



HAL
open science

Étude structurale de la calcyanine, nouvelle protéine impliquée dans la biominéralisation intracellulaire chez les cyanobactéries

Geoffroy Gaschignard

► **To cite this version:**

Geoffroy Gaschignard. Étude structurale de la calcyanine, nouvelle protéine impliquée dans la biominéralisation intracellulaire chez les cyanobactéries. Biochimie, Biologie Moléculaire. Sorbonne Université, 2023. Français. NNT : 2023SORUS272 . tel-04264752

HAL Id: tel-04264752

<https://theses.hal.science/tel-04264752>

Submitted on 30 Oct 2023

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

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



Sorbonne Université

École doctorale Complexité du Vivant – ED 515

Thèse préparée à l'Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie.

(Équipes BIBIP et BIOMIN)

Etude structurale de la calcyanine, une nouvelle protéine impliquée dans la biominéralisation intracellulaire chez les cyanobactéries.

Par Geoffroy Gaschignard

Thèse de doctorat de Biologie Structurale / Bioinformatique

Dirigée par Isabelle Callebaut, Fériel Skouri-Panet et Manuela Dezi

Présentée et soutenue publiquement le 25/09/23

Frédéric MARIN – Directeur de recherche / Université de Bourgogne – Rapporteur

Guillaume LENOIR – Maître de conférences / Université Paris-Saclay – Rapporteur

Anne LOPES – Maîtresse de conférences / Université Paris-Saclay - Examinatrice

Ingrid LAFONTAINE – Professeure / Sorbonne Université - Examinatrice, Présidente

Isabelle CALLEBAUT – Directrice de recherche / Sorbonne Université – Directrice de thèse

Manuela DEZI – Maîtresse de conférences / Sorbonne Université – Co-encadrante de thèse

Membre invité :

Fériel SKOURI-PANET – Ingénieure de recherche / Sorbonne Université – Co-encadrante de thèse

Remerciements :

Je tiens tout d'abord à remercier les rapporteurs, Frédéric Marin et Guillaume Lenoir, pour avoir relu et évalué ce manuscrit, et pour leurs retours constructifs sur ce dernier. Je les remercie aussi, ainsi qu'Ingrid Lafontaine et Anne Lopes, d'avoir accepté d'être les jurys pour la défense de cette thèse. J'ai grandement apprécié la qualité de la discussion qui a suivi la présentation.

Je veux aussi exprimer ma gratitude envers Isabelle Callebaut, Fériel Skouri-Panet et Manuela Dezi pour le grand investissement et la bienveillance dont elles ont fait preuve dans leur encadrement tout au long de cette thèse, ainsi que pour leur volonté de faire de cette thèse un espace d'apprentissage et de développement de mon projet professionnel. Grâce à elles j'ai énormément appris pendant ce projet, tant sur le plan humain que scientifique.

Ce doctorat n'aurait pu avoir lieu sans l'apport financier, humain et scientifique de l'ANR Harley. Je remercie ses membres, et plus spécifiquement son porteur, Karim Benzerara, pour m'avoir fait confiance, et m'avoir accompagné scientifiquement tout au long de cette thèse.

Je tiens à adresser des remerciements spéciaux à Catherine Berthomieu, Nicolas Bremond, Morgane Recuerda et Edern Pamart pour m'avoir très bien accueilli au BIAM à Cadarache, et pour leurs apports scientifiques. Cela a été l'occasion de découvrir un autre environnement de travail et de me former sur de nouvelles expériences, ce qui a été précieux dans ma formation.

Merci aux équipes BIOMIN et BIBIP de l'IMPMC, qui m'ont eu en garde partagée. Durant ces 3 ans et demi en leur compagnie, j'ai toujours reçu l'aide et le soutien dont j'avais besoin.

Un immense merci à Maxime Millet de m'avoir expliqué comment fonctionnait un ordinateur, et pour son flux continu de critiques non-constructives sur mes expériences. J'ai adoré avoir un co-bureau avec qui je puisse travailler et exposer mes idées, des plus stupides aux plus intelligentes, sans me faire (trop) juger, tout en pouvant rire.

Merci à Elodie Duprat pour sa gentillesse et surtout son expertise scientifique qui nous a permis d'aboutir à mon premier article en tant que premier auteur !

Merci à Cynthia Travert pour sa gentillesse et son aide pour toutes les expériences que j'ai faites. Sans elle, ce projet serait resté bloqué à l'étape « Trouver son chemin dans le laboratoire ».

Merci à Slavica Jonic de m'avoir formé à l'analyse d'image de Cryo-EM, je n'ai pas eu l'occasion de me servir de cette compétence dans ce projet, mais je ne désespère pas de le faire un jour !

Merci à Stéphanie Finet d'être restée (plusieurs fois) jusqu'à 3h du matin un samedi soir au Synchrotron SOLEIL pour me permettre d'étudier mes échantillons au SAXS.

Merci à Neha Mehta pour ses conseils avisés sur le monde de la recherche et pour son investissement pour faire vivre la vie scientifique du laboratoire. Par ailleurs j'ai énormément apprécié sa présence pour me guider et me donner confiance lors de ma première conférence scientifique internationale.

Merci à Kamel Bouazza, Helin Elhan et Si Min Montagnie d'être venus en renfort pour cette thèse lors de leur stage respectif, et de m'avoir permis de satisfaire mes pulsions dictatoriales.

Merci à Catherine Vénien-Bryan, Gabriel Brandt et à tout les autres membres du couloir 22-23 avec qui j'ai pu discuter entre deux expériences.

Merci enfin à tout le personnel administratif pour leur implication et pour leur bienveillance même quand je fournissais mes ordres de mission la veille de mon départ.

Je remercie aussi tous les partenaires qui ont pu prendre part à ce projet. Notamment je remercie chaleureusement Julien Henri pour son expertise scientifique sur la partie cristallisation de ce projet, mais aussi pour avoir été un excellent tuteur de thèse, toujours prêt à m'écouter avec bienveillance sur mes questionnements scientifiques ou professionnelles.

Merci aussi à Aurélie Di Cicco pour m'avoir formé à la microscopie électronique, à Christophe Marchand pour toutes ses analyses de spectrométrie de masse qui nous ont permis d'éviter de ré-étudier une protéine déjà connue et Benjamin Bailleul pour ses remarques constructives au cours des différents comités de suivi de thèse.

Je tiens de même à remercier la légion des doctorants/post-docs de l'IMPMC qui ont assuré un super soutien moral à travers le Covid, les expériences qui plantent et les périodes de stress.

Je voudrais vraiment exprimer ma plus grande gratitude à Juliette Gaëtan qui a toujours été présente pour discuter et me soutenir. Sans elle et sa capacité d'écoute, ma thèse n'aurait pas été la même, et j'aurais eu beaucoup moins de plaisir à venir au laboratoire chaque jour.

Je voudrais dire aussi merci à : Dania Marisol Zuniga pour avoir animé tout le couloir 22-23, pour sa bonne humeur dès le matin et pour m'avoir apporté beaucoup d'excellents conseils scientifiques;

Juliette Debrie pour ses très nombreuses critiques pas constructives sur ma thèse, mes choix vestimentaires et ma vie en générale;

Jeanne Caumartin pour sa gentillesse au quotidien;

Apolline Bruley, pour m'avoir appris à naviguer les eaux incertaines de la bio-informatique et de l'administration de l'école doctorale;

Amandine Hecquet pour son sourire, pour ses supers gâteaux et pour m'avoir fait découvrir l'IEES;

Divine Vangu pour son investissement pour les autres, et pour ses incroyables pâtisseries;

Julie Aufort pour ses discours de motivation qui m'ont beaucoup aidé à écrire ce manuscrit;

Rémi Vuillemot pour son sourire, et pour m'avoir débloqué mon ordinateur quand je pensais l'avoir définitivement verrouillé et perdu 2 ans de thèse;

Romain Bolzonni pour m'avoir accueilli à Cadarache, et pour son stock inépuisable de nourriture;

Anne-Elisabeth Marceline, Quentin Bollaert, Martin Demoucron, Ramón Messias, Öykü Ataytür, Octave Duros pour avoir toujours été présents et plein de bonnes humeurs pour les déjeuners, les gouters ou allers prendre des bières au Baker.

Merci enfin à Cécile Bidaud, Laura Galezowski, Andreas Zoumpoulakis, Carlos Fernandes, Rafael Veloso, Baptiste Truffet, Romain Taureau, Léon Ambriambarijaona, Mohamad Harastani, Aashini Rajpal et à tous les autres membres du laboratoire présents ou passés que je pourrais avoir oublié.

Je voudrais ensuite prendre le temps de remercier tout les gens qui n'ont pas directement pris part à cette thèse, mais qui m'ont permis de maintenir un bon équilibre vie personnelle - vie professionnelle.

A ce titre je voudrais dire merci à tous mes amis de l'ESPCI, de la prépa Clémenceau, des mes années de collège-lycée et d'ailleurs, qui, parfois en dépit de la distance ou des années, sont toujours là quand j'ai besoin d'eux. Je voudrais adresser des remerciements spéciaux à Emile qui me supporte aux quotidiens et qui m'écoute râler (et aussi qui fait le ménage de l'appartement quand j'oublie). De même je voudrais dire merci à Hugo qui a eu le courage héroïque de me soutenir durant toutes ma rédaction et la préparation de ma soutenance, à travers mes coups de stress et mes pertes de motivation.

Enfin, et parce que c'est grâce à leur soutien depuis ma naissance que j'ai pu faire des études puis une thèse, je remercie du fond de mon cœur mes parents qui ont toujours été là pour moi. Je remercie aussi mon frère Amaury et mes sœurs, Margaux, qui m'aura volé le titre de premier docteur de la fratrie et Camille. J'espère qu'avec ma défense ils auront enfin compris quel était mon sujet de thèse.

Table des matières

| | |
|---|-----------|
| Liste des abréviations..... | 9 |
| Introduction..... | 11 |
| La biominéralisation est un phénomène répandu dans le vivant..... | 11 |
| La biominéralisation du carbonate de calcium..... | 16 |
| Les protéines impliquées dans la biominéralisation contrôlée du CaCO ₃ | 18 |
| <i>Transport des ions Ca²⁺</i> | 20 |
| <i>Transport des ions HCO₃⁻</i> | 21 |
| <i>Transport des ions H⁺ et contrôle du pH</i> | 21 |
| <i>Autre transports d'ions (Na⁺, K⁺, Cl⁻)</i> | 22 |
| <i>Initiation de la formation du biominéral</i> | 22 |
| <i>Sélection du polymorphe et de la cristallinité</i> | 23 |
| <i>Inhibition de la précipitation</i> | 23 |
| <i>Contrôle de la forme du biominéral</i> | 24 |
| <i>Importance des protéines acides et des LCD (Low Complexity Domain)</i> | 25 |
| <i>L'inconnue des protéines de la biominéralisation</i> | 25 |
| La biominéralisation intracellulaire d'ACC par les cyanobactéries..... | 26 |
| <i>Découverte de la calcyanine</i> | 29 |
| Projet de thèse..... | 35 |
| Chapitre 1 : Analyse de la modélisation de la structure 3D de la calcyanine de <i>S. calcipolaris</i> | 39 |
| <i>Les modèles de la calcyanine complète de <i>S. calcipolaris</i> produits par AlphaFold2 et ESMFold, appuient nos premières modélisations par homologie et donnent un premier aperçu de l'ensemble de la structure 3D de la protéine, mais demandent à être complétés</i> | 39 |
| <i>Le domaine CoBaHMA est une nouvelle famille dans la superfamille HMA, au sein du repliement ferredoxin-like et présente une signature structurale et fonctionnelle unique</i> | 45 |
| <i>Les GlyZips forment des hélices antiparallèles en épingle à cheveux, avec une face hydrophobe et une face hydrophile. Leurs orientations relatives au sein du domaine (GlyZip)₃ restent inconnues</i> | 51 |
| <i>L'expérience est indispensable pour accéder à la structure de l'ensemble de la calcyanine de <i>S. calcipolaris</i></i> | 56 |
| Chapitre 2 : Étude expérimentale de la calcyanine de <i>S. calcipolaris</i> | 59 |
| <i>Le taux d'expression et le rendement de purification de la calcyanine de <i>S. calcipolaris</i> chez <i>E. coli</i> sont très faibles</i> | 59 |
| <i>La calcyanine purifiée forme des homo-oligomères de haut poids moléculaires très hétérogènes</i> | 61 |

| | |
|---|------------|
| La protéolyse limitée a permis d'identifier un fragment stable et homogène, exploitable pour la cristallogénèse..... | 65 |
| Chapitre 3 : Les protéines à domaine CoBaHMA..... | 79 |
| Le domaine CoBaHMA est présent sur d'autres architectures protéiques, principalement associé à des domaines membranaires..... | 79 |
| Conclusion et Discussion..... | 87 |
| Perspectives..... | 95 |
| Matériels et Méthodes..... | 101 |
| 1. Bioinformatique..... | 101 |
| 1.1. Pymol, Chimera : visualisation, hydrophobicité, RMSD..... | 101 |
| 1.2. Modélisation..... | 101 |
| 1.3. Outils d'évaluation des structures 3D..... | 103 |
| 1.4. FoldSeek..... | 104 |
| 1.5. Téléchargement de structure 3D..... | 104 |
| 1.6. Alignements de séquences..... | 105 |
| 1.7. Prédiction de l'agrégation..... | 105 |
| 2. Experimental..... | 106 |
| 2.1. Différentes constructions de calcyanine étudiées..... | 106 |
| 2.2. Gène et expression de la calcyanine..... | 108 |
| 2.3. Lyse cellulaire..... | 109 |
| 2.4. Purification des constructions de calcyanine..... | 110 |
| 2.5. Expression et purification de la protéase TEV étiquetée 6His..... | 114 |
| 2.6. Protéolyse limitée..... | 114 |
| 2.7. Spectroscopie de masse..... | 115 |
| 2.8. Small Angle X-rays Scattering : expériences et analyses..... | 116 |
| 2.9. Essais de cristallogénèse..... | 117 |
| Références..... | 119 |
| Annexes..... | 152 |

Liste des abréviations :

a.a : acides aminés
ACC : Amorphous Calcium Carbonate
cryo-EM : cryo-Electron Microscopy
C_{ter} : C-terminal
EDTA : Ethylenediaminetetraacetic acid
EPS : ExoPolySaccharide
HCA : Hydrophobic Cluster Analysis
HMA : Heavy Metal Associated
HMM : Hidden Markov Model
iACC : intracellular Amorphous Calcium Carbonate
iHMA : integrated Heavy Metal Associated
IPTG : Isopropyl β -D-1-thiogalactopyranoside
LCD : Low Complexity Domain
MALS : Multiple Angle Light Scattering
MEB : Microscope Electronique à Balayage
MET : Microscope Electronique à Transmission
MFS : Major Facilitator Transporter
MS : Spectroscopie de Masse
MSA : Multiple Sequence Alignment
N_{ter} : N-terminal
PA : Phosphatidic Acid
PAE : Predicted Aligned Error
PCC : Pasteur Culture collection of Cyanobacteria
PDB : Protein Database
PEEM : PhotoEmission Electron Microscope
pET : plasmid for expression by T7 RNA polymerase
PG : Phosphatidylglycerol
pLDDT : predicted Local Distance Difference Test
RMN: Résonance Magnétique Nucléaire
RMSD : Root-Mean-Square-Deviation
SAXS : Small Angle X-Ray Scattering
SEC : Size Exclusion Chromatography
SQDG : Sulfoquinovosyldiacylglycerol
TEV : Tobacco Etched Virus

Introduction :

La biominéralisation est un phénomène répandu dans le vivant.

Les minéraux et les êtres vivants ont coévolué (Hazen et al., 2008) depuis au moins 3,4 milliards d'années (Javaux, 2019). Cette connexion très ancienne a mené à un haut degré d'interaction entre eux. L'un des exemples les plus frappants de ce type d'interaction est la biominéralisation, qui est souvent définie comme l'ensemble de processus qui mènent à la formation de minéraux, appelés biominéraux, par des êtres vivants (Görge et al., 2021; Skinner, 2005). L'International Mineralogical Association définit le minéral de la façon suivante : « *a mineral is an element or chemical compound that is normally crystalline and which has been formed as a result of geological processes.* ». Cette définition s'accompagne d'une exclusion stricte des produits formés par des êtres vivants sans intervention d'un processus géologique (Nickel, 1995). De plus, les solides amorphes comme les ACC (Amorphous Calcium Carbonate) (Martignier et al., 2017), ou les cristaux organiques, comme l'hemozoïne (cristal d'hème) (Matz et al., 2020), sont exclus de cette définition, alors qu'ils sont étudiés dans le cadre de la biominéralisation. Pour ces raisons, la définition d'un biominéral est généralement plus large que celle d'un minéral et inclut tous les composés cristallins, organiques ou inorganiques, ainsi que les précipités amorphes inorganiques. De plus, les biominéraux peuvent être un mélange d'éléments organiques (protéines, polysaccharides ...) et inorganiques, comme la coquille en calcite des mollusques qui contient une matrice organique (Marin, 2020).

La biominéralisation est très répandue dans le vivant. On trouve des biominéraux chez les eucaryotes, comme par exemple l'hydroxyapatite ($\text{Ca}_5(\text{PO}_4)_3(\text{OH})$) des os et des dents des vertébrés (Sharma et al., 2021), l'aragonite (CaCO_3) du squelette des coraux (Figure 1.D; C. A. Schmidt et al., 2022) ou encore la silice (SiO_2) que l'on peut trouver dans les parois cellulaires chez certaines plantes (Figure 1.C; Bauer et al., 2011; Zexer & Elbaum, 2022). De même, la biominéralisation est aussi répandue chez les procaryotes (Cosmidis & Benzerara, 2022; Görge et al., 2021), comme chez les bactéries magnétotactiques (Figure 1.A; Blakemore, 1975) qui forment des magnétosomes, des organelles constituées de la magnétite (Fe_3O_4) ou de la greigite (Fe_3S_4) (Faivre & Schüler, 2008), ou certaines espèces de bactéries qui stockent du soufre élémentaire (Nims et al., 2019).

La biominéralisation peut répondre à de nombreux impératifs biologiques. Certains d'entre eux sont bien connus, comme la structuration et la résistance aux chocs apportés par le squelette des vertébrés ou la protection fournie par la coquille des mollusques et des coquillages, mais il en existe des plus exotiques. Les bactéries magnétotactiques s'alignent dans le champ magnétique terrestre grâce à leurs magnétosomes, ce qui leur permet de s'orienter dans l'espace (Blakemore, 1975). La calcite de l'exosquelette des fourmis coupe-feuille aurait des propriétés antifongiques (H. Li et al., 2020). Enfin le parasite *Plasmodium falciparum* transforme l'hème libre issu de la digestion de l'hémoglobine humaine, qui lui est extrêmement toxique, en hemozoïne (cristal d'hème) qui ne lui est pas toxique (Matz et al., 2020).

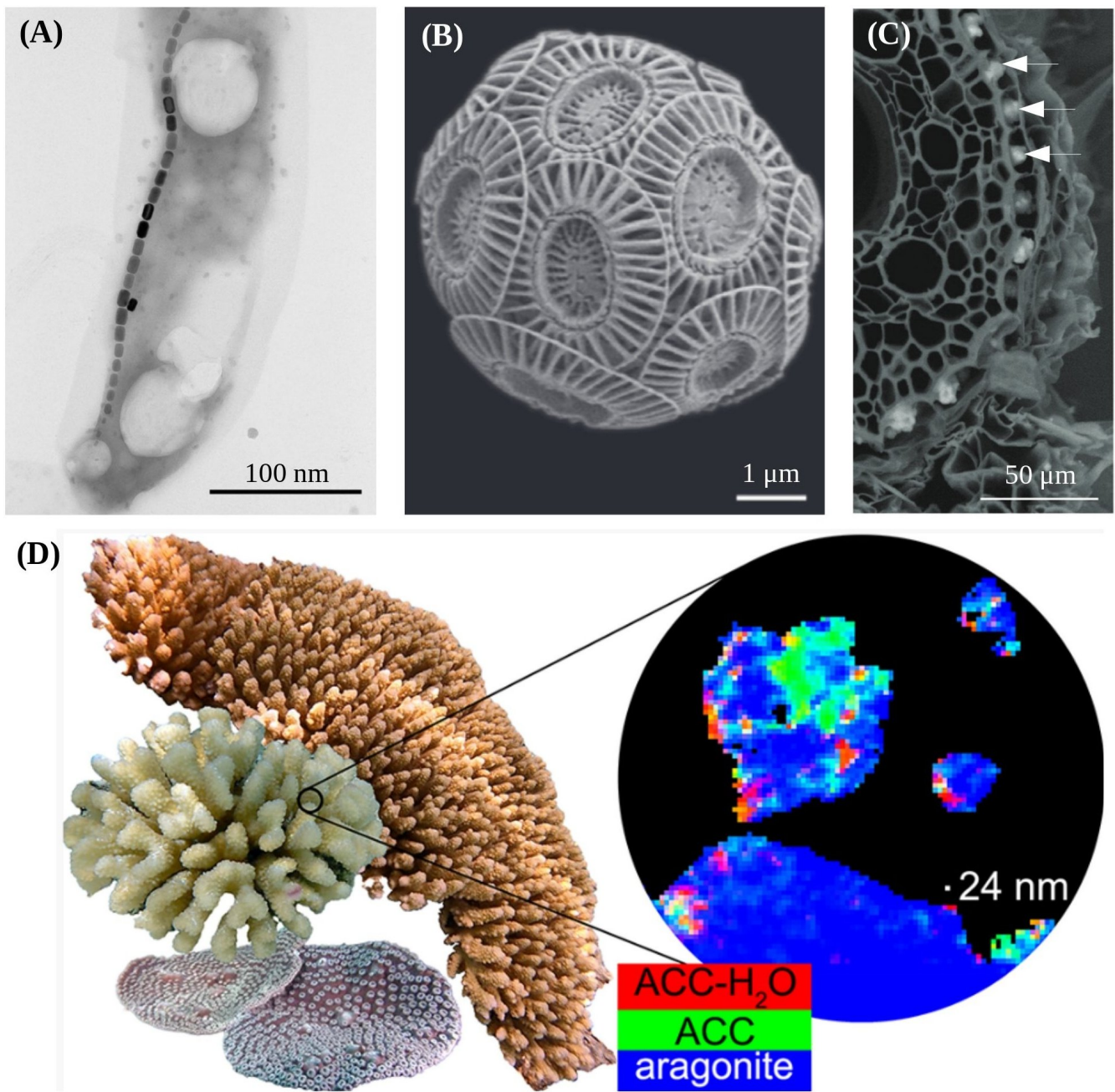


Figure 1 : Exemples de biominéralisation.

Photos en microscopie électronique (A) d'une bactérie magnétotactique, du genre *Vibrio*, et de sa chaîne de magnétosomes (en noir) (Favre & Schüler, 2008); (B) de l'algue coccolithophore *Emiliana huxleyi* (Skeffington & Scheffel, 2018); (C) d'agrégats de silice (indiqués avec des flèches blanches) dans les racines du Sorgho *Sorghum bicolor* L. (Zexer & Elbaum, 2022). (D) Image de squelettes de coraux (C. A. Schmidt *et al.*, 2022) avec, en vignette, une carte de composition des minéraux établie grâce à un PEEM (PhotoEmission Electron Microscope).

Toutefois, si la biominéralisation peut être bénéfique pour les organismes impliqués, elle peut aussi avoir des effets délétères. Chez l'être humain, les calculs rénaux (Sivaguru et al., 2021) ou la calcification des tendons (Sansone et al., 2018) sont 2 pathologies dues à la formation de biominéraux. De la même manière, les micro-organismes qui biominéralisent extracellulairement peuvent se retrouver piégés dans leurs propres biominéraux (Cosmidis & Benzerara, 2022).

Comme le montrent ces exemples, la grande diversité d'organismes qui biominéralisent s'accompagne d'une grande diversité d'espèces de biominéraux formés : des carbonates, des phosphates, des sulfures... Cette diversité de minéraux est illustrée dans la table 1 de l'article de Skinner, reprise ici en Figure 2, qui répertorie un grand nombre de biominéraux inorganiques connus (Skinner, 2005). Ce tableau n'inclut pas les biominéraux constitués de cristaux organiques comme l'hemozoïne déjà mentionnée (Matz et al., 2020), qui étendent encore cette diversité.

La biominéralisation est généralement scindée en 2 sous catégories, selon que le processus est contrôlé par l'être vivant ou non (Dupraz et al., 2009; Görgen et al., 2021; Weiner, 2003).

Dans le premier cas, la biominéralisation est appelée contrôlée (Figure 3.C). Elle fait intervenir des processus biologiques spécifiques impliquant des protéines, des tissus, des organelles... C'est un processus qui a donc un coût énergétique, ne serait-ce que pour transporter les ions, souvent contre leur gradient, jusqu'au site de minéralisation pour pouvoir former le biominéral même dans un environnement sous-saturé (Gilbert et al., 2022).

Les exemples non-pathologiques de biominéralisation chez les eucaryotes que nous avons cités, ainsi que les bactéries magnétotactiques sont des exemples de biominéralisation contrôlée.

Dans le 2nd cas, la biominéralisation n'est pas contrôlée par l'organisme, mais est la conséquence de son activité dans un environnement qui est sursaturé en phase minérale. Ce cas là est parfois lui-même subdivisé en 2 cas. Dans le cas où l'organisme, ou sa matrice extracellulaire, n'est qu'un site de nucléation pour le minéral, la biominéralisation est dite influencée (Figure 3.A). Ce type de minéralisation peut alors avoir lieu sur un organisme biologiquement inactif, voire mort.

La précipitation du CaCO_3 sur la matrice extracellulaire des biofilms de cyanobactéries en est un exemple (Arp et al., 1999).

Dans le cas où la précipitation est un produit secondaire du métabolisme de l'organisme, la biominéralisation est dite induite (Figure 3.B). C'est le cas des bactéries sulfato-réductrices, qui, comme leur nom l'indique, réduisent le sulfate (SO_4^{2-}) de leur environnement en sulfure (H_2S), qui peut alors réagir avec des métaux comme le Fer pour former un précipité de sulfure de métal (Barton & Tomei, 1995).

L'étude de la biominéralisation a donc des impacts dans de nombreux domaines, tels que l'environnement ou la médecine.

