que cette protéine soit capable de réaliser sa fonction aussi vite et bien en milieu froid qu'ailleurs lui permet de compenser les potentielles erreurs de repliements d'autres protéines qui, protégées par le CCT, subissent moins de pression de sélection et ne présentent donc pas de traces d'adaptations au froid. De manière générale, la stabilité des protéines enzymatiques des milieux froids est plus faible que celle des milieux tempérées, bien que des exceptions existent [306], [308], [318]. Cette augmentation de flexibilité peut se faire via l'insertion de domaines entiers dans les régions en dehors du site actif chez les enzymes. Cela est par exemple observé de manière conservée dans la version psychrophile de la protéine PEPT1, dont la structure primaire présente une région polymorphique codée par un seul exon consistant en une répétition d'un même domaine de sept acides aminés [319]. Concernant les structures secondaires, il y a une faible tendance à la diminution de la proportion de feuillets β [320]. L'interprétation des modifications de structure primaire observées est plus complexe. L'adaptation des protéomes au froid ne semble en effet pas passer par des substitutions massives de certains acides aminés par d'autres. Ces évènements seraient plutôt ponctuels et repliement ou fonction-dépendants, et non des résultats de convergence de stratégies d'adaptation [312], [321]. Chaque repliement a ainsi au cours de son histoire évolutive eu sa propre façon d'augmenter sa flexibilité face à ce type de contrainte, avec des modifications à plusieurs échelles [306], [322]. Enfin, les substitutions des résidus n'ont pas le même impact selon leur position dans la structure primaire. Ils vont néanmoins souvent dans le sens d'une diminution de la polarité ou une d'augmentation de l'hydrophobicité [306], [312]. Dans les enzymes, ils ont souvent lieu à des positions importante pour la mobilité pendant la catalyse [323]. Les prolines de certaines protéines psychrophiles sont moins fréquentes dans les boucles et dans certaines structures secondaires localisées, mais augmentent dans les hélices internes pas ailleurs [306], [318], [322]. La fréquence des valines tend à diminuer [320]. Selon les protéines, des asparagines peuvent être remplacées par des acides aspartiques [142] ou au contraire voir leur fréquence augmenter [320], dans les deux cas à des sites exposés. Toutes ces substitutions peuvent avoir comme conséquence un excès de charge négative et une diminution du nombre de liaisons par ponts salins à la surface de la structure, une diminution du nombre de liaisons faibles et donc des interactions entre domaines ou régions du repliement, une diminution de la taille, de l'hydrophobicité relative et du niveau d'enfouissement des clusters de résidus apolaires (notamment les glycines) [297], [306], [322], [324], [325], [326]. Elle peuvent également faciliter les interactions avec le solvant, diminuant ainsi les contraintes structurales subies par la structure et augmentant de sa flexibilité [324]. Dans les structures secondaires, les arginines et glutamines des hélices α ont tendance à être remplacées par des lysines et des alanines, ce qui n'est pas le cas dans les feuillets [320]. La flexibilité en moyenne plus élevée des repliements psychrophiles est associée à une plus grande diversité d'états non repliés et des conformères avec une plus haute énergie libre globale séparés par des barrières énergétiques basses [327]. Ils subissent en général une forte pression de sélection vers des modes de repliements laissant moins d'opportunités à des erreures, par exemple

des repliements coopératifs sans intermédiaires pour les protéines multidomaines [327]. Les repliements coopératifs permettent de limiter les pertes énergétiques pouvant découler d'un processus segmenté en limitant la probabilité d'atteindre des conformations dont les repliements ne sont pas optimisés, comme c'est le cas des structures mésophiles qui ont plusieurs minima locaux [327]. L'augmentation de leur flexibilité génère aussi une baisse d'affinité vis-à-vis de leur substrat, qui est compensée par une augmentation de l'efficacité catalytique [322], [325].

V.4. variabilité des protéines planctoniques

Les espèces du milieu marin, et plus spécifiquement celles appartenant au plancton, sont les plus à mêmes de présenter de telles adaptations. D'abord parce que la plupart des unicellulaires ne peuvent vivre que dans le milieu liquide, le milieu aérien étant déshydratant ; mais aussi parce que, pour les espèces qui peuvent se protéger de la déshydratation, le milieu aérien est beaucoup plus isolant que le milieu aquatique. Les êtres vivants qui y évoluent ne peuvent donc dans la majorité de cas réguler que modérément leur température interne. Les espèces planctoniques, transportées passivement dans une diversité importante de milieux, peuvent ainsi voir leur température interne varier, particulièrement lors du passage d'un biome à un autre. Ils peuvent aussi avoir à faire face à des modifications de pression et de composition du solvant, qui sont connus pour impacter les liaisons au sein des repliements [323], [328]. Il est donc probable que leurs protéomes aient subis une sélection particulièrement intense et disposent d'une plasticité phénotypique suffisante pour pouvoir maintenir leurs fonctions dans tous ces environnements. De manière globale, la température optimale d'une grand nombre d'enzymes planctoniques (comme les esterases par exemple) est corrélée à la température moyenne annuelle du milieu [329]. Les structures des protéines des milieux marins à forte variabilité thermique (delta de température de près de 15°C) sont aussi plus stables que celles des milieux à faible variabilité (delta de température compris entre 6 et 8°C) [329] (Figure 30).

Ces modifications de températures optimales reposent probablement entre autre sur des modifications de structure primaire des enzymes concernées. Chez *Bathycoccus prasinos*, les populations Arctique se distinguent des autres par des substitutions originales et conservées d'acides aminés (Figure 31 A). C'est particulièrement le cas pour l'EPSP synthase dans laquelle une asparagine est substituée par un acide aspartique à une position de surface. L'acide aspartique, acide aminé chargé négativement contrairement à l'asparagine, polaire mais neutre, y est responsable de l'apparition d'un nouveau cluster de charge négatives, qui, comme on l'a vu, peut participer à diminuer la rigidité d'une protéine [142]. Chez SAR11 (Bactérie du groupe des Pelagibacterales, caractérisée écologiquement par son abondance très élevée dans l'Océan), c'est la concentration en azote dans le milieu qui est à l'origine d'une pression de sélection différentielle sur la structure de la Glutamine Synthase (GS). Plus la concentration en azote est élevée et plus la pression de sélection sur la GS est relâchée, résultant en une augmentation du nombre de mutations non synonymes sur des sites importants du site actif qui peuvent modifier la distance au ligand et l'accessibilité relative au solvant. La diminution de la disponibilité en azote génère au contraire une augmentation de la sélection purifiante sur l'enzyme, et donc un déplacement des mutations non synonymes vers des sites sans impact sur la distance au ligand et l'accessibilité relative au solvant [330]. Ces mutations ont néanmoins le potentiel pour apporter un avantage sélectif notable à leur porteur si elles résultent en une augmentation des propriétés catalytiques de la GS en milieu pauvre en azote. Ce comportement pourrait être un trait commun à toutes les enzymes cores de SAR11 : celles subissant des mutations à proximités de leurs centres réactionnels auraient une probabilité beaucoup plus élevée d'être contre-sélectionnées dans les milieux où la pression de sélection sur SAR11 est forte [330]. En effet, les populations de SAR11 des courants froids ont tendance à avoir plus de variants protéiques que celles des courants chauds, oligotrophes, où la pression de sélection causée par le manque de disponibilité

en nutriment est plus forte [331] (Figure 31 B). Les variants sont plus fréquemment hydrophiles, mais certains acides aminés hydrophobes, jouant un rôle important dans la stabilité de la structure, sont également enrichis dans certains contextes, en réponse à des contraintes environnementales [331].

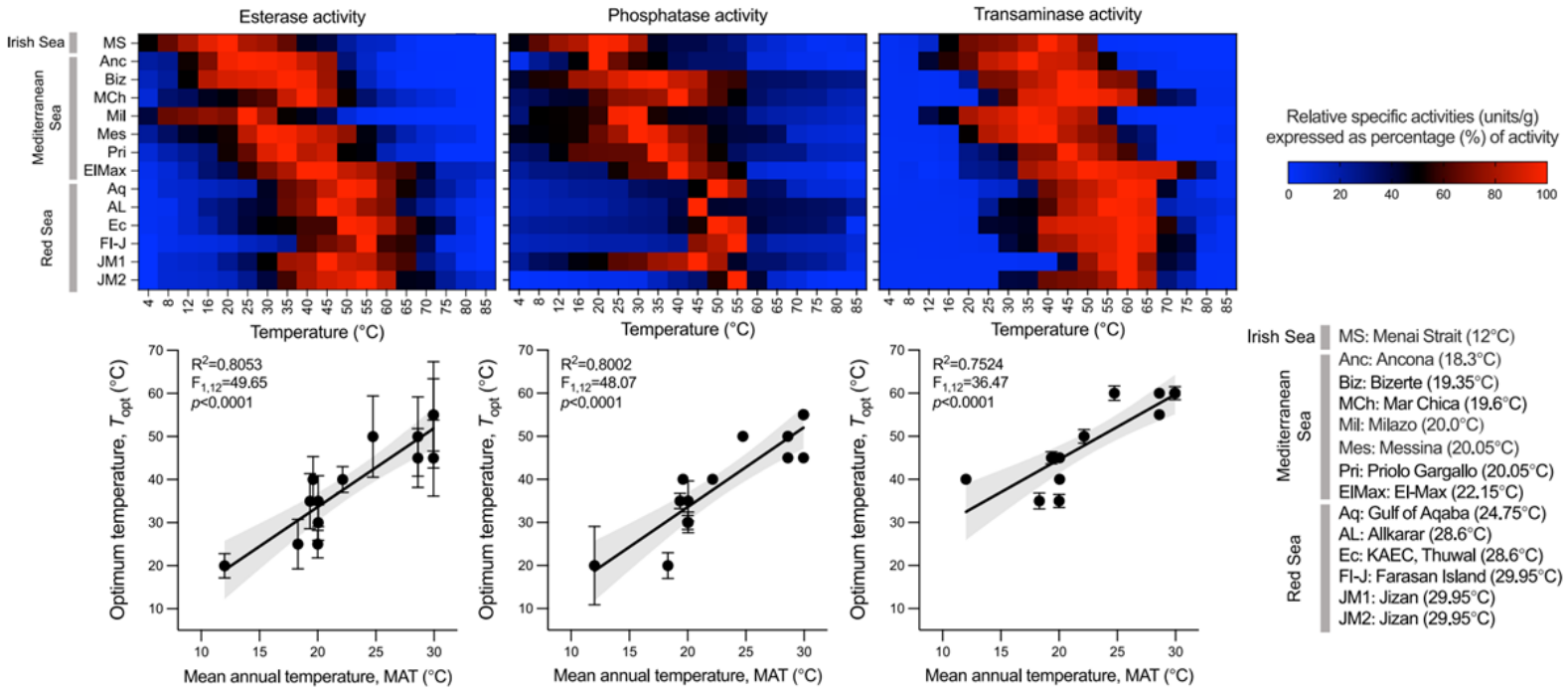


Figure 30. Lien entre température du milieu et température optimale de différentes familles d'enzymes. Adaptation thermique d'isozymes des communautés microbiennes de sédiments marins. En haut : profil thermique pour **(A)** les esterases **(B)** les phosphatases **(C)** les transaminases dans leurs formes actives, extraites à partir des sédiments de différents milieux allant de la mer d'Irlande à la mer Rouge en passant par la Mediterrannée. Les heatmaps représentent le pourcentage relatif d'activité spécifique à chaque température calculée comme le rapport entre le taux enzymatique initial (en unité.mg^{-1}) et l'activité maximale (en moyenne sur trois réplicats). Le gradient de couleur va du bleu foncé (aucune activité = 0%) au rouge clair (100% d'activité). En dessous de chaque heatmap, la relation entre la température du milieu d'échantillonnage (°C) et la température d'activité enzymatique maximale (T_{opt}) exprimée en moyenne ± SD (n = 3) est représentée. Des régressions linéaires simples sont affichées en lignes noires, les zones grises représentant l'intervalle de confiance à 95%. Extrait de Marasco *et al.* [329].

Dans l'ensemble, les folds tendent donc plutôt à être soit évolvables, c'est-à-dire qu'ils peuvent tolérer de nombreuses substitutions dans leur séquence primaire sans perdre leur repliement, soit « rigides » du point de vue évolutif, c'est-à-dire qu'un faible nombre de mutations sont suffisante pour leur faire perdre leur structure. Ces deux grandes catégories permettent de poser des hypothèses sur les forces évolutives agissant sur les folds dans un contexte environnemental donné et leurs conséquences sur la réalisation de leurs fonctions. Les propriétés d'évolvabilité des folds résultent à la fois de leur fonction et de leur histoire évolutive, des caractéristiques intrinsèques de l'assemblage des résidus, et des liaisons qu'ils forment et enfin, de leur repliement en lui-même. À noter que ces propriétés d'évolvabilité sont aussi en partie responsables de la taille des familles de domaines. Dans les folds évolvables, la sélection environnementale agi à l'échelle de la structure primaire, qui peut subir des subsitutions sans altérations de fonction. Les substitutions vers des résidus modifiant la flexibilité du fold (dans le sens d'une augmentation en milieu froid et d'une diminution en milieu chaud) sont sélectionnés en fonction du contexte. Au contraire, les folds évolutivement « stables » ou

« rigides » ne peuvent pas supporter une grande variabilité de séquence sans perdre leur fonction. Dans un contexte de pression de sélection abiotique, il est possible que la pression de la sélection sur la fonction des protéines dont les domaines adoptent de tels folds entraîne un changement de fold qui soit capable de réaliser ou de participer à la réalisation de la même fonction. Il n'existe cependant pas de littérature sur ce sujet dans le contexte océanique et pour les repliements observés dans les protéomes des microbiomes marins ; c'est dans cette perspective que cette thèse s'attachera à mettre en évidence, au moyen de l'analyse de la distribution des folds dans l'environnement, que cette stratégie adaptative existe aussi probablement.

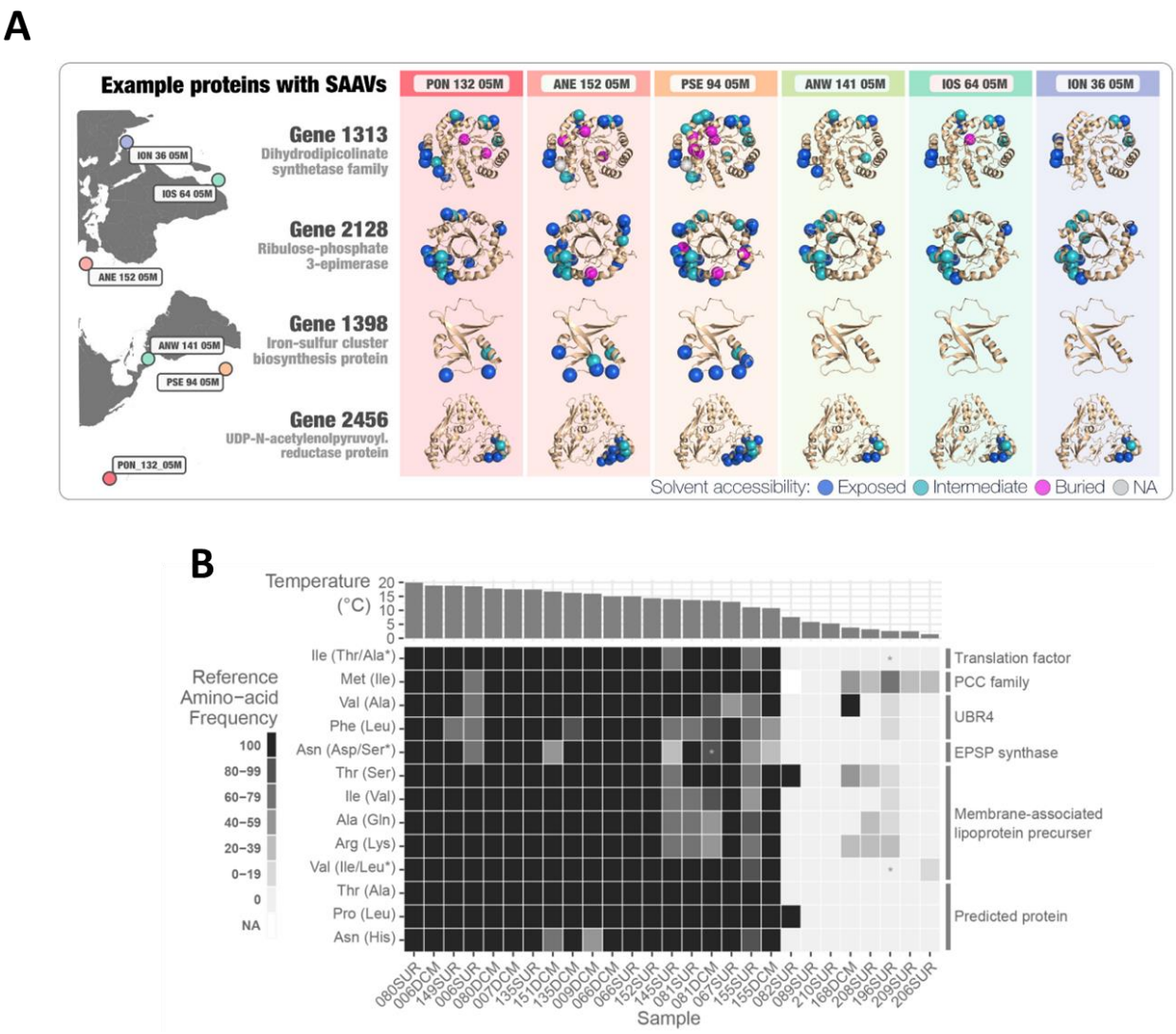


Figure 31. Lien entre fréquence des variants d'acide aminé et milieu de vie pour certaines protéines planctoniques. (A) Chez SAR11. Représentation de la localisation des variants d'acide aminé (SAAVs) sur les structures prédites de quatre protéines codées par des gènes du noyau du sous groupe Ia.3.V de SAR11 dans six métagénomes provenant de milieux distants géographiquement. Extrait de Delmont *et al.* [331]. **(B)** Chez *Bathycoccus prasinos*. La heatmap représente la fréquence d'acides aminés dans 27 échantillons différents à 13 positions génomiques qui séparent les populations en fonction de la température. Les acides aminés alternatifs sont indiqués entre parenthèses à gauche. Les échantillons sont ordonnés en fonction de la température qui est indiquée en haut. L'identité des six protéines contenant les 13 variants est précisée à droite. Extrait de Leconte *et al.* [142].

Les processus adaptatifs déployés par les espèces du plancton marin en réponse à la variabilité des pressions de sélections auxquelles elles doivent faire face lors de leur advection par les courants impliquent-ils les folds ? Dit autrement, la pression de sélection subie par le plancton marin, résulte-t-elle en des réponses à l'échelle du foldome, ou uniquement à l'échelle des séquences ? Est-il possible d'établir des relations entre l'échelle nanoscopique des folds et celle, planétaire, de la distribution des espèces planctoniques dans les océans ?

PARTIE 2. MATÉRIEL ET MÉTHODE

Afin d'étudier le plancton, de recenser sa diversité et mieux comprendre l'organisation de ses communautés, deux campagnes d'échantillonnage à l'échelle globale ont été menées entre 2009 et 2013 : *Tara* Oceans (2009-2012) et Polar Circle (2013).

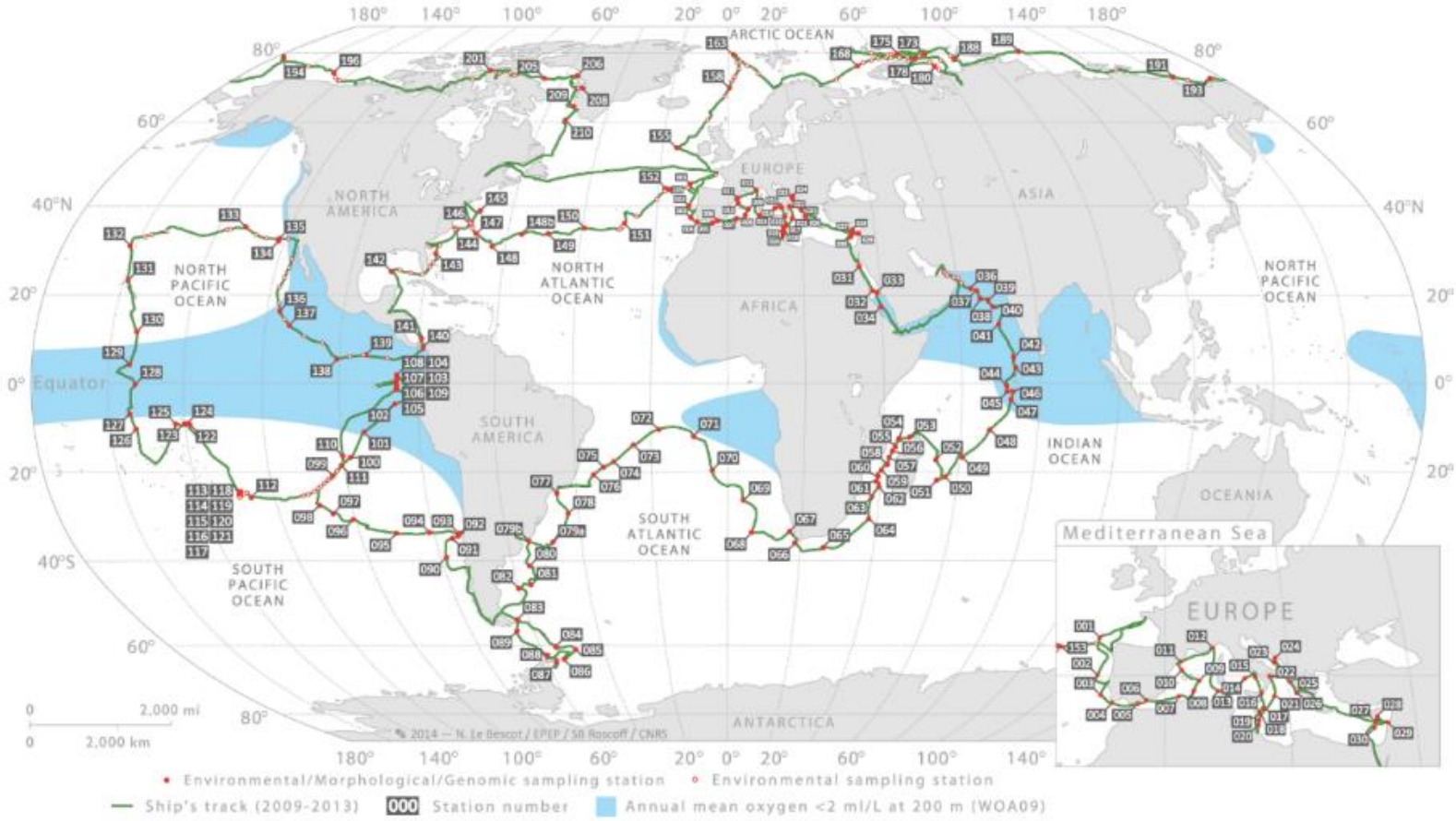


Figure 32. La goélette Tara et son trajet lors des expéditions *Tara* Oceans et Polar Circle. Le trajet est figuré en vert sur la carte du monde, avec les différentes stations d'échantillonnage et leurs numéros. Les aires bleutées correspondent aux régions où la moyenne annuelle de concentration en dioxygène est inférieure à 2mL/L (WOA09), et où les concentrations en CO_2 et le pH sont généralement respectivement hauts et bas. En haut : Photo de © Sacha Bollet. https://fondationtaraocean.org/kiosque/tara-oceans-chroniques-expedition-scientifique/. Carte extraite de Pesant *et al*. [9].

Au cours de ces deux expéditions, des échantillonnages de plancton ont été réalisés à plusieurs profondeurs dans la colonne d'eau à une échelle globale, donc dans de nombreux systèmes océanographiques et contextes environnementaux qui sont dans une certaine mesure représentatifs de l'Océan global. L'une des originalités des protocoles utilisés pour ces échantillonnages est l'utilisation de filtres de mailles de tailles différentes, donnant un aperçu complet de la diversité dans différentes fractions de tailles du pico- (<0.02µm) au mega- (m) plancton.

Une des façons d'analyser ces échantillons est par la reconstruction de Génomes Assemblés à partir de Métagénomes (MAGs). Comme leur nom l'indique, ces génomes proviennent de métagénomes environnementaux, c'est-à-dire de l'ensemble de l'ADN de tous les êtres vivants et des virus d'échantillons prélevés directement dans un milieu (océan, terre, flore intestinale, etc). Après avoir été extrait, cet ADN est séquencé. Dans le cas d'un séquençage par technologie Illumina ©, une grande quantité de lectures courtes (entre 50 et 300 paires de bases selon le modèle) est produite. L'analyse de ces lectures commence ensuite par un traitement par des algorithmes d'assemblage. Leur objectif est de reconstruire des séquences nucléotidiques les plus longues possibles en assemblant les lectures courtes entre elles, en utilisant les similarités de séquence entre leurs extrémités. Les séquences longues produites à l'issue de cette phase sont appelées contigs. Leurs extrémités n'ont plus aucune similarité de séquence avec celles de tous les autres contigs, ce qui empêche de les assembler. Plusieurs contigs peuvent cependant appartenir à la même espèce et représenter des fragments de son génome. L'étape suivante consiste donc à grouper les contigs ensemble pour recréer des génomes morcelés, chacun appartenant potentiellement à une espèce. Elle utilise en général deux métriques qui sont généralement espèce-spécifiques : la fréquence tétranucleotidique ou, pour des échantillons prélevés à plusieurs localisations différentes, la co-variation d'abondance. Les échantillons prélevés lors de l'expédition *Tara* Oceans ont permis de reconstruire les génomes environnementaux d'environ 700 Eucaryotes (Figure 33), 1900 Procaryotes et 31000 NCLDV [5], [332]. En utilisant l'information du nombre de lectures appartenant à chacun de ces MAGs, et le nombre total de lectures dans l'échantillon, il est possible d'obtenir une abondance relative, dite métagénomique, pour chacun de ces génomes. Ces données permettent d'analyser la distribution de différents groupes d'espèces planctoniques dans les stations *Tara* Oceans et Polar Circle.

À partir de génomes, qu'ils soient environnementaux ou non, plusieurs étapes sont ensuite possibles. La première est la prédiction de gènes. Une fois les gènes prédits, il est également possible de prédire des protéines et donc de reconstruire des protéomes. Toutes ces étapes sont beaucoup plus complexes chez les Eucaryotes en raison des particularités de leurs gènes, notamment l'existence d'introns. L'une des bases de données regroupant la plus grande diversité de protéomes Eucaryotes actuellement est EukProt [8] (Figure 34). À partir de protéomes, des structuromes peuvent être reconstruits par annotation structurale des protéines. C'est à cette étape que commence le matériel et méthode de cette thèse.

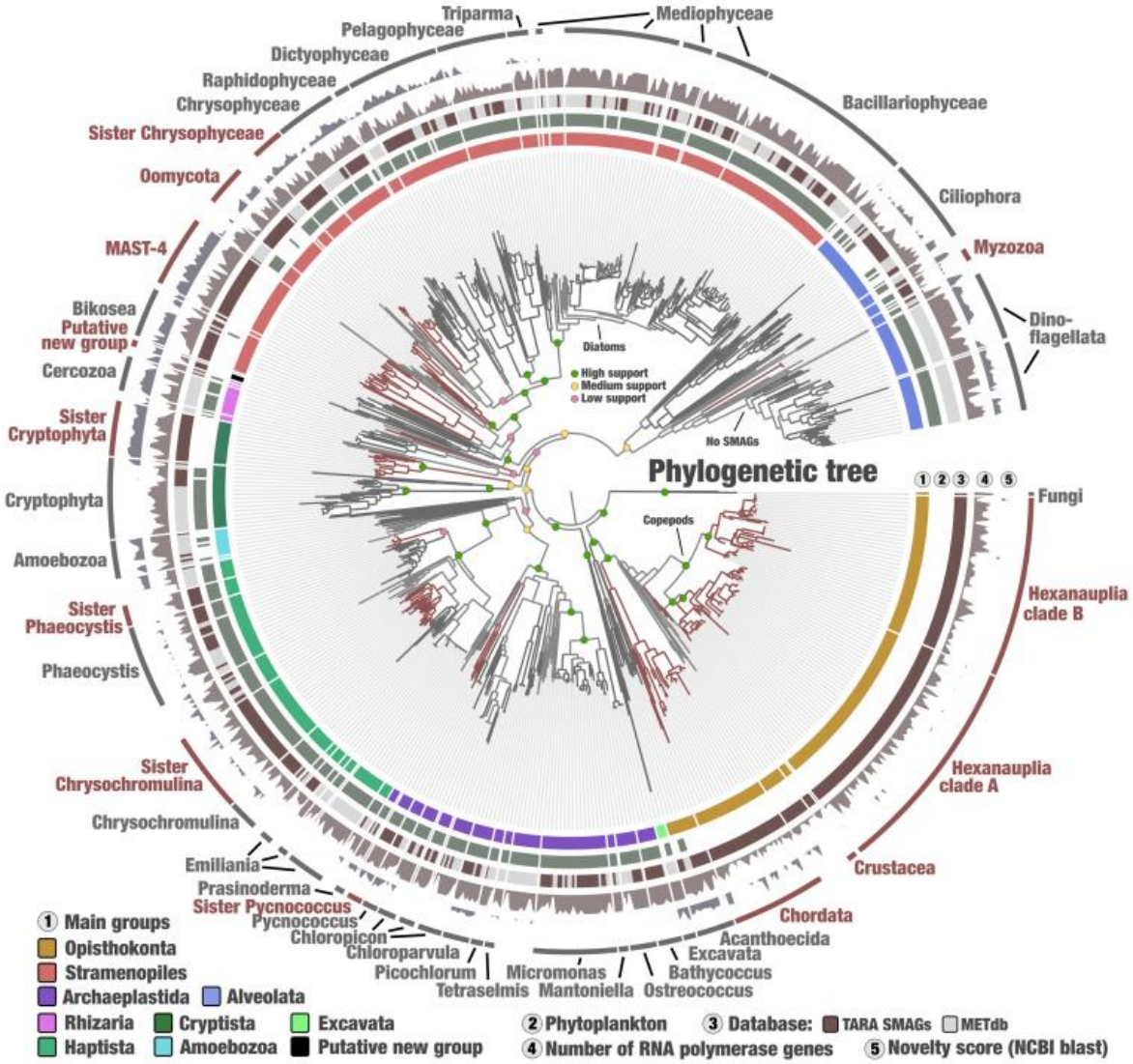


Figure 33. Diversité des MAGs Eucaryotes reconstruits à partir des échantillons de l'expédition *Tara* Oceans. Extrait de Delmont *et al*. [5].

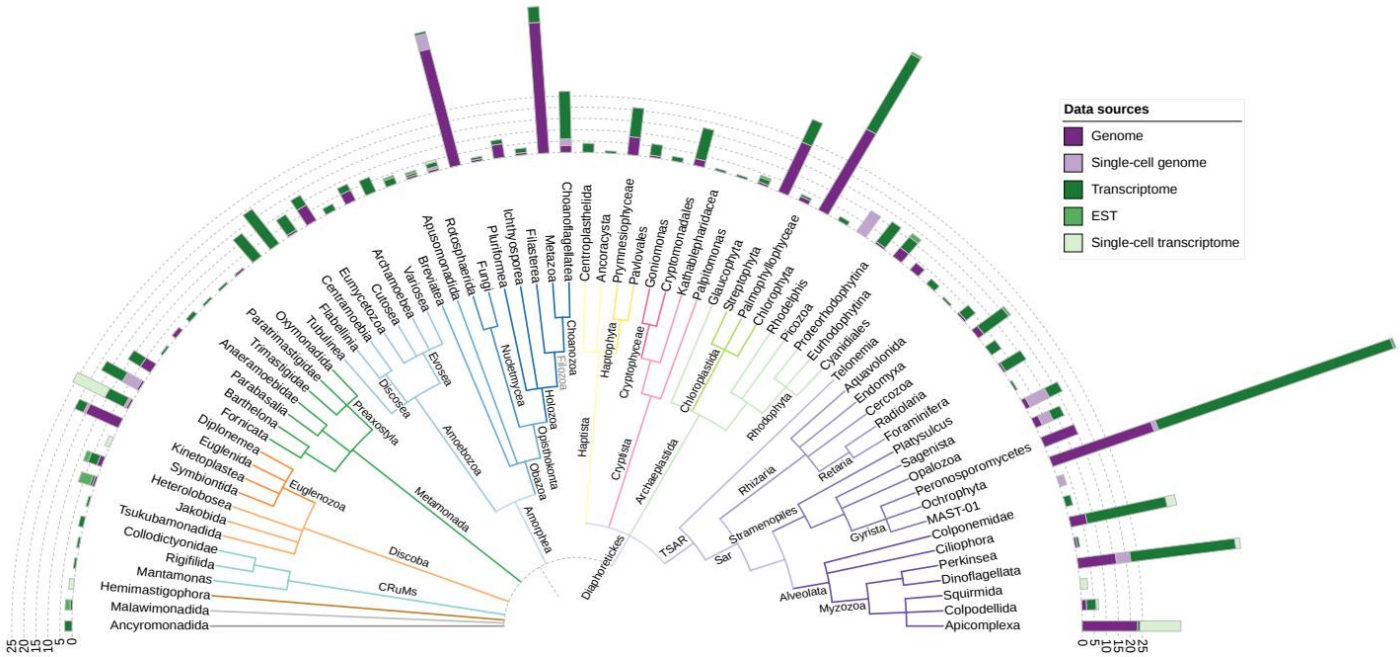


Figure 34. Diversité des Eucaryotes dans EukProt. Extrait de Richter *et al*. [8].

annotation structurale des protéomes

L'annotation structurale des protéomes a été réalisée avec CATH [2], [3] sur les protéomes des MAGs Eucaryotes, Procaryotes et de NCLDV provenant des échantillons TO (713 Eucaryotes, 1888 Procaryotes [5], 30802 NCLDV [332]) et sur les 990 protéomes de référence (RPs) d'Eukprot [8]. Les lignées d'Eucaryotes (MAGs et RPs), et de Bactéries utilisées pour l'étude sont :

- 9 chez les Bactéries (Actinobactéries, Bacteroidota, Bdellovibrionota, Chloroflexota, Cyanobacteriota, Marinisomatota, Myxococcota, Planctomycetota, Proteobacteria, Verrucomicrobiota)
- 14 chez les Eucaryotes, qui ont parfois dû être harmonisés entre les MAGs et les RPs basé sur les espèces dans chacune des bases de données : Choanoflagellés (11 MAGs « Choanozoa » et 27 RPs), Arthropodes (219 MAGs et 19 RPs), Chordés (44 MAGs et 22 RPs), Chlorophytes (64 MAGs et 80 RPs), Ciliophores (30 MAGs et 45 RPs), Haptophytes (92 MAGs et 27 RPs), Cryptophytes (8 MAGs et 19 RPs « Cryptophycés »), MAST-4 (19 MAGs et 4 RPs), Myzozoa (1 MAG MALV-I + 5 MAGs Myzozoa et 121 RPs), Bigyra (13 MAGs et 3 RPs « Bicosoecida »), Heterokontophytes (16 MAGs et 15 RPs « Core Oomycetes »), Bacillariophytes (52 MAGs et 83 RP « Diatomées ») et Ochrophytes (51 MAG et 52 RPs « Chrysophycés » « Eustigmatophycés » « Pelagophycés » « Dictyophycés »).

La première étape consiste en un alignement des Hidden Markov Models (HMMs) des Functional Families de CATH (funfam-hmm3.lib from 2020/09/22 [333], [334]) sur les différents protéomes avec la commande *hmmsearch* de *« hmm-3.3.2 »* [335]. Ces profiles HMM sont globalement représentatifs de la diversité des domaines des bases de données de séquences protéiques (principalement Uniprot [197] et Ensembl [198]). Si l'alignement est fructueux, le domaine est annoté avec un identifiant à quatre nombres. Cet identifiant reflète la structuration hiérarchique de la base de données en quatre niveaux: Classe (proportion de boucles α et de feuillets β), Architecture (arrangement brut des structures secondaires entre elles), Topologie (similarité structurales fine en termes de liens et d'arrangement des structures secondaires) et Homologie (similarité importante de séquence, impliquant potentiellement une fonction commune et lien évolutif étroit). Ces niveaux seront nommés respectivement C, CA, CAT et CATH dans le reste de l'étude. Le niveau correspondant à la définition d'une structure de domaine protéique, ou fold, est le niveau de la Topologie (CAT). Le taux d'annotation CATH a été calculé comme le rapport entre le nombre de gènes annotés structuralement sur le nombre de gènes total de chaque génome. 14834616 de gènes ont été annotés structuralement, avec environ 9 millions de gènes annotés dans EukProt et 5,5 millions dans les MAG.

Des comparaisons de taux d'annotation CATH ont été effectuées entre domaines du vivant (Fig.35) ainsi qu'entre MAGs et les RPs par phyla (Fig.37) à l'aide d'un test de Wilcoxon au seuil de confiance 1%, réalisé avec la fonction *geom_signif* du package R « ggsignif » [336]. Tous les tests de Wilcoxon des analyses suivantes ont été réalisé avec cette fonction. Les corrélations entre taux d'annotation CATH des MAGs et plusieurs propriétés génomiques ont été mesurées avec un coefficient de Pearson avec test de significativité au seuil de confiance 1% [337] (Fig.36). Les propriétés génomiques sélectionnées pour ces tests sont le score de complétion BUSCO, la longueur du génome, le nombre total de gènes (métadonnées des publications des ressources [5], [31]) et le nombre de gènes avec une annotation structurale.

occurrences des folds dans les protéomes

La valeur d'occurrence (OV) de chaque Homologie (niveau CAT**H**) dans un protéome correspond au nombre de domaines protéiques de ce protéome adoptant cette Homologie. À partir de cette valeur, l'occurrence de chaque fold (niveau CA**T**) correspond à la somme de toutes ses Homologies, et ainsi de suite jusquà la Classe (niveau **C**). L'OV de chaque fold au niveau d'un domaine du vivant ou d'un phylum correspond simplement à la somme des OVs de ce fold dans chacun des protéomes lui appartenant. L'ensemble des OVs de chaque fold dans un MAG ou un RP constitue son foldome. Les foldomes des différents domaines du vivant et phyla Eucaryotes (MAGs) ont été représentés à l'aide de la fonction *treemap* du package « treemap » [338] (Fig.39 ; Fig.41 ; Fig.42) et avec Krona [339] (https://doi.org/10.5281/zenodo.14935989). Les métriques en présence/absence proviennent directement des OVs, les OVs supérieures ou égale à un étant assignées à 1 et les autres à 0. L'ensemble des folds présents dans un domaine du vivant, un phylum ou un protéome constitue son répertoire de folds, qui est donc bien à distinguer de son foldome qui correspond à l'ensemble des folds d'un protéome (donc l'OV de chacun de ces folds). Les comparaisons qualitatives des répertoires de folds entre MAGs et RPs chez les Eucaryotes aux niveaux du domaine et des phyla ont été représentées avec le package « ggVennDiagram » [340] (Fig.38 ; Fig.40). Enfin, le nombre de combinaisons d'un fold indique le nombre de protéines multidomaines dans lesquelles il est trouvé ; et le nombre de partenaire, le nombre de fold différents en combinaison avec lui dans toutes les combinaisons. Ces grandeurs sont utilisées dans les Fig.52, 54 et 66.

clustering des MAGs basé sur leurs répertoires de folds

Les MAGs Eucaryotes dont la complétion BUSCO est supérieure à 50% ont été clusterisés en d'après leur répertoire de folds (Fig.44). Pour réaliser le clustering, une matrice de distance sur les présence absences de folds a d'abord été calculée avec la fonction *vegdist* du package « vegan » avec la méthode « jaccard » [341], puis le clustering a été réalisé avec la fonction *hclust*, méthode « ward.D2 » [342] et exporté à l'aide de la fonction *write.tree* du packet « ape » [343]. Il a ensuite été affiché sur Anvi'o [344] avec la fonction *anvi-interactive*. Les couches de métadonnées (phylum et groupe fonctionnel tel que défini dans Delmont *et al.* [5]) sont importées dans l'interface interactive à l'aide de la fonction *anvi-import-misc-data*. Des bins de folds ont ensuite été créés manuellement puis exportés avec la fonction *anvi-export-collection*. La proportion d'Architectures dans chacun de ces bins a ensuite été calculée, puis les bins ont été clusterisés de la même manière que précédemment (sauf méthode « euclidean » pour *vegdist*). La représentation graphique a été faite avec « ggplot2 » [345] (Fig.45).

comparaison des foldomes des Eucaryotes unicellulaires et pluricellulaires

Les RPs ont été utilisés pour évaluer les différences qualitatives et quantitatives des foldomes entre unicellulaires et pluricellulaires. Trois lignées pluricellulaires (Embryophytes, 24 RPs; Phaeophycées, 5 RPs; Metazoaires, 65 RPs) et trois lignée unicellulaires (Chlorophytes, 80 RPs; Diatomées, 83 RPs; Choanoflagellés, 27 RPs) ont été sélectionnées pour cette étudeLes différences ont été représentées comme décrit précédemment en utilisant 'ggVennDiagram' et 'ggplot2' [340], [345] (Fig.46 ; Fig.47 ; Fig.48).

modèles linéaires de la distributions des occurrences des folds dans les foldomes

Les OVs des folds dans les MAGs et les RPs peuvent être analysées par le prisme de leur distribution dans les foldomes (nombre de folds ayant une OV spécifique). Six phyla ont été sélectionnés pour cette étude : les Chordés et les Choanoflagellés pour les Unicontes et les

Chlorophytes, Bacillariophytes, MAST-4 et Haptophytes pour les Bicontes, car ce sont ceux avec le plus grand nombre moyen de MAGs par station TO.

Les OVs ont dans un premier temps été transformées en valeurs relatives (pour permettre les comparaisons entre MAGs et RPs), puis regroupées dans des bins logarithmique à l'aide de la fonction *logspace* du package « ramify » [346] afin de réduire le bruit statistique causée par la complétion variable des MAGs [347]. La loi de puissance (loi puissance) :

$$y = ax^{-\alpha} \quad (1)$$

a été utilisé pour modéliser les distributions. Comme elle suit une droite sur une double échelle logarithmique, les régressions linéaires ont été réalisées avec la fonction *lm* de R [342] puis représentées sur une double échelle logarithmique avec le package « ggplot2 » [345] (Fig.49; Fig.50; Fig.51; Fig.53). L'existence d'une potentielle relation de type loi puissance entre OVs et nombre de partenaires (nombre de fold différent trouvés en combinaison avec le fold) ou le nombre de combinaisons (nombre total de protéines mutidomaines dont un des domaines adopte le fold) a aussi été testée avec la même méthode (Fig.52 ; Fig.54).

<u>abondance des folds dans les communautés</u>

Afin d'étudier la distribution biogéographie des folds, leur abondance dans l'environnement a été calculée. Concrètement, une valeur d'abondance (AV) a été calculée pour chaque CATH par station et phylum en tenant compte à la fois de l'abondance relative de chaque MAGs dans la station et le phylum considéré ainsi que des occurrences du CATH dans ces MAGs, pondérée par la longueur de chacun d'entre eux en Mbp (Mega Paire de Base). Elle peut donc être considérée comme la somme des occurrences du fold pondérée par la couverture verticale de chaque espèce ayant ce fold. Elle a été calculée comme suit :

$$\text{vertical coverage MAG}_a\ \text{station}_j\ (\text{VC})_{a,j} = \frac{\text{AV of MAG}_a\ \text{station}_j}{\text{genome length MAG}_a\ \times\ 10^{-6}}$$

$$\text{abund CATH}_i\text{station}_j\text{phylum}_k = \sum_{a=1}^{\text{nb of MAG phylum}_k\ \text{in station}_j\ \text{having CATH}_i} (\text{VC})_{a,j} \times\ \text{OV CATH}_i\ \text{in MAG}_a \quad (2)$$

qui peut être généralisée pour tous les CATH et les stations pour chaque phylum comme suit, avec *m* le nombre de MAGs dans le phylum *k* et *n* la taille du répertoire en fold du phylum *k* (nombre total de folds différents dans tous les foldomes des MAGs du phylum *k*) :

$$\text{abund CATH phylum}_k = \begin{pmatrix} \text{VC}_{1,1} & \cdots & \text{VC}_{1,m} \\ \vdots & \ddots & \vdots \\ \text{VC}_{89,1} & \cdots & \text{VC}_{89,m} \end{pmatrix} \times \begin{pmatrix} \text{OV}_{1,1} & \cdots & \text{OV}_{1,n} \\ \vdots & \ddots & \vdots \\ \text{OV}_{m,1} & \cdots & \text{OV}_{m,n} \end{pmatrix} \quad (3)$$

Et pour la communauté Eucaryote complète (les 683 MAGs) :

$$\text{abund CATH all Eucaryotes} = \begin{pmatrix} \text{VC}_{1,1} & \cdots & \text{VC}_{1,683} \\ \vdots & \ddots & \vdots \\ \text{VC}_{89,1} & \cdots & \text{VC}_{89,683} \end{pmatrix} \times \begin{pmatrix} \text{OV}_{1,1} & \cdots & \text{OV}_{1,908} \\ \vdots & \ddots & \vdots \\ \text{OV}_{683,1} & \cdots & \text{OV}_{683,908} \end{pmatrix} \quad (4)$$

L'utilisation de la couverture verticale pour pondérer les valeurs d'occurrence permet de prendre en compte la complétion de chaque MAG, qui a un impact sur les OVs. L'AV de chaque Topologie (CAT) est la somme des AVs de toutes ses Homologies (CATH). Les AVs des MAGs dans une station sont compositionnelles (proportions au sein d'un échantillon qui ne représente qu'une partie de la communauté complète). Ce biais existe toujours pour les AVs des folds puisqu'elles sont basées sur celles des MAGs. Elles ont donc été transformée pour certaines analyses exploitant des méthodes d'écologie numérique (Fig.56 ; Fig.57 A ; Fig.62 ; Fig.67-77) à l'aide d'un Centered-Log Ratio (CLR), calculé avec la fonction *decostand* et la méthode "rclr" du package "vegan" [341] de R [342]. Cette méthode est une façon de limiter le biais compositionnel [348].

L'analyse de la distribution des folds dans les océans a été réalisée sur les échantillons de surface de la fraction de taille 0.8-2000µm (enrichie en Eucaryotes) [9], [349]. Pour représenter les abondances par heatmap (Fig.56; Fig.57 A; Fig.68), celles-ci ont été clusterisées. Le clustering est basé sur une matrice de distance calculée avec la fonction *vegdist*, méthode "bray". Il a été réalisé avec la fonction *hclust*, méthode "average". Ces deux fonctions proviennent du package R "vegan" [341]. La même méthode a été déployée pour les AVs des MAGs (Fig.57 B). La carte globale de similarité des stations basées sur l'abondance des folds a été réalisée avec une PCoA (fonction *cmdscdale*, paramètres k=3, eig=T et add=T) de R [342] calculée à partir de la même matrice de distance que celle utilisée pour le clustering des stations présenté au début de ce paragraphe. Les coordonnées des stations sur les trois premiers axes de la PCoA ont été converties en code RGB (Red Green Blue) à l'aide de la fonction *rgb* de R [342], puis représentées sur la carte du monde "world" du package R "mapdata" [350](Fig.62). Cette technique colorimétrique pour représenter une distance biologique par une distance colorimétrique avait été développée dans l'analyse biogéographique métagénomique de Richter *et al.* [121].

Les distributions des OVs et AVs de la communauté ont d'abord été ajustées à un modèle linéaire. Pour cette analyse, les AVs ne sont pas tranformées avec un CLR. Les deux grandeurs ont été log-binnées dans 100 bins. Le modèle utilisé pour l'ajustement est l'équation *(1)*, en suivant la même méthode que dans « modèles linéaires de la distribution des OVs dans les protéomes ». Les ajustements ont été calculés avec la fonction *linregress* du package « SciPy » [351] et représentés à l'aide de matplotlib [352] (Fig.58 A ; https://doi.org/10.5281/zenodo.14935989) dans Python [353].

modèles de Pareto type II de la distribution des abondance des folds

La PDF de la loi de Pareto type II (aussi appelée distribution Lomax [354], [355] et "PII" ensuite):

$$y = \frac{\alpha\left(\frac{k+x-\mu}{k}\right)^{-1-\alpha}}{k} \quad (5)$$

a ensuite été utilisée pour modéliser la distribution des AVs des folds dans les communautés de Bactéries et d'Eucaryotes (tous les MAGs Eucaryotes ensemble puis les MAGs de chacun des six phyla de la partie « modèles linéaires de la distributions des occurrences des folds dans les foldomes » p.93) (Fig.58 B,C ; https://doi.org/10.5281/zenodo.14935989).

Les ajustements de PII ont été réalisés à l'aide d'un script Python utilisant la fonction *curve_fit* du package « scipy.optimize » [57] avec les paramètres par défaut, à l'exception des limites (entre 0 et l'occurrence maximale pour *k* et *mu*) et de maxfev=50000. Une gamme de valeurs pour le paramètre α de l'équation *(5)* comprises entre 0 et 4,1 a été testée et l'erreur quadratique moyenne (RMSE) ainsi que la variance de l'estimation des paramètres ont été utilisées pour choisir la meilleure combinaison de paramètres. Le modèle présentant la valeur de variance la plus faible dans le premier quartile des valeurs de RMSE a été sélectionné comme étant le meilleur. Si la valeur *k* du meilleur ajustement est moins de 10 fois supérieure à la valeur d'occurrence la plus faible, ce qui se produit généralement

lorsque les valeurs d'occurrence faibles contiennent des valeurs aberrantes, un ajustement est fait en supprimant la valeur la plus faible, puis le processus recommence.

Une fois que la combinaison de paramètres conduisant au meilleur ajustement est identifiée, un test de Kolmogorov-Smirnov est alors exécuté à l'aide de la fonction *ks_2samp* du packet Python « scipy.stats » [351], afin de quantifier son adéquation avec les données observées. L'hypothèse H_0 pour chaque test est « la distribution ne dévie pas significativement d'une loi de Pareto type II ». Le seuil de confiance est de 1% pour chacun des tests. Les modèles ont été représentés à l'aide de du package « matplotlib » [352] (Fig.58 B,C). Les résultats du test (valeur de KS et *p*-value) (Table 1) ont été représentés à l'aide de la fonction *geom_density_ridges* du package « ggridges » [356] (Fig.59) puis sur une carte du monde pour chaque phylum à l'aide de la carte « world » du package R « mapdata » [350] (Fig.60).

Les corrélations avec les paramètres environnementaux et non environnementaux ont été testées en utilisant la fonction *cor* de R [342] avec les paramètres par défaut, ainsi que la fonction *cor.mtest* du package « corrplot » [357] avec le paramètre conf.level=0.99. Elles ont été représentées avec *ggcorrplot* du package « ggcorrplot » [337] (Fig.61). La plupart des paramètres non environnementaux (complétion BUSCO, longueur du génome, longueur de la partie codante du génome) proviennent de la publication des MAGs [3]. L'indice de Shannon des MAGs par phylum a été calculé à l'aide de la fonction *diversity* (paramètres par défaut) du package « vegan » [341]. Les paramètres environnementaux sélectionnés pour cette analyse sont la durée d'ensoleillement (SSD), la température de surface (SST) au moment de l'échantillonnage, la salinité, la concentration en dioxygène, la concentration en azote sous trois formes différentes (NH4, NO3 et NO2), la concentration en fer, en phosphate et en silicate et enfin la médiane, l'amplitude et l'écart-type de la SST mensuelle. Les paramètres mesurés lors de l'expédition TO ont été publiés par Pesant *et al*. [9]. Les concentrations en fer et en phosphate, les variables liées aux SST mensuelles et la salinité n'ont pas été mesuré lors de l'expédition TO. Les concentrations en fer à l'échelle globale ont été fournies par le modèle biogéochimique PISCES-v2 [111], [358]. Les SST mensuelles, transformées en médiane annuelle et en écart-type, les concentrations de phosphate et la salinité ont été extraites de WOA18 [10]. Les valeurs des paramètres du WOA13 ont une forte corrélation avec celles des mesures TO *in situ* (r^2 = 0,96, 0,89 et 0,83 respectivement) [111]. Les valeurs manquantes de WOA18 et PISCES-v2 aux stations TO ont été obtenues par extrapolation, en prenant la valeur au point avec la valeur existante à la longitude la plus proche.

définition des classes d'abondances de folds

Les valeurs du paramètre *k* de l'équation *(5)* ont été utilisées comme seuils pour séparer les folds en fonction de leurs AVs dans trois catégories pour chaque phylum. Les AVs des folds la première catégorie sont inférieures à *k* dans toutes les communautés ; les AVs de ceux de la deuxième catégorie sont toujours supérieures à *k*. Les folds dont les AVs ne sont pas systématiquement supérieures ou inférieures à *k* (à trois stations près) sont classé dans la catégorie «intermédiaire». Le noyau de ces catégories correspond aux folds dans la même catégorie dans les six phyla de l'étude (Fig.63; Table 2).

Les propriétés des folds dans chacune des catégories (nombre de Topologies, Architectures, nombre de partenaires) ont été représentées à l'aide de la fonction *upset* du package « ComplexUpset » [359], [360] ainsi que des fonctions du package « ggplot2 » [345] (Fig.64 ; Fig.65). Les différences significatives de nombres de partenaires des folds appartenant aux différentes catégories par phylum ont été testées avec le test de Wilcoxon au seuil de confiance de 1 % (Fig.66) [336].

Enfin, l'appartenance moyenne de chaque folds aux trois catégories environnementales dans les six phyla a été établie à l'aide de catégories arbitraires: « in most phyla » indique que le fold appartient à une catégorie donnée dans au moins quatre embranchements ; “when present” fait référence au fait que certains folds ne font pas partie du répertoire certains phyla. La mention « absent from the selected phyla » indique que les folds ne sont pas présents dans le répertoire des six phyla sélectionnés pour établir les catégories (Fig.68).

variabilité et biogéographie des folds dans les différentes catégories

La significativité des différences de d'AVs (avec transformation CLR) entre folds des différentes catégories d'abondance et de leurs noyaux a été testée pour chaque phylum (ainsi que pour une catégorie « 6 phyla together » qui correspond simplement à la somme des abondances des folds dans les six phyla de l'étude) à l'aide d'un test de Wilcoxon au seuil de confiance 1% (Fig. 67) [336].

L'α-diversité des folds dans les communautés a ensuite été mesurée à l'aide de l'indice de Shannon [88] par catégorie. Elle a été calculée de la même façon que pour les MAGs, donc sur le CLR des AVs des folds avec la fonction *diversity* (paramètres par défaut) du paquet « vegan » [341]. La représentation graphique a été avec la fonction *geom_smooth* du package « ggplot2 » [68] avec les paramètres par défaut (Fig.69). L'indice de Shannon a ensuite été comparé entre stations polaires et non polaires. Les premières sont situées à des latitudes absolues supérieures à 60°, et toutes les autres sont non polaires. La significativité des différence a été testée pour chaque phylum et catégorie d'abondance de la même manière que dans le paragraphe précédent, donc à l'aide d'un test de Wilcoxon avec un seuil de confiance de 1% (Fig.70 ; Table 3) [336]. La même chose a été faite avec les AVs des MAG (Fig.71 ; Table 3).

L'existence d'une structuration de la distribution biogéographique des communautés de chaque catégorie de folds dans chaque phylum a ensuite été testée à l'aide d'un dbRDA. Cette analyse a besoin de deux matrices : une matrice de données environnementale et une matrice de distribution. Les données environnementales utilisées pour ces modèles sont les mêmes que pour les corrélations de la Fig.61. Les analyses de dbRDA sont sensibles aux variables corrélées entre elles ; seules des variables non corrélées ont donc été utilisées (concentrations en fer et en phosphate, médiane et écart type des températures mensuelles). Des Moran Eigenvector Map (MEM) ont été rajoutées pour prendre en compte la structuration géographique des stations TO. Ils ont été calculés à l'aide de la fonction *dbmem* (paramètre MEM.autocor = non-null) du package « adespatial » de R [361] en utilisant la distance dans l'eau entre les stations TO [98]. Pour la matrice de distribution, une matrice de distance est d'abord calculée à partir des AVs (transformées en CLR) à l'aide de la fonction *dist* de R [342] avec les paramètres par défaut. Une PCoA est ensuite réalisée à l'aide de la fonction *pcoa* du package « ape » [343] avec les paramètres par défaut. La matrice formée par les coordonnées des stations dans l'espace PCoA est la matrice de distribution utilisée par la dbRDA. L'analyse de dbRDA est ensuite lancée. Afin de ne conserver que les MEM pertinents, la fonction *ordistep* de « vegan » [341] avec le paramètre direction réglé sur « both » et 10000 permutations sur un modèle nul et un modèle prenant utilisant toutes les MEM est intégrée à l'analyse. Seules les MEMs avec une *p*-value inférieure à 0,01 sont retenues. Le coefficient de détermination ajusté des dbRDAs est évalué à l'aide de la fonction *RsquareAdj* du package « vegan » [341]. L'existence de différences significatives de coefficients de détermination ajusté entre catégories d'abondances de folds a été évaluée à l'aide de la fonction *Anova* du package « car » [362]. Les coordonnées des stations dans les deux premiers axes de la dbRDA ont ensuite été converties en code RG à l'aide de la fonction *rgb* de R [342] puis les stations avec leurs couleurs ont été représentées sur une carte du monde à l'aide de la carte « world » du package R « mapdata » [350] (Fig.72 ; Table 4 ; Table 5 ; Table 6 ; https://doi.org/10.5281/zenodo.14935989). La

même analyse a été réalisée sur les AVs des MAGs par phylum (Table 4 ; Fig.73 ; Table 5 ; Table 6 ; https://doi.org/10.5281/zenodo.14935989).

Enfin, les AVs (transformées en CLR) des folds du noyau de la catégorie intermédiaire dans les Chlorophytes ont été représentées après avoir été d'abord regroupées par bassins océaniques puis clusterisées dans chaque bassin (matrice de distance calculée avec la fonction *vegdist*, méthode « Bray » [341] ; clustering avec la fonction *hclust*, méthode « average » [342]). Pour compléter la représentation, les AVs des MAGs Chlorophytes ont été affichées avec les stations dans le même ordre que pour les folds, ainsi que les OVs non transformées des folds du noyau de la catégorie intermédiaire dans les Chlorophytes (Fig.73).

modèle de distribution des folds à l'échelle de l'Océan global

La distribution de chaque fold à l'échelle de l'océan a été estimée en utilisant Climap. Climap est un outil interne non publié inspiré de la méthode développée par Frémont *et al*. [111]. Il a pour but de prédire la distribution d'une espèce ou d'un fold dans tous les océans ainsi que les principaux facteurs qui en sont à l'origine en se basant sur des mesures d'abondances dans des stations au contexte environnemental connu (en l'occurrence les stations *Tara* Oceans et Polar Circle) puis en extrapolant. Cette extrapolation est réalisée à l'aide d'un ensemble model (algorithme d'agrégation des prédictions de plusieurs méthodes, dans le but d'améliorer la capacité de généralisation et la robustesse du modèle) exploitant des méthodes d'apprentissage supervisé de la librairie scikit.learn [363]. Les modèles produits par Climap sont construits à partir de trois méthodes : Random Forest, Gradient Boosting et MLP (MultiLayer Perceptron). Le principe de l'ensemble model est de compenser les faiblesses des méthodes individuelles et limiter le surapprentissage (phénomène rendant le modèle trop spécifique aux données d'entraînement et incapable de généraliser). L'algorithme de Random Forest entraîne un ensemble d'arbres de décision, qui traitent chacun un sous-ensemble différent des données d'entraînement. Le Gradient Boosting entraîne un modèle additif sur un ensemble de prédicteurs faibles, à travers plusieurs étapes d'ajustement d'arbres de régression. Le principe de ces deux méthodes est similaire, dans la mesure où l'assemblage de prédicteurs faibles est censé renforcer la précision du modèle. Enfin, le perceptron multicouche (MLP) est un réseau de neurones à propagation directe, composé d'au moins trois couches : une couche d'entrée, une ou plusieurs couches cachées et une couche de sortie. Ici, la couche d'entrée contient autant de neurones que de variables explicatives et l'utilisateur module le nombre de couches cachées et de neurones qu'elles contiennent. La construction et la validation des modèles s'effectuent sur une grille d'hyper-paramètres : pour toutes les combinaisons d'hyper-paramètres possibles, un modèle est entraîné et testé par validation croisée. Climap contient deux méthodes pour construire le modèle final : le stacking et le vote. Le stacking est une méthode de « généralisation par empilement » qui consiste à associer les prédictions des modèles de base pour entraîner un dernier modèle [364]. Le vote consiste à moyenner les prédictions des trois modèles constitutif de l'ensemble model. La métrique finale de qualité choisie pour les modèles est un Root Mean Squared Error (RMSE)(Fig.74).

Climap contient un module d'analyse d'importance des variables d'un modèle faisant appel à deux méthodes: les tests de permutation et la méthode SHAP. La permutation consiste à changer les associations entre stations et valeurs pour chaque variable de manière aléatoire, puis de mesurer les conséquences de ces modifications sur l'erreur de prédiction du modèle à l'aide d'un R^2 (test statistique sur la significativité de la différence d'erreur avant et après permutation). Plus ce R^2 est élevé, plus l'impact de la variable sur la prédiction est important (Fig.75 ; Fig.76). La méthode SHAP [365] est basée sur les valeurs de Shapley, un concept relatif à la théorie des jeux. Dans ce cas, il s'agit d'un jeu coopératif où les variables sont des joueurs contribuant chacun au « gain » final, ici la prédiction. Les fonctions du modèle agissent comme les règles d'un jeu dans cette analogie. Le principe

est d'estimer la contribution de chaque facteur à la prédiction selon la valeur qu'il prend. On obtient ainsi une valeur de SHAP par variable et par prédiction pour chaque modèle. Cette valeur quantifie de combien la variable considérée fait dévier la prédiction (ici d'abondance) de la valeur d'abondance moyenne prédite. Elle est dans la même unité que l'abondance. SHAP produit un graphique récapitulatif de l'impact de chaque variable selon les différentes valeurs qu'elle peut prendre. Si l'amplitude des valeurs de Shapley pour une variable est élevée, on considère que cette variable contribue significativement au modèle. On obtient ainsi une estimation visuelle de l'ordre d'importance des variables. Pour faciliter la lisibilité, seules les SHAP values associées au 31 premières prédictions sont représentées (Fig.77 ; Fig.78 ; https://doi.org/10.5281/zenodo.14935989). Ces deux méthodes sont sensibles aux variables corrélées. Pour une analyse robuste, il est nécessaire d'entraîner les modèles sur des facteurs environnementaux non corrélés. Dans le cas des données environnementales marines, certains paramètres sont connus pour être fortement corrélés, par exemple la concentration en dioxygène et la température de surface. Des variables non corrélées ont donc été sélectionnées, à savoir les concentrations en fer et en nitrate ainsi que la température annuelle médiane et l'amplitude annuelle de température. Le jeu de paramètres environnementaux pour les prédictions rassemble des valeurs pour ces quatre variables dans tous les océans à une résolution de 1°. Les valeurs de températures et de fer proviennent des mêmes sources que celles présentées dans la partie « Pareto type II fits on the distribution of abundances of the folds in the communities per phylum ». Les valeurs de nitrate à cette granulosité ont été extraites du WOA18 [10]. La corrélation avec les valeurs de nitrate mesurées au cours de l'expédition TO [9] est très forte (0.9, $p < 0.01$).

Le projet d'analyser les folds protéiques dans les données de *Tara* Oceans a émergé au cours de discussions avec Youri Timsit et Magali Lescot (MIO) avant le démarrage de ma thèse. Les annotations structurales ainsi que certains résultats préliminaires, non présentés ici, ont été produits par Magali Lescot et Caroline Vernette (MIO). Les chapitres I et II regroupent mes travaux ainsi que ceux de deux stagiaires que j'ai co-encadré avec Olivier Jaillon, Louis Joigneaux et Alexandre Labesse. Les modèles du chapitre II ont été réalisés en collaboration avec Emanuele Pigani et Daniele Iudicone (SZN). Le chapitre III regroupe mes travaux ainsi que ceux produits par Manon Depaty, stagiaire que j'ai co-encadré avec Olivier Jaillon et Margaux Crédeville. L'outil utilisé au cours de son stage, Climap, a été développé par Téo Lemane et est une réécriture de la méthode publiée par Paul Frémont [111].

PARTIE 3. RÉSULTATS

CHAPITRE 1. FOLDOMES DES GÉNOMES ENVIRONNEMENTAUX ET DE RÉFÉRENCE

Sommaire

Les principales questions posées par ce chapitre sont :

- *La qualité des MAGs est-elle suffisante pour une étude à l'échelle des folds ?*

- *Quelles sont les caractéristiques des foldomes des génomes environnementaux et des protéomes de référence ?*

- *Existe-t-il un effet de la pluricellularité sur la distribution des folds dans les foldomes ? Si oui, peut-il être quantifié ?*

Pour y répondre, je présenterai dans un premier temps des statistiques relatives à la qualité de l'annotation structurale des MAGs et des RPs. Je décrirai ensuite les foldomes des différentes lignées des jeux de données. Je proposerai ensuite une classification des MAGs basée sur leurs répertoires de folds. Enfin, j'évaluerai l'effet de la pluricellularité sur la distribution des folds dans les foldomes à l'aide des RPs.

1/ statistiques de l'annotation structurale des protéomes

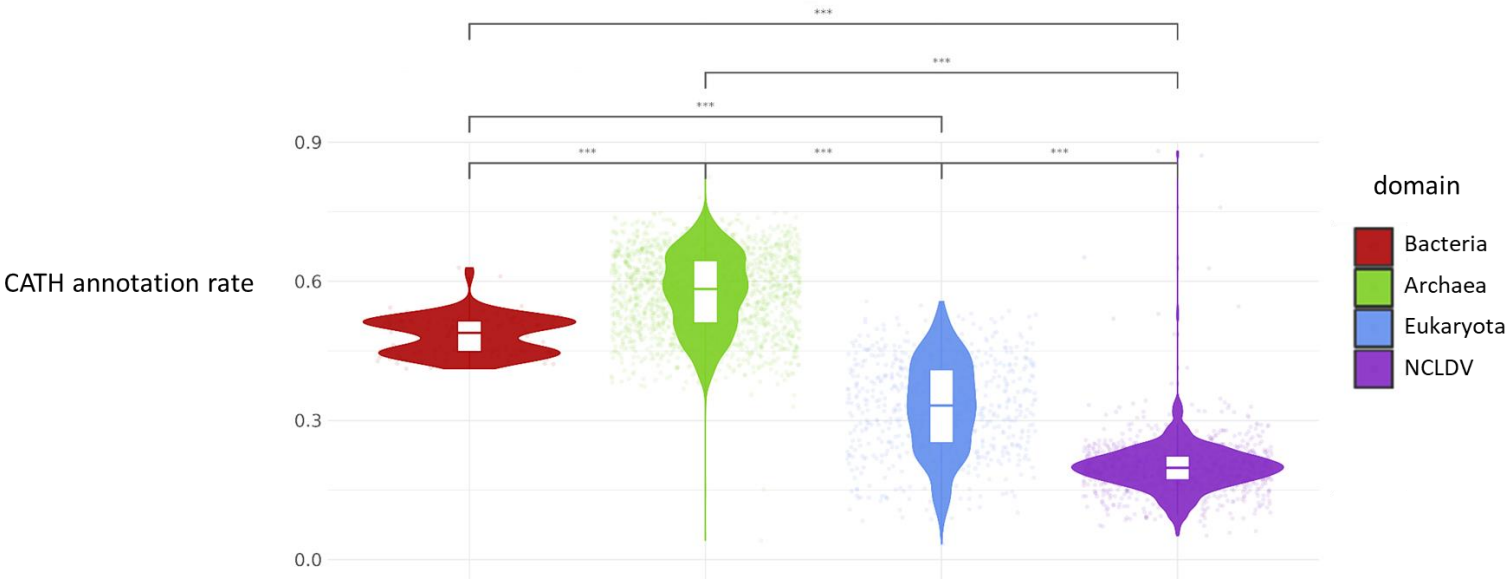


Figure 35. Taux d'annotation CATH par domaine du vivant. Résultat pour les MAGs. Le taux d'annotation CATH est la proportion de gènes annotés structuralement sur le nombre total de gènes de chaque génome. Les barres en haut de la figure représentent les résultats des tests de Wilcoxon (*** : *p*-value < 0.01).

Les taux d'annotation CATH (Figure 35) sont significativement différents entre domaines du vivant. Les NCLDVs ont le taux d'annotation moyen le plus bas, inférieur à 30%. Cela s'explique en partie par l'absence de représentation des génomes de ce groupe dans les bases de données avec lesquelles les HMMs de CATH sont construits. En outre, la distribution des valeurs de taux dans ce groupe est très unimodale, indiquant que la proportion de gènes annotés est globalement la même pour tous les protéomes. Les quelques valeurs aberrantes (proches de 90%) correspondent à des génomes très courts (quelques dizaines de gènes).

Les Eucaryotes sont le deuxième groupe avec le plus bas taux d'annotation, avec une valeur moyenne proche de 30%. Cela est attendu au vu du taux de complétion des MAGs de ce domaine (environ 38%), de sa diversité et du nombre de génomes planctoniques appartenant à des groupes sans représentants dans les bases de données. La complexité de leurs génomes et le grand nombre de gènes inconnus, même chez les espèces avec des représentants, doit également participer au faible taux d'annotation. Enfin, les Bactéries et Archées ont les meilleurs taux d'annotation moyens (entre 0.6 et 0.5), ce qui peut s'expliquer par plusieurs facteurs. Le premier est leur complétion moyenne qui est plus élevée que dans les autres groupes (87% et 83% pour les Bactéries et les Archées, respectivement). La proportion de protéines monodomaine dans ces groupes est également plus élevée que chez les Eucaryotes [214]. Les gènes codant pour ces protéines chez les Bactéries et Archées ont donc plus de chance d'être complets et donc de trouver une homologie dans CATH. Il y a aussi plus de transferts de gènes dans ces groupes, donc plus de gènes universels ; si ces gènes codent pour des proteines avec des homologies dans CATH, ils constituent un groupe de gènes systématiquement annotés, augmentant *in fine* le taux d'annotation global. La distribution bimodale des valeurs chez les Bactéries indique la possible existence de deux grands types de répertoires de folds dans les génomes de ce groupe, dont l'un comporterait potentiellement un nombre plus élevé de domaines inconnus, ou en tout cas sans homologues dans CATH. Ces deux types pourraient également être liés à la taxonomie

des MAGs bactériens. En effet, 75% des MAGs de ce groupe avec un taux d'annotation CATH supérieur à 58% sont des Proteobacteria. La diversité taxonomique des protéomes avec un taux d'annotation inférieur à 58% est au contraire beaucoup plus importante. Les principaux groupes y sont les Bacteroidota (27%), les Proteobacteria (20%), les Planctomycetota (18%) et les Verrucomicrobiota (13%). La taxonomie explique donc bien une partie du résultat observé, les Proteobacteria étant en moyenne mieux annotés que les autres.

À noter que de façon générale, les foldomes des MAGs ne représentent donc qu'une fraction de leur foldome total. Chez les Eucaryotes, cette fraction est d'environ 10% et peut descendre jusqu'à 0.1% pour certains MAGs particulièrement incomplets.

Pour mieux comprendre les facteurs à l'origine des variations du taux d'annotation CATH entre génomes, des corrélations avec différents paramètres génomiques ont été testées. Les corrélations pour les Bactéries et les Archées (Figure 36) montrent les mêmes tendances, à savoir que plus le génome est grand et le nombre de gènes élevés et plus le taux d'annotation CATH est bas. Cela indique peut-être que les Bactéries et Archées avec de grands génomes ont aussi tendances à avoir plus de gènes inconnus qui ne peuvent pas être annotés structuralement que ceux avec des petits génomes. De façon surprenante, cet effet n'est pas observé chez les Eucaryotes chez qui c'est la complétion qui est la plus corrélée avec le taux d'annotation. L'importance de la complétion traduit peut être le découplage entre longueur du génome et nombre de gènes chez les Eucaryotes, chez qui la proportion de génome non codant peut atteindre 98% [366]. De plus, une part importante de leurs gènes sont dupliqués ; dans certains groupes, ces duplications ont causé des expansions de certains gènes avec ou sans homologies dans CATH, ce qui est probablement à l'origine de l'absence de corrélations claires. La complétion est au contraire moins affectée par la longueur du génome non codant et le taux de duplication des gènes, expliquant peut-être la corrélation positive observée. L'absence de corrélation nette chez les NCLDV résulte probablement des valeurs aberrantes observées dans la Fig.35.

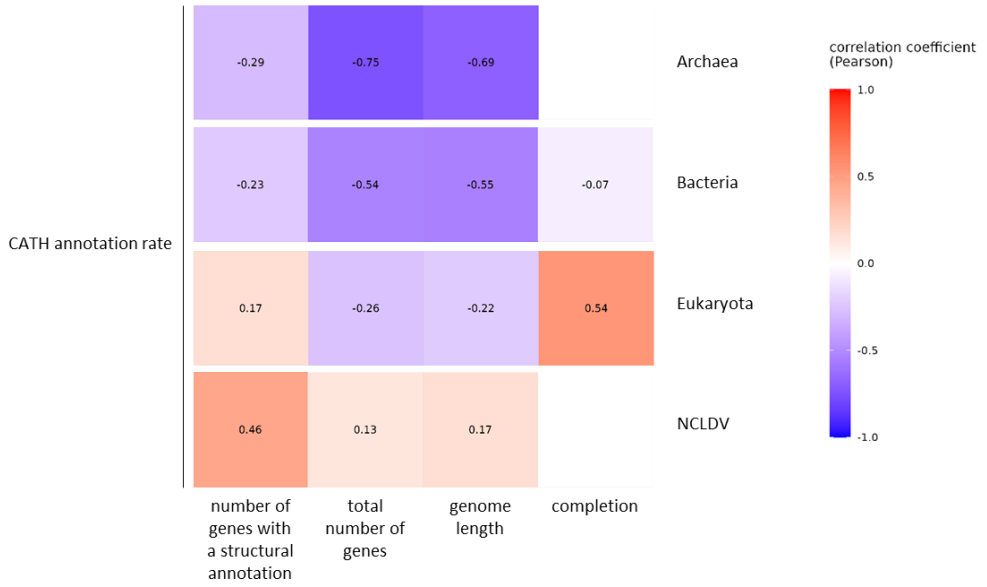


Figure 36. Corrélation entre taux d'annotation CATH par domaine du vivant et propriétés génomiques. Résultat pour les MAGs. Chaque ligne correspond à un domaine du vivant (indiqué à droite). Chaque case représente le résultat d'une corrélation entre le taux d'annotation CATH et la propriété génomique indiquée sur l'axe X (completion = BUSCO score). La couleur et la valeur de chaque case indiquent la valeur du coefficient de corrélation de Pearson. Les corrélations dans les cases vides sont non significatives ; toutes les autres le sont au seuil de confiance 1%.

Les MAGs Eucaryotes ont ensuite été comparés aux RPs pour évaluer plus finement l'effet de la complétion et les possibles biais taxonomiques associés sur les taux d'annotation CATH (Figure 37). Pour les MAGs, ce taux varie selon les phyla : les Chlorophytes et MAST-4 ont le plus élevé, proche de 45% en moyenne, contre 15% en moyenne chez les Ciliophores et les Myzozoa qui ont le plus faible. Ces deux derniers phyla se trouvent être les seuls représentant des Alvéolés dans les MAGs, un groupe sur lequel la méthode de reconstruction des génomes environnementaux est connue pour présenter des limitations importantes. Dans les RPs, les valeurs de taux moyen les plus élevées sont chez les Chordés et Arthropodes (55 et 75% respectivement), et la plus basse chez les Haptophytes (environ 25%).

Des différences significatives sont observées entre MAGs et RPs au sein d'un même phylum. Chez les Arthropodes, Chordés et Ciliophores, les taux d'annotation sont en moyenne deux fois plus élevés dans les RPs. Cela n'est pas surprenant puisque les RPs de ces phyla correspondent majoritairement à des organismes de référence ; il y a donc beaucoup plus de chances que leurs domaines aient des homologues dans CATH. Les RPs Chordés contiennent beaucoup de Vertébrés de référence comme *H.sapiens* par exemple. Les MAGs Chordés ne contiennent au contraire que quelques Appendiculaires, mais aucun Vertébrés. Les MAGs Arthropodes sont essentiellement des Copépodes. La qualité des MAGs n'est donc pas entièrement responsable des différences observées pour ces phyla. Les MAGs et RPs des Choanoflagellés, Chlorophytes, Cryptophytes, Bigyra, Bacillariophytes et Ochrophytes ont au contraire des valeurs de taux d'annotation moyen qui ne sont significativement pas différentes. L'annotation CATH est donc équivalente entre MAGs et RPs voire légèrement meilleure pour les MAGs de ces groupes, ce qui confirme leur qualité.

Il n'y a donc à ce stade pas de contre-indications majeures à l'utilisation de génomes incomplets pour des études à l'échelle des folds, bien qu'ils ne représentent qu'une fraction du foldome complet.

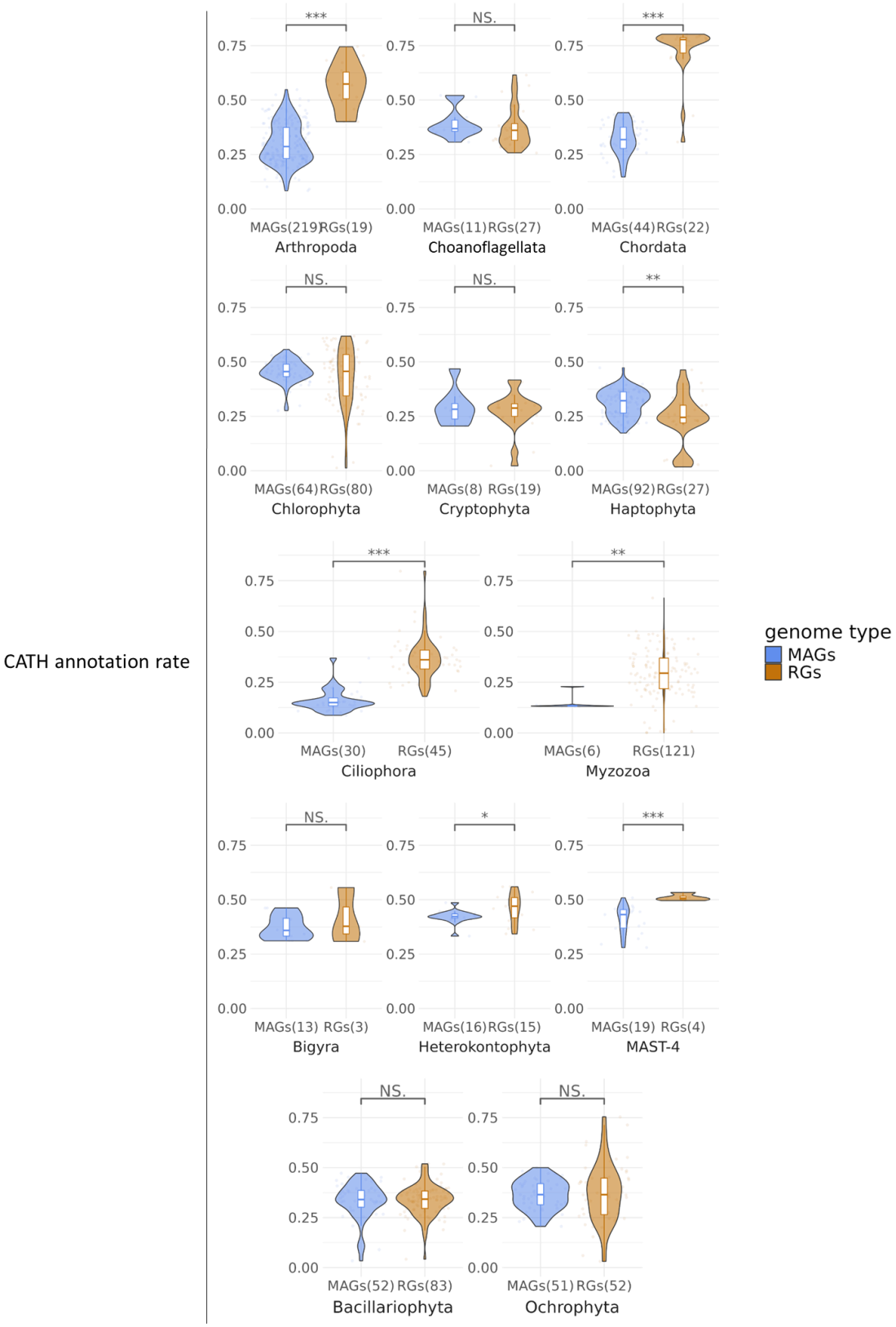


Figure 37. Comparaison des taux d'annotation CATH entre MAGs et RPs par phylum. Dans chaque sous-figure, le nombre de génomes/protéomes est indiqué entre parenthèses et le nom du phylum est en dessous. Les barres au-dessus de chaque sous-figure représentent le résultat du test de Wicoxon entre MAGs et RPs pour chaque phylum (*** : *p*-value < 0.01 ; ** : *p*-value < 0.05 ; * : *p*-value < 0.1 ; NS. : *p*-value > 0.1).

2/ comparaison des foldomes des génomes environnementaux et des protéomes de référence

Après avoir vérifié les taux d'annotations CATH des MAGs et des RPs, le premier objectif est de déterminer le contenu en fold des MAGs et d'identifier de potentielles spécificités d'usages de folds des groupes planctoniques du point de vue qualitatif. Pour cela, les répertoires de folds des MAGs Eucaryotes (tous ensemble et par phylum) et des RPs ont été comparés.

a. différences des répertoires entre MAGs et RPs

Le premier constat concernant la diversité en folds des répertoires des MAGs et des RPs est que celle des seconds est plus élevée que celle des premiers (1061 folds contre 885) (Figure 38 A). Seuls deux folds sont exclusifs aux MAGs :

- 1.10.1330 (*Type 1 dockerin domain*; une Homologie : 1.10.1330.10 *Dockerin domain*). L'Homologie 1.10.1330.10 est associée aux EC 3.2.1 et 3.1.1 (fonction de Glycosidase et Carboxylic Ester Hydrolase quand elle est adoptée par une enzyme). Ce fold n'est trouvé que dans trois Haptophytes avec une seule occurrence. La complétion de ces MAGs est de 18, 54 et 59% (la complétion médiane des MAGs Haptophytes est de 39%). La longueur de leurs génomes est comprise entre 1Mpb et 36Mpb. Le MAG avec la complétion la plus faible a 5636 gènes, celui avec la plus élevée en a 21435 (la médiane du nombre de gènes pour l'ensemble des Haptophytes est de 15408). Enfin, le taux d'annotation CATH moyen de ces trois MAGs est de 35% contre une médiane à 32% pour le phylum. Il n'y a donc rien du point de vue des propriétés génomique ou de l'annotation CATH qui puisse expliquer la présence de ce fold dans ces MAGs.

- 3.90.430 (*Activator of Metallothioenin 1 ; Chain A*; une Homologie : 3.90.430.10 *Copper Fist DNA-Binding domain*). L'Homologie 3.90.430.10 est trouvée à 99.9% dans des plantes et champignons, et n'est associée à aucun EC. Dans les MAGs, ce fold n'est présent que dans quatre Arthropodes avec une seule occurrence et dans l'unique Ascomycète avec trois occurrences. Compte tenu des informations fournies par CATH, il était attendu de trouver ce fold dans le MAG Ascomycète. Sa présence dans quatre MAGs Arthropodes est plus surprenante. Leur score BUSCO est compris entre 46 et 59, donc au-delà du troisième quartile de la distribution des complétions dans ce phylum. Le MAG Ascomycète a quant à lui une complétion de 93.7%. Les propriétés génomiques des quatres Arthropodes sont plutôt remarquables. Ils ont de très longs génomes (entre 76Mpb et 98Mpb, le troisième quartile pour la longueur des génomes dans ce phylum étant de 56Mpb), un nombre de gènes élevés (entre 21639 et 49137, le troisième quartile pour le nombre de gènes dans ce phylum étant de 19256) et un répertoire de folds plutôt important (entre 549 et 610, le troisième quartile pour la taille du répertoire de folds étant de 535 dans ce phylum). Leurs valeurs de taux d'annotation CATH sont cependant plutôt faibles par rapport à la médiane du phylum, qui est de 29% (contre entre 18 et 31% pour ces quatre MAGs). La présence de ce fold dans les quatre MAGs Arthropodes pourrait donc s'expliquer en partie par leurs propriétés génomiques particulières.

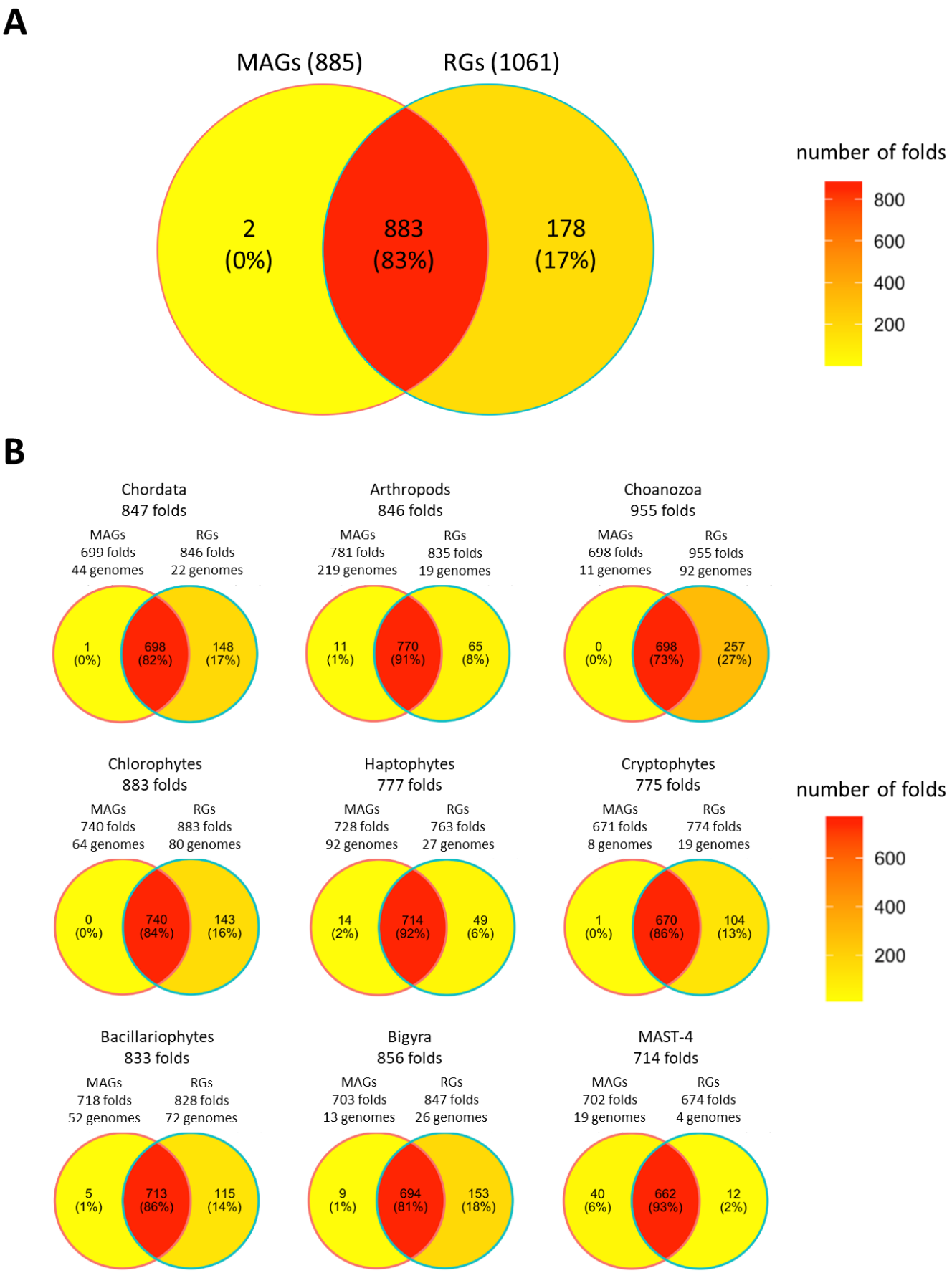


Figure 38. Différences des répertoires de folds entre MAGs et RPs. Le nombre total de folds dans chaque base de données est indiqué entre parenthèses. Les nombres et les couleurs indiquent le nombre de folds dans chaque sous ensemble. **(A)** À l'échelle des Eucaryotes. **(B)** Par phylum Eucaryote.

Concernant les 178 folds spécifiques aux RPs, trois ont plus de 10 occurrences dans tous les protéomes de certains taxa:

- 1.10.1220 (*Arc repressor Mutant*; dix Homologies pour 246 domaines) dont les occurrences sont supérieures à 10 dans tous les Vertébrés

- 1.10.1660 (*Multidrug-efflux Transporter Regulator, Chain : A; Domain 2*; cinq Homologies pour 90 domaines) dont les occurrences sont supérieures à 10 dans les Chrysophycés, les Filastères, les Métamonades, les Euglènozoés ainsi que dans une Diatomée (*Thalassiosira minuscula*, 26 occurrences).

- 1.10.3470 (*ABC transporter involved in vitamin B12 uptake, BtuC*; une Homologie pour 20 domaines) dont les occurrences sont supérieures à 10 dans les Chrsyophycés, quatre Amoebozoés, deux Filastètres, une Metamonade, *Thalassiosira minuscula*, et une Phaeophycée.

Ils ont tous les trois une Architecture Orthogonal Bundle (1.10), qui est l'Architecture adoptée par le plus grand nombre de folds d'après CATH (290 contre 224 pour la deuxième, le 2-layer sandwich 3.30). Parmis les folds avec des occurrences élevées dans les RPs non retrouvés dans les MAGs, certains adoptent l'Architecture 4.10 (4.10.1140, 4.10.770, 4.10.780, 4.10.81) et d'autre la 6.10 (6.10.10, 6.10.30, 6.20.10, 6.20.200, 6.20.240, 6.20.40). Ces deux Architectures correspondent à des Classes décrivant des folds qui s'écartent de la définition de folds globulaires, et qui sont généralement plutôt retrouvés chez les Métazoaires et plus spécifiquement les Vertébrés ; il n'est donc pas étonnant de les trouver exclusivement dans des RPs.

Au niveau du phylum, les répertoires des RPs sont généralement plus grands que ceux des MAGs (Figure 38 B). Ce n'est pas le cas pour les MAST-4, mais il n'y a que quatre RPs de ce phylum. Les MAGs Haptophytes, Arthropodes et dans une moindre mesure, Bigyra et Bacillariophytes recèlent une diversité de fold importante et non représentée dans les RPs du même phylum, mais trouvées au moins dans un RP puisqu'il n'y a au final que deux folds propres aux MAGs (Figure 38 A), comme détaillé ci-dessus. Dans l'ensemble, ces différences s'expliquent probablement essentiellement par des considérations techniques. Les MAGs ont peut-être en moyenne à la fois moins de gènes et moins de gènes avec une annotation CATH que les RPs, résultant en des taux d'annotation CATH relativement proche entre MAGs et RPs pour les phyla autres que les Arthropodes, Chordés et Ciliophores (Fig.37). La différence globale de nombre de gènes apparaît au final dans les différences qualitatives entre répertoires de folds des MAGs et RPs.

b. différences des foldomes entre MAGs et RPs

Les foldomes des neuf phyla combinés sont très similaires entre MAGs et RPs au niveau des proportions de Classes et d'Architectures et, dans une moindre mesure, de Topologies et d'Homologies (Figure 39). Le Rossmann fold (le fold le plus fréquent dans le vivant) représente environ 15% des occurrences dans les deux ressources. Les principales divergences proviennent de la Classe 2 et plus précisément des folds de l'Architecture Ribbon (2.10), dont les proportions relatives sont de 0.94% et 1.3% dans les MAGs et RPs respectivement, ainsi que les Classes 4 et 6 qui occupent des proportions plus importantes dans les RPs (1.3% et 0.14% contre 0.95% et 0.08% pour les RPs et les MAGs, respectivement). Ces différences proviennent probablement de la présence de Vertébrés dans les RPs des Chordés. Beaucoup de folds, en particulier dans les Classes 4 et 6, et de combinaisons inédites de folds sont apparus au cours de l'évolution de cette lignée [262], [264], [367]. En outre, les Copépodes, et plus généralement les Ecdysozoaires, ont des folds spécifiques impliqués dans la structure de l'hémocyanine [264]. Il n'est donc pas si surprenant d'observer des différences entre MAGs et RPs concernant les Classes 4 et 6.

Figure 39. Comparaison des foldomes entre MAGs et RPs. Les Treemaps sont construites à partir de l'occurrence des Homologies (niveau CATH) dans les neuf phyla utilisés pour la Fig.38. Chaque rectangle correspond à une Homologie ; la surface est proportionnelle au rapport entre somme des occurrences de l'Homologie dans tous les protéomes et somme des occurrences de toutes les Homologies dans tous les protéomes. Les nombres et leur taille correspondent à un niveau hiérarchique dans la classification CATH. Du plus grand au plus petit : Classe, Architecture, Topologie, Homologie. Les couleurs indiquent la Classe : en nuances de marron, Classe 1 ; en nuances de violet, Classe 2 ; en nuances de vert, Classe 3 ; en nuances de rose, Classe 4 ; en nuances de bleu, Classe 6.

3/ différences de répertoires et de foldomes des MAGs

Après les comparaisons entre MAGs et RPs pour les Eucaryotes, qui ont confirmé que l'annotation structurale des MAGs Eucaryotes était relativement satisfaisante, tous les MAGs ont été utilisés pour comparer les différences de répertoires et de foldomes entre espèces planctoniques des différents domaines du vivant.

a. comparaisons des répertoires de folds entre domaines du vivant

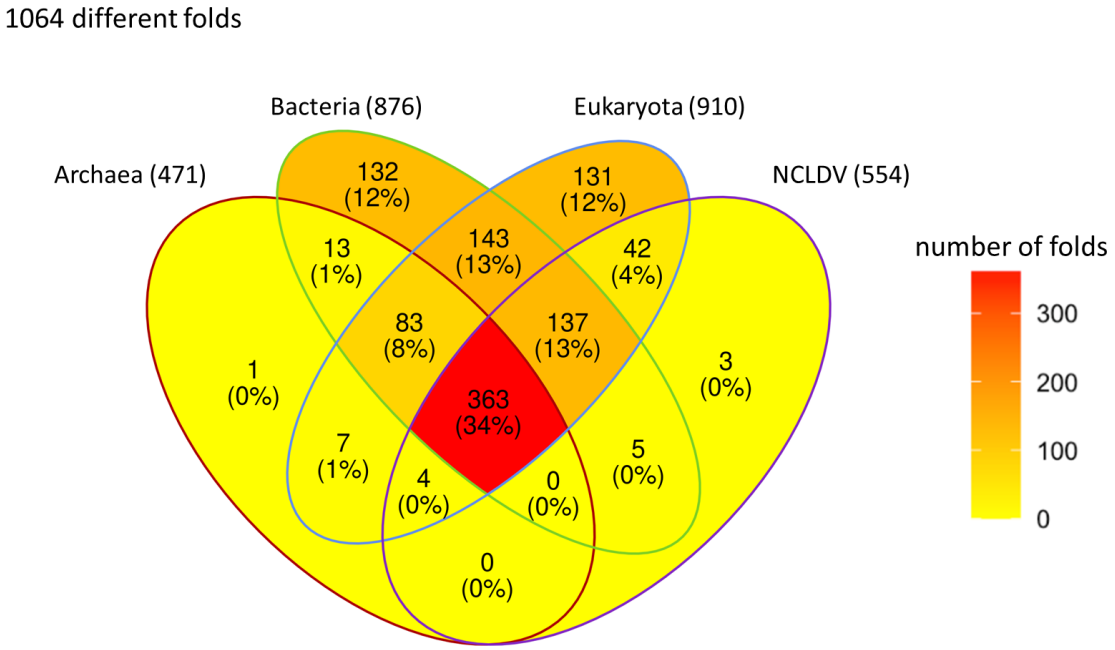


Figure 40. Spécificité des répertoires par domaine du vivant. Il y a 1064 folds en tout. Le répertoire de chaque domaine du vivant est constitué de la combinaison de l'ensemble des répertoires des espèces lui appartenant ; sa taille est précisée entre parenthèses après le nom du domaine. Les nombres et couleurs indiquent le nombre de folds dans chaque sous-ensemble.

Les répertoires de folds ont d'abord été comparés (Figure 40). Tout d'abord les Eucaryotes ont le plus grand nombre de folds différents (910), suivi par les Bactéries (876), puis les NCLDVs (554) et les Archées (471), représentant un répertoire total de 1064 folds. Près d'un tiers de ces folds sont adoptés par les domaines protéiques d'au moins un MAG dans chaque domaine du vivant ; une partie de ces 363 folds constitue un noyau commun à tout le vivant. Il est particulièrement intéressant de comparer ce chiffre avec la proportion de domaines uniques partagés par tous les domaines du vivant qui n'est que de 1.8% (correspondant à 186 domaines) [254], témoignant bien de la stabilité des folds au cours des temps évolutifs. Les Eucaryotes et Bactéries ont pratiquement le même nombre de folds exclusifs. Ce n'est pas le cas pour leurs domaines protéiques: 40% de la diversité totale des domaines uniques est spécifique aux Eucaryotes contre 25% pour les Bactéries (Figure 25) [254]. Cela montre que les innovations de domaines protéiques chez les Eucaryotes n'ont été que rarement accompagnées par des innovations de folds, mais bien que le mode d'évolution principal dans ce groupe est le réemploi de folds déjà existants mis en combinaison de manière innovante avec d'autres [214]. 4% des folds sont partagés exclusivement entre Eucaryotes et NCLDV, une proportion très élevée en comparaison de celle de domaines uniques partagés exclusivement entre Eucaryotes et Virus

qui est de 0.7% (représentant néanmoins environ 72 domaines) [254]. Cela s'explique peut-être par le fait que les NCLDVs infectent spécifiquement les Eucaryotes, et souligne potentiellement l'importance des transferts de gènes entre ces deux domaines au cours de leur coévolution.

Pour compléter ces observations, les foldomes des différents domaines du vivant ont été comparés (Figure 41)

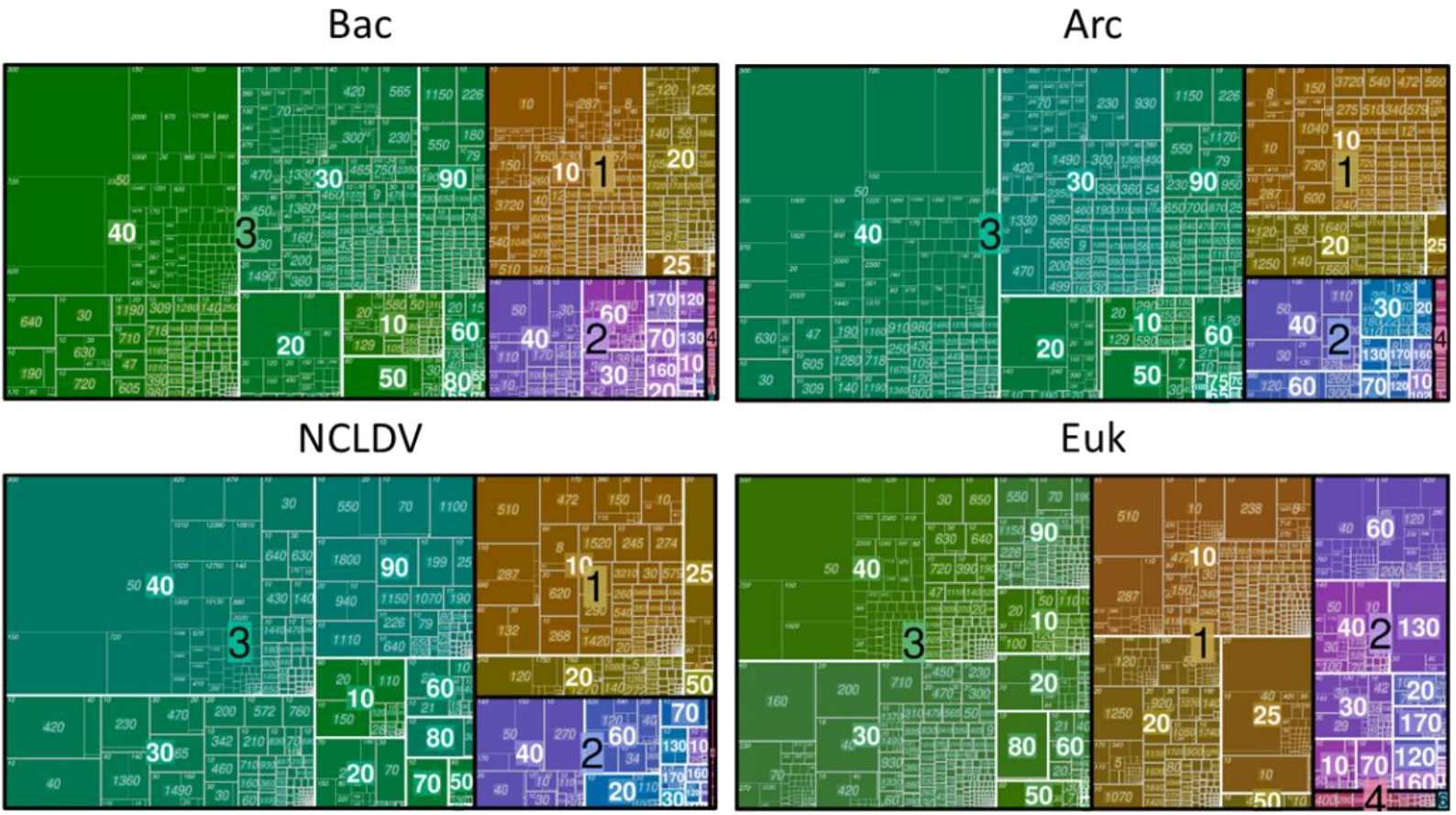


Figure 41. Comparaison des foldomes des différents domaines du vivant. Chaque rectangle correspond à une Homologie ; la surface est proportionnelle au rapport entre somme des occurrences de l'Homologie dans tous les protéomes et somme des occurrences de toutes les Homologies dans tous les protéomes De haut en bas et de gauche à droite : Bactéries, Archées, NCLDV, Eucaryotes. Les nombres et leurs tailles correspondent aux différents niveaux hiérarchique de CATH. La Classe 1 est en nuances de marron, la 2 en nuances de violet, la 3 en nuances de vert, la 4 en nuances de rouge, la 6 en nuances de bleu. Une visualisation interactive en Krona des foldomes est accessible à cette adresse : https://doi.org/10.5281/zenodo.14935989

Les différentes Classes de folds occupent des proportions différentes selon le domaine du vivant considéré. Entre 65 à 70% des folds des NCLDVs et des Procaryotes appartiennent à la Classe Alpha Beta, contre 50% chez les Eucaryotes. La Superfamille la plus fréquente dans tous les domaines du vivant est la P-loop containing nucleotide triphosphate hydrolases (3.40.50.300). Sa fonction principale est la production d'énergie via hydrolyse des ATP ou GTP ou oxydation des NADH ou NADPH [214], [257]. Le fold le plus fréquent est le Rossmann fold (3.40.50). Il s'agit d'un superfold qui est probablement apparu plusieurs fois au cours de l'évolution et qui est adopté par une grande diversité de domaine dans des protéines de fonctions très variées [241], [242]. Il représente environ 15% des occurrences chez les Eucaryotes contre 25% en moyenne chez les Procaryotes et NCLDVs. Cette différence provient essentiellement d'une utilisation plus importante des folds all Beta et all Alpha

chez les Eucaryotes. En effet, environ 10% des folds chez les Eucaryotes ont une Architecture up down bundle (1.20), contre 5% en moyenne chez les autres. La proportion plus élevée de folds de Classe 2 chez les Eucaryotes est principalement la conséquence de l'utilisation de folds Immunoglobulin-like (2.60.40) qui sont fréquemment impliqués dans des voies de signalisation cellulaire spécifiques à ce domaine [245]. Les folds de la Classe Few Secondary Structures sont également plus fréquents chez les Eucaryotes (environ 1% des occurrences) que chez les autres (0,6, 0,7 et 0,2% chez les Bactéries, les Archées et les NCLDVs respectivement), de même que les repliements de la Classe Special (0,1% des occurrences chez les Eucaryotes contre 0,01 % chez les Bactéries et absent chez les autres). Les folds spéciaux et désordonnés chez les Eucaryotes sont surtout adoptés par des domaines de protéines impliquées dans des fonctions de régulation et de signalisation, qui sont beaucoup plus fréquentes et diversifiées dans ce domaine du vivant que dans les autres [240], [245].

Dans l'ensemble, les résultats de cette partie montrent que les tendances d'usage de folds connues grâce aux études basées sur des génomes de référence sont retrouvées avec les MAGs. Ils confirment la divergence des Eucaryotes autant au niveau de leurs foldomes que de leurs répertoires de folds, malgré leur diversité en folds plus faible que dans les RPs. Les MAGs ont aussi permis d'identifier certains folds partagés probablement exclusivement par les NCLDVs et les Eucaryotes. Les Bactéries possèdent quant à elles des répertoires très diversifiés mais des foldomes relativement similaires à ceux des NCLDVs et Archées.

b. comparaison des foldomes des phyla Eucaryotes

Le reste des résultats de ce Chapitre sont restreints aux Eucaryotes, qui seront également au centre des analyses des Chapitres suivants. L'objectif est de décrire les foldomes des différents phyla afin d'identifier de potentielles spécificités d'usage. Les phyla sélectionnés dans cette sous partie sont les Chordés, les Choanoflagellés, les Arthropodes, les Chlorophytes, les Bacillariophytes, les MAST-4, les Bigyra, les Cryptophytes et les Haptophytes.

Les foldomes de MAGs de ces neuf phyla Eucaryotes ont dans un premier temps été visualisés pour avoir un premier aperçu de leur variabilité taxonomique (Fig.42). Dans l'ensemble, elle est faible et chaque foldome est globalement similaire à celui des Eucaryotes (Fig.41). Dans tous les phyla, les folds Alpha Beta et mostly Alpha représentent environ la moitié et un tiers des occurrences, respectivement. Les folds mostly Beta occupent une proportion d'occurrence significativement différente entre Unicontes et Bicontes (15.3% en moyenne contre 21.7% en moyenne, respectivement ; Wilcoxon test, *p*-value <0.05). Les Classes 4 et 6 représentent des proportions d'occurrence significativement plus élevée dans les Arthropodes et Chordés que dans les autres phyla (2% des occurrences relatives en moyenne pour la Classe 4 chez les Unicontes contre 0.64% pour les Bicontes ; Wilcoxon test, *p*-value <0.05). Comme cela a été dit plus haut, les protéines dont les domaines adoptent ces Classes de fold ont tendance à avoir des fonctions de signalisations et de régulations ; il n'est donc pas surprenant qu'ils occupent une plus grande part des occurrences relatives chez les pluricellulaires que chez les unicellulaires.

Figure 42. Foldomes des MAGs Eucaryotes par phylum. Le nombre de folds et de MAGs par phylum est indiqué entre parenthèses. Chaque treemap contient l'ensemble des MAGs appartenant à un phylum donné, dont les occurrences de chaque Homologie ont été sommées. La taille des rectangles représente la proportion d'occurrence occupée par l'Homologie correspondante. Les nombres indiquent l'identifiant CATH, leur taille les niveaux de la classification hiérarchique. Dans l'ordre de taille décroissante : Classe, Architecture, Topologie et Homologie. Les couleurs correspondent aux différentes classes : en nuances de marron, Clases 1 ; en bleu-violet, Classe 2 ; en vert, Classe 3 ; en rosee, Class 4 ; en vert-bleu, Classe 6. Une visualisation interactive par Krona des foldomes sont accessibles à cette adresse : https://doi.org/10.5281/zenodo.14935989

Enfin, il n'y a pas de tendances taxonomiques au sein des Bicontes à l'échelle des Classes: le foldome des MAST-4 ressemble plus à celui des Cryptophytes qu'à ceux des Bacillariophytes et Bigyra par exemple, alors qu'ils sont plus proches du point de vue taxonomique. À l'échelle des folds, les proportions d'occurrences relatives occupées par le Rossmann fold sont légèrement variables entre Bicontes et Unicontes : en moyenne 17% pour les Bicontes et 10,7% chez les Unicontes (Wilcoxon test, *p*-value <0.05). Il est intéressant de noter que les Choanoflagellés, qui font partie des Unicontes mais sont unicellulaires, ont un foldome qui pourrait être qualifié d'« intermédiaire » entre les Bicontes et les Métazoaires, notamment parce que les proportions d'occurrences de leurs folds de Classe 2 et 4 sont plus proches de celles des Métazoaires, mais la proportion d'occurrence du Rossmann fold est en revanche plus proche de celle des Biconte. Les Arthropodes présentent également une spécificité au niveau de leurs Homologies : c'est le seul phylum dans lequel l'Homologie 3.30.160.60 (Classic Zinc Finger, 359 domaines) occupe plus d'occurrences relatives que la 3.40.50.300 (6% contre 4%, respectivement).

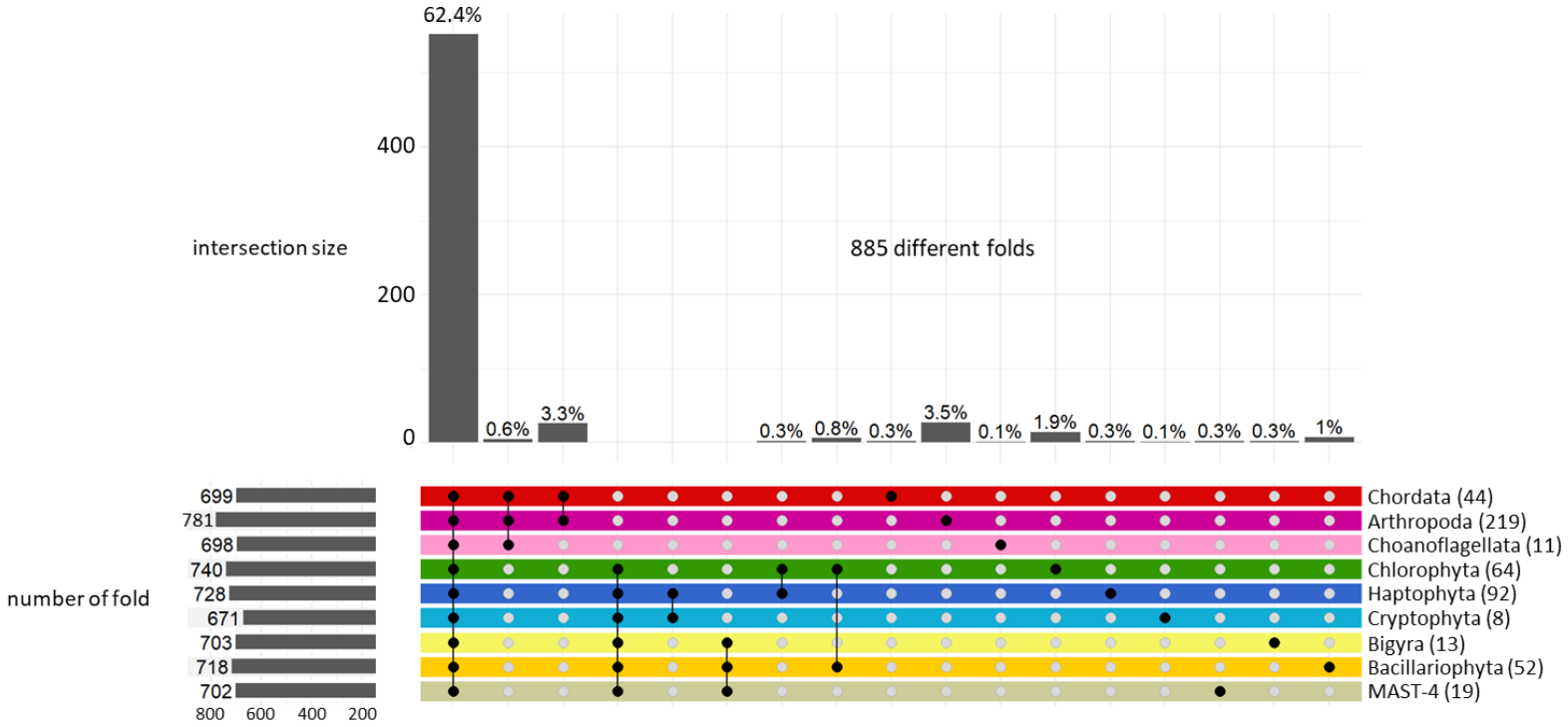


Figure 43. Spécificité des répertoires de folds par phylum. Le répertoire des neuf phyla combinés contient 885 folds différents au total. Au centre de la figure, dans les bandes colorées, les points noirs indiquent les phyla utilisés pour réaliser les intersections dont les tailles sont représentées par les barres verticales en haut. Chaque couleur correspond à un phylum. Les barres horizontales à gauche de la figure montrent le nombre de folds par phylum. Le nombre de MAG dans chaque phylum est indiqué entre parenthèses à droite.

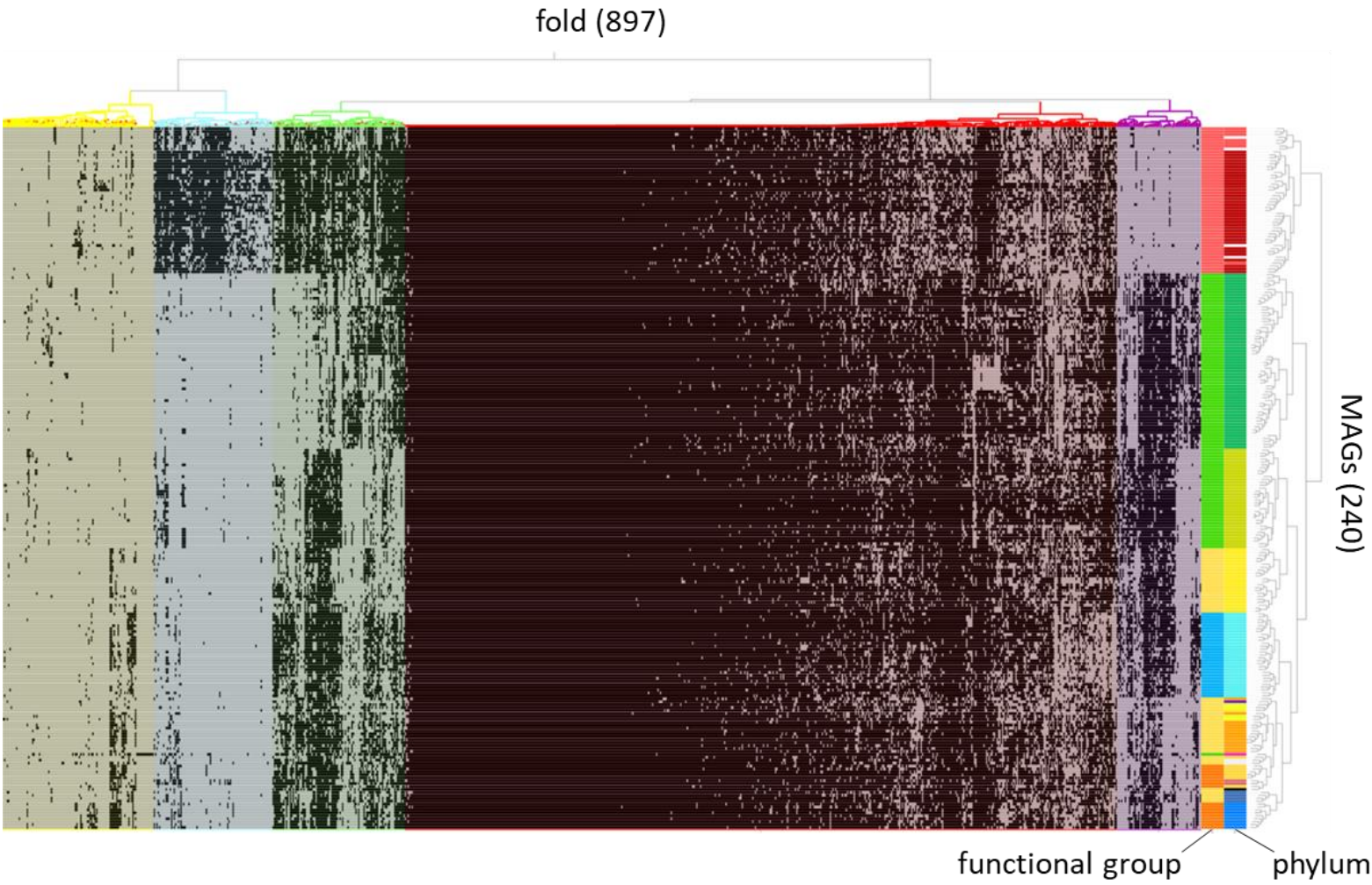
Après avoir comparé les foldomes des différents phyla Eucaryotes, il en a été fait de même pour leurs répertoires de folds (Figure 43). Ils partagent la majorité de leurs folds (62% soit 552 sur les 885 différents). Les Chordés et Arthropodes, deux seuls phyla pluricellulaires des MAGs, partagent 29 folds entre eux qui ne sont pas retrouvés dans les autres, ce qui en fait les deux phyla ayant le plus grand nombre de folds partagés spécifiquement. Cela est probablement lié à la fois à leur proximité taxonomique et à la pluricellularité. En outre, les Arthropodes sont le groupe avec le plus de folds qui leurs sont spécifiques (environ 31), alors que les Chordés n'ont que trois folds spécifiques, une

conséquence probable de leur faible diversité dans les MAGs. Les Unicontes en général (deuxième barre verticale en partant de la gauche) partagent cinq folds spécifiques. Il n'y a au contraire aucun folds spécifique partagé par tous les Bicontes. Au sein de ce groupe, des phyla proches taxonomiquement comme les Haptophytes et les Cryptophytes, ou les MAST-4, les Bacillariophytes et les Bigyra ne partagent pas non plus de folds spécifiques. Les Chlorophytes sont le deuxième groupe après les Arthropodes avec le plus grand nombre de fold spécifiques ; ils sont aussi le groupe des MAGs avec la meilleure complétion moyenne et le taux d'annotation CATH le plus élevé (Figure 37). En revanche, les Chlorophytes et les Bacillariophytes, qui partagent des niches écologiques plus similaires que les Bacillariophytes avec les MAST-4 par exemple, partagent sept fold spécifique [5]. Le même type d'observation est fait entre les Chlorophytes et les Haptophytes, qui partagent trois folds spécifiques. Cela montre qu'en dehors des folds universels, l'usage des folds est peut-être en partie influencé par l'écologie en plus de l'histoire évolutive (en tout cas chez les Bicontes et pour les Unicontes pluricellulaires).

4/ clustering des MAGs basé sur leurs répertoires de folds

Cette partie présente une classification d'une part des MAGs basée sur leurs répertoires de folds. Il est connu que ces répertoires peuvent être utilisés pour classer le vivant, mais cela n'avait jusqu'ici été fait qu'avec des génomes de référence et ne s'appliquait donc globalement pas à la diversité observée dans le plancton [20]. Ici, cette classification a aussi pour objectif d'évaluer à quel point l'écologie et la taxonomie des MAGs impactent leurs répertoires de folds.

Le clustering des répertoires de folds est réalisé avec tous les MAGs dont la complétion est supérieure à 50%, incluant donc des phyla supplémentaires par rapport aux Fig.42 et 43 (Figure 44). Le répertoire total des 240 MAGs en question contient treize folds de plus que celui des neuf phyla des figures précédentes, dont certains sont spécifiques à l'unique MAG Ascomycète. Le clustering se révèle être très cohérent avec la taxonomie, à l'exception de la position des Choanoflagellés, des Bacillariophytes et des Ochrophytes. Ces deux derniers se retrouvent dans une branche distincte des Heterokontophytes, Bigyra et MAST-4 alors qu'ils en sont taxonomiquement proches. Ces trois taxons de Stramenopiles sont au contraire groupés avec les Choanoflagellés, Cryptophytes, sister Cryptophytes, Cercozoa et Ascomycota. Ces incohérences semblent s'expliquer par des similarités fonctionnelles, indiquées par la couche « functional group ». Les groupes fonctionnels ont été définis par Delmont *et al*. en utilisant les occurrence d'environ 28000 fonctions [5]. Les Bacillariophytes se trouvent appartenir au même groupe fonctionnel que les Chlorophytes ; il n'est donc pas surprenant que ces deux groupes partagent certains folds de manière exclusive. En outre, les Bacillariophytes possèdent également des folds spécifiques des Straménopiles. La conséquence de l'existence de ces deux types de folds dans le répertoire des Bacillariophytes est probablement à l'origine de leur position intermédiaire, entre Chlorophytes et Stramenopiles, dans le clustering. La position des MAGs Ochrophytes est également intéressante puisqu'ils se retrouvent sur la même branche que les Bacillariophytes alors que leur groupe fonctionnel (groupe 2) est celui des autres Straménopiles. La position des Choanoflagellés s'explique aussi par leur appartenance au groupe fonctionnel 4 dont font partie les sister Cryptophyta et les Heterokonthophytes. Enfin la séparation des Haptophytes et Cryptophytes s'explique probablement en partie par le fait qu'ils n'appartiennent pas au même groupe fonctionnel.



phylum
- Arthropoda (39)
- Ascomycota (1)
- Bacillariophyta (34)
- Bigyra (4)
- Cercozoa (1)
- Chlorophyta (60)
- Choanoflagellata (2)
- Chordata (7)
- Cryptophyta (4)
- Haptophyta (29)
- Heterokontophyta (7)
- MAST-4 (13)
- Ochrophyta (23)
- Putative New Branch (1)
- Sister Cryptophyta (9)
- NA (6)

functional group
- Group 1 (29)
- Group 2 (49)
- Group 3 (95)
- Group 4 (17)
- Small Animals (50)

fold group
- I (113)
- II (89)
- III (99)
- IV (533)
- V (63)

Figure 44. Clustering des MAGs basé sur leur répertoire de folds. Les 240 MAGs dont la complétion est supérieure à 50% sont sur l'axe des ordonnées. Le dendrogramme résultant de leur clustering est à droite sur cet axe. Le nombre de MAGs dans chaque phylum et groupe fonctionnel est indiqué dans la légende à gauche, sous les catégories « phylum » et « functional group ». Les groupes fonctionnels sont ceux définis par Delmont *et al*. à partir des occurrences d'environ 28000 fonctions différentes [5]. Les folds sont sur l'axe des abscisses. Chaque point dans la heatmap indique la présence (point noir) ou l'absence (point gris) d'un fold. Ils sont clusterisés sur l'axe des abscisses et le dendrogramme associé est représenté en haut. Il a été manuellement découpé en cinq clusters dont les branches et les zones correspondantes de la heatmap sont colorées avec les couleurs indiquées dans « fold group » (dans la légende à gauche). Le nombre de folds dans chaque groupe est indiqué entre parenthèse dans cette catégorie. La proportion d'Architectures (niveau CATH) dans chaque catégorie est détaillée dans la Fig.45.