TABLE 1. Biominerals*.

| Class | Chemical designation | Example |
|-----------|------------------------|---|
| 1 | Native elements/alloys | S |
| 2,3 | Sulphides/arsenides | FeS ₂ pyrite, marcasite FeFe ₂ S ₄ , greigite (Fe,Ni)S _{0.9} , mackinawite Fe _(1-x) S, pyrrhotite where $x = 0-0.2$ FeS·nH ₂ O, hydrotroilite ZnS, sphalerite, wurtzite PbS, galena Ag ₂ S, acanthite As ₂ S ₃ , orpiment |
| 4-8 | Oxides/hydroxides | FeFe ₂ O ₄ , magnetite Fe ₂ O ₃ , amorphous Fe ³⁺ TiO ₂ , ilmenite α -FeOOH, goethite γ -FeOOH, lepidocrocite 5Fe ₂ O ₃ ·9H ₂ O, ferrihydrite Mn ₃ O ₄ , amorphous (Mn ²⁺ Ca,Mg) Mn ⁴⁺ ₂ ·H ₂ O, todorokite Na ₂ Mn ₂ O ₂ ·9H ₂ O, birnesite |
| 9-12 | Halogenides | CaF ₂ , fluorite K ₂ SiF ₆ , hieratite Cu ₂ Cl(OH) ₃ , atacamite |
| 13-17 | Carbonates | CaCO ₃ , calcite, aragonite, vaterite, amorphous CaCO ₃ ·H ₂ O, monohydrocalcite (Ca _(1-x) Mg _x)CO ₃ , magnesian calcite (Ca,Mg)CO ₃ , 'protodolomite' Pb ₃ (CO ₃) ₂ (OH) ₂ , hydrocerrusite |
| 28-32 | Sulphates | CaSO ₄ ·2H ₂ O, gypsum KFe ³⁺ (SO ₄) ₂ (OH) ₆ , jarosite SrSO ₄ , celestite CaSO ₄ , baryte |
| 37-43 | Phosphate/arsenates | Ca ₅ (PO ₄) ₃ (OH,F,Cl), calcium apatite group: (Ca,X) ₁₀ (PO ₄ ,CO ₃) ₆ (OH,CO ₃) ₂ , 'bioapatite' X = cations CaHPO ₄ ·2H ₂ O, brushite Ca ₈ H ₂ (PO ₄) ₆ ·6H ₂ O, octacalcium phosphate Ca ₁₈ H ₂ (Mg ²⁺ , Fe ²⁺)(PO ₄) ₁₄ , whitlockite Fe ²⁺ (PO ₄) ₂ ·8H ₂ O, vivianite Mg(NH ₄)(PO ₄)·6H ₂ O, struvite |
| Silicates | | SiO ₂ ·nH ₂ O opal, non-crystalline or disordered cristobalite, tridymite |

* See Lowenstam and Weiner (1989) Table 2.1 pp. 8-15 for a listing of the phyla where these mineral species have been found.

Figure 2 : Liste de biominéraux connus

Extrait de Skinner, 2005.

La 1^{ère} colonne indique la/les classe(s) des minéraux de la ligne telle(s) que définie(s) dans *Dana's new mineralogy* (Dana *et al.*, 1997).

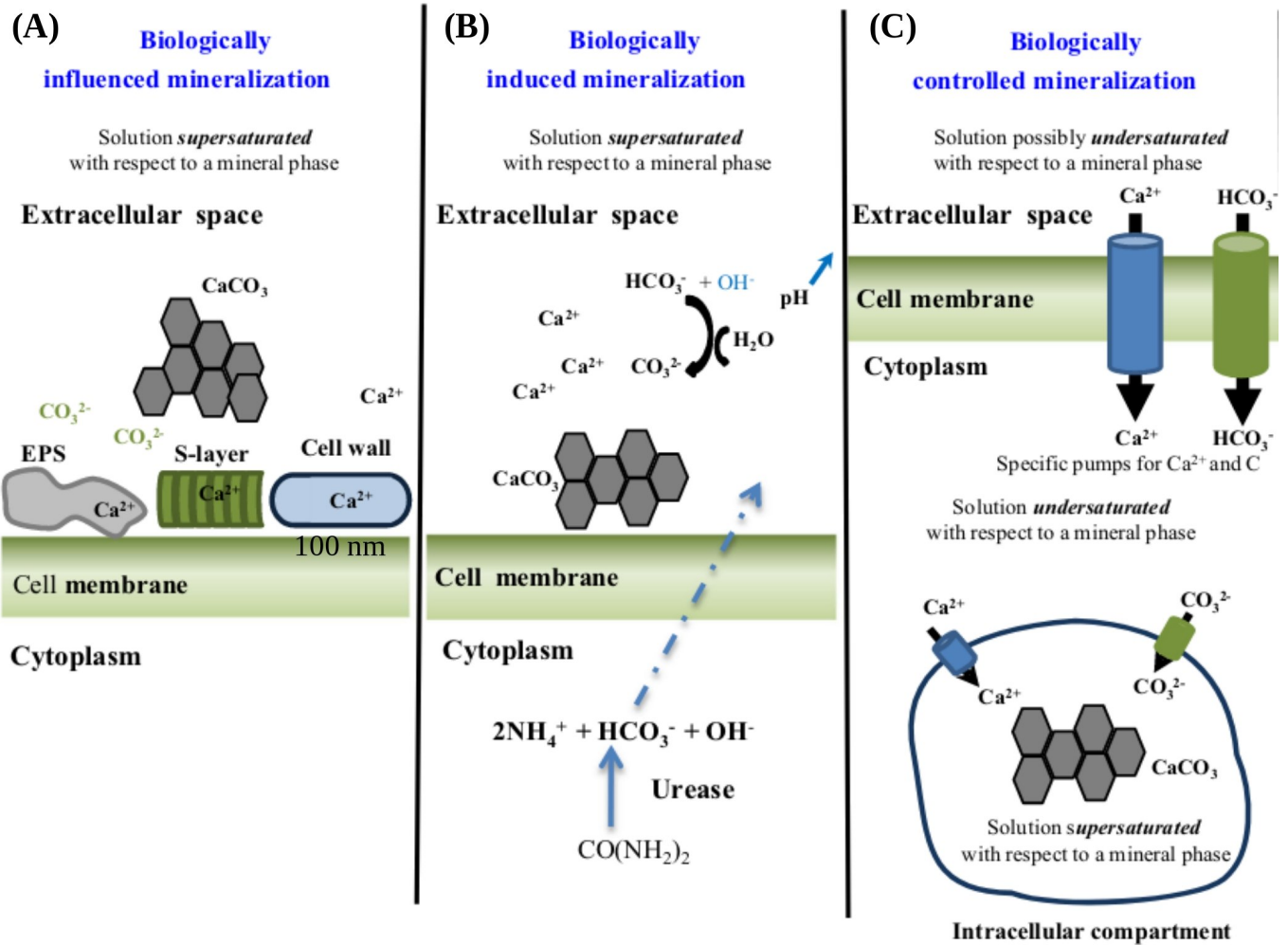


Figure 3 : Classification de la biominéralisation.

Extrait de Görden *et al.*, 2021.

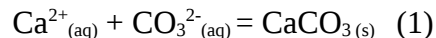
Schéma récapitulatif des catégories de biominéralisation, illustrées dans le cas du carbonate de calcium. (A) Biominéralisation influencée. (B) Biominéralisation induite. (C) Biominéralisation contrôlée.

La biominéralisation du carbonate de calcium.

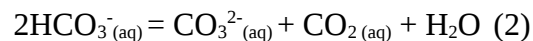
Le carbonate de calcium est un minéral qui est retrouvé en abondance dans l'environnement, principalement sous forme de calcaire ou de craie. Il peut se trouver naturellement sous 6 formes : 3 formes cristallines (aragonite, vaterite, calcite), 2 formes hydratées métastables (monohydrocalcite et ikaite), et sous forme amorphe (ACC) (Sekkal & Zaoui, 2013). Les ACC sont eux-mêmes subdivisés en 2 catégories selon qu'ils soient hydratés ou non (Innocenti Malini et al., 2017). Les ACC sont métastables thermodynamiquement. Ils cristallisent à forte température (Koga et al., 1998) ou à forte pression (Yoshino et al., 2012). Dans le cas de la biominéralisation, les ACC sont souvent des précurseurs de phase de CaCO₃ cristallisée (Gilbert et al., 2022).

Le calcium est un alcalino-terreux (2nde colonne du tableau périodique) qui forme donc des cations divalents, Ca²⁺, en solution. Le CaCO₃ peut accepter des substitutions du calcium par d'autres alcalino-terreux comme le magnésium, le strontium, le baryum ou le radium. Ces substitutions sont polymorphes-dépendantes : la calcite s'accommode bien des ions Mg²⁺ mais très mal du Sr²⁺ et du Ba²⁺, au contraire de l'aragonite, alors que les ACC peuvent intégrer tous les alcalino-terreux grâce à leur absence de structure cristalline. Il dépend aussi du processus de formation du CaCO₃ : la calcite cristallisée à partir d'un ACC dopé au strontium ou au barium est capable d'accommoder bien plus de strontium et de barium qu'une calcite qui a cristallisé sans intermédiaire (A. Saito et al., 2020).

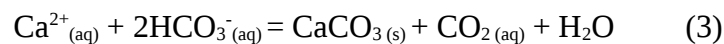
Le CaCO₃ précipite dans des conditions de sur-saturation, locale ou globale, selon l'équation :



Le carbonate CO₃²⁻ est la base conjuguée du bicarbonate HCO₃⁻ (pKa = 10,32) qui est lui-même la base conjuguée du CO₂ (pKa = 6,37). L'océan, où se fait l'essentiel de la production de CaCO₃ (Morse et al., 2007), a un pH compris entre 8 et 8,25 (L.-Q. Jiang et al., 2019). Dans la majeure partie des cas, il faut donc aussi prendre en compte la réaction de dissociation du bicarbonate :



Ce qui donne la réaction la plus courante pour la précipitation du CaCO₃ :



Un grand nombre d'eucaryotes dont les coraux (Figure 1.D), les oursins, les poules, les étoiles de mers ou encore les coccolithophores (Figure 1.B) biominéralisent le CaCO₃ (Gilbert et al., 2022). De même, c'est une occurrence commune chez les procaryotes (Görgen et al., 2021), qui sont notamment responsables de la formation de larges macro-dépôts de CaCO₃ comme certains stromatolithes (Saghai et al., 2016).

La biominéralisation du carbonate de calcium fait l'objet d'une attention toute particulière de la part de la communauté scientifique pour plusieurs aspects.

Des études s'intéressent à l'impact de l'acidification des océans consécutive à l'augmentation du CO₂ dans l'atmosphère due à l'activité humaine (L.-Q. Jiang et al., 2019), sur les capacités de

biominéralisation des organismes marins, comme le corail (C. A. Schmidt et al., 2022) ou les coccolithophores (Meyer & Riebesell, 2015). Cela pourrait modifier durablement les cycles du carbone et du calcium sur lesquels ces organismes agissent (Riebesell et al., 2009; Zavarzin, 2002). Si ces organismes ont de fortes chances d'être durablement affectés par le changement climatique, ils pourraient aussi amener des solutions pour lutter contre celui-ci. Des organismes comme l'iroko (un arbre originaire d'Afrique de l'Ouest) ou les cyanobactéries sont étudiés pour piéger de vaste quantité de CO₂ sous forme de CaCO₃ (Gwenzi, 2019). De même des stratégies de stockage du carbone s'inspirent de la biominéralisation, notamment en utilisant l'anhydrase carbonique, une protéine qui catalyse la réaction :



et qui est considérée comme centrale dans les processus de biominéralisation (Gilbert et al., 2022), comme catalyseur pour accélérer la précipitation du CO₂ sous forme de carbonates (Gwenzi, 2019).

Les organismes biominéralisants pourraient aussi être intéressants pour lutter contre d'autres formes de pollution, en piégeant dans leur biominéraux des radio-éléments ou des métaux toxiques (Gadd & Pan, 2016; Mehta et al., 2019).

Par ailleurs, alors que les parties organiques d'un organisme se dégradent rapidement, les biominéraux sont très présents dans le registre fossile et constituent de précieux indices pour l'étude des formes ancestrales de vie. Par exemple, les microbialithes, des roches formées par l'action de micro-organismes, sont considérés comme étant les plus anciennes traces de vie sur Terre (Wacey, 2012). De plus, les espèces qui biominéralisent sont plus représentées dans les dépôts géologiques que les espèces qui ne forment pas de biominéraux (J. Li et al., 2013).

Les protéines impliquées dans la biominéralisation contrôlée du CaCO₃.

La biominéralisation contrôlée du CaCO₃ chez les procaryotes est pour l'heure assez mal comprise, principalement faute d'exemples. Il n'y en a à ce jour que 3 de répertoriés : chez les gamma-protéobactéries *Achromatium* (Figure 4.A; Benzerara et al., 2021), chez certaines cyanobactéries (Couradeau et al., 2012) dont il sera sujet plus tard dans cette introduction, et chez certaines alphaproteo-bactéries magnétotactiques (Figure 4.B; Monteil et al., 2021). Signe, de la méconnaissance de ce type de biominéralisation, la nature exacte du biominéral formé par *Achromatium*, calcite ou ACC, ainsi que sa localisation, cytoplasmique ou periplasmique, faisaient encore objet de débats en 2021 (Benzerara et al., 2021), près d'un siècle et demi après la première description de ce phénotype en 1893 (Schewiakoff, 1893).

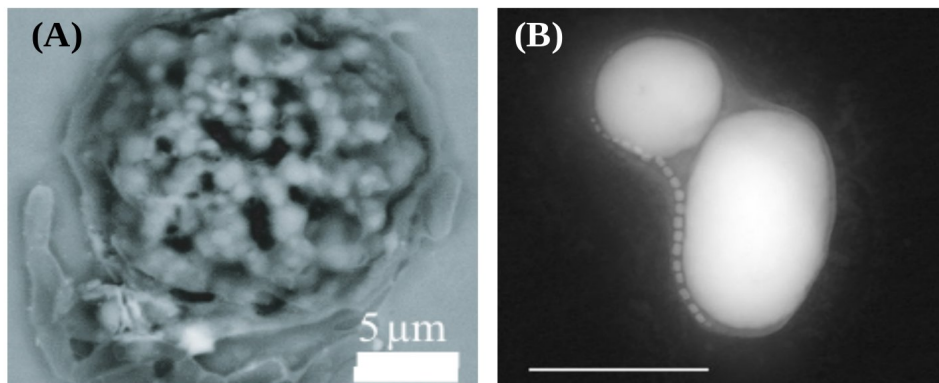


Figure 4 : Biominéralisation contrôlée du CaCO₃ chez les procaryotes.

(A) Cellule d'*Achromatium* du lac Pavin, observée au MEB (Microscope Electronique à Balayage) (Benzerara *et al.*, 2021). (B) Image MET (Microscope Electronique à Transmission) d'une bactérie magnétotactique du lac Pavin. Les larges inclusions blanches sont des ACC. Les petites inclusions en chaîne sont des magnétosomes (Monteil *et al.*, 2021).

A l'inverse, la biominéralisation contrôlée du CaCO₃ chez les eucaryotes est bien mieux comprise et à même fait l'objet d'un modèle intégré par Gilbert *et al.* en 2022 pour les organismes marins. Pour résumer très succinctement ce modèle, chez ces organismes biominéralisants, le CaCO₃ se trouve dans un compartiment très contrôlé biologiquement, nommé « espace privilégié ». Ce compartiment est isolé de l'environnement extérieur par des cellules épithéliales ou des phospholipides selon la taille de l'espace et/ou les organismes. Le CaCO₃ est tout d'abord précipité sous forme d'ACC dans des vésicules dont la concentration en Ca²⁺, en CO₃²⁻ et le pH sont augmentés jusqu'à saturation. Les ACC sont relargués dans l'espace privilégié puis s'agglomèrent ensuite sur le biominéral déjà formé puis cristallisent, si le biominéral est un cristal. Les espaces libres entre les ACC sont, en parallèle, comblés par croissance du cristal par attachement ionique (Gilbert et al., 2022).

Ces études menées chez les eucaryotes ont pu montrer l'importance des protéines dans ces mécanismes de biominéralisation. Il serait trop long d'en faire une liste exhaustive, mais il est pertinent de décrire en partie la diversité des fonctions qui peuvent être impliquées dans la formation d'un biominéral, afin de nourrir des hypothèses pour la calcanine, la protéine dont il sera sujet dans cette thèse. Les protéines qui seront abordées ici, sont schématisées sur la Figure 5.

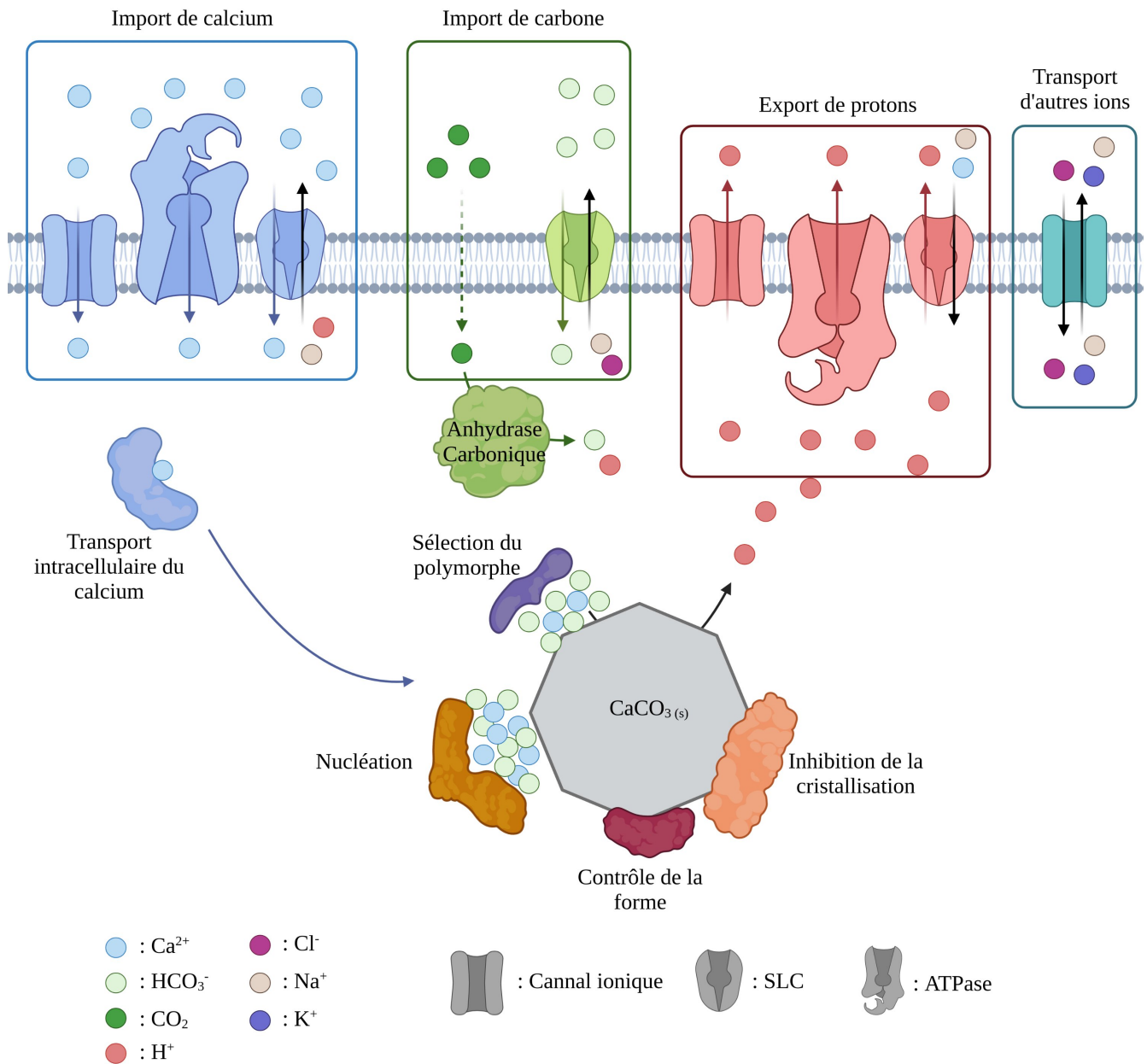


Figure 5 : Les protéines de la biominéralisation du CaCO_3 .

Schéma de quelques classes de protéines qui ont été identifiées comme intervenant dans des biominéralisations contrôlées, et qui sont décrites dans cette thèse. La forme ainsi que la taille des protéines représentées ne sont pas représentatives de la réalité. L'espace contenant le biominéral est un compartiment. Figure produite avec www.BioRender.com

Tout d'abord, pour pouvoir déclencher la précipitation du CaCO_3 il faut atteindre la super-saturation au niveau du site de biominéralisation, en important ou exportant les ions nécessaires au biominéral. Ces ions sont généralement transportés à travers les membranes par des protéines de transports. Il en existe 4 grandes classes :

- _ Les protéines ABC (ATP Binding Cassette) qui hydrolysent l'ATP pour transporter leur substrat à travers la membrane;
- _ Les ATPases qui sont subdivisées en 3 types : F-, V- et P-. Les F-types utilisent le gradient des ions pour générer de l'ATP, quand les V- et P-types ATPases hydrolysent l'ATP pour transporter leur substrat à travers la membrane en effectuant d'important changements conformationnels;
- _ Les SLC (SoLute Carrier) qui transportent un grand nombre de solutés (ions, vitamines, acides aminés...) à travers la membrane soit selon leur gradient, soit contre leur gradient mais en s'appuyant dans ce cas sur le transport d'un autre ion selon son gradient (symport ou antiport). Ces protéines n'utilisent pas l'hydrolyse de l'ATP comme source d'énergie pour le transport;
- _ Les canaux ioniques qui sont des transporteurs passifs fournissant un passage sélectif aux ions (Pizzagalli et al., 2021).

Transport des ions Ca^{2+} .

Comme les ions Ca^{2+} sont essentiels à un grand nombre de processus biologiques (Remick & Helmann, 2023), il est parfois difficile de distinguer les transporteurs du Ca^{2+} impliqués dans la formation du biominéral de ceux qui sont utiles pour d'autre processus. Par exemple, chez le coccolithophore *Emiliana huxleyi* (visible en Figure 1.B), la Ca^{2+} -ATPase ECA2 est sur-exprimée dans les cellules diploïdes biominéralisantes par rapport aux diploïdes non-biominéralisantes. Cependant, cette même protéine est 2 fois plus exprimée dans les coccolithophores haploïdes, qui sont tous non-biominéralisants, par rapport aux cellules diploïdes, biominéralisantes ou non. Il paraît donc y avoir une implication de cette protéine à la fois dans la formation des biominéraux, mais aussi dans d'autres processus (Mackinder et al., 2011). De même, Yarra *et al.* n'ont pas été capables de détecter de sur-expression de protéines de transport du Ca^{2+} dans des bivalves biominéralisants, en dépit de la formation de carbonate de calcium (Yarra et al., 2021).

Il existe toutefois des exemples concrets de protéines de transport d'ions Ca^{2+} qui ont été identifiées comme associées à la biominéralisation du CaCO_3 .

Sur les 4 grandes classes de transporteurs, on trouve des transporteurs du calcium impliqués dans la biominéralisation dans au moins 3 d'entre elles : ATPases, SLCs et canaux ioniques, ce qui montre la variété des mécanismes moléculaires employés.

Chez *E. huxleyi*, le phénotype de biominéralisation s'accompagne de la sur-expression de l'échangeur cation/proton CAX3, un antiporteur $\text{H}^+/\text{Ca}^{2+}$ (Mackinder et al., 2011).

La formation des œufs chez les oiseaux exige un transfert important de calcium vers les cellules de l'utérus depuis le plasma sanguin, puis de ces cellules vers le fluide utérin où a lieu la formation de l'œuf. Ceci a permis de mettre en avant la potentielle implication d'un grand nombre de protéines de transport du calcium :

- _ Pour l'entrée du Ca^{2+} dans les cellules, les protéines TRPV2, 3 et 6 (Transient Receptor Potential Vanilloid), des canaux ioniques et/ou ATP2C2, une P-type ATPase de transport du calcium de type 2C seraient utilisés ;
- _ Le transport du calcium dans la cellule se ferait grâce à la calbindin, une protéine capable de chelater les ions. D'autres protéines comme l'ATP2A2/3, des P-type ATPases endoplasmiques du transport du calcium, ou l'ITPR1/2/3, des canaux ioniques à Ca^{2+} activés par l'inositol trisphosphate pourraient aussi être impliqués ;
- _ Enfin l'export de la cellule vers le fluide utérin aurait lieu grâce à l'antiporteur $\text{Ca}^{2+}/\text{Na}^+$ SLC8A1-3 et la P-type ATPase ATP2B1-B2 (Gautron et al., 2021; Nys & Le Roy, 2018; Sah et al., 2018).

Transport des ions HCO_3^- .

La production des ions bicarbonates n'est pas forcément co-localisée avec le site de précipitation du biominéral, comme c'est le cas dans la formation des œufs de poule (Gautron et al., 2021). Dès lors, il est aussi nécessaire de transporter ces ions. Ce sont, à notre connaissance, quasi exclusivement des protéines du groupe SLC qui assurent ce processus. Plus spécifiquement, la famille SLC4, qui rassemblent des transporteurs de bicarbonate ou de carbonate, semble être la plus représentée des SLC. Dans les SLC4, le transfert de carbonate/bicarbonate s'accompagne généralement d'un transport d'ions Na^+ ou Cl^- , sous forme de symport ou d'antiport, afin de profiter d'un gradient chimique favorable pour aller à l'encontre de celui, défavorable, du bicarbonate (Romero et al., 2013).

Plus précisément, il est possible de citer SLC4A4, A5 et A10, des symporteurs $\text{Na}^+/\text{HCO}_3^-$, ainsi que SLC26A9, un antiporteur $\text{HCO}_3^-/\text{Cl}^-$ chez la poule (Nys & Le Roy, 2018); AEL1 (Anion Exchangeur like 1), un membre de la famille SLC4 dont le mécanisme n'est pas encore bien connu, et qui est sur-exprimé dans les coccolithophores *E. huxley* biominéralisants (Mackinder et al., 2011); SpiSLC4y, une SLC4 localisée uniquement sur le derme calcifiant chez le corail (Zoccola et al., 2015); enfin les bivalves qui biominéralisent sur-expriment une SLC4, ainsi qu'une SLC26A qui pourrait transporter des HCO_3^- (Yarra et al., 2021).

Transport des ions H^+ et contrôle du pH.

La précipitation du CaCO_3 (réaction (3) qui peut être considérée comme $\text{Ca}^{2+} + \text{HCO}_3^- = \text{CaCO}_3 + \text{H}^+$) s'accompagne d'une libération de protons et d'une acidification de l'organisme. Il est donc nécessaire d'évacuer ces protons pour maintenir un pH basique favorable à la précipitation du biominéral, ainsi que l'homéostasie de l'organisme biominéralisant. Ainsi, Taylor *et al.* ont calculé qu'en l'absence d'export de protons, le pH cytosolique d'un coccolithophore biominéralisant diminuerait de 0,3 pH/min (Taylor et al., 2011). De fait, des transporteurs de protons ont été identifiés comme impliqués dans la biominéralisation.