Le clustering des folds sur l'axe des abscisses révèle cinq groupes :

- groupe I : présents de manière quasi aléatoire avec une dizaine de fold présent assez fréquemment chez tous les unicellulaires non Chlorophytes et Bacillariophytes

- groupe II : présents surtout chez les Metazoaires avec quelques folds présents également dans un certain nombre d'unicellulaires non Chlorophytes, en particulier des Bacillariophytes

- groupe III : présents de manière assez universelle, quasiment tous chez les Metazoaires et les unciellulaires non Chlorophytes et Bacillariophytes. Ces deux derniers n'en ont chacun qu'environ une moitié mais qui n'est pas la même entre les deux phyla

- groupe IV : les folds universels et très fréquents, dont l'absence ne résulte probablement dans l'immense majorité des cas que de la complétion (bien que certaines absences semblent être spécifiques de certains groupes à une résolution taxonomique inférieure au phylum)

- groupe V : globalement absents des Métazoaires à l'exception de quelques-uns. Une partie partagée par les trois groupes photosynthétiques (groupe fonctionnel 1, 3 et les Ochrophytes) avec une dizaine de folds spécifiques aux Chlorophytes dont certains sont partagés avec les Haptophytes. Les unicellulaires n'appartenant pas aux groupes 1, 3 ou aux Ocrophytes ont également un répertoire spécifique dans ce groupe

Au final, l'information fournie par les répertoires de folds donne un clustering « intermédiaire » entre classification par groupe fonctionnel et la taxonomie, avec une position particulièrement intéressante pour les Bacillariophytes. Certaines portions des répertoires de fold sont partagées entre taxons de par leur proximité phylogénétique (certains fold des groupe III et V) alors que d'autres le sont de par la proximité fonctionnelle (également dans les groupes III et V principalement).

Le clustering basé sur les répertoires est plus cohérent avec la taxonomie et les groupes fonctionnels que celui basé sur les foldomes (non présenté ici), dans lequel la métrique est une valeur d'occurrence. Avec les MAGs, cela est principalement lié au fait que les occurrences sont impactées par les complétions et taux d'annotation CATH ; en outre, des analyses similaires faites sur des génomes de référence ont abouti aux mêmes conclusions, parce que les occurrences sont aussi impactées par les taux de duplication qui peuvent être espèce-spécifiques au sein d'une lignée, ce qui peut brouiller le signal taxonomique et fonctionnel [260].

Afin de mieux comprendre les différences observées entre groupes de folds dans la Fig.44, leur contenu en Architecture a été représenté (Figure 45). Les cinq groupes sont assez différents à ce niveau de classification CATH. Les folds de Classe 6 ne sont présents que dans les catégorie I, II et IV. La catégorie avec la proportion de folds de l'Architecture 4.10 la plus élevée est la II, dont la majorité des folds ne sont présents que chez les Métazoaires. C'est aussi la catégorie avec la plus faible proportion de folds de Classe 3, ce qui est cohérent avec les observations de la Fig.42. La catégorie avec la plus faible proportion de folds de Classe 2 est la V qui est pratiquement absente des Métazoaires. Ce groupe, ainsi que le II, sont les plus divergents (notamment à cause de leurs proportions de folds de Classe 2 et de Classe 4), et ce sont aussi les deux groupes les plus spécifiques du point de vue taxonomique : le groupe II est pratiquement spécifique des Métazoaires alors que le V est plutôt propre aux unicellulaires.

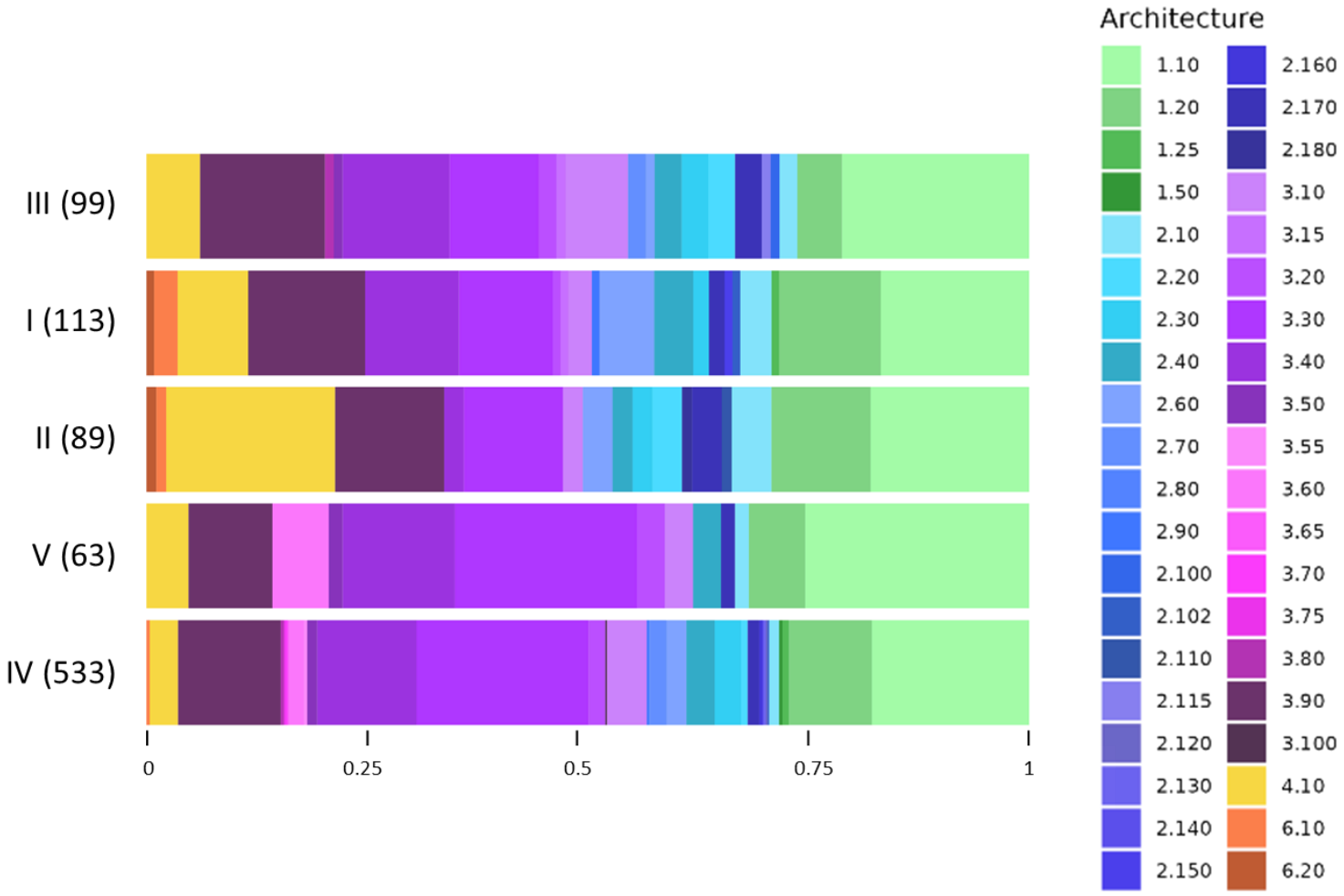


Figure 45. Proportion d'Architectures dans chaque groupe de folds défini dans la Fig.44. Les différents groupes sont indiqués à gauche, avec le nombre de folds dans chacun d'eux entre parenthèses. Les couleurs correspondent aux Architectures ; en vert, celles de la Classe 1 ; en bleu, celles de la Classe 2 ; en violet, celles de la Classe 3 ; en jaune celles de la Classe 4 (4.10) ; en orange celles de la Classe 6.

5/ différences de répertoires de folds entre unicellulaires et pluricellulaires

Comme cela a été observé dans les résultats précédents et détaillé dans la littérature [264], le passage de l'unicellularité à la pluricellularité s'accompagne de l'acquisition d'un grand nombre de fonctions, et donc potentiellement de nouveaux folds. Comme le passage de l'unicellularité à la pluricellularité a eu lieu plusieurs fois en parallèle au cours de l'évolution du vivant, la question ici était de vérifier s'il existait malgré cela des folds partagés universellement par les pluricellulaires. Les MAGs ne permettant pas d'explorer cette question, ce sont les RPs qui ont été utilisés ici, puisqu'ils représentent une plus grande diversité d'Eucaryotes que les MAGs et comprennent des Bicontes pluricellulaires, à savoir des Embryophytes et des Phaeophycés. Les Homologies et les Topologies seront utilisées dans cette partie, afin d'avoir une vision la plus précise possible du phénomène étudié.

Les taux d'annotation CATH des différents phyla utilisés ici sont significativement plus élevés pour les pluricellulaires que les unicellulaires (0.6 en moyenne pour les pluricellulaires : 0.57 pour les Arthropodes, 0.75 pour les Chordés, 0.61 pour les Embryophytes et 0.43 pour les Phaeophycés, contre 0.38 en moyenne pour les unicellulaires ; Wilcoxon test ; *p*-value<0.01). Cela provient probablement du fait qu'une part importante des domaines dans CATH sont des domaines de Chordés, et que les unicellulaires y sont minoritaires.

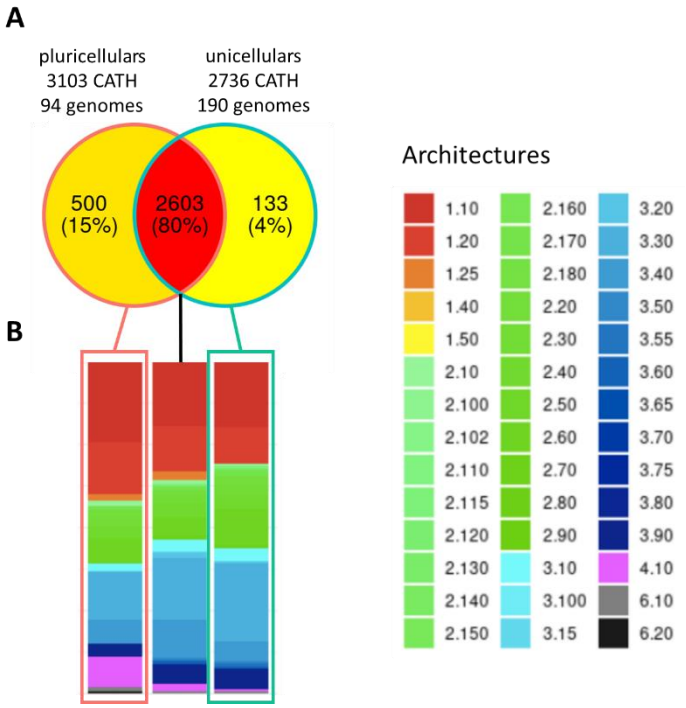


Figure 46. Spécificités et composition des répertoires CATH entre RPs unicellulaires et pluricellulaires. Les trois groupes de pluricellulaires sont les Embryophytes, les Phaeophycés et les Metazoaires. Les trois groupes unicellulaires sont les Chlorophytes, les Bacillariophytes et les Choanoflagellés. Leurs répertoires combinés contiennent au total 3236 Homologies différentes. **(A)** Les nombres et couleurs indiquent le nombre et la proportion d'Homologies dans chaque sous-ensemble. **(B)** Proportion d'Architectures dans chaque sous-ensemble. En rouge-jaune, Classe 1 ; en vert, Classe 2 ; en bleu, celle de la Classe 3 ; en rose, la Classe 4 (4.10) ; en gris et noir, la Classe 6.

Les répertoires d'Homologies ont d'abord été comparés entre unicellulaires et pluricellulaires (Figure 46 A). 15% de toutes les Homologies sont propres aux pluricellulaires contre 4% pour les unicellulaires, en partie en lien avec la meilleure représentativité des pluricellulaires dans CATH mais confirmant également l'importance de l'innovation de nouveaux domaines protéiques lors de l'acquisition de la pluricellularité [254]. Les différences qualitatives (Figure 46 B) concernent surtout les Classes 4 et 6, adoptées par une proportion beaucoup plus importante de domaines chez les pluricellulaires que chez les unicellulaires. Comme évoqué précédemment, les protéines ayant des domaines adoptant ces folds sont plus souvent impliquées dans des fonctions de signalisation et de régulation propres aux pluricellulaires [264].

Différentes comparaisons ont ensuite été faites au niveau des répertoires de folds, d'abord entre phyla unicellulaires et phyla pluricellulaire, puis au sein d'une lignée entre unicellulaires et pluricellulaires (Figure 47). Les trois phyla pluricellulaires partagent moins de folds (64%) que les trois unicellulaires (83%) (Figure 47 A). Le repertoire de folds des trois phyla pluricellulaires combinés contient 1017 folds, celui des unicellulaires 941 folds. Dans les pluricellulaires, les Métazoaires ont un nombre spécifique de folds particulièrement élevé (131) par rapport aux autres (entre 51 et 17). Ils sont le groupe avec le plus haut taux d'innovation de domaine au cours de leur évolution mais aussi le groupe vers lequel CATH est le plus biaisé [264]. Concernant les comparaisons entre unicellulaires et pluricellulaires, les résultats sont différents selon les phyla (Figure 47 B). Il y a plus de folds spécifiques aux Chlorophytes (149) qu'aux Embryophytes (22). Le même type de tendance est observé pour les Stramenopiles (73 pour les Diatomés contre 53 pour les Phaeophycés). Dans les deux cas, le nombre de RPs unicellulaires est plus important que celui de pluricellulaires, ce qui explique probablement en partie ces écarts. Au sein des Choanozoés, il y a plus de folds spécifiques aux Métaozaires (120) qu'aux Choanoflagellés (77), ce qui est cohérent avec l'histoire évolutive des Métazoaires.

Ces différences pourraient également être en partie liées au temps de divergence entre lignées pluricellulaires et unicellulaires au sein de chacun des taxa, qui sont différents. La divergence entre Chlorophytes et Embryophytes est la plus ancienne et est estimée entre 1850 et 1600Ma [368]. Celle entre Bacillariophytes et Xantophycés (le groupe frère des Phaeophycés) est daté à d'environ 1000Ma [369], [370]. Enfin, la divergence entre Choanoflagellés et Métazoaires est la plus récente, et a eu lieu entre 937 et 746 Ma [371].

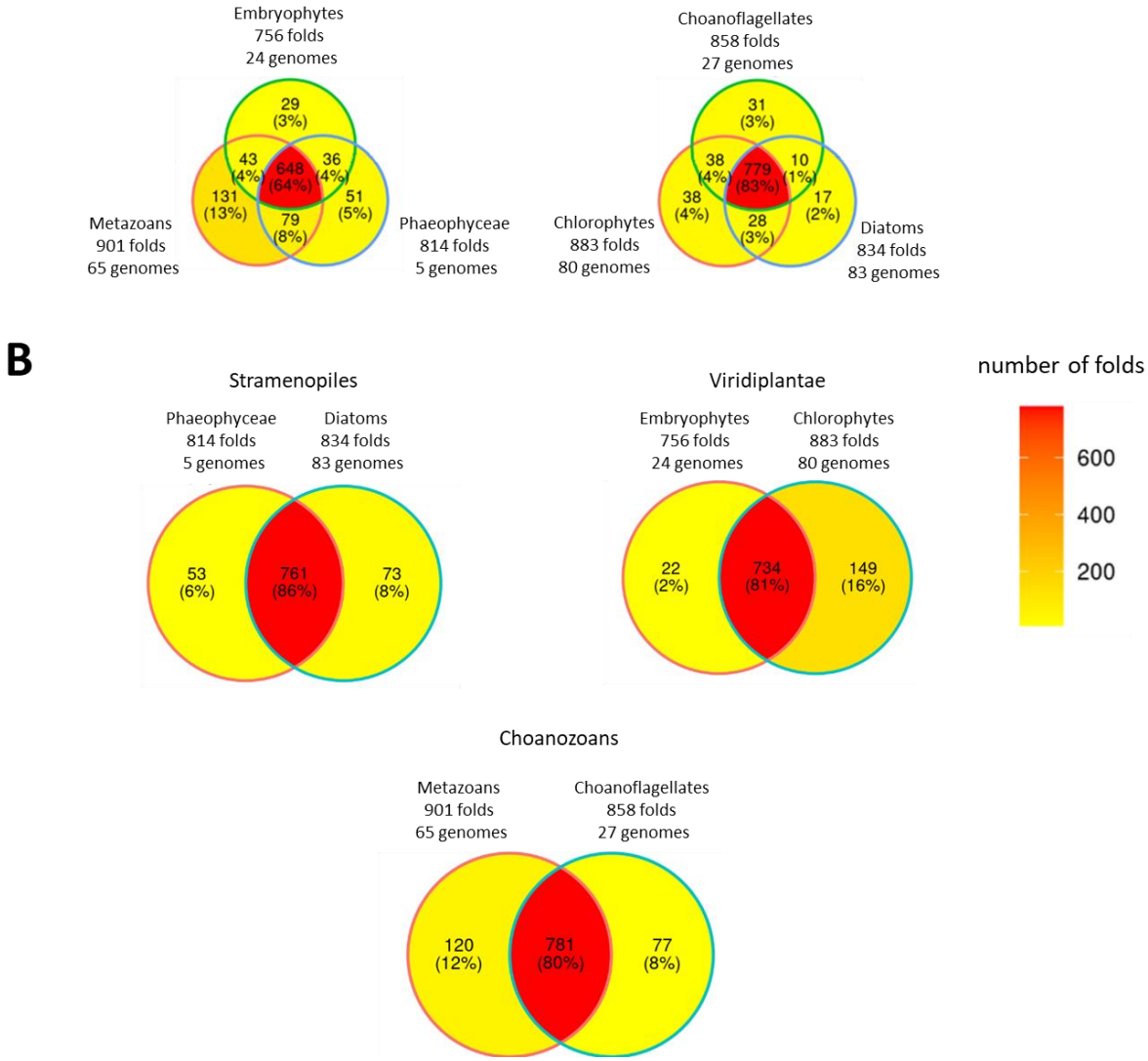


Figure 47. Spécificité des répertoires de folds entre unicellulaires et pluricellulaires. Les nombres et couleurs indiquent le nombre et la proportion de folds dans chaque sous-ensemble. **(A)** Entre phylum pour tous les pluricellulaires (à gauche) et tous les unicellulaires (à droite). Le répertoire de folds des trois pluricellulaires contient 1017 folds, celui des trois unicellulaires 941 folds. **(B)** Entre pluricellulaires (à gauche de chaque diagramme) et unicellulaires (à droite de chaque diagramme) par clade.

Afin d'identifier à une échelle plus fine les Homologies partagées entre les différents taxa pluricellulaires, les occurrences des Homologies partagées par au moins deux phyla ont été représentées (Figure 48). Par définition, les Homologies sont partagées par des groupes monophylétiques ; les pluricellulaires représentant un groupe polyphylétique, il était attendu de ne pas en trouver commune aux quatre. Il y a néanmoins quatre homologies partagées par plus d'un génome dans trois groupes:

- 1.10.437.10 (*Blc2-like*): les Blc2 sont une famille de protéine régulatrice, liée à des fonctions d'apoptose (inhibition ou activation).

- 1.10.10.1180 (*MAN1, winged helix domain*) : MAN1 est une protéine impliquée dans le contrôle de la prolifération cellulaire.

- 4.10.365.10 (*p27*) : p27 est une protéine impliquée dans la régulation du cycle cellulaire, et plus précisément du passage de la phase G1 à S [372]

- 2.90.20.10 (*Plasmodium vivax P25 domain*) : 13 domaines adoptent cette Homologie ; au moins un d'entre eux est spécifique à *Plasmodium vivax* [373]. Ici, elle a les occurrences les plus importantes dans tous les génomes chez qui elle est présente (près de 300 occurrences chez *Hordeum vulgare*).

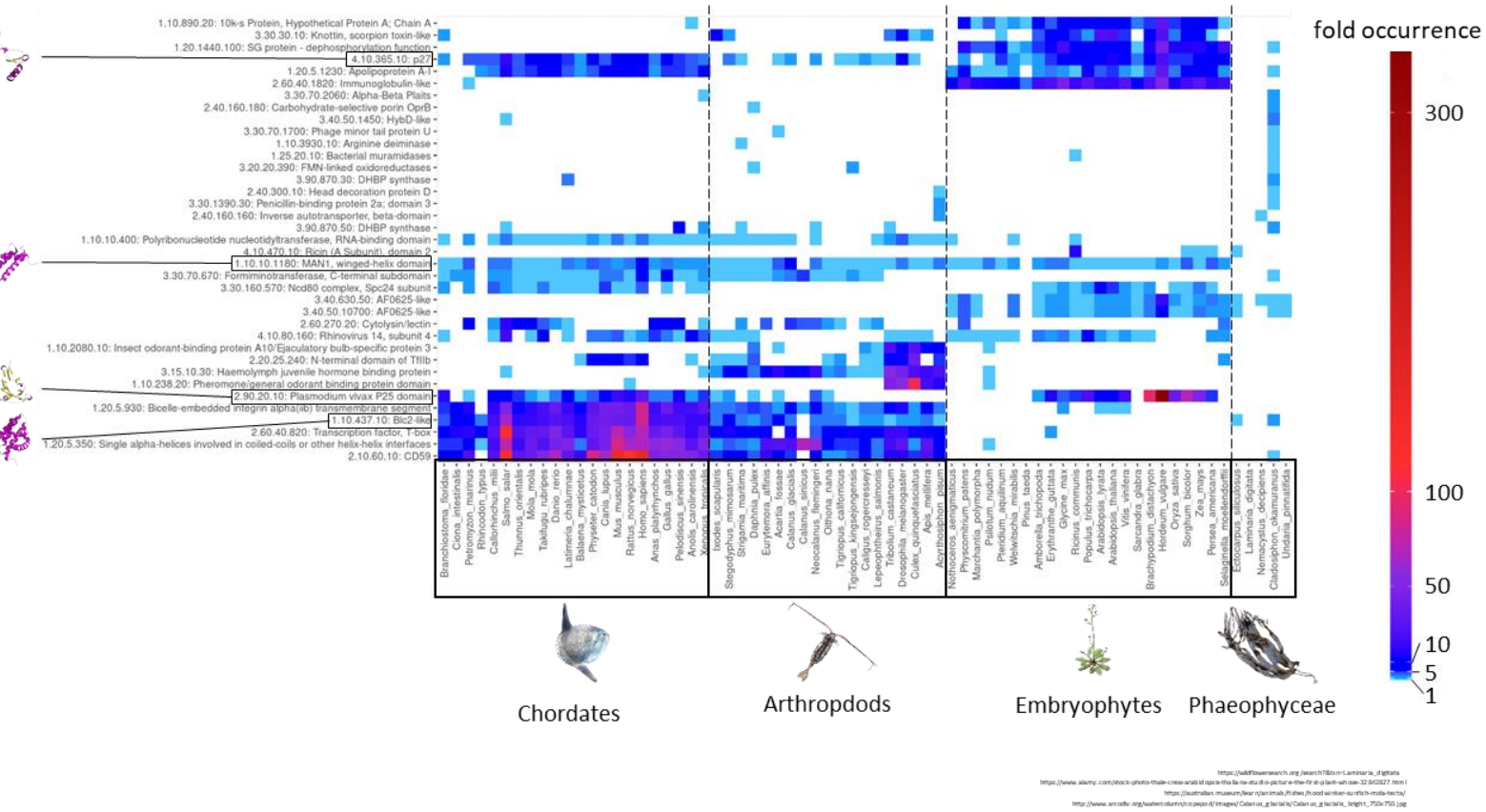


Figure 48. Homologies partagées par au moins deux phyla pluricellulaires. Les cases de la heatmap représentent l'occurrence des 38 Homologies présentes dans au moins deux phyla pluricellulaires. Leurs couleurs indiquent la valeur d'occurrence, comme indiqué à droite de la figure. Les noms et identifiants des CATH correspondants sont indiqués à gauche ; les CATH encadrés et illustrés sont partagés par trois phyla et présent dans pratiquement tous les protéomes de ces phyla. Les espèces auxquelles appartiennent les protéomes sont indiquées en bas et classées par phylum.

6/ conclusion

Dans ce chapitre, trois questions ont été posées. La première était « *La qualité des MAGs est-elle suffisante pour une étude à l'échelle des folds ?* ». Pour y répondre, j'ai comparé les taux d'annotation CATH dans les quatres domaines du vivant chez les MAGs et pour des protéomes Eucaryotes de référence. Le taux d'annotation CATH moyen est de 30%, avec une certaine variabilité selon les domaines. Il est aussi globalement plus élevé pour les protéomes de référence que pour les MAGs chez les Eucaryotes. La qualité des MAGs est néanmoins suffisante pour nos analyses.

La question suivante était « *Quelles sont les caractéristiques des foldomes des génomes environnementaux et des protéomes de référence ?* ». J'ai d'abord vérifié que les foldomes des MAGs Eucaryotes et des protéomes de référence étaient globalement similaires. Ensuite, grâce à la diversité taxonomique représentée par les MAGs, j'ai pu constater que les Eucaryotes avaient des foldomes relativement différents de ceux des Procaryotes et des *Nucleocytoviricota*, notamment du point de vue de la proportion des cinq Classes de CATH dans ces différents domaines. Les folds de la Classe 3 représentent en effet une part moins importante des occurrences chez les Eucaryotes que chez les autres. À une échelle taxonomique plus fine, des différences sont aussi observées chez les Eucaryotes, notamment entre les Métazoaires et les autres groupes. Enfin, j'ai vérifié que les répertoires de folds des MAGs permettaient globalement bien de les classer en cohérence avec leur taxonomie, une propriété bien connue pour des analyses faites sur des génomes de référence.

La dernière question de ce chapitre était « *Existe-t-il un effet de la pluricellularité sur la distribution des folds dans les foldomes ? Si oui, peut-il être quantifié ?* ». Pour répondre à cette question, j'ai utilisé les protéomes de référence pour comparer les foldomes de quatre groupes de pluricellulaires avec ceux d'unicellulaires taxonomiquement proches. Cela a montré que les répertoires d'unicellulaires et pluricellulaires proches possédaient chacun un nombre important de folds spécifiques. Enfin, il n'existe pas de folds spécifiques et universels de pluricellulaires, en tout cas dans notre jeu de données.

Dans l'ensemble, ce chapitre aura permis de valider l'utilisation des MAGs pour une étude à l'échelle des folds et aura montré que les répertoires de certains phyla Eucaryotes peu représentés dans les bases de données de références ont des propriétés interessantes. Il aura également ouvert des pistes pour une étude approfondie du lien entre pluricellularité et foldome, qui nécessitera un jeu de données plus complet et plus exhaustif pour étudier le phénomène en profondeur.

CHAPITRE 2. MODÉLISATION DE LA DISTRIBUTION DES FOLDS DANS LES PROTÉOMES ET DANS LES COMMUNAUTÉS PLANCTONIQUES

Sommaire

L'usage des folds dans les génomes se fait de telle façon que quelques superfolds sont adoptés par une très grande majorité de domaine, et une grande diversité de folds ne sont adoptés que par un seul ou quelques domaines. La distribution des occurrences des folds dans les protéomes peut donc être modélisée par une loi de puissance. La pertinence de ce type de modèle a jusqu'ici été validée sur des organismes de référence [254], [258], [261]. La distribution des folds dans l'environnement n'a quant à elle jamais été étudiée.

Dans ce contexte, les objectifs de ce chapitre sont multiples. Il s'agira dans un premier temps d'étudier la distribution des folds dans les foldomes des MAGs, et de les comparer avec ceux de protéomes de référence. La question principale est :

- *Le modèle de loi puissance de la distribution des folds dans les foldomes est-il pertinent dans le cas de génomes environnementaux incomplets ? À quel point est-il universel dans l'arbre du vivant ?*

Une fois le modèle de loi puissance testé dans les génomes, il sera appliqué à la distribution des folds dans les communautés planctoniques avec la question :

- *Le modèle de loi puissance est-il pertinent à l'échelle d'une communauté de protéomes d'espèces en interaction avec des dynamiques écologiques propres aux communautés planctoniques ?*

Pour répondre à ces questions, je commencerai par tester le modèle de loi puissance dans les foldomes des MAGs au niveau des domaines du vivant et des phyla. Je comparerai les résultats sur les phyla avec des modèles sur les foldomes des RPs. Enfin, je comparerai les modèles entre unicellulaires et pluricellulaires des RPs pour émettre des hypothèses concernant l'effet du taux de duplication des gènes sur les paramètres de la loi puissance.

Dans un deuxième temps, j'utiliserai la loi de puissance pour modéliser la distribution des abondances des folds dans les stations TO chez les Bactéries et les Eucaryotes, et vérifierai la pertinence de ce modèle en le comparant avec une autre loi.

1/ modélisation de la distribution des occurrences des folds dans les foldomes : universalité de la validité de la loi puissance et impact des duplications de gènes

a. comparaison des modèles pour les Eucaryotes entre MAGs et RPs

Pour vérifier que la complétion des MAGs n'empêche pas leur utilisation pour étudier les propriétés des distributions des occurrences de folds dans les foldomes, les modèles appliqués aux MAGs Eucaryotes ont dans un premier temps été comparés avec ceux des RPs (Fig.49).

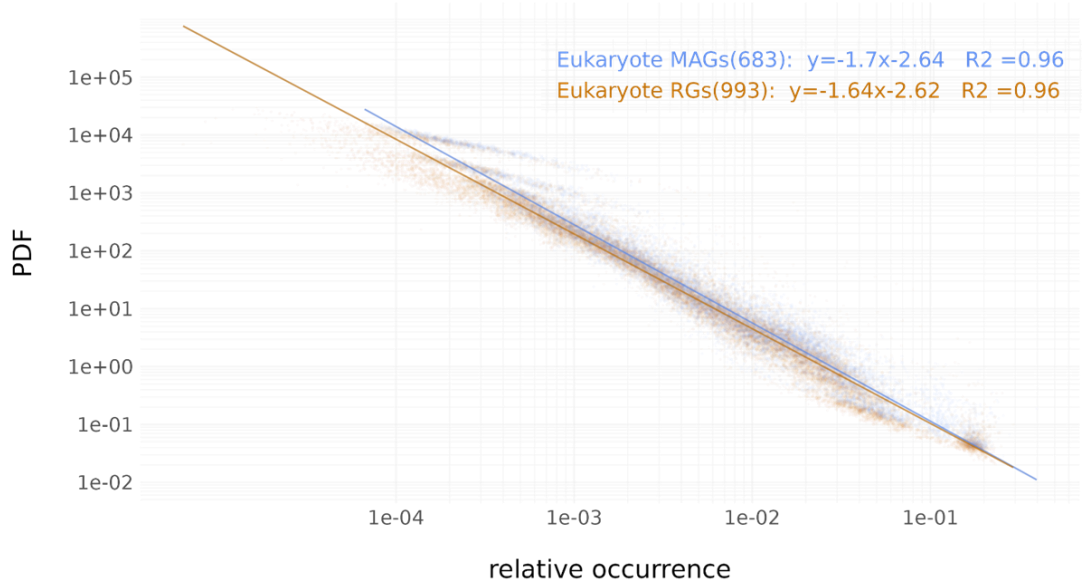


Figure 49. Modèles de loi puissance de la distribution des occurrences des folds dans les MAGs et les RPs Eucaryotes. Les MAGs sont en bleu, les RPs en orange. Chaque point représente un bin de valeurs d'occurrence relative dans un foldome. Les droites sont encadrées par des zones grises indiquant l'écart type. Les coefficients directeurs des droites ainsi que les coefficients de régressions sont indiqués en haut à droite de la figure. Les axes sont en échelle logarithmique. Chaque modèle de régression est significatif avec une *p*-value inférieure à 0.01.

Les R^2 associés aux deux distributions sont de 0.96, indiquant la pertinence de la loi de puissance pour les modéliser. Les deux droites sont très proches, ce qui confirme que l'utilisation des foldomes de génomes incomplets est globalement pertinente pour ce type d'étude. La complétion plus ou moins importante de chaque MAG n'est pas à l'origine d'une déviation par rapport au modèle de la loi puissance car ces lois ont une propriété d'invariance d'échelle, qui implique qu'un échantillon tiré au hasard dans un ensemble suivant une loi puissance suivra la même loi puissance à une constante près. Ici même si la reconstruction des génomes et l'annotation structurale n'est pas équivalente à une sélection aléatoire de gènes au sein d'un génome, l'effet des différentes étapes de réduction (complétion des MAGs et taux d'annotation CATH) ne suffit pas à invalider la propriété d'invariance d'échelle.

Pour vérifier si la similarité importante entre droite de régressions de la Fig.49 entre MAGs et RPs résulte d'un effet global observé dans tous les taxa ou non, les mêmes modèles ont été appliqués au foldomes des différents phyla (Figure 50).

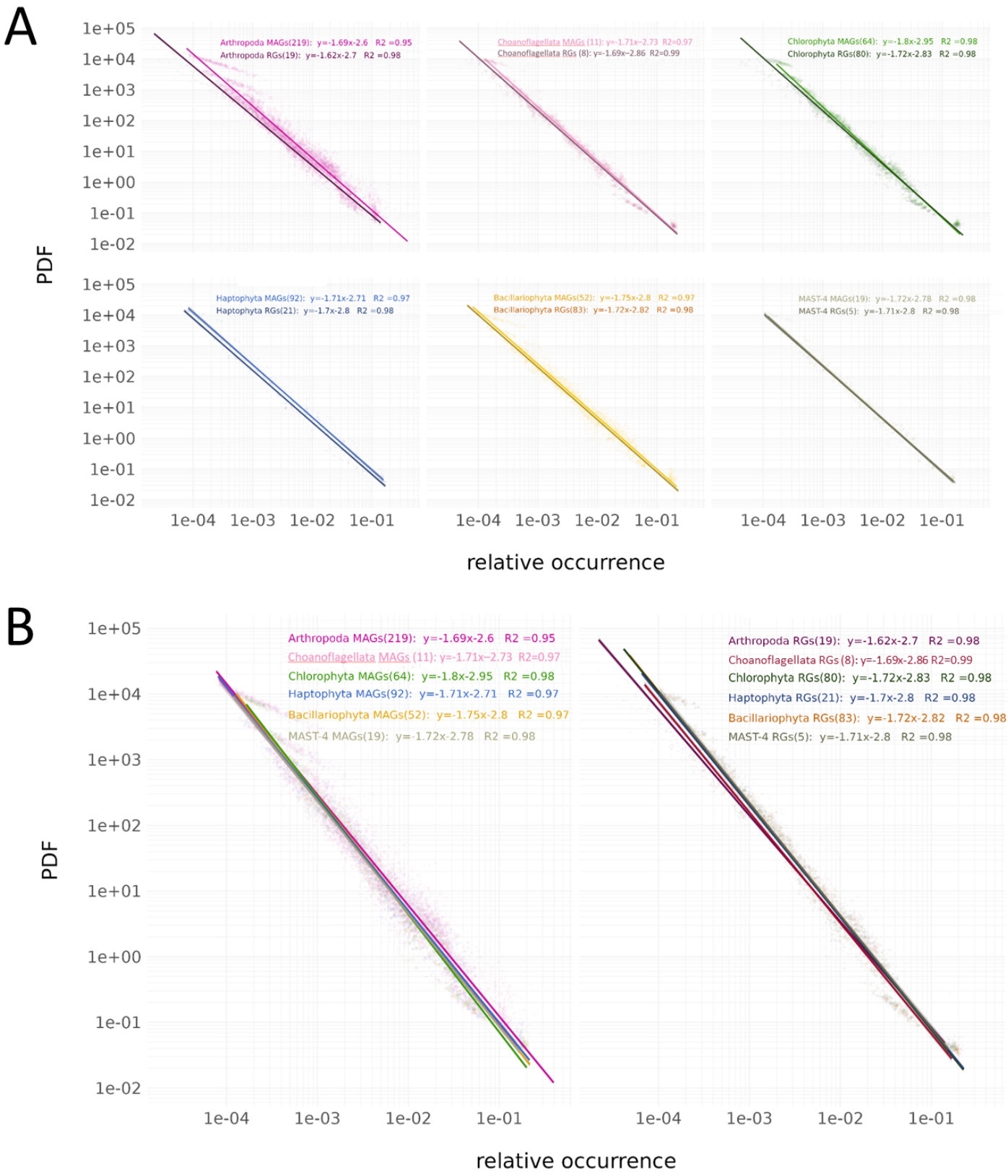


Figure 50. Comparaison des modèles de loi puissance par phylum entre MAGs et RPs. Chaque point représente un bin d'occurrence relative pour un foldome, et sa couleur correspond à un phylum. Ils sont représentés sur une double échelle logarthmique. Tous les modèles sont significatifs au niveau de confiance 0.01. **(A)** Comparaisons pour les six phyla sélectionnés (de gauche à droite et haut en bas : Arthropodes, Choanoflagellés, Chlorophytes, Haptophytes, Bacillariophytes, MAST-4). **(B)** Comparaisons des modèles pour les six phyla entre eux, avec les MAGs à gauche et les RPs à droite.

Pour cette analyse, seuls six phyla ont été sélectionnés en raison de la complétion moyenne des MAGs de ces phyla et le nombre moyen de MAGs de ces phyla par station TO. De façon globale, les coefficients directeurs des droites en valeur absolue sont plus élevés avec les MAGs qu'avec les RPs, entrainant une légère divergence des droites pour les valeurs faibles d'occurrences relatives (Figure 50 A). Dans cette gamme d'occurrences, à valeur d'occurrence relative fixée, il y a donc plus de folds ayant cette valeur dans les MAGs que les RPs. Cela résulte probablement de l'incomplétion des MAGs : si dans un MAG un nombre important de gènes codant pour un fold donné est manquante en comparaison des RPs uniquement parce que ce MAG est incomplet, alors le fold en question se retrouvera dans un bin d'occurrence plus bas que dans les RPs. Les différences les plus importantes de coefficients directeurs entre MAGs et RPs sont observées chez les Arthropodes (MAGs : -1.69 ; RPs : -1.62), les Chlorophytes (MAGs : -1.8 ; RPs : -1.72) et dans une moindre mesure les Bacillariophytes (MAGs : -1.75 ; RPs : -1.72). Les différences pour les Arthropodes proviennent encore une fois probablement des différences de diversité taxonomique dans les deux bases de données (les MAGs ne contiennent que des Copépodes alors les RPs sont représentatifs de toute la diversité du phylum). Les MAST-4 sont le seul phylum pour lequel l'équation de régression est pratiquement la même entre MAGs et RPs. Cela s'explique peut-être en partie par le fait que ce phylum est le seul pour lequel il y a plus de MAGs que de RPs, et que le répertoire de folds des premiers est plus diversifié que celui des seconds (Figure 38). Les écarts entre modèles au sein d'un phylum ne sont probablement pas causés par les différences de taux d'annotation CATH entre MAGs et RPs observés dans la Fig.37. Il n'y avait en effet pas de différence significative de taux d'annotation entre MAGs et RPs pour les Choanoflagellés, les Chlorophytes et les Bacillariophytes, or il y a des différences dans les équations de régressions de ces trois groupes. Inversement, les MAST-4 ont des différences significatives de taux d'annotation CATH entre MAGs et RPs alors que les équations de régression de leurs modèles sont quasiment identiques. Concernant les comparaisons des droites de régressions entre phyla au sein d'une base de donnée, les résultats sont sensiblement différents selon si l'on s'intéresse aux MAGs ou au RPs (Figure 50 B). Avec les RPs, les différences entre pluricellulaires, Unicontes et Bicontes sont visibles. Les Arthropodes, pluricellulaires, ont le coefficient directeur le plus bas en valeur absolue, suivi par les Choanoflagellés (groupe le plus proche du point de vu phylogénétique). Enfin, les Bicontes ont des coefficients directeurs très similaires entre eux, plus élevés en valeur absolue que ceux des Unicontes. Ces tendances sont moins évidentes à visualiser avec les MAGs, bien que les Arthropodes y aient également le plus petit coefficient directeur mais avec un plus petit écart.

De façon générale, Les écarts observés entre droites autant dans la Fig.49 que dans la Fig.50 B pourraient résulter de différences de taux de duplication de gènes. Dans la Fig.49, les RPs contiennent des Vertébrés qui ont connus d'importantes duplications de gènes au cours de leur histoire évolutive [374]. Ces duplications peuvent avoir des impacts particulièrement importants sur le coefficient directeur de la droite de régression. Par exemple, si une famille de gènes sans homologue dans CATH est fortement dupliquée, aucun de ses gènes n'aura d'annotation CATH, ce qui peut résulter en un biais important sur les valeurs d'occurrence. Dans la Fig.50 B pour les RPs, le taux de duplication différentiel entre phyla est probablement aussi à l'origine des variations de coefficients directeurs. Les Arthropodes ayant le plus petit en valeur absolue, le taux de duplication de leurs gènes a probablement été le plus élevé des six phyla de l'analyse. Cette hypothèse est aussi appuyée par le fait qu'il s'agit du seul phylum pluricellulaire de cette analyse. Chez les MAGs, l'effet des duplications est en partie masqué par la complétion. Il est plus délicat d'interpréter la valeur de l'ordonnée à l'origine, qui doit être impactée par la valeur minimale d'occurrence relative et sa variation entre les génomes. En théorie, cette valeur devrait être la même pour tous, c'est-à-dire 1 ; ici en revanche, les données sont rendues continues pour permettre de comparer les génomes entre eux indépendamment de leur complétion, rendant impossible l'interprétation des différences de ce paramètre.

b. modèles de distribution dans les différents domaines du vivant

Après avoir observé l'effet de la complétion des MAGs sur les modèles de distribution des folds dans les foldomes, l'importance de cet effet a été vérifiée au niveau des domaines du vivant avec les MAGs (Figure 51).

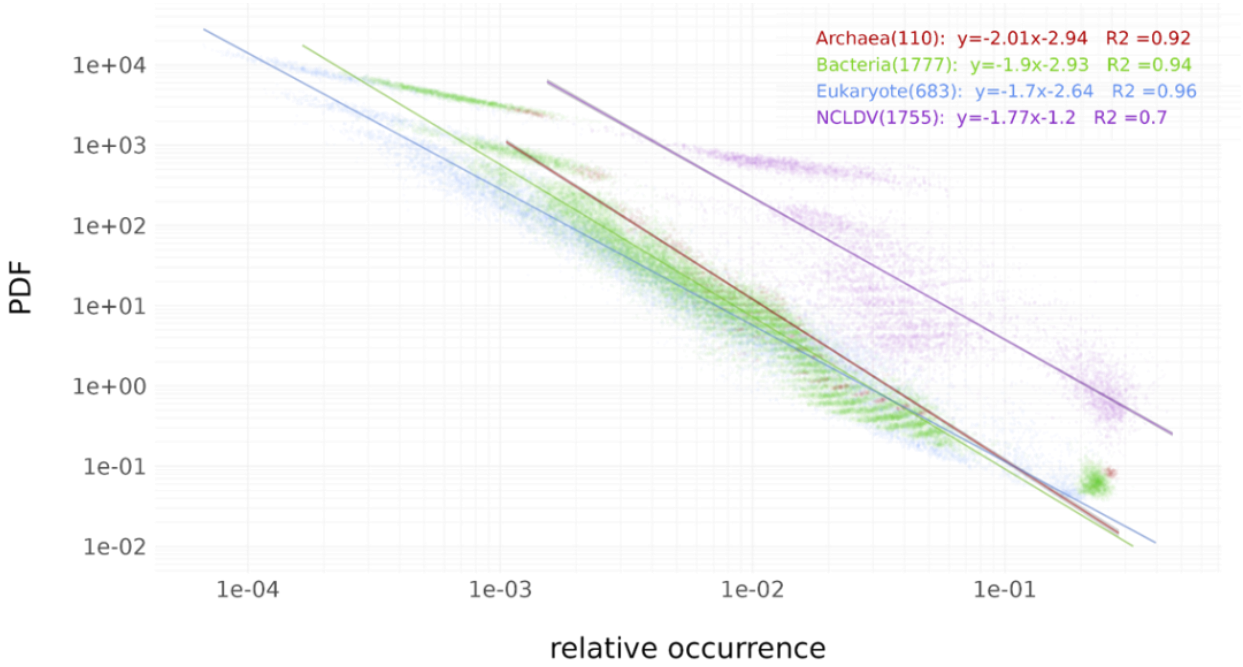


Figure 51. Modèles de loi puissance des occurrences à l'échelle des domaines du vivant dans les MAGs. Chaque point représente un bin d'occurrence relative pour un foldome. Ils sont colorés, ainsi que les droites de régression, en fonction du domaine du vivant auquel ils correspondent. Ils sont représentés sur une double échelle logarithmique. Le nom du domaine et le nombre de MAG entre parenthèses, l'équation de régression et le R^2 sont indiqués en haut à droite. Toutes les régressions sont significatives au seuil de confiance 1%.

La distribution des occurrences chez les Procaryotes peut être correctement modélisée par une loi de puissance, en cohérence avec des résultats d'études faites sur des génomes de référence [254], [255], [258]. Ici, le coefficient directeur du modèle des Bactéries est plus proche de celui des Archées que des Eucaryotes. Des résultats sur la distribution de la taille des familles de domaines ont montré une plus grande proximité de coefficients directeurs entre Bactéries et Eucaryotes, qu'entre Bactéries et Archées (Figure 24) [254]. Le R^2 du modèle pour les NCLDVs est moins satisfaisant que celui pour les autres domaines du vivant, ce qui est peut-être une conséquence des valeurs extrêmes dz taux d'annotation CATH observées dans la Fig.35. Il est donc probable que la distribution des folds dans ce groupe suive bien une loi de puissance, comme cela a déjà été observé dans d'autres groupes de Virus [254].

En excluant les NCLDVs, les Archées ont le coefficient directeur le plus élevé en valeur absolue, suivies par les Bactéries puis les Eucaryotes. Ces différences vont dans le sens d'une importance du taux de duplication de gènes sur le coefficient de régression, ce qui avait été émis comme hypothèse pour expliquer les différences entre MAGs et RPs chez les Eucaryotes plus haut (Figure 50). D'un point de vue théorique, cet effet s'explique par la façon dont les occurrences des folds évoluent dans les génomes. Le mode d'évolution principal est le modèle neutre, dans lequel les duplications de gènes

ont plus de chance de survenir sur un gène dont il existe déjà un grand nombre de copies, donc qui a déjà subi de nombreuses duplications. Ces gènes fortement dupliqués codent généralement pour des protéines ayant au moins un domaine adoptant un superfold, comme le Rossmann fold. Ainsi plus les évènements de duplication ont été fréquents dans un génome et plus le delta d'occurrence entre le fold le plus fréquent et les folds adoptés uniquement par quelques domaines ou un seul (unifolds et mesofolds) est important. Plus ce delta est grand, plus l'écart sur l'axe des abscisses entre le fold le plus fréquent et ces folds est grand, résultant en un coefficient directeur plus faible de la droite de régression [254], [258].

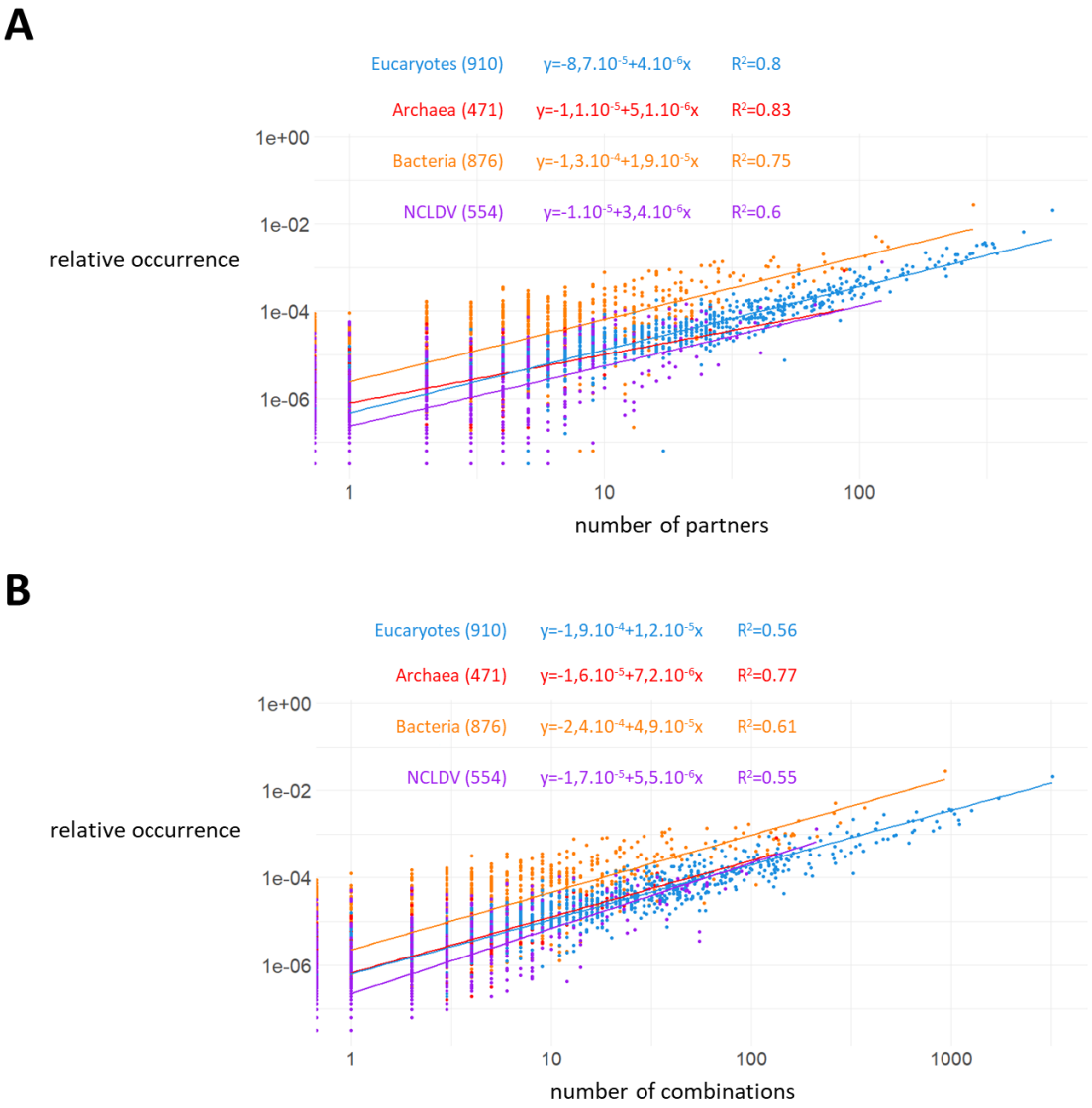


Figure 52. Relation entre occurrence relative et propriétés de connectivité des folds par domaine du vivant. Chaque point représente un fold dans un génome environnemental et est coloré en fonction du domaine du vivant. Les résultats des modèles linéaires sont indiqués en haut de chaque figure, avec de gauche à droite le nom du domaine du vivant et le nombre de folds, les équations de régression et les R^2 associés. Chaque régression est significative avec une *p*-value < 0.01. **(A)** Relation entre occurrence relative et nombre de partenaires (folds différents avec lesquels le fold est trouvé en combinaison dans un génome donné) par folds. **(B)** Relation entre occurrence relative et nombre de combinaisons par folds.

Afin de tester la pertinence du modèle de loi puissances sur d'autres grandeurs caractéristiques associées aux folds, celle-ci a dans un deuxième temps été utilisée pour modéliser les relations entre occurrence relative de chaque fold dans un foldome et nombre de combinaisons ou de partenaires (Figure 52). Le nombre de partenaires correspond au nombre de folds différents observés en combinaison avec le fold en question dans toutes les protéines multidomaines où le fold est présent (Figure 52 A). Il est moins élevé que le nombre de combinaisons puisque beaucoup de combinaisons impliquent des folds courants (Figure 52 B). Ces deux grandeurs permettent de décrire en partie les réseaux d'interactions formés par les folds, qui sont principalement organisés autour du principe d'attachement préférentiel. Les modéliser permet donc dans une certaine mesure de quantifier l'intensité de ce phénomène et surtout de mettre en évidence de potentielles différences entre domaines du vivant.

Les R^2 des modèles sont compris entre 0.83 et 0.55, et sont globalement meilleurs pour les nombres de partenaires que de combinaisons (R^2 moyen à 0.6225 contre 0.745 pour les nombres de combinaisons et de partenaires, respectivement). Les différences de qualité sont domaines-spécifiques : les moins bons modèles sont pour les NCLDVs et les meilleurs pour les Archées dans les deux cas. Les Eucaryote sont le domaine du vivant avec les folds ayant les plus grands nombres de partenaires et de combinaisons, comme attendu, suivi par les Bactéries. Les résultats ici sont cependant probablement fortement biaisés par la complétion et le taux d'annotation CATH. En effet, le nombre de partenaires d'un fold étant plus limité que le nombre de combinaisons dans lequel il est observé, il y a plus de chance dans un protéome incomplet d'être proche du nombre véritable de partenaires que du nombre véritable de combinaisons pour chaque folds. Cela coïncide avec les écarts à la loi puissance qui sont plus élevés pour les nombres de combinaisons que de partenaires. Il est en revanche surprenant d'observer de meilleurs R^2 pour les Eucaryotes que pour les Bactéries en terme de nombre de partenaires, d'autant plus que les Bactéries ont un meilleur R^2 pour le nombre de combinaisons et que les Bactéries ont une bien meilleure complétion moyenne et une meilleur taux d'annotation CATH (Fig.35). Cela est possiblement lié au fait que les Eucaryotes sont le domaine avec le plus de combinaisons de folds uniques mais qui impliquent dans la majorité des cas des folds universels [214]. Dans ce domaine du vivant, le nombre de partenaires uniques augmente potentiellement moins vite avec la complétion et les occurrences relatives que le nombre de combinaisons uniques, entraînant le décrochage observé pour les nombres de combinaisons par rapport aux Bactéries. En outre, les écarts pourraient également résulter de différences dans les bases de données concernant ces deux domaines du vivant. Il est par exemple imaginable que les combinaisons des Eucaryotes, en moyenne plus longues que celles des Bactéries, aient moins de chance d'être annotées structuralement par CATH, mais que la diversité de partenaires associés à un fold est représentée dans des combinaisons plus courtes et plus faciles à annoter. Dans l'ensemble, il n'est pas possible de dire si la relation entre occurrences et nombre de partenaires et de combinaisons est universellement de type loi de puissance.

c. effets du taux différentiel de duplication de gènes entre phyla eucaryotes sur les modèles de distribution

Dans la Fig.50 B, des différences entre modèles de distribution des occurrences des Arthropodes et des autres phyla ont été constatées, et s'expliquent probablement par des différences de taux de duplication de gènes globales entre pluricellulaires et unicellulaires. Afin de vérifier l'universalité de ce résultat, d'autres paires de phyla unicellulaires et pluricellulaires ont été sélectionnées. Ce sont les mêmes que celles de « différences de répertoires de folds entre unicellulaires et pluricellulaires » (p.122).

Des modèles de loi puissance ont donc été appliqués aux distributions des folds dans les RPs des Phaeophycés, Bacillariophytes, Embryophytes, Chlorophytes, Chordés, Arthropodes et Choanoflagellés (Figure 53). Les équations de régression obtenues se révèlent être dans l'ensemble cohérentes avec la division entre unicellulaires et pluricellulaires, à l'exception des Phaeophycés dont la droite de régression n'est pas significativement plus proche de celle des unicellulaires ou des pluricellulaires. Cela est probablement la conséquence du faible nombre de RPs dans ce phylum. Les Embryophytes et les Chordés ont des droites de régression très proches, avec les plus faibles coefficients directeurs en valeur absolue. De façon surprenante, le coefficient directeur du modèle pour les Arthropodes est plus élevé que celui des Embryophytes et des Chordés, alors que, étant plus proche taxonomiquement des Chordés, il aurait été attendu que leurs deux modèles soient les plus similaires. Les coefficients de régression des modèles des unicellulaires sont les plus élevés en valeur absolue et sont très proches, en particulier pour les Chlorophytes et les Bacillariophytes. La distribution des occurrences relatives des folds est donc impactée par la pluricellularité, qui entraîne globalement une baisse du coefficient directeur de la droite de régression du modèle. Dans la continuité des hypothèses émises précédemment, cela indiquerait que les taux de duplication de gènes pourraient être plus élevés chez les pluricellulaires que les unicellulaires. Les duplications de gènes et les duplications de génomes entiers précèdent souvent la diversification des lignées, et sont suivies de phases de réductions fonction-spécifiques. L'importance des fonctions de régulations chez les pluricellulaires génère une force qui limite la contre-sélection de gènes dupliqués, qui vont avoir tendance à créer une nouvelle fonction de régulation plutôt que disparaître [375]. À long terme, le taux de duplication de leurs gènes pourrait donc apparaître comme plus élevé, en partie par ce mécanisme. Les folds, intermédiaire entre gènes et fonctions, peuvent donc apporter un éclairage supplémentaire sur les grandes questions sur de l'évolution des génomes.

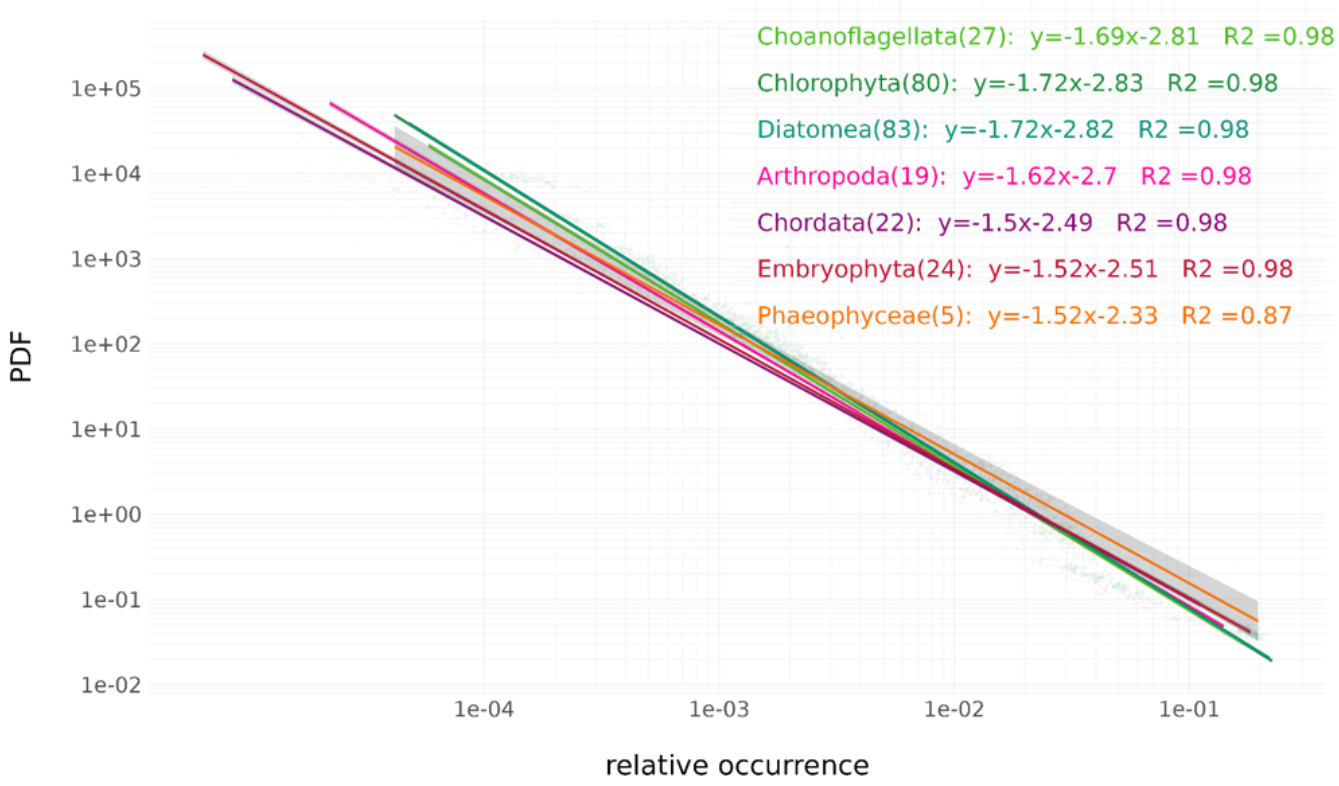


Figure 53. Comparaison des modèles de loi puissance de la distribution des folds dans les foldomes de phyla unicellulaires et pluricellulaires. Chaque point représente un bin d'occurrence relative dans un génome et est représenté sur une double échelle logarithmique. Dans le coin en haut à droite, pour chaque phylum : nom du phylum et nombre de génomes entre parenthèses, équation de régression et R^2. Les phyla en vert sont unicellulaires et ceux allant de l'orange au rose sont pluricellulaires. Les droites de régressions ssont encadrées par une zone grise qui indique l'erreur standard. Chaque modèle est significatif au seuil de confiance de 1%.

La relation entre occurrences et nombre de combinaisons et de partenaires a ensuite été modélisée avec la loi puissance sur les mêmes phyla (Figure 54).

Contrairement aux observations faites dans la Fig.52, le modèle de loi puissance est de bien meilleure qualité pour la relation avec le nombre de combinaisons (Figure 54 B) plutôt qu'avec le nombre de partenaires (Figure 54 A). Il semblerait donc que l'incomplétion et les taux d'annotation CATH aient biaisé les résultats de la Fig.52, et qu'ils affectent plus le nombre de combinaisons que le nombre de partenaires.

Ici, au vu des R^2, il est possible de dire que le nombre de combinaisons augmente à la puissance avec le nombre d'occurrences chez les pluricellulaires, alors que l'augmentation du nombre de partenaire ne suit de loi de puissance ni chez les unicellulaires ni chez les pluricellulaires.