Les protéines qui semblent les plus importantes dans ce contexte sont les V-types H^+ ATPases, des ATPases qui transportent des protons. En effet, ces protéines sont sur-exprimées chez les bivalves et les coccolithophores biominéralisants (Mackinder et al., 2011; Yarra et al., 2021). De même, chez les foraminifères, la diminution du pH autour de ces organismes qui accompagne l'évacuation des protons

issus de la biominéralisation peut être annulée par l'ajout d'un inhibiteur de V-types H⁺ ATPases (Toyofuku et al., 2017). Enfin, chez les larves d'oursin, l'inhibition des V-types H⁺ ATPases diminue les capacités de régulation du pH intracellulaire des cellules biominéralisantes lors de la formation du squelette (Hu et al., 2020).

Ce ne sont toutefois pas les seuls transporteurs de H⁺ qui ont été répertoriés. Les bivalves sur-expriment aussi un canal à proton dépendant du voltage pendant la biominéralisation (Yarra et al., 2021), canal que l'on retrouve chez les coccolithophores (Taylor et al., 2011), mais dont l'utilité chez ces organismes ne semble pas être liée à la biominéralisation (Mackinder et al., 2011). Par ailleurs, chez les larves d'oursins, le SLC Na⁺/H⁺, un antiporteur, est aussi associé à la régulation du pH en période de recalcification du squelette (Hu et al., 2020).

Autre transports d'ions (Na⁺, K⁺, Cl⁻).

Très succinctement, le transport des ions essentiels à la biominéralisation est parfois réalisé avec des antiporteurs ou des symporteurs, comme décrit précédemment. Pour maintenir l'homéostasie des cellules, la biominéralisation nécessite parfois l'implication de transporteur d'ions Na⁺, K⁺ ou Cl⁻ comme cela a pu être illustré chez les bivalves ou dans la formation de l'œuf (Nys & Le Roy, 2018; Yarra et al., 2021).

Initiation de la formation du biominéral.

La formation d'un précipité se fait en au moins 2 étapes : la formation d'un noyau, puis la croissance de ce noyau. Ces 2 étapes peuvent s'accompagner d'autres évènements, comme l'agrégation de plusieurs précipités entre eux, un changement de structure, par exemple pour passer d'un précipité amorphe à cristallin (Carino et al., 2017). La formation du noyau est souvent considérée comme l'étape la plus complexe. Bien qu'en théorie la précipitation peut se faire dès lors que la saturation est atteinte, dans la pratique il faut être en état de sursaturation pour voir l'apparition d'une phase solide (Carino et al., 2017; McPherson, 2017). Dans le cas des ACC, Carino *et al.* ont modélisé que la précipitation avait lieu à une sursaturation égale à ~3,5 fois le seuil de saturation (Carino et al., 2017).

Les être vivants biominéralisants ont développé des stratégies impliquant des protéines pour atteindre plus facilement la sursaturation et déclencher la précipitation de leur biominéral.

In vitro, l'anhydrase carbonique accélère la formation d'ACC, comparé à une solution contrôle sans protéine, en augmentant la concentration d'ions bicarbonate en solution en catalysant la réaction (4), et donc la sursaturation (Rodriguez-Navarro et al., 2019). Cette réaction fait de l'anhydrase carbonique l'une des protéines les plus importantes dans la biominéralisation du carbonate (Gilbert et al., 2022; Le Roy et al., 2014).

Une autre stratégie pour les protéines consiste à interagir avec les ions Ca²⁺ ou HCO₃⁻ afin de les concentrer localement. Ainsi, une étude de dynamique moléculaire a montré que le lysozyme, une protéine présente dans le blanc d'œuf connue pour ses propriétés antibactériennes et qui est étudiée

dans la synthèse de biominéraux à base d'or (Bakshi et al., 2010) ou de titane (Y. Wu et al., 2023), favoriserait la précipitation des ACC grâce à un triplet d'arginine. Ce triplet permettrait de concentrer localement les ions carbonates (Rani & Saharay, 2019).

Cependant, ce sont plus souvent les a.a (acides aminés) acides, donc chargés négativement, qui sont considérés comme sites de nucléation, via une interaction avec les ions Ca^{2+} . Par exemple, dans le cas de l'huître perlière *Pinctada fucata*, la partie riche en acide aspartique de l'aspeïn serait responsable de la nucléation de la calcite (Takeuchi et al., 2008).

Enfin, il s'est aussi développé récemment l'idée que la nucléation du CaCO_3 pouvait être considérée comme une séparation de phase liquide - liquide, un processus dans lequel une solution homogène se sépare en 2 phases liquides contenant des concentrations distinctes de solutés. Des protéines ont été identifiées comme capable de diriger une séparation de phase liquide-liquide. De fait, ces protéines, comme Pif80 trouvée chez *Pinctada funcata*, pourraient s'avérer être des candidates très intéressantes pour expliquer la nucléation du CaCO_3 . Cette idée est encore peu développée, et donc manque encore d'exemples et de théorie, mais selon Tarczewska *et al*, elle pourrait être une piste de recherche très intéressante dans le futur (Tarczewska et al., 2022).

Sélection du polymorphe et de la cristallinité.

Les organismes qui biominéralisent le CaCO_3 cherchent généralement à contrôler le polymorphe du CaCO_3 qui est formé. Un exemple intéressant est l'huître perlière *P. fucata* déjà évoquée. Sa coquille est formée de deux couches de biominéraux : une couche d'aragonite, et une autre de calcite. Ces 2 couches seraient formées depuis une même solution riche en Mg^{2+} , une condition qui favorise fortement la formation d'aragonite. La précipitation de calcite serait due à l'aspeïn exprimée par *P. fucata*, une protéine qui permet la précipitation de ce minéral même dans un milieu qui favorise la précipitation abiotique de l'aragonite (Takeuchi et al., 2008).

De même chez le corail, la protéine CARP3 permet de favoriser, dans un milieu ne contenant pas de Mg, la précipitation de la vatérite vis à vis de la calcite, alors que la vatérite est un polymorphe instable du carbonate de calcium (Laipnik et al., 2020).

Enfin, ces protéines peuvent aussi prévenir la formation de cristaux en stabilisant des ACC, une phase thermodynamiquement métastable du carbonate de calcium. Le domaine riche en glycine de la protéine SM50, une SMP (Spicule Matrix Protein) que l'on trouve dans la matrice des spicules de la larve d'oursin a cette capacité, par exemple (Rao et al., 2013).

Inhibition de la précipitation.

Le contrôle de la biominéralisation implique d'être capable de déclencher la formation du biominéral, mais aussi de l'arrêter pour contrôler sa taille, et éviter la formation spontanée de biominéraux.

Ce rôle a été reporté pour au moins 2 protéines différentes chez l'abalone : la perlinhibine, une protéine riche en cystéine, en histidine et en arginine, et la perlwapine, une protéine qui possède 3 domaines

WAP (Whey Acidic Domain), qui s'attachent à la surface de la calcite pour inhiber sa croissance (Mann et al., 2007; Treccani et al., 2006).

Contrôle de la forme du biominéral.

Les biominéraux peuvent adopter des formes très complexes. Les coccolithes en sont un exemple frappant (une image de coccolithes est visible en Figure 1.B). La finesse et la diversité de leurs motifs en font des sujets d'applications en nano-fluidique, en exploitant les canaux présent sur les coccolithes de certaines espèces, ou en optique pour moduler les propriétés de biréfringences de la calcite (Skeffington & Scheffel, 2018). Dans un registre moins impressionnant, mais tout aussi pertinent, la calcite de la coquille d'œuf de la poule croît selon une direction privilégiée, là où un cristal abiotique, en l'absence de contrainte croîtra dans toutes les directions de l'espace (Gautron et al., 2021).

Les protéines responsables de la forme du biominéral, sont généralement des protéines de la matrice du biominéral, quand une telle matrice existe.

Dans le cas de l'œuf de poule, la protéine Ovocleidin 116, aussi connue sous le nom de MEPE (Matrix Extracellular Phosphoglycoprotein), une protéine de la matrice de la coquille, joue un rôle crucial dans cette morphogénèse. Elle est associée à la formation de micro-cavités dans la coquille, ainsi qu'à l'organisation de la calcite. De même, elle est impliquée dans l'élasticité, l'épaisseur et la forme de la coquille d'œuf. Enfin, elle est aussi impliquée dans l'arrêt de la biominéralisation (Sah et al., 2018).

De même, la protéine Starmaker qui est impliquée dans la formation des otolithes des poisson-zèbres, permettrait d'intercaler la matrice organique avec les cristaux de carbonate de calcium en interagissant à la fois avec le calcium et avec d'autres protéines. Ce serait des a.a acides régulièrement répartis sur sa chaîne carbonée qui permettraient à la protéine d'interagir avec le calcium. Par ailleurs, l'interaction des ions calcium avec les a.a acides de Starmaker permettrait de diminuer la répulsion électrostatique entre ces derniers. Cela expliquerait l'augmentation de la compacité de la protéine en présence de ces ions qui a été observée. Il a été postulé que cette diminution de l'entropie de la protéine faciliterait son interaction avec d'autres protéines (Kapłon et al., 2009).

La structuration du biominéral peut aussi faire appel au cytosquelette de l'organisme. Chez les coccolithophores, l'inhibition de la polymérisation des microtubules entraîne une malformation des coccolithes, ce qui suggère une implication du réseau de microtubules dans le contrôle de la forme du biominéral (Durak et al., 2017). De même, chez les foraminifères, la F-actine est très impliquée dans la structuration de la zone de biominéralisation (Tyszka et al., 2019). Enfin, il a été identifié chez le bivalve *Atrina rigida*, une chitine synthase qui possède un domaine myosin motor head. Cela suggère un lien entre la synthèse de la chitine, qui a été supposée avoir un rôle crucial dans la formation des biominéraux chez les bivalves (Weiss & Schönitzer, 2006), et le cytosquelette d'actine (Weiss et al., 2006).

Par ailleurs, il a été montré chez les coccolithophores, que la disruption du cytosquelette d'actine entraîne un arrêt de la biominéralisation et de l'extrusion des écailles (Durak et al., 2017). Si ce n'est pas à proprement parler un contrôle de la forme du biominéral, cela démontre chez cet organisme un rôle crucial du cytosquelette dans la biominéralisation.

Importance des protéines acides et des LCD (Low Complexity Domain).

En dépit de cette très grande diversité de protéines, certaines propriétés communes ont émergé dans l'étude des protéines de la biominéralisation.

Tout d'abord, les protéines acides, mais aussi les polysaccharides acides, sont régulièrement mis en avant par les articles traitant de la biominéralisation, notamment pour leur importance dans la matrice organique des biominéraux (Aizenberg et al., 1996; Gilbert et al., 2022; Marin & Luquet, 2007; Tyszka et al., 2019). Leur importance est souvent liée à l'hypothèse qu'elles interagissent avec les ions Ca^{2+} pour organiser et structurer le biominéral. Les a.a basiques sont beaucoup moins mis en avant. Par exemple, selon Innocentini *et al.*, l'arginine, un a.a basique, est peu présente dans les protéines biominéralisantes, même si il faut préciser que les auteurs ne proposent pas de données chiffrées pour appuyer leur propos. D'après eux, l'arginine pourrait avoir une trop grande affinité pour les ACC, plus grande que les acide glutamique et aspartique, ce qui verrouillerait les protéines riches en arginines sur les ACC, et empêcherait l'intervention d'autres protéines dans la maturation du biominéral; ce qui n'est pas souhaitable par les organismes biominéralisants (Innocenti Malini et al., 2017).

De même, une autre grande caractéristique qui a émergé de l'étude des protéines de la biominéralisation, est la prévalence de LCD (Low Complexity Domains, ou domaines de faible complexité). Ces domaines ont une composition fortement biaisée vers un ou quelques a.a. Dans le cas de la biominéralisation, ces a.a sont souvent l'acide aspartique, la serine, la thréonine, la glycine ou l'alanine, bien que des LCD biaisés vers l'arginine ou la lysine puissent aussi être rencontrés (Gilbert et al., 2022; Marin, 2020). Ces domaines sont souvent désordonnés (Marin, 2020). Le désordre étant par ailleurs probablement très important pour ces protéines pour augmenter leur surface de contact avec le biominéral ou d'autres protéines, et changer rapidement de conformation pour s'adapter aux changements du biominéral (Kalmar et al., 2012).

L'inconnue des protéines de la biominéralisation.

Nous avons ici cherché à illustrer quelques fonctions qui nous paraissent importantes dans la biominéralisation et qui pourraient nourrir des hypothèses pertinentes sur la fonction de la calcyanine qui est l'objet de cette thèse. Cependant il faut garder à l'esprit qu'il existe encore une quantité impressionnante de fonctions très diverses de protéine associées à la formation de biominéraux. A ce propos, dans une revue sur le shellome des mollusques, l'ensemble des protéines associées aux coquille, F. Marin rapporte qu'il peut être trouvé associé à la coquille des mollusques des domaines protéiques typiques d'enzymes (tyrosinase, protéase, peroxidase..), de fonctions immunitaires et d'inhibiteurs de protéases (serpin, TIMP, lipocain-like...), de fixation aux saccharides... Soit une grande variété de fonctions protéiques. De plus, comme l'illustrent ces exemples, une part importante de ces domaines n'ont pas d'utilité évidente pour la biominéralisation, indiquant ainsi qu'il y a encore beaucoup à apprendre sur les mécanismes biochimiques de formation de biominéraux (Marin, 2020).

Enfin, il est tout à fait possible qu'il existe encore des fonctions protéiques cruciales pour la biominéralisation qui sont encore inconnues.

Par ailleurs, Polowczyk *et al.* ont montré que l'action combinée du lysozyme et de l'ovalbumine sur le CaCO_3 était très différente de l'action du lysozyme ou de l'ovalbumine seuls (Polowczyk *et al.*, 2016). Ce n'est donc pas seulement la connaissance des protéines individuelles qui est pertinente dans le cadre de la biominéralisation, mais aussi celle de leurs interactions, ce qui complexifie d'autant leur étude.

Enfin, certaines protéines impliquées dans la biominéralisation ont été reconnues comme ayant plusieurs fonctions potentielles. C'est le cas de l'anhydrase carbonique qui catalyse la formation des ions bicarbonates, accélère la transition des ACC vers la calcite et peut s'auto-assembler pour s'intégrer à la matrice du biominéral (Rodriguez-Navarro *et al.*, 2019). Il faut donc aussi prendre en compte que ces protéines peuvent être multifonctionnelles pour comprendre pleinement leur importance.

La biominéralisation intracellulaire d'ACC par les cyanobactéries.

Parmi les organismes les plus impliqués dans la biominéralisation du CaCO_3 se trouvent les cyanobactéries. Les cyanobactéries sont des bactéries qui sont présentes dans une très grande diversité d'environnements : eau douce, eau salée, sources géothermales, déserts, régions polaires, microbialithes (Whitton, 2012), ou encore l'intestin humain (Soo *et al.*, 2017). Elles sont souvent considérées comme étant toutes capables de photosynthèse, bien que la découverte récente de bactéries qui sont reliées génétiquement aux cyanobactéries, mais qui sont dépourvues de systèmes photosynthétiques, laisse penser que ce trait n'est peut-être pas tout à fait commun à toutes les cyanobactéries (Soo *et al.*, 2014, 2017). Quoiqu'il en soit, celles qui en sont capables ont eu un impact majeur sur la Terre. Tout d'abord, les cyanobactéries pourraient être les responsables du « Great Oxygenation Event », qui a vu la proportion d' O_2 dans l'atmosphère passer de 10^{-5} % à 1-10% de la pression atmosphérique actuelle, il y a ~2,4 milliards d'années (Olejarz *et al.*, 2021).

Ensuite, l'endosymbiose d'une cyanobactérie par une cellule eucaryote est la théorie la plus populaire pour expliquer l'existence des chloroplastes qui permettent aux plantes de faire de la photosynthèse (Raven & Allen, 2003; Sato, 2021).

Ce sont aussi ces capacités photosynthétiques qui expliquent une partie de la biominéralisation du CaCO_3 faite par les cyanobactéries. En effet, pour pouvoir fixer le CO_2 avec l'enzyme RuBisCO, le HCO_3^- est transformé en CO_2 par l'anhydrase carbonique ce qui augmente le pH de l'environnement extra-cellulaire et favorise la précipitation du CaCO_3 . Cette biominéralisation induite s'accompagne aussi d'une biominéralisation influencée. Il a été constaté que le CaCO_3 peut aussi nucléer et précipiter sur l'EPS (ExoPolysaccharide) ou la membrane des cyanobactéries, bien que les mécanismes sous-jacents ne soient pas toujours bien compris (Görgen *et al.*, 2021).

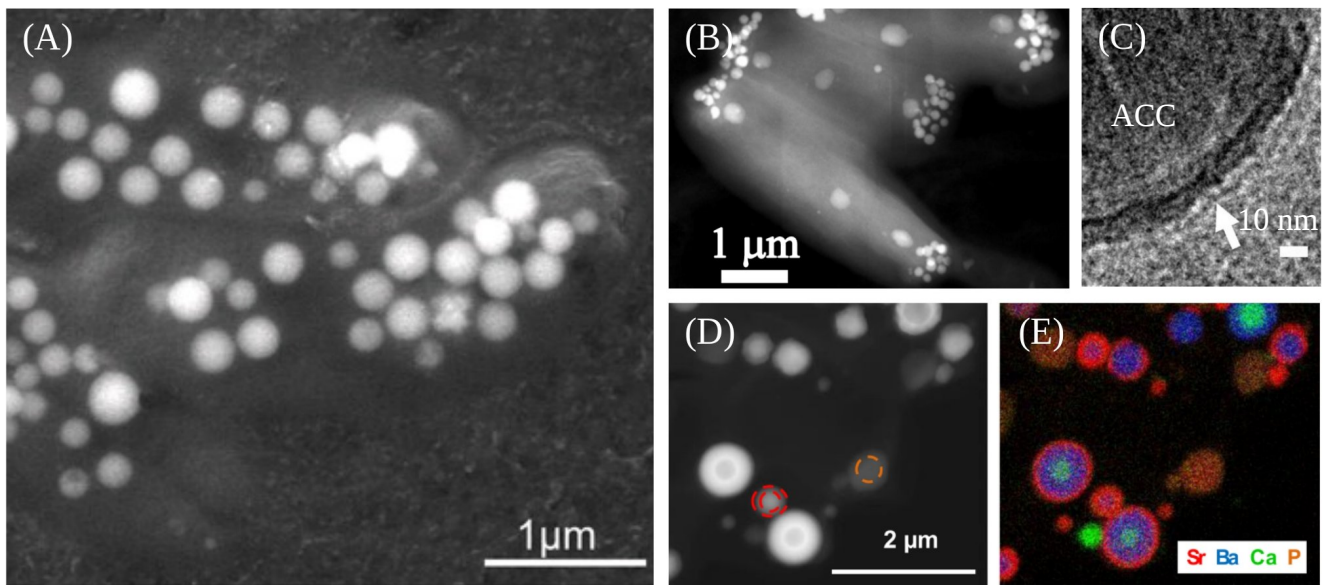


Figure 6 : La biominéralisation intracellulaire chez les cyanobactéries.

(A) Image MEB d'une cyanobactérie biominéralisante (Couradeau *et al.*, 2012). Les inclusions blanches sont des ACC. (B) Image MET de *Synechococcus calcipolaris* avec les iACC localisés aux pôles de la cellule (Benzerara *et al.*, 2014). (C) Image de cryo-EM (cryo-Electron Microscopy) de l'enveloppe entourant les iACC de *Gloeomargarita lithophora* (indiquée par une flèche blanche) (Blondeau *et al.*, 2018). (D) Image MET des iACC de *G. lithophora*. (E) Cartographie chimique EDXS (Energy Dispersive X-ray Spectroscopy) des iACC montrés en D. La structure en anneaux témoigne de la sélectivité vis-à-vis de l'incorporation d'alcalino-terreux de la souche lors de la formation de ces inclusions (Cam *et al.*, 2016).

Figure 8 : Alignements de séquence du domaine (GlyZip)₃ et du domaine CoBaHMA des calcyanines.

Extraits de Benzerara *et al.*, 2022.

Alignement de séquence du domaine (A) (GlyZip)₃ et du domaine (B) CoBaHMA des calcyanines. Les positions sont surlignées en noir. Les autres positions conservées sont colorées selon le type d'a.a.

Pendant longtemps, il a été considéré qu'il n'existait pas de biominéralisation contrôlée du CaCO₃ par les cyanobactéries. Toutefois, en 2012, il a été observé au laboratoire pour la première fois la présence d'ACC intracellulaires chez la cyanobactérie *Gloeomargarita lithophora* (Figure 6.A; Couradeau *et al.*, 2012), une cyanobactérie isolée depuis le lac alcalin d'Alchichica au Mexique (Moreira *et al.*, 2017). Cette souche a aussi été identifiée par Ponce-Toledo *et al.* (2017), grâce à des analyses phylogénétiques, comme le plus proche voisin de l'ancêtre bactériens des plastides, dont font partie les chloroplastes. Ces ACC ont été nommés iACC (pour ACC intracellulaire).

Le nombre d'espèces de cyanobactéries où des iACC ont été observés a fortement augmenté par la suite. Benzerara *et al.* en dénombrait 19 en 2022, Mehta *et al.* ont répertorié 4 *Microcystis* supplémentaires formant des iACC, enfin Gaëtan *et al.* ont rapporté la présence d'iACC dans des prélèvements environnementaux de *Microcystis* (Benzerara *et al.*, 2014, 2022; De Wever *et al.*, 2019; Gaëtan *et al.*, 2023; Mehta *et al.*, 2022).

Ces iACC mesurent entre 60 nm et 870 nm. Ils sont entourés par une enveloppe, dont la nature est pour l'heure inconnue, mais qui a une épaisseur cohérente avec une mono-couche lipidique ou une couche protéique (Figure 6.C; Blondeau et al., 2018). Deux distributions spatiales ont été observées pour la localisation de ces iACC : soit dispersés dans la cellule (Figure 6.A), soit localisés à ses pôles (Figure 6.B), ce second cas évoquant une implication de la machinerie de division cellulaire et/ou du cytosquelette (Benzerara et al., 2014; J. Li et al., 2016). De même, 2 comportements distincts ont été observés quant à l'incorporation des alcalino-terreux dans les iACC. Dans la majorité des espèces de cyanobactéries qui forment des iACC, la proportion d'alcalino-terreux autres que le calcium dans le biominéral est proche de celui de l'environnement. A l'inverse, *Gloeomargarita lithophora* privilégie les alcalino-terreux les plus lourds pour la formation de ses iACC : en présence de baryum, de strontium et de calcium, la cellule prélève d'abord le baryum, puis le strontium, et une fois ces 2 éléments épuisés, le calcium (Figure 6.D&E; Cam et al., 2016). Pour l'heure, seulement une autre souche de cyanobactérie, *Gloeomargarita ahusahtiae*, a été reportée comme ayant possiblement une sélectivité similaire (Bacchetta et al., 2022). Les expériences de précipitation abiotique d'ACC ne permettent pas d'expliquer ce comportement, ce qui indique une implication de *G. lithophora* dans la sélection de ces alcalino-terreux (Cam et al., 2015).

Les souches qui forment des iACC sont des hyper-accumulatrices d'alcalino-terreux. La masse de calcium contenue dans la cellule comparée à la masse sèche de la cellule est 10 à 100 fois plus élevées chez ces cyanobactéries, par rapport aux procaryotes non biominéralisants. Parmi les cyanobactéries biominéralisantes, *G. lithophora* fait partie de celles qui hyper-accumulent le plus. Par ailleurs, certaines de ces souches ont besoin de plus de Ca^{2+} dans leur environnement que les souches non-biominéralisantes pour ne pas être limitées dans leur croissance (De Wever et al., 2019).

Cette hyper-accumulation de la part de *G. lithophora* couplée à la sélectivité de cette espèce pour les alcalino-terreux lourds et à sa tolérance à la radioactivité, en font un organisme très prometteur pour isoler les isotopes radioactifs ^{226}Ra et ^{90}Sr des milieux aqueux (Mehta et al., 2019).

Les cyanobactéries formant des iACC sont capables de biominéraliser dans une solution sous-saturée en CaCO_3 (Cam et al., 2018). De plus, les mesures rapportées dans la littérature des conditions chimiques intracellulaires pour les cyanobactéries indiquent des pH cytosoliques compris entre 6,4 et 8,4 (Sekine et al., 2023), et des concentrations en Ca^{2+} qui vont de 100-200nM à 3 μM (Barrán-Berdón et al., 2011; Torrecilla et al., 2000), des valeurs qui sont incompatibles avec une saturation en CaCO_3 (Cam et al., 2018). Ces mesures n'ont pas été effectuées sur des souches qui forment des iACC, mais, à moins que les conditions intracellulaires des souches biominéralisantes ne soient spectaculairement différentes de celles dans lesquelles ont été faites les mesures, la précipitation de CaCO_3 ne peut pas avoir lieu spontanément dans une cyanobactérie. Il y a donc un contrôle actif de la part des cyanobactéries formant des iACC pour accumuler et concentrer localement dans un compartiment le Ca^{2+} et le CO_3^{2-} nécessaires à la précipitation du CaCO_3 .

Les raisons de ce phénotype de biominéralisation sont pour l'heure inconnues. Plusieurs hypothèses ont été soulevées par les auteurs ayant travaillé sur ce sujet. Pour reprendre le résumé établi par De Wever et al. (De Wever et al., 2019), il a été proposé que :

- _ Les iACC lestent la cellule;
- _ Les iACC servent à tamponner le pH de la cellule, et à compenser la formation d'ions OH^- lors de la transformation de HCO_3^- en CO_2 pendant la photosynthèse;
- _ Les iACC sont un réservoir de carbone inorganique.

Découverte de la calcyanine.

Afin de comprendre les mécanismes à l'œuvre derrière la formation des iACC, une analyse de génomique comparative a été effectuée entre 6 souches de cyanobactéries formant des iACC, et 50 n'en formant pas. Un seul et unique gène a été identifié comme partagé par toutes les souches biominéralisantes, et absents des autres. A l'inverse, il n'a été observé aucun gène spécifique des souches non-biominéralisantes mais absent des souches biominéralisantes (Diop, 2016, Rapport de M2).

Ce gène spécifique des souches formant des iACC était totalement inconnu à l'époque, et a été nommé *ccyA*. L'étude de la phylogénie du gène *ccyA* semble indiquer que ce gène est ancestral chez les cyanobactéries. *CcyA* code pour une protéine qui a été nommée calcyanine. Cette protéine n'a jamais été décrite, et son rôle est donc inconnu.

Des premières analyses bioinformatiques ont été conduites, résumées dans l'article de Benzerara *et al.* (2022), qui se trouve en Annexe 1.