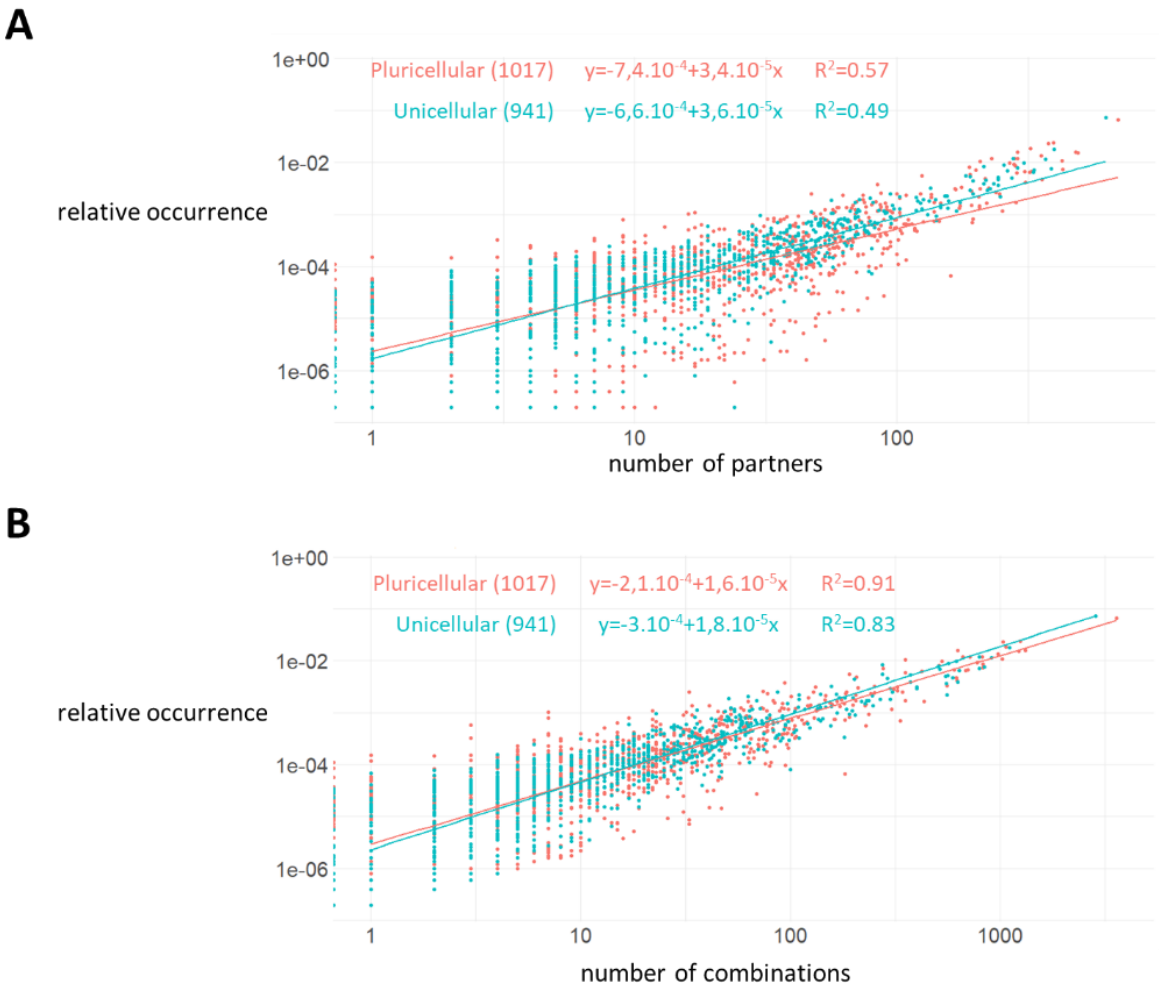


Figure 54. Relation entre occurrence relative et métriques quantifiant la capacité de combinaison des folds entre unicellulaires et pluricellulaires. Chaque point représente un fold dans un RP et est coloré selon si ce RP est unicellulaire ou pluricellulaire. Les résultats des régressions linéaires sont indiqués en haut de chaque figure, avec de gauche à droite la catégorie et le nombre de folds entre parenthèses, les équations de régression et les R^2 associés. Chaque régression est significative au seuil de confiance de 1%. **(A)** Relation entre occurrence relative et nombre de partenaires (folds différents avec lesquels le fold est trouvé en combinaison dans un génome donné) par folds. **(B)** Relation entre occurrence relative et nombre de combinaisons par folds.

Sous l'hypothèse que la variabilité des modèles de loi puissance est essentiellement causée par des différences de taux de duplications de gènes, la diversité des MAGs a été utilisée pour évaluer la variabilité de ce taux entre différents phyla Bactérien (Figure 55).

Dans l'ensemble, les modèles sont satisfaisants (R^2 toujours supérieur à 0.92). Les phyla les plus divergents du point de vue de l'équation de leur droite de régression sont les Marinisomatota qui ont le plus grand coefficient de régression en valeur absolue et les Myxococcota, qui ont le plus petit. Les Myxococcota pourraient avoir eu un taux de duplication de gènes particulièrement importants au cours de leur évolution, alors que les Marinisomatota auraient au contraire eu les taux les plus faibles ou subi des évènements de réductions de génomes plus importants que les autres phyla Bactérien. Tous les autres phyla auraient eu des taux de duplication relativement similaires au cours de leur évolution.

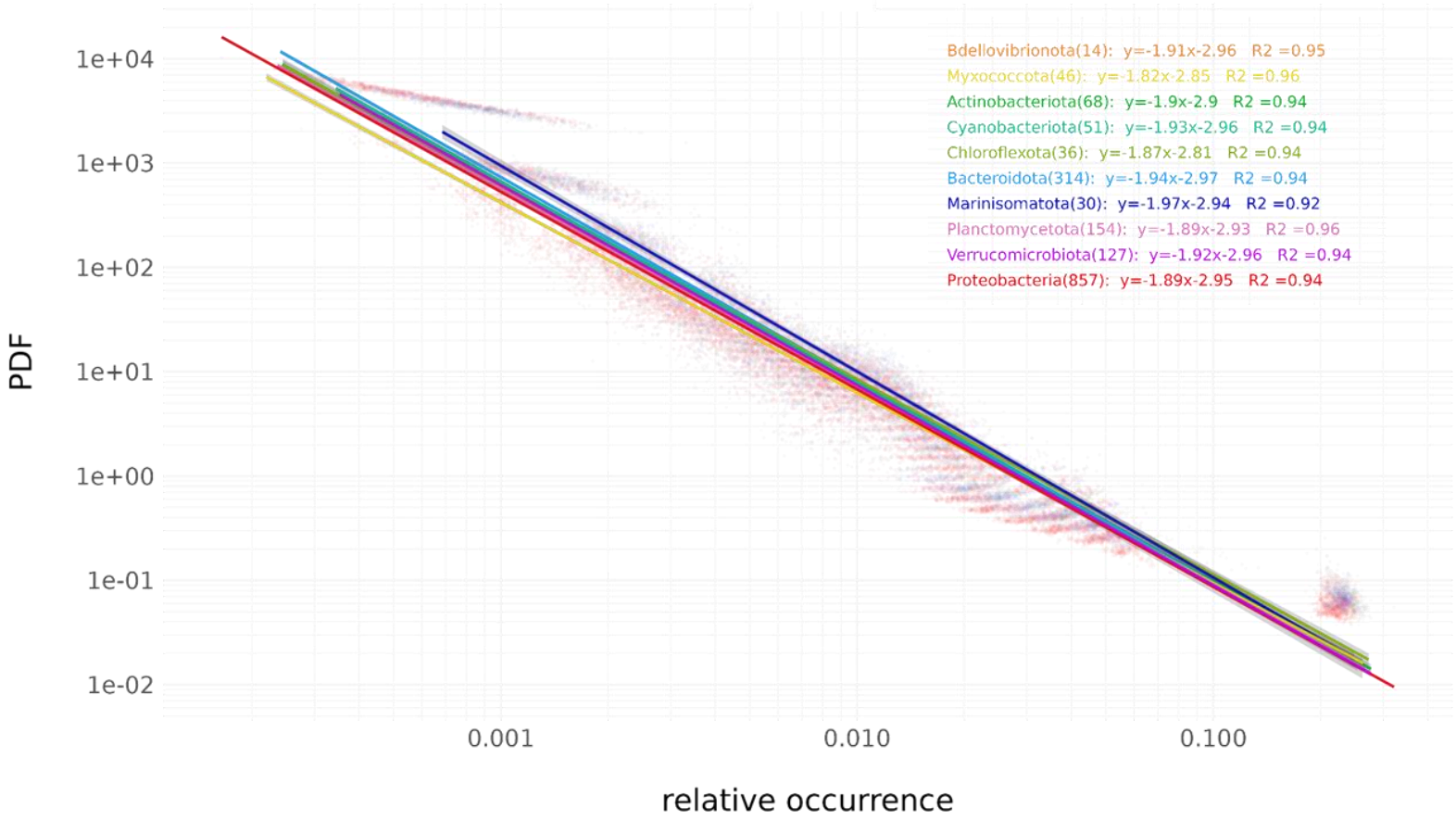


Figure 55. Modèles de loi puissance pour les distributions des occurrences de folds dans les différents phyla bactériens des MAGs. Chaque point correspond à un bin d'OV dans un foldome et est coloré selon le phylum auquel le foldome appartient. La couleur associée à chaque phylum est indiquée en haut à droite. Les points sont représentés sur une double échelle logarithmique. Chaque droite est encadrée par une zone grise indiquant l'écart type. Toutes les modèles sont significatifs au seuil de confiance 1%.

2/ modélisation des abondances des folds dans l'Océan : validité de la loi de Pareto II comme une propriété émergente des communautés planctoniques

Il a été montré dans la partie précédente que la distribution des folds dans les foldomes des MAGs suivait bien une loi de puissance, malgré leurs complétions variables. La validation de cette propriété permet maintenant d'utiliser les MAGs pour étudier la distribution des folds dans l'environnement, grâce à la combinaison des valeurs d'abondance des MAGs dans les stations TO et d'occurrence des folds. L'enjeu principal ici est donc d'étudier la distribution des abondances des folds dans l'environnement et d'évaluer si l'universalité de la loi puissance se maintient à cette échelle.

a. distribution globale des abondance des folds dans les stations TO

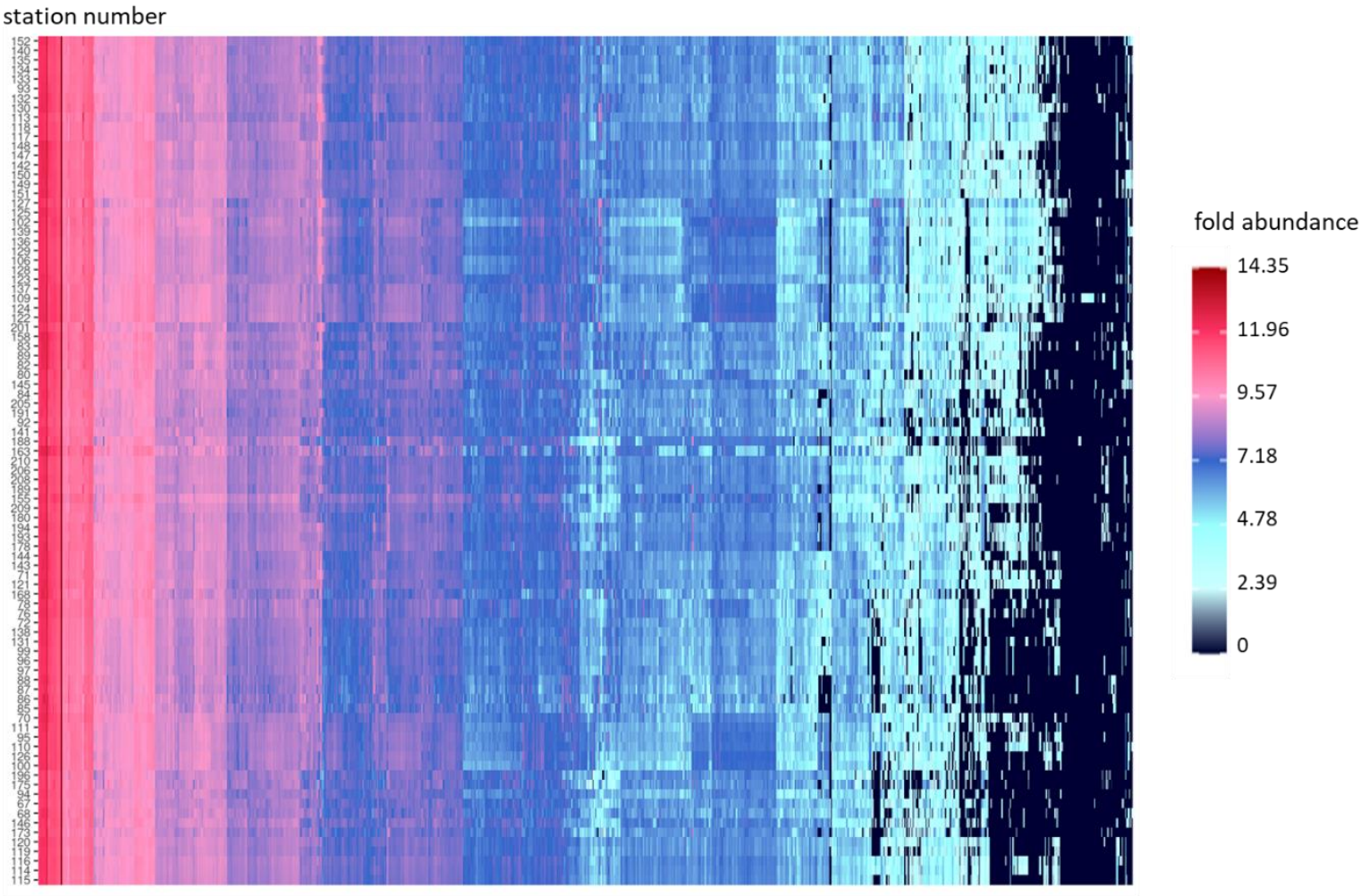


Figure 56. Distribution des folds dans les stations TO. AVs transformées en CLR des 910 folds eucaryotes dans les échantillons de la fraction de taille 0.8-2000µm et de surface. Les couleurs représentent les AVs, l'échelle est indiquée à droite. Les stations sont sur l'axe des ordonnées (numéros de stations à gauche), les folds sur l'axe des abscisses. L'ordre des stations est celui de leur clustering qui est basé sur les AVs des folds. Les folds sont ordonnés selon le clustering réciproque, dont les clusters principaux ont été réordonnés manuellement pour la lisibilité.

La première étape était de visualiser l'abondance de tous les folds dans toutes les stations TO (Figure 56). Globalement, au moins trois catégories de folds semblent se distribuer différemment dans les stations. Un certain nombre de folds sont très abondants partout (tout à gauche de la figure). Parmi eux sont présents le Rossmann fold, (3.40.50) et le 3.30.160 (Double Stranded RNA Binding Domain, dont l'Homologie 3.30.160.60 *Classic Zing Finger* est celle étant adoptée par le plus de domaines chez les Arthropodes alors que c'est la 3.40.50.300 *P-loop containing nucleotide triphosphate hydrolases* dans les autres phyla) font partie des folds très abondants. Une majorité de folds sont présent partout avec des variations importantes d'abondance, voir des absences dans quelques stations (milieu de la figure). Leurs abondances étant globalement élevées, ces absences sont probablement le résultat d'un manque d'exhaustivité de l'échantillonnage, et ne correspondent donc pas nécessairement à une réalité biologique. Enfin, près d'un tiers des folds sont absents de la majorité des stations et ont des abondances faibles quand ils sont présents (à droite de la figure). Ces folds sont probablement présents uniquement dans une seule lignée dont la distribution est restreinte à quelques stations, expliquant leur nombre élevé d'absences. Seuls les folds dont l'abondance est variable (ceux au milieu de la figure) pourraient avoir une distribution structurée biogéographiquement. En effet, les folds les plus abondants (à gauche) semblent avoir presque les mêmes AVs partout et sont pour certains, comme le Rossmann fold, associés à des fonctions universelles et fondamentales, qui ne sont probablement pas affectées par la géographie. Les plus rares ont pour la plupart un signal de présence-absence, qui provient probablement du fait qu'ils ne sont présents que dans un petit nombre de MAG, de façon plus ou moins phylum-spécifique, ou que leurs AVs sont très affectées par la complétion des données. Le fait que seule une très petite fraction de la diversité des folds (environ 10%) ait les valeurs d'abondance les plus élevées laisse penser que la distribution des AVs dans les stations TO est probablement de type loi de puissance, comme cela a été observé pour les OVs dans la partie précédente.

Le clustering de la Fig.56 laisse apparaitre deux groupes principaux de stations différenciés par leurs SST (Figure 57 A). La plupart des stations ayant une SST inférieure à 14°C, dans les océans Arctique et Austral, une station du Pacifique Sud (station 113) ainsi que quelques stations de l'Atlantique Sud sont plus similaires entre elles qu'avec le reste des stations. À titre de comparaison, la distribution des AVs des MAGs de la fraction de taille 0.8-2000µm (Figure 57 B) est très différente de celle des folds et consiste principalement en un signal de présence absence (avec une certaine variabilité selon les phyla). Dans chaque phylum, les MAGs présents dans le biome polaire sont globalement différents de ceux dans les biomes non polaires. Chez les Arthropodes, les MAGs Antarctiques (stations 85 à 88) sont également différents de ceux dans l'Arctique. La diversité des MAGs Chlorophytes et Haptophytes semble plus élevée dans les biomes non polaires par rapport au biome polaire, et inversement pour les Bacillariophytes. Plus généralement, il est attendu d'observer des communautés planctoniques différentes entre les biomes polaires et non polaires [74], [90], [91], [98]. Certains folds étant des synapomorphies [20], il n'est pas surprenant que des communautés d'espèces différentes conduisent à des foldomes différents entre biomes polaire et non polaire.

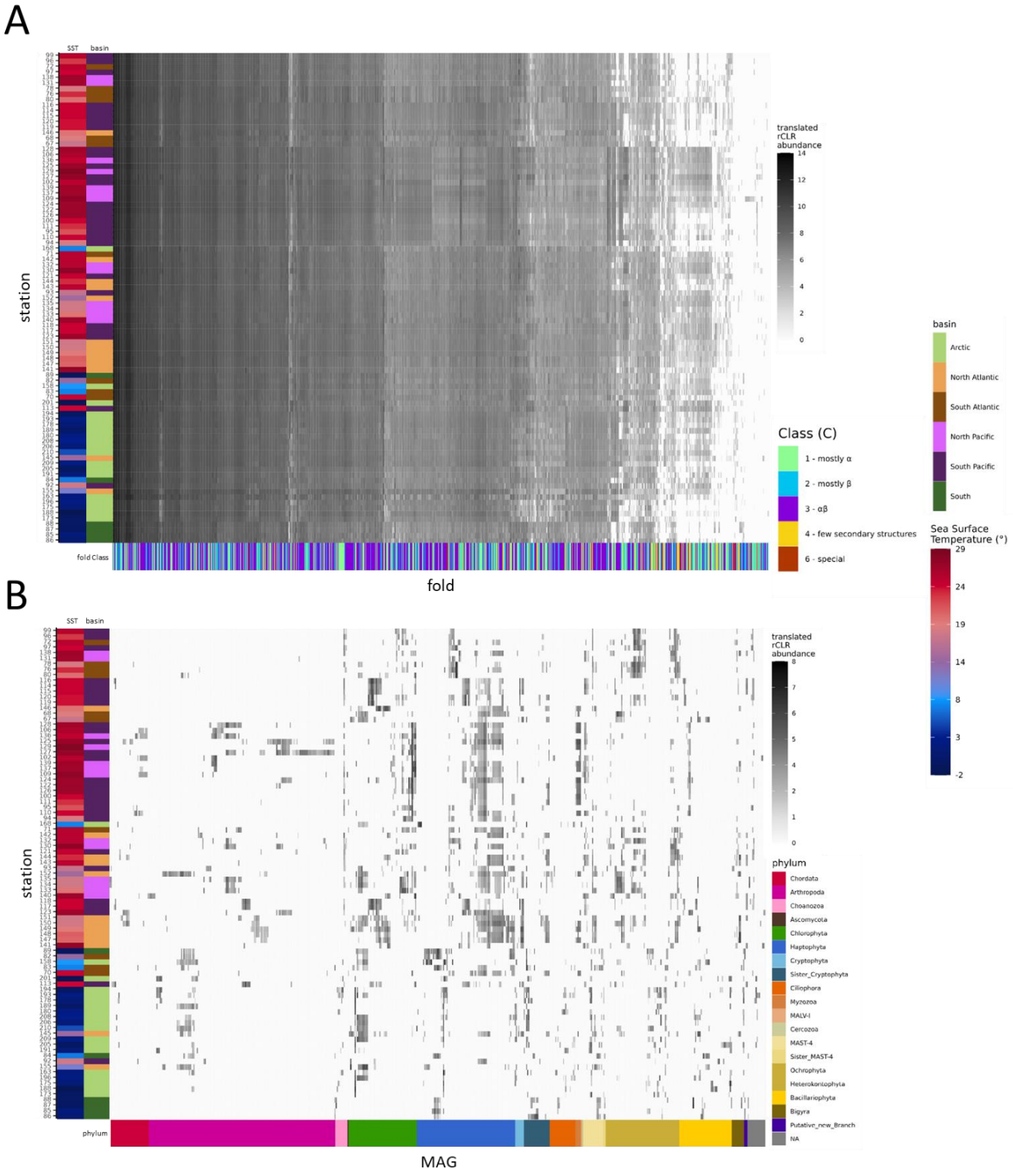


Figure 57. Caractéristiques des stations et des communautés de MAGs de la Fig.56. Les stations (axe des ordonnées) sont dans le même ordre que dans la Fig.56. **(A)** Fig.56 avec propriétés des stations (SST, bassin) sur l'axe des ordonnées et Classe des folds sur l'axe des abscisses. **(B)** Heatmap des AVs des MAGs avec les MAGs sur l'axe des abscisses. Ils sont groupés par phyla puis clusterisés en fonction de leurs AVs au sein de chaque station. Les stations sur l'axe des ordonnées sont dans le même ordre que dans la Fig.56.

b. modèles de loi de Pareto II de la distribution des abondances des folds

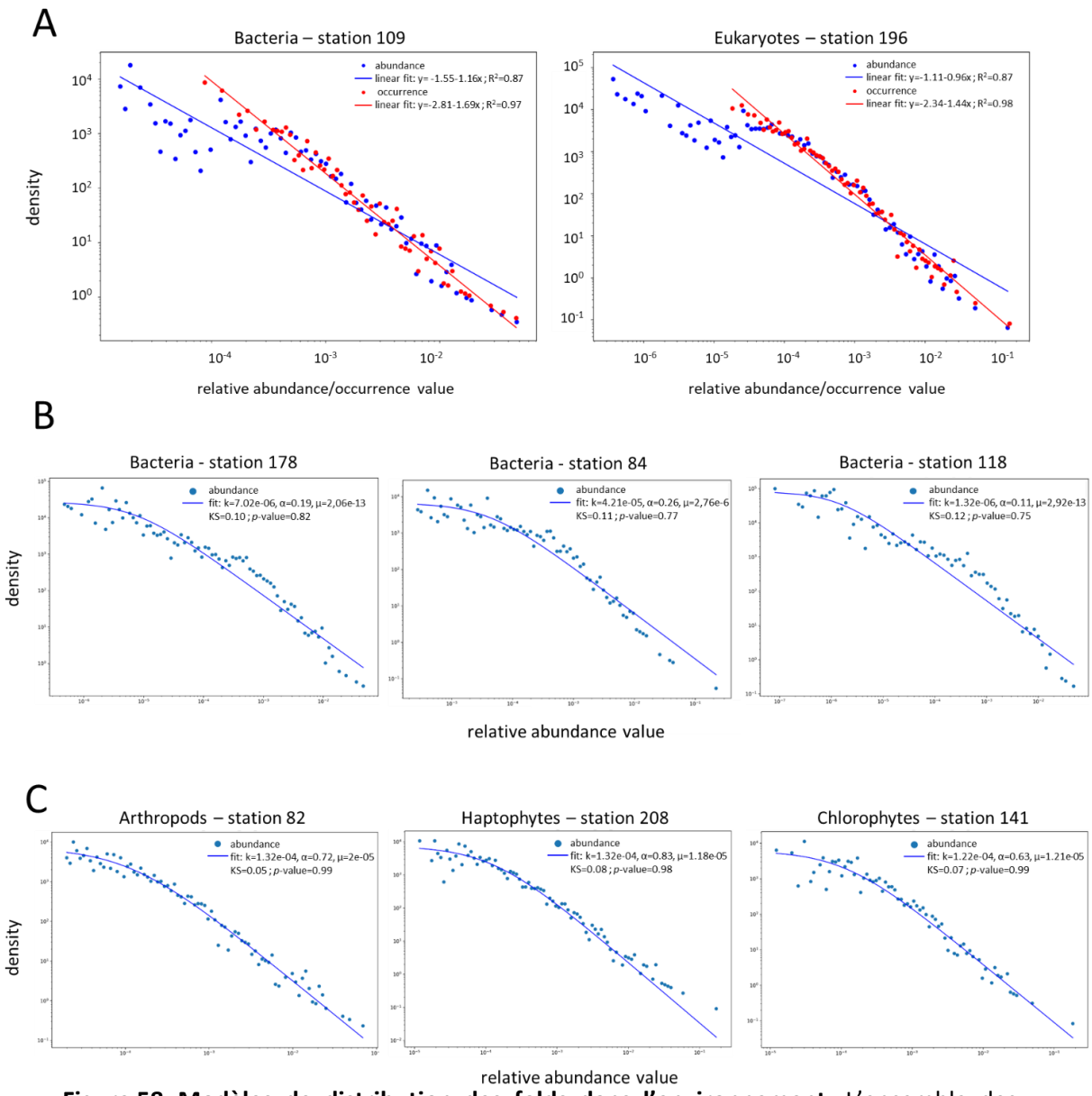


Figure 58. Modèles de distribution des folds dans l'environnement. L'ensemble des modèles sont accessibles à cette adresse : https://doi.org/10.5281/zenodo.14935989 **(A)** Exemples de modèles linéaires pour les occurrences (en rouge) et les abondances (en bleu) des folds chez les Bactéries et les Eucaryotes dans les stations 109 et 196. Les équations de régression et R^2 associés sont indiqués dans le coin en haut à droite de chaque figure. Les modèles sont tous satisfaisant au seuil de confiance 1%. **(B-C)** Modèles de Pareto type II pour la distribution des AVs des folds. Les valeurs des paramètres du modèle (k, α, μ), ainsi que de la valeur du test de KS et la p-value associée sont indiqués dans le coin en haut à droite pour chaque modèle. **(B)** Modèles pour les Bactéries dans les stations 178, 84 et 118. **(C)** Modèles pour les Arthropodes dans la station 82, les Haptophytes dans la station 208 et les Chlorophytes dans la station 141.

Après avoir visualisé la distribution globale des abondances des folds dans toutes les stations TO, les distributions de leurs AVs et de leurs OVs (pour chaque fold, somme des ses occurrences dans les différents MAGs présents dans la station considérée) dans chacune de ces stations ont été

modélisées avec la loi puissance (Figure 58 A) , avec la même méthode que pour les OVs dans les foldomes de la partie précédente (Fig.50, 51, 53, 55).

Les modèles sont en général satisfaisants au seuil de confiance 1% ($R^2 > 0.85$, *p*-value < 0.01), bien que dans la plupart des cas, les R^2 de ceux des AVs sont inférieurs à ceux des OVs. La comparaison de leurs distributions révèle donc un impact non négligeable de l'abondance des MAGs sur celle des folds. L'abondance des MAGs étant en partie déterminée par des mécanismes écologiques, l'abondance des folds pourrait également l'être dans une certaine mesure, et donc ne pas découler entièrement des propriétés génomiques des MAGs présents dans la communauté.

Pour mieux modéliser les AVs des folds, des tests ont été réalisés avec d'autres lois au sein de la famille des lois de Pareto, dont la loi de puissance est un cas particulier. Les lois mathématiques de cette famille sont en général bien adaptées pour modéliser des données avec une distribution à « queue longue ». Au sein de cette famille, la loi de Pareto de type II (PII) représente un bon compromis entre précision et complexité. Il compte parmi ses paramètres un paramètre d'échelle (appelé «*k*» ici) qui représente la valeur sur l'axe des abscisses à laquelle la distribution change d'échelle, ce qui correspond au phénomène de déviation observé dans les modèles de loi puissance des AVs des folds. Cette loi a donc été utilisée pour modéliser les distributions des AVs des folds de Bactéries, d'Eucaryotes et de six phyla différents au sein des Eucaryotes (Figure 58 B,C). Dans l'ensemble, les modèles sont satisfaisants d'après les valeurs du test de KS et les *p*-values associées, autant chez les Eucaryotes que les Bactéries. Cela indique que ces observations ne résultent probablement pas d'un biais Eucaryote-spécifique.

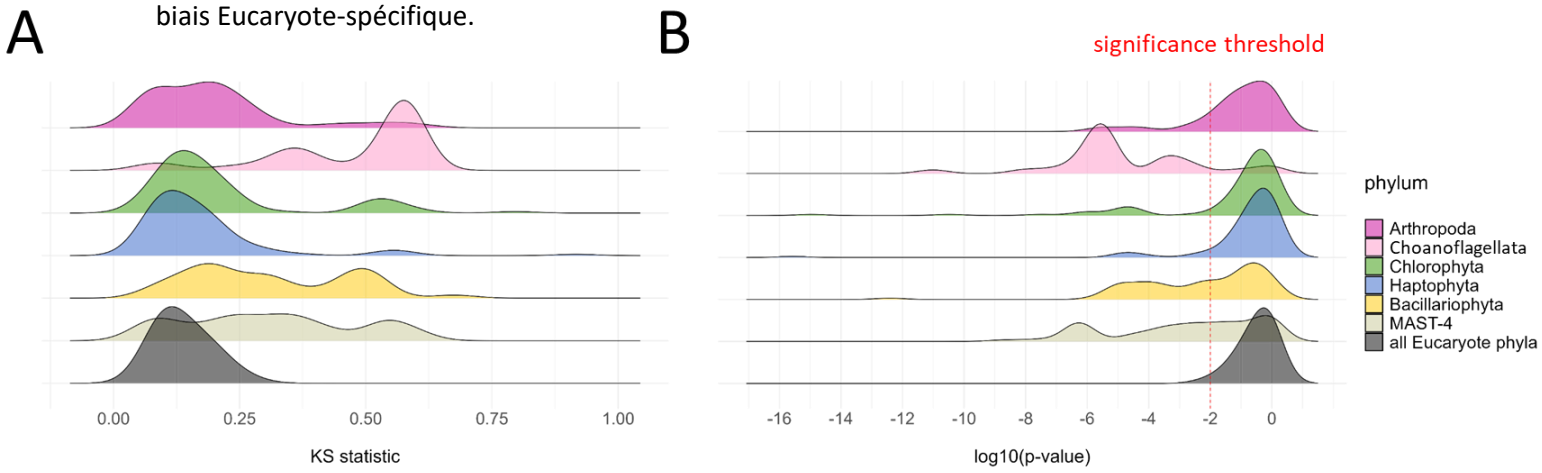
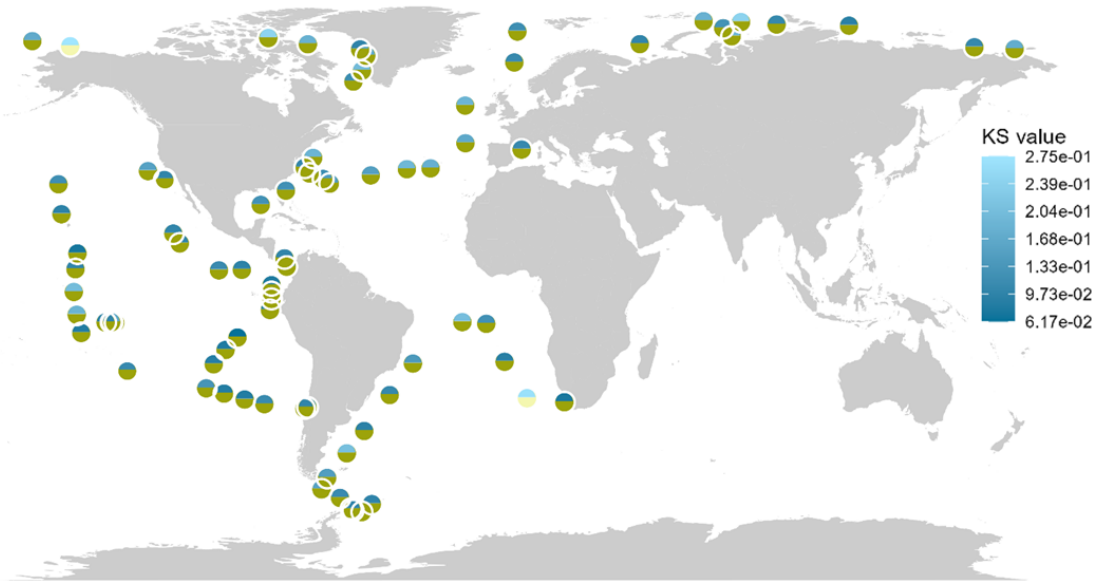


Figure 59. Significativité des modèles de PII de la distribution des AVs des folds. Le test de KS est structuré comme suit : H_o = la distribution des données ne dévie significativement pas du modèle de PII fourni ; H_1 = la distribution des données dévie significativement du modèle de PII fourni. En conséquence, les « bons » modèles de PII sont ceux pour lesquels H_o ne peut pas être rejeté, donc pour lesquels la *p*-value est au-dessus du seuil de confiance. Il y a un test de KS par station et phylum et les résultats sont regroupés par phylum. **(A)** Distribution des valeurs des tests de KS. Plus la valeur est basse et meilleur est l'ajustement. **(B)** Distribution des *p*-values des tests de KS. Le seuil de confiance est de 1%, soit -2 en log10 (indiqué par la ligne pointillée rouge).

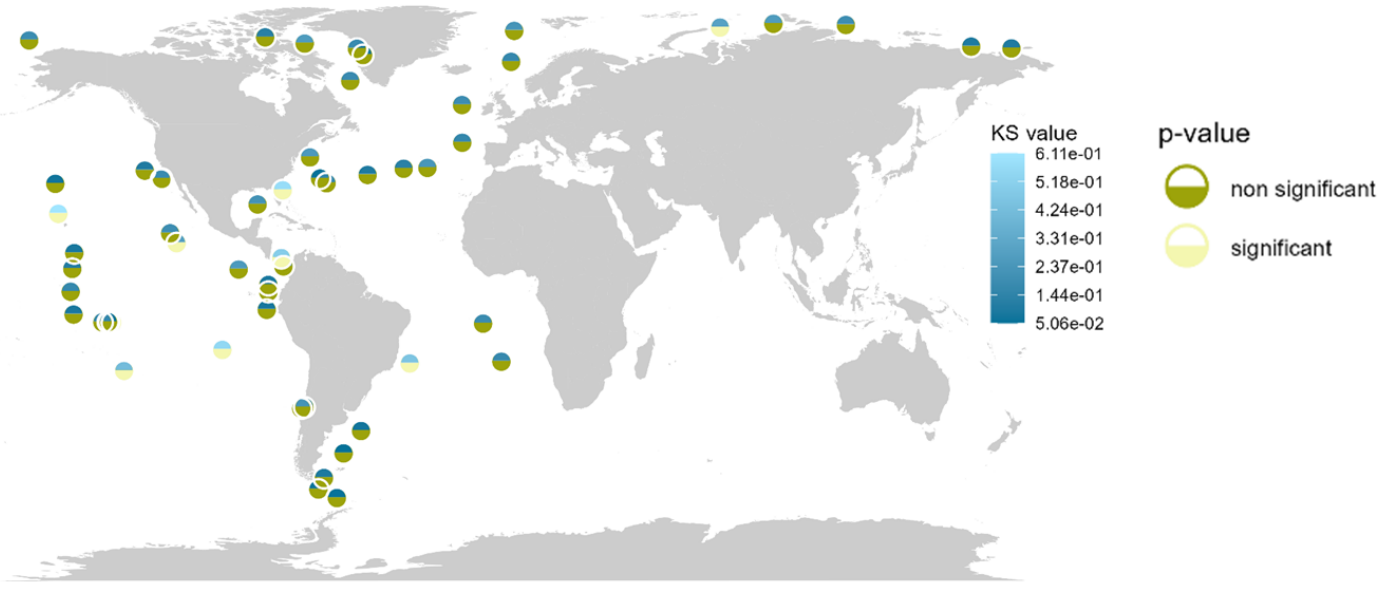
La pertinence des modèles présente cependant une certaine variabilité selon les stations et les phyla chez les Eucaryotes (Figure 59). La distribution des AVs suit une loi de PII dans la plus grande partie des stations pour tous les Eucaryotes réunis ainsi que les Arthropodes, les Chlorophytes et les Haptophytes, avec des valeurs de KS majoritairement inférieures à 0,2. Au contraire, beaucoup de modèles dévient significativement d'une loi de PII avec des valeurs KS élevées (supérieures à 0,3) chez les Bacillariophytes, MAST-4 et surtout les Choanoflagellés.

c. variabilité de la validité des modèles de distirbution avec la loi de Pareto II

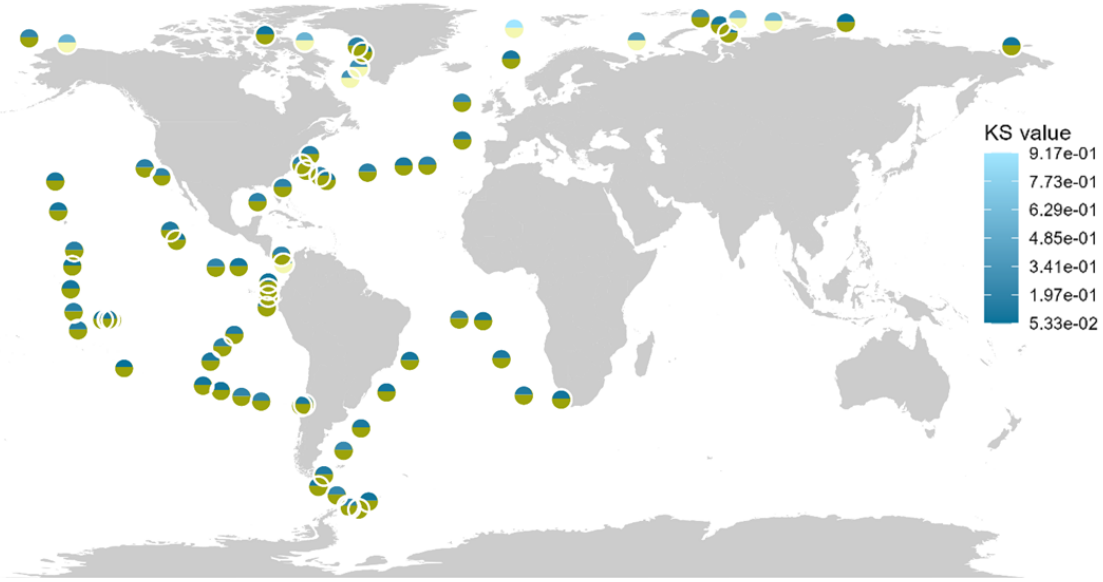
all Eucaryotes



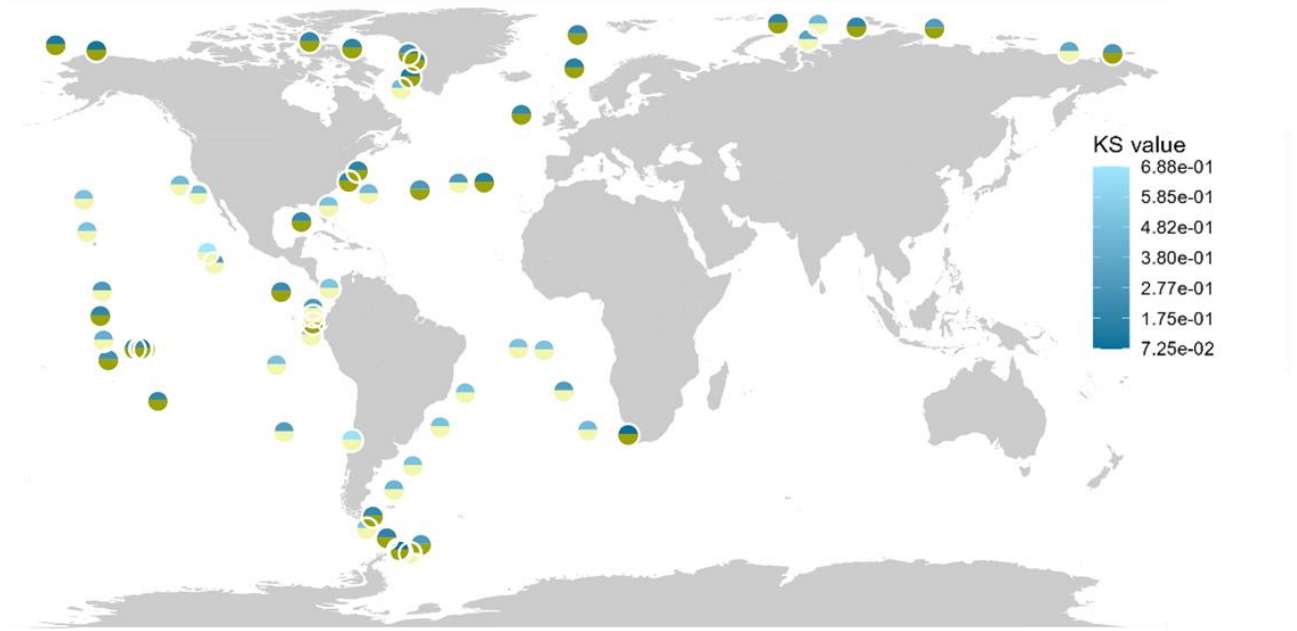
Arthropods



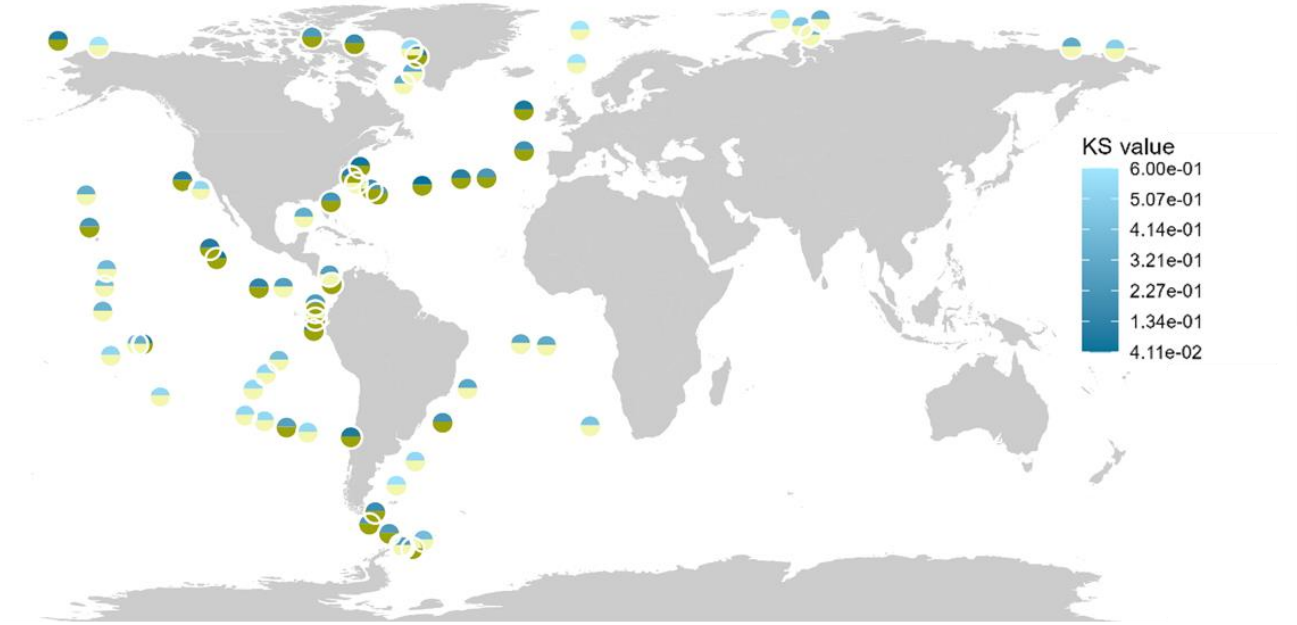
Haptophytes



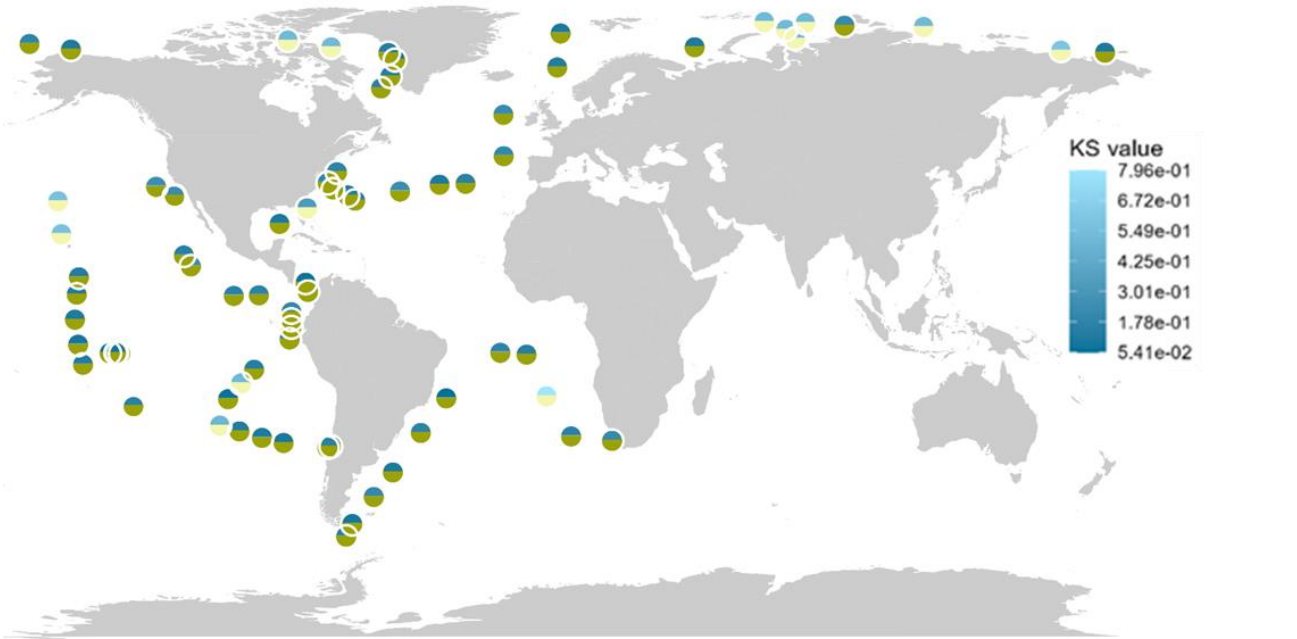
Bacillariophyta



MAST-4



Chlorophytes



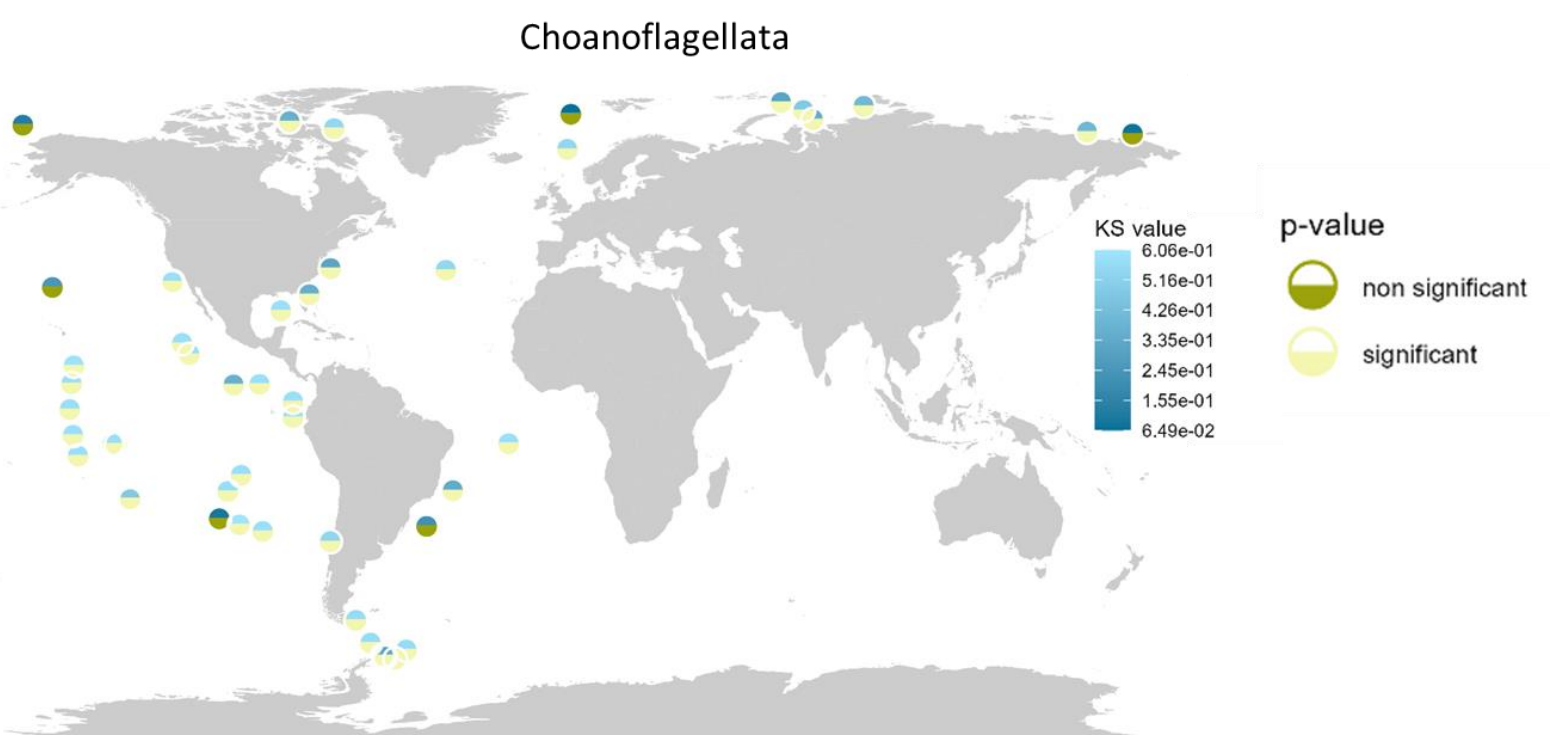


Figure 60. Résultats des tests de KS pour les modèles de PII dans chaque station TO. Chaque cercle représente une station TO. La couleur du demi-cercle du haut indique la valeur du test de KS, celle du demi-cercle du bas la *p*-value, avec en vert foncé les *p*-value supérieures à 0.01 et en vert clair celles inférieures à 0.01. Ici, l'hypothèse H_0 du test est « la distribution des points ne dévie pas significativement du modèle de Pareto type II ». Les tests « positifs » dans le sens de l'analyse sont donc ceux ne rejetant pas H_0 (avec une *p*-value supérieure à 0.01 au seuil de confiance 1%), dans une optique de recherche de stations où la loi de Pareto type II ne peut pas être statistiquement rejetée. Ces stations correspondent aux « good models » dans la Table 1.

phylum	number of good models	total number of models	proportion of good models
Haptophyta	80	89	90%
Chlorophyta	70	85	82%
Arthropoda	51	60	85%
Bacillariophyta	38	74	51%
MAST-4	36	81	44%
Choanoflagellata	6	56	11%

Table 1. Nombre et proportion de bon modèles dans chaque phylum. Les « good models » sont les distributions pour lesquelles le modèle de Pareto type II ne peut pas être statistiquement rejeté avec le test de KS (voir Fig.60) Les lignes sont ordonnées en fonction des valeurs dans la colonne « number of good models ».

Pour mieux comprendre la variabilité taxonomique de la pertinence des modèles de PII, les résultats des tests de KS ont été représentés sur une carte du monde (Figure 60; Table 1).

Avec tous les Eucaryotes, seules les distributions des AVs dans les stations 196 et 68 ne suivent pas la loi de PII. Avec seulement les Haptophytes ou les Chlorophytes, la plupart des distributions déviant significativement de la loi de PII sont localisées dans l'Arctique, à l'exception de trois stations non polaires (station 70, 131 et 132) chez les Chlorophytes. Ce sont les deux groupes avec le plus grand nombre de stations dans lesquelles la distribution des AVs suit la loi de PII (80 et 70, respectivement) La tendance inverse est observée chez les Bacillariophytes, chez qui seulement la moitié des modèles suivent la loi de PII, et, dans une moindre mesure, chez les Arthropodes. La distribution des modèles

suivant la loi de PII chez les Choanoflagellés ne présente pas de structuration géographique puisque seuls 11% des modèles sont valides, ce qui traduit probablement un échantillonnage insuffisant.

Figure 61. Corrélations entre résultats des tests de KS et potentielles variables explicatives. Les colonnes surmontées d'un triangle indiquent le résultat du test de corrélation avec le log10 de la *p*-value, celles avec une étoile la valeur du KS. Les variables en vert et en jaune correspondent respectivement à des moyennes et des écarts types. Les cases sont vides quand les tests de corrélation ne sont pas significatifs au seuil de confiance de 1%. **(A)** Corrélations avec des variables génomiques. « CAT repertoire » : taille moyenne du répertoire de tous les MAGs dans une station. « CAT occurrence » : moyenne du nombre total de folds dans le foldome de chaque MAG dans une station. **(B)** Corrélation avec des variables environnementales. "T(sampling)": SST mesurée lors de l'échantillonnage. "T": moyenne annuelle ou écart-type de la moyenne annuelle de SST. "T(amplitude)": amplitude annuelle de SST.

Dans le but de mieux comprendre les proportions variables de bons modèles de PII par phylum, des corrélations entre valeur de test de KS ou *p*-value et différentes variables explicatives possibles (aussi bien génomiques qu'environnementales) ont été testées (Figure 61).

Les corrélations significatives les plus fortes (coefficient de corrélation de Pearson supérieur à 0,6) sont en général observées avec les paramètres non-environnementaux (Figure 61 A). Les meilleures corrélations sont entre les valeurs de KS/*p*-value et la valeur de l'indice de Shannon ainsi que le nombre de MAGs par station chez les Bacillariophytes: plus le nombre de MAG et l'indice de Shannon sont élevés, plus la *p*-value est importante et la valeur de KS faible. Les meilleurs modèles sont donc localisés dans les stations où les MAGs de ce phylum sont les plus abondants et/ou diversifiés. La même relation est observée chez les Arthropodes et les MAST-4 mais elle est plus faible, avec des coefficients inférieurs à 0.5. Chez les Chlorophytes, seule la valeur de KS est corrélée négativement avec les paramètres non environnementaux (nombre de MAG, longueur moyenne du génome des MAG, occurrence moyenne des CAT, taille moyenne du répertoire des CAT). Chez les Haptophytes et tous les MAGs Eucaryotes ensemble, les seules corrélations significatives sont respectivement avec le pourcentage de lectures alignées et l'indice de Shannon, mais avec des coefficients proches de 0.3. En conséquence, la qualité des modèles de PII dans ces deux groupes semble indépendante des propriétés génomiques de la communauté. Cela n'est pas si surprenant pour le groupe « Eucaryotes », qui correspond à un regroupement de MAGs de phyla différents ayant chacun des folds dont la distribution des OVs dans les génomes suit une loi de puissance dont les paramètres sont dans une certaine mesure phylum-spécifique (Fig.50). Le fait de considérer la totalité de la diversité disponible, et d'intégrer une information écologique (les AVs), génère nécessairement un écart quasi systématique de la loi de puissance vers une loi de PII. En revanche, ce résultat est plus surprenant pour les Haptophytes, en particulier en les comparant aux Chlorophytes. En effet, la distribution géographique des bons modèles dans les deux phyla semblant suivre la même tendance, il aurait été attendu d'observer le même type de corrélations.

De façon générale, les corrélations avec les paramètres environnementaux sont faibles (toutes inférieures à 0.5), indiquant un fort découplage avec la qualité des modèles (Figure 61 B). Les meilleures corrélations sont avec la température pour les Bacillariophytes, ce qui est dans une certaine mesure cohérent avec les résultats de la Fig.60. En effet, les meilleurs modèles ont tendance à être localisés dans les stations où les Bacillariophytes sont les plus diversifiés et abondants, et la diversité et l'abondance de ce groupe est connue pour être particulièrement élevée dans les milieux froid, d'où la corrélation entre la qualité des modèles de PII pour les AVs des folds et la température [91], [99], [139]. Ici les propriétés biogéographiques des communautés de Bacillariophytes impactent donc bien les propriétés de la distribution de leurs folds à une échelle globale. Il aurait été également attendu d'obtenir des corrélations avec les paramètres environnementaux pour les Haptophytes et les Chlorophytes, puisque leur distribution semble suivre une structuration géographique, mais ce n'est pas le cas. Le très faible nombre de corrélations significatives pour les Choanoflagellés et les MAST-4 est probablement la conséquence du faible nombre de MAGs appartenant à ces phylums dans l'ensemble des données (trois MAGs Choanoflagellés en moyenne par station). Cette analyse avait principalement pour but d'apporter des éléments de compréhension aux paramètres à l'origine des lois de PII observées mais il est possible que de simples corrélations ne soient pas suffisante. Il est également possible que le log10 de la *p*-value et la valeur de KS ne permettent que de conclure sur la validité ou non du test mais que leur variabilité ne soit pas du tout lié aux paramètres testés ici.

Malgré l'absence de corrélations fortes pour la majorité des phyla, les résultats des modèles de PII semblent indiquer qu'il existe un lien entre composition de la communauté au niveau des espèces et propriétés de la communauté au niveau des folds. Pour les Bacillariophytes, il semble clair que les communautés de folds sont différentes entre milieux polaires et non polaires. C'est probablement le cas pour les Chlorophytes et les Haptophytes également, chez qui les communautés d'espèces entre les biomes polaires et non polaires sont différentes [98], [111]. Concernant les Arthropodes, les résultats obtenus jusqu'ici ne permettent pas vraiment de conclure, en partie parce que la

structuration de leurs communautés d'espèces est connue pour être différente de celle du phytoplancton [98], [111]. Plus généralement, ces résultats aident à comprendre la pertinence des modèles de loi de PII pour la distribution des AVs plutôt que les OVs dans certaines stations TO selon les phyla Eucaryotes. Au niveau des communautés de folds, les OVs dans les stations sont simplement la somme des OVs de chaque fold dans chaque MAG présent dans chaque station. En fin de compte, cette distribution ne s'écarte pas de la loi de puissance, ce qui indique que quelle que soit la distance évolutive et fonctionnelle entre les espèces au sein d'une communauté, la distribution des OVs de tous les folds qui la compose résulte du processus qui régit la distribution des occurrences dans les génomes, à savoir l'attachement préférentiel.

Ce n'est pas le cas pour les AVs des folds, qui donnent des indications dans certains phyla (principalement ceux du phytoplancton) sur les propriétés écologiques et fonctionnelles de la communauté d'espèces. Les stations ayant peu de MAGs ou des MAGs ayant à peu près les mêmes abondances ont tendance à avoir des folds dont la distribution des AVs suit plutôt une loi puissance. Dans ce cas, il y a peu de différences avec la distribution des OVs car, en l'absence de variabilité dans les AVs des MAGs, la variabilité entre les OVs des folds est trop faible pour générer un écart à la loi de puissance. Au niveau des espèces, les communautés de ces stations ont une dynamique écologique plutôt dominée par des phénomènes d'acclimatation, associés à une macrodiversité plus faible et une persistance de la dominance des mêmes espèces tout au long de l'année et sur de vastes échelles géographiques. Les communautés fonctionnant avec ce type de dynamique sont plutôt localisées dans les milieux non polaires, en particulier dans les milieux tropicaux [90].

Au contraire, les AVs des folds dans les stations avec de nombreux MAGs dont certains dominent la communauté alors que d'autres sont très rares suivent plutôt la loi de PII. La variabilité des AVs entre MAGs rares et abondants génère une déviation de la loi puissance pour la distribution des AVs des folds. Les effets de cette déviation sont surtout importants pour les folds rares, dont les AVs sont moins faibles sous une loi de PII que ce qu'elles auraient été sous une loi de puissance. Cet effet est probablement causé par le fait que les AVs des folds très rares dans les MAGs très abondants remontent au niveau des AVs de folds non rares chez des MAGs rares à l'échelle de la station. Il est également possible que les MAGs très abondants le soient également en partie grâce aux valeurs d'occurrence de certains folds qui pourraient leur apporter un avantage sélectif en leur permettant de réaliser de façon plus efficace que d'autres MAGs certaines fonctions, et qu'il existe donc un lien entre occurrence de certains folds et abondance des MAGs. Au niveau des espèces, la dynamique écologique des communautés est dans ce cas plutôt dominée par des remplacements d'espèces en fonction des saisons et sur de courtes échelles géographiques avec un vaste répertoire d'espèces rares et une macrodiversité élevée. Ce type de dynamique correspond plutôt aux stations polaires, en lien avec la règle de Rapoport [90].

Le passage d'une loi de puissance à une loi de PII pourrait donc refléter le potentiel fonctionnel de la communauté d'espèces, les stations dans lesquelles les AVs des folds suivent une loi de puissance ayant peut-être un nombre plus restreint de niches fonctionnelles que celles dans lesquelles les AVs des folds suivent une loi de PII. Ce nombre plus restreint de niches fonctionnelles pourrait être à l'origine d'une pression de sélection importante à l'échelle des structures de domaines protéiques, générant une certaine forme d'homogénéisation des foldomes des MAGs.

À noter que l'échantillonnage sur lequel se basent ces résultats est cependant loin d'être exhaustif, puisque l'ensemble des MAGs Eucaryotes d'un échantillon ne recrute qu'environ 30% des lectures métagénomiques [5]. Il est possible qu'avec un séquençage plus profond, et une reconstruction de plus de MAGs conduisant à un taux de recrutement plus élevé, la structuration géographique des distributions des communautés de folds suivant la loi de PII serait plus claire.

3/ conclusion

Ce chapitre était organisé autour de deux questions principales, la première étant « *Le modèle de loi puissance de la distribution des folds dans les foldomes est-il pertinent dans le cas de génomes environnementaux incomplets ? À quel point est-il universel dans l'arbre du vivant ?* ». J'ai d'abord vérifié que les modèles de loi puissances sur les MAGs et les RPs étaient très proches. Les paramètres de ces modèles reflètent en partie l'histoire évolutive des différents groupes ; ce phénomène est observé plus fortement avec les RPs qu'avec les MAGs, mais cela est peut-être aussi la conséquence de la diversité taxonomique des MAGs en comparaison des RPs. J'ai ensuite pu valider l'universalité de la loi de puissance pour la distribution des occurrences des folds chez les MAGs Eucaryotes et Procaryotes, avec une déviation notable chez les *Nucleocytoviricota* qui est probablement liée au jeu de données.

Dans un deuxième temps, la question était « *Le modèle de la puissance est-il pertinent à l'échelle d'une communauté de protéomes d'espèces en interaction avec des dynamiques écologiques telles que celles propres aux communautés planctoniques ?* ». Pour répondre à cette question, j'ai testé la loi de puissance sur la distribution des abondances des folds dans les communautés, et observé une déviation. Pour la prendre en compte, j'ai utilisé un autre modèle, appelé loi de Pareto type II. Chez les Eucaryotes, la distribution des abondances ne dévie pas systématiquement de la loi de puissance ; par exemple chez les Chlorophytes et les Haptophytes, les déviations semblent apparaître plutôt dans les stations polaires, alors que c'est l'inverse pour les Bacillariophytes. Cela semble indiquer que la déviation du modèle de loi puissance est liée au contexte écologique de la communauté locale, qui influe donc sur la distribution des abondances des folds qui, sans cet effet écologique, est simplement la résultante de l'évolution des foldomes. L'étude de l'abondance des folds dans l'environnement révèle donc dans une certaine mesure comme écologie et évolution peuvent impacter la distribution d'objets biologiques et comment ces deux forces évolutives interagissent.

Dans l'ensemble, ce chapitre a donc montré comment étudier les foldomes des espèces planctoniques pouvait aider à mieux comprendre l'écologie et l'évolution de ce groupe.

CHAPITRE 3. STRUCTURATION BIOGÉOGRAPHIQUE DE LA DISTRIBUTION DES FOLDS DANS LES COMMUNAUTÉS PLANCTONIQUES MARINES

Sommaire

Maintenant que les principales propriétés de la distribution des folds dans les protéomes planctoniques ainsi que dans les stations TO sont connues, la principale question posée par ce chapitre est :

- *La distribution des folds dans les océans est-elle structurée biogéographiquement, à l'instar de celle des communautés et des espèces planctoniques ?*

Pour y répondre, je commencerai par montrer la dissimilarité des stations TO du point de vue du foldome de leurs communautés planctoniques. Je distinguerai ensuite plusieurs catégories d'abondances de folds à l'aide des modèles de Pareto présentés dans le chapitre précédent. Je présenterai leur composition, puis testerai l'existence d'une structuration biogéographique de la distribution des folds dans chacune des catégories. Dans l'optique de mieux comprendre les résultats obtenus, j'utiliserai un modèle d'apprentissage automatique pour évaluer l'importance de différents paramètres environnementaux sur la distribution des folds dans chacune des catégories d'abondance.

1/ similarité des foldomes des communautés planctoniques

Afin d'évaluer l'existence d'une potentielle structuration biogéographique de la distribution globale des folds des communautés planctoniques, la similarité des stations TO du point de vue de leur composition en fold a été représentée sur une carte du monde (Figure 62).

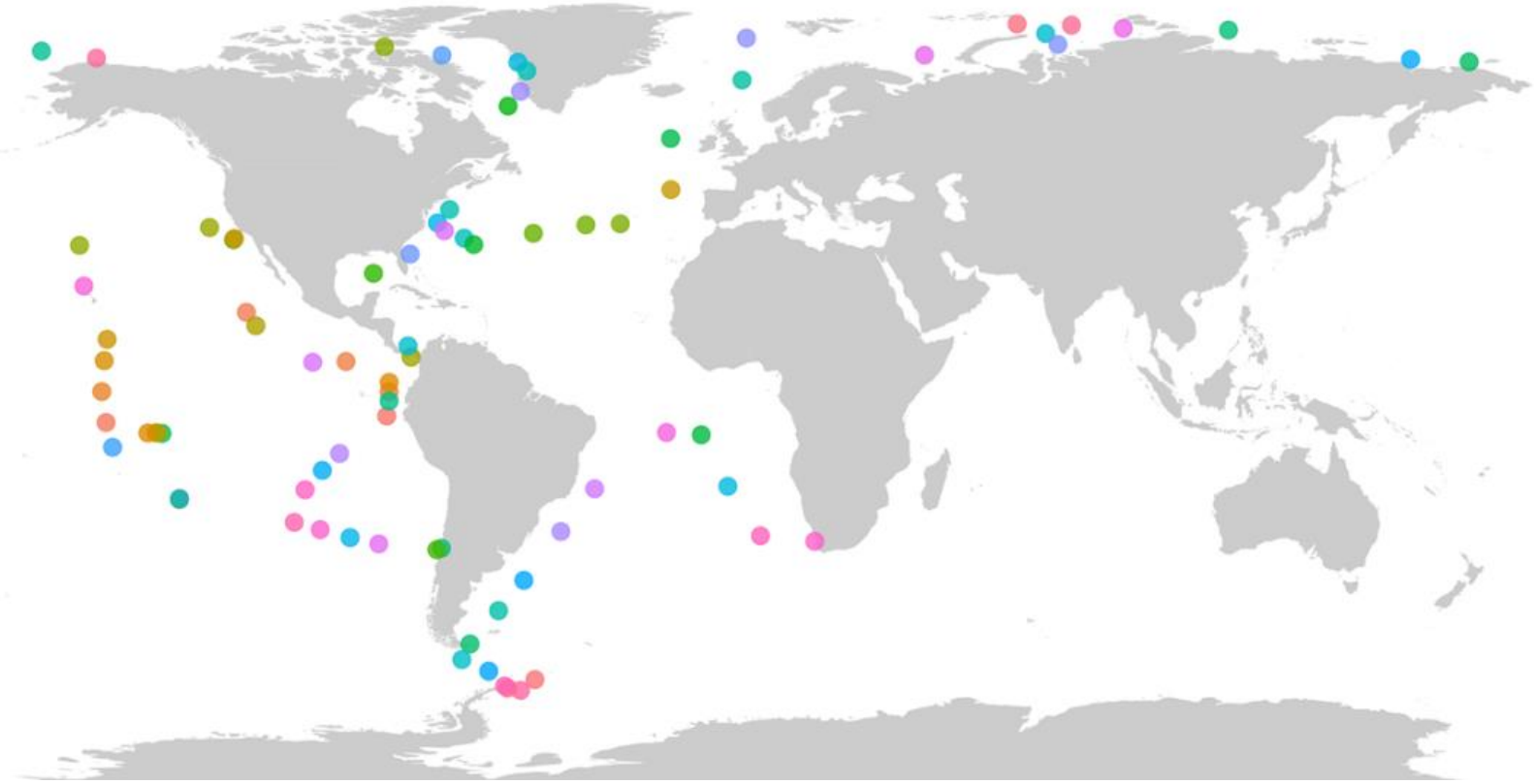


Figure 62. Similarité des foldomes des stations TO. Le foldome d'une station est l'abondance de tous les folds présents dans la station (par analogie avec le foldome d'une espèce qui correspond à l'ensemble des folds, avec une métrique d'occurrence). Les AVs relatives transformées en CLR sont celles des 908 folds eucaryotes dans les échantillons de la fraction de taille 0.8-2000µm et de surface. La couleur des points reflète la similarité entre stations. Une PCoA a été calculée à partir des AVs des folds dans toutes les stations. Les coordonnées des stations dans les trois premières dimensions de cette PCoA sont converties en un code RGB puis représentées sur une carte du monde. Les stations de couleurs proches sont donc proches dans l'espace de la PCoA et ont des foldomes plus similaires entre elles qu'avec les autres.

Cette représentation fait apparaître des similarités plus fortes entre les stations du Pacifique Nord et de l'Atlantique Nord qu'avec les autres stations, ainsi qu'entre une partie des stations du Pacifique Sud, de l'Atlantique Sud et de l'Arctique. À l'exception de ces observations, il n'y a pas de structuration biogéographique claire. Celle-ci ne peut donc pas être observée lorsque tous les folds Eucaryotes sont considérés en même temps, et ce principalement à cause de leur distribution dans laquelle quelques folds ont des abondances très élevées partout, alors que la majorité ont des abondances faibles. C'est pourtant bien la distribution de ces folds qui est potentiellement structurée biogéographiquement puisque les abondances de certains d'entre eux semblent varier de façon importante d'une station à une autre (Fig.56).

2/ définition de trois catégories d'abondances de folds

Afin d'extraire les folds dont les abondances varient significativement, une classification des folds basée sur leurs abondances et les modèles de PII présentés dans la Partie 3 du Chapitre 2 (p.140) a été proposée. Elle n'a été appliquée qu'aux folds des six phyla Eucaryotes utilisés dans cette même partie (Arthropodes, Choanoflagellés, Chlorophytes, Haptophytes, Bacillariophytes et MAST-4).

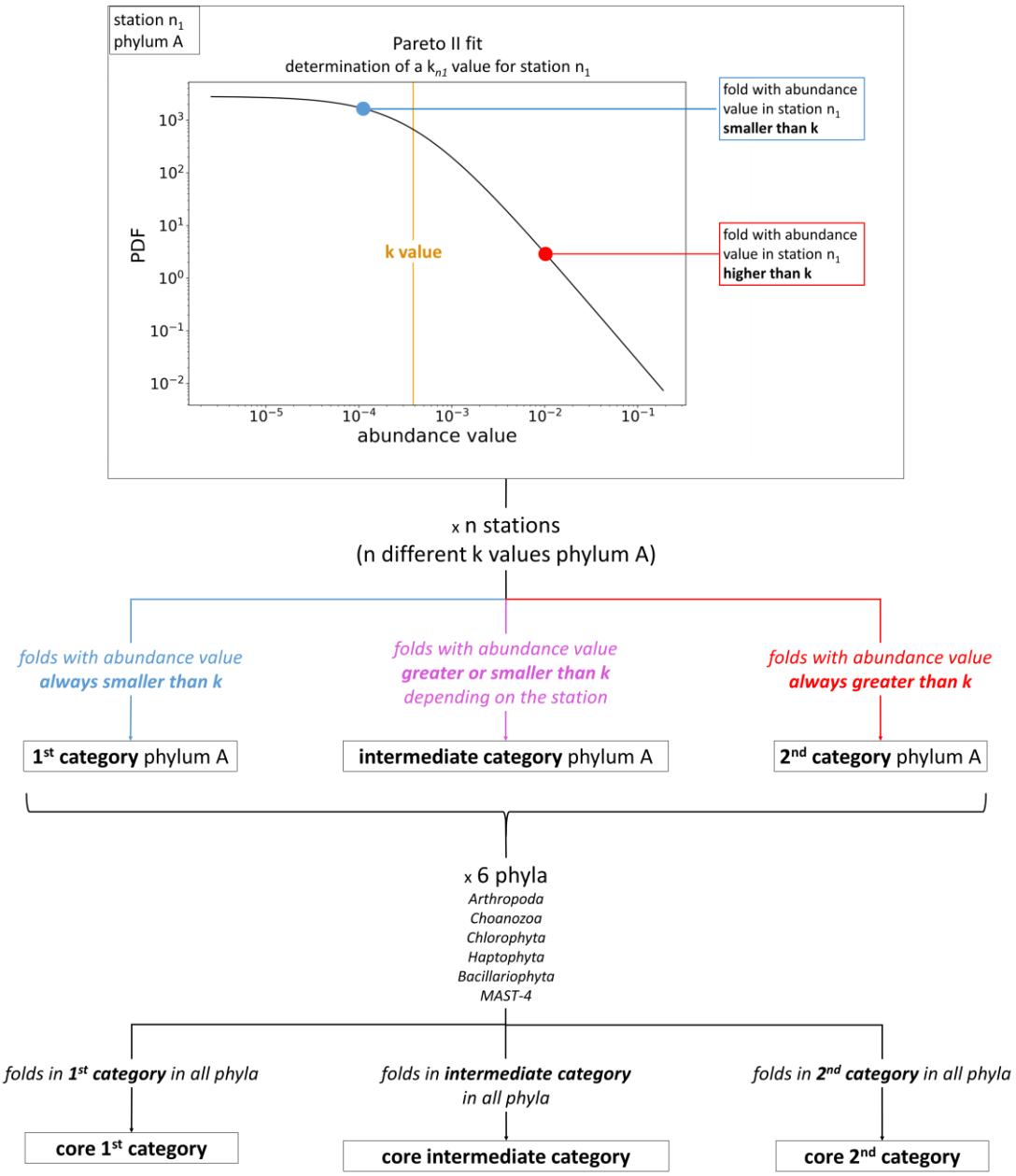


Figure 63. Définition schématique des trois catégories et système de classification des folds.

Grâce à la valeur du paramètre *k* et ses variations selon les stations et phyla, les modèles de PII permettent de classer de façon objective les folds en fonction de leur position dans la distribution des AVs dans trois catégories d'abondance (Figure 63). Ces trois catégories sont :

- la première, dans laquelle les folds ont des AVs inférieures à *k* dans toutes les stations

- l'intermédiaire, dans laquelle les folds ont des AVs inférieures à *k* dans certaines stations et supérieures à k dans d'autres
- la seconde dans laquelle les folds ont des AVs supérieures à *k* dans toutes les stations

Le noyau de chacune de ces catégories rassemble tous les folds appartenant à la catégorie dans les six phyla Eucaryotes utilisés pour cette partie de l'étude. Ils permettent de comparer les catégories entre phyla dans les analyses biogéographiques ultérieures.

La composition en fold des trois catégories dans chacun des phyla a ensuite été comparée entre phyla (Figure 64, Figure 65)

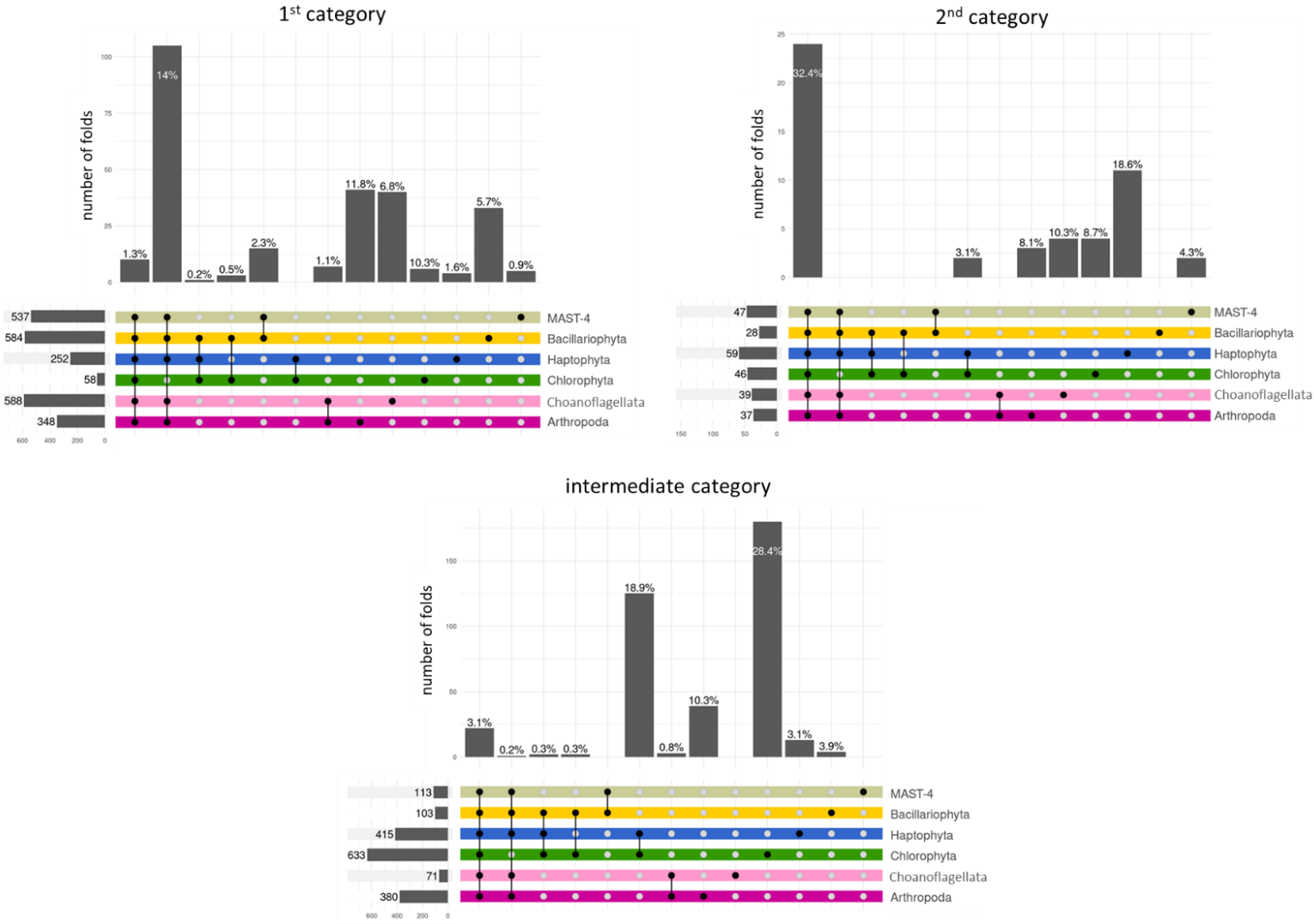


Figure 64. Comparaison des différentes catégories d'abondance de folds par phylum. Chaque couleur correspond à un phylum. Les barres verticales au-dessus des barres colorées horizontales indiquent la proportion de folds appartenant à chaque intersection, qui sont explicitées par les points noirs au milieu des barres colorées. Le noyau de chaque catégorie est représenté par l'intersection la plus à gauche dans chaque figure. Les barres horizontales grises en bas à gauche de chaque figure donnent le nombre de fold de la catégorie décrite par la figure dans le phylum sur la même ligne.

Curieusement, il y a une nette distinction en terme de proportion de folds appartenant au noyau selon les catégories : 30% des folds de deuxième catégorie, 1.3% pour les folds intermédiaires 3% pour les folds de première catégorie. Les différences entre phyla sont importantes. Chez les Chlorophytes, 86% des folds appartiennent à la catégorie intermédiaire et seulement 7,9% à la première. La même tendance, mais moins accentuée, est aussi observée chez les Haptophytes (57% et 35% respectivement). Chez les Arthropodes, les proportions de folds appartenant à la première catégorie et à la catégorie intermédiaire sont presque les mêmes (45% et 50% respectivement). Dans les autres embranchements, la majorité des folds appartiennent à la première catégorie (82% chez les Bacillariophytes, 84% chez les Choanoflagellés et 77% chez MAST-4), et moins d'un quart appartiennent à la catégorie intermédiaire (14% chez les Bacillariophytes, 10% chez les Choanoflagellés et 16% chez MAST-4). Il semblerait donc qu'il existe un lien entre le nombre de folds dans chaque catégories d'abondance et la proportion de modèles de PII valides dans chaque phylum (Table 1). Les Chlorophytes et les Haptophytes, qui ont le plus grand nombre de modèles valides, ont également les plus grandes catégories intermédiaires. Les Arthropodes, qui ont des modèles valides dans moins de stations que ces deux derniers phyla (mais autant en proportion, 85%), ont pratiquement autant de folds dans la première catégorie et la catégorie intermédiaire. Enfin, les trois derniers phyla ont à la fois les plus petites catégories intermédiaires et le plus petit nombre de bons modèles de PII.

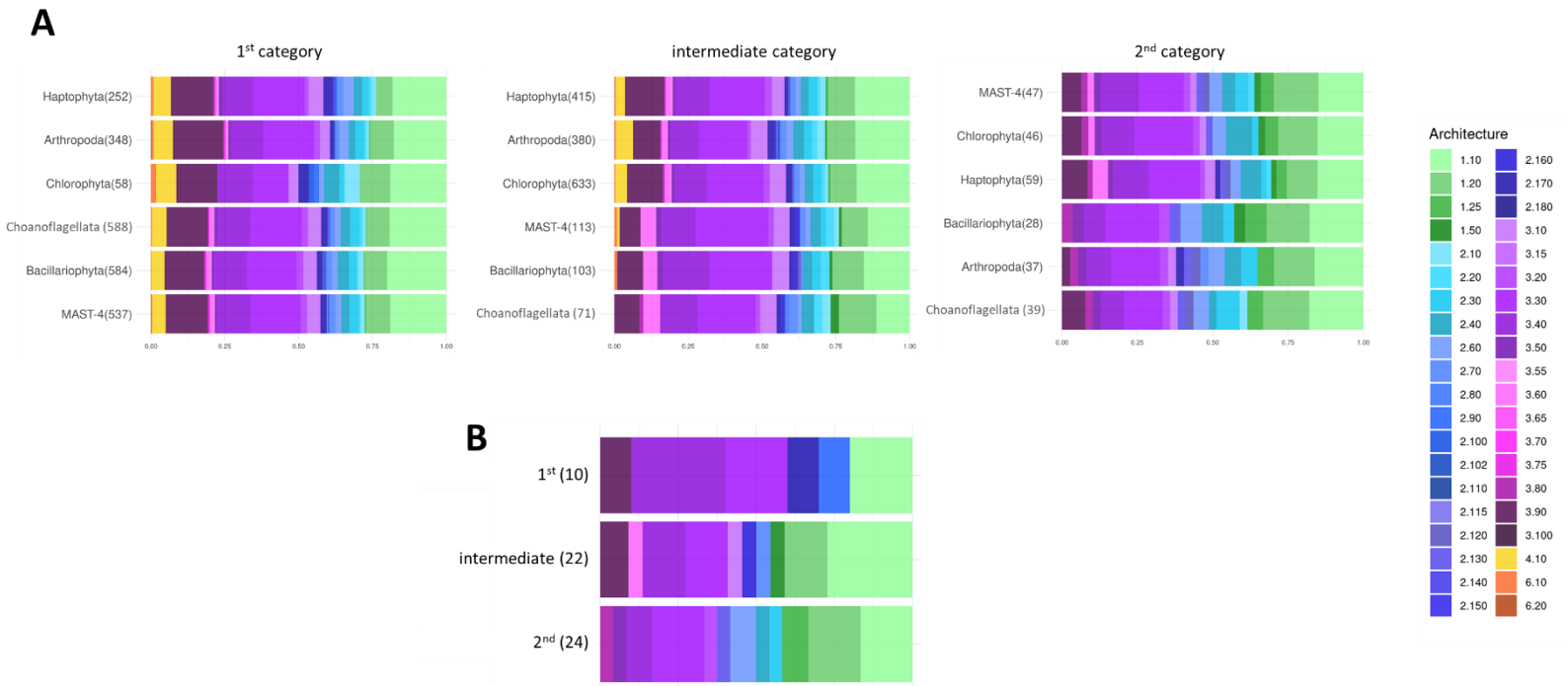


Figure 65. Composition en Architecture de chaque catégorie d'abondance de fold. La proportion de folds appartenant à chaque Architecture est indiquée sur l'axe des abscisses. Les couleurs indiquent les Architectures, avec en dégradé de vert la Classe 1, de bleu la Classe 2, de violet la Classe 3, de jaune la Classe 4, d'orange la Classe 6. **(A)** variabilité de la composition des catégories par (de gauche à droite: première catégorie, catégorie intermédiaire et deuxième catégorie). L'ordre des phyla sur l'axe des ordonnées suit les similarités en composition. Le nom de chaque phylum est suivi par le nombre de fold dans cette catégorie et ce phylum entre parenthèse. **(B)** Composition en Architecture du noyau de chaque catégorie. Le nombre de folds dans chaque noyau est indiqué entre parenthèses.

Du point de vue qualitatif, la proportion des Classes est stable entre catégorie et phyla (Figure 65): 25-30% des folds sont majoritairement Alpha, 10% sont majoritairement Beta et 40% sont Alpha Beta. L'Architecture Irrégulière (4.10) et la Classe Spéciale (6) sont représentées dans des proportions plus variables. Elles sont toutes les deux absentes de la deuxième catégorie mais peuvent regrouper ensemble jusqu'à 10 % des folds au sein de certaines catégories, avec une forte variabilité taxonomique. Par exemple, dans la catégorie intermédiaire, il n'y a pas de 4.10 chez les des Bacillariophytes et les Choanoflagellés, qui n'ont pas non plus de Classe 6, alors que ces deux Classes sont présentes dans les autres phyla. La similarité des catégories au niveau des Architectures entre phyla n'est également pas la même selon les catégories. Chez les Choanoflagellés par exemple, la catégorie rare est plus proche de celle des Bacillariophytes et des MAST-4 (ce qui est plus cohérent du point de vue fonctionnel), alors que la seconde catégorie est plus proche de celle des Arthropodes (ce qui est plus cohérent du point de vu taxonomique). La première catégorie et l'intermédiaire des Arthropodes sont plus proches de celles des Chlorophytes et des Haptophytes que de celles des Choanoflagellés. Les différentes catégories des Chlorophytes et des Bacillariophytes sont relativement dissimilaires, alors que les espèces de ces deux phyla sont proches fonctionnellement [5]. Le principal facteur responsable de la similarité semble en fait être le nombre de folds dans chaque catégorie (c'est particulièrement frappant pour la catégorie intermédiaire, dans laquelle les Haptophyteset Chlorophytes ont des compositions assez similaires et ont au moins trois fois plus de fold dans cette catégorie que les autres), nombre qui est lié à celui de bons modèles de PII.

Concernant le noyau de chaque catégorie, seules les trois Classes principales (principalement Alpha, principalement Beta, Alpha Beta) y sont représentées (Figure 65 B ; Table 2). Dans l'ensemble, la diversité en Architectures est légèrement supérieure dans la deuxième catégorie (douze) que dans la catégorie intermédiaires (dix) et la première (six). La proportion des différentes Architectures varie cependant en fonction des catégories: environ 40 % des folds intermédiaires et de deuxième catégorie sont principalement Alpha, contre environ 20% dans la première. Cette proportion est également plus élevée que les 25 % trouvés en moyenne dans la catégorie intermédiaire dans les différents phyla. L'Architecture Orthogonal Bundle (1.10) est représentée dans chaque catégorie. Il n'y a pas de folds avec l'Architecture Up-Down Bundle (1.20) dans la première catégorie, alors qu'il y en a dans la catégorie intermédiaire et la deuxième. Le fold Glycosyltransférase (1.50.10) est spécifique de la catégorie intermédiaire, tandis que les folds Alpha Horseshoe (1.25) ne sont présents que dans la deuxième catégorie. La proportion en folds majoritairement Beta est plus élevée dans le noyau des premières et deuxièmes catégories (environ 25%) par rapport à la catégorie intermédiaire (environ 10%), et la diversité en Architecture est maximale dans la deuxième catégorie (quatre différentes). Aucune Architecture de Classe 2 n'est partagée entre les trois catégories. Les proportions de folds de classe Alpha Beta sont également variables entre catégories, représentant 30 % des folds dans la deuxième, environ 40 % dans l'intermédiaire et près de 60 % dans la première. Les Architectures 2-Layer Sandwich (3.30) et 3-Layer(aba) Sandwich (3.40) sont présentes dans toutes les catégories noyaux, en proportions variables. Trois folds ont l'Architecture 3.40 dans les catégories intermédiaires et première et deux dans la deuxième, l'un d'entre eux étant le Rossmann fold (3.40.50) dans toutes les catégories. Deux Architectures sont spécifiques de la catégorie intermédiaire (Roll (3.10) et 4-Layer Sandwich (3.60)), contre trois dans la deuxième (Alpha-Beta Barrel (3.20), 3-Layer(bba) Sandwich (3.50) et Alpha-Beta Horseshoe (3.80)).

core category	CAT id	Class	Architecture	Topology
2nd	1.10.10	Mainly Alpha	Orthogonal Bundle	Arc Repressor Mutant, subunit A
2nd	1.10.287	Mainly Alpha	Orthogonal Bundle	Helix Hairpins
2nd	1.10.510	Mainly Alpha	Orthogonal Bundle	Transferase(Phosphotransferase); domain 1
2nd	1.10.8	Mainly Alpha	Orthogonal Bundle	Helicase, Ruva Protein; domain 3
2nd	1.20.120	Mainly Alpha	Up-down Bundle	Four Helix Bundle (Hemerythrin (Met), subunit A)
2nd	1.20.1250	Mainly Alpha	Up-down Bundle	Growth Hormone; Chain: A;
2nd	1.20.5	Mainly Alpha	Up-down Bundle	Single alpha-helices involved in coiled-coils or other helix-helix interfaces
2nd	1.20.58	Mainly Alpha	Up-down Bundle	Methane Monooxygenase Hydroxylase; Chain G, domain 1
2nd	1.25.10	Mainly Alpha	Alpha Horseshoe	Leucine-rich Repeat Variant
2nd	1.25.40	Mainly Alpha	Alpha Horseshoe	Serine Threonine Protein Phosphatase 5, Tetratricopeptide repeat
2nd	2.130.10	Mainly Beta	7 Propeller	Methylamine Dehydrogenase; Chain H
2nd	2.30.30	Mainly Beta	Roll	SH3 type barrels.
2nd	2.40.50	Mainly Beta	Beta Barrel	OB fold (Dihydrolipoamide Acetyltransferase, E2P)
2nd	2.60.120	Mainly Beta	Sandwich	Jelly Rolls
2nd	2.60.40	Mainly Beta	Sandwich	Immunoglobulin-like
2nd	3.20.20	Alpha Beta	Alpha-Beta Barrel	TIM Barrel
2nd	3.30.200	Alpha Beta	2-Layer Sandwich	Phosphorylase Kinase; domain 1
2nd	3.30.40	Alpha Beta	2-Layer Sandwich	Herpes Virus-1
2nd	3.30.420	Alpha Beta	2-Layer Sandwich	Nucleotidyltransferase; domain 5
2nd	3.30.70	Alpha Beta	2-Layer Sandwich	Alpha-Beta Plaits
2nd	3.40.30	Alpha Beta	3-Layer(aba) Sandwich	Glutaredoxin
2nd	3.40.50	Alpha Beta	3-Layer(aba) Sandwich	Rossmann fold
2nd	3.50.50	Alpha Beta	3-Layer(bba) Sandwich	FAD/NAD(P)-binding domain
2nd	3.80.10	Alpha Beta	Alpha-Beta Horseshoe	Leucine-rich repeat, LRR (right-handed beta-alpha superhelix)
intermediate	1.10.1070	Mainly Alpha	Orthogonal Bundle	Phosphatidylinositol 3-kinase Catalytic Subunit; Chain A, Domain 5
intermediate	1.10.132	Mainly Alpha	Orthogonal Bundle	Topoisomerase I; Chain A, domain 4
intermediate	1.10.20	Mainly Alpha	Orthogonal Bundle	Histone, subunit A
intermediate	1.10.220	Mainly Alpha	Orthogonal Bundle	Annexin V; domain 1
intermediate	1.10.30	Mainly Alpha	Orthogonal Bundle	DNA Binding (I), subunit A
intermediate	1.10.730	Mainly Alpha	Orthogonal Bundle	Isoleucyl-tRNA Synthetase; Domain 1
intermediate	1.20.1050	Mainly Alpha	Up-down Bundle	Glutathione S-transferase Yfyf (Class Pi); Chain A, domain 2
intermediate	1.20.1110	Mainly Alpha	Up-down Bundle	Calcium-transporting ATPase, transmembrane domain
intermediate	1.20.1740	Mainly Alpha	Up-down Bundle	Amino acid/polyamine transporter I
intermediate	1.50.10	Mainly Alpha	Alpha/alpha barrel	Glycosyltransferase
intermediate	2.160.20	Mainly Beta	3 Solenoid	Pectate Lyase C-like
intermediate	2.70.150	Mainly Beta	Distorted Sandwich	Calcium-transporting ATPase, cytoplasmic transduction domain A
intermediate	3.10.110	Alpha Beta	Roll	Ubiquitin Conjugating Enzyme
intermediate	3.30.1360	Alpha Beta	2-Layer Sandwich	Gyrase A; domain 2
intermediate	3.30.310	Alpha Beta	2-Layer Sandwich	TATA-Binding Protein
intermediate	3.30.559	Alpha Beta	2-Layer Sandwich	Chloramphenicol Acetyltransferase
intermediate	3.40.1110	Alpha Beta	3-Layer(aba) Sandwich	Calcium-transporting ATPase, cytoplasmic domain N
intermediate	3.40.140	Alpha Beta	3-Layer(aba) Sandwich	Cytidine Deaminase; domain 2
intermediate	3.40.250	Alpha Beta	3-Layer(aba) Sandwich	Oxidized Rhodanese; domain 1
intermediate	3.60.15	Alpha Beta	4-Layer Sandwich	Metallo-beta-lactamase; Chain A
intermediate	3.90.640	Alpha Beta	Alpha-Beta Complex	Actin; Chain A, domain 4
intermediate	3.90.79	Alpha Beta	Alpha-Beta Complex	Nucleoside Triphosphate Pyrophosphohydrolase
1st	1.10.1500	Mainly Alpha	Orthogonal Bundle	Probable Glutaminase Ybgj; Chain: A, domain 2
1st	1.10.60	Mainly Alpha	Orthogonal Bundle	Diphtheria Toxin Repressor; domain 2
1st	2.170.8	Mainly Beta	Beta Complex	Phosphoenolpyruvate Carboxykinase; domain 2
1st	2.90.10	Mainly Beta	Orthogonal Prism	Agglutinin, subunit A
1st	3.30.10	Alpha Beta	2-Layer Sandwich	Trypsin Inhibitor V; Chain A
1st	3.30.2170	Alpha Beta	2-Layer Sandwich	archaeoglobus fulgidus dsm 4304 fold
1st	3.40.1080	Alpha Beta	3-Layer(aba) Sandwich	Glutaconate Coenzyme A-transferase
1st	3.40.109	Alpha Beta	3-Layer(aba) Sandwich	NADH Oxidase
1st	3.40.449	Alpha Beta	3-Layer(aba) Sandwich	Phosphoenolpyruvate Carboxykinase; domain 1
1st	3.90.330	Alpha Beta	Alpha-Beta Complex	Nitrile Hydratase; Chain A

Table 2. Identité des folds dans le noyau de chacune des catégories d'abondance. La colonne « Topology » indique le nom du domaine représentatif de la Topologie en question d'après CATH.

Afin de relier catégorie d'abondance des folds et distribution dans les foldomes, les différences de nombre de partenaires entre folds des différentes catégories d'abondances ont été évaluées (Figure 66). Le nombre moyen de partenaires par fold dans les trois catégories est significativement différent dans tous les phyla, les folds de deuxième catégorie ayant entre 250 et 400 partenaires différents, les intermédiaires entre 25 et 80 et ceux de la première catégorie entre 7 et 20. Les catégories d'abondance distinguent donc trois types de folds qui, en plus d'avoir des gammes d'abondances très différentes dans l'environnement, ont aussi des propriétés combinatoires différentes. La classification environnementale est donc dans une certaine mesure cohérentes avec la classification des folds en Superfolds, mesofolds et unifolds réalisée à partir des occurrences des folds dans les foldomes [251].

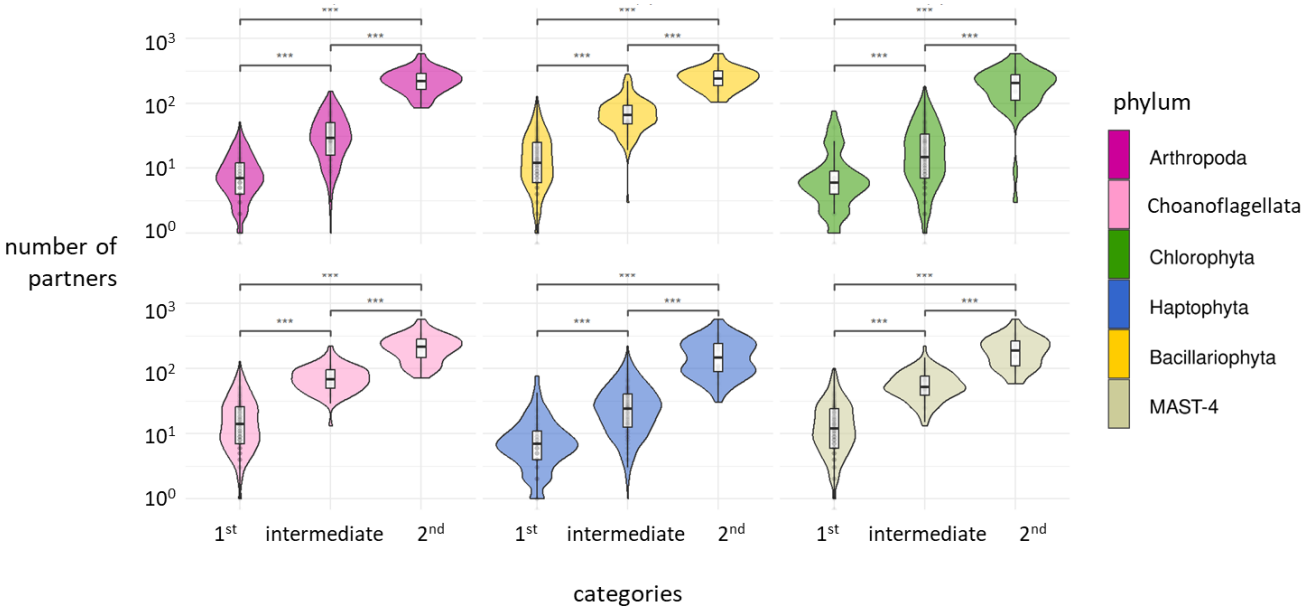


Figure 66. Nombre de partenaires des folds par catégorie d'abondance de folds. Les categories d'abondance sont représentées sur l'axe des abscisses, le nombre de partenaire de chaque fold sur l'axe des ordonnées. La significativité des différences entre catégories est testée pour chaque phylum avec un test de Wilcoxon (***: p-value < 0.01).

3/ différences de structuration biogéographique de la distribution des folds dans les stations Tara Oceans en fonction de leur classe d'abondance

Une fois les folds classés en fonction de leurs abondances, le niveau de structuration de leurs distributions biogéographiques a été évalué et comparé entre les différentes catégories d'abondance.

a. différences d'abondances des folds en fonction de leur classe d'abondance et phylum

Dans un premier temps, les AVs des folds dans les différentes catégories ont été comparées pour mieux comprendre les propriétés de leur distribution dans l'environnement. Comme attendu, les AVs moyennes des folds de chaque catégorie sont significativement différentes (Figure 67). Les folds de première catégorie sont systématiquement les moins abondants, avec des absences dans tous les phyla. Les AVs des folds de catégories intermédiaires sont plus faibles que celles des folds de deuxième catégorie. Les premières et secondes catégories seront donc renommées « rares » et « abondantes », respectivement, pour le reste de l'étude.

De façon intéressante, certains folds des Arthropodes et Choanoflagellés sont absents dans certaines stations mais suffisamment abondants dans d'autres pour être classés parmi les intermédiaires. C'est le cas par exemple des folds 1.10.132, 3.30.1360, 3.30.310 et 3.40.250 chez les Arthropodes de la station 131. Il n'y a cependant qu'un seul MAG appartenant à ce phylum dans cette station, dont le répertoire de folds est dépourvu de ces quatre folds. Ce MAG est présent dans d'autres stations et n'y est pas particulièrement rare. Il n'est donc pas possible de savoir si l'absence de ces quatre folds dans ce MAG est un véritable signal biologique ou un artefact de reconstruction du MAG et/ou d'annotation structurale.

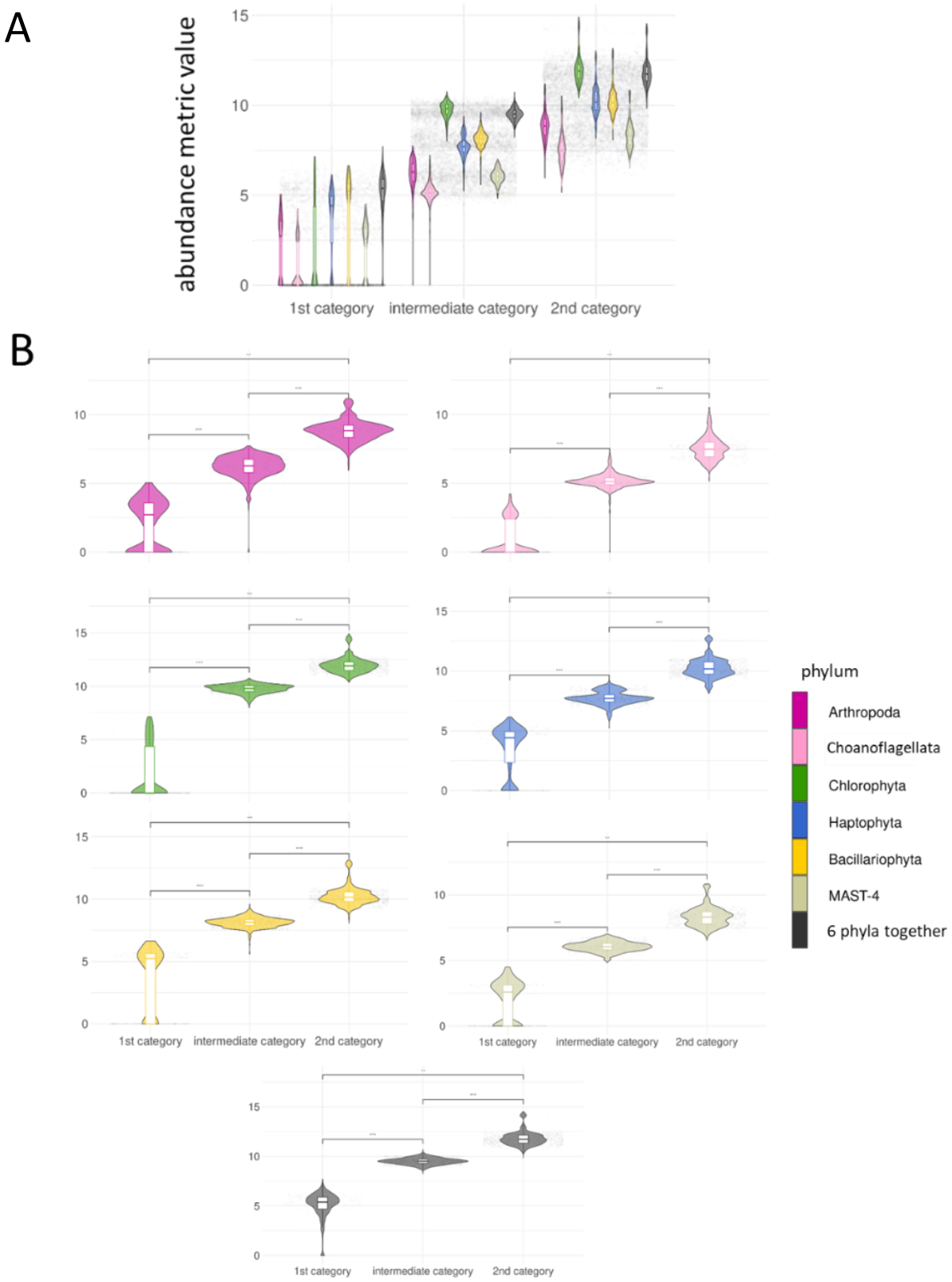


Figure 67. AVs des folds dans les différentes catégories d'abondance par phylum. La catégorie "six phyla together" correspond à la somme des AVs de chaque fold dans les six phyla. (A) AVs des folds de chaque catégorie non noyau par phylum. Les violin plots sont organises d'abord par catégorie d'abondance puis par phylum. L'AV est indiquée sur l'axe des ordonnées. **(B)** Différence d'AVs entre catégories par phylum. La significativité des différences est estimée avec un test de Wilcoxon, au seuil de confiance 1% (***: p-value < 0.01).

Afin de visualiser la distribution des catégories à l'échelle globale, une couche d'information indiquant la catégorie moyenne de chaque fold a été ajoutée à la Fig.56 (Figure 68).

Elle montre que le regroupement arbitraire des folds de la Fig.56 est dans la plupart des cas cohérent avec les catégories d'abondances définies ici. Les folds dans la partie gauche de la heatmap, dont les abondances sont élevées et stables dans toutes les stations, appartiennent à la catégorie « abondants » dans tous les phyla ou dans la majorité d'entre eux. Les folds strictement intermédiaire ou intermédiaires dans la majorité des phyla et abondants dans les autres correspondent à des folds dont les AVs sont élevées avec une variabilité importante. Dans la moitié droite de la figure, les folds sont globalement intermédiaires stricts ou intermédiaires rares. Dans la partie la plus à droite, les folds sont majoritairement rares stricts, à deux exception près: le 2.40.330 qui n'est présent que dans un phylum et y est abondant, et le 1.20.141 qui est soit abondant, soit rare.

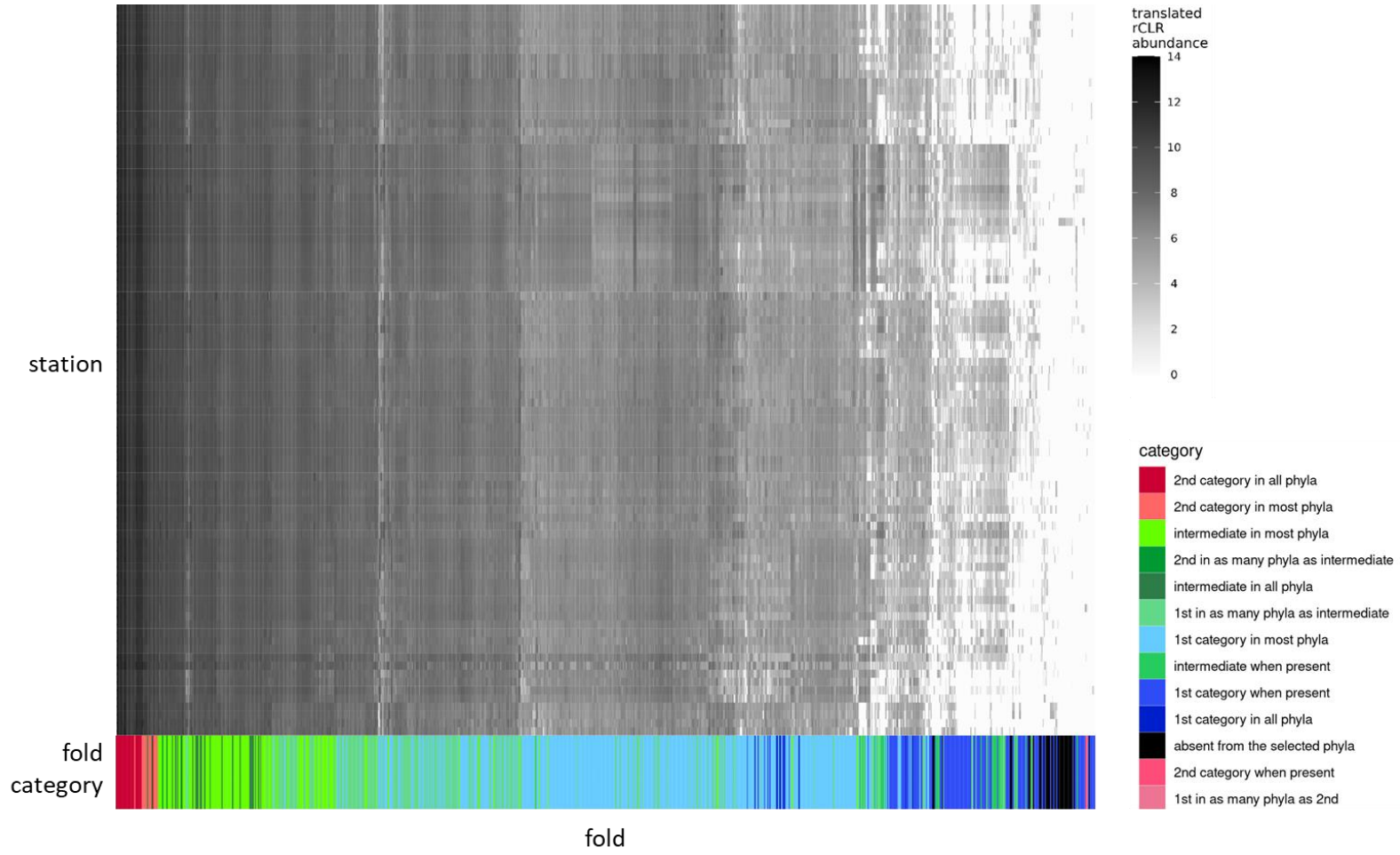


Figure 68. Distribution des catégories d'abondance à l'échelle globale. Fig.56 avec sur l'axe des abscisses la catégorie à laquelle le fold appartient dans les six phyla («in all phyla » : strictement dans une catégorie ; « in most phyla » : dans au moins quatre phyla sur six ; « in as many phyla » : dans une catégorie dans trois phyla et dans une autre dans les trois autres ; « when present » : le fold n'est pas trouvé dans les six phyla puisque la Fig.56 a été faite avec tous les MAGs mais appartient à la catégorie indiquée dans les phyla où il est présent ; « absent » : fold absent des six phyla sélectionnés).

b. différences d'α-diversité des folds en fonction de leur classe d'abondance et phylum

L'existence de différences significatives d'AVs entre les catégories d'abondance a ensuite conduit à interroger la variabilité de la complexité des communautés de folds en fonction de la latitude en analysant leur α-diversité avec l'indice de Shannon (Figure 69; Figure 70; Table 3). L'analyse est réalisée sur les catégories ainsi que leurs noyaux. L'utilisation des deux (catégories par phylum et noyau des catégorie) est nécessaire car l'utilisation uniquement des catégories par phylum conduirait à certains biais. En particulier, les différences de valeurs d'indice de Shannon pour les folds d'une catégorie entre phyla à une station donnée résultent principalement du nombre différent de folds dans cette catégorie dans les différents phyla, ce qui n'est pas le cas pour les folds du noyau qui sont par définition les mêmes dans chaque phylum.

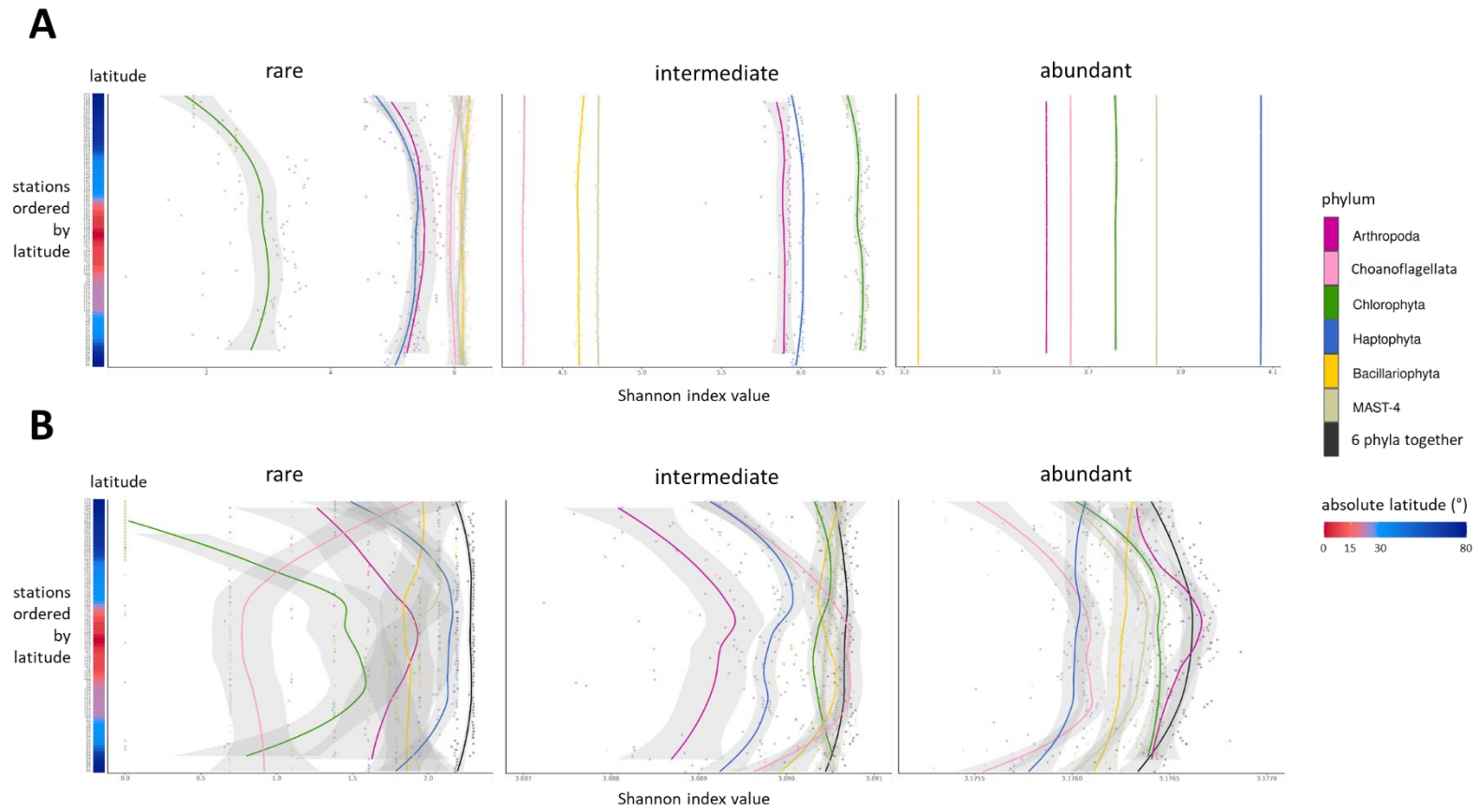


Figure 69. A-diversité des catégories d'abondance et de leurs noyaux. Les stations sont indiquées sur l'axe des ordonnées et classées par latitude. La valeur de l'indice de Shannon est en abscisses. Les bandes grises autour des courbes indiquent l'écart type. La catégorie « 6 phyla together » correspond aux indices de Shannon calculés sur la somme des abondances de chaque fold dans les six phyla. **(A)** Pour chaque catégorie. **(B)** Pour le noyau de chaque catégorie.

Il ne semble y avoir aucune variabilité latitudinale pour les folds abondants et intermédiaires quel que soit le phylum, à l'exception des folds intermédiaires chez les Haptophytes dont l'α-diversité semble diminuer très légèrement aux pôles (Figure 69 A). La variabilité latitudinale est beaucoup plus importante pour les folds rares, en cohérence avec la variabilité du nombre de folds dans cette catégorie dans les différents phyla. Le gradient latitudinal est le plus fort chez les Chlorophytes, Haptophytes et Arthropodes qui ont les plus petites catégories rares. Il ne semble donc pas y avoir de variation significative d'α-diversité dans les catégories de chaque phylum. Concernant le noyau des différentes catégories (Figure 69 B), les variations de l'indice de Shannon se font globalement avec la latitude. L'amplitude de la variation n'est cependant pas du tout la même entre le noyau rare, pour lequel il est d'environ 10^{-1}, et les noyaux intermédiaires et abondants pour lesquels il n'est que de 10^{-3} et 10^{-4}, respectivement. Malgré l'écart entre ces ordres de grandeur, la tendance est la même quand les six phyla sont considérés ensemble, à savoir une diminution de l'α-diversité vers les pôles. Le même type de profil a déjà été observé dans des études publiées à l'échelle des espèces [64]. Il correspond à la définition d'une distribution antitropicale, symétrique par rapport à l'équateur [134]. Les résultats sont plus variables quand les phyla sont considérés séparément. Seuls les folds du noyau des trois catégories chez les Arthropodes présentent la même tendance à la baisse d'α-diversité plus la latitude augmente. C'est également le phylum avec la valeur la plus élevée d'α-diversité (en excluant les six phyla pris ensemble) dans le noyau de la catégorie abondante, avec un maximum atteint sous les tropiques. Les Copépodes, qui sont un groupe très diversifié d'organismes pluricellulaires, ont eu un taux de duplication de gènes très élevé au cours de leur histoire évolutive [376]; dans le Chapitre II (Fig.50), cela avait été identifié comme étant probablement le facteur le plus important pour expliquer la pente de la droite de distribution des OVs dans leurs génomes. Ce taux de duplication a particulièrement affecté les folds qui sont ici regroupés dans le noyau de la catégorie des folds abondants (globalement, les superfolds) ; il est donc probable qu'il soit indirectement responsable des valeurs d'α-diversité particulièrement importantes observées ici. Chez les Haptophytes et Chlorophytes, la tendance à la décroissance de l'α-diversité vers les pôles s'observe dans le noyau des catégories rares et intermédiaires, et rares et abondants (en particulier dans l'hémisphère nord pour les folds abondants) respectivement. Enfin chez les Bacillariophytes, l'α-diversité augmente systématiquement vers l'Arctique bien que l'amplitude de la variation soit globalement plus faible que dans les autres groupes. C'est aussi dans cette région que l'α-diversité de ce phylum au niveau des espèces est la plus élevée [55], [139].

Afin de vérifier l'intensité de ces gradient latitudinaux d'α-diversité, des tests de significativité sur les différences d'AVs entre stations polaires et non polaires ont été réalisés (Figure 70; Table 3). Comme supposé avec la Fig.69 B, l'amplitude de la variation latitudinale d'α-diversité des noyaux de toutes les catégories chez les Bacillariophytes est trop faible pour être significative au seuil de confiance de 1% ; c'est aussi le cas pour les six phyla réunis (Figure 70 A,B,C ; Table 3). Cette observation confirme que la relation entre α-diversité et latitude pour le noyau de toutes les catégories est différente en fonction des phyla, résultant en un effet compensatoire et *in fine* une absence de différences significatives quand les six sont réunis. L'absence de différences d'α-diversité chez les Bacillariophytes est néanmoins surprenante, surtout en comparaison des différences significatives observées pour les folds du noyau des catégories rares et abondantes et rares et intermédiaires chez les Chlorophytes et les Haptophytes, respectivement. Cela indique qu'il n'y a probablement pas de convergence visible au niveau du noyau des catégories pour le phytoplancton dans l'usage des folds en réponse au passage du biome polaire au biome non polaire, en tout cas avec les MAGs.

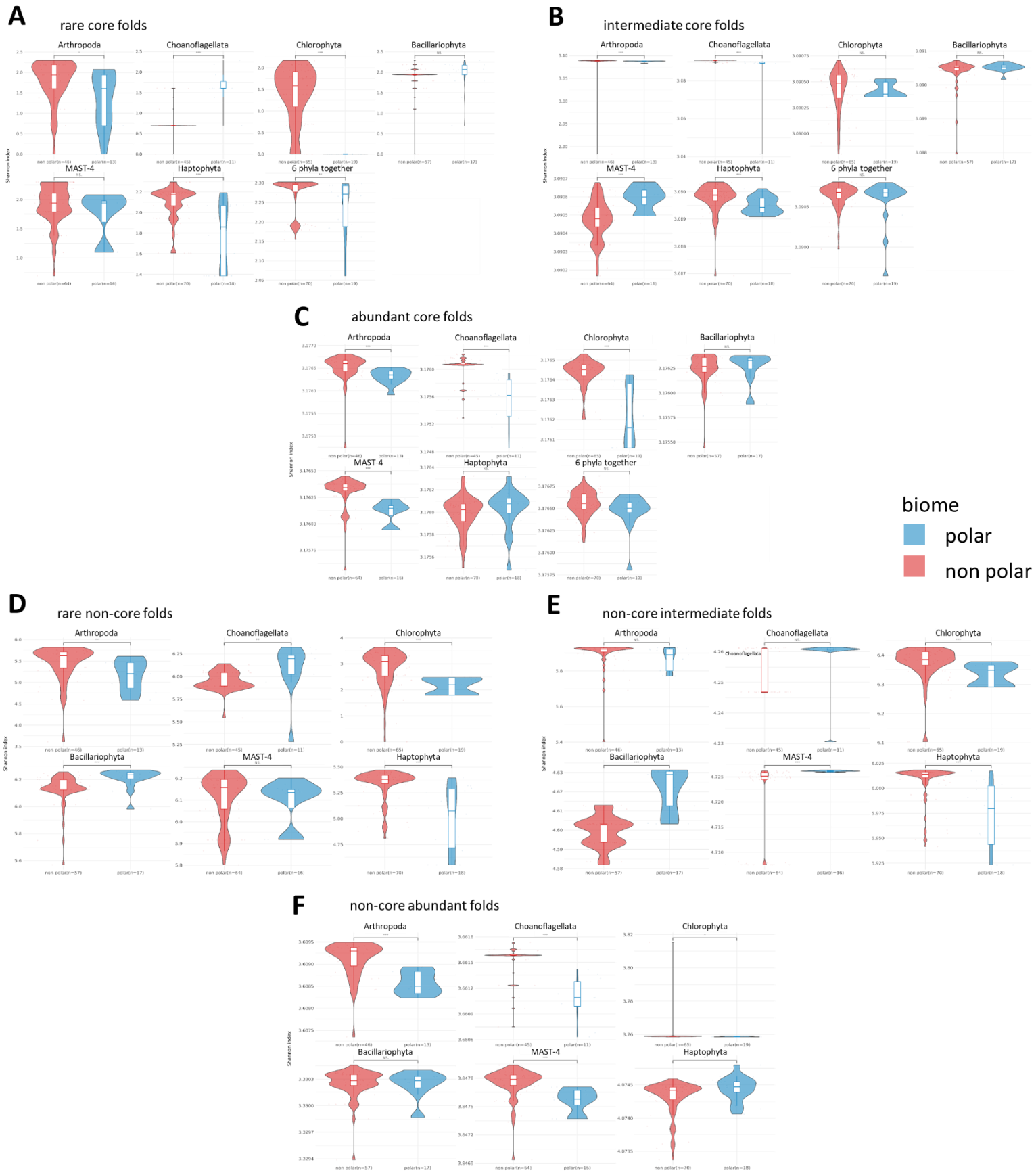


Figure 70. Différences d'α-diversité des catégories entre biomes polaires et non polaires. Le biome polaire rassemble toutes les stations de latitude absolue supérieure à 60°. Les valeurs entre parenthèses sur l'axe des abscisses indiquent le nombre de stations dans chacun des biomes avec des folds dans le phylum et la catégorie en question. La significativité des différences est mesurée avec un test de Wilcoxon (***: p-value < 0.01; **: p-value < 0.05; *: p-value < 0.1; NS: p-value ≥ 0.1). **(A-C)** Fold du noyau de chaque catégorie **(A)** rare **(B)** intermédiaire **(C)** abondant. **(D-F)** Fold de chaque catégorie non noyau **(D)** rare **(E)** intermédiaire **(F)** abondant.