Les séquences de calcyanine qui ont été répertoriées ont tout d'abord été soumises à une analyse HCA (Hydrophobic Cluster Analysis). Cette approche, utilise une représentation 2D des séquences en a.a, reprenant la trame d'une hélice α dupliquée. Les amas d'a.a hydrophobes forts (V I L F M Y W) sont entourés pour être mis en évidence. Cette méthode permet de mettre en lumière des informations structurales, par l'intermédiaire des amas hydrophobes dont il a été montré qu'ils correspondent majoritairement aux structures secondaires régulières présents dans les domaines repliés. En sus de mettre en évidence des structures secondaires, l'analyse HCA permet de révéler la présence d'éléments répétés ou de régions désordonnées (caractérisées par l'absence d'amas) (Bitard-Feildel *et al.*, 2018; Callebaut *et al.*, 1997).

Hormis les calcyanine à domaine Z (en bleu sur la Figure 7), les séquences de calcyanine contiennent sur la quasi totalité de leur séquence des amas hydrophobes. Les tailles et les contenu en a.a de ces amas sont caractéristiques de domaines globulaires de protéines (Lamiable *et al.*, 2019). Des similitudes très nettes sont visibles entre les régions C_{ter} (C-terminal) des différentes calcyanines (encadrés orange sur la Figure 7). A l'inverse, les amas des régions N_{ter} (N-terminal) divergent entre les différentes séquence. Le fait que la région C_{ter} soit conservée, alors que la région N_{ter} soit variable a permis de découper la calcyanine en 2 domaines.

Le domaine C_{ter} de la calcyanine est hautement conservé (Figure 8.A). Il est formé d'une triple répétition d'un motif glycine zipper long de ~ 50 a.a, défini par la répétition de la séquence GxxxG (en jaune sur la Figure 7), et s'accompagne d'une répétition tous les 3/4 résidus également d'a.a hydrophobes. Il est interrompu en son milieu par un motif GP conservé. Bien que le motif glycine zipper ait été largement décrit dans la littérature, celui des calcyanines est bien plus long (~ 12 répétitions du GxxxG) que la majorité de ceux déjà reportés (Kim *et al.*, 2005; Martin *et al.*, 2012; Teese & Langosch, 2015). Enfin, les glycine zippers des calcyanines ont des positions polaires conservées à leurs extrémités, souvent occupées par des a.a acides (en rose sur la Figure 7).

La forme horizontale des amas hydrophobes intégrés aux glycine zippers (en vert sur la Figure 7) est indicatrice de structures hélicoïdales de type amphiphatique, la structure 3D de chacun de ces 3 motifs glycine zippers a donc été prédite comme un motif en épingle à cheveux formé de 2 hélices

antiparallèles, le motif GP formant l'épingle. Les 3 motifs ont été nommés GlyZip1, GlyZip2 et GlyZip3. Le domaine complet a été nommé (GlyZip)₃.

Des recherches de similarité de séquences ont pu montrer une proximité entre le GlyZip2 et le glycine zipper localisé en C_{ter} du domaine OmpA-like de la famille de protéines PdsO (Proteobacterial sortase system peptidoglycan-associated protein). Selon l'entrée de la banque de domaines InterPro (Paysan-Lafosse et al., 2023) associée à PdsO (IPR022511), cette famille de protéines est spécifique des Protéobactéries, mais sa fonction, et donc celle de son glycine zipper, est pour l'heure inconnue. En revanche aucune similarité de séquence n'a été détectée pour le domaine (GlyZip)₃ complet.

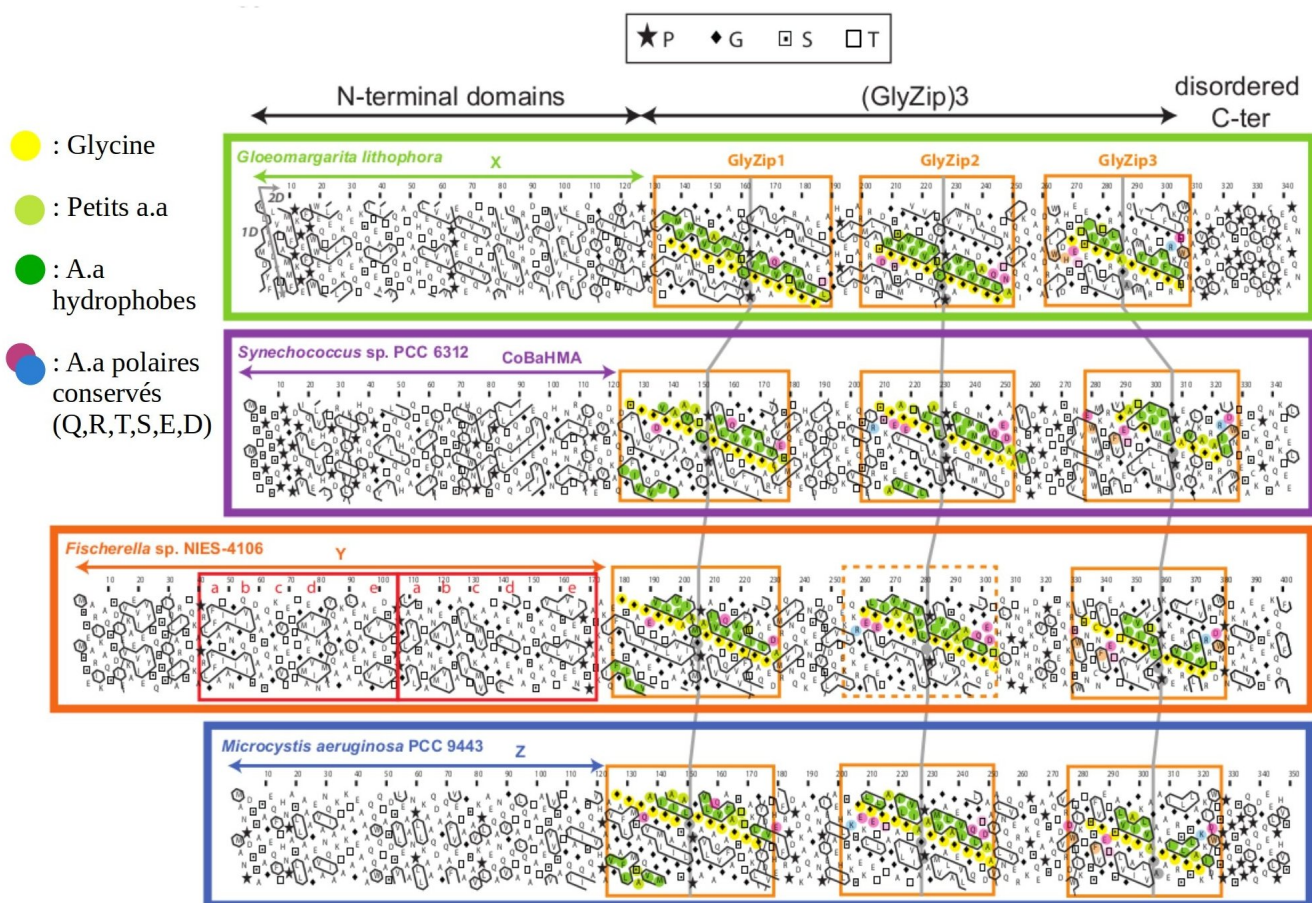


Figure 7 : Tracés HCA des 4 variants de la calcanine actuellement décrits.

Extrait de Benzerara *et al.*, 2022.

Les séquences en a.a sont inscrites sur une trame α -hélicoïdale dupliquée. Les a.a hydrophobes forts (V I L F M Y W) sont entourés, formant des amas correspondant majoritairement aux structures secondaires régulières. Les symboles utilisés pour représenter certains a.a sont repris dans l'encart du haut. Le découpage en domaine est illustré avec des flèches. Les 3 glycines zippers sont encadrés en orange. Les a.a conservés sont surlignés en couleur. La duplication du domaine Y est mise en valeur par des cadres rouges.

A l'inverse, le domaine N_{ter} a plusieurs variants répertoriés. 4 d'entre eux, nommés X, Y, Z et CoBaHMA, ont été décrits dans Benzerara *et al.*, 2022, mais il se pourrait qu'il en existe d'autres encore. Les 4 variants décrits pour le moment sont radicalement différents les uns des autres en termes de séquence, au point qu'il n'est pour l'heure pas possible de détecter de similarité de séquence entre eux. Ces 4 domaines n'avaient jamais été décrits avant l'étude des calcyanines. De plus, 3 d'entre eux, X, Y et Z, ne présentent pas de similarité de séquence avec des domaines déjà connus.

A l'inverse, le domaine CoBaHMA partage des similarités de séquence avec des domaines du repliement « ferredoxin-like » qui feront l'objet d'une plus ample description dans la suite de cette thèse. Le domaine CoBaHMA est aussi caractérisé par une très forte conservation de 4 a.a basiques, qui forment un motif HxxxxRxRxR (Figure 8.B).

Il est aussi intéressant de remarquer que le domaine Y est constitué d'une duplication de séquence, une caractéristique qui pourrait être partagée avec le domaine X.

Par ailleurs, la majorité des calcyanines à domaine Y n'ont pas le GlyZip2 dans leur domaine (GlyZip)₃, ce qui laisserait entendre que ce GlyZip n'est pas essentiel à la fonction de la calcyanine.

Il semblerait que la sélectivité dans le choix des alcalino-terreux que présente *G. lithophora* (Cam et al., 2016) et que semble partager *G. ahouahtiae* (Bacchetta et al., 2022) dans la formation des iACC soit liée au domaine X que ces 2 souches ont sur leur calcyanine. Il est cependant difficile de conclure sur le lien entre un domaine et un phénotype avec seulement 2 exemples.

Enfin, la partie C_{ter} désordonnée, dont les séquences sont visibles sur la Figure 8.A, semble peu conservée entre les différents variants de la calcyanine. Par exemple, elle est très riche en prolines chez *G. lithophra* (~1/3 des a.a) alors que celle *Fischerella sp. NIES-4106*, n'a qu'une seule proline (Figure 1). De plus, au sein même des variants de la calcyanine, il ne semble pas y avoir de conservations évidentes pour cette partie de la protéine.

Comme le N_{ter} de la calcyanine est variable, la recherche de nouvelles calcyanines a été réalisée sur la base du domaine (GlyZip)₃ C_{ter}. Un profil de ce domaine a été utilisé pour rechercher le gène *ccya* dans 594 génomes de cyanobactéries. 27 d'entre elles possédaient ce gène. Sur ces 27, 17 ont pu être observées au microscope. 13 de ces souches observées, formaient des iACC. Ce gène est donc un marqueur génétique de cette biominéralisation. Par la suite, il a été utilisé pour repérer de nouvelles souches de cyanobactéries biominéralisantes. Ce gène est répandu chez les cyanobactéries, il se trouve dans plusieurs de ses genres comme *Microcystis*, *Synechococcus* ou encore *Fischerella* (Benzerara et al., 2022).

GlyZip3



| | | |
|---|-----|--|
| <i>Gloeomargarita lithophora Alchichica-D10</i> | 260 |ADRLVAI...DHSGETTSELIGGTVGRVLVAGA..OGQMYGARLQTY.....FGRKISWQNAAGVPDPLATAAPISPTDLPDLPE (+11*) |
| <i>Cyanothece sp. PCC 7425</i> | 290 | EAAATPEEMASKTNSWLEKTAAGFVGETASATIGGALGSLALGP..QGKEVGMKLGNR.....ISRFIDWEGKEKPSDPKPAPELPPPA* |
| <i>Synechococcus calcipolaris G9</i> | 269 | ..IPTDLSVNIIVLEWMMKTSRAFVGETALATLGGLLSRVILGP..QAEVGLRACFR.....VGRLDWNGQDGGKTKEMAETPTTETITTSPE (+07*) |
| <i>Synechococcus sp. PCC 6312</i> | 266 | TAPTSSSVNIIVLEWMMKTSRAFVGETALATLGGLLMRILGP..QAEAVGLKAGSR.....AGRIIDWNTESASCKQPQSNQAKTEEL* |
| <i>Thermosynechococcus elongatus BP-1</i> | 249 | ..GPPPSLHFFLEWVVKTSQTFKGETLLATLGGIVARGILGP..QAEAAQIRACGR.....MSRHLDROQSNTAAKEKKV* |
| <i>Chroococcidiopsis thermalis PCC 7203</i> | 255 | TATNPORLPATTOQFVQTTQKQFGETATVTVGGAVGKIVLGS..PGQOMGLKLCGR.....ISKVIEWDTTPHVRVDITPTPQLLGETQTGDRQSY* |
| <i>Neosynechococcus sphagnicola syl</i> | 264 | ..HPSSGNVLQDLQSHIDRTSRDFAGETSAVLGSLGLFLGT..SGRDTRRIGSY.....VGRRLNWRDPSPPPILEATTSPASPAPESTAIAS (+14*) |
| <i>Scytonema millei</i> | 255 | SSATLQKSPATTOQLVNTTQKQFGETATATVGGAVGKIIILGS..PGQOMGLQLCGR.....ISKVVDWNTATTHQVNIPTPQLSVENPQSDR* |
| <i>Synechococcus lividus PCC 6715</i> | 249 | ..DPTPPSVDVLEWMLKASRSFVGETLWATLGGLLARLTLGP..HAEAVGTRACFR.....ISRQDMSEPALTTAPQKKEV* |
| <i>Synechococcus PCC 6716</i> | 249 | ..DPTPPSVDVLEWMLKASRSFVGETLWATLGGVLRARLTLGP..HAEAVGTRACFR.....ISRQDMSEPALTTAPQKKEV* |
| <i>Synechococcus PCC 6717</i> | 249 | ..DPTPPSVDVLEWMLKASRSFVGETLWATLGGVLRARLTLGP..QAEAVGTRACFR.....ISRQDMSEPALTTAPQKKEV* |
| <i>Thermosynechococcus vulcanus NIES-2134</i> | 249 | ..GPPPSLHFFLEWVVKTSQTFKGETLLATLGGIVARGILGP..QAEAAQIRACGR.....MSRHLDROQSNTAAKEKKV* |
| <i>Thermosynechococcus sp. NK55a</i> | 249 | ..EPSPPSLHFFLEWVVKTSQTFKGETLLGTLGGIVARMLGP..QAEAEQIRACFR.....ISRHLDMQSQSNTAPGKKNYKVIAYGDVQRDRS* |
| <i>Microcystis sp. T1-4</i> | 264 | PSEEAASPEQTDVDRDWTIKTGSSFFVGETGSALVGGAIKGLVAGA..AGAEIGRSLGAI.....AGKIIDWTSQSPSSPPQPPDQGENPHVS* |
| <i>Microcystis aeruginosa PCC 9806</i> | 264 | PSQEAASPAQTDVDRDWTIKTGSSFFVGETGSALVGGAIKGLVAGA..AGAEIGRSLGAI.....AGKIIDWTSQSPSSPPQPPDQGENPHVS* |
| <i>Microcystis aeruginosa PCC 9432</i> | 264 | PSQEAASPAQTDVDRDWTIKTGSSFFVGETGSALVGGAIKGLVAGA..AGAEIGRSLGAI.....AGKIIDWTSQSPSSPPQPPDQGENPHVS* |
| <i>Microcystis aeruginosa CACIAM O3</i> | 264 | PSQEAASPAQTDVDRDWTIKTGSSFFVGETGSALVGGAIKGLVAGA..AGAEIGRSLGAI.....AGKIIDWTSQSPSSPPQPPDQGENPHVS* |
| <i>Microcystis aeruginosa PCC 9443</i> | 264 | PSQEAASPAQTDVDRDWTIKTGSSFFVGETGSALVGGAIKGLVAGA..AGAEIGRSLGAI.....AGKIIDWTSQSPSSPPQPPDQGENPHVS* |
| <i>Microcystis aeruginosa PCC 9808</i> | 264 | PSQEAASPAQTDVDRDWTIKTGSSFFVGETGSALVGGAIKGLVAGA..AGAEIGRSLGAI.....AGKIIDWTSQSPSSPPQPPDQGENPHVS* |
| <i>Microcystis aeruginosa PCC 7941</i> | 264 | PSQEAASPAQTDVDRDWTIKTGSSFFVGETGSALVGGAIKGLVAGA..AGAEIGRSLGAI.....AGKIIDWTSQSPSSPPQPPDQGENPHVS* |
| <i>Microcystis aeruginosa NaRes975</i> | 264 | PSQEAASPAQTDVDRDWTIKTGSSFFVGETGSALVGGAIKGLVAGA..AGAEIGRSLGAI.....AGKIIDWTSQSPSSPPQPPDQGENPHVS* |
| <i>Microcystis aeruginosa SPC777</i> | 264 | PSQEAASPAQTDVDRDWTIKTGSSFFVGETGSALVGGAIKGLVAGA..AGAEIGRSLGAI.....AGKIIDWTSQSPSSPPQPPDQGENPHVS* |
| <i>Microcystis aeruginosa PCC 7806SL</i> | 264 | PSQEAASPAQTDVDRDWTIKTGSSFFVGETGSALVGGAIKGLVAGA..AGAEIGRSLGAI.....AGKIIDWTSQSPSSPPQPPDQGENPHVS* |
| <i>Microcystis aeruginosa PCC 9717</i> | 264 | PSQEAASPAQTDVDRDWTIKTGSSFFVGETGSALVGGAIKGLVAGA..AGAEIGRSLGAI.....AGKIIDWTSQSPSSPPQPPDQGENPHVS* |
| <i>Microcystis aeruginosa PCC 9807</i> | 264 | PSQEAASPAQTDVDRDWTIKTGSSFFVGETGSALVGGAIKGLVAGA..AGAEIGRSLGAI.....AGKIIDWTSQSPSSPPQPPDQGENPHVS* |
| <i>Synechococcus sp. CB0101</i> | 225 |QVGRVSNLVRLEDDQAGEREALLGGLGRGLSGQEWGAQLTRSLGAS.....LGRKIDWASWSRHLVRIKPRVAPPAS* |
| <i>Synechococcus sp. Lanier</i> | 198 |WRNPKLRTTQGRSAAEAASQNLPLGLLGAAILON..PGRKAEKLCGLY.....LGRINWQNLKPLTN* |
| <i>Synechococcus sp. RS9917</i> | 231 |GVGKRRWFNSMVDDSDAGEALSELAAIRIGSLLGNNPARIASVCMR.....VGRKINWRASVEQRHLVNLRLQVPT* |
| <i>Fischerella sp. NIES-4106</i> | 318 | DNVKPLSSSQQKGSWLGNTACNFLGETGTAVIGGIVGTVLGP..KGKEVGRKIGMF.....VGRRTDWNPNTKIADHVEAGSEGEKQYF* |
| <i>Hastigocladus laminosus UU774</i> | 248 | SNVKPLSSSQQKGSWLGNTACNFLGETGTAVIGGIVGTVLGP..KGKEVGRKIGMF.....VGRRTDWNPNTKIADHVEAGSEGEKQYF* |
| <i>Fischerella muscicola PCC 74</i> | 248 | CDVKPVPSSSQQKGSWLGNTACNFLGETGTAVIGGIVGTVLGP..KGKEVGRKIGMF.....VGRRTDWNPNTKIADHVEAGSEGEKQYF* |
| <i>Fischerella sp. NIES-3754</i> | 248 | SNVKPLSSSQQKGSWLGNTACNFLGETGTAVIGGIVGTVLGP..KGKEVGRKIGMF.....VGRRTDWNPNTKIADHVEAGSEGEKQYF* |
| <i>Chlorogloeopsis fritschii PCC 9212</i> | 248 | YNVKPSSSQQKGSWLGNTACNFLGETGTAVIGGIVGTVLGP..KGKEVGRKIGMF.....VGRRTDWNPNTKIADHVEAGSEGEKQYF* |
| <i>Fischerella major NIES-592</i> | 248 | SNVKPLSSSQQKGSWLGNTACNFLGETGTAVIGGIVGTVLGP..KGKEVGRKIGMF.....VGRRTDWNPNTKIADHVEAGSEGEKQYF* |
| <i>Chlorogloeopsis fritschii PCC 6912</i> | 248 | YNVKPSSSQQKGSWLGNTACNFLGETGTAVIGGIVGTVLGP..KGKEVGRKIGMF.....VGRRTDWNPNTKIADHVEAGSEGEKQYF* |

(B)

| | | | |
|---|---|---|-----|
| <i>Cyanothece sp. PCC 7425</i> | 1 | MTSTSVNCSPGYESADAARSLIPAFLEHAIATGRRLRLRIERLRRODKAYGVCCQQOIQSLAGVVSIRVNPE..AGSIVVDYQVVMGQE..AQTAEELKLVLEVCNLHVQFPKGS | 109 |
| <i>Synechococcus calcipolaris G9</i> | 1 |MPKPSDDEESLPPVAELVHLTRDRRLRLRPLPKKDDPYGRYQDYLKPIPGITVEVRLNLQ..AASLSIHYALD.....LITPQLLALIERWGDVQIIGQGKH | 96 |
| <i>Synechococcus sp. PCC 6312</i> | 1 | ..MSSTSPSSSQPPGDPVIOAELVHLTPDRRLRLKIPHLRQDQPGYGTYSQQHLQAQTGITEVRLNST..AQSLTLHWNPO.....VISLPLLTLQAIIGDLEVIQGNH | 102 |
| <i>Thermosynechococcus elongatus BP-1</i> | 1 |MATAETAPIAELVHLTGDRRLRLRIQELKTDVGRFRDS..TAYLKTIRGIRSVHVNPL..AASSTIEVHRE.....QITPQLLAAIQFWGEVQIIGQGNR | 91 |
| <i>Thermosynechococcus sp. NK55a</i> | 1 |MATAETAPIAELVHLTGDRRLRLRIQELKTDVGRFRDS..TAYLKTIRGIRSVHVNPL..AASSTIEVHRE.....QITPQLLAAIQFWGEVQIIGQGNR | 91 |
| <i>Synechococcus lividus PCC 6715</i> | 1 |MATAETAPIAELVHLTGDRRLRLRIQELKTDVGRFRDS..TAYLKTIRGIRSVHVNPL..AASSTIEVHRE.....QITPQLLAAIQFWGEVQIIGQGNR | 91 |
| <i>Thermosynechococcus vulcanus NIES-2134</i> | 1 |MATAETAPIAELVHLTGDRRLRLRIQELKTDVGRFRDS..TAYLKTIRGIRSVHVNPL..AASSTIEVHRE.....QITPQLLAAIQFWGEVQIIGQGNR | 91 |
| <i>Synechococcus PCC 6716</i> | 1 |MAAADTAPIAEVHLTGDRRLRLRIQELKRDPAFRDS..VAYLHLEGRIRAVHCNPL..AASSTIEVHRE.....QITPQLLAAIQFWGEVQIIGQGNR | 91 |
| <i>Synechococcus PCC 6717</i> | 1 |MAAADTAPIAEVHLTGDRRLRLRIQELKRDPAFRDS..VAYLHLEGRIRAVHCNPL..AASSTIEVHRE.....QITPQLLAAIQFWGEVQIIGQGNR | 91 |
| <i>Chroococcidiopsis thermalis PCC 7203</i> | 1 |MSLAVELDAEESTVVKVRSIAIAVAVAGRLRLRLRLLQPIEDAAAYA..OEEQIRSLAFVESIRVNNT..VGCITITVEVTFQFASTPPLPVELLAAIAQVSNLQIDLGTN.. | 104 |
| <i>Scytonema millei VB511283</i> | 1 |MSLAAELDATEITVLIKVRGAIAAVAVAGRLRLRLRLLQPIEDAAAYA..OEEQIRSLAFVESIRVNNT..VGCITITVEVTFQFASTPPLPVELLAAIAQVSNLQIDLGTN.. | 104 |
| <i>Neosynechococcus sphagnicola syl</i> | 1 |MPDVLQVGVYIVHSTLGRMRRLRIRYRLTLDSEYAQOEEQILRSQVWVSMRINAA..AASLIVESGSOLSP..AEAEARLVAVMAQFIVTKISPPGA | 95 |
| <i>Synechococcus sp. RS9917</i> | 1 |MNPRSNTLQVVEHLPGRRLRLRYPADLTAEERQ..GOSMLLAEPPWVAALRCSSA..SRSLVITLAAAG.....CTAVRNOIALAAMGWLLSDGHHAS | 87 |
| <i>Synechococcus sp. CB0101</i> | 1 |MSSAQQLMHLATNQRRLRLRYSLELMPDAAEAIAAALWGPQWVHALQORRA..SRSLVLELEPG.....CTFVRWHLALADLGCGLVDRRR.. | 83 |
| <i>Synechococcus sp. Lanier</i> | 1 |MLPKRIRLKLIIH.....GSYKNVIEVASSLGGHWHGA..SRSLIVLPP.....TLQVANSALYNQGWLLQAPHPLL | 65 |

Identical aa ■ Conserved aa: ■ V,I,L,F,M,Y,W (A,C,T,S) - hydrophobic ■ W,Y,F - aromatic ■ G,A (C,S,T) - small ■ R,K,H - basic ■ D,E,Q,N,S,T - polar, non basic ■ P

La calcyanine semble donc jouer un rôle central dans la biominéralisation des iACC chez les cyanobactéries. Pour comprendre précisément ce rôle, ainsi que le mécanisme dans lequel la protéine est impliquée, plusieurs projets ont été initiés sur la calcyanine, pour la plupart rassemblés au sein de l'ANR HARLEY, qui a financé cette thèse.

Tout d'abord les souches biominéralisantes ont été étudiées d'un point de vue génétique. Pour cela l'équipe de F. Chauvat (I2BC) a dû tout d'abord développer des outils pour pouvoir transformer plusieurs espèces de cyanobactéries biominéralisantes pour lesquelles aucun outil n'était disponible. Cette étape indispensable a été laborieuse et a mis en évidence les limites de ce type de manipulation pour des souches polyploïdes, à croissance lente, naturellement résistantes aux antibiotiques. De fait les résultats présentés ci-dessous doivent être considérés d'un œil critique.

Les essais d'inactivation (KO) du gène *ccyA* par insertion d'une cassette dans des souches de cyanobactéries biominéralisantes (*Cyanothece* sp. PCC 7425 et *Synechococcus* sp. PCC 6312, calcyanine à domaine CoBaHMA) n'ont pas pu aboutir. Une des hypothèses serait que le gène *ccyA* est essentiel à ces bactéries, et que sa délétion entraîne la mort des mutants.

L'essai inverse, qui a consisté à exprimer le gène *ccyA* (calcyanine à domaine X ou domaine CoBaHMA) dans des souches ne formant pas d'iACC (*Synechocystis* sp. PCC 6803 et *Synechococcus elongatus* PCC 7942) n'a pas permis de voir l'apparition d'un phénotype de biominéralisation intracellulaire. Par contre, le métabolisme du calcium de ces cellules semblait impacté par cette transformation. En effet, les mutants de *S. elongatus* PCC 7942 incorporaient plus de calcium que la souche contrôle, avec une accumulation de calcium dans les polyphosphates. Si cela peut être le signe de cellules stressées ou mourantes, cela pourrait être aussi expliqué par le fait que la calcyanine a un impact sur l'import de calcium dans les cyanobactéries.

Enfin, ces expériences de génétique ont été l'occasion de tenter de localiser la calcyanine dans les cyanobactéries. Pour cela, la calcyanine a été exprimée fusionnée à l'eGFP (enhanced Green Fluorescent Protein), une protéine fluorescente. Avec ce type de fusion, il est possible de localiser la protéine dans une cellule vivante, par étude de la fluorescence au microscope optique. Dans la souche biominéralisante *Synechocystis* sp. PCC 6312, la calcyanine (à domaine CoBaHMA) fusionnée à l'eGFP semblait être localisée aux membranes. Par contre, dans la souche non-biominéralisante *Synechococcus elongatus* PCC 7942, l'expression des calcyanines de *Cyanothece* PCC 7425 (CoBaHMA-(GlyZip)₃) et *G. lithophora* (X-(GlyZip)₃) fusionnées à l'eGFP faisait apparaître une fluorescence diffuse dans le cytoplasme. Ni les interférences entre la fluorescence de la GFP et l'autofluorescence des cyanobactéries ni l'impact de l'eGFP sur la calcyanine n'ont été quantifiés dans ces expériences, ce qui limite leur interprétation. Toutefois, ces résultats pourraient indiquer que la calcyanine à domaine CoBaHMA est localisée à la membrane, mais qu'elle a besoin de partenaire pour être adressée correctement (P. S. Görden, 2020, Thèse).

Au risque de se répéter, ces résultats doivent être considérés avec un regard critique car les expériences ont été complexes, et tout les contrôles n'ont pas pu être effectués. Néanmoins cette approche génétique dessine pour la calcyanine à domaine CoBaHMA, l'image d'une protéine vitale pour les cyanobactéries biominéralisantes, localisée aux membranes, en interaction avec des partenaires, et impliquée directement ou indirectement dans l'accumulation du calcium.

Toujours en relation avec ces aspects génétiques, la co-occurrence de *ccyA* avec certains autres gènes a été étudiée. Comme nous l'avons décrit plus en amont dans la thèse, les processus de biominéralisation nécessitent un grand nombre de protéines, notamment impliquées dans le transport d'ions. De Wever *et al.* ont montré que dans les génomes de 7 souches biominéralisantes, se trouvaient systématiquement les gènes de plusieurs antiporteurs : un échangeur $\text{Ca}^{2+}/\text{H}^{+}$, UPF0016 un gène codant possiblement pour un antiporteur $\text{Ca}^{2+}/\text{Cation}$ et un échangeur $\text{Na}^{+}/\text{H}^{+}$ qui pouvait se comporter comme un antiporteur $\text{Ca}^{2+}/\text{H}^{+}$ à pH élevé. De même, le gène d'un canal mécano-sensible a été répertorié (De Wever *et al.*, 2019). Une recherche dans 602 génomes cyanobactériens a montré la co-occurrence de la calcyanine avec un antiporteur $\text{Ca}^{2+}/\text{H}^{+}$ ainsi que le SLC26 BicA un transporteur d'ions HCO_3^- dépendant du Na^{+} (Benzerara *et al.*, 2022). A notre connaissance, BicA n'a encore jamais été reliée à un processus de biominéralisation contrôlé. Par contre, c'est un transporteur qui a été identifié comme étant impliqué dans les processus de concentration du carbone pour la photosynthèse (Price & Howitt, 2011). De même sur le méga plasmide de *Fischerella* sp NIES-4106, la calcyanine est localisée dans une région où l'on trouve un antiporteur du calcium, des ATPases transporteur de calcium ainsi qu'une anhydrase carbonique, des protéines qui, comme décrit précédemment, peuvent être liées à la biominéralisation (Benzerara *et al.*, 2022).

Projet de thèse.

Pour résumer ce qui a été présenté ci dessus, les études menées jusqu'à présent dans l'équipe ont pu montrer, pour la première fois, l'existence d'un processus de biominéralisation intracellulaire chez certaines souches de cyanobactéries. L'observation de la précipitation de carbonate de calcium dans un environnement, le cytosol, en pratique sous-saturé, a guidé les chercheurs vers la recherche de gènes responsables de ce phénotype. L'analyse de génomique comparative a, en effet, permis d'associer la calcyanine, codé par le gène nommé *ccyA*, à la présence d'iACC. La découverte de *ccyA* chez les souches biominéralisantes et l'observation de iACC dans des souches porteuse du gène *ccyA* dans leur génome, a permis, sans équivoque, d'identifier la calcyanine comme le marqueur de la biomineralisation intracellulaire des cyanobactéries (Benzerara *et al.*, 2022).

Comme décrit dans cette introduction, la biominéralisation du CaCO_3 implique un grand nombre de protéines avec des fonctions distinctes. De fait, comme résumé en Figure 9, il existe beaucoup d'hypothèses fonctionnelles pour la calcyanine.

Les études de génétique réalisées dans le but de comprendre le rôle physiologique de la production d'iACC dans ces souches, semblaient indiquer que la calcyanine a un rôle dans l'accumulation intracellulaire du calcium. Cela expliquerait le lien qui semble être présent entre le domaine X et la sélectivité dans l'incorporation des alcalino-terreux.

L'analyse génomique de l'environnement génomique des gènes *ccyA* a montré la co-occurrence de la calcyanine avec des protéines de transport du calcium et du carbonate. La présence dans le génome des souches biominéralisantes de protéines d'import du calcium indique que cette fonction est déjà assurée pour ces souches, et que donc ce n'est probablement pas là le rôle de la calcyanine.

Cette dernière pourrait alors être plus impliquée dans le transport intracellulaire, par exemple pour emmener les alcalino-terreux de la membrane jusqu'au site de nucléation ou encore pour importer les alcalino-terreux dans les organelles où a lieu la précipitation. Enfin, comme le gène *ccyA* est spécifique

de la formation d'iACC qui sont entourés d'une enveloppe, possiblement protéique, la calcyanine pourrait former cette enveloppe.

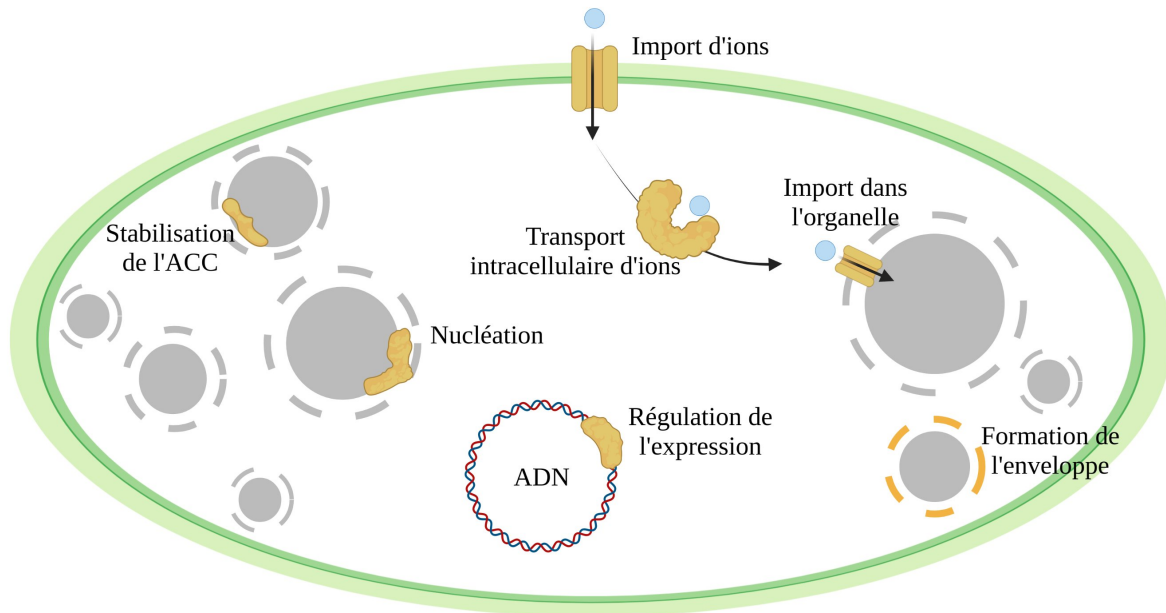


Figure 9 : Hypothèses sur la fonction de la calcyanine.

En jaune sont illustrées quelques hypothèses pour la/les possible(s) fonction(s) de la calcyanine dans la biominéralisation. Figure produite avec www.BioRender.com.

Dans cette thèse, nous avons cherché à poursuivre et étendre l'étude de la calcyanine pour apporter des éléments pour comprendre sa fonction, en nous intéressant à ses caractéristiques structurales.

Nous nous sommes concentrés sur un seul des 4 variants de la calcyanine, celui portant un domaine CoBaHMA. En effet, c'est le plus répandu et *a priori* le plus ancestral des 4 variants de la calcyanine (Benzerara et al., 2022). Son étude devrait apporter des réponses sur le mécanisme moléculaire à l'œuvre dans ce phénotype de biominéralisation ainsi que son évolution. Parmi les souches de cyanobactéries présentant ce variant, nous avons étudié *Synechococcus calcipolaris*, une souche isolée depuis des microbialites du lac Alchichica, un lac alcalin situé au Mexique. C'est une des premières souches biominéralisantes à avoir été décrite, elle était donc déjà bien connue dans notre laboratoire. Ses iACC sont localisés au pôle de la cellule (Benzerara et al., 2014), et elle ne présente pas de sélectivité particulière vis-à-vis des alcalino-terreux.

Pour étudier cette calcyanine nous avons fait dialoguer 2 approches complémentaires relevant de la bioinformatique et de la biochimie.

D'un point de vue bioinformatique, nous nous sommes tout d'abord attachés à prédire la structure 3D de la calcyanine de *S. calcipolaris* et de ses différents domaines. Des approches de modélisation comparative ont été mises en œuvre afin de réaliser un modèle de la structure 3D du domaine CoBaHMA, seul domaine de la calcyanine à présenter des similarités de séquences avec des structures 3D connues. Nous avons profité du développement de méthodes de prédiction de structure 3D basées sur l'intelligence artificielle, qui ont vu le jour durant le déroulement de cette thèse, pour compléter

cette première approche par plusieurs modélisations de la structure 3D de la calcyanine complète. L'analyse de ces modèles à la lumière de la conservation des séquences de cette famille de protéines nous a donné de premières informations et hypothèses quant aux positions critiques pour la structure et la/les fonction(s).

Cependant ces modèles sont incomplets, le modèle du domaine (GlyZip)₃ ne pouvant être prédit de manière complètement satisfaisante. De plus, ils ne fournissaient pas d'information sur l'état oligomérique ou les interactions entre les domaines CoBaHMA et (GlyZip)₃.

Nous avons alors cherché à suppléer à ces limitations par une approche expérimentale. Nous avons exprimé la calcyanine dans le système hétérologue *E. coli*, puis cherché à la purifier pour pouvoir résoudre sa structure expérimentale. La calcyanine entière s'est avérée trop hétérogène pour espérer former des cristaux, étape indispensable à la diffraction des rayons X. Par contre, par protéolyse limitée, et à l'aide du modèle de la structure 3D, nous avons identifié et isolé un fragment de la calcyanine qui s'est avéré beaucoup plus propice à la détermination structurale. L'expression et la purification de ce fragment ont permis d'aboutir à des cristaux très prometteurs.

Par ailleurs nous avons recherché la présence éventuelle du domaine CoBaHMA dans d'autres protéines, pour préciser le contexte biologique dans lequel la ou les fonction(s) du domaine CoBaHMA s'expriment. Ces recherches ont également bénéficié de l'apport des nouvelles méthodes de prédiction de structures 3D, reposant sur l'intelligence artificielle, qui fournissent des modèles de structures 3D à un niveau de précision inégalé à l'échelle de protéomes entiers. Ces modèles ont ainsi été largement utilisés pour accroître la spécificité des recherches de similarités, ainsi que pour décrypter plus en avant l'organisation modulaire des protéines de la famille CoBaHMA, et proposer de nouvelles hypothèses quant à la fonction de ce domaine.

Chapitre 1 : Analyse de la modélisation de la structure 3D de la calcyanine de *S. calcipolaris*.

Les modèles de la calcyanine complète de *S. calcipolaris* produits par AlphaFold2 et ESMFold, appuient nos premières modélisations par homologie et donnent un premier aperçu de l'ensemble de la structure 3D de la protéine, mais demandent à être complétés.

La modélisation de la structure 3D de la calcyanine de *S. calcipolaris* s'est faite en 2 temps. Dans un premier temps, courant 2020, nous avons établi un premier modèle par homologie (modélisation comparative) de son domaine CoBaHMA. L'absence de protéines dont les séquences sont similaires à celles du domaine (GlyZip)₃ et dont la structure 3D ait été résolue expérimentalement, nous a empêché de faire de même pour ce domaine. De fait, nous ne pouvions produire de modèle pour la séquence complète de la calcyanine.

Cependant, le champ de la modélisation des structures 3D de protéines a été révolutionné en 2021 par l'émergence de méthodes reposant sur l'intelligence artificielle, avec le logiciel AlphaFold2 (Jumper et al., 2021) qui permet de prédire les contacts entre résidus, à partir de coévolutions observées au sein d'alignement de séquences apparentées. D'autres logiciels comme RoseTTAFold (Baek et al., 2021) ou ESMFold (Lin et al., 2023) ont été développés par la suite, reposant sur des technologies différentes. Par exemple ESMFold ne nécessite pas d'alignement de séquences pour prédire une structure 3D, ce qui lui permet de modéliser des séquences qui ont peu ou pas de séquences apparentées, contrairement à AlphaFold2.

Nous avons alors modélisé la calcyanine complète à l'aide de 2 de ces logiciels : AlphaFold2 et ESMFold.

Par souci de synthèse et de clarté, j'ai choisi de présenter ces résultats de modélisation dans un ordre non chronologique. Je commencerai par présenter les modèles de la structure complète, pour ensuite me focaliser sur les domaines, avec, pour le domaine CoBaHMA, une mise en regard de la modélisation comparative avec celles réalisées par intelligence artificielle.

Nous avons donc modélisé la structure 3D de la calcyanine de *S. calcipolaris* avec 2 logiciels : AlphaFold2 (Figure 10.A) et ESMFold (Figure 10.C). Comme anticipé par analyse des séquences, la structure 3D de la calcyanine est modélisée sous la forme de 2 domaines structurés (CoBaHMA et (GlyZip)₃) et d'une queue désordonnée en C_{ter} (Benzerara et al., 2022). Avant d'analyser les caractéristiques de ces modèles, nous avons tout d'abord évalué leur qualité selon 4 critères :

- _ Le Z-score fourni par ProSA (Wiederstein & Sippl, 2007), qui évalue le profil énergétique de la structure.
- _ Le pLDDT (predicted Local Distance Difference Test) (Jumper et al., 2021), qui évalue la confiance locale de chaque logiciel en sa prédiction.
- _ La stéréochimie des acides aminés fourni par PROCHECK (Laskowski et al., 1993).
- _ Le PAE (Mirdita et al., 2021) qui quantifie l'erreur de positionnement relatives des paires d'a.a de la séquence.

Les évaluations ProSA des 2 modèles de la calcyanine conduisent à des Z-scores (Sippl, 1993) de respectivement -6,02 (Figure 11.A) et -7,31 (Figure 11.B). Ces valeurs se situent dans la gamme de Z-score obtenus par des structures expérimentales de longueurs de séquence similaires, indiquant que ces modèles sont fiables d'un point de vue énergétique.

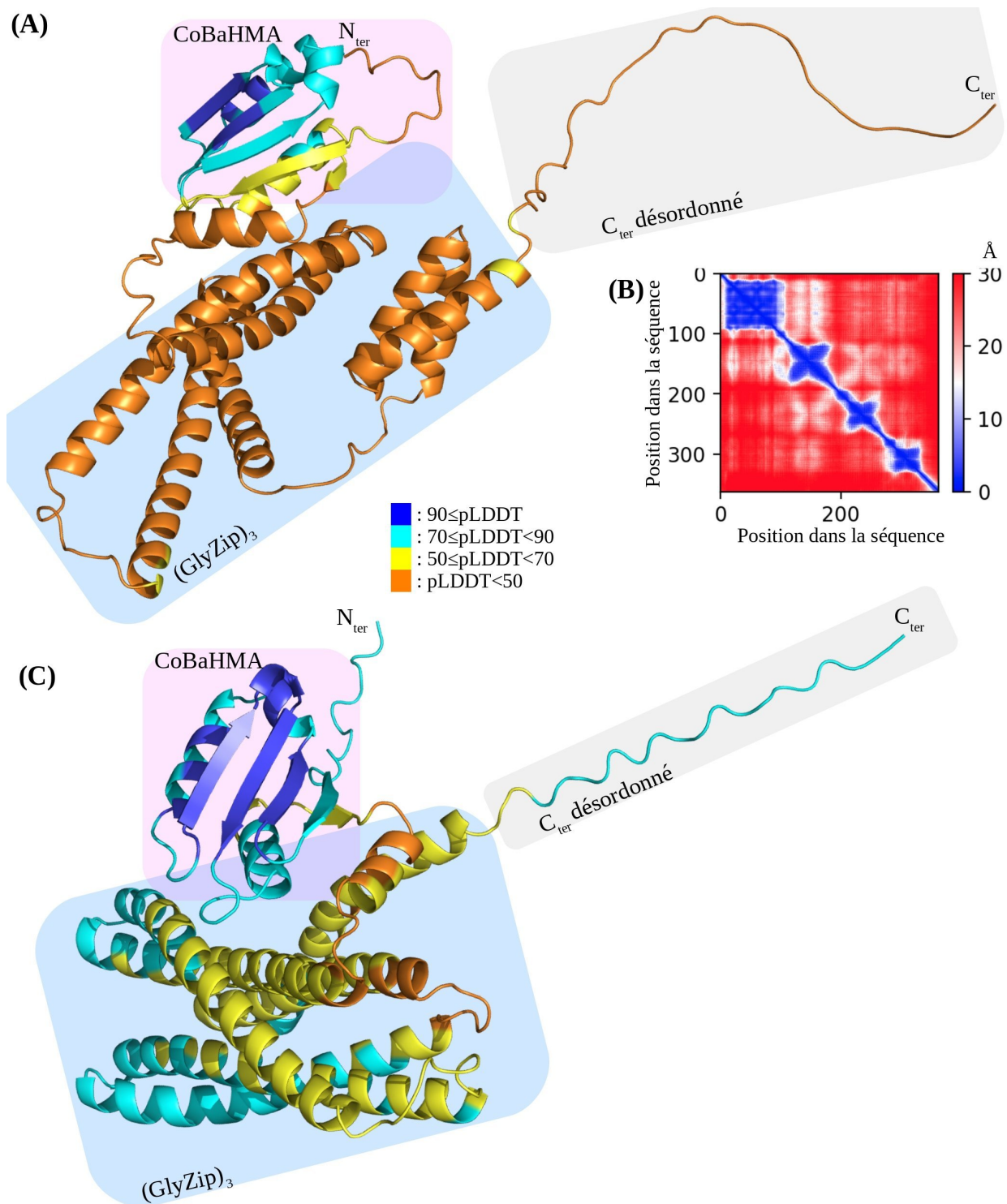


Figure 10 : Modèles de la structure 3D de la calyculine de *S. calcipolaris* conçus par AlphaFold2 et ESMFold.

Modèles conçus par (A) AlphaFold2 et (C) ESMFold, colorés selon le pLDDT. Le découpage en domaines est mis en évidence par des rectangles colorés. (B) PAE du modèle AlphaFold2 qui illustre l'erreur de positionnement relatif des a.a de la séquence 2 à 2, et permet de juger de la pertinence de la prédiction des contacts intra- et inter-domaines.

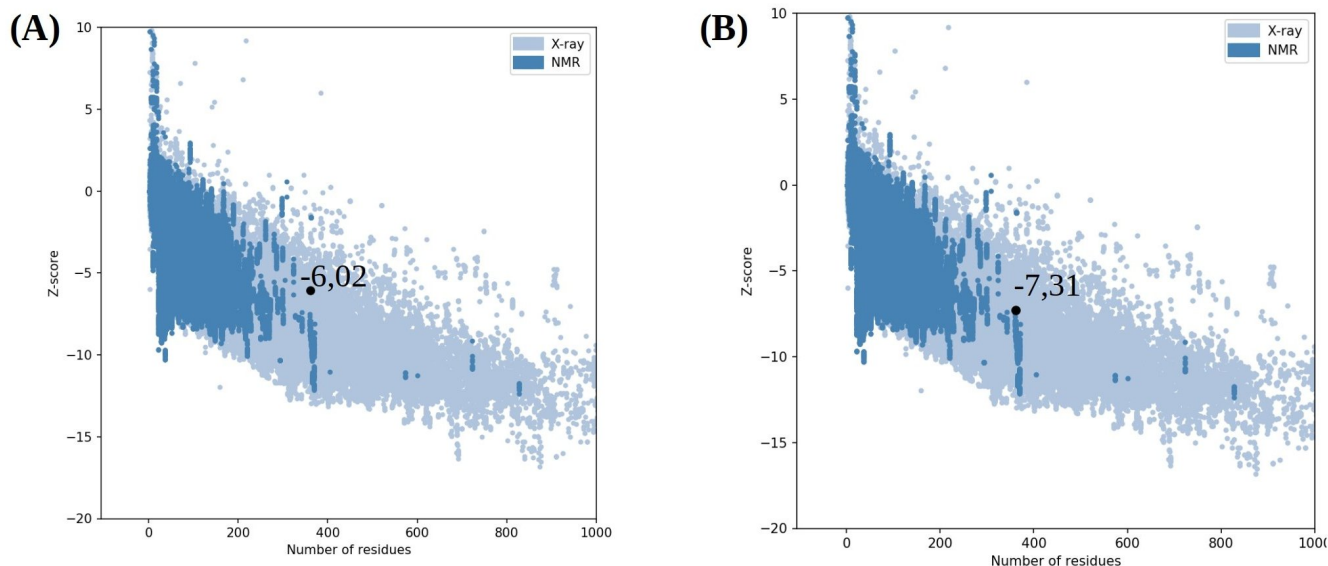


Figure 11 : Evaluations ProSA des modèles de la structure 3D de la calcyanine de *S. calcipolaris* conçus par AlphaFold2 et ESMFold.

Comparaison des Z-scores de structures expérimentales, avec les Z-scores des modèles conçus par (A) AlphaFold2 et (B) par ESMFold.

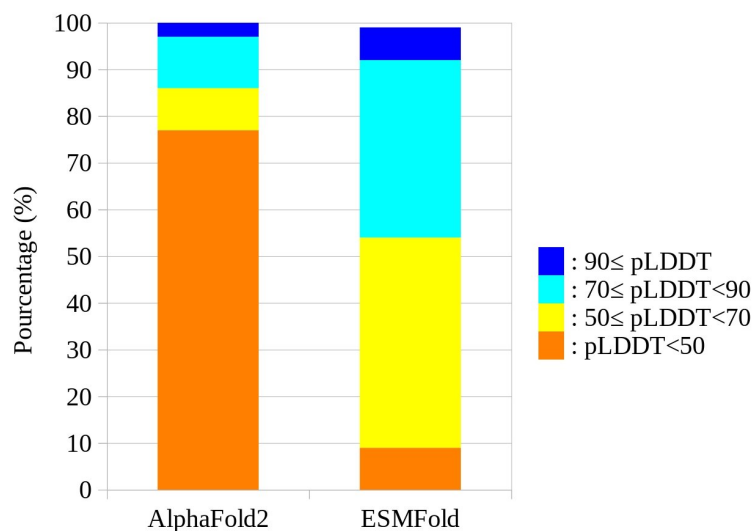


Figure 12 : Comparaison des distribution du pLDDT des modèles de la structure 3D de la calcyanine de *S. calcipolaris* conçus par AlphaFold2 et ESMFold.

Comme illustré en Figure 12, pour le modèle AlphaFold2, 77% des a.a se trouvent sous le seuil d'interprétabilité ($pLDDT < 50$). Ces a.a sont principalement sur le domaine (GlyZip)₃ pour lequel une structure ordonnée est proposée, mais dont la vraisemblance ne peut être étayée; et dans la région C_{ter} qui été prédite comme désordonnée. Pour le modèle ESMFold, seuls 9% des a.a se trouvent sous le seuil d'interprétabilité. Ces a.a se trouvent soit à proximité, soit dans des boucles désordonnées. Le domaine (GlyZip)₃ a un pLDDT très majoritairement > 50 , voir même > 70 . Étrangement, en dépit du fait que le désordre ait été relié à un $pLDDT < 50$ dans les modèles de structure 3D d'AlphaFold2 (Tunyasuvunakool et al., 2021; C. J. Wilson et al., 2022), dans ce modèle ESMFold, la partie C_{ter} est bien prédite désordonnée, mais avec un $pLDDT > 70$.

Selon le pLDDT, ESMFold semble donc plus confiant sur sa prédiction que ne l'est AlphaFold2. Il est possible d'avancer au moins 3 hypothèses pour expliquer cela.

La première, de loin la plus problématique, est que les valeurs de pLDDT ne sont pas comparables entre les 2 logiciels. Comme les deux logiciels reposent sur des méthodologies différentes, il est envisageable que ce calcul de pLDDT soit réalisé différemment dans les 2 cas. La gestion du C_{ter} désordonné est, à ce titre, particulièrement intéressante. Le désordre a été relié à un pLDDT < 50 chez AlphaFold2 (Tunyasuvunakool et al., 2021; C. J. Wilson et al., 2022), cependant le modèle ESMFold attribue un pLDDT > 70 à cette zone désordonnée. Cela semble indiquer une gestion différente de cette métrique dans les 2 logiciels. Donc il ne serait pas possible de l'utiliser pour comparer ces 2 modèles.

La question de la pertinence de l'utilisation du pLDDT comme outil de comparaison de modèles issus de logiciels différents a été déjà soulevée par Aubel *et al.* (Aubel et al., 2023) avec pour exemple la prédiction de la structure de la protéine Goddard avec 4 logiciels de prédiction (AlphaFold2, ESMFold, OmegaFold (R. Wu et al., 2022) et RGN2 (Chowdhury et al., 2022)). Dans cet exemple, en dépit de prédictions similaires, les pLDDT différaient significativement, interrogeant son utilité comme élément comparatif de prédictions opérées par différents logiciels (Aubel et al., 2023).