En revanche, l'étude des catégories non noyau révèle des différences significatives entre phyla au sein du phytoplancton et phyla non phytoplanctoniques, qui sont globalement cohérentes avec la distribution connue de l'α-diversité au niveau des espèces [64], indiquant que les zones à forte diversité d'espèces sont, comme attendu, des zones à forte diversité de folds (Figure 70 D,E,F ; Table 3). Les différences significatives ne sont cependant pas observées dans les mêmes catégories de folds en fonction des phyla. L'α-diversité des catégories rares et intermédiaire des trois phyla phytoplanctoniques est significativement différente entre biomes. Chez les Bacillariophytes, elle est plus élevée aux pôles, et inversement pour les Haptophytes et Chlorophytes (Figure 70 D,E ; Table 3). Chez les trois phyla non phytoplanctoniques, c'est l'α-diversité de la catégorie abondante qui est significativement différente entre biomes, et va toujours dans le sens d'une augmentation en-dehors des pôles. Ces différences entre phyla phytoplanctoniques et non phytoplanctonique sont particulièrement intéressantes: si dans les catégories non noyau l'α-diversité des folds n'était impactée que par l'α-diversité des MAGs, alors la réponse serait la même quelle que soit la catégorie, ce qui n'est pas le cas. De plus, la catégorie abondante est la plus homogène en terme de nombre de folds à travers les phyla (Fig.64); la différence entre phyla non phytoplanctoniques et phytoplanctoniques ne peut donc pas non plus être simplement expliquée par cela. Il semblerait donc qu'il y ait une forme de convergence dans le type de catégorie d'abondance impacté par la transition du front polaire en fonction du type trophique, les catégories rares et intermédiaires étant plus impactées chez les phototrophes et la catégorie abondante étant plus impactée chez les hétérotrophes. La complétion de ces MAGs étant globalement plus basse que celle des MAGs phytoplanctoniques, il est également possible que ce soit elle plus que le type trophique qui soit responsable de ces résultats. La convergence observée entre groupes phytoplanctoniques, notamment entre Chlorophytes et Bacillariophytes, est cohérente à la fois avec la distribution connue de leurs α-diversités au niveau des espèce, mais aussi le fait qu'ils partagent le même groupe fonctionnel avec des distributions biogéographiques différentes [5], [55]. Les différences de résultats observées avec les folds de chaque catégorie et ceux uniquement de leurs noyaux sont également intrigantes. Par construction, l'amplitude des variations d'AVs des folds dans le noyau des catégories est nécessairement plus faible que l'amplitude des variations d'AVs des folds qui ne font pas partie du noyau (à cause de leur variabilité, ils peuvent appartenir à une autre catégorie dans un autre phylum et donc ne pas appartenir au noyau). Il est donc possible que les différences de résultats observées soient causées par le fait que le signal latitudinal d'α-diversité des folds des noyaux soit masqué dans les catégories non noyaux par ces folds.

		Arthropoda	Choanoflagellata	Chlorophyta	Haptophyta	Bacillariophyta	MAST-4	6 phyla together
core	rare	NS	***	***	***	NS	NS	NS
	intermediate	***	***	NS	***	NS	***	NS
	abundant	***	***	***	NS	NS	***	NS
non-core	rare	NS	NS	***	***	***	NS	/
	intermediate	NS	NS	***	***	***	***	
	abundant	***	***	NS	NS	NS	***	
MAGs		NS	NS	NS	***	NS	NS	***

Table 3. Résultat des tests de Wilcoxon des Fig.70-71. Les cellules colorées en rouges avec "NS" indiquent les tests non significatifs au seuil de confiance 1% (*p*-value ≥ 0.01). Les cellules en vert avec "***" correspondent aux tests significatif au même seuil de confiance (*p*-value < 0.01).

Afin de vérifier si cet effet résultait uniquement de l'α-diversité des MAGs, celle-ci a été représentée de la même façon que pour les folds (Figure 71). Bien que des différences significatives existent pour les Haptophytes, ce n'est pas le cas au seuil de confiance 1% pour tous les autres phyla, ce qui indique qu'il n'y a globalement pas de lien direct entre les deux échelles (sauf potentiellement pour les Haptophytes). Les différences d'α-diversité observées proviennent donc bien surtout des folds eux-mêmes, et sont accentuées dans les folds par rapport aux MAGs.

Dans l'ensemble, tous ces résultats pourraient indiquer que l'α-diversité des folds dans le noyau de la catégorie abondant est plus liée à l'histoire évolutive des phyla, alors que ce sont des mécanismes écologiques qui seraient plutôt responsables de celle du noyau de la catégorie intermédiaire. Enfin, la catégorie rare et son noyau seraient surtout impactés par le taux de complétion et d'annotation structurale des MAGs.

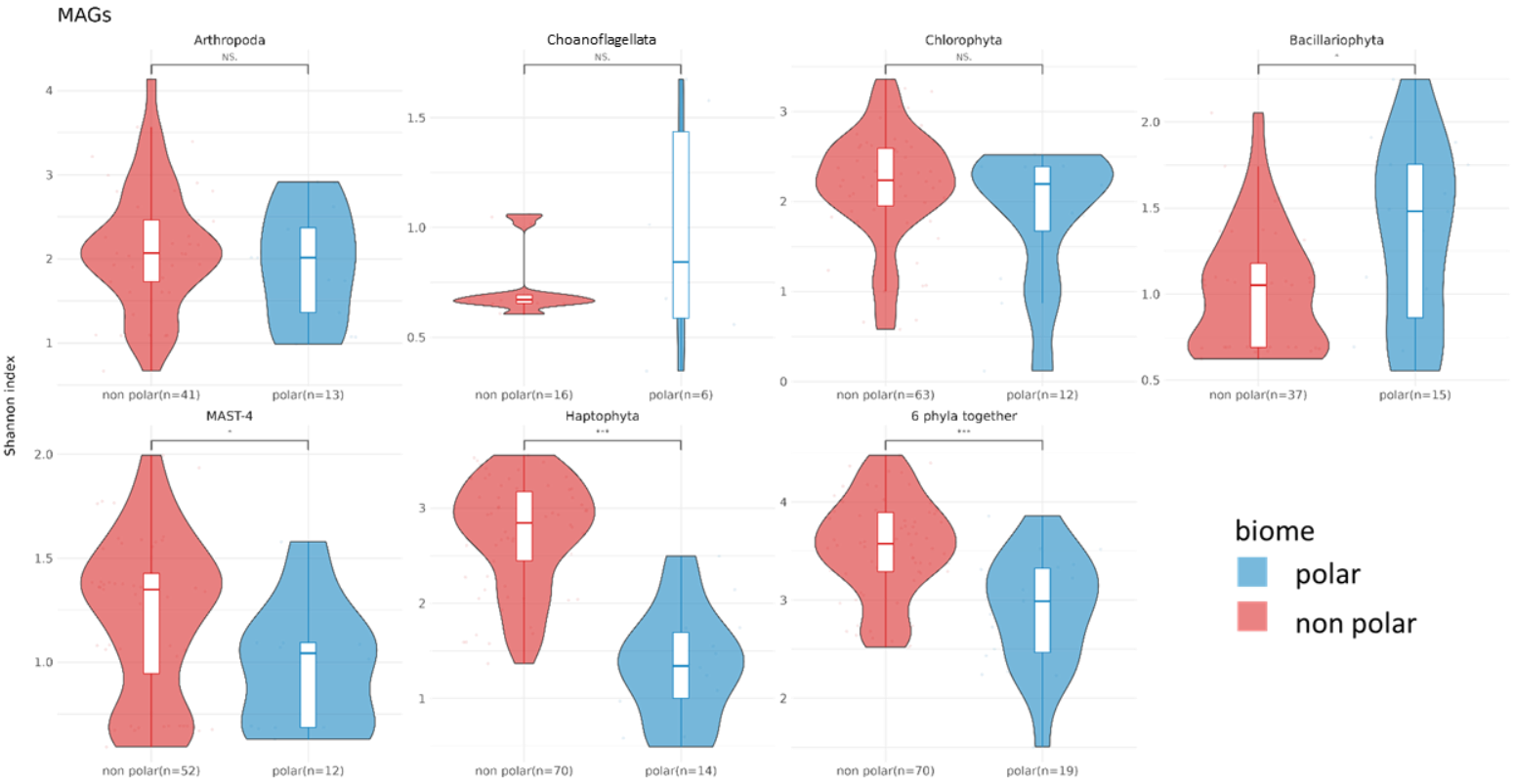


Figure 71. A-diversité en espèces par phylum dans les biomes polaires et non polaires. Les stations contenant moins de trois MAGs dans le phylum d'intérêt ont été enlevées de l'analyse. Le nombre de stations dans chaque biome par phylum est indiqué entre parenthèses en abscisses. Les stations dont la latitude absolue est supérieure à 60° sont classées comme étant polaire, les autres non polaires. Les tests de significativité sont des tests de Wilcoxon (***: *p*-value < 0.01; **: *p*-value < 0.05; *: *p*-value < 0.1; NS: *p*-value ≥ 0.1).

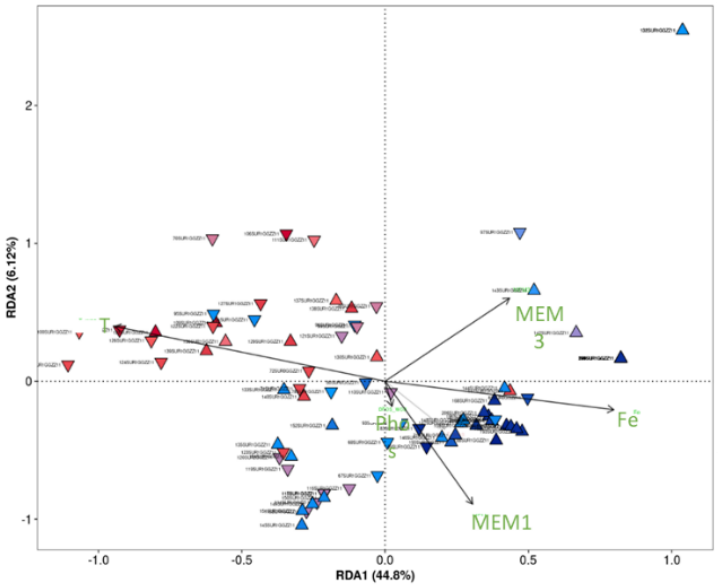
c. différences de niveau de structuration biogéographique de la distribution des folds en fonction de leur classe d'abondance et phylum

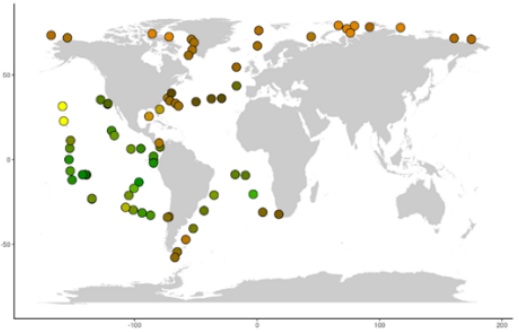
Afin de mieux comprendre les résultats d'α-diversité de la partie précédente, la distribution biogéographique des catégories et de leurs noyaux ont été étudiés à l'aide de RDAs. Leur r^2 ajusté permet de quantifier la structuration biogéographique de la distribution (voir Matériel et Méthode). Dans la plupart des phyla, la distribution des folds de chaque catégorie d'abondance et de leurs noyaux ne présente pas de structuration biogéographique (Table 3,4). Aucune catégorie n'est systématiquement plus ou moins explicative dans tous les phyla ou n'est statistiquement associée à un coefficient de détermination plus ou moins élevé (ANOVA [86] : F=0,32, p=0,73). La distribution des folds abondants, des folds du noyau des catégories abondante et intermédiaire chez les Chlorophytes présente néanmoins une certaine structuration biogéographique, principalement sur la première dimension des RDAs (jusqu'à 48%). La part de la variance expliquée par les variables explicatives choisies peut aller jusqu'à 50% dans ce phylum. De la structuration est aussi observée chez les Choanoflagellés dans les catégories intermédiaires et abondants et leurs noyaux, et dans une moindre mesure chez les Haptophytes et les MAST-4 dans le noyau des catégories intermédiaire et abondant et la catégorie intermédiaire, respectivement (Table 4). Pour les dbRDA avec des coefficients de détermination supérieurs à 0.3, la température annuelle médiane est toujours significative. La concentration en phosphate l'est également de façon quasi systématique. La concentration en fer n'est significative que pour les Choanoflagellés et les Chlorophytes. L'écart type de la température annuelle n'est significatif que chez les Choanoflagellés et les Haptophytes (Table 5). Dans ces combinaisons phylum-catégorie, les paramètres environnementaux participent donc dans une certaine mesure à la structuration observée. Il n'y a pas forcément de cohérence entre les résultats de structuration et les différences significatives d'α-diversité observées entre biomes polaires et non polaires (Fig.70 ; Table 3); par exemple, la catégorie abondante des Chlorophytes présente une structuration mais pas de différences significatives d'α-diversité entre stations polaires et non polaires. Parmi toutes les combinaisons phylum-catégorie dont la distribution environnementale présente un certain degré de structuration, les Moran Eigenvector Maps (MEMs) ne sont significativement explicatifs que chez les Chlorophytes (Figure 72; Table 6). Les MEMs permettent de convertir la distribution géographique des stations en vecteur de valeurs, chaque MEM représentant un axe de variation (par exemple la distribution est-ouest, nord-sud ou par bassin). Cela veut dire la distribution des folds de Chlorophytes de la catégorie abondant, ainsi que la distribution de ceux de son noyau et du noyau de la catégorie intermédiaire est structurée en partie par la géographie.

Afin de vérifier si ce phénomène provient des MAGs eux-mêmes, l'existence d'une potentielle structuration de leur distribution biogéographique a également été testée (Table 3 ; Table 4). Au final, la distribution des MAGs des phyla sélectionnés ici ne présente pas de structuration biogéographique. À noter que ces résultats sont en contradiction avec ceux d'autres publications obtenus à partir des abondances relatives estimées par métabarcoding, qui montraient que les communautés de Bacillariophytes et de Mamiellophyceae sont structurées au moins par les fronts Arctiques et Antarctiques [98]. La distribution des folds du noyau de la catégorie intermédiaire chez les MAGs Chlorophytes est cependant cohérente avec ces résultats. Il est donc possible que la distribution biogéographique des MAGs ne soit pas complètement représentative de celle des communautés complètes, mais que pour les Chlorophytes, cela ne fait pas disparaître le signal à l'échelle de certains folds.

A

dbRDA with folds from the core intermediate category (22 folds; adjusted r^2=0.491)





dbRDA with folds from the core abundant category (24 folds; adjusted r^2=0.473)

significant (p<0.01) a

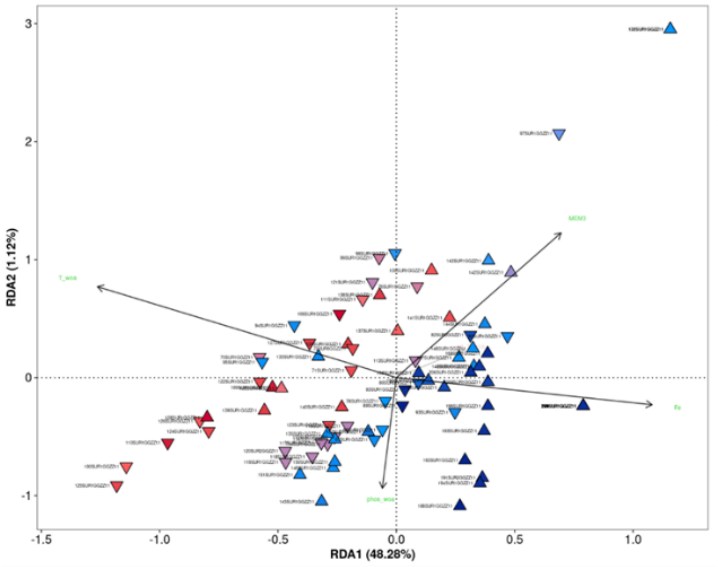
not significant (p≥0.01) a

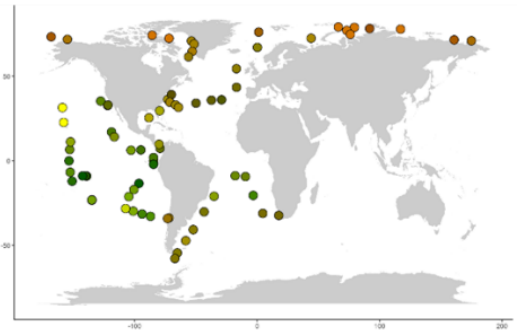
north △

south ▽

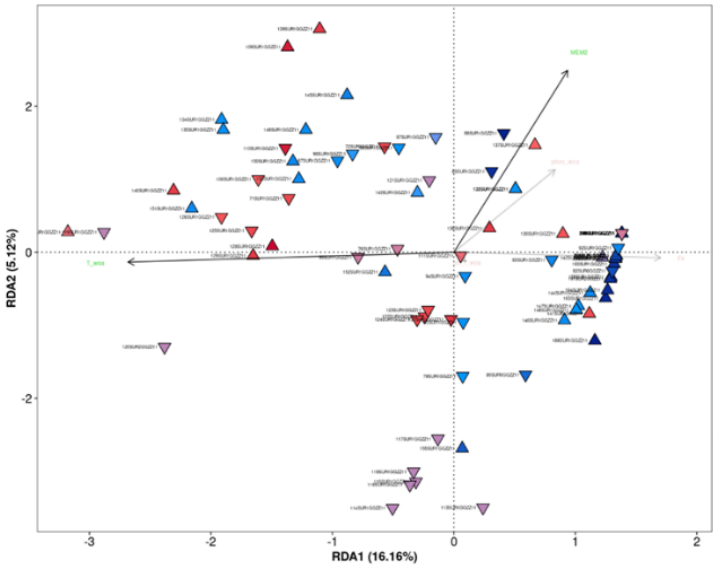
absolute latitude (°)

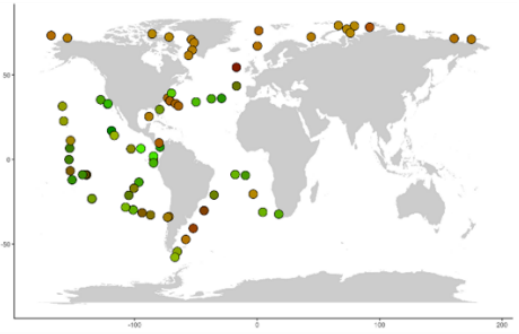
0 15 30 80





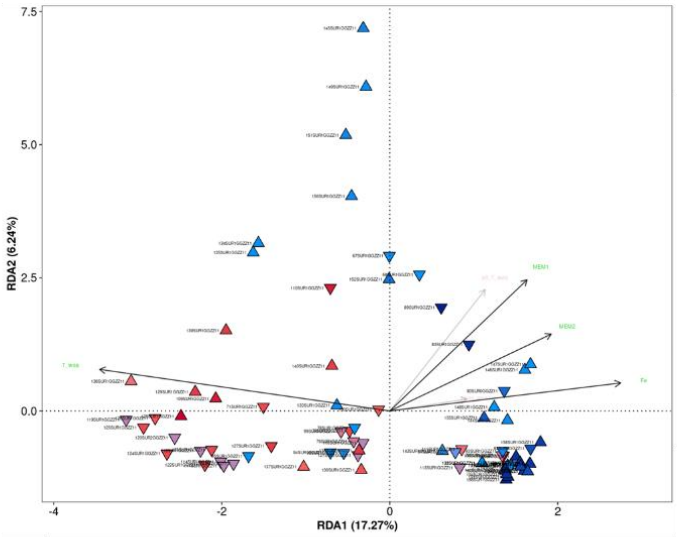
dbRDA with folds from the core rare category (10 folds; adjusted r^2=0.192)

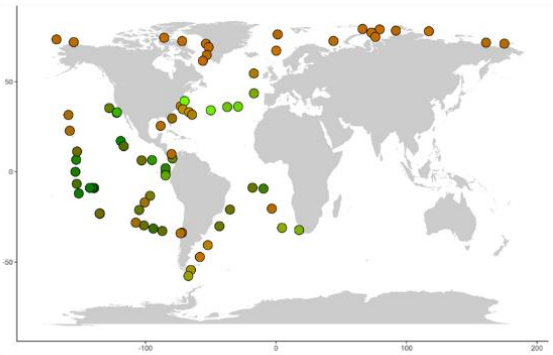




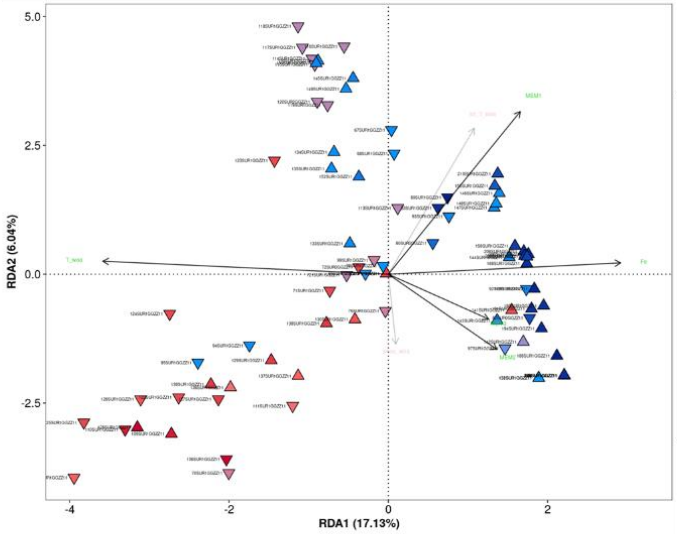
B

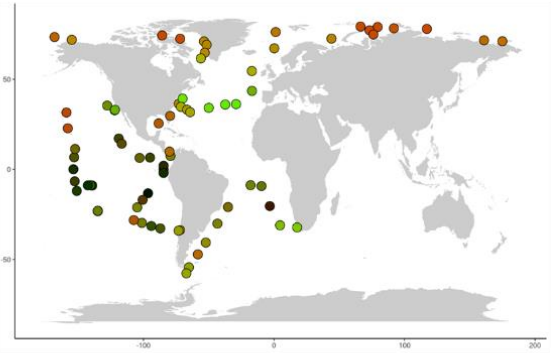
dbRDA with folds from the non-core rare category (58 folds; adjusted r^2=0.238)





dbRDA with folds from the non-core intermediate category (633 folds; adjusted r^2=0.227)





significant (p<0.01) a

not significant (p≥0.01) a

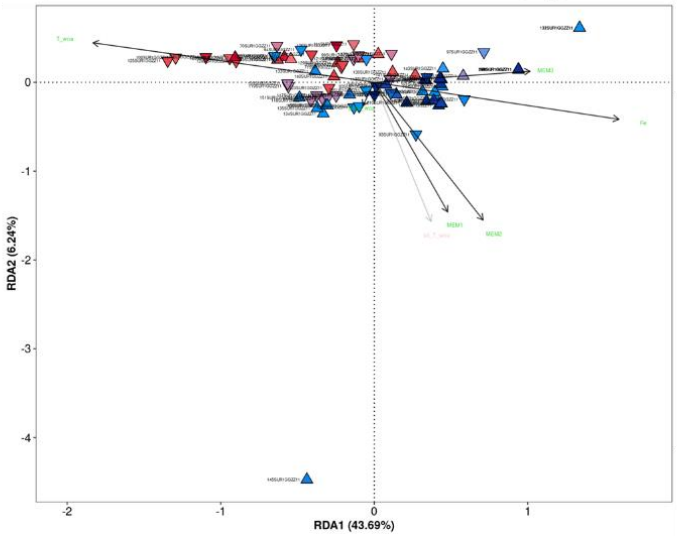
north △

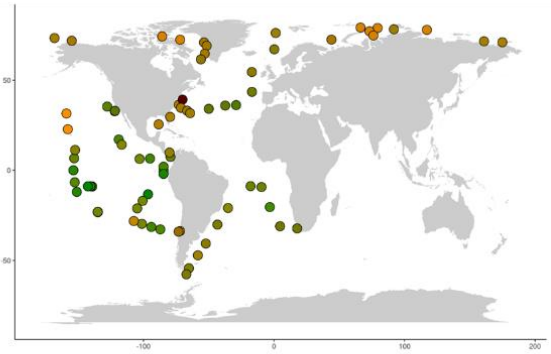
south ▽

absolute latitude (°)

0 15 30 80

dbRDA with folds from the non-core abundant category (46 folds; adjusted r^2=0.499)





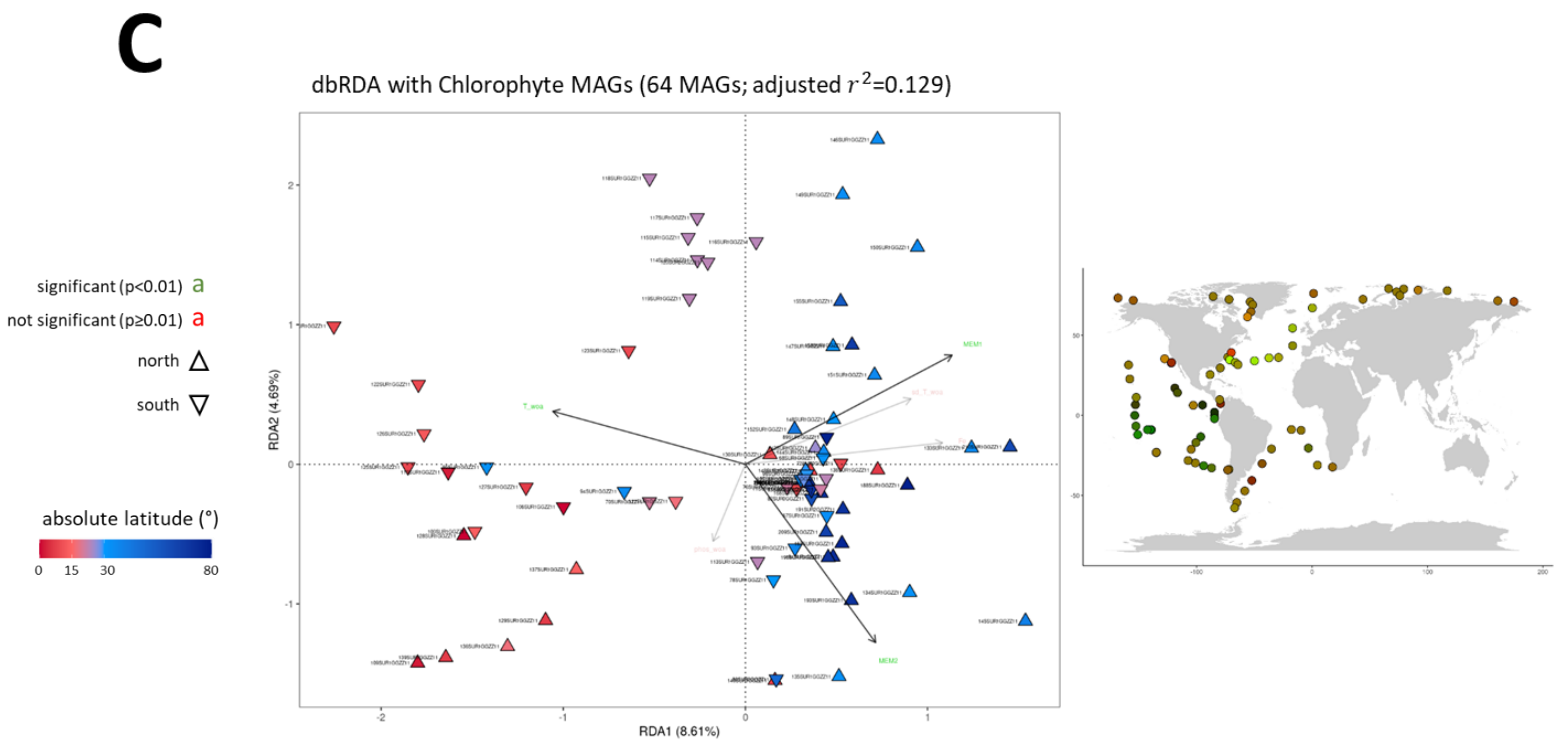


Figure 72. Structuration de la distribution biogéographique des folds et des MAGs chez les Chlorophytes. (A) Pour les folds dans le noyau de chaque catégorie. **(B)** Pour les folds dans chaque catégorie. **(C)** Pour les MAGs. **(A-C)** dbRDA biplot (à gauche) et carte RG (Red-Green) correspondante (à droite). Le r^2 ajusté indiqué dans le sous-titre de chaque biplot correspond au coefficient de détermination ajusté (proportion de variance de la matrice d'AVs des folds expliquée par les variables sélectionnées). Dans les biplot, les triangles représentent les stations et sont colorés en fonction de la latitude absolue. Ceux qui pointent vers le haut correspondent à des stations de latitude positive et ceux vers le bas à des stations de latitude négative. Les flèches noires pointant vers du texte en vert indiquent les variables significatives au seuil de confiance 1%. Les cartes RG montrent les station TO colorées en fonction de leur position dans les deux premières dimensions du biplot correspondant (la dimension 1 est associée à un degré de rouge et la dimension 2 à un degré de vert). La troisième dimension n'étant jamais significative, elle n'est pas utilisée et il n'y a pas de bleu. Les résultats pour les autres phyla peuvent être retrouvés dans les Table 4,5 et 6. Les RDAs et cartes RGs associées pour tous les autres phyla sont accessibles à ce lien : https://doi.org/10.5281/zenodo.14935989

category	phylum	Arthropoda	Choanoflagellata	Chlorophyta	Haptophyta	Bacillariophyta	MAST-4
core	rare	0,132	0,215	0,192	0,113	0,196	0,137
	intermediate	0,139	0,337	0,491	0,372	0,106	0,14
	abundant	0,076	0,456	0,473	0,303	0,126	0,066
non-core	rare	0,094	0,288	0,238	0,132	0,142	0,191
	intermediate	0,141	0,445	0,227	0,23	0,171	0,356
	abundant	0,073	0,431	0,499	0,251	0,125	0,095
MAGs		0,097	/	0,129	0,169	0,088	0,175

Table 4. Valeurs des coefficients de determination des dbRDA pour chaque catégorie de folds, leurs noyaux et les MAGs. Le « / » dans la case des MAGs x Choanoflagellata indique qu'il n'y a pas assez de MAGs de ce phylum dans certaines stations pour réaliser une dbRDA à l'échelle des espèces. L'ensemble des dbRDAs et leurs cartes RG associées sont accessibles à ce lien : https://doi.org/10.5281/zenodo.14935989

A

category	phylum	Arthropoda	Choanoflagellata	Chlorophyta	Haptophyta	Bacillariophyta	MAST-4
core	rare	14,89	22,6	16,16	12,69	17,29	10,68
	intermediate	15,1	29,17	44,8	34,46	10,15	12,82
	abundant	9,1	47,16	48,28	25,77	13,89	7,32
non-core	rare	10,35	23,72	17,27	11,3	10,3	12,49
	intermediate	12,25	42,18	17,13	18,28	14,51	33,32
	abundant	9,32	44,63	43,69	22,49	13,74	9,49
MAGs		8,38	0	8,61	8,4	6,39	10,59

B

category	phylum	Arthropoda	Choanoflagellata	Chlorophyta	Haptophyta	Bacillariophyta	MAST-4
core	rare	2,41	3,06	5,12	2,02	3,5	5,01
	intermediate	2,66	8,01	6,12	4,02	2,8	3,08
	abundant	4,42	1,95	1,12	6,71	2,21	3,15
non-core	rare	2,29	6,1	6,24	3,28	4,12	5,81
	intermediate	4,08	4,76	6,04	6,27	3,77	2,83
	abundant	3,51	2,04	6,24	4,89	2,15	3,53
MAGs		3,75	0	4,69	6,04	4,76	6,84

Table 5. Pourcentage de variance expliqué par les deux premiers axes des dbRDAs. (A) Pourcentage de variance dans la première dimension. **(B)** Pourcentage de variance dans la deuxième dimension. L'ensemble des dbRDAs et leurs cartes RG associées sont accessibles à ce lien : https://doi.org/10.5281/zenodo.14935989

category		Arthropoda	Choanoflagellata	Chlorophyta	Haptophyta	Bacillariophyta	MAST-4
core	rare	T, Fe	T, P	MEM2, T	T	T, Fe, P	MEM1, T
	intermediate	T	T, Fe, P	MEM1, MEM3, T, Fe	T, sd(T), P	T, sd(T)	T, P
	abundant	T	T, Fe, P	MEM3, T, Fe, P	T, sd(T), P	T	/
non-core	rare	T, Fe	T, sd(T), P	MEM1, MEM2, T, Fe	T, sd(T), Fe, P	MEM1, T, Fe, P	MEM1, T, Fe, P
	intermediate	T, sd(T), Fe	T, Fe, P	MEM1, MEM2, MEM3, T, Fe	T, P	T, Fe, P	T, P
	abundant	T	T, Fe, P	MEM1, MEM2, MEM3, T, Fe, P	T, sd(T), P	T	T, P
MAGs		MEM1, T	/	MEM1, MEM2, T	MEM1, MEM2, T, sd(T), P		MEM1, T, P

Table 6. Variable significativement explicative de la structuration des dbRDAs. La significativité est au seuil de confiance 1%. Les Moran Eigenvector Maps (MEM) sont indiquées en jaune, la médiane de la température annuelle (T) en rouge, l'écart-type de la SST annuelle (sd(T)) en violet, la concentration en fer (Fe) en gris, la concentration en phosphate en vert. Il n'y a pas assez de MAGs Choanoflagellés pour réaliser une dbRDA. L'ensemble des dbRDAs et leurs cartes RG associées sont accessibles à ce lien : https://doi.org/10.5281/zenodo.14935989

Pour mieux comprendre la structuration de la distribution biogéographique des 24 folds du noyau de la catégorie intermédiaire chez les Chlorophytes, la distribution de chacun d'entre eux dans les stations et les génomes a été représentée (Figure 73).

Pratiquement tous les folds présentent une certaine variabilité latitudinale avec des AVs globalement plus faibles aux pôles et en particulier l'Arctique, à l'exception des folds 3.30.559, 1.10.30, 1.10.20, 3.10.110 et 3.60.15. Le fold 1.20.1740 a la plus forte variabilité d'AV et atteint son abondance maximale à la station 125, située dans le Pacifique tropical (Figure 73 B). Pour vérifier si cet effet est dû à la présence de communautés d'espèces distinctes entre les régions méridionales et septentrionales, dans lesquelles les OVs de ces folds seraient différentes, les AVs des MAGs dans les stations TO sont représentées (Figure 73 A). Les MAGs ont une distribution caractérisée par un fort signal de présence absence, la plupart d'entre eux étant présents dans moins de la moitié des stations. Du point de vue taxonomique, les communautés d'espèces sont différentes selon les bassins. Celles des pôles ne sont composées que de *Prasinoderma,* de *Micromonas* et de *Bathycoccus* qui sont aussi trouvées sous des latitudes plus basses. Les communautés non polaires rassemblent tous les autres genres dont la plupart sont complètement absents du biome polaire. Le Pacifique Sud rassemble une diversité particulièrement importante de MAGs de *Micromonas*, *Ostreococcus*, *Chloroparvula, Sister Pycnococcus* et *Chloropicon*. Le Pacifique Nord présente plus de similarités en terme de composition de la communauté avec l'Atlantique Sud. Enfin la communauté de l'Atlantique Nord partage des MAGs à la fois avec les pôles et l'Atlantique Sud. Les communautés d'espèces de Chlorophytes semblent donc être principalement structurées par le bassin, puis par la latitude. Quant aux OVs des folds du noyau de la catégorie intermédiaire dans les MAGs Chlorophytes (Figure 73 C), leur variabilité entre MAGs est relativement faible mais peut être multipliée par un facteur d'au moins dix pour certain folds (par exemple 1.20.1740 (Amino Acid/polyamine transporter I), qui peut passer d'une OV de 28 à 4 dans certains MAGs). Ce phénomène ne semble pas être lié à la taxonomie mais plutôt à la complétion des MAGs, en particulier pour les valeurs extrêmes (l'OV la plus élevée du fold 1.10.20 qui vaut 28 est par exemple atteinte dans le MAG TARA_ARC_108_MAG_00063, un *Micromonas* avec la complétion la plus élevée de tous les MAG de Chlorophytes (91%)). Seule une partie des folds ayant des AVs avec une variabilité latitudinale élevée ont également des OVs qui varient de façon importante dans les MAGs (principalement 3.40.1110 (calcium-transporting ATPase, cytoplasmic domain N), 1.20.1110 (calcium-transporting ATPase, transmembrane domain), 2.70.150 (calcium-transporting ATPase, cytoplasmic transduction domain A) et 1.20.1050).

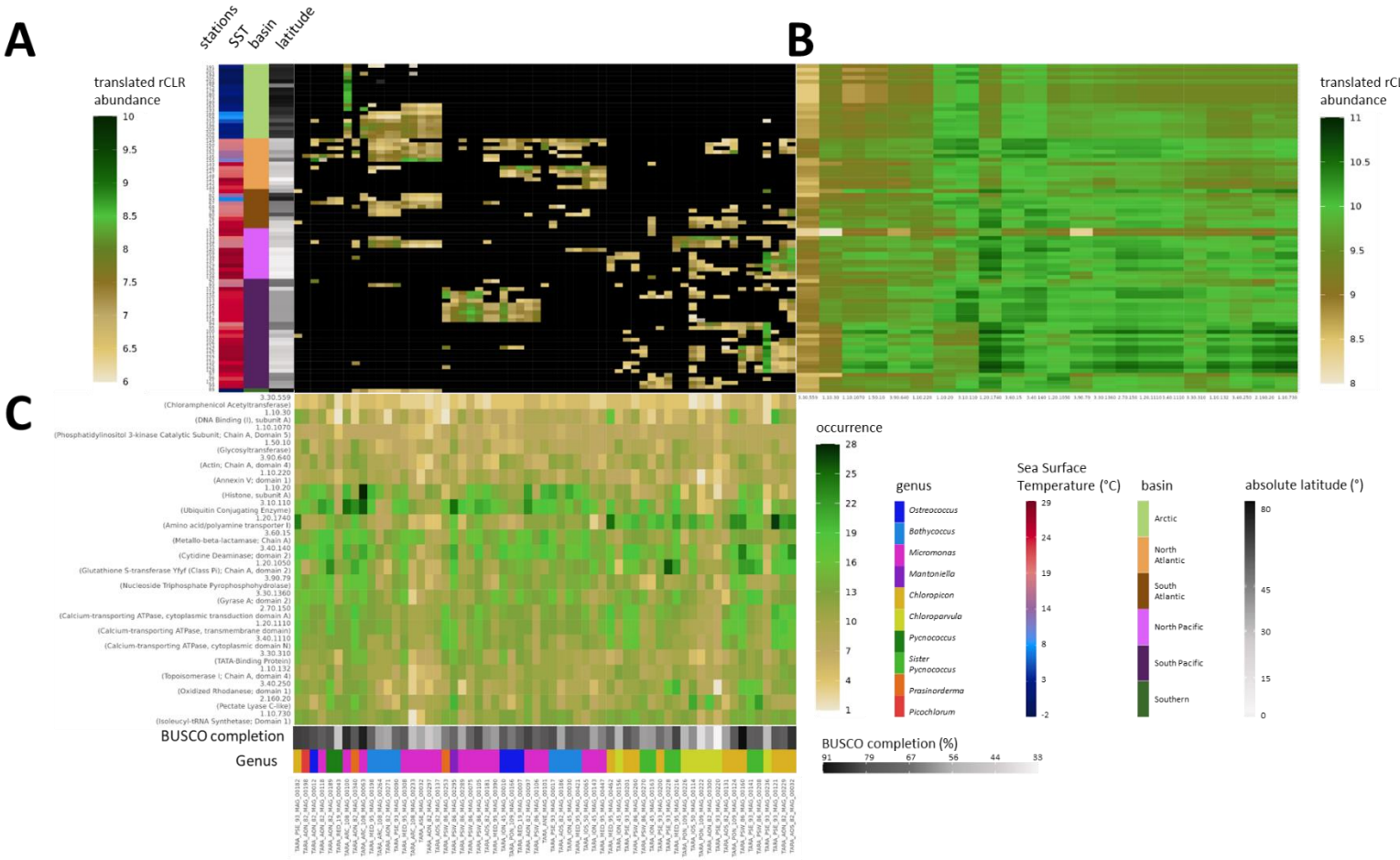


Figure 73. Distribution des MAGs, AVs des folds dans les stations et OVs des folds du noyau de la catégorie intermédiaire dans les MAGs Chlorophytes. (A-B) AVs des MAGs et des folds dans les stations. L'ordre des stations (en ordonées) provient du clustering sur les AVs des folds, après groupement par bassin. **(A)** AVs des MAGs dans les stations. L'identifiant du MAG est indiqué sur l'abscisse de **(C)**. Les abondances transformées en CLR et translatées sont indiquées à gauche. Les couches d'informations supplémentaires à gauche sont, de gauche à droite, la SST lors du prélèvement, le bassin et la latitude absolue (voir valeurs en bas à droite). **(B)** AVs des folds dans les stations. L'identifiant CAT des folds est indiqué en abscisses. Les abondances transformées en CLR et translatées sont indiquées à droite. **(C)** OVs des 22 folds du noyau de la catégorie intermédiaire. L'identifiant CAT du fold ainsi que le nom fourni par CATH sont indiqués sur l'axe des ordonnées, l'identifiant du MAG est sur l'axe des abscisses. Les couches en bas sont de haut en bas la BUSCO complétion (%) et le genre auquel le MAG appartient (voir valeurs en bas à droite

4/ prédiction des facteurs environnementaux influençant la distribution biogéographique des folds à l'aide d'approches d'apprentissage automatique

Dans cette partie, la distribution de chaque fold des trois noyaux (Table 1) a été prédite à l'échelle de l'ensemble des océans dans le but d'identifier le facteur impactant le plus cette distribution à une échelle globale. Les prédictions ont été réalisées avec *Climap*, un outil développé par Téo Lemane qui rassemble plusieurs algorithmes d'apprentissage automatique et est basé sur l'approche développée par Frémont *et al*. [111]. Il exploite la diversité des contextes environnementaux des stations TO et les AVs des 56 folds des noyaux des trois catégories dans les stations TO de de surface dans la fraction de taille 0.8-2000µm, et ce pour chacun des phyla sélectionnés dans les paragraphes précédents. Son but principal dans ce contexte est d'identifier des liens potentiels entre valeurs de paramètres environnementaux et abondances relatives. La répétition de l'analyse dans les six phyla pour chaque fold a à la fois pour but de limiter des potentiels biais qui seraient propre à un seul phylum, mais aussi de poser la question sous l'angle de la convergence. En effet, l'idée est d'identifier des folds qui auraient été sélectionnés indépendamment de leur fonction mais bien pour leurs propriétés biophysiques. De tels folds auraient en théorie la même sensibilité face aux variations de paramètres environnementaux dans tous les phyla, ou, dit autrement, auraient les mêmes facteurs les plus explicatifs de leur distribution dans tous les phyla.

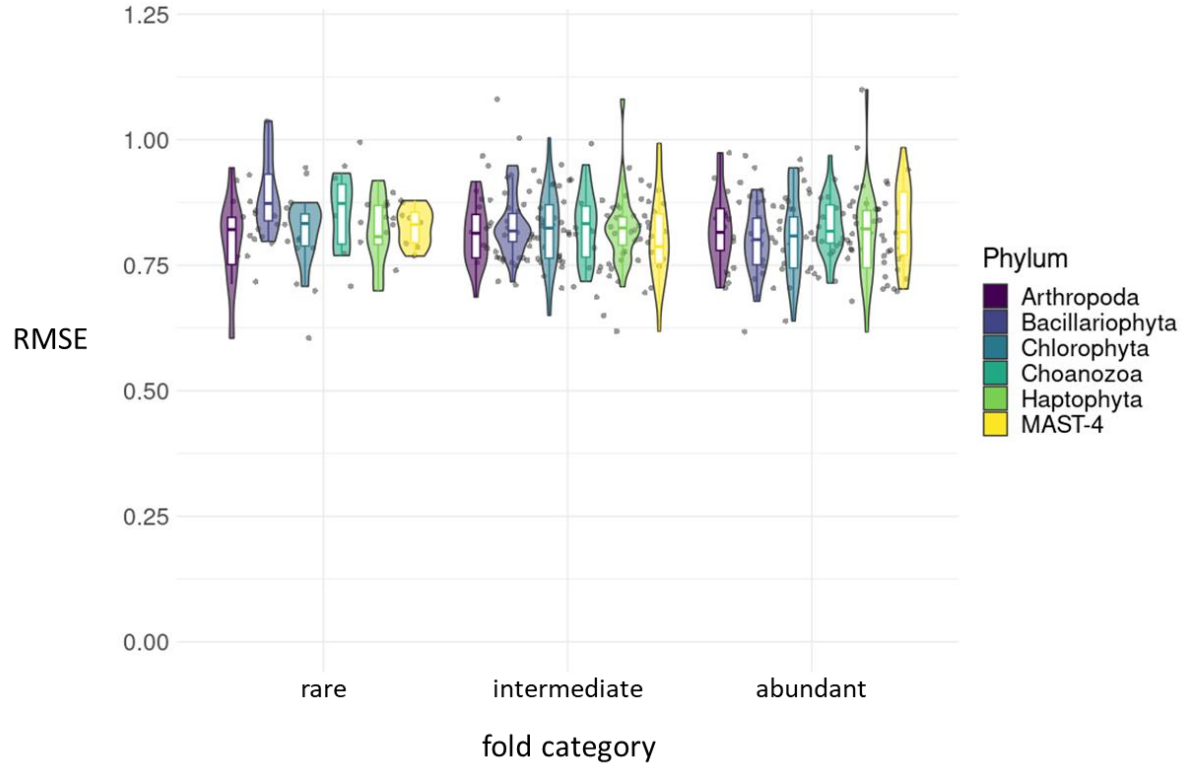


Figure 74. Racine de l'Erreur Quadratique Moyenne (RMSE) de chaque modèle par catégorie et phylum. Chaque point représente un modèle. Les violin plots sont organisés de gauche à droite par catégorie environnementale, puis par phylum dans chaque catégorie.

L'entraînement de chaque modèle (une combinaison fold-phylum) a mis entre trois et six heure, et il y a au total 336 modèles avec chacun 9600 combinaisons d'hyper-paramètres. Chacune de ces

combinaisons retourne un RMSE (Erreur Quadratique Moyenne) pour le modèle ; il y a donc par modèle 9600 valeurs de RMSE. Cette grandeur permet de quantifier l'écart entre valeurs prédites par le modèle et valeurs observées utilisées pour l'entraînement. Elle peut être interprétée comme le nombre d'écart-types séparant en moyenne les données prédites et les données d'entraînement ; si elle est inférieure à un, cela veut dire qu'il y a moins d'un écart type en moyenne, donc qu'il n'y a pas de différences significatives entre valeurs prédites et valeurs de l'entraînement. Il vaut donc mieux choisir pour chaque modèle la combinaison d'hyper-paramètres conduisant au RMSE le plus faible. Les valeurs des hyper-paramètres de chaque modèle sont extrêmement variables, sans tendance notable par catégorie ou phylum. Par exemple pour les réseaux de neurones, des structures allant d'une à quatre couches cachées, chacune contenant un à cinq neurones ont été sélectionnées. La comparaison des modèles entre eux, basée sur leurs valeurs de RMSE, révèle qu'il n'y a pas non plus de différences significatives de cette grandeur entre catégories et, au sein de chaque catégorie, entre phyla (Figure 74). Le RMSE moyen se situe entre 0.76 et 0.87. Les meilleurs modèles, dont les RMSE sont inférieurs à 0.7 (11 modèles), ne correspondent jamais au même fold dans plusieurs phyla. La majorité appartient à la catégorie abondant. Le modèle avec le RMSE le plus faible correspond néanmoins au fold 3.40.449 (rare) chez les Arthropodes.

Le but principal de ces modèles, une fois leur pertinence validée, est d'évaluer l'effet de chaque paramètre environnemental sélectionné sur la prédiction et d'identifier le paramètre avec le pouvoir prédictif le plus fort. D'un point de vue biologique, il pourrait donner une indication sur la variable ayant le plus d'impact sur la l'abondance d'un fold donné dans un écosystème, et donc potentiellement permettre de mieux comprendre les résultats sur les tests de structurations présentés dans les Tables 3,4.

Dans un premier temps, l'importance des variables a été mesurée à l'aide de tests de permutations. Les valeurs associées à une variable d'entraînement sont mélangées et l'entraînement est relancé sur ces nouvelles valeurs. La comparaison des R^2 permet de quantifier l'écart entre les valeurs prédites avec le nouveau modèle et l'ancien. En répétant l'opération pour chaque variable environnementale, il est possible d'évaluer celle ayant le plus d'impact sur les résultats du modèle. Les résultats sont ici assez clairs (Figure 75, Figure 76). Chaque variable a une importance significativement différente des autres, avec par ordre d'importance décroissant, la concentration en Fer, la température médiane, l'amplitude de température et la concentration en nitrate. Cet ordre est le même dans tous les phyla et pour toutes les catégories de folds.

Dans un second temps, l'importance des variables a été mesurée à l'aide du module SHAP [365] (Figure 77 ; Figure 78). Il quantifie, pour chaque prédiction, de combien la valeur de chaque paramètre environnemental à la localisation de la prédiction fait dévier la valeur de celle-ci par rapport à la valeur médiane des prédictions. Le module ne permet pas d'obtenir des résultats globaux, contrairement aux tests de permutations, mais permet de comprendre plus en détail la façon dont chaque variable affecte la valeur de la prédiction ; les deux approches sont donc complémentaires.

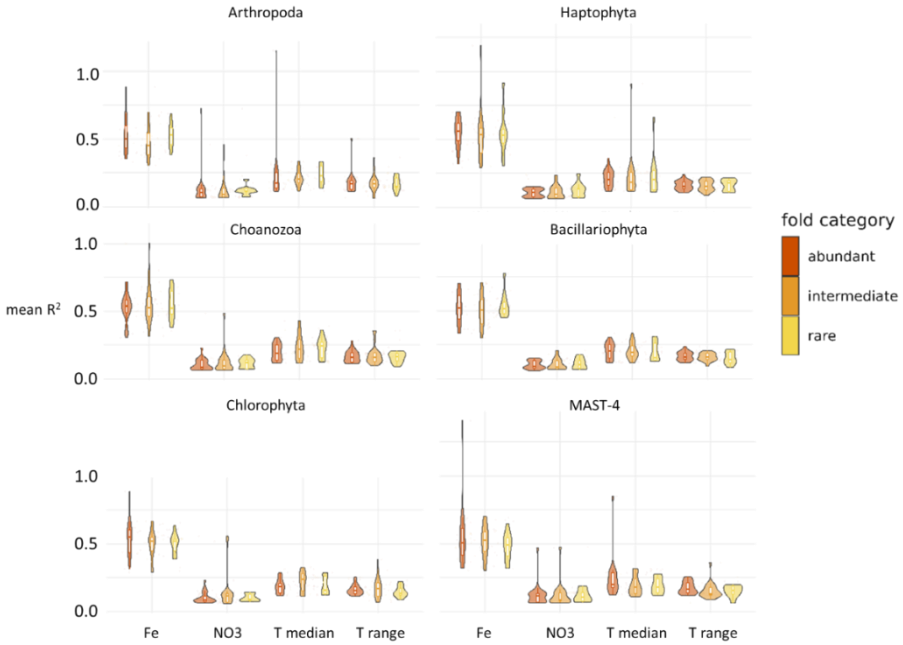


Figure 75. Écart de R^2 moyen des tests de permutation pour chaque variable environnementale par catégorie et phylum. L'écart de R^2 moyen quantifie l'erreur moyenne réalisée par le modèle lors de différentes permutations de valeurs de chaque paramètre. Plus il est élevé et plus l'erreur est grande, donc plus la variable a d'impact sur les résultats du modèle.

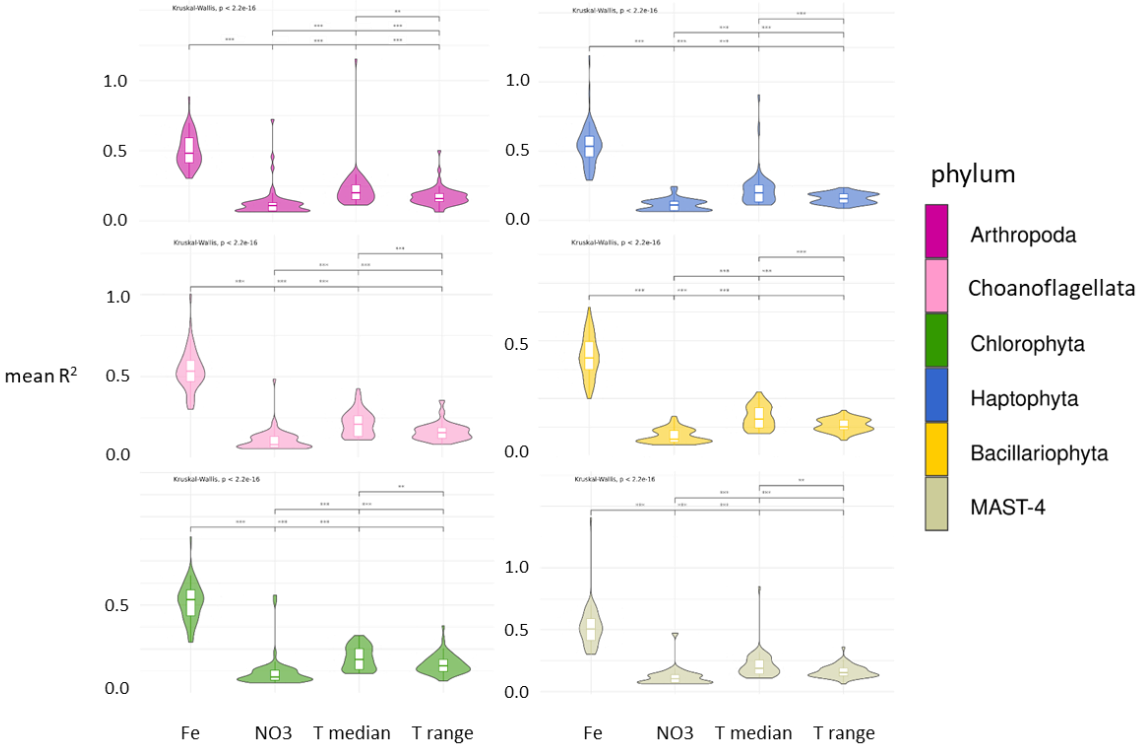
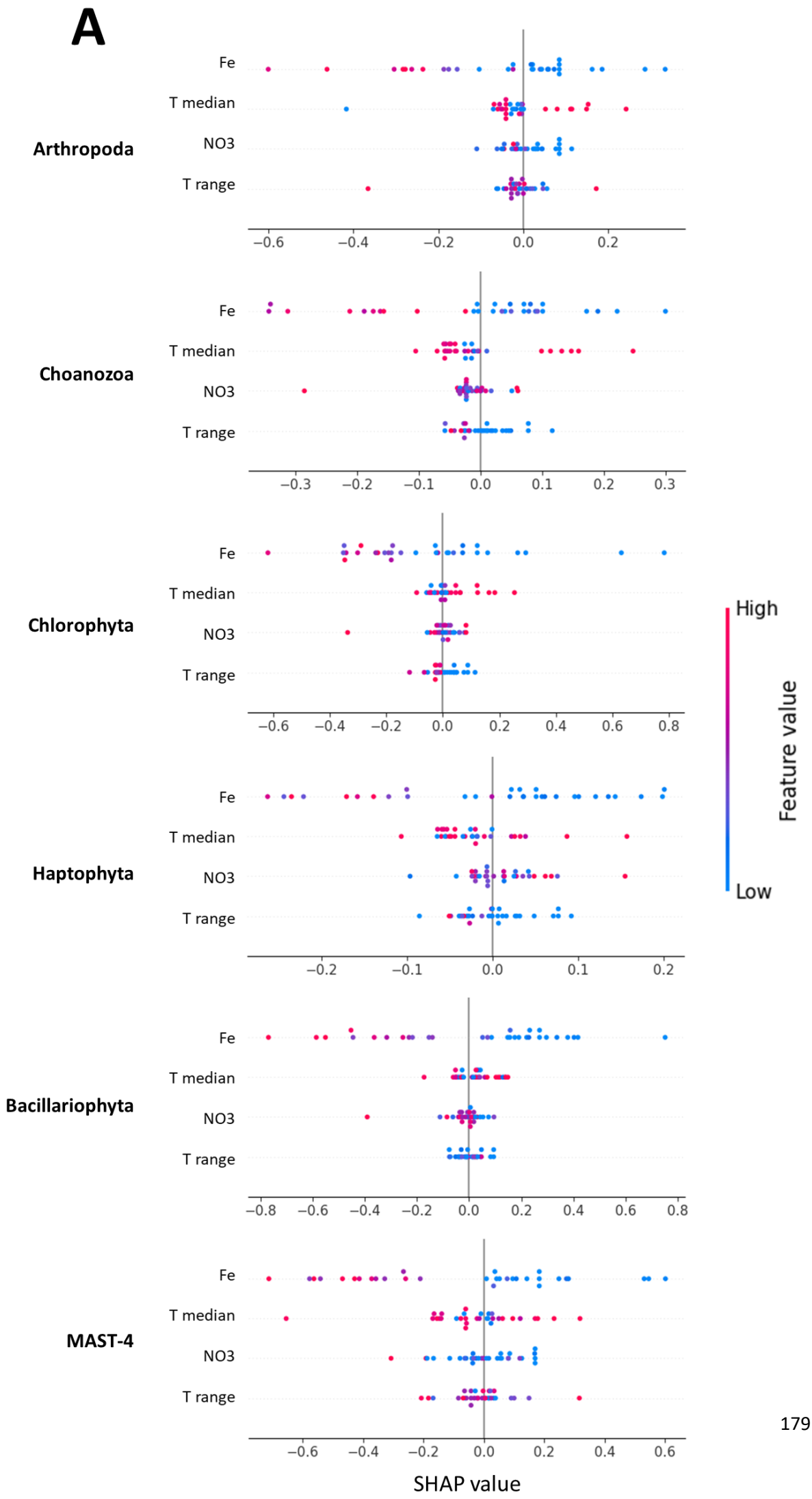
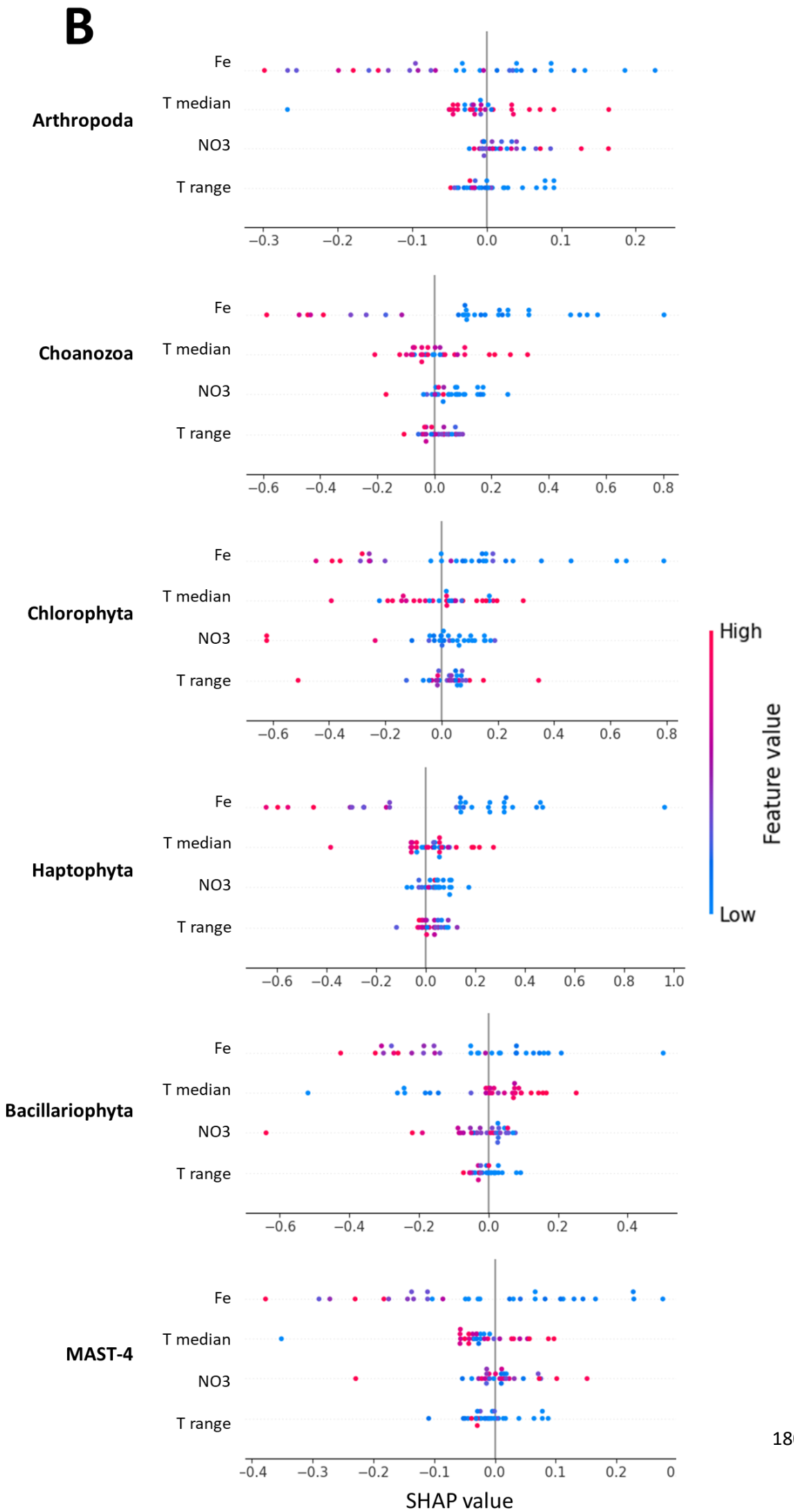


Figure 76. Significativité des écarts de R^2 moyen des tests de permutation pour chaque variable environnementale par phylum. L'écart de R^2 moyen quantifie l'erreur moyenne réalisée par le modèle lors de différentes permutations de valeurs de chaque paramètre ; plus il est élevé et plus l'erreur est grande, donc plus la variable a d'impact sur les résultats du modèle. L'existence de différences significative de R^2 entre variable a d'abord été testée avec un test de Kruskall-Wallis puis les comparaisons par paires ont été réalisées avec un test de Wilcoxon (*** : *p*-value < 0.01).