Une autre hypothèse repose sur l'ensemble d'entraînement d'ESMFold. ESMFold a été entraîné sur les données d'entraînement d'AlphaFold2 (des structures 3D expérimentales issues de la banque de structures 3D expérimentales PDB (Protein DataBase), antérieures à Avril 2018), auxquelles ont été ajoutés 12 millions de modèles AlphaFold2 de séquences issues de UniRef 50 (Hsu et al., 2022; Lin et al., 2023, 2023 supplementary data). Dans le cas d'AlphaFold2, il a été montré que le pLDDT était, en moyenne, supérieur lorsque la protéine à modéliser avait un homologue de structure 3D connue dans la PDB. Dit autrement, le pLDDT d'un modèle augmente quand AlphaFold2 a déjà été entraîné sur la structure 3D de la protéine à modéliser (D. T. Jones & Thornton, 2022). Pour l'heure, il n'y a pas eu de démonstration similaire pour ESMFold, mais il est probable qu'un effet semblable soit observable. Comme évoqué plus haut, nos recherches pour identifier des homologues au domaine (GlyZip)₃ dans la PDB n'ont pas abouti. AlphaFold2 n'a donc pas pu être entraîné sur une structure 3D expérimentale de référence pour ce domaine. Cependant, il pourrait y avoir des modèles de longs glycine zippers dans l'ensemble d'entraînement d'ESMFold incluant des modèles d'AlphaFold2. Cela permettrait à ESMFold de partir d'une base connue pour construire son modèle, et augmenterait sa précision. Par exemple, les protéines possédant un domaine PdsO, qui présente des similarités de séquences avec le GlyZip2 des calcyanine (Benzerara et al., 2022), sont présentes dans UniRef50 et pourraient avoir été intégrées en tant que modèles AlphaFold2 à cet ensemble d'entraînement.

Enfin, la dernière hypothèse repose sur le fait qu'AlphaFold2 requiert un MSA (Multiple Sequence Alignment) pour pouvoir modéliser une structure 3D quand ESMFold se contente de la séquence seule. En l'absence de MSA, les performances d'AlphaFold2 chutent drastiquement (Jumper et al., 2021; Lin et al., 2023). Le manque de séquences similaires au domaine (GlyZip)₃ des calcyanines ne permet pas d'avoir un alignement suffisamment profond pour ce domaine, ce qui limite certainement les capacités d'AlphaFold2. ESMFold n'est pas affecté par cette limite, ce qui pourrait justifier sa plus grande confiance en sa prédiction.

L'évaluation par PROCHECK de la stéréochimie de ces modèles, reposant sur la considération des angles φ - ψ associés à chaque a.a, a révélé des imperfections apparentes dans la prédiction (Figure 13.A). Selon PROCHECK, un bon modèle doit avoir 90% de ses a.a dans la région les plus favorisées du diagramme de Ramachandran. Or, pour le modèle construit par AlphaFold2 seul 72,8% des a.a sont

dans ces régions. De plus 4,9% des a.a sont dans la région « Interdite ». De même pour le modèle ESMFold, 85,4% des a.a sont dans des régions plus favorisées et 0,3% dans la région « Interdite ».

Toutefois une étude plus approfondie des diagrammes de Ramachandran de ces deux modèles de structure 3D de la calcyanine (Figure 13.B&C) révèle que les zones structurées (CoBaHMA et les GlyZips) ont une très bonne stéréochimie, mais que les zones désordonnées ont de nombreux a.a dans les zones interdites du diagramme. Il est à noter que les logiciels AlphaFold2 et ESMFold reposent sur un apprentissage vis à vis des structures déjà connues, sans l'implication de modèle physique. Dès lors quand ces logiciels sont confrontés à des séquences pour lesquelles leur apprentissage n'est pas adapté, notamment pour les segments désordonnées, il est probable qu'ils agissent de manière incohérente, sans respect de physique et la stéréochimie. Pour AlphaFold2, cela se traduit par une distribution des a.a dans le diagramme de Ramachandran de plus en plus éloignée de celle attendue à mesure que le pLDDT diminue (D. T. Jones & Thornton, 2022).

Il est, là encore, pertinent de noter qu'en dépit d'un pLDDT > 70, la partie désordonnée du modèle ESMFold a une mauvaise stéréochimie, ce qui renforce l'idée qu'AlphaFold2 et EMSFold ne traitent pas le pLDDT de la même manière. De même il est intéressant de remarquer que les GlyZips prédits par AlphaFold2 ont beau avoir un pLDDT < 50, leur stéréochimie est correcte. Il y a donc probablement une base solide à cette prédiction.

Enfin, le dernier critère d'évaluation utilisé pour ces modèles est le PAE (Predicted Aligned Error), une prédiction de l'erreur de positionnement relative de tous les couples d'a.a de la séquence (Mirdita et al., 2021). Celui-ci, uniquement disponible pour le modèle d'AlphaFold2, indique des erreurs de positionnement relatives faibles au sein des GlyZip et du domaine CoBaHMA, mais une grande erreur entre ces différents éléments (Figure 10.B). Ainsi, le positionnement relatif des éléments de structures secondaires paraît fiable au sein du domaine CoBaHMA et des GlyZips, au contraire des positionnements des GlyZips entre eux, et du CoBaHMA vis à vis du domaine (GlyZip)₃. Il n'est donc pas possible de prédire les interactions inter-domaines, ni inter-GlyZips dans ce modèle.

De cette première analyse, nous pouvons constater qu'ESMFold et AlphaFold2 fournissent tous 2 des modèles cohérents de la structure des différents domaines de la calcyanine, avec des valeurs de pLDDT élevées pour le domaine CoBaHMA, et plus faibles pour le domaine (GlyZip)₃. Néanmoins, afin de pleinement évaluer la pertinence de ces modèles nous nous attacherons ci-après à une analyse détaillée des 2 domaines de la calcyanine.

Comme ESMFold est globalement plus confiant qu'AlphaFold2 dans sa prédiction, nous nous servirons principalement de son modèle pour illustrer la suite de cette thèse.

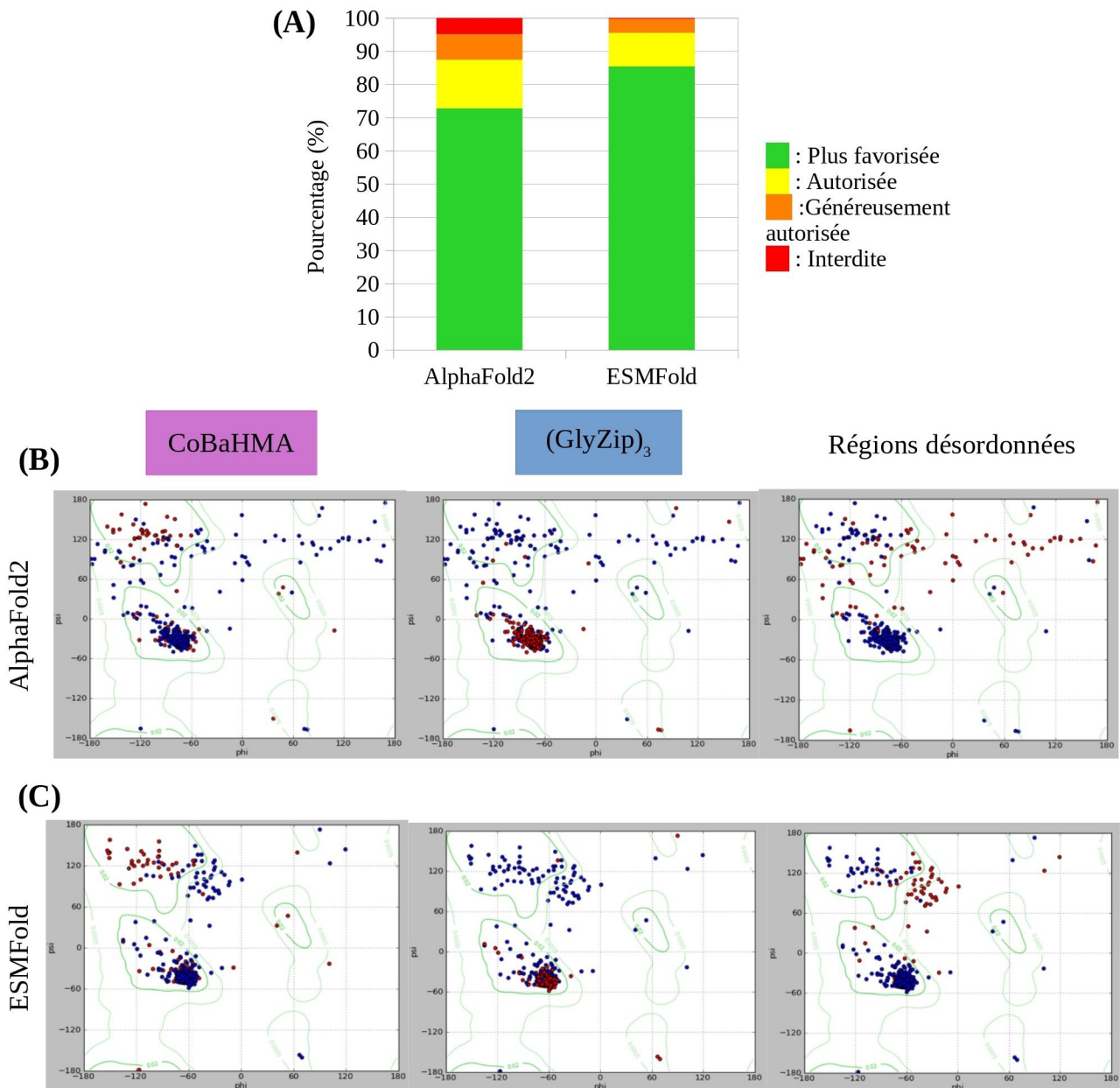


Figure 13 : Evaluations PROCHECK de la stéréochimie des modèles de la structure 3D de la calcyanine de *S. calcipolaris*.

(A) Comparaison de la répartition des a.a dans les différentes régions du diagramme de Ramachandran, pour les 2 modèles. (B-C) Diagrammes de Ramachandran de chacune des régions des modèles conçus par (B) AlphaFold2 et (C) ESMFold. En rouge sont colorés les a.a du domaine (de gauche à droite) CoBaHMA ([15-104]) / des motifs GlyZip ([116-185], [198-263], [274-326]) / des régions désordonnées (pour AlphaFold 2 : [1-14], [105-116], [187-195], [265-274], [324-362]; pour ESMFold : [342-362]). Les a.a des autres parties sont colorés en bleu.

Le domaine CoBaHMA est une nouvelle famille dans la superfamille HMA, au sein du repliement ferredoxin-like et présente une signature structurale et fonctionnelle unique.

AlphaFold2 et ESMFold proposent des modèles semblables du domaine CoBaHMA de la calcyanine de *S. calcipolaris*. En amont de l'utilisation de ces outils d'intelligence artificielle, nous avons par ailleurs construit un modèle de la structure 3D du le domaine la calcyanine de CoBaHMA de *Synechococcus* RS9917 par modélisation comparative. En effet, les recherches réalisées avec l'outil HH-PRED (alignement profils/profils) (Zimmermann et al., 2018), ont permis d'identifier des similitudes de séquence significatives entre le domaine CoBaHMA des calcyanines et 3 familles de domaines de la superfamille HMA (Heavy Metal Associated) : la famille HMA, la famille iHMA et la famille YAM. Les taux d'identité de séquence observés sont compris entre 8 % et 18 % (Probabilités comprises entre 96.5 et 91.3) (Figure 14).

Les domaines HMA (Bull & Cox, 1994) (parfois nommés MBD pour Metal Binding Domains) sont des domaines cytosoliques que l'on peut trouver seuls (Hearnshaw et al., 2009), ou associés à d'autres domaines tels que les P1B-type ATPases (Gourdon et al., 2011). Ces domaines sont impliqués dans la régulation de la concentration de cations métalliques. Ils sont caractérisés par un motif conservé CxxC, parfois étendu à MxCxxC (Bull & Cox, 1994). Il a été montré que ce site pouvait interagir directement avec les ions Cu^+ (Hearnshaw et al., 2009) et les ions Zn^{2+} (Banci et al., 2003) (Figure 15.B).

Les domaines iHMA (pour integrated HMA, parfois nommés RATX1) sont des domaines cytosoliques que l'on trouve dans des protéines de résistance aux pathogènes chez les plantes. Les domaines iHMA permettent la reconnaissance de pathogènes (Cesari et al., 2013; De la Concepcion et al., 2018) (Figure 15.D).

Enfin, le domaines YAM est un domaine cytosolique du YajR, une protéine membranaire que l'on trouve dans des souches de bactéries Gram-négative (*E. coli*, *Acinetobacter calcoaceticus*, *Cronobacter turicensis*...), membre de la superfamille MFS (Major Facilitator Transporter). Son (ses) rôle(s) est/sont pour l'heure mal compris, mais il pourrait être impliqué dans la régulation de la fonction de YajR par oligomérisation et/ou en réponse à des variations de pH ou d'ions (D. Jiang et al., 2013, 2014) (Figure 15.C).

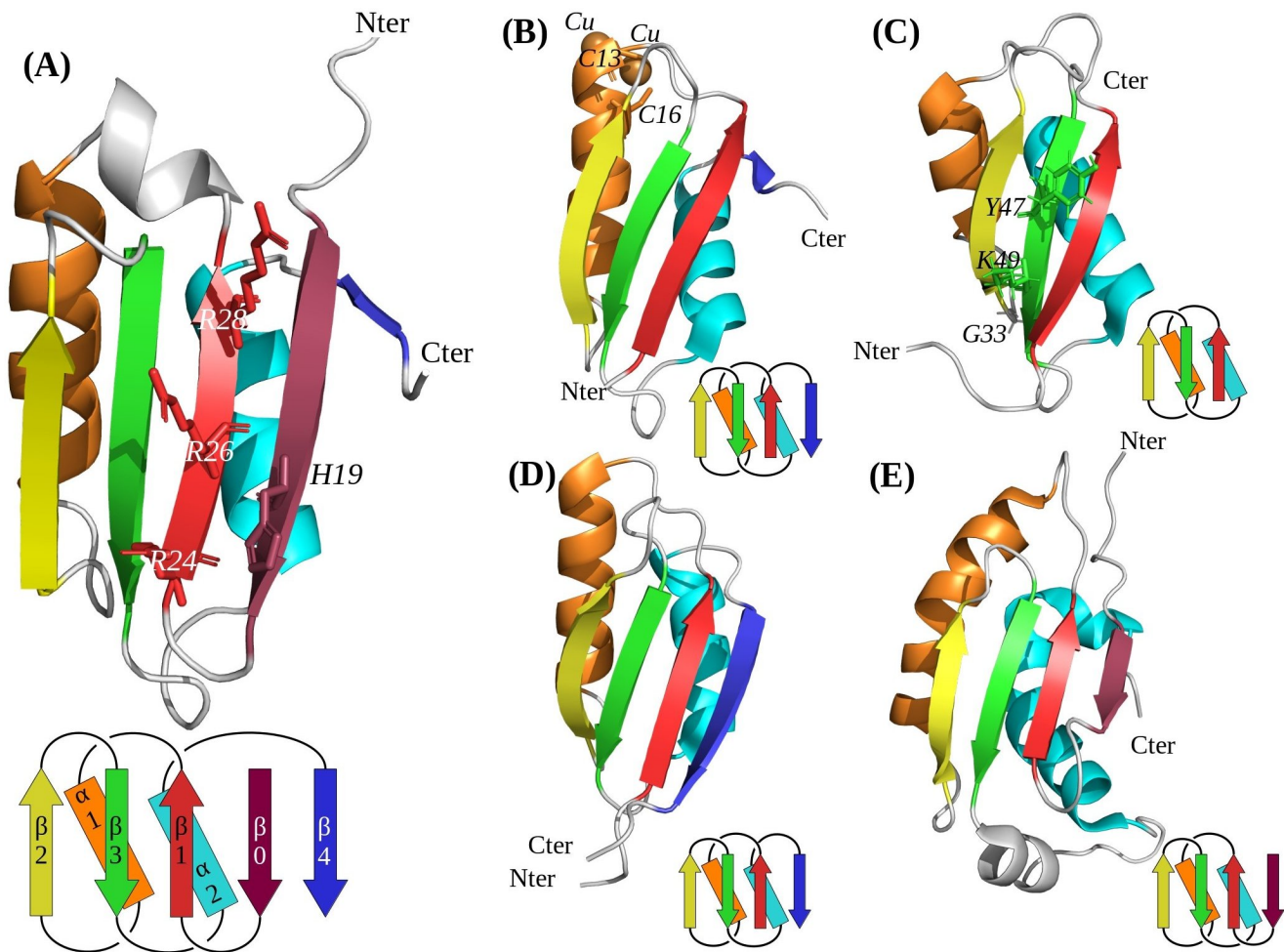


Figure 15 : Modèle 3D de la structure du domaine CoBaHMA, et structure expérimentale du domaine HMA, YAM, iHMA et au domaine N_{ter} de KipI.

(A) Modèle du domaine CoBaHMA construit avec ESMFold. La chaîne latérale des a.a basiques conservés est représentée. La topologie du domaine est affichée sous la structure 3D. (B) Structure et topologie du domaine HMA de la protéine CopZ (Hearnshaw et al., 2009; ID PDB RCSB : 2QIF). Ce domaine a la structure canonique du repliement « ferredoxin-like ». La chaîne latérale des cystéines du motif CxxC de liaison des ions Cu²⁺ typique des HMA est affichée. (C) Structure et topologie du domaine YAM de la protéine YajR (Jiang et al., 2014; ID PDB RCSB : 2RU9). La chaîne latérale des a.a conservés est affichée sur la structure. (D) Structure et topologie du domaine iHMA de la protéine Pikm-1 (De la Concepcion et al., 2018; ID PDB RCSB : 6FU9). (E) Structure et topologie du domaine N_{ter} de KipI (Jacques et al. 2010; ID PDB RCSB : 2KWA).

Ces 3 domaines adoptent un repliement « ferredoxin-like » caractérisé par un enchaînement de structures secondaires $\beta\alpha\beta\beta\alpha\beta$ (β_1 - α_1 - β_2 - β_3 - α_2 - β_4) avec une topologie de feuillet β $2\uparrow 3\downarrow 1\uparrow 4\downarrow$ (topologie basée sur (Richardson, 1981), les flèches indiquant l'orientation des brins). Il y a toutefois des variations entre ces familles, notamment au niveau du brin β_4 : le domaine YAM ne possède pas ce brin (D. Jiang et al., 2014), quand les iHMA (De la Concepcion et al., 2018) ont un β_4 très long et conservé. Sur la base de cette similitude de séquences, nous avons pu modéliser le domaine CoBaHMA avec la structure expérimentale de ces domaines comme référence grâce à Modeller (Webb & Sali, 2016). Cependant l'analyse du MSA du domaine CoBaHMA (Figure 14) indique la présence d'un élément de structure secondaire additionnel en amont du brin β_1 , incluant une histidine très conservée. Cette séquence N_{ter} du domaine CoBaHMA a été modélisée comme un brin β additionnel nommé β_0 , en prenant comme référence le domaine N_{ter} de KipI (ID PDB RCSB : 2KWA). En effet, des similitudes de séquence, plus faibles (12 %), ont également été détectées avec ce domaine grâce à l'outil HH-PRED. Le domaine N_{ter} de KipI est formé d'un enchaînement de structures secondaires $\beta\beta\alpha\beta\alpha$ (Jacques et al., 2011) (Figure 15.E), avec pour le feuillet β une topologie : $2\uparrow 3\downarrow 1\uparrow 0\downarrow$. Une superposition des structures 3D de KipI et du domaine YAM de YajR indique leur compatibilité (RMSD 2.1 Å sur 55 C α). La topologie du repliement est donc conservée, mais avec une transformation topologique simple (β_4 remplacée par le β_0) conduisant à des positions différentes des extrémités N- et C-terminales du domaine. Ce type de transformation est un processus fréquemment observé au cours de l'évolution des domaines de protéines (Grishin, 2001).

Le modèle par homologie ainsi obtenu a un enchaînement de structure secondaire $\beta\beta\alpha\beta\beta\alpha(\beta)$, avec un brin β_4 présent ou non selon les structures de référence utilisées. Le feuillet β adopte une topologie $2\uparrow 3\downarrow 1\uparrow 0\downarrow(4\downarrow)$.

Le même repliement, dérivé du repliement ferredoxin-like, est observé dans les 2 modèles AlphaFold2 et ESMFold, qui se superposent parfaitement (RMSD 0.33 Å sur les 80 atomes C α sur les a.a [12-93] qui couvrent tout le domaine CoBaHMA). Le pLDDT est > 50 pour tout le domaine dans ces deux modèles, avec une large portion d'a.a avec un pLDDT > 70, voir > 90 surtout dans le modèle ESMFold, ce qui indique une bonne, voire une très bonne confiance des logiciels dans leur modèle.

Toutefois les 3 modèles (par homologie, AlphaFold2, ESMfold) divergent sur les a.a présents sur le brin β_0 . Il y a un décalage d'un 1 a.a entre le modèle établi par modélisation comparative et les 2 autres, ce qui induit une différence sur le positionnement de l'histidine conservée du β_0 (Figure 14). La modélisation comparative l'oriente vers le cœur hydrophobe, quand les modèles conçus par l'intelligence artificielle l'orientent vers l'extérieur du domaine (Figure 15.A), ce qui paraît plus cohérent au regard des propriétés hydrophiles de cet a.a. L'utilisation d'AlphaFold2 et ESMFold a non seulement permis de confirmer le modèle obtenu par comparaison du domaine CoBaHMA, mais a aussi permis de l'affiner sur le brin β_0 .

Ces modèles, ainsi que les similitudes de séquence détectées lors de la recherche de moules expérimentaux pour la modélisation comparative nous permettent ainsi d'affirmer que le domaine CoBaHMA adopte un repliement « ferredoxin-like ». De plus, la proximité de séquence avec les familles HMA, iHMA et YAM le place dans la superfamille HMA (superfamille SCOP d58.17, <https://scop.berkeley.edu/sunid=55008>).

Cependant, le domaine CoBaHMA a une signature en a.a distincte des 3 autres familles, qui le place dans une famille indépendante. En effet, le domaine CoBaHMA n'a pas les signatures des autres familles de la superfamille HMA, comme le motif CxxC des HMA (Benzerara et al., 2022; Bull & Cox, 1994) ou le motif YxK conservé sur le brin β_3 des YAM (D. Jiang et al., 2014). Par contre, il présente un patch d'a.a basiques à la surface de son feuillet β , constitué d'une histidine sur le brin β_0 et de 3 arginines (plus rarement de lysines) sur le brin β_1 qui lui sont spécifiques (Figure 14). Il est donc

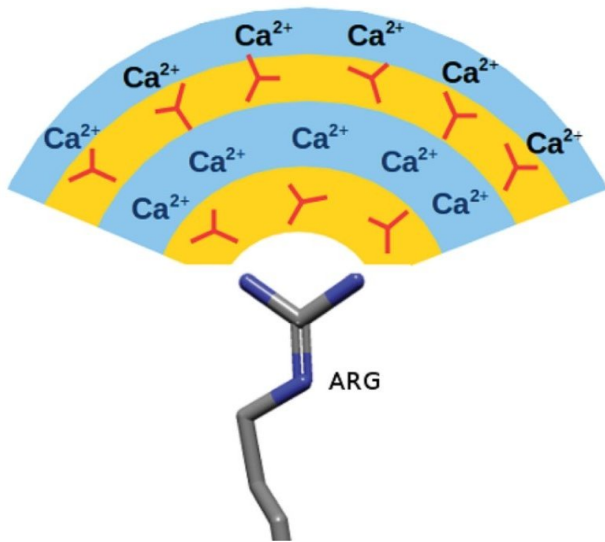


Figure 16 : L'arginine comme site de nucléation du carbonate de calcium.

Extrait de Rani and Saharay, 2019.

En rouge les molécules de carbonate.

protéine impliquée dans la biominéralisation des coquilles d'œuf, c'est cette interaction des ions carbonates avec un trio d'arginines spatialement proches qui serait à l'origine de la précipitation d'ACC (Amorphous Calcium Carbonate), possible précurseur de la calcite de la coquille (Rani & Saharay, 2019). La dynamique moléculaire a aussi montré une forte affinité des arginines pour les ACC (Innocenti Malini et al., 2017).

Nous pouvons donc supposer que le domaine CoBaHMA de la calcyanine interagit avec des ions carbonates via ses arginines conservées lors de la formation des iACC chez les cyanobactéries, par exemple pour favoriser leur précipitation.

L'histidine conservée pourrait elle aussi interagir avec les ions carbonates via sa charge positive. Néanmoins, il y a peu de données sur une possible interaction carbonate-histidine. Tout au plus, nous pouvons évoquer l'interaction de l'histidine avec la calcite (cristal de carbonate de calcium) reportée par Kumar *et al.*, mais dans un environnement acide (Kumar et al., 2020).

Cependant, ce n'est pas la seule hypothèse que nous pouvons formuler concernant cet ensemble d'a.a basiques conservés. En effet de nombreuses protéines interagissent avec les têtes chargées négativement de phospholipides via des a.a basiques (Lemmon, 2008; Stace & Ktistakis, 2006). Nous pouvons citer les domaines FYVE qui se fixent sur les phosphatidylinositol-3-phosphates via un motif RR/KHHCR (Kutateladze, 2006), ou encore le domaine PX de p47^{phox} qui interagit avec les acides phosphatidiques via une histidine, une arginine et une lysine (Karathanassis et al., 2002). La présence d'a.a basiques

probable que le domaine CoBaHMA a une fonction distincte de celles reportées pour les autres membres de la superfamille.

C'est cette analyse qui a conduit à l'appellation « CoBaHMA », pour « domain with CONserved BASic residues in the HMA superfamily ».

En considérant le phénotype de biominéralisation dans lequel est impliquée la calcyanine ainsi que les a.a conservés, nous pouvons poser quelques hypothèses concernant la fonction du domaine CoBaHMA.

Des simulations de dynamique moléculaire ont montré qu'en présence d'ions Ca^{2+} et CO_3^{2-} , les arginines avaient une forte proportion à interagir avec les ions carbonates (Innocenti Malini et al., 2017; Kumar et al., 2020; Rani & Saharay, 2019) (Figure 16). Dans le cas du lysozyme, une

conservés sur un feuillet β rappelle plus spécifiquement l'un des 2 sites d'interaction avec les phosphatidylserines et les phosphoinositides des domaines eucaryotiques C2, illustré sur la Figure 17 (Guerrero-Valero et al., 2009; Guillén et al., 2013). Nous pouvons ainsi poser l'hypothèse que le domaine CoBaHMA interagit avec un lipide chargé négativement. Comme le lipidome des cyanobactéries, contrairement à ceux des bactéries et des eucaryotes, est constitué quasiment exclusivement de glycérolipides (Boudière et al., 2014; Wada & Murata, 2004), ces lipides pourraient être le phosphatidylglycerol (PG) ou le sulfoquinovosyldiacylglycerol (SQDG). Éventuellement, cela pourrait aussi être les acides phosphatidiques (PA), dont la présence n'est pas systématiquement mesurée, mais qui sont chargés négativement et qui sont les précurseurs probables des autres glycérolipides des cyanobactéries (Lem & Stumpf, 1984; Murata & Nishida, 1987; Petroustos et al., 2014).

Grâce à cette modélisation, nous avons donc pu obtenir un modèle robuste de la structure 3D du domaine CoBaHMA. Ce modèle, ainsi que les recherches de similitudes de séquence, nous ont permis de placer le domaine CoBaHMA dans la superfamille HMA du repliement « ferredoxin-like », incluant les familles HMA, YAM et iHMA. La signature fonctionnelle du domaine CoBaHMA est unique dans cette superfamille, et suggère une interaction avec les ions carbonates ou les phospholipides.

Ces résultats ont été, pour partie, publiés en 2022 dans l'article « A New Gene Family Diagnostic for Intracellular Biomineralization of Amorphous Ca Carbonates by Cyanobacteria » par Benzerara, Gaschignard *et al.* que nous avons placé en Annexe 1 de cette thèse. Cela concerne la modélisation comparative et la modélisation par AlphaFold2 du domaine CoBaHMA, ainsi que la classification du domaine CoBaHMA dans la superfamille HMA du repliement « ferredoxin-like ». La modélisation par ESMFold a été réalisée ultérieurement.

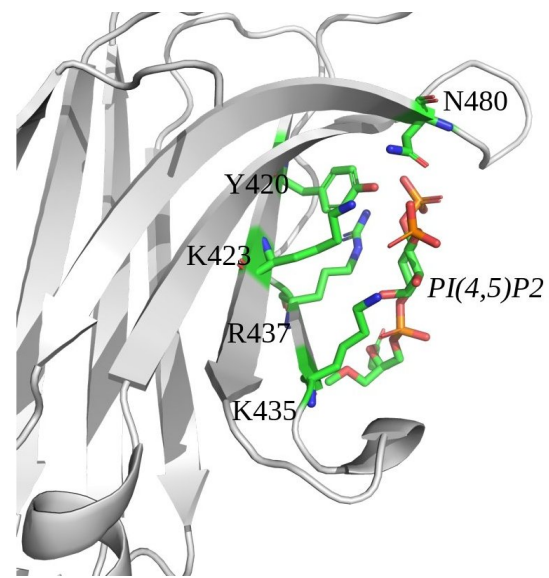


Figure 17 : Interaction entre le domaine C2 de la Rabphilin 3A, et les phosphates de PI(4,5)P2.

Inspiré de Guillén *et al.*, 2013. ID PDB RCSB : 4NS0.

Les a.a impliqués dans l'interaction avec PI(4,5)P2 sont colorés selon leur atome (C:vert; N:bleu; O:rouge; P:orange), et nommés.

Les GlyZips forment des hélices antiparallèles en épingle à cheveux, avec une face hydrophobe et une face hydrophile. Leurs orientations relatives au sein du domaine (GlyZip)₃ restent inconnues.

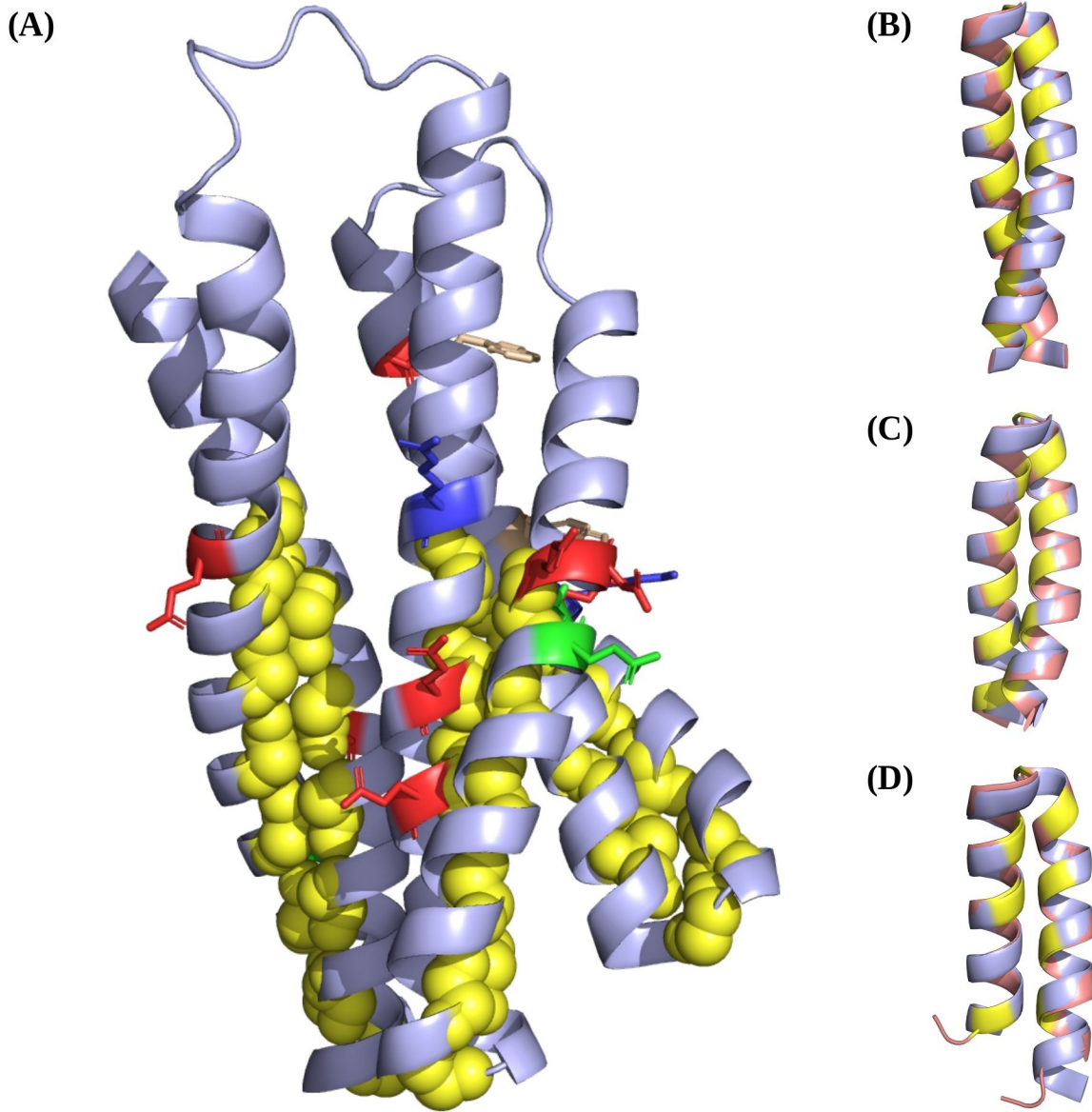


Figure 18 : Structure 3D du domaine (GlyZip)₃ modélisé par ESMFold.

(A) Domaine (GlyZip)₃ complet. Les glycines des GlyZip sont représentées par des sphères jaunes. Les chaînes latérales des a.a polaires conservés sont affichées, et colorées selon leur type (acide : rouge; basique : bleu; polaire non chargé : vert; aromatique : beige). Superposition des prédictions réalisées par ESMFold (bleu pâle) et AlphaFold2 (rouge pâle) avec les glycines colorés en jaune, pour (A) le GlyZip1 (RMSD sur les C α = 0.297Å); (B) le GlyZip2 (RMSD sur les C α = 0.299); (C) le GlyZip3 (RMSD sur les C α = 0.429).

Contrairement au domaine CoBaHMA, la structure 3D du domaine (GlyZip)₃ s'est avérée complexe à modéliser. Comme indiqué initialement, il n'a pas été possible de réaliser un travail de modélisation comparative pour le domaine (GlyZip)₃ faute de références structurales expérimentales. Le domaine (GlyZip)₃ a une signature de séquence unique : une triple répétition d'un long glycine zipper, où chaque répétition a ses conservations propres en dehors du motif commun répété périodiquement, constitué de résidus hydrophobes et de résidus de faible encombrement stérique (glycine, alanine). Il y a notamment plusieurs a.a acides hautement conservés aux extrémités de ces glycines zippers.

ESMFold et AlphaFold2 prédisent tout les 2 les 3 GlyZips comme des double hélices en épingle à cheveux (Figure 18.A). Ceci est, là encore, conforme aux prédictions que nous avons faites sur ce domaine, sur la base de l'analyse de tracés HCA (Benzerara et al., 2022). Si l'arrangement relatif des 3 GlyZips diffère entre ces deux modèles, les prédictions individuelles de chacun des GlyZips par les 2 logiciels sont parfaitement superposables (Figure 18.B&C&D), avec des RMSD (Root-Mean-Square-Deviation) sur C α < 0.43Å. La concordance de ces prédictions indique une robustesse de cette prédiction, en dépit de scores de confiance faibles (pLDDT < 70). Il est intéressant de noter que dans le cas du modèle produit par AlphaFold2, le pLDDT est < 50 sur une zone structurée, alors qu'il a été montré qu'un tel score de confiance était un bon indicateur de désordre (Tunyasuvunakool et al., 2021; C. J. Wilson et al., 2022). Les GlyZips sont donc un bonne illustration que le pLDDT n'est pas un indicateur suffisant pour prédire les zones désordonnées, ainsi que commenté par à l'échelle de protéomes entiers dans l'article de Bruley et al., 2022. Les faibles scores observés relèvent ici vraisemblablement d'un manque de données expérimentales relatives à cette structure lors des phases d'apprentissage, suggérant que l'on est en présence d'une structure 3D originale, non encore décrite. Les structures 3D prédites des 3 GlyZips individuels sont très proches, comme illustré pour le modèle ESMFold sur la Figure 19.A, avec un RMSD maximum de 0.75Å sur la chaîne carbonée, lors des comparaisons 2 à 2. L'écart le plus grand étant entre le GlyZip1 et le GlyZip3, qui correspondent aux GlyZips les plus éloignés en termes de séquence.

Comme dans de nombreux exemples motifs de glycine zippers plus courts (Kim et al., 2005; Martin et al., 2012), les glycines sont positionnées à l'interface entre les deux hélices (Figure 18.A). Le très faible encombrement stérique induit par l'absence de chaîne latérale carbonée chez la glycine, permet d'avoir un contact très proche entre les hélices. Ceci maximise les interactions de Van der Waals et facilite la formation de liaisons hydrogène, ce qui stabilise l'interaction entre les hélices (Teese & Langosch, 2015). Au niveau du coude entre les 2 hélices du motif GlyZip se trouve une proline fortement conservée, précédée d'une glycine.

L'analyse de la structure des GlyZips individuels, modélisés de façon cohérente par les approches AlphaFold2 et ESMFold, montre qu'en dépit de l'absence de références structurales expérimentales, les déterminants structuraux qui mènent au modèle de 2 hélices en épingle à cheveux, en orientation antiparallèle, sont forts.

Une recherche d'homologues structuraux du GlyZip1, qui est le plus régulier des 3 glycines zippers, a conduit à observer que celui-ci est structurellement similaire à 5 parties de protéines appartenant à 4 structures de la PDB100 (7ojf, structure de Slyb13-Bama d'*E.coli*; 7xdi, structure de la queue du bactériophage SSV19; 7a0g, structure du pore SmhB de *Serratia marcescens*; 6h2f, structure du pré-pore Ah1B de *Aeromonas hydrophilia*), avec toutefois des e-value $\geq 8,76e-1$ et des identités de séquence $\leq 31,8\%$.

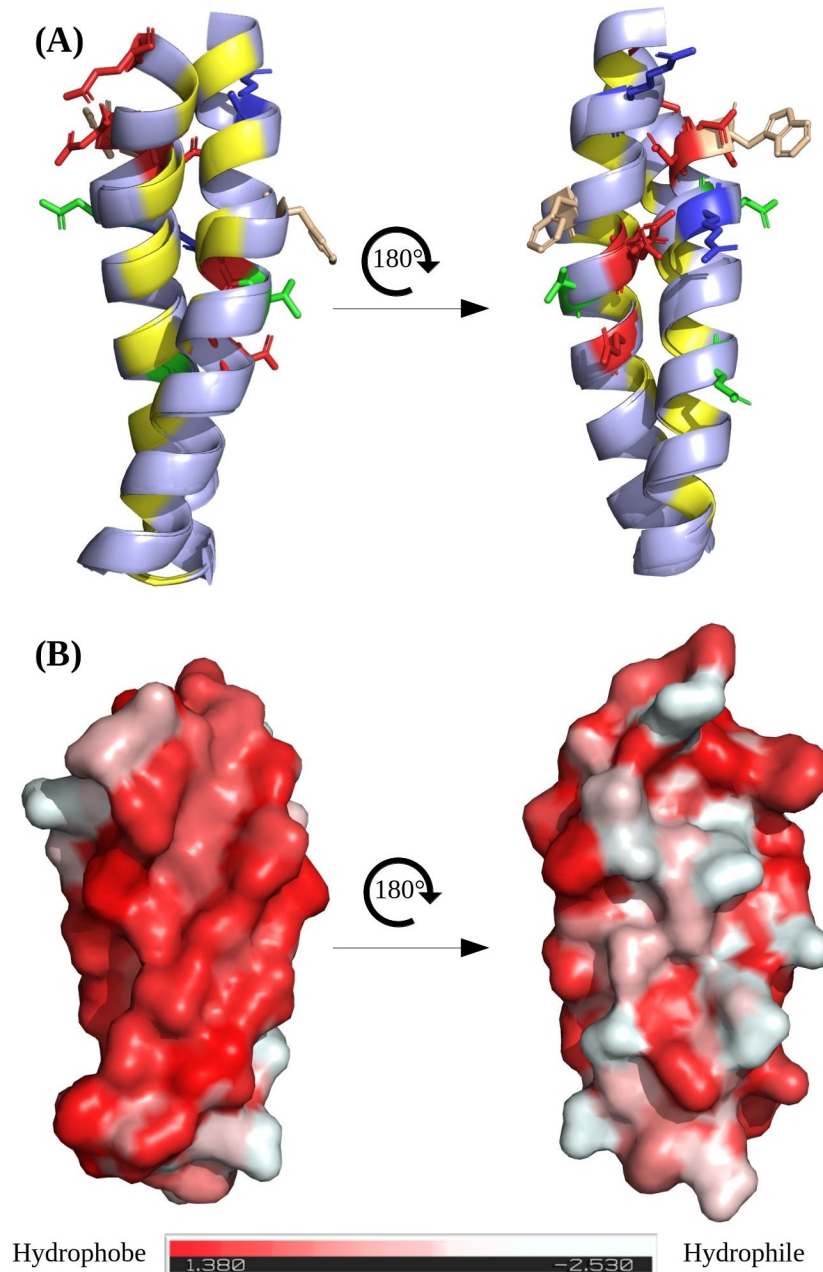
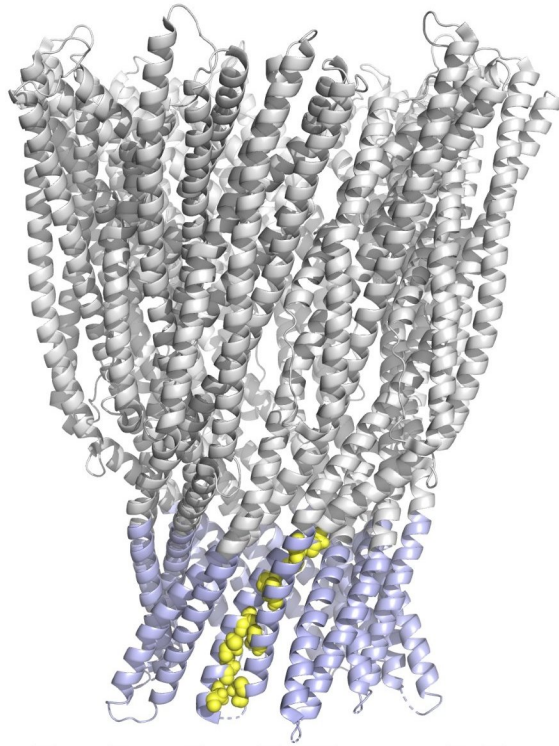


Figure 19 : Superposition de la structure 3D des 3 GlyZips individuels modélisées par ESMFold, et hydrophobicité des GlyZips.

(A) Les 2 faces de la superposition des structures des 3 GlyZips individuels modélisées par ESMFold. Les glycines des GlyZips sont colorées en jaune. Les chaînes latérales des a.a polaires conservés des GlyZips sont affichées, et colorées selon le type d'a.a (acide : rouge; basique : bleu; polaire non chargé : vert; aromatique : beige). (B) Surface des modèles représentés en A., colorée selon l'hydrophobicité des a.a.

Ces 5 chaînes polypeptidiques possèdent toutes un « small zipper » étendu de taille similaire à GlyZip1, avec des répétitions périodiques d'a.a peu encombrants mais sans les signatures polaires conservées des glycines zippers des calcyanines, ni le motif GP présent sur son épingle à cheveux. Le small zipper est une version plus générale du glycine zipper : tout comme le glycine zipper, il est basé sur un motif de type *xxx* avec * = G/S/T/A/V, des petits acides aminés. Ses propriétés sont souvent considérées comme similaires à celle du glycine zipper (Kim et al., 2005; Teese & Langosch, 2015). L'une des ces structures est illustrée en Figure 20.A.

(A)



(B)

GlyZip1 : GFVVGGQIG--DVVGGVVGGTAGGVF-MGPAGMLMGAQVGTF--VGGVIGGRLG
 7a0g : GAIAGIVVGGLLVIGGAIIVTAIGAVAGLTSTPVVMGGIAMMTAGAGGVVIGGAIV

(C)

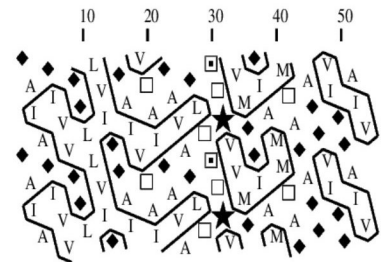


Figure 20 : Structure de la partie B de SmhABC de *Serratia marcescens*.

(A) La structure est un decamère. En bleu, la partie de la partie de la structure identifiée comme similaire au GlyZip1 de la calcyanine par FoldSeek. Sur un des monomères est représenté, par des sphères jaunes, le small zipper qui fait l'interface entre les 2 hélices en épingle à cheveux. ID PDB RCSB : 7a0g. (B) Alignement de la séquence de 7a0g et de GlyZip1. Les positions identiques sont écrites en blanc sur fond noir. (C) Tracé HCA de la portion de 7a0g qui s'aligne sur le domaine GlyZip1.

La modélisation des GlyZips sous forme de double hélices en épingle à cheveux paraît donc cohérente avec ces structures expérimentales. Cette cohérence est remarquable dans le cas d'AlphaFold2, car ce dernier n'a pas pu être entraîné sur les structures mentionnées qui ont été publiées après Avril 2018, date limite pour faire partie du set d'entraînement (Jumper et al., 2021). Là encore, c'est un signe que les déterminants structuraux qui mènent à la modélisation en double hélices en épingle à cheveux sont forts.

La double hélice des GlyZips présente une face hydrophobe et une face hydrophile (Figure 19.B). Les a.a chargés/polaires conservés qui forment la signature de chacun des glycines zippers, sont tous sur la face hydrophile de la double hélice. Il est probable que cette face ait une importance fonctionnelle alors que la face hydrophobe remplit plus vraisemblablement un rôle dans l'architecture de cette région. Les modèles complets du domaine (GlyZip)₃ tels que proposés par ESMFold et AlphaFold2, exposent de larges portions de ces faces hydrophobes au solvant. Dans le cas d'ESMFold, les GlyZip sont assemblés en faisceau avec les faces hydrophobes orientées vers l'intérieur, tandis qu'AlphaFold2 n'aboutit pas de manière fiable à un assemblage. Les positionnements relatifs des différents GlyZip est donc probablement erronés. C'est d'ailleurs supporté dans le cas du modèle AlphaFold2 par le PAE associé. Il indique clairement de grandes incertitudes sur le positionnement relatif des différents GlyZips. Il nous manque donc des informations vis à vis de cet assemblage, nous empêchant par conséquent d'aboutir à un modèle complet du domaine (GlyZip)₃.

Pour modéliser correctement ce domaine, il manque probablement des partenaires. Sur les 5 homologues structuraux détectés, 3 sont des small zippers localisés dans une membrane (structure : 7aog (2 small zippers); 6h2f (1 small zipper)). Dans ces 3 résultats, plusieurs unités de small zippers de double hélices en épingle à cheveux, sont assemblés et forment la partie membranaire d'un pore (Churchill-Angus et al., 2021; J. S. Wilson et al., 2019). Ces exemples illustrent 2 potentiels partenaires d'interaction qui pourraient manquer pour aboutir à un modèle pertinent du domaine (GlyZip)₃ : les lipides et d'autres unités de GlyZip.

Premièrement, le domaine (GlyZip)₃ de la calcyanine pourrait nécessiter un environnement membranaire pour se structurer en un assemblage cohérent, par exemple pour orienter les faces hydrophobes des GlyZips vers l'extérieur de la protéine, contrairement ce que fait actuellement ESMFold et AlphaFold2. Le motif GxxxG est un motif surreprésenté dans les hélices transmembranaires (Senes et al., 2000), et il a été abondamment décrit pour son importance dans leurs interactions (Teese & Langosch, 2015). Par ailleurs il a été montré sur plusieurs exemples qu'en dépit de leur prétention, AlphaFold2 et ESMFold pouvaient avoir de réels problèmes à modéliser les protéines membranaires, précisément car ils ne peuvent produire les contraintes issues des interactions lipides-protéines (Azzaz et al., 2022; Téletchéa et al., 2023). Ceci expliquerait leurs difficultés communes à modéliser le domaine (GlyZip)₃. Cette hypothèse que le

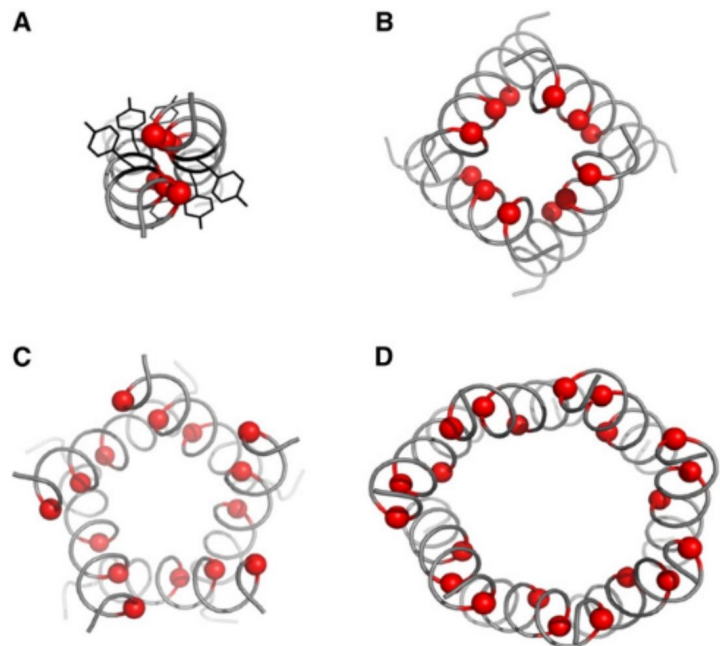


Figure 21 : Exemples d'assemblage de glycine zippers

Extrait de Martins *et al.*, 2012.

En gris la chaîne carbonée. En rouge, les glycines. Les ID PDB RCSB sont écrits entre parenthèses. (A) Survival Motor Neurons YG-box (4GLI). (B) Hélices formant le pore du canal à potassium KcsA (1BL8). (C) Canal mécano-sensible MscL (1MSL). (D) Canal mécano-sensible MscS (1MXM).

domaine (GlyZip)₃ serait membranaire va de pair avec l'hypothèse que le domaine CoBaHMA interagirait avec des lipides via ses a.a basiques conservés.

Secondement, les GlyZips du domaine (GlyZip)₃ pourraient s'assembler avec d'autres GlyZips par la formation d'homooligomères de calcyanine. En effet, comme évoqué précédemment, les glycines zippers ont été abondamment décrits comme motifs d'interaction pour des hélices α transmembranaires ou cytosoliques. Un certains nombres d'hélices α porteuses de glycine zipper interagissent entre elles, soit en mettant face à face 2 glycine zippers (Figure 21.A) pour former des dimères; soit en positionnant la face porteuse du glycine zipper d'une hélice sur la face opposée au glycine zipper de l'autre, pour former des pores (Figure 21.B-D) (Kim et al., 2005; Martin et al., 2012; Teese & Langosch, 2015). Des essais de modélisation en trimère de la calcyanine ont été menés avec ESMFold et AlphaFold2 pour tester cette hypothèse. Malheureusement, ils n'ont pas abouti, les modèles ne proposant aucune interaction entre les différents monomères de calcyanine.

En l'absence de modèle pertinent pour l'assemblage du domaine (GlyZip)₃, il n'est pour l'heure pas possible de formuler efficacement des hypothèses quant à la fonction de ce domaine. Il est toutefois pertinent de mentionner que la séquence du domaine (GlyZip)₃ est riche en glycines, mais aussi en acides aspartiques et glutamiques. Or, c'est une caractéristique récurrente des protéines impliquées dans la biominéralisation du carbonate de calcium que d'être enrichies en a.a acides et en a.a petits (Aizenberg et al., 1996; Kalmar et al., 2012; Laipnik et al., 2020). Ces protéines peuvent jouer plusieurs rôles comme l'inhibition de la cristallisation du carbonate de calcium pour favoriser les ACC (Aizenberg et al., 1996) ou la sélection d'un polymorphe (Laipnik et al., 2020). Il paraît donc possible que ce domaine soit directement impliqué dans la biominéralisation des ACC.

L'expérience est indispensable pour accéder à la structure de l'ensemble de la calcyanine de S. calcipolaris.

Le champ de la modélisation des structures 3D a spectaculairement progressé grâce, entre autres, à ESMFold et AlphaFold2. La calcyanine est cependant un excellent exemple de l'intérêt mais aussi des limites de ces logiciels. AlphaFold2 et ESMFold ne sont pas encore en état de modéliser toutes les protéines existantes, surtout quand celles-ci ont peu de points communs avec ce qui a déjà été décrit.

Par ailleurs, la calcyanine montre aussi la nécessité d'aborder le pLDDT avec un regard critique. En effet, comme décrit précédemment dans les comparaisons de modélisation, il paraît très probable qu'ESMFold et AlphaFold2 n'abordent pas cette métrique de la même manière. Les scores issus de ces 2 logiciels ne sont donc a priori pas comparables. De même, la calcyanine est aussi un bon exemple que le lien entre le désordre et un pLDDT < 50 n'est pas absolu, que ce soit dans AlphaFold2 ou ESMFold.

Il apparaît clairement que la modélisation de la calcyanine de *S. calcipolaris* n'est pas triviale et que cette protéine a des propriétés structurales originales. De plus, même pour le domaine CoBaHMA dont la structure a été modélisée avec robustesse, la question de la fonction reste ouverte. Dès lors, une approche expérimentale apparaît nécessaire pour caractériser plus en avant ses propriétés structurales et éventuellement progresser dans la connaissance de sa/ses fonction(s).