A
Arthropoda
Choanozoa
Chlorophyta
Haptophyta
Bacillariophyta
MAST-4
Fe
T median
NO3
T range
High
Feature value
Low
SHAP value


B
Arthropoda
Choanozoa
Chlorophyta
Haptophyta
Bacillariophyta
MAST-4
Fe
T median
NO3
T range
High
Low
Feature value
SHAP value

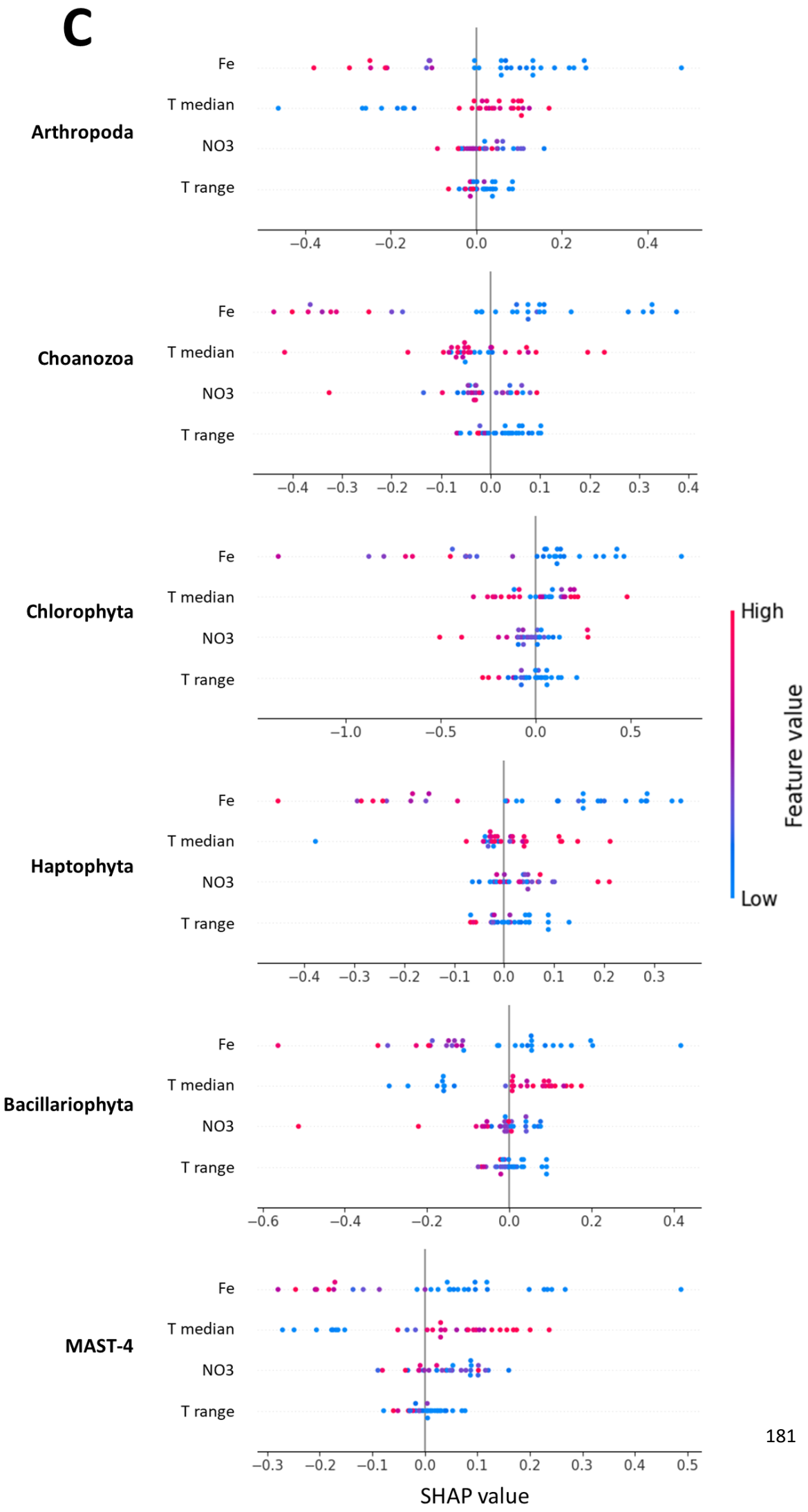

C
Arthropoda
Choanozoa
Chlorophyta
Haptophyta
Bacillariophyta
MAST-4
Fe
T median
NO3
T range
High
Low
Feature value
SHAP value

Figure 77. Exemple de résultats du module SHAP sur un fold de chaque catégorie environnementale par phylum. Les variables environnementales sont classées par amplitude décroissante de valeur SHAP sur l'axe des ordonnées, celle ayant la plus grande étant en haut. La valeur SHAP (abscisses) quantifie l'impact de la variable sur le modèle, comme indiqué par la barre de couleur à droite de chaque figure (intitulée « Feature value »). La couleur des points correspond à la valeur des paramètres pour chaque prédiction. Comme il y a 37348 prédictions par folds et phylum, un nombre réduit de points ont été représentés dans un souci de lisibilité. **(A)** Valeurs SHAP pour le fold rare 1.10.1500. **(B)** Valeurs SHAP values pour le fold intermédiaire 2.160.20. **(C)** Valeurs SHAP pour le Rossmann fold. Les figures SHAP pour tous les folds du noyau des trois catégories d'abondance dans tous les phyla sont accessibles à cette adresse : L'ensemble des dbRDAs et leurs cartes RG associées sont accessibles à ce lien : https://doi.org/10.5281/zenodo.14935989.

Les deux méthodes retournent des résultats globalement cohérents concernant l'importance des variables : par ordre d'importance décroissant, la concentration en Fer est toujours la plus importante de loin, suivie par la température médiane, la concentration en nitrate et l'amplitude des températures. En outre, des valeurs extrêmes de R^2 sont observées pour la température ou la concentration en nitrate pour certains folds dans certains phyla. Les résultats de SHAP ne sont pas toujours cohérents avec les tests de permutation.

Chez les Arthropodes, la concentration en fer n'est pas la variable ayant le plus grand R^2 et la plus grande valeur SHAP pour un fold:

- 1.20.1740 (Amino acid/polyamine transporter I. 1 superfamille, 3 domaines), pour lequel c'est la concentration en nitrate qui a le plus grand R^2 et la plus grande valeur absolue SHAP. Plus cette concentration est élevée et plus la valeur SHAP est négative, donc diminue la valeur d'abondance relative par rapport à l'abondance relative médiane (Figure 77 B). Les trois domaines adoptant ce fold sont associés à des fonctions de transporteur d'acides aminés et de polyamines. Il n'est donc pas forcément surprenant que la distribution de ce fold soit surtout impactée par la concentration en nitrate.

Chez les Chlorophytes, la distribution d'un seul fold est plus impactée par la concentration en nitrate que la concentration en fer :

- 3.40.1110 (Calcium-transporting ATPase, cytoplasmic domain N. 1 superfamille, 171 domaines). D'après SHAP, les concentrations élevées de nitrate sont responsables d'une déviation de jusqu'à -1 de l'abondance prédite de ce fold, alors que les concentrations fortes en fer entraînent une déviation de -0.6 et les concentrations faibles de +0.5. Curieusement, c'est aussi la seule combinaison fold-phylum dans laquelle l'amplitude de température peut avoir plus d'effet sur la valeur prédite que la concentration en fer, puisque des valeurs élevées d'amplitudes peuvent être responsables d'une déviation de l'abondance relative de -0.7 ou +0.6 (Figure 77 D).

La concentration en fer est le facteur le plus important dans la distribution de tous les folds testés chez les Haptophytes, les MAST-4, les Choanoflagellés et les Bacillariophytes. Les valeurs extrêmes de R^2 pour la concentration en nitrate observées dans la Fig.75 ne dépassent en fait jamais la valeur de R^2 associée à la concentration en fer pour le même fold.

De façon globale, les valeurs élevées de concentration en fer semblent systématiquement causer une déviation négative de la valeur SHAP et donc dans une certaine mesure de l'abondance relative (pouvant aller jusqu'à -1 unité d'abondance relative) (Figure 77, Figure 78). Concernant la température

médiane, les valeurs élevées sont associées à des valeurs SHAP positives, et les valeurs faibles à des valeurs SHAP négatives dans certains cas des exemples de la Fig.78: pour 3.40.50, pour les Arthropodes, Bacillariophtyes et MAST-4 ; pour 2.160.20, pour tous les phyla sauf les Bacillariophytes ; pour 1.10.1500, pour tous les phyla. Les tendances sont moins nettes pour les deux autres variables, pour lesquelles les valeurs fortes sont associées aussi bien à des valeurs SHAP négatives que positives.

La température était attendue comme la variable influant le plus sur la distribution des structures de domaines protéiques. Elle peut influencer directement la flexibilité ou la rigidité des protéines [142], [377]. Dans le milieu marin, une corrélation positive existe entre température optimale de fonctionnement des enzymes et températures moyenne [329]. La température a donc un impact très important sur la séquence. Peut-être que la concentration en fer est plus importante à l'échelle du fold. En particulier, le fer est retrouvé dans des clusters fer-souffre qui participent aux réactions d'oxydoréduction du métabolisme cellulaire [296]. Ils peuvent être trouvés dans certaines Homologies du Rossmann fold, notamment la P-loop, qui est la plus abondante dans l'environnement. Le Rossmann fold étant retrouvé en combinaison avec une grande diversité de folds, l'importance systématique de la concentration en fer dans la prédiction de la distribution des autres folds est peut-être indirectement liée uniquement à son importance pour ce fold. Il est connu que des concentrations élevées en zinc dans le milieu peuvent être corrélées à une augmentation de la fréquence des doigts de zinc dans les folds de certaines Diatomées [21].

Il est également possible que ces résultats ne soient vrais que pour les 56 folds des noyaux des trois catégories d'abondance, mais que ce ne soit pas le cas pour les autres. Enfin, il est aussi possible que cet effet soit très biaisé par la qualité des modèles. Leur entraînement n'est réalisé que sur 89 stations qui ne représentent pas l'intégralité des contextes environnementaux marins, bien qu'elles aient été sélectionnées pour minimiser cet effet. De plus, la concentration en fer utilisée ici est le seul paramètre environnemental dont les valeurs sont issues de modèles et ne sont pas comparables à des données *in situ* (sa concentration n'a pas été mesurée lors de l'expédition TO). Cependant, ces valeurs ont déjà été utilisées dans d'autres analyses sans poser de problèmes notables, mais l'attention n'avait pas particulièrement été porté sur l'importance des variables [111]. Dans tous les cas, ces modèles montrent qu'il est possible de prédire la distribution biogéographique des folds à l'échelle globale mais que les données ne permettent pas d'identifier un signal clair et robuste concernant l'importance des variables.

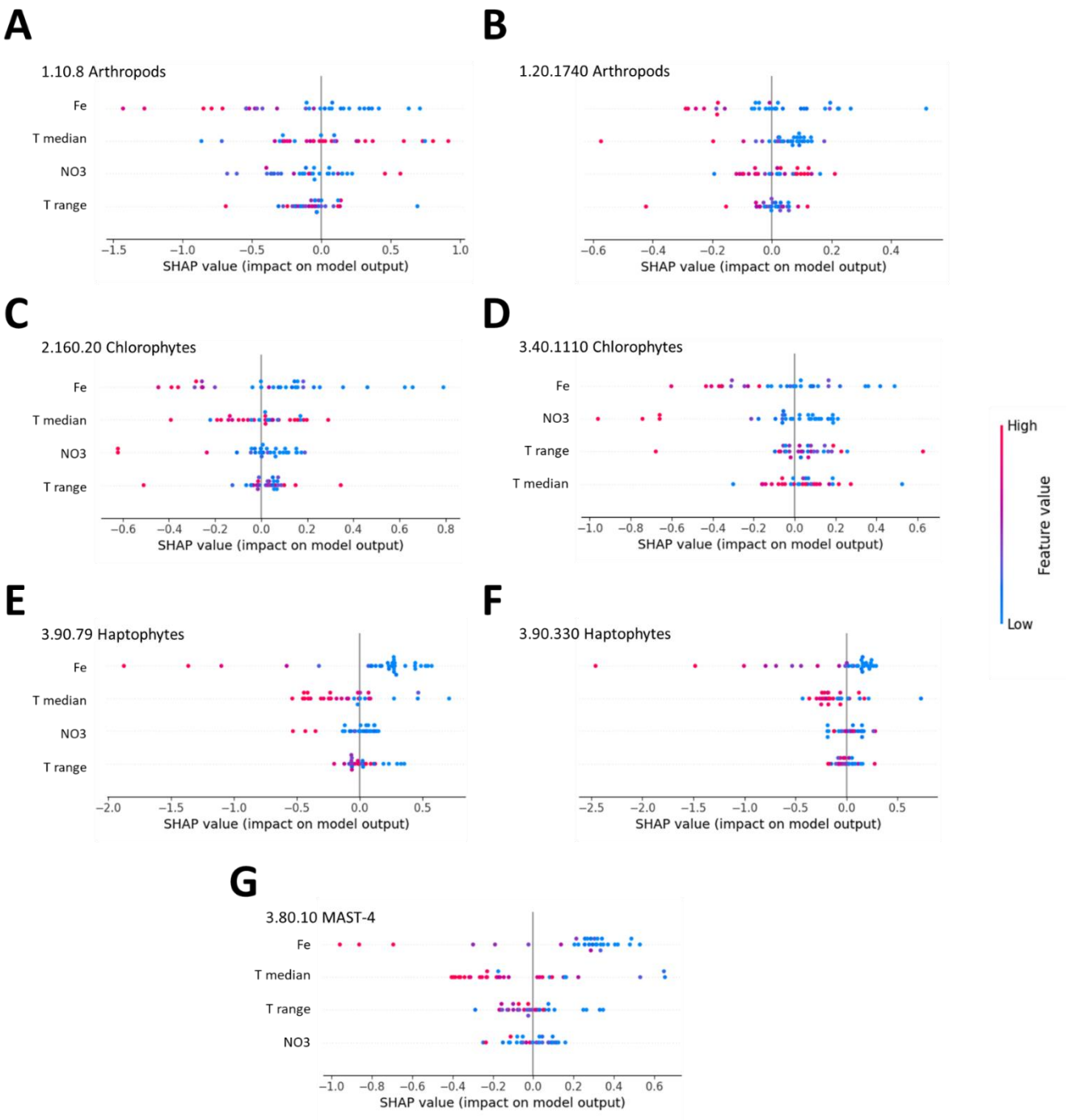


Figure 78. Valeurs SHAP pour les folds avec des valeurs extrêmes dans les tests de permutations. Les variables environnementales sont classées par amplitude décroissante de valeur SHAP value sur l'axe des ordonnées, celle ayant la plus grande étant en haut. La valeur SHAP (abscisses) quantifie l'impact de la variable sur le modèle, comme indiqué par la barre de couleur à droite de chaque figure (intitulée « Feature value »). La couleur des points correspond à la valeur des paramètres. **(A)** Valeurs SHAP pour le fold 1.10.8 chez les Arthropodes. **(B)** Valeurs SHAP pour le fold 1.20.1740 chez les Arthropodes. **(C)** Valeurs SHAP pour le fold 2.160.20 chez les Chlorophytes. **(D)** Valeurs SHAP pour le fold 3.40.1110 chez les Chlorophytes. **(E)** Valeurs SHAP pour le fold 3.90.701 chez les Haptophytes. **(F)** Valeurs SHAP pour le fold 3.90.330 chez les Haptophytes. **(G)** Valeurs SHAP pour le fold 3.80.10 chez les MAST-4. Les figures SHAP pour tous les folds du noyau des trois catégories d'abondance dans tous les phyla sont accessibles à cette adresse : L'ensemble des dbRDAs et leurs cartes RG associées sont accessibles à ce lien : https://doi.org/10.5281/zenodo.14935989.

5/ conclusion

La question principale posée par ce chapitre était « *la distribution des folds dans les océans est-elle structurée biogéographiquement, à l'instar de celle des communautés et des espèces planctoniques ?* »

Pour y répondre, j'ai d'abord étudié la distribution globale des 910 folds Eucaryotes du jeu de donnée dans les stations *Tara* Oceans. À cette échelle, aucune structuration particulière n'est observée. En revanche, les observations sur la Figure 56 dans le Chapitre II (p.137) ont montré qu'il semblait y avoir trois catégories distinctes de folds dans ces stations. À l'aide des modèles de Pareto, j'ai pu classer les folds des MAGs de six phyla Eucaryotes différents dans trois catégories –abondant, intermédiaire et rare-.

J'ai d'abord étudié l'α-diversité de ces catégories en fonction des phyla et des biomes, et montré qu'elle variait significativement dans certains phyla entre les stations polaires et non polaires. Ces différences sont particulièrement observables chez les phyla phytoplanctoniques mais dans des sens différents : l'α-diversité des folds rares et intermédiaires des Haptophytes et Chlorophytes est significativement plus élevée dans les biomes non polaires par rapport aux polaires, contrairement aux Bacillariophytes.

J'ai ensuite évalué le niveau de structuration biogéographique des folds dans les différentes catégories et phyla, et testé à quel point certains paramètres environnementaux pouvaient expliquer ces structurations. Je n'ai observé de structuration que pour les folds des noyaux des catégories intermédiaires et abondants et pour la catégorie abondante chez les Chlorophytes, et dans une moindre mesure pour les folds des noyaux des catégories intermédiaires et abondants chez les Haptophytes. Chez les Chlorophytes, la structuration est en partie expliquée par la médiane de la température annuelle, la concentration en fer et la localisation géographique des stations d'échantillonnage.

Enfin, dans l'optique de mieux comprendre ces résultats, j'ai utilisé un modèle d'apprentissage automatique environnemental pour tenter de prédire l'importance de ces mêmes paramètres sur la distribution de chaque fold des noyaux des trois catégories, pour les six phyla. Cette analyse a révélé que le fer avait une importance supérieure à celle de la température pour expliquer les distributions de chacun de ces folds.

Dans l'ensemble, ces résultats ont montré qu'il n'était pas possible de conclure de façon définitive avec nos données sur la question de la biogéographie des folds. Ils laissent cependant penser que la distribution de certains folds est effectivement structurée dans l'environnement par certains paramètres environnementaux comme la température et la concentration en fer, et que ce phénomène est particulièrement important pour le phytoplancton.

PARTIE 4. CONCLUSIONS ET PERSPECTIVES

Sommaire

1/ résultats principaux

Le cadre de cette thèse a été d'étudier les microbiomes marins à l'échelle des **structures tridimensionnelles de domaines protéiques**, et donc de réunir avec une approche originale la biologie structurale et l'écologie numérique. Ces deux disciplines, bien que complémentaires, sont rarement associées en raison des différences d'échelles de taille des objets qu'elles étudient, mais aussi des différences d'échelles espace-temps dans lesquelles ces derniers évoluent. Pourtant, les structures tridimensionnelles des domaines protéiques constituent un **niveau intermédiaire entre la séquence et la fonction**, et sont à ce titre un niveau biologique soumis à la **pression de sélection** qui participe à façonner l'évolution des espèces.

Avec ce cadre, l'objectif principal de cette thèse était d'explorer l'hypothèse selon laquelle **les variations environnementales qui modulent la composition des communautés planctoniques exercent aussi une pression à l'échelle de leurs folds**, et que **les mécanismes adaptatifs d'au moins une partie espèces planctoniques face à ces variations environnementales incluent les folds**. Analyser **la distribution des folds dans ces communautés** pourraient donc être une façon pertinente d'étudier la modulation des communautés planctoniques en réponse à la pression de sélection abiotique.

Les résultats présentés dans cette thèse semblent confirmer cette hypothèse et montrent que **l'échelle des folds représente est informative et pertinente pour caractériser l'organisation et la dynamique des communautés planctoniques**.

Plus précisément, cette étude a abouti à **deux résultat principaux** ayant chacun plusieurs implications.

résultat 1 - Universalité de la Loi Puissance et Émergence de la Loi de Pareto type II

1. universalité de la loi puissance pour modéliser la distribution des folds dans les protéomes

Sur le plan technique, l'utilisation de génomes environnementaux incomplets [5], [31] pour étudier la distribution des folds dans les protéomes a été globalement validée, en se basant sur des comparaisons avec des protéomes de référence [8]. Cela a permis d'obtenir des conclusions sur un vaste pan de la diversité planctonique, inaccessible avec des génomes complets.

Grâce à cette validation, la pertinence de la loi puissance pour modéliser la distribution des folds dans les protéomes a été vérifiée, et ce à travers une diversité taxonomique très large, incluant des Eucaryotes, des Bactéries, des Archées et des NCLDVs. Cela n'avait jusqu'ici été observé que sur des protéomes d'espèces modèles [254], [255], [258]. Le maintien de la pertinence de la loi puissance dans des lignées très éloignées de celles de ce type d'organisme suggère qu'elle constitue un principe universel de structuration des protéomes.

2. impact direct des duplications de gènes sur les paramètres du modèle de loi puissance

Les différents modèles de loi puissance présentés fournissent des arguments en faveur d'un impact direct du taux de duplication de gènes sur leurs paramètres. Plus une lignée a subi de duplications de gènes au cours de son histoire évolutive, plus la valeur absolue du coefficient directeur de la linéarisation du modèle de loi puissance est faible. Cela résulte du fait que les duplications affectent surtout les superfolds, augmentant l'écart entre leurs occurrences et celles des autres folds. Chez les Eucaryotes, cette tendance se traduit par des différences entre Bicontes et Unicontes, et, au sein des

Unicontes, entre les Choanoflagellés et les Métazoaires (qui ont le coefficient le plus faible en valeur absolue).

3. **déviation de la loi puissance vers la loi de Pareto II pour modéliser les abondances des folds dans les communautés planctoniques**

En utilisant l'information d'abondance relative des MAGs, il a été possible d'évaluer l'abondance des folds dans les stations TO. La notion d'abondance de folds dans l'environnement n'avait jusqu'ici jamais été proposée. La majorité des modèles de loi puissance de la distribution de cette grandeur se sont révélés être moins satisfaisant que pour la distribution des occurrences dans les protéomes à cause d'une déviation dans la partie correspondant aux folds les moins abondants. La loi de Pareto type II a été utilisée pour mieux prendre en compte cette déviation dans les modèles. Cette loi est habituellement utilisée dans des disciplines comme l'économie, la physique ou les sciences sociales pour décrire des systèmes complexes caractérisés par des interactions asymétriques et hiérarchisées. Contrairement à la loi puissance, qui décrit une hiérarchie statique entre les éléments d'un système, la loi de Pareto II reflète un mécanisme d'accumulation ou d'échange asymétrique entre ces entités. En économie, elle peut être utilisée pour modéliser la répartition des richesses, où une fraction limitée de la population détient une part disproportionnée des ressources. Par analogie, cela suggèrerait que certains folds dominants jouent un rôle central dans l'organisation des communautés planctoniques de manière cosmopolite, tandis que les autres, représentant une large majorité, sont « redistribués » ou « réattribués » en fonction des dynamiques écologiques et évolutives propres aux différents contextes environnementaux rencontrés par les communautés planctoniques.

4. **propriété émergente de la communauté planctonique**

La pertinence de l'utilisation de la loi de Pareto type II pour les modèles de distribution des abondances des folds a été observée à la fois chez les Eucaryotes et les Bactéries dans la majorité des stations TO. Cela suggère qu'il s'agit d'une propriété émergente des communautés microbiennes « complètes » (par le prisme des MAGs qui ne représentent que la part des génomes les plus abondants d'une communauté) et non d'une simple conséquence de l'organisation génomique individuelle. À l'échelle des lignées Eucaryotes, la pertinence de ce modèle ne s'observe cependant pas dans toutes les stations, avec une distribution qui semble avoir un sens écologique en particulier pour les lignées phytoplanctoniques analysées. Il pourrait donc constituer une indication concernant les dynamiques écologiques propres aux communautés de chaque lignée en fonction du contexte environnemental.

Dans l'ensemble, ces observations constituent un résultat totalement nouveau, renforçant l'idée que les principes d'organisation des communautés biologiques peuvent être décrits à travers des modèles issus d'autres disciplines.

résultat 2 - Structuration Biogéographique des Folds et Classification en Catégories d'Abondance

1. **classification des folds en plusieurs catégories grâce aux modèles de distribution des abondances**

La variabilité des paramètres de la loi de Pareto type II en fonction des stations TO a permis de définir trois classes d'abondance de folds dans chacune des lignées planctoniques testées. Ces classes sont globalement analogues à certaines utilisées à l'échelle des espèces [131], à savoir abondant, rare et intermédiaire (rare dans certaines stations et abondant dans d'autres).

2. **structuration biogéographique de la distribution des communautés de folds dans chacune des catégories d'abondances**

L'ensemble des folds dans une catégorie d'abondance forme, de façon analogue aux espèces, une communauté dont la distribution dans les stations TO peut être étudiée. Cette distribution a révélé une structuration biogéographique dans certaines catégories et dans certains phyla. C'est particulièrement le cas des folds intermédiaires et abondants chez les Chlorophytes et dans une moindre mesure chez les Haptophytes, dont la distribution est structurée par bassins et par biomes, et est en partie influencée par la moyenne annuelle de la température de surface et la concentration en fer.

3. **importance de la concentration en fer dans la distribution biogéographique de certains folds**

Dans le but de comprendre plus en détail les résultats de structuration présentés dans le paragraphe précédent, la distribution de biogéographique de certains folds a été modélisée à l'aide d'un outil d'apprentissage automatique développé par un collaborateur au sein de l'équipe. Cet outil permet de prédire la distribution d'objets biologiques avec une abondance dans plusieurs stations océaniques à l'échelle de l'Océan global, en tentant de détecter des associations entre contexte environnemental et valeur d'abondance de l'objet dans ces stations. Il a révélé, de façon surprenante, un effet prépondérant (et bien supérieur à la moyenne annuelle de la température de surface) de la concentration en fer dans la distribution de tous les folds testés. L'effet de la concentration en métaux dans l'environnement sur le protéome avait déjà été observé dans le cas du zinc chez une espèce de Diatomée [21]. Indépendamment de ce résultat, le fait que la qualité globale des modèles de distribution biogéographique des folds soit satisfaisante montre que cette approche ouvre la voie à l'utilisation de modèles basés sur l'intelligence artificielle pour cartographier et anticiper les dynamiques des microbiomes marins à plusieurs niveaux biologiques.

Dans l'ensemble, les résultats de cette thèse suggèrent que les dynamiques écologiques et évolutives qui façonnent la répartition des espèces planctoniques influencent également la distribution des folds. Cela implique que **les variations environnementales qui modulent la composition des communautés planctoniques exercent aussi probablement une pression à l'échelle de leurs folds**, et donc que l'hypothèse fondatrice de thèse est certainement en partie valide.

2/ limites de l'étude, perspectives et directions futures

limites liées à l'utilisation des MAGs

L'utilisation des MAGs a permis d'explorer la diversité des microbiomes marins échantillonnés lors des expéditions *Tara* Oceans et Polar Circle, bien qu'ils soient en majorité partiellement complets et fragmentés. Malgré la validation de leur utilisation pour une étude à l'échelle des folds, principalement par comparaison des répertoires entre génomes environnementaux et de référence, cette incomplétion peut conduire à des biais, en particulier pour les folds de la classe d'abondance rare, dont la diversité et les valeurs d'occurrence sont très probablement trop basses par rapport à la réalité. Des corrections ont été mises en place pour minimiser cet effet (filtrage des MAGs dont la BUSCO est inférieure à 50% dans certaines analyses, prise en compte de la longueur des génomes pour le calcul de la métrique d'abondance, transformation CLR), mais il ne peut néanmoins être négligé.

La complétion des MAGs est surtout insuffisante pour les études incluant l'importance des duplications de gènes qui jouent un rôle dans l'évolution des folds, résultant en des modèles de distribution des occurrences significativement différents entre génomes environnementaux et de référence.

Cet effet est cependant moins problématique à l'échelle des communautés dans l'environnement, qui constituaient le cœur des analyses présentées ici. Bien que les MAGs utilisés dans cette étude ne sont que les espèces les plus abondantes au sein des différentes communautés planctoniques, cette part correspond aux espèces dont la fitness est la plus élevée dans un contexte environnemental donné. Puisque les folds contribuent en partie à cette valeur de fitness, les foldomes des espèces les plus abondantes sont donc probablement les plus pertinents à étudier pour essayer de détecter des associations entre folds et contexte environnemental.

limites des annotations structurales (CATH) et perspectives offertes par AlphaFold

L'annotation structurale des protéomes repose dans cette thèse sur la base de données CATH [2], [3], qui n'est pas exhaustive et fonctionne par homologie de séquence. De fait, un pourcentage élevé de protéines des MAGs ne dispose pas d'annotation structurale, leurs domaines n'ayant pas d'homologues connus.

L'émergence de modèles prédictifs comme AlphaFold, qui permettent d'inférer la structure tridimensionnelle des protéines à partir des séquences, représente une opportunité majeure pour améliorer ces annotations, notamment par l'intermédiaire de la TED [184]. Étant donné que cette base de donnée recense un total de près de 365 millions domaines avec de nouveaux folds, contre 151 million dans CATH, il est probable que le taux d'annotation structurale avec des MAGs avec cette nouvelle base de donnée soit meilleur. Il est néanmoins peu probable que cette amélioration ne change fondamentalement les conclusions présentées ici, car elles ne feraient probablement qu'apporter de la diversité dans les folds rares, mais ne changeraient pas de façon considérable les écarts ces derniers et les folds abondants qui sont adoptés par des domaines ayant majoritairement des homologues connus.

limites liées aux propriétés fonctionnelles des folds

Un même fold peut être impliqué dans différentes fonctions, et la relation entre structure et fonction n'est pas toujours triviale. Dans les analyses environnementales, il n'est pas possible de dire avec les résultats présentés ici si la distribution biogéographique des folds est liée à leurs propriétés thermodynamique et à leur cinétique de repliement, ou à leur rôle fonctionnel.

Par ailleurs, les résultats obtenus dans cette thèse ouvrent plusieurs **axes de recherche** :

perspective 1 - extensions des modèles de prédictions de la distribution biogéographique des folds

La première perspective serait d'étendre les modèles de distribution biogéographique à l'ensemble de folds du jeu de donnée, puisque seul une cinquantaine d'entre eux ont été utilisés pour l'analyse présentée ici. Cela permettrait déjà de vérifier si l'importance du fer est vraiment partagée par tous les folds ou non, et donc de mieux interpréter ce résultat. Plus généralement, les modèles prédictifs intégrant les facteurs environnementaux pourraient atteindre de meilleures performances et permettre de mieux comprendre comment les conditions océaniques, telles que la température ou la disponibilité en nutriments, influencent la dynamique de la distribution des folds au sein des communautés planctoniques essentiellement en augmentant la taille du jeu d'entraînement à l'aide d'autres jeux de données.

perspective 2 - intégration des annotations fonctionnelles

Au vue de la dernière limitation évoquée, l'un des axes d'amélioration essentiels consisterait donc à combiner l'annotation fonctionnelle et structurale des protéines afin d'affiner la compréhension des mécanismes évolutifs. Il s'agit notamment de déterminer à quel niveau s'exerce principalement la pression de sélection : agit-elle sur les séquences nucléotidiques, sur des résidus fonctionnels spécifiques, ou bien directement sur la structure tridimensionnelle des folds ? L'une des limites de ce type d'approche serait néanmoins que l'utilisation exclusivement de protéines ayant à la fois une annotation structurale et fonctionnelle pourrait résulter en une réduction encore plus importante du jeu de donnée, qui est déjà relativement incomplet.

Un autre objectif serait d'évaluer si certaines fonctions biologiques particulières sont corrélées à la structuration biogéographique des folds, en identifiant d'éventuels motifs fonctionnels spécifiques à certaines zones océaniques.

perspective 3 - exploitation des données issues des métatranscriptomes

L'étude des métatranscriptomes, qui renseignent sur l'expression active des gènes dans un écosystème donné, offrirait une opportunité d'analyser l'expression des folds en réponse aux variations des conditions environnementales. Cette approche permettrait de vérifier si leur abondance fluctue en fonction des conditions locales et saisonnières, et si la loi de Pareto type II est toujours pertinente modéliser leur répartition à ce niveau d'organisation.

Enfin, il serait intéressant de tester si certains folds sont exprimés de manière préférentielle dans des contextes écologiques particuliers, ce qui pourrait révéler des stratégies adaptatives spécifiques et fournir des indices sur le lien entre pression environnementale, sélection naturelle et évolution structurale des protéines.

perspective 4 - étendre l'analyse à d'autres bases de données

Des jeux de données issues d'autres expéditions (Tara Pacific, Mission Microbiome, Tara Europa, OMRGC) sont déjà disponibles. Leur intégration permettrait de vérifier si les tendances observées avec les données TO et Polar Circle restent vraies avec plus grand nombre d'échantillons. L'analyse devrait également être étendue à d'autres lignées planctoniques, notamment Procaryotes, et à d'autres profondeurs dans la colonne d'eau, en particulier dans les milieux mésopélagiques dont les communautés planctoniques et les conditions physico-chimiques sont très différentes du milieu épipélagique.

perspective 5 - analyse des combinaisons de folds dans les protéines multidomaines

Les folds ont ici été étudiés indépendamment des combinaisons qu'ils forment entre eux. Ces combinaisons ont cependant un rôle fonctionnel et structural, et leur étude dans le plancton pourrait permettre d'identifier des combinaisons inconnues. Estimer l'abondance de certaines combinaisons spécifiques dans l'environnement pourrait également permettre de mieux interpréter certains résultats à l'échelle des folds présentés dans cette thèse.

3/ conclusion générale

Cette thèse a mis en évidence que l'échelle des **structures de domaines protéiques** permet d'analyser la **distribution biogéographique des communautés planctoniques**. L'intégration des modèles mathématiques a permis d'en détecter certaines **propriétés émergentes**, et d'identifier la **loi de Pareto II** comme un bon modèle pour les étudier.

Toutefois, certaines limites méthodologiques, notamment liées à la **complétion des MAGs** et à l'**annotation structurales de leurs protéines**, doivent être surmontées pour **affiner la compréhension des liens entre biologie structurale, écologie et évolution**.

L'ensemble de ces résultats ouvre de nouvelles perspectives pour mieux comprendre comment les dynamiques écologiques et évolutives influencent l'organisation des protéines dans les génomes microbiens.

Dans le futur, ces travaux pourront être enrichis par l'intégration d'annotations structurales améliorées et d'annotations fonctionnelles ainsi de nouvelles bases de données, et l'amélioration de modèles prédictifs par des méthodes d'apprentissage automatique. Ainsi, le croisement entre la biologie structurale et de l'écologie permettra de mieux comprendre **les stratégies adaptatives des microbiomes marins** et d'approfondir **l'étude des mécanismes évolutifs influençant l'organisation des communautés planctoniques.**

Bibliographie

[1] A. M. Eren *et al.*, “Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data,” *Methods in Ecology and Evolution*, vol. 4, no. 12, pp. 1111–1119, Dec. 2013, doi: 10.1111/2041-210X.12114.

[2] I. Sillitoe, N. Bordin, N. Dawson, and C. A. Orengo, “CATH: Increased Structural Coverage of Functional Space,” *Nucleic Acids Research*, 2021.

[3] C. A. Orengo, A. Michie, S. Jones, D. T. Jones, M. Swindells, and J. M. Thornton, “CATH - a hierarchic classification of protein domain structures,” *Structure*, 1997.

[4] A. Kolmogorov, “Sulla determinazione empirica di una legge di distribuzionc,” *1st. Ital. Attuari.*, vol. 4, pp. 1–11, 1933a.

[5] T. O. Delmont *et al.*, “Functional Repertoire Convergence of Distantly Related Eukaryotic Plankton Lineages Revealed by Genome-Resolved Metagenomics,” *Cell Genomics*, 2022.

[6] R. R. Sokal, “The Principles and Practice of Numerical Taxonomy,” *Taxon*, vol. 12, no. 5, pp. 190–199, 1963, doi: 10.2307/1217562.

[7] E. L. L. Sonnhammer, S. R. Eddy, and R. Durbin, “Pfam: A comprehensive database of protein domain families based on seed alignments,” *Proteins: Structure, Function, and Bioinformatics*, vol. 28, no. 3, pp. 405–420, Jul. 1997, doi: 10.1002/(SICI)1097-0134(199707)28:3<405::AID-PROT10>3.0.CO;2-L.

[8] D. J. Richter *et al.*, “EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes,” *Peer Community Journal*, vol. 2, 2022, doi: 10.24072/pcjournal.173.

[9] S. Pesant *et al.*, “Open science resources for the discovery and analysis of Tara Oceans data,” *Scientific Data*, vol. 2, no. 1, p. 150023, May 2015, doi: 10.1038/sdata.2015.23.

[10] J. R. Reagan, T. P. Boyer, H. E. Garcia, and D. Dukhovskoy, “World Ocean Atlas 2023.,” *NOAA National Centers for Environmental Information*, 2024, [Online]. Available: NCEI Accession 0270533.

[11] C. Sardet, *Plancton: Aux origines du vivant*, Ulmer. 2022.

[12] S. Pallacks *et al.*, “Anthropogenic acidification of surface waters drives decreased biogenic calcification in the Mediterranean Sea,” *Communications Earth & Environment*, vol. 4, no. 1, p. 301, Aug. 2023, doi: 10.1038/s43247-023-00947-7.

[13] G. Li, L. Cheng, J. Zhu, K. E. Trenberth, M. E. Mann, and J. P. Abraham, “Increasing ocean stratification over the past half-century,” *Nature Climate Change*, vol. 10, no. 12, pp. 1116–1123, Dec. 2020, doi: 10.1038/s41558-020-00918-2.

[14] M. T. Kavanaugh, “Seascape Ecology: a Review,” 2018.

[15] “Oceans,” in *Baas Becking’s Geobiology*, 2015, pp. 92–102. doi: 10.1002/9781118295472.ch9.

[16] R. de Wit and T. Bouvier, “‘Everything is everywhere, but, the environment selects’; what did Baas Becking and Beijerinck really say?,” *Environmental Microbiology*, 2006.

[17] M. A. O’Malley, “The nineteenth century roots of ‘everything is everywhere,’” *Nature Reviews Microbiology*, 2007.

[18] M. W. Beyerinck 1851-1931 (viaf)19794581, *De infusies en de ontdekking der bakteriën*. Amsterdam : Muller, 1913. [Online]. Available: http://lib.ugent.be/catalog/rug01:000134114

[19] A. Longhurst, “Ecological Geography of the Sea,” *Academic Press*, 1997.

[20] M. Romei, G. Sapriel, P. Imbert, and M. Carpentier, “Protein folds as synapomorphies of the tree of life,” *Evolution*, 2022.

[21] T. Mock *et al.*, “Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus,” *Nature*, vol. 541, no. 7638, pp. 536–540, Jan. 2017, doi: 10.1038/nature20803.

[22] M. Domnauer *et al.*, “Proteome plasticity in response to persistent environmental change,” *Molecular Cell*, vol. 81, no. 16, pp. 3294-3309.e12, Aug. 2021, doi: 10.1016/j.molcel.2021.06.028.

[23] F. Lombard *et al.*, “Globally Consistent Quantitative Observations of Planktonic Ecosystems,” *Frontiers in Marine Science*, vol. 6, 2019, doi: 10.3389/fmars.2019.00196.

[24] Christian K. Sieracki, Michael E. Sieracki, and Charles S. Yentsch, “An imaging-in-flow system for automated analysis of marine microplankton,” *Mar Ecol Prog Ser*, vol. 168, pp. 285–296, 1998.
[25] G. Gorsky *et al.*, “Digital zooplankton image analysis using the ZooScan integrated system,” *Journal of Plankton Research*, vol. 32, no. 3, pp. 285–303, Mar. 2010, doi: 10.1093/plankt/fbp124.
[26] D. A. Siegel, K. O. Buesseler, S. C. Doney, S. F. Sailley, M. J. Behrenfeld, and P. W. Boyd, “Global assessment of ocean carbon export by combining satellite observations and food-web models,” *Global Biogeochemical Cycles*, vol. 28, no. 3, pp. 181–196, Mar. 2014, doi: 10.1002/2013GB004743.
[27] C. A. Hostetler, M. J. Behrenfeld, Y. Hu, J. W. Hair, and J. A. Schulien, “Spaceborne Lidar in the Study of Marine Systems,” *Annual Review of Marine Science*, vol. 10, no. Volume 10, 2018. Annual Reviews, pp. 121–147, 2018. doi: https://doi.org/10.1146/annurev-marine-121916-063335.
[28] M. Luo, Y. Ji, D. Warton, and D. W. Yu, “Extracting abundance information from DNA-based data,” *Molecular Ecology Resources*, vol. 23, no. 1, pp. 174–189, Jan. 2023, doi: 10.1111/1755-0998.13703.
[29] B. M. Satinsky, S. M. Gifford, B. C. Crump, and M. A. Moran, “Chapter Twelve - Use of Internal Standards for Quantitative Metatranscriptome and Metagenome Analysis,” in *Methods in Enzymology*, vol. 531, E. F. DeLong, Ed., Academic Press, 2013, pp. 237–250. doi: 10.1016/B978-0-12-407863-5.00012-5.
[30] S. Sunagawa, L. P. Coelho, and S. Chaffron, “Structure and Function of the Global Ocean Microbiome,” *Science*, 2015.
[31] M. Gaïa *et al.*, “Mirusviruses link herpesviruses to giant viruses,” *Nature*, vol. 616, no. 7958, pp. 783–789, Apr. 2023, doi: 10.1038/s41586-023-05962-4.
[32] C. A. Suttle, “Marine Viruses - Major Players in the Global Ecosystem,” *Nature Reviews*, 2009.
[33] F. E. Angly *et al.*, “The Marine Viromes of Four Oceanic Regions,” *PLOS Biology*, vol. 4, no. 11, p. e368, Nov. 2006, doi: 10.1371/journal.pbio.0040368.
[34] H. Endo *et al.*, “Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions,” *Nature Ecology & Evolution*, vol. 4, no. 12, pp. 1639–1649, Dec. 2020, doi: 10.1038/s41559-020-01288-w.
[35] H. Kaneko, R. Blanc-Mathieu, H. Endo, and H. Ogata, “Eukaryotic Virus Composition can Predict the Efficiency of Carbon Export in the Global Ocean,” *IScience*, 2021.
[36] “Global drivers of eukaryotic plankton biogeography in the sunlit ocean,” 2021.
[37] T. Cordier *et al.*, “Patterns of Eukaryotic Diversity from the Surface to the Deep-Ocean Sediment,” *Science Advances*, 2022.
[38] C. De Vargas, S. Audie, and N. Henry, “Eukaryotic plankton diversity in the sunlit ocean,” *Science*, vol. 348, no. 6237, Art. no. 6237, 2015.
[39] C. R. Woese, O. Kandler, and M. L. Wheelis, “Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya.,” *Proceedings of the National Academy of Sciences*, vol. 87, no. 12, pp. 4576–4579, Jun. 1990, doi: 10.1073/pnas.87.12.4576.
[40] P. Forterre, “The universal tree of life: an update,” *Frontiers in Microbiology*, vol. 6, 2015, doi: 10.3389/fmicb.2015.00717.
[41] F. Burki, A. J. Roger, M. W. Brown, and A. G. B. Simpson, “The New Tree of Eukaryotes,” *Trends in Ecology & Evolution*, vol. 35, no. 1, pp. 43–55, Jan. 2020, doi: 10.1016/j.tree.2019.08.008.
[42] C. Hulo *et al.*, “ViralZone: a knowledge resource to understand virus diversity,” *Nucleic Acids Research*, vol. 39, no. suppl_1, pp. D576–D582, Jan. 2011, doi: 10.1093/nar/gkq901.
[43] A. Nasir and G. Caetano-Anollés, “A phylogenomic data-driven exploration of viral origins and evolution,” *Science Advances*, vol. 1, no. 8, p. e1500527, doi: 10.1126/sciadv.1500527.

[44] D. Faktorová, E. Dobáková, P. Peña-Diaz, and J. Lukeš, “From simple to supercomplex: mitochondrial genomes of euglenozoan protists [version 2; peer review: 2 approved],” *F1000Research*, vol. 5, no. 392, 2016, doi: 10.12688/f1000research.8040.2.
[45] P. Rotkewicz, Nature Center, Ithaca, NY, Mar. 2004. [Online]. Available: https://www.pirx.com/droplet/about.html
[46] B. S. C. Leadbeater, Q. Yu, J. Kent, and D. J. Stekel, “Three-dimensional images of choanoflagellate loricae,” *Proceedings of the Royal Society B: Biological Sciences*, vol. 276, no. 1654, pp. 3–11, Aug. 2008, doi: 10.1098/rspb.2008.0844.
[47] Roscoff Culture Collection, [Online]. Available: https://roscoff-culture-collection.org/rcc-strain-details/1718
[48] C. McKay, “Foraminifera.” Department of Geology, Lund University, 2014. [Online]. Available: https://www.researchmagazine.lu.se/2014/07/22/foraminifera/
[49] Y. Tsukii, “Protist Information Server.” [Online]. Available: http://protist.i.hosei.ac.jp/
[50] E. Mitani *et al.*, “Fatty acid composition profiles of 235 strains of three microalgal divisions within the NIES Microbial Culture Collection,” *Microbial Resources and Systematics*, vol. 33, pp. 19–29, Jul. 2017.
[51] D. R. Maddison and K.-S. Schulz, “Tree of Life Web Project. Image number: 3243. Author: Celeste Leander.” 2007. [Online]. Available: http://tolweb.org
[52] Luka \vSupraha, “Phenotypic evolution and adaptive strategies in marine phytoplankton (Coccolithophores),” 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:14538936
[53] A. K. Tice *et al.*, “Expansion of the molecular and morphological diversity of Acanthamoebidae (Centramoebida, Amoebozoa) and identification of a novel life cycle type within the group,” *Biology Direct*, vol. 11, no. 1, p. 69, Dec. 2016, doi: 10.1186/s13062-016-0171-0.
[54] S. Sunagawa, S. G. Acinas, P. Bork, and C. De Vargas, “Tara Oceans: towards global ocean ecosystems biology,” *Nature Reviews*, 2020.
[55] J. J. Pierella Karlusich, F. M. Ibarbalz, and C. Bowler, “Phytoplankton in the Tara Ocean,” *Annual Review of Marine Science*, vol. 12, no. Volume 12, 2020. Annual Reviews, pp. 233–265, 2020. doi: https://doi.org/10.1146/annurev-marine-010419-010706.
[56] Y. M. Bar-On and R. Milo, “The Biomass Composition of the Oceans: A Blueprint of Our Blue Planet,” *Cell*, vol. 179, no. 7, pp. 1451–1454, Dec. 2019, doi: 10.1016/j.cell.2019.11.018.
[57] C. H. Wigington *et al.*, “Re-examination of the relationship between marine virus and microbial cell abundances,” *Nature Microbiology*, vol. 1, no. 3, p. 15024, Jan. 2016, doi: 10.1038/nmicrobiol.2015.24.
[58] X. Irigoien, J. Huisman, and R. P. Harris, “Global biodiversity patterns of marine phytoplankton and zooplankton,” *Nature*, vol. 429, no. 6994, pp. 863–867, Jun. 2004, doi: 10.1038/nature02593.
[59] M. J. Behrenfeld, E. Boss, D. A. Siegel, and D. M. Shea, “Carbon-based ocean productivity and phytoplankton physiology from space,” *Global Biogeochemical Cycles*, vol. 19, no. 1, Mar. 2005, doi: 10.1029/2004GB002299.
[60] P. G. Falkowski, R. T. Barber, and V. Smetacek, “Biogeochemical Controls and Feedbacks on Ocean Primary Production,” *Chemistry and Biology of the Oceans*, 1998.
[61] S. M. Vallina, M. Follows, S. Dutkiewicz, J. M. Montoya, and M. Loreau, “Global Relationship between Phytoplankton Diversity and Productivity in the Ocean,” *Nature Communications*, 2014.
[62] M. Lévy, O. Jahn, S. Dutkiewicz, and M. Follows, “Phytoplankton Diversity and Community Structure Affected by Oceanic Dispersal and Mesoscale Turbulence,” *Fluids & Environments*, 2014.
[63] L. Dlugosch, A. Poehlein, B. Wemheuer, and M. Simon, “Significance of Gene Variants for the Functional Biogeography of the Near-Surface Atlantic Ocean Microbiome,” *Nature Communications*, 2022.

[64] F. M. Ibarbalz, N. Henry, M. C. Brandão, C. Bowler, and L. Zinger, “Global Trends in Marine Plankton Diversity across Kingdoms of Life,” *Cell*, 2019.
[65] E. Villarino *et al.*, “Large-scale ocean connectivity and planktonic body size,” *Nature Communications*, vol. 9, no. 1, p. 142, Jan. 2018, doi: 10.1038/s41467-017-02535-8.
[66] R. Massana and R. Logares, “Eukaryotic versus prokaryotic marine picoplankton ecology,” *Environmental Microbiology*, vol. 15, no. 5, pp. 1254–1261, May 2013, doi: 10.1111/1462-2920.12043.
[67] P. Flombaum *et al.*, “Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus.,” *Proc Natl Acad Sci U S A*, vol. 110, no. 24, pp. 9824–9829, Jun. 2013, doi: 10.1073/pnas.1307701110.
[68] J. R. Casey, R. M. Bolteau, M. K. Engqvist, and M. Follows, “Basin-scale biogeography of marine phytoplankton reflects cellular-scale optimization of metabolism and physiology,” *Science Advances*, 2022.
[69] R. Massana, “Eukaryotic Picoplankton in Surface Oceans,” *Annual Review of Microbiology*, vol. 65, no. Volume 65, 2011. Annual Reviews, pp. 91–110, 2011. doi: https://doi.org/10.1146/annurev-micro-090110-102903.
[70] B. A. Ward, S. Dutkiewicz, O. Jahn, and M. J. Follows, “A size-structured food-web model for the global ocean,” *Limnology and Oceanography*, vol. 57, no. 6, pp. 1877–1891, 2012, doi: https://doi.org/10.4319/lo.2012.57.6.1877.
[71] T. Hirata *et al.*, “Synoptic relationships between surface Chlorophyll- a and diagnostic pigments specific to phytoplankton functional types,” *Biogeosciences*, vol. 8, pp. 311–327, 2011.
[72] J. J. Walsh *et al.*, “Wind events and food chain dynamics within the New York Bight ,” *Limnology and Oceanography*, vol. 23, no. 4, pp. 659–683, Jul. 1978, doi: 10.4319/lo.1978.23.4.0659.
[73] N. S. R. Agawin, C. M. Duarte, and S. Agustí, “Nutrient and temperature control of the contribution of picoplankton to phytoplankton biomass and production,” *Limnology and Oceanography*, vol. 45, no. 3, pp. 591–600, May 2000, doi: 10.4319/lo.2000.45.3.0591.
[74] R. El Hourany, J. Pierella Karlusich, L. Zinger, H. Loisel, M. Levy, and C. Bowler, “Linking satellites to genes with machine learning to estimate phytoplankton community structure from space,” *Ocean Science*, vol. 20, no. 1, pp. 217–239, 2024, doi: 10.5194/os-20-217-2024.
[75] B. A. Ward, “Temperature-Correlated Changes in Phytoplankton Community Structure Are Restricted to Polar Waters,” *PLOS ONE*, vol. 10, no. 8, pp. 1–15, Aug. 2015, doi: 10.1371/journal.pone.0135581.
[76] S. Dutkiewicz, P. Cermeno, O. Jahn, and B. A. Ward, “Dimensions of Marine Phytoplankton Diversity,” *Biogeosciencs*, 2019.
[77] D. G. Capone, J. P. Zehr, H. W. Paerl, B. Bergman, and E. J. Carpenter, “Trichodesmium, a Globally Significant Marine Cyanobacterium,” *Science*, vol. 276, no. 5316, pp. 1221–1229, May 1997, doi: 10.1126/science.276.5316.1221.
[78] J. P. Zehr and P. J. Turner, “Nitrogen fixation: Nitrogenase genes and gene expression,” in *Methods in Microbiology*, vol. 30, Academic Press, 2001, pp. 271–286. doi: 10.1016/S0580-9517(01)30049-1.
[79] D. Bombar *et al.*, “Measurements of nitrogen fixation in the oligotrophic North Pacific Subtropical Gyre using a free-drifting submersible incubation device,” *Journal of Plankton Research*, vol. 37, pp. 727–739, 2015.
[80] P. H. Moisander, T. Serros, R. W. Paerl, R. A. Beinart, and J. P. Zehr, “Gammaproteobacterial diazotrophs and nifH gene expression in surface waters of the South Pacific Ocean,” *The ISME Journal*, vol. 8, no. 10, pp. 1962–1973, Oct. 2014, doi: 10.1038/ismej.2014.49.
[81] C. R. Loescher *et al.*, “Facets of diazotrophy in the oxygen minimum zone waters off Peru,” *The ISME Journal*, vol. 8, no. 11, pp. 2180–2192, Nov. 2014, doi: 10.1038/ismej.2014.71.

[82] F. Vincent *et al.*, “Viral infection switches the balance between bacterial and eukaryotic recyclers of organic matter during coccolithophore blooms,” *Nature Communications*, vol. 14, no. 1, p. 510, Jan. 2023, doi: 10.1038/s41467-023-36049-3.
[83] D. M. Needham, R. Sachdeva, and J. A. Fuhrman, “Ecological Dynamics and Co-Occurrence among Marine Phytoplankton, Bacteria and Myoviruses Shows Microdiversity Matters.,” *The ISME Journal*, 2017.
[84] C. Bunse and J. Pinhassi, “Marine Bacterioplankton Seasonal Succession Dynamics,” *Cell Press*, 2017.
[85] G. Bourdin, L. Karp-Boss, F. Lombard, G. Gorsky, and E. Boss, “Dynamic island mass effect from space. Part I: detecting the extent,” *EGUsphere*, vol. 2024, pp. 1–38, 2024, doi: 10.5194/egusphere-2024-2670.
[86] M. Messié *et al.*, “The Delayed Island Mass Effect: How Islands can Remotely Trigger Blooms in the Oligotrophic Ocean,” *Geophysical Research Letters*, vol. 47, no. 2, p. e2019GL085282, Jan. 2020, doi: 10.1029/2019GL085282.
[87] F. Milke, J. Meyerjürgens, and M. Simon, “Ecological mechanisms and current systems shape the modular structure of the global oceans’ prokaryotic seascape,” *Nature Communications*, 2023.
[88] C. E. Shannon, “A mathematical theory of communication,” *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948, doi: 10.1002/j.1538-7305.1948.tb01338.x.
[89] J. Ladau *et al.*, “Global marine bacterial diversity peaks at high latitudes in winter,” *The ISME Journal*, vol. 7, no. 9, pp. 1669–1677, Sep. 2013, doi: 10.1038/ismej.2013.37.
[90] G. Salazar, L. Paoli, A. Alberti, M. B. Sullivan, P. Wincker, and S. Sunagawa, “Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome,” *Cell*, 2019.
[91] F. M. Ibarbalz, N. Henry, F. Mahé, and L. Karp-Boss, “Pan-Arctic plankton community structure and its global connectivity,” *ELEMENTA: Science of the Anthropocene*, 2023.
[92] S. Chaffron, E. Delage, M. Budinich, and D. Eveillard, “Environmental Vulnerability of the Global Ocean Epipelagic Plankton Community Interactome,” *Science Advances*, 2021.
[93] J. A. Fuhrman, J. A. Steele, I. Hewson, and J. H. Brown, “A Latitudinal Diversity Gradient in Planktonic Marine Bacteria,” *PNAS*, 2008.
[94] W. J. Sul, T. A. Oliver, H. W. Ducklow, L. A. Amaral-Zettler, and M. L. Sogin, “Marine bacteria exhibit a bipolar distribution,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 6, pp. 2342–2347, Feb. 2013, doi: 10.1073/pnas.1212424110.
[95] D. Righetti, M. Vogt, N. Gruber, A. Psomas, and N. E. Zimmermann, “Global Pattern of Phytoplankton Diversity Driven by Temperature and Environmental Variability,” *Science Advances*, 2019.
[96] E. J. Raes, L. Bodrossy, J. van de Kamp, and A. M. Waite, “Oceanographic Boundaries Constrain Microbial Diversity Gradients in the South Pacific Ocean,” *PNAS*, 2018.
[97] A. C. Gregory, A. A. Zayed, N. Conceição-Neto, and M. B. Sullivan, “Marine DNA Viral Macro- and Microdiversity from Pole to Pole,” *Cell*, 2019.
[98] G. Sommeria-Klein *et al.*, “Global drivers of eukaryotic plankton biogeography in the sunlit ocean,” *Science*, vol. 374, no. 6567, pp. 594–599, Oct. 2021, doi: 10.1126/science.abb3717.
[99] S. Malviya, E. Scalco, S. Audic, and C. Bowler, “Insights into global diatom distribution and diversity inthe world’s ocean,” *PNAS*, 2016.
[100] E. Villar, G. K. Farrant, M. Follows, and D. Iudicone, “Environmental Characteristics of Agulhas Rings Affect Interocean Plankton Transport,” *Nature*, 2015.
[101] D. R. Mende, J. A. Bryant, F. O. Aylward, and E. F. Delong, “Environmental Drivers of a Microbial Genomic Transition Zone in the Ocean’s Interior,” *Nature Microbiology*, 2017.
[102] G. Salazar *et al.*, “Global diversity and biogeography of deep-sea pelagic prokaryotes,” *The ISME Journal*, vol. 10, no. 3, pp. 596–608, Mar. 2016, doi: 10.1038/ismej.2015.137.
[103] G. Busseni, L. Caputi, R. Piredda, and D. Iudicone, “Large scale patterns of marine diatom richness: Drivers and trends in a changing ocean,” *Global Ecology and Biogeography*, 2020.