Chapitre 2 : Étude expérimentale de la calcyanine de *S. calcipolaris*.

Le taux d'expression et le rendement de purification de la calcyanine de *S. calcipolaris* chez *E. coli* sont très faibles.

Nous avons exprimé la calcyanine chez *E. coli* comme une protéine recombinante fusionnée à une étiquette 6His (6 Histidines) (Wood, 2014) ou Strep II (T. G. M. Schmidt et al., 1996) pour permettre son identification et sa purification. Les gènes synthétiques utilisés ont été optimisés pour l'expression chez *E. coli*.

Nous avons réussi à exprimer la calcyanine dans des souches d'*E. coli* réputées pour leur capacité à exprimer des protéines considérées comme difficiles d'expression ou toxiques :

_ BL21 (DE3) pLysS (Figure 22.D) et C41 (DE3) (Figure 22.A) (Miroux & Walker, 1996) pour la protéine étiquetée 6His en C_{ter} (Görgen, 2017, Rapport de M2);

_ C41 (DE3) et C43 (DE3) (Figure 22.B&C) (Miroux & Walker, 1996) pour la protéine étiquetée Strep II en C_{ter}.

A l'inverse, les essais d'expression de la protéine étiquetée 6His dans BL21(DE3) codons plus, une souche qui n'est pas spécifiquement adaptée à l'expression de protéine difficile d'expression, ont échoué (Görgen, 2017, Rapport de M2).

La calcyanine semble donc être une protéine difficile à produire pour *E. coli*, et qui nécessite une souche adaptée.

Nous avons choisi la souche *E. coli* C43 (DE3) pendant cette thèse qui donnait les meilleurs résultats quelle que soit la construction.

Toutefois, même avec cette souche, l'expression restait faible. Les valeurs de rendement seront discutées plus en détails quand nous aborderons la purification, mais avec cette souche et la calcyanine de *S. calcipolaris*, nous étions en mesure d'isoler 500µg de protéine par litre de culture bactérienne.

Comme les rendements étaient faibles avec la calcyanine de *S. calcipolaris*, nous avons essayé d'exprimer d'autres calcyanines à domaine CoBaHMA. Nous nous sommes focalisés sur des protéines venant de souches thermophiles (*S. calcipolaris* est une souche mésophile (croissance optimale à 45°C > T > 20°C)). En effet les protéines d'organismes thermophiles présentent une stabilité thermique accrue, grâce à diverses stratégies comme un plus grand nombre d'interactions ioniques, plus de liaisons hydrogène, un cœur hydrophobe plus gros... tout en conservant la structure de leurs homologues mésophiles (Zhou et al., 2008). De plus elles sont globalement meilleures que leurs homologues mésophiles à résister à l'agrégation (dans le sens, un assemblage de protéines mal repliées), en enfouissant les régions propices à former des agrégats dans leur cœur hydrophobe ou en les entourant d'a.a chargés ou de prolines (Thangakani et al., 2012). Ces calcyanines étaient donc potentiellement plus aisées à manipuler car plus stables.

Parmi toutes les souches de cyanobactéries répertoriées qui ont une calcyanine à domaine CoBaHMA, il y avait plusieurs souches thermophiles (Benzerara et al., 2022) (croissance optimale à 80°C > T > 45°C), telle que *Thermosynechococcus elongatus BP-1* (Nakamura, 2002).

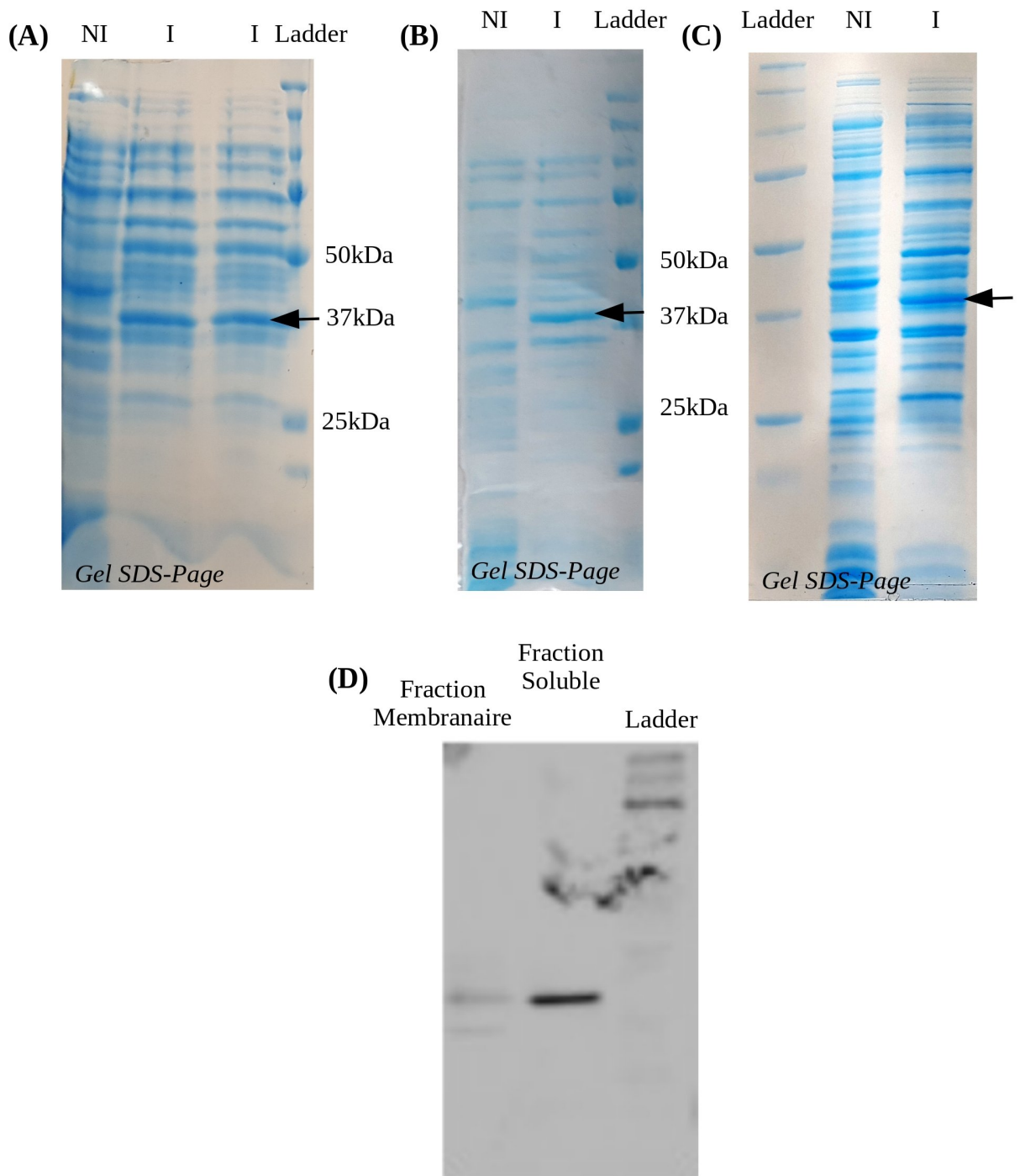


Figure 22 : Expression de la calcanine de *S. calcipolaris* et fractionnement cellulaire.

NI = Non Induit; I = Induit. La flèche noire indique la bande d'expression de la calcanine dans les gels concernés.

Gel SDS-PAGE de l'expression de la calcanine de *S. calcipolaris* (A) étiquetée 6His dans *E. coli* C41 (DE3) à 37°C avec une induction à 0,5mM IPTG. (B) étiquetée Strep II dans *E. coli* C41 (DE3) à 37°C avec une induction à 0,5mM IPTG; (C) étiquetée Strep II dans *E. coli* C43 (DE3) à 37°C avec une induction à 0,5mM IPTG. (D) Western-Blot du fractionnement cellulaire de l'expression de la calcanine de *S. calcipolaris* dans *E. coli* BL21 (DE3) pLysS. La protéine a été identifiée avec un anticorps anti-étiquette 6His (Görge *et al.*, 2017, Rapport de M2).

Nous avons choisi les séquences des calcyanines de *T. elongatus BP-1*, *Synechococcus lividus sp.* PCC (Pasteur Culture collection of Cyanobacteria) 6716 et *Synechococcus lividus sp.* PCC 6717. Toutes ces calcyanines ont une architecture de type CoBaHMA - (GlyZip)₃. Leurs séquences sont identiques à plus de 50%, et similaires à plus de 65% à celle de *S. calcipolaris*. Ces protéines sont donc proches en termes de séquence de celle de *S. calcipolaris*.

Toutefois, en dépit d'un grand nombre de conditions testées (5 souches différentes d'*E. coli* (BL21 pLysS, C41 (DE3), C43 (DE3), C44 (DE3), C45 (DE3) (Angius et al., 2018)) à 2 températures différentes (37°C et 20°C)) nous n'avons pas réussi à exprimer ces protéines. Nous avons donc continué à travailler avec la calcyanine de *S. calcipolaris*.

Il est assez difficile d'expliquer cette difficulté d'expression des calcyanines à domaine CoBaHMA. Une possibilité serait que la calcyanine serait produite en trop grande quantité par *E. coli*. Comme cela a été montré, une accumulation de protéines recombinantes chez *E. coli* peut avoir des effets délétères voire létaux pour la cellule, par exemple en induisant une perte de capacité à synthétiser des protéines (Dong et al., 1995), ou en saturant le système de translocation à la membrane de la bactérie (Wagner et al., 2008). En théorie, les souches d'*E. coli* C41, C43, C44, C45 évitent ce problème en diminuant l'expression de leurs protéines recombinantes, ce qui induit une diminution du rendement d'expression par cellule, mais augmente la survie des bactéries et donc le rendement global (Angius et al., 2018; Kwon et al., 2015; Wagner et al., 2008). Cependant, dans le cas de la calcyanine, cela pourrait n'être pas suffisant.

La calcyanine pourrait aussi être toxique en elle-même pour *E. coli*. Les long motifs glycine zipper constitutifs de son domaine (GlyZip)₃ présentés précédemment évoquent ceux des toxines CdzC/CdzD (García-Bayona et al., 2017) ou ceux du Type VI secretion system (Ali et al., 2022). Dans ces 2 exemples, le glycine zipper est vu comme un élément central de la toxicité des protéines, car il s'insérerait dans les membranes de la cellules cible et les perméabiliserait (Ali et al., 2022; García-Bayona et al., 2017). La calcyanine pourrait avoir un effet similaire dans *E. coli*.

Toutefois cette 2nde hypothèse paraît peu probable, car lors des expériences de fractionnement cellulaire, nous avons retrouvé la calcyanine de *S. calcipolaris* majoritairement dans la fraction cytosolique des protéines d'*E. coli* (Figure 22.D) et non dans les fractions membranaires.

La calcyanine purifiée forme des homo-oligomères de haut poids moléculaires très hétérogènes.

Comme décrit dans la partie précédente, nous avons exprimé la calcyanine associée à 2 étiquettes distinctes 6His et Strep II, et ce afin de s'assurer que l'étiquette n'impactait pas ou peu la calcyanine.

Nous avons purifié ces 2 constructions par une chromatographie d'affinité suivie d'une SEC (Size Exclusion Chromatography) sur colonne Superdex 200. Les résultats étaient très similaires pour les 2 constructions. En dépit d'une gamme de séparation de 10-600 kDa, le profil de chromatographie était très complexe, avec la présence d'un large pic multiple. L'analyse du contenu de fractions correspondant à ce pic multiple par gel SDS-PAGE montrait la présence de la calcyanine, quasiment pure, dans tout le pic à partir du volume mort de la colonne (Figure 23.A&B&C).

La complexité du pic indiquait que la calcyanine ne formait pas un, mais plusieurs objets de tailles différentes en solution, dont les plus gros allait jusqu'au volume mort de la colonne Superdex 200. D'après les informations fournies par le constructeur

(<https://cdn.cytivalifesciences.com/api/public/content/digi-13947-pdf>), ce volume mort se situe pour des protéines de taille 600kDa (~15 fois la masse de la calcyanine).

Cette observation a été confirmée par dépôt des différentes fractions d'éluion sur gel natif. Ces fractions formaient plusieurs bandes sur gel natif, bandes allant du monomère de la BSA (~66 kDa), jusqu'en haut du gel, avec la présence d'un quasi continuum de bande en haut du gel (Figure 23.D). Une population majoritaire semblait toutefois se distinguer aux alentours du tétramère de la BSA (~264kDa). Mais cette population n'était jamais seule sur gel natif. De plus, en l'absence de pic clairement défini sur le chromatogramme, il nous paraissait difficile de l'isoler.

Nous avons tout de même essayé de prélever cette population en rassemblant les fractions de la 2eme moitié du pic, où elle est enrichie (Figure 23.A, cadre en pointillé).

Les analyses menées en SEC-SAXS (Small Angle X-ray Scattering) sur ces fractions nous ont montré un R_g (Rayon de giration) décroissant sur la largeur du pic quelque soit l'étiquette (Figure 24.B), ce qui témoigne d'une hétérogénéité de la taille des objets en solution. Il était possible d'identifier tout de même une petite zone de R_g stable en fin de pic. Le R_g de cette zone est 59 ± 4 Å. En se basant sur la formule $\log(R_g) = -0.27 + 0.39 \log(M_w)$ (coefficient pour une expérience SAXS), développée par Smilgies et Folta-Stogniew pour des protéines globulaires, cela correspondrait, dans l'hypothèse d'une protéine globulaire, à un objet de ~ 470 kDa (~12 unités de calcyanine) (Smilgies & Folta-Stogniew, 2015). Comme les 2 déterminations de la masse, le gel natif et le R_g , sont imprécises, il est possible que cette population corresponde à celle qui ait été observée sur gel natif.

La décroissance du R_g en amont de la zone stable indique que la calcyanine forme un grand nombre d'objets de poids moléculaires supérieurs au ~ 470 kDa estimé.

Il n'était matériellement pas possible d'isoler cette zone stable de R_g . La calcyanine s'exprimait déjà très mal : en prenant l'intégralité du pic complexe visible sur le chromatogramme, nous ne pouvions isoler que 500µg de protéine par L de culture bactérienne. La fraction du pic avec un R_g stable ne représentait qu'une infime portion de ce pic complexe.

Les essais de cristallogénèse demandent 1 mg de protéine monodisperse, une quantité et des conditions qui n'étaient pas envisageables pour la calcyanine de *S. calcipolaris*.

Il faut aussi noter que nous avons observé en fin de purification, de façon faible mais systématique, un contaminant à ~75kDa (Figure 23.B&C). Ceci était surprenant, notamment dans le cas de la construction étiquetée Strep II, cette étiquette étant connue pour sa forte spécificité à la Strep-Tactine ce qui permet une grande pureté dès la sortie de la chromatographie d'affinité (Lichy et al., 2005). Pour que ce contaminant passe les 2 étapes de purification avec les 2 étiquettes, il fallait envisager une interaction directe entre ce dernier et la calcyanine. Une analyse de MS (Mass Spectroscopy) de cette bande a révélé que ce contaminant était probablement la DnaK (H. Saito & Uchida, 1977), une chaperonne d'*E. coli*. Parmi ses différentes fonctions identifiées, on trouve l'aide au repliement des protéines, ainsi que la désintégration des agrégats puis le repliement des protéines dénaturées (Rosenzweig et al., 2019). La présence de la DnaK est peut-être une indication que la calcyanine était, en partie au moins, dans un état dénaturé ou mal replié.

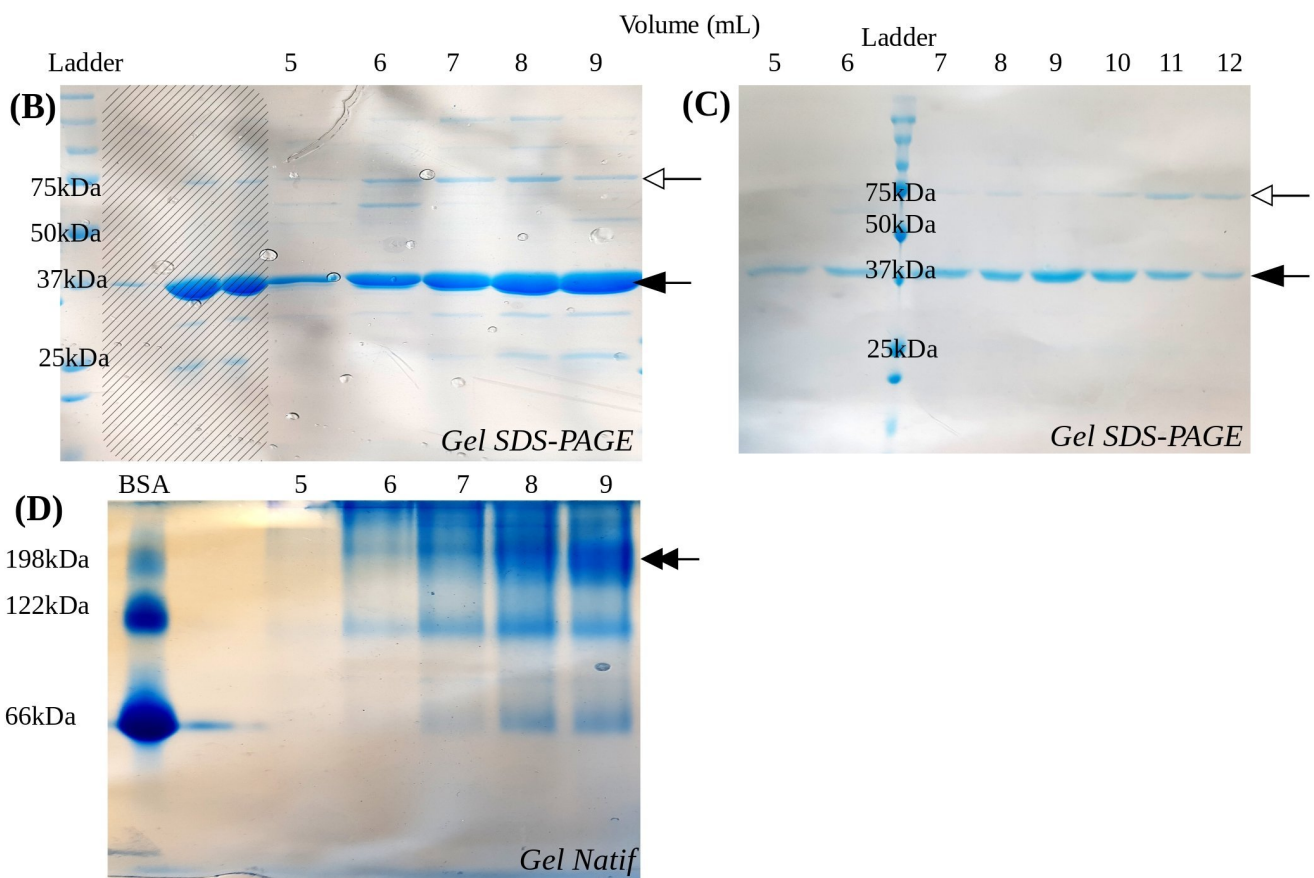
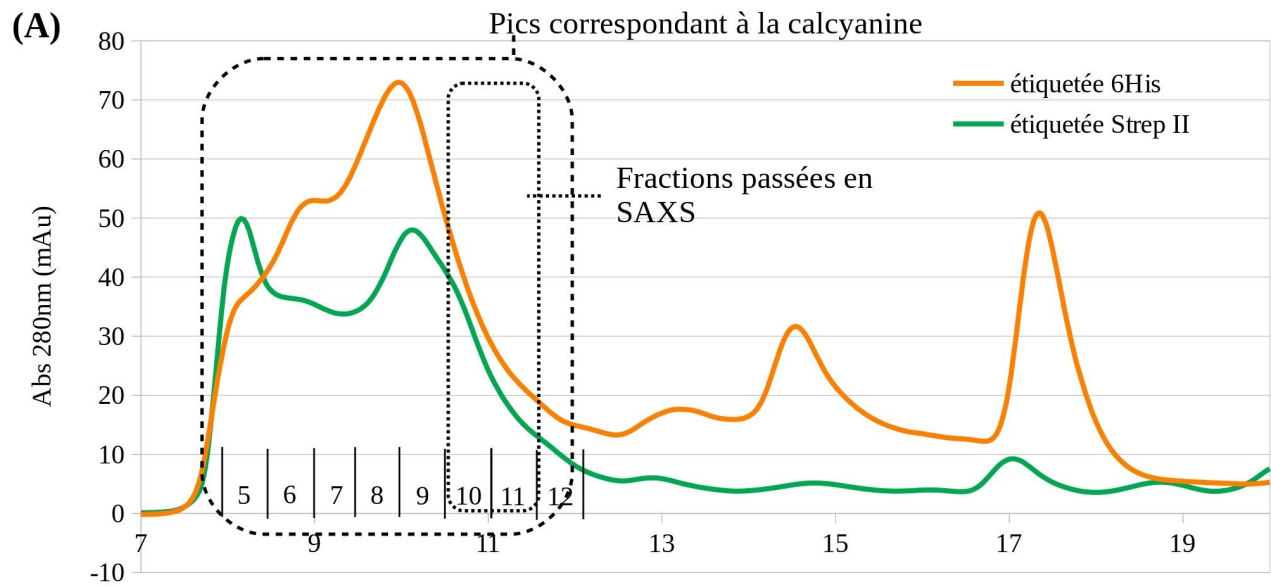


Figure 23 : Purification de la calyculine de *S. calcipolaris* étiquetée 6His et Strep II.

La flèche noire indique la bande de la calyculine sur le gel SDS-PAGE, la flèche blanche celle de la DnaK. La double flèche noire indique la population majoritaire sur gel natif.

(A) Profil d'éluion de la calyculine étiquetée 6His et StrepII, en sortie de SEC sur Superdex 200 10/300 HR. (B)&(C) Gel SDS-PAGE des fractions d'éluion pour la calyculine (B) étiquetée 6His (C) étiquetée Strep II. (D) Gel natif des fractions présentes sur le gel SDS-PAGE (B).

Les numéros au dessus des puits correspondent aux fractions d'éluion de la SEC.

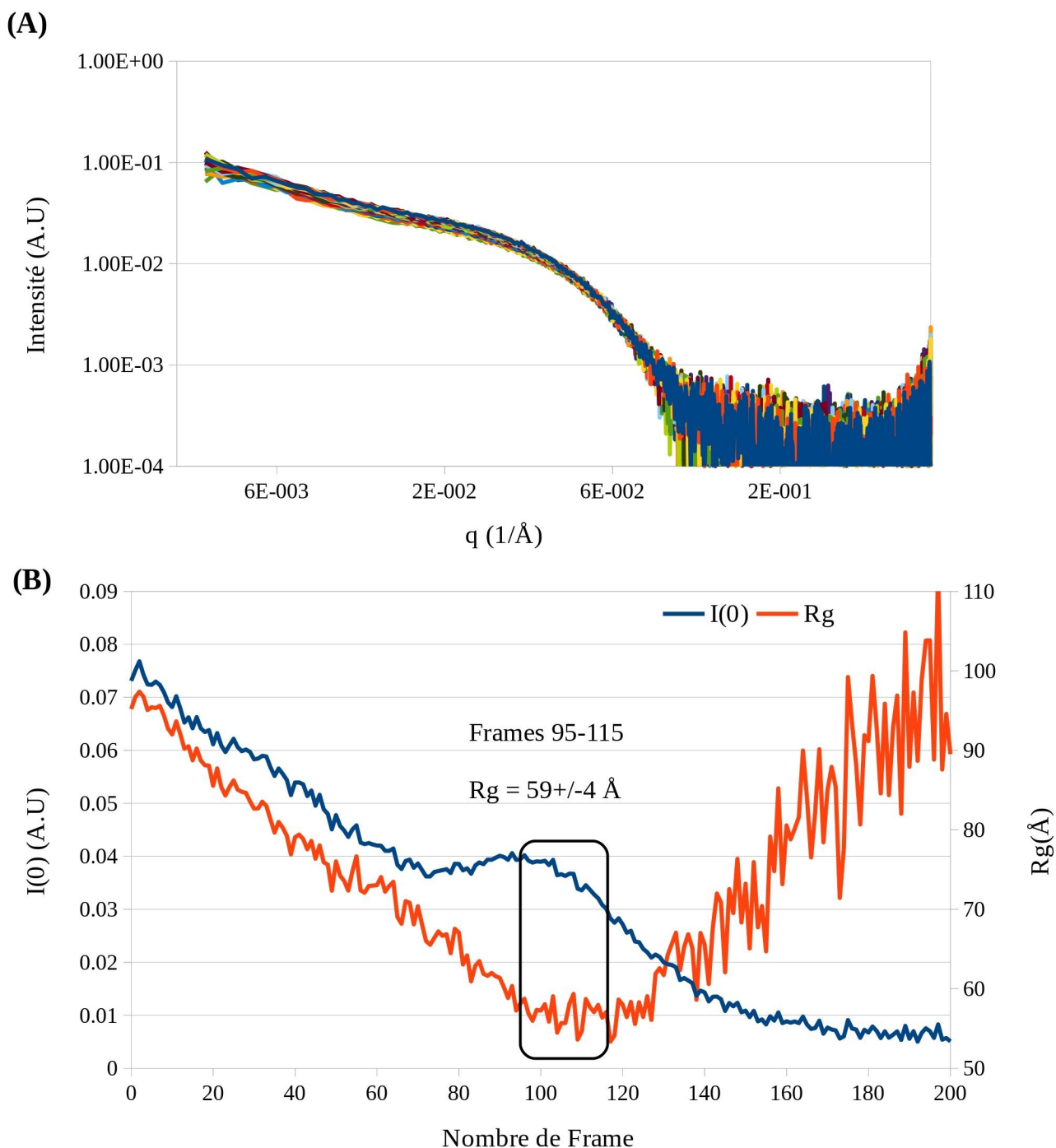


Figure 24 : Analyse SAXS de la calcyanine de *S. calcipolaris* étiquetée StrepII.

(A) Courbe $I(q) = f(q)$ (cf partie 2.8 du Matériels & Méthodes) des frames correspondant à la population avec un R_g stable. (B) Courbe du R_g et $I(0)$. La partie des courbes correspondant à la population avec un R_g stable est encadrée en noir. Le calcul du R_g s'est fait en sélectionnant la partie $q \in [0.015; 0.023] \text{Å}^{-1}$ des frames du cadre noir.

Devant ces résultats, nous avons établi qu'étudier la calcyanine de *S. calcipolaris* entière n'était pas une stratégie viable pour déterminer la structure de la calcyanine. Nous avons alors cherché à identifier un fragment de la calcyanine qui soit plus stable et plus propice à l'étude structurale par protéolyse limitée.

La protéolyse limitée a permis d'identifier un fragment stable et homogène, exploitable pour la cristallogénèse.

La protéolyse limitée consiste à exposer