[104] C. Pedros-Alio, “The Rare Bacterial Biosphere,” *Annual Reviews Marine Science*, 2012.
[105] T. Rodríguez-Ramos, E. Marañón, and P. Cermeño, “Marine nano- and microphytoplankton diversity: redrawing global patterns from sampling-standardized data,” *Global Ecology and Biogeography*, vol. 24, no. 5, pp. 527–538, 2015, doi: https://doi.org/10.1111/geb.12274.
[106] J. Grilli, “Macroecological laws describe variation and diversity in microbial communities,” *Nature Communications*, 2020.
[107] M. Mestre and J. Höfer, “The Microbial Conveyor Belt: Connecting the Globe through Dispersion and Dormancy,” *Trends in Microbiology*, 2020.
[108] L. Alonso-Sáez, L. Díaz-Pérez, and X. A. G. Morán, “The hidden seasonality of the rare biosphere in coastal marine bacterioplankton,” *Environmental Microbiology*, vol. 17, no. 10, pp. 3766–3780, Oct. 2015, doi: 10.1111/1462-2920.12801.
[109] T. POMMIER *et al.*, “Global patterns of diversity and community structure in marine bacterioplankton,” *Molecular Ecology*, vol. 16, no. 4, pp. 867–880, Feb. 2007, doi: 10.1111/j.1365-294X.2006.03189.x.
[110] G. K. Farrant *et al.*, “Delineating ecologically significant taxonomic units from global patterns of marine picocyanobacteria,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 24, pp. E3365–E3374, Jun. 2016, doi: 10.1073/pnas.1524865113.
[111] P. Frémont *et al.*, “Restructuring of Genomic Provinces on Surface Ocean Plankton under Climate Change,” 2022.
[112] A. Bertrand, D. Grados, F. Colas, and R. Fablet, “Broad Impacts of Fine-scale Dynamics on Seascape Structure from Zooplankton to Seabirds,” *Nature Communications*, 2014.
[113] H. Kaneko, H. Endo, N. Henry, and H. Ogata, “Predicting global distributions of eukaryotic plankton communities from satellite data,” *ISME*, 2023.
[114] V. J. Coles, M. R. Stukel, M. T. Brooks, and R. R. Hood, “Ocean Biogeochemistry Modeled with Emergent Trait-Based Genomics,” *Science*, 2017.
[115] E. F. Delong, C. M. Preston, T. Mincer, and D. M. Karl, “Community Genomics among Stratified Microbial Assmeblages in the Ocean’s Interior,” *Science*, 2007.
[116] J.-F. Ghiglione, P. E. Galand, T. Pommier, and A. E. Murray, “Pole-to-pole Biogeography of Surface and Deep Marine Bacterial Communities,” *PNAS*, 2012.
[117] D. Wilkins, E. van Sebille, S. R. Rintoul, F. M. Lauro, and R. Cavicchioli, “Advection shapes Southern Ocean microbial assemblages independent of distance and environment effects,” *Nature Communications*, vol. 4, no. 1, p. 2457, Sep. 2013, doi: 10.1038/ncomms3457.
[118] E. Laiolo, I. Alam, M. Uludag, and C. M. Duarte, “Metagenomic probing toward an atlas of the taxonomic and metabolic foundations of the global ocean genome,” *Frontiers in Science*, 2024.
[119] J. Raes, I. Letunic, T. Yamada, and P. Bork, “Toward Molecular Trait-based Ecology through Integration of Biogeochemical, Geographical and Metagenomic Data,” *Molecular Systems Biology*, 2011.
[120] B. K. Swan, B. Tupper, A. Sczyrba, and R. Stepanauskas, “Prevalent Genome Streamlining and Latitudinal Divergence of Planktonic Bacteria in the Surface Ocean,” *PNAS*, 2013.
[121] D. Richter, R. Watteaux, and T. Vannier, “Genomic evidence for global ocean plankton biogeography shaped by large-scale current systems,” Dec. 2020, doi: 10.7554/eLife.78129.
[122] Y. Seeleuthner, S. Mondy, V. Lombard, and P. Wincker, “Single-cell genomics of multiple uncultured Stramenopiles reveals underestimated functional diversity across oceans,” *Nature Communications*, 2018.
[123] J. Rigonato, M. Budinich, A. A. Murillo, and O. Jaillon, “Ocean-wide comparisons of mesopelagic planktonic community structures,” *ISME*, 2023.
[124] C. R. Giner *et al.*, “Marked changes in diversity and relative activity of picoeukaryotes with depth in the world ocean,” *The ISME Journal*, vol. 14, no. 2, pp. 437–449, Feb. 2020, doi: 10.1038/s41396-019-0506-9.
[125] H. Endo, R. Blanc-Mathieu, Y. Li, and H. Ogata, “Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions,” *Nature Ecology and Evolution*, 2020.

[126] C. A. Hanson, J. A. Fuhrman, M. C. Horner-Devine, and J. B. H. Martiny, “Beyond biogeographic patterns: processes shaping the microbial landscape,” *Nature Reviews Microbiology*, 2012.
[127] S. J. Giovannoni and K. L. Vergin, “Seasonality in Ocean Microbial Communities,” *Science*, 2012.
[128] N. Joli, A. Monier, R. Logares, and C. Lovejoy, “Seasonal patterns in Arctic prasinophytes and inferred ecology of Bathycoccus unveiled in an Arctic winter metagenome,” *The ISME Journal*, vol. 11, no. 6, pp. 1372–1385, Jun. 2017, doi: 10.1038/ismej.2017.7.
[129] M. Wutkowska, A. Vader, R. Logares, and T. M. Gabrielsen, “Linking extreme seasonality and gene expression in Arctic marine protists,” *Scientific Reports*, 2023.
[130] E. G. Hutchinson, “Concluding remarks.,” *Cold Spring Harbor Symposia on Quantitative Biology.*, vol. 22, pp. 415–427, 1957, doi: https://doi.org/10.1101/SQB.1957.022.01.039.
[131] D. Rabinowitz, “29. Seven Forms of Rarity (1981),” in *Classic Papers with Commentaries*, F. A. Smith, J. L. Gittleman, and J. H. Brown, Eds., Chicago: University of Chicago Press, 2014, pp. 480–494. doi: doi:10.7208/9780226115504-033.
[132] A. S. Amend *et al.*, “Macroecological patterns of marine bacteria on a global scale,” *Journal of Biogeography*, vol. 40, no. 4, pp. 800–811, 2013, doi: https://doi.org/10.1111/jbi.12034.
[133] “Zoogeography of the Sea,” *BioScience*, vol. 3, no. 2, pp. 17–17, Apr. 1953, doi: 10.1093/aibsbulletin/3.2.17-e.
[134] HUBBS C. L., “Antitropical distribution of fishes and other organisms. Symposium on problems of bipolarity and of pantemperate faunas,” *Proc. Seventh Pac. Sci. Congr. (Pac. Sci. Assoc.)*, vol. 3, pp. 324–329, 1952.
[135] F. O. Aylward, J. M. Eppley, J. M. Smith, and E. F. Delong, “Microbial Community Transcriptional Networks are Conserved in Three Domains at Ocean Basin Scales,” *PNAS*, 2015.
[136] G. H. Tilstone, P. K. Lange, A. Misra, R. J. W. Brewin, and T. Cain, “Micro-phytoplankton photosynthesis, primary production and potential export production in the Atlantic Ocean,” *Progress in Oceanography*, vol. 158, pp. 109–129, Nov. 2017, doi: 10.1016/j.pocean.2017.01.006.
[137] T. Li *et al.*, “Eukaryotic plankton community assembly and influencing factors between continental shelf and slope sites in the northern South China Sea,” *Environmental Research*, vol. 216, p. 114584, Jan. 2023, doi: 10.1016/j.envres.2022.114584.
[138] W. Foissner, “Biogeography and Dispersal of Micro-organisms: A Review Emphasizing Protists,” *Acta Protozoologica*, 2006.
[139] J. J. Pierella Karlusich, K. Cosnier, L. Zinger, and C. Bowler, “Patterns and drivers of diatom diversity and abundance in the global ocean,” *bioRxiv*, 2024.
[140] N. Guérin, M. Ciccarella, E. Flamant, and Q. Carradec, “Genomic adaptation of the picoeukaryote Pelagomonas calceolata to iron-poor oceans revealed by a chromosome-scale genome sequence,” *Communications Biology*, 2022.
[141] T. Vannier, J. Leconte, Y. Seeleuthner, and O. Jaillon, “Survey of the green picoalga Bathycoccus genomes in the global ocean,” 2016.
[142] J. Leconte *et al.*, “Equatorial to Polar Genomic Variability of the microalgae Bathycoccus prasinos,” *bioRxiv*, 2021.
[143] C. Bachy *et al.*, “Viruses infecting a warm water picoeukaryote shed light on spatial co-occurrence dynamics of marine viruses and their hosts,” *The ISME Journal*, vol. 15, no. 11, pp. 3129–3147, Nov. 2021, doi: 10.1038/s41396-021-00989-9.
[144] M.-A. Madoui, J. Poulain, K. Sugier, and P. Wincker, “New Insights into Global Biogeography, Population Structure and Natural Selection from the Genome of the Epipelagic Copepod Oithona,” *Molecular Ecology*, 2017.
[145] P. Cermeno, T. Rodríguez-Ramos, and S. M. Vallina, “Species richness in marine phytoplankton communities is not correlated to ecosystem productivity,” *Mar Ecol Prog Ser*, 2013.

[146] G. C. Stevens, “The Latitudinal Gradient in Geographical Range: How so Many Species Coexist in the Tropics,” *The American Naturalist*, vol. 133, no. 2, pp. 240–256, Feb. 1989, doi: 10.1086/284913.
[147] D. J. Currie *et al.*, “Predictions and tests of climate-based hypotheses of broad-scale variation in taxonomic richness,” *Ecology Letters*, vol. 7, no. 12, pp. 1121–1134, Dec. 2004, doi: 10.1111/j.1461-0248.2004.00671.x.
[148] L. E. Holman, M. de Bruyn, S. Creer, and M. Rius, “Animals, protists and bacteria share marine biogeographic patterns,” *Nature Ecology and Evolution*, 2021.
[149] E. R. Abraham, “The generation of plankton patchiness by turbulent stirring,” *Nature*, vol. 391, no. 6667, pp. 577–580, Feb. 1998, doi: 10.1038/35361.
[150] A. D. Barton, B. A. Ward, R. G. Williams, and M. J. Follows, “The impact of fine-scale turbulence on phytoplankton community structure,” *Limnology and Oceanography: Fluids and Environments*, vol. 4, no. 1, pp. 34–49, Apr. 2014, doi: 10.1215/21573689-2651533.
[151] W. Wu *et al.*, “Contrasting the relative importance of species sorting and dispersal limitation in shaping marine bacterial versus protist communities,” *The ISME Journal*, vol. 12, no. 2, pp. 485–494, Feb. 2018, doi: 10.1038/ismej.2017.183.
[152] M. Vellend, *The Theory of Ecological Communities*. Princeton University Press, 2017. doi: 10.1515/9781400883790.
[153] R. Logares *et al.*, “Disentangling the mechanisms shaping the surface ocean microbiota,” *Microbiome*, vol. 8, no. 1, p. 55, Apr. 2020, doi: 10.1186/s40168-020-00827-8.
[154] B. A. Ward, B. B. Cael, S. Collins, and C. R. Young, “Selective constraints on global plankton dispersal,” *PNAS*, 2021.
[155] F. L. Hellweger, E. van Sebille, and N. D. Fredrick, “Biogeographic Patterns in Ocean Microbes Emerge in a Neutral Agent-based Model,” *Microbial Ecology*, 2014.
[156] L. Zinger *et al.*, “Global Patterns of Bacterial Beta-Diversity in Seafloor and Seawater Ecosystems,” *PLOS ONE*, vol. 6, no. 9, pp. 1–11, Sep. 2011, doi: 10.1371/journal.pone.0024570.
[157] V. F. Farjalla *et al.*, “Ecological determinism increases with organism size,” *Ecology*, vol. 93, no. 7, pp. 1752–1759, Jul. 2012, doi: 10.1890/11-1144.1.
[158] A. Isabwe, H. Yao, S. Zhang, Y. Jiang, M. F. Breed, and X. Sun, “Spatial assortment of soil organisms supports the size-plasticity hypothesis,” *ISME Communications*, vol. 2, no. 1, p. 102, Oct. 2022, doi: 10.1038/s43705-022-00185-6.
[159] E. Ser-Giacomi, L. Zinger, S. Malviya, and S. De Monte, “Ubiquitous abundance distribution of non-dominant plankton across the global ocean,” *Nature Ecology and Evolution*, 2018.
[160] P. V. Martin, A. Bucek, T. Bourguignon, and S. Pigolotti, “Ocean currents promote rare species diversity in protists,” *Science Advances*, 2020.
[161] E. Marañón, P. Cermeño, M. Huete-Ortega, D. C. López-Sandoval, B. Mouriño-Carballido, and T. Rodríguez-Ramos, “Resource Supply Overrides Temperature as a Controlling Factor of Marine Phytoplankton Growth,” *PLOS ONE*, vol. 9, no. 6, p. e99312, Jun. 2014, doi: 10.1371/journal.pone.0099312.
[162] A. Régimbeau, O. Aumont, C. Bowler, and D. Eveillard, “Towards modeling genome-scale knowledge in the global ocean,” *bioRxiv*, 2023.
[163] M. Caffin *et al.*, “N2 fixation as a dominant new N source in the western tropical South Pacific Ocean (OUTPACE cruise),” *Biogeosciences*, vol. 15, no. 8, pp. 2565–2585, 2018, doi: 10.5194/bg-15-2565-2018.
[164] R. P. Abernathey and J. Marshall, “Global surface eddy diffusivities derived from satellite altimetry,” *Journal of Geophysical Research: Oceans*, vol. 118, no. 2, pp. 901–916, Feb. 2013, doi: 10.1002/jgrc.20066.
[165] E. Pigani, B. Hay Mele, L. Campese, and S. Suweis, “Deviation from neutral species abundance distributions unveils geographical differences in the structure of diatom communities,” *Science Advances*, 2024.

[166] S. P. Hubbell, *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, 2001. doi: 10.1515/9781400837526.
[167] P. A. Abrams, “A world without competition,” *Nature*, vol. 412, no. 6850, pp. 858–859, Aug. 2001, doi: 10.1038/35091120.
[168] J. H. Brown, J. F. Gillooly, A. P. Allen, V. M. Savage, and G. B. West, “TOWARD A METABOLIC THEORY OF ECOLOGY,” *Ecology*, vol. 85, no. 7, pp. 1771–1789, Jul. 2004, doi: 10.1890/03-9000.
[169] M. Levitt and C. Chothia, “Structural patterns in globular proteins,” *Nature*, vol. 261, no. 5561, pp. 552–558, Jun. 1976, doi: 10.1038/261552a0.
[170] B. A. CUNNINGHAM, P. D. GOTTLIEB, M. N. PFLUMM, and G. M. EDELMAN, “Immunoglobulin Structure: Diversity, Gene Duplication, and Domains**Supported by grant GB 8371 from the National Science Foundation and by grants AM 04256 and AI 09273 from the National Institutes of Health.,” in *Progress in Immunology*, B. AMOS, Ed., Academic Press, 1971, pp. 3–24. doi: 10.1016/B978-0-12-057550-3.50007-7.
[171] D. B. Wetlaufer, “Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins,” *Proceedings of the National Academy of Sciences*, vol. 70, no. 3, pp. 697–701, Mar. 1973, doi: 10.1073/pnas.70.3.697.
[172] R. D. Schaeffer and V. Daggett, “Protein folds and protein folding.,” *Protein Eng Des Sel*, vol. 24, no. 1–2, pp. 11–19, Jan. 2011, doi: 10.1093/protein/gzq096.
[173] J. S. Richardson, “The Anatomy and Taxonomy of Protein Structure,” in *Advances in Protein Chemistry*, vol. 34, C. B. Anfinsen, J. T. Edsall, and F. M. Richards, Eds., Academic Press, 1981, pp. 167–339. doi: 10.1016/S0065-3233(08)60520-3.
[174] G. Apic, J. Gough, and S. Teichmann, “An insight into domain combinations,” *Bioinformatics*, vol. 17 Suppl 1, pp. S83-9, 2001, doi: 10.1093/bioinformatics/17.suppl_1.s83.
[175] N. Bordin *et al.*, “AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms,” *Communications Biology*, vol. 6, no. 1, p. 160, Feb. 2023, doi: 10.1038/s42003-023-04488-9.
[176] C. Zhang and C. DeLisi, “Estimating the Number of Protein Folds,” *JMB*, 1998.
[177] Y. I. Wolf, N. V. Grishin, and E. V. Koonin, “Estimating the number of protein folds and families from complete genome data11Edited by J. Thornton,” *Journal of Molecular Biology*, vol. 299, no. 4, pp. 897–905, Jun. 2000, doi: 10.1006/jmbi.2000.3786.
[178] C. Chothia, “One thousand families for the molecular biologist,” *Nature*, vol. 357, no. 6379, pp. 543–544, Jun. 1992, doi: 10.1038/357543a0.
[179] S. Govindarajan, R. Recabarren, and R. A. Goldstein, “Estimating the total number of protein folds.,” *Proteins*, vol. 35, no. 4, pp. 408–414, Jun. 1999.
[180] X. Liu, K. Fan, and W. Wang, “The Number of Protein Folds and Their Distribution Over Families in Nature,” *Proteines: Structure, Function, and Bioinformatics*, 2004.
[181] X. Liu, B. Lv, and W. Guo, “The size distribution of protein families within different types of folds,” *Biochemical and Biophysical Research Communications*, 2011.
[182] M. Varadi *et al.*, “AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences.,” *Nucleic Acids Res*, vol. 52, no. D1, pp. D368–D375, Jan. 2024, doi: 10.1093/nar/gkad1011.
[183] D. R. Armstrong *et al.*, “PDBe: improved findability of macromolecular structure data in the PDB,” *Nucleic Acids Research*, vol. 48, no. D1, pp. D335–D343, Jan. 2020, doi: 10.1093/nar/gkz990.
[184] A. M. Lau *et al.*, “Exploring structural diversity across the protein universe with The Encyclopedia of Domains,” *Science*, vol. 386, 2024, doi: 10.1126/science.adq4946.
[185] R. Day, D. A. C. Beck, R. S. Armen, and V. Daggett, “A consensus view of fold space: Combining SCOP, CATH, and the Dali Domain Dictionary,” *Protein Science*, vol. 12, no. 10, pp. 2150–2160, Oct. 2003, doi: 10.1110/ps.0306803.

[186] N. K. Fox, S. E. Brenner, and J.-M. Chandonia, “SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures,” *Nucleic Acids Research*, vol. 42, no. D1, pp. D304–D309, Jan. 2014, doi: 10.1093/nar/gkt1240.
[187] J.-M. Chandonia, L. Guan, S. Lin, C. Yu, N. K. Fox, and S. E. Brenner, “SCOPe: improvements to the structural classification of proteins – extended database to facilitate variant interpretation and machine learning,” *Nucleic Acids Research*, vol. 50, no. D1, pp. D553–D559, Jan. 2022, doi: 10.1093/nar/gkab1054.
[188] A. Andreeva, E. Kulesha, J. Gough, and A. G. Murzin, “The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures,” *Nucleic Acids Research*, vol. 48, no. D1, pp. D376–D382, Jan. 2020, doi: 10.1093/nar/gkz1064.
[189] H. Cheng *et al.*, “ECOD: An Evolutionary Classification of Protein Domains,” *PLOS Computational Biology*, vol. 10, no. 12, p. e1003926, Dec. 2014, doi: 10.1371/journal.pcbi.1003926.
[190] H. Cheng, Y. Liao, R. D. Schaeffer, and N. V. Grishin, “Manual classification strategies in the ECOD database,” *Proteins: Structure, Function, and Bioinformatics*, vol. 83, no. 7, pp. 1238–1251, Jul. 2015, doi: 10.1002/prot.24818.
[191] R. D. Schaeffer, Y. Liao, H. Cheng, and N. V. Grishin, “ECOD: new developments in the evolutionary classification of domains,” *Nucleic Acids Research*, vol. 45, no. D1, pp. D296–D302, Jan. 2017, doi: 10.1093/nar/gkw1137.
[192] L. Holm, “DALI and the persistence of protein shape,” *Protein Science*, vol. 29, no. 1, pp. 128–140, Jan. 2020, doi: 10.1002/pro.3749.
[193] V. P. Waman *et al.*, “CATH 2024: CATH-AlphaFlow Doubles the Number of Structures in CATH and Reveals Nearly 200 New Folds,” *Journal of Molecular Biology*, vol. 436, no. 17, p. 168551, Sep. 2024, doi: 10.1016/j.jmb.2024.168551.
[194] G. Postic, Y. Ghouzam, R. Chebrek, and J.-C. Gelly, “An Ambiguity Principle for Assigning Protein Structural Domains,” *Science Advances*, 2017.
[195] J. Wells, A. Hawkins-Hooker, N. Bordin, I. Sillitoe, B. Paige, and C. Orengo, “Chainsaw: protein domain segmentation with fully convolutional neural networks,” *Bioinformatics*, vol. 40, no. 5, p. btae296, May 2024, doi: 10.1093/bioinformatics/btae296.
[196] K. Zhu, H. Su, Z. Peng, and J. Yang, “A unified approach to protein domain parsing with inter-residue distance matrix,” *Bioinformatics*, vol. 39, no. 2, p. btad070, Feb. 2023, doi: 10.1093/bioinformatics/btad070.
[197] The UniProt Consortium, “UniProt: a worldwide hub of protein knowledge,” *Nucleic Acids Research*, vol. 47, no. D1, pp. D506–D515, Jan. 2019, doi: 10.1093/nar/gky1049.
[198] A. D. Yates *et al.*, “Ensembl 2020,” *Nucleic Acids Research*, vol. 48, no. D1, pp. D682–D688, Jan. 2020, doi: 10.1093/nar/gkz966.
[199] G. Csaba, F. Birzele, and R. Zimmer, “Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis,” *BMC Structural Biology*, vol. 9, no. 1, p. 23, Apr. 2009, doi: 10.1186/1472-6807-9-23.
[200] L. L. Porter and L. L. Looger, “Extant fold-switching proteins are widespread,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 23, pp. 5968–5973, Jun. 2018, doi: 10.1073/pnas.1800168115.
[201] A. Cuff *et al.*, “The CATH Hierarchy Revisited—Structural Divergence in Domain Superfamilies and the Continuity of Fold Space,” *Structure*, vol. 17, no. 8, pp. 1051–1062, Aug. 2009, doi: 10.1016/j.str.2009.06.015.
[202] V. Alva, K. K. Koretke, M. Coles, and A. N. Lupas, “Cradle-loop barrels and the concept of metafolds in protein classification by natural descent,” *Curr Opin Struct Biol*, vol. 18, no. 3, pp. 358–365, Jun. 2008, doi: 10.1016/j.sbi.2008.02.006.
[203] Y. Ghouzam, G. Postic, A. G. de Brevern, and J.-C. Gelly, “Improving Protein Fold Recognition with Hybrid Profiles Combining Sequence and Structure Evolution,” *Structural Bioinformatics*, 2015.

[204] A. S. Konagurthu *et al.*, “Universal Architectural Concepts Underlying Protein Folding Patterns,” 2021.
[205] C. O. Mackenzie, J. Zhou, and G. Grigoryan, “Tertiary alphabet for the observable protein structural universe,” 2016.
[206] E. Haber and C. B. Anfinsen, “Regeneration of Enzyme Activity by Air Oxidation of Reduced Subtilisin-Modified Ribonuclease,” *Journal of Biological Chemistry*, vol. 236, no. 2, pp. 422–424, Feb. 1961, doi: 10.1016/S0021-9258(18)64379-0.
[207] S. Pechmann, J. W. Chartron, and J. Frydman, “Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP in vivo,” *Nature Structural & Molecular Biology*, vol. 21, no. 12, pp. 1100–1105, Dec. 2014, doi: 10.1038/nsmb.2919.
[208] D. N. Ivankov, S. O. Garbuzynskiy, E. Alm, K. W. Plaxco, D. Baker, and A. V. Finkelstein, “Contact order revisited: Influence of protein size on the folding rate,” *Protein Science*, vol. 12, no. 9, pp. 2057–2062, Sep. 2003, doi: 10.1110/ps.0302503.
[209] S. Sacquin-Mora, “Fold and Flexibility: What Can Proteins’ Mechanical Properties Tell us About Their Folding Nucleus ?,” *The Royal Society*, 2015.
[210] J. W. A. te Velthuis and P. C. Bagowski, “Linking Fold, Function and Phylogeny: A Comparative Genomics View on Protein (Domain) Evolution,” *Current Genomics*, vol. 9, no. 2, pp. 88–96, 2008, doi: 10.2174/138920208784139537.
[211] S. E. Jackson, “How do small single-domain proteins fold?,” *Folding and Design*, vol. 3, no. 4, pp. R81–R91, Aug. 1998, doi: 10.1016/S1359-0278(98)00033-9.
[212] N. Tokuriki and D. S. Tawfik, “Protein Dynamism and Evolvability,” *Science*, 2009.
[213] J.-H. Han, S. Batey, A. A. Nickson, S. A. Teichmann, and J. Clarke, “The folding and evolution of multidomain proteins,” *Nature Reviews Molecular Cell Biology*, vol. 8, no. 4, pp. 319–330, Apr. 2007, doi: 10.1038/nrm2144.
[214] G. Apic, J. Gough, and S. A. Teichmann, “Domain combinations in archaeal, eubacterial and eukaryotic proteomes11Edited by G. von Heijne,” *Journal of Molecular Biology*, vol. 310, no. 2, pp. 311–325, Jul. 2001, doi: 10.1006/jmbi.2001.4776.
[215] M. Bashton and C. Chothia, “The geometry of domain combination in proteins11Edited by J. Thornton,” *Journal of Molecular Biology*, vol. 315, no. 4, pp. 927–939, Jan. 2002, doi: 10.1006/jmbi.2001.5288.
[216] I. Sorokina, A. R. Mushegian, and E. V. Koonin, “Is Protein Folding a Thermodynamically Unfavorable, Active, Energy-Dependent Process?,” *International Journal of Molecular Sciences*, 2022.
[217] J. L. England and E. I. Shakhnovich, “Structural Determinant of Protein Designability,” *Phys. Rev. Lett.*, vol. 90, no. 21, p. 218101, May 2003, doi: 10.1103/PhysRevLett.90.218101.
[218] S. Kumar, B. Ma, C.-J. Tsai, N. Sinha, and R. Nussinov, “Folding and binding cascades: Dynamic landscapes and population shifts,” *Protein Science*, vol. 9, no. 1, pp. 10–19, Jan. 2000, doi: 10.1110/ps.9.1.10.
[219] C.-J. Tsai, S. Kumar, B. Ma, and R. Nussinov, “Folding funnels, binding funnels, and protein function,” *Protein Science*, vol. 8, no. 6, pp. 1181–1190, Jan. 1999, doi: 10.1110/ps.8.6.1181.
[220] A. K. Dunker *et al.*, “Intrinsically disordered protein.,” *J Mol Graph Model*, vol. 19, no. 1, pp. 26–59, 2001, doi: 10.1016/s1093-3263(00)00138-8.
[221] J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones, “Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life,” *Journal of Molecular Biology*, vol. 337, no. 3, pp. 635–645, Mar. 2004, doi: 10.1016/j.jmb.2004.02.002.
[222] B. Xue, A. K. Dunker, and V. N. Uversky, “Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life.,” *J Biomol Struct Dyn*, vol. 30, no. 2, pp. 137–149, 2012, doi: 10.1080/07391102.2012.675145.
[223] G. Caetano-Anollés and D. Caetano-Anollés, “An Evolutionarily Structured Universe of Protein Architecture,” *Genome Research*, 2003.

[224] G. D. Rose, P. J. Fleming, J. R. Banavar, and A. Maritan, “A backbone-based theory of protein folding,” *Proceedings of the National Academy of Sciences*, vol. 103, no. 45, pp. 16623–16633, Nov. 2006, doi: 10.1073/pnas.0606843103.
[225] M. D. Yoder and F. Jurnak, “The parallel β helix and other coiled folds,” *The FASEB Journal*, vol. 9, no. 5, pp. 335–342, Mar. 1995, doi: 10.1096/fasebj.9.5.7896002.
[226] G. M. Salem, E. G. Hutchinson, C. A. Orengo, and J. M. Thornton, “Correlation of Observed Fold Frequency with the Occurrence of Local Structural Motifs,” *Journal of Molecular Biology*, vol. 287, no. 5, pp. 969–981, 1999, doi: https://doi.org/10.1006/jmbi.1999.2642.
[227] Á. Tóth-Petróczy and D. S. Tawfik, “The robustness and innovability of protein folds,” *Current Opinion in Structural Biology*, vol. 26, pp. 131–138, Jun. 2014, doi: 10.1016/j.sbi.2014.06.007.
[228] E. Dellus-Gur, A. Toth-Petroczy, M. Elias, and D. S. Tawfik, “What Makes a Protein Fold Amenable to Functional Innovation? Fold Polarity and Stability Trade-offs,” *Journal of Molecular Biology*, vol. 425, no. 14, pp. 2609–2621, Jul. 2013, doi: 10.1016/j.jmb.2013.03.033.
[229] M. Osadchy and R. Kolodny, “Maps of protein structure space reveal a fundamental relationship between protein structure and function,” *Proceedings of the National Academy of Sciences*, vol. 108, no. 30, pp. 12301–12306, Jul. 2011, doi: 10.1073/pnas.1102727108.
[230] X. Han, A. Sit, C. Christoffer, S. Chen, and D. Kihara, “A global map of the protein shape universe,” *PLoS Computational Biology*, 2019.
[231] A. Pascual-García, D. Abia, Á. R. Ortiz, and U. Bastolla, “Cross-Over between Discrete and Continuous Protein Structure Space: Insights into Automatic Classification and Networks of Protein Structures,” *PLOS Computational Biology*, vol. 5, no. 3, p. e1000331, Mar. 2009, doi: 10.1371/journal.pcbi.1000331.
[232] J. Hou, G. E. Sims, C. Zhang, and S.-H. Kim, “A global representation of the protein fold space,” *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2386–2390, Mar. 2003, doi: 10.1073/pnas.2628030100.
[233] J. Hou, S.-R. Jun, C. Zhang, and S.-H. Kim, “Global mapping of the protein structure space and application in structure-based inference of protein function,” *Proceedings of the National Academy of Sciences*, vol. 102, no. 10, pp. 3651–3656, Mar. 2005, doi: 10.1073/pnas.0409772102.
[234] E. Zuckerkandl and L. Pauling, “Molecules as documents of evolutionary history,” *Journal of Theoretical Biology*, vol. 8, no. 2, pp. 357–366, Mar. 1965, doi: 10.1016/0022-5193(65)90083-4.
[235] A. C. Martin, C. A. Orengo, E. G. Hutchinson, and J. M. Thornton, “Protein folds and functions,” *Structure*, 1998.
[236] M. M. Konaté, G. Plata, J. Park, and D. Vitkup, “Molecular function limits divergent protein evolution on planetary timescales,” *eLife*, 2019.
[237] A. G. Murzin, “Metamorphic Proteins,” *Science*, vol. 320, no. 5884, pp. 1725–1726, Jun. 2008, doi: 10.1126/science.1158868.
[238] L. C. James and D. S. Tawfik, “Conformational diversity and protein evolution – a 60-year-old hypothesis revisited,” *Trends in Biochemical Sciences*, vol. 28, no. 7, pp. 361–368, Jul. 2003, doi: 10.1016/S0968-0004(03)00135-X.
[239] R. L. Tuinstra, F. C. Peterson, S. Kutlesa, E. S. Elgin, M. A. Kron, and B. F. Volkman, “Interconversion between two unrelated protein folds in the lymphotactin native state,” *Proceedings of the National Academy of Sciences*, vol. 105, no. 13, pp. 5057–5062, Apr. 2008, doi: 10.1073/pnas.0709518105.
[240] A. K. Dunker and Z. Obradovic, “The protein trinity—linking function and disorder,” *Nature Biotechnology*, vol. 19, no. 9, pp. 805–806, Sep. 2001, doi: 10.1038/nbt0901-805.
[241] K. E. Medvedev, L. N. Kinch, R. D. Schaeffer, and N. V. Grishin, “Functional analysis of Rossmann-like domains reveals convergent evolution of topology and reaction pathways,” *PLOS Computational Biology*, vol. 15, no. 12, p. e1007569, Dec. 2019, doi: 10.1371/journal.pcbi.1007569.

[242] K. E. Medvedev, L. N. Kinch, R. D. Schaeffer, J. Pei, and N. V. Grishin, “A Fifth of the Protein World: Rossmann-like Proteins as an Evolutionarily Successful Structural unit,” 2020.
[243] L. Aravind, R. Mazumder, S. Vasudevan, and E. V. Koonin, “Trends in protein evolution inferred from sequence and structure analysis,” *Current Opinion in Structural Biology*, vol. 12, no. 3, pp. 392–399, Jun. 2002, doi: 10.1016/S0959-440X(02)00334-2.
[244] R. K. Wierenga, “The TIM-barrel fold: a versatile framework for efficient enzymes,” *FEBS Letters*, vol. 492, no. 3, pp. 193–198, Mar. 2001, doi: 10.1016/S0014-5793(01)02236-0.
[245] D. W. A. Buchan, A. J. Shepherd, and C. A. Orengo, “Gene3D: Structural Assignment for Whole Genes and Genomes Using the CATH Domain Structure Database,” 2002.
[246] E. Ferrada and A. Wagner, “Protein robustness promotes evolutionary innovations on large evolutionary time-scales.,” *Proc Biol Sci*, vol. 275, no. 1643, pp. 1595–1602, Jul. 2008, doi: 10.1098/rspb.2007.1617.
[247] R. Unger, S. Uliel, and S. Havlin, “Scaling law in sizes of protein sequence families: From super-families to orphan genes,” *Proteins: Structure, Function, and Bioinformatics*, vol. 51, no. 4, pp. 569–576, Jun. 2003, doi: 10.1002/prot.10347.
[248] S. Burke and R. Elber, “Super folds, networks, and barriers,” *Proteins*, vol. 80, pp. 463–470, 2012, doi: https://doi.org/10.1002/prot.23212.
[249] A. Magner, W. Szpankowski, and D. Kihara, “On the Origin of Protein Superfamilies and Superfolds,” *Scientific Reports*, 2015.
[250] C. A. Orengo, D. T. Jones, and J. M. Thornton, “Protein superfamilies and domain superfolds,” 1994.
[251] A. F. W. Coulson and J. Moult, “A Unifold, Mesofold, and Superfold Model of Protein Fold Use,” 2002.
[252] A. V. Finkelstein, A. M. Gutun, and A. Y. Badretdinov, “Why are the same protein folds used to perform different functions?,” *FEBS Letters*, vol. 325, 1993, [Online]. Available: https://api.semanticscholar.org/CorpusID:6892409
[253] V. A. Kuznetsov, “Protein Domain Statistics in Proteomes,” in *Scale-Dependent Statistics of the Numbers of Transcripts and Protein Sequences Encoded in the Genome*, 2002.
[254] S. K. Forslund, M. Kaduk, and E. L. L. Sonnhammer, “Evolution of Protein Domain Architectures,” *Method Mol Biol.*, pp. 469–504, 2019.
[255] E. V. Koonin, Y. I. Wolf, and G. P. Karev, “The structure of the protein universe and genome evolution,” *Nature*, 2002.
[256] G. P. Karev, Y. I. Wolf, A. Y. Rzhetsky, F. S. Berezovskaya, and E. V. Koonin, “Birth and death of protein domains: A simple model of evolution explains power law behavior,” *BMC Evolutionary Biology*, 2002.
[257] S. Abeln and C. M. Deane, “Fold Usage on Genomes and Protein Fold Evolution,” *Proteines: Structure, Function, and Bioinformatics*, 2005.
[258] J. Qian, N. M. Luscombe, and M. Gerstein, “Protein Family and Fold Occurrence in Genomes: Power-law Behaviour and Evolutionary Model,” *Journal of Molecular Biology*, 2001.
[259] V. A. Kuznetsov, “Family of skewed distributions associated with the gene expression and proteome evolution,” *Signal Processing*, vol. 83, no. 4, pp. 889–910, 2003, doi: https://doi.org/10.1016/S0165-1684(02)00481-4.
[260] S. Yang, R. F. Doolittle, and P. E. Bourne, “Phylogeny Determined by Protein Domain Content,” *PNAS*, 2005.
[261] E. V. Koonin *et al.*, “A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes,” *Genome Biology*, vol. 5, no. 2, p. R7, Jan. 2004, doi: 10.1186/gb-2004-5-2-r7.
[262] M. Itoh, J. C. Nacher, K. Kuma, and M. Kanehisa, “Evolutionary history and functional implications of protein domains and their combinations in eukaryotes,” *Genome Biology*, 2007.
[263] X.-C. Zhang *et al.*, “Evolutionary dynamics of protein domain architecture in plants.,” *BMC Evol Biol*, vol. 12, p. 6, Jan. 2012, doi: 10.1186/1471-2148-12-6.

[264] M. Romei, M. Carpentier, J. Chomilier, and G. Lecointre, “Origins and Functional Significance of Eukaryotic Protein Folds,” *Journal of Molecular Evolution*, vol. 91, no. 6, pp. 854–864, Dec. 2023, doi: 10.1007/s00239-023-10136-x.
[265] M. Wang, L. S. Yafremava, D. Caetano-Anollés, and G. Caetano-Anollés, “Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world,” *Genome Research*, 2007.
[266] Y. I. Wolf, S. E. Brenner, P. A. Bash, and E. V. Koonin, “Distribution of Protein Folds in the Three Superkingdoms of Life,” *Genome Research*, 1999.
[267] I. Budimir, E. Giampieri, E. Saccenti, and C. Sala, “Intraspecies characterization of bacteria via evolutionary modeling of protein domains,” *Scientific Reports*, 2022.
[268] J. Lin and M. Gerstein, “Whole-Genome Trees Based on the Occurrence of Folds and Orthologs: Implications for Comparing Genomes on Different Levels,” *Genome Research*, 2000.
[269] G. J. P. Naylor and W. M. Brown, “Structural biology and phylogenetic estimation,” *Nature*, vol. 388, no. 6642, pp. 527–528, Aug. 1997, doi: 10.1038/41460.
[270] T. A. Williams, C. J. Cox, P. G. Foster, G. J. Szöllősi, and T. M. Embley, “Phylogenomics provides robust support for a two-domains tree of life,” *Nature Ecology & Evolution*, vol. 4, no. 1, pp. 138–147, Jan. 2020, doi: 10.1038/s41559-019-1040-x.
[271] C. Alvarez-Carreño, R. J. Gupta, A. S. Petrov, and L. D. Williams, “Creative destruction: New protein folds from old,” *Proceedings of the National Academy of Sciences*, vol. 119, no. 52, p. e2207897119, Dec. 2022, doi: 10.1073/pnas.2207897119.
[272] N. V. Grishin, “Fold Change in Evolution of Protein Structures,” *Journal of Structural Biology*, vol. 134, no. 2, pp. 167–185, May 2001, doi: 10.1006/jsbi.2001.4335.
[273] B. A. Cunningham, J. J. Hemperly, T. P. Hopp, and G. M. Edelman, “Favin versus concanavalin A: Circularly permuted amino acid sequences,” *Proceedings of the National Academy of Sciences*, vol. 76, no. 7, pp. 3218–3222, Jul. 1979, doi: 10.1073/pnas.76.7.3218.
[274] M. Véron, F. Falcoz-Kelly, and G. N. Cohen, “The Threonine-Sensitive Homoserine Dehydrogenase and Aspartokinase Activities of Escherichia coli K12,” *European Journal of Biochemistry*, vol. 28, no. 4, pp. 520–527, Aug. 1972, doi: 10.1111/j.1432-1033.1972.tb01939.x.
[275] B. Rost, “Enzyme Function Less Conserved than Anticipated,” *Journal of Molecular Biology*, vol. 318, no. 2, pp. 595–608, Apr. 2002, doi: 10.1016/S0022-2836(02)00016-5.
[276] L. Yu, D. K. Tanwar, E. D. S. Penha, and M. K. Basu, “Grammar of protein domain architectures,” *PNAS*, 2019.
[277] M. Y. Galperin and E. V. Koonin, “Divergence and Convergence in Enzyme Evolution,” *Journal of Biological Chemistry*, 2012.
[278] E. J. Deeds, B. Shakhnovich, and E. I. Shakhnovich, “Proteomic Traces of Speciation,” 2003.
[279] M. Y. Galperin, D. R. Walker, and E. V. Koonin, “Analogous Enzymes: Independent Inventions in Enzyme Evolution,” *Genome Research*, 1998.
[280] M. V. Omelchenko, M. Y. Galperin, Y. I. Wolf, and E. V. Koonin, “Non-homologous isofunctional enzymes: A systematic analysis of alternative solutions in enzyme evolution,” *Biology Direct*, vol. 5, pp. 31–31, 2010.
[281] A. R. Buller and C. A. Townsend, “Intrinsic evolutionary constraints on protease structure, enzyme acylation, and the identity of the catalytic triad,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 8, pp. E653–E661, Feb. 2013, doi: 10.1073/pnas.1221050110.
[282] K. Tomii, Y. Sawada, and S. Honda, “Convergent evolution in structural elements of proteins investigated using cross profile analysis,” *BMC Bioinformatics*, vol. 13, no. 1, p. 11, Jan. 2012, doi: 10.1186/1471-2105-13-11.
[283] H. M. Beyer *et al.*, “The Convergence of the Hedgehog/Intein Fold in Different Protein Splicing Mechanisms.,” *Int J Mol Sci*, vol. 21, no. 21, Nov. 2020, doi: 10.3390/ijms21218367.

[284] V. A. Kuznetsov, “Validation of Random Birth-Death Model of Evolution of Proteome Complexity,” 2004.
[285] A. W. R. Serohijos, S. Y. R. Lee, and E. I. Shakhnovich, “Highly Abundant Proteins Favor More Stable 3D Structures in Yeast,” *Biophysical Journal*, vol. 104, no. 3, pp. L1–L3, Feb. 2013, doi: 10.1016/j.bpj.2012.11.3838.
[286] S. H. White and R. E. Jacobs, “Statistical distribution of hydrophobic residues along the length of protein chains. Implications for protein folding and evolution,” *Biophysical Journal*, vol. 57, no. 4, pp. 911–921, Apr. 1990, doi: 10.1016/S0006-3495(90)82611-4.
[287] S. White and R. Jacobs, “The evolution of proteins from random amino acid sequences. I. Evidence from the lengthwise distribution of amino acids in modern protein sequences,” *J Mol Evol*, vol. 36, no. 1, pp. 79–95, Jan. 1993, doi: 10.1007/bf02407307.
[288] S. H. White, “The evolution of proteins from random amino acid sequences: II. Evidence from the statistical distributions of the lengths of modern protein sequences,” *Journal of Molecular Evolution*, vol. 38, no. 4, pp. 383–394, Apr. 1994, doi: 10.1007/BF00163155.
[289] H. Li, R. Helling, C. Tang, and N. Wingreen, “Emergence of Preferred Structures in a Simple Model of Protein Folding,” *Science*, vol. 273, no. 5275, pp. 666–669, Aug. 1996, doi: 10.1126/science.273.5275.666.
[290] K. M. Kim and G. Caetano-Anollés, “The evolutionary history of protein fold families and proteomes confirms that the archaeal ancestor is more ancient than the ancestors of other superkingdoms,” *BMC Evolutionary Biology*, vol. 12, no. 1, p. 13, Jan. 2012, doi: 10.1186/1471-2148-12-13.
[291] F. Mughal, A. Nasir, and G. Caetano-Anollés, “The origin and evolution of viruses inferred from fold family structure,” *Archives of Virology*, 2020.
[292] M. Fayez Aziz and G. Caetano-Anollés, “Evolution of networks of protein domain organization,” *Nature Scientific Reports*, 2021.
[293] H. F. Winstanley, S. Abeln, and C. M. Deane, “How old is your fold?,” *Bioinformatics*, vol. 21, no. suppl_1, pp. i449–i458, Jun. 2005, doi: 10.1093/bioinformatics/bti1008.
[294] A. F. Moutinho, F. F. Trancoso, and J. Y. Dutheil, “The Impact of Protein Architecture on Adaptive Evolution,” *Molecular Biology of Evolution*, 2019.
[295] H.-F. Ji, L. Chen, Y.-Y. Jiang, and H.-Y. Zhang, “Evolutionary formation of new protein folds is linked to metallic cofactor recruitment.,” *Bioessays*, vol. 31, no. 9, pp. 975–980, Sep. 2009, doi: 10.1002/bies.200800201.
[296] J. Liu *et al.*, “Metalloproteins Containing Cytochrome, Iron–Sulfur, or Copper Redox Centers,” *Chem. Rev.*, vol. 114, no. 8, pp. 4366–4469, Apr. 2014, doi: 10.1021/cr400479b.
[297] M. Dumontier, K. Michalickova, and C. W. Hogue, “Species-specific protein sequence and fold optimizations,” *BMC Bioinformatics*, vol. 3, no. 1, p. 39, Dec. 2002, doi: 10.1186/1471-2105-3-39.
[298] K. B. Zeldovich, I. N. Berezovsky, and E. I. Shakhnovich, “Protein and DNA Sequence Determinants of Thermophilic Adaptation,” *PLOS Computational Biology*, vol. 3, no. 1, p. e5, Jan. 2007, doi: 10.1371/journal.pcbi.0030005.
[299] D. B. Sauer and D.-N. Wang, “Predicting the optimal growth temperatures of prokaryotes using only genome derived features,” *Bioinformatics*, vol. 35, no. 18, pp. 3224–3231, Sep. 2019, doi: 10.1093/bioinformatics/btz059.
[300] S. E. Jensen, L. C. Johnson, T. Casstevens, and E. S. Buckler, “Predicting protein domain temperature adaptation across the Prokaryote-Eukaryote divide,” *bioRxiv*, 2021.
[301] J. Gu and V. J. Hilser, “Sequence-Based Analysis of Protein Energy Landscapes Reveals Nonuniform Thermal Adaptation within the Proteome,” *Molecular Biology and Evolution*, vol. 26, no. 10, pp. 2217–2227, Oct. 2009, doi: 10.1093/molbev/msp140.
[302] S.-J. Chen *et al.*, “Protein folds vs. protein folding: Differing questions, different challenges,” *Proceedings of the National Academy of Sciences*, vol. 120, no. 1, p. e2214423119, Jan. 2023, doi: 10.1073/pnas.2214423119.

[303] A. Szilágyi and P. Závodszky, “Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey,” *Structure*, vol. 8, no. 5, pp. 493–504, May 2000, doi: 10.1016/S0969-2126(00)00133-7.
[304] G. N. Somero, “The physiology of climate change: how potentials for acclimatization and genetic adaptation will determine ‘winners’ and ‘losers,’” *The Journal of Experimental Biology*, 2009.
[305] P. A. Fields, Y. Dong, X. Meng, and G. N. Somero, “Adaptation of Protein Structure and Function to Temperature: There is more than one way to ‘skin a cat,’” *The Company of Biologists*, 2015.
[306] C. Struvay and G. Feller, “Optimization to Low Temperature Activity in Psychrophilic Enzymes,” *International Journal of Molecular Sciences*, vol. 13, no. 9, pp. 11643–11665, 2012, doi: 10.3390/ijms130911643.
[307] S. Radestock and H. Gohlke, “Protein rigidity and thermophilic adaptation,” *Proteins: Structure, Function, and Bioinformatics*, vol. 79, no. 4, pp. 1089–1108, Apr. 2011, doi: 10.1002/prot.22946.
[308] H.-K. S. Leiros *et al.*, “Structure of Phenylalanine Hydroxylase from Colwellia psychrerythraea 34H, a Monomeric Cold Active Enzyme with Local Flexibility around the Active Site and High Overall Stability *,” *Journal of Biological Chemistry*, vol. 282, no. 30, pp. 21973–21986, Jul. 2007, doi: 10.1074/jbc.M610174200.
[309] Y.-C. Chao, M. Merritt, D. Schaefferkoetter, and T. G. Evans, “High-throughput quantification of protein structural change reveals potential mechanisms of temperature adaptation in Mytilus mussels,” *BMC Evolutionary Biology*, vol. 20, no. 1, p. 28, Feb. 2020, doi: 10.1186/s12862-020-1593-y.
[310] D. Kültz, “MOLECULAR AND EVOLUTIONARY BASIS OF THE CELLULAR STRESS RESPONSE,” *Annual Review of Physiology*, vol. 67, no. Volume 67, 2005. Annual Reviews, pp. 225–257, 2005. doi: https://doi.org/10.1146/annurev.physiol.67.040403.103635.
[311] A. Gutiérrez-Preciado, B. Dede, B. A. Baker, L. Eme, D. Moreira, and P. López-García, “Extremely acidic proteomes and metabolic flexibility in bacteria and highly diversified archaea thriving in geothermal chaotropic brines,” *Nature Ecology & Evolution*, vol. 8, no. 10, pp. 1856–1869, Oct. 2024, doi: 10.1038/s41559-024-02505-6.
[312] A. A.-T. Weber, A. F. Hugall, and T. D. O’Hara, “Convergent Evolution and Structural Adaptation to the Deep Ocean in the Protein-Folding Chaperonin CCTalpha,” *GBE*, 2020.
[313] P. A. Fields and G. N. Somero, “Hot spots in cold adaptation: Localized increases in conformational flexibility in lactate dehydrogenase A4 orthologs of Antarctic notothenioid fishes,” *Proceedings of the National Academy of Science*, 1998.
[314] S. Chakravarty and R. Varadarajan, “Elucidation of Factors Responsible for Enhanced Thermal Stability of Proteins: A Structural Genomics Based Study,” *Biochemistry*, vol. 41, no. 25, pp. 8152–8161, Jun. 2002, doi: 10.1021/bi025523t.
[315] L.-L. Yang, S.-K. Tang, Y. Huang, and X.-Y. Zhi, “Low Temperature Adaptation Is Not the Opposite Process of High Temperature Adaptation in Terms of Changes in Amino Acid Composition,” *Genome Biology and Evolution*, vol. 7, no. 12, pp. 3426–3433, Dec. 2015, doi: 10.1093/gbe/evv232.
[316] D. P. Kreil and C. A. Ouzounis, “Identification of thermophilic species by the amino acid compositions deduced from their genomes,” *Nucleic Acids Research*, vol. 29, no. 7, pp. 1608–1615, Apr. 2001, doi: 10.1093/nar/29.7.1608.
[317] P. L. Privalov, “Cold Denaturation of Protein,” *Critical Reviews in Biochemistry and Molecular Biology*, vol. 25, no. 4, pp. 281–306, Jan. 1990, doi: 10.3109/10409239009090612.
[318] N. J. Russell, “Toward a molecular understanding of cold activity of enzymes from psychrophiles,” *Extremophiles*, vol. 4, no. 2, pp. 83–90, Apr. 2000, doi: 10.1007/s007920050141.

[319] A. Rizzello, A. Romano, G. Kottra, and M. Maffia, “Protein Cold Adaptation Strategy via a Unique Seven-Amino Acid Domain in the Icefish (Chionodraco hamatus) PEPT1 Transporter,” *PNAS*, 2013.
[320] G. Gianese, P. Argos, and S. Pascarella, “Structural adaptation of enzymes to low temperatures,” *Protein Engineering, Design and Selection*, vol. 14, no. 3, pp. 141–148, Mar. 2001, doi: 10.1093/protein/14.3.141.
[321] C. Berthelot, J. Clarke, T. Desvignes, and M. S. Clark, “Adaptation of Proteins to the cold in Antarctic Fish: A Role for Methionine ?,” *Genome Biology and Evolution*, vol. 11, no. 1, pp. 220–231, 2018, doi: doi: 10.1093/gbe/evy262.
[322] A. O. Smalås, H.-K. S. Leiros, V. Os, and N. P. Willassen, “Cold adapted enzymes.,” *Biotechnology annual review*, vol. 6, pp. 1–57, 2000.
[323] G. N. Somero, “Protein Adaptations to Temperature and Pressure: Complementary Roles of Adaptive Changes in Amino Acid Sequence and Internal Milieu,” *Comparative Biochemistry and Physiology*, 2003.
[324] G. Feller, D. d’Amico, and C. Gerday, “Thermodynamic stability of a cold-active alpha-amylase from the Antarctic bacterium Alteromonas haloplanctis,” *Biochemistry*, vol. 38, no. 14, pp. 4613–4619, Apr. 1999, doi: 10.1021/bi982650+.
[325] S. D’Amico, C. Gerday, and G. Feller, “Structural Determinants of Cold Adaptation and Stability in a Large Protein*,” *Journal of Biological Chemistry*, vol. 276, no. 28, pp. 25791–25796, Jul. 2001, doi: 10.1074/jbc.M102741200.
[326] G. Gianese, F. Bossa, and S. Pascarella, “Comparative structural analysis of psychrophilic and meso- and thermophilic enzymes,” *Proteins: Structure, Function, and Bioinformatics*, vol. 47, no. 2, pp. 236–249, May 2002, doi: 10.1002/prot.10084.
[327] G. Feller, “Protein stability and enzyme activity at extreme biological temperatures,” *Journal of Physics: Condensed Matter*, vol. 22, no. 32, p. 323101, Jul. 2010, doi: 10.1088/0953-8984/22/32/323101.
[328] R. JAENICKE, “Protein stability and molecular adaptation to extreme conditons,” *European Journal of Biochemistry*, vol. 202, no. 3, pp. 715–728, Dec. 1991, doi: 10.1111/j.1432-1033.1991.tb16426.x.
[329] R. Marasco, M. Fusi, C. Coscolin, and D. Daffonchio, “Enzyme adaptation to habitat thermal legacy shapes the thermal plasticity of marine microbiomes,” *Nature Communications*, 2023.
[330] E. Kiefl, O. C. Esen, S. E. Miller, and A. Murat Eren, “Structure-informed microbial population genetics elucidate selective pressures that shape protein evolution,” *bioRxiv*, 2022.
[331] T. O. Delmont, E. Kiefl, O. Kilinc, and E. A. Murat, “Single-Amino Acid Variants Reveal Evolutionary Processes that Shape the Biogeography of a Global SAR11 Subclade,” *eLife*, 2019.
[332] S. Kijima, H. Hikida, T. O. Delmont, M. Gaïa, and H. Ogata, “Complex Genomes of Early Nucleocytoviruses Revealed by Ancient Origins of Viral Aminoacyl-tRNA Synthetases,” *Molecular Biology and Evolution*, vol. 41, no. 8, p. msae149, Aug. 2024, doi: 10.1093/molbev/msae149.
[333] S. Das, D. Lee, I. Sillitoe, N. L. Dawson, J. G. Lees, and C. A. Orengo, “Functional classification of CATH superfamilies: a domain-based approach for protein function annotation,” *Bioinformatics*, vol. 31, no. 21, pp. 3460–3467, Nov. 2015, doi: 10.1093/bioinformatics/btv398.
[334] S. Das *et al.*, “CATH FunFHMMer web server: protein functional annotations using functional family assignments,” *Nucleic Acids Research*, vol. 43, no. W1, pp. W148–W153, Jul. 2015, doi: 10.1093/nar/gkv488.
[335] S. R. Eddy, *hmmscan :: search sequence(s) against a profile database*. (Nov. 2020). [Online]. Available: http://hmmer.org/
[336] C. Ahlmann-Eltze and P. Indrajeet, “ggsignif: R Package for Displaying Significance Brackets for ’ggplot2.,” *PsyArxiv*, 2021, doi: 10.31234/osf.io/7awm6.
[337] A. Kassambara, “ggcorrplot: Visualization of a Correlation Matrix using ‘ggplot2’. R package version 0.1.4.1.,” 2023, [Online]. Available: https://CRAN.R-project.org/package=ggcorrplot

[338] M. Tennekes, “treemap: Treemap Visualization. R package version 2.4-4.,” 2023, [Online]. Available: https://CRAN.R-project.org/package=treemap
[339] B. D. Ondov, N. H. Bergman, and A. M. Phillippy, “Interactive metagenomic visualization in a Web browser,” *BMC Bioinformatics*, vol. 12, no. 1, p. 385, Sep. 2011, doi: 10.1186/1471-2105-12-385.
[340] C.-H. Gao *et al.*, “ggVennDiagram: intuitive Venn diagram software extended.,” *iMeta*, vol. 3, no. 69, 2024, doi: 10.1002/imt2.177. Features.
[341] J. Oksanen, G. L. Simpson, G. Blanchet, and J. Weedon, “vegan: Community Ecology Package. R package version 2.6-4.,” 2022, [Online]. Available: https://CRAN.R-project.org/package=vegan
[342] R Core Team, “R: A Language and Environment for Statistical Computing.,” *R Foundation for Statistical Computing, Vienna, Austria.*, 2023, [Online]. Available: https://www.R-project.org/
[343] E. Paradis and K. Schliep, “ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R.,” *Bioinformatics*, vol. 35, pp. 526–528, 2019, doi: 10.1093/bioinformatics/bty633.
[344] A. M. Eren *et al.*, “Community-led, integrated, reproducible multi-omics with anvi’o,” *Nature Microbiology*, vol. 6, no. 1, pp. 3–6, Jan. 2021, doi: 10.1038/s41564-020-00834-3.
[345] H. Wickham, “ggplot2: Elegant Graphics for Data Analysis.,” *Springer-Verlag New York*, 2016, [Online]. Available: https://ggplot2.tidyverse.org
[346] B. Greenwell, “ramify: Additional Matrix Functionality,” 2016, doi: 10.32614/CRAN.package.ramify.
[347] S. Milojević, “Power law distributions in information science: Making the case for logarithmic binning,” *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12, pp. 2417–2425, Dec. 2010, doi: 10.1002/asi.21426.
[348] A. Fuschi, Alessandra Merlotti, and D. Remondini, “Correlation measures in metagenomic data: the blessing of dimensionality,” *bioRxiv*, 2024, doi: https://doi.org/10.1101/2024.02.29.582875.
[349] E. Karsenti *et al.*, “A Holistic Approach to Marine Eco-Systems Biology,” *PLOS Biology*, vol. 9, no. 10, p. e1001177, Oct. 2011, doi: 10.1371/journal.pbio.1001177.
[350] R. A. Becker, A. R. Wilks, and R. Brownrigg, “mapdata: Extra Map Databases. R package version 2.3.1.,” 2022, [Online]. Available: https://CRAN.R-project.org/package=mapdata
[351] P. Virtanen, R. Gommers, T. E. Oliphant, and P. van Mulbregt, “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python.,” *Nature Methods*, vol. 17, no. 3, pp. 216–272, doi: 10.1038/s41592-019-0686-2.
[352] J. D. Hunter, “Matplotlib: A 2D Graphics Environment.,” *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007, doi: 10.1109/MCSE.2007.55.
[353] G. Van Rossum and F. L. Drake, “Python 3 Reference Manual,” *CreateSpace*, 2009.
[354] W. Glänzel, “Characteristic scores and scales: A bibliometric analysis of subject characteristics based on long-term citation observation,” *Journal of Informetrics*, vol. 1, no. 1, pp. 92–102, Jan. 2007, doi: 10.1016/j.joi.2006.10.001.
[355] Q. L. Burrell, “Extending Lotkaian informetrics,” *Information Processing & Management*, vol. 44, no. 5, pp. 1794–1807, Sep. 2008, doi: 10.1016/j.ipm.2008.03.002.
[356] C. O. Wilke, “ggridges: Ridgeline Plots in ‘ggplot2’. R package version 0.5.6.,” 2024, [Online]. Available: https://CRAN.R-project.org/package=ggridges
[357] T. Wei and V. Simko, “R package ‘corrplot’: Visualization of a Correlation Matrix (Version 0.92).,” 2021, [Online]. Available: https://github.com/taiyun/corrplot
[358] O. Aumont, C. Ethé, A. Tagliabue, L. Bopp, and M. Gehlen, “PISCES-v2: an ocean biogeochemical model for carbon and ecosystem studies,” *Geoscientific Model Development*, vol. 8, no. 8, pp. 2465–2513, 2015, doi: 10.5194/gmd-8-2465-2015.
[359] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister, “UpSet: Visualization of Intersecting Sets.,” *IEEE Transactions on Visualization and Computer Graphics (InfoVis ’14)*, vol. 20, no. 12, pp. 1983–1992, 2014.

[360] M. Krassowski, “krassowski/complex-upset,” *Zenodo*, 2020, doi: http://doi.org/10.5281/zenodo.3700590.
[361] S. Dray *et al.*, “adespatial: Multivariate Multiscale Spatial Analysis. R package version 0.3-23.”, [Online]. Available: https://CRAN.R-project.org/package=adespatial
[362] J. Fox and S. Weisberg, “An R Companion to Applied Regression.,” 2019, [Online]. Available: https://socialsciences.mcmaster.ca/jfox/Books/Companion/
[363] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, no. null, pp. 2825–2830, Nov. 2011.
[364] D. H. Wolpert, “Stacked generalization,” *Neural Networks*, vol. 5, no. 2, pp. 241–259, Jan. 1992, doi: 10.1016/S0893-6080(05)80023-1.
[365] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Neural Information Processing Systems*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:21889700
[366] E. Sarı and O. Erbaş, “Non-Coding RNA and Functions,” 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:244525108
[367] M. K. Basu, L. Carmel, I. B. Rogozin, and E. V. Koonin, “Evolution of protein domain promiscuity in eukaryotes.,” *Genome research*, vol. 18 3, pp. 449–61, 2008.
[368] Z. Yang *et al.*, “Phylotranscriptomics unveil a Paleoproterozoic-Mesoproterozoic origin and deep relationships of the Viridiplantae,” *Nature Communications*, vol. 14, no. 1, p. 5542, Sep. 2023, doi: 10.1038/s41467-023-41137-5.
[369] J. W. Brown and U. Sorhannus, “A Molecular Genetic Timescale for the Diversification of Autotrophic Stramenopiles (Ochrophyta): Substantive Underestimation of Putative Fossil Ages,” *PLOS ONE*, vol. 5, no. 9, p. e12759, Sep. 2010, doi: 10.1371/journal.pone.0012759.
[370] S. M. Porter, “The fossil record of early eukaryotic diversification,” *The Paleontological Society Papers*, vol. 10, pp. 35–50, 2004, doi: 10.1017/S1089332600002321.
[371] H. Liu *et al.*, “A taxon-rich and genome-scale phylogeny of Opisthokonta,” *PLOS Biology*, vol. 22, no. 9, p. e3002794, Sep. 2024, doi: 10.1371/journal.pbio.3002794.
[372] M. Bury, B. Le Calvé, G. Ferbeyre, V. Blank, and F. Lessard, “New Insights into CDK Regulators: Novel Opportunities for Cancer Therapy,” *Trends in Cell Biology*, vol. 31, no. 5, pp. 331–344, May 2021, doi: 10.1016/j.tcb.2021.01.010.
[373] A. K. Saxena *et al.*, “The essential mosquito-stage P25 and P28 proteins from Plasmodium form tile-like triangular prisms,” *Nature Structural & Molecular Biology*, vol. 13, no. 1, pp. 90–91, Jan. 2006, doi: 10.1038/nsmb1024.
[374] G. Panopoulou *et al.*, “New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes.,” *Genome research*, vol. 13 6A, pp. 1056–66, 2003.
[375] S. Maere *et al.*, “Modeling gene and genome duplications in eukaryotes,” *Proceedings of the National Academy of Sciences*, vol. 102, no. 15, pp. 5454–5459, Apr. 2005, doi: 10.1073/pnas.0501102102.
[376] L.-G. Lundin, “Gene duplications in early metazoan evolution,” *Cell & Developmental Biology*, vol. 10, pp. 523–530, 1999, doi: https://doi.org/10.1006/scdb.1999.0333.
[377] G. N. Somero, “Adaptation of enzymes to temperature: searching for basic ‘strategies,’” *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, vol. 139, no. 3, pp. 321–333, Nov. 2004, doi: 10.1016/j.cbpc.2004.05.003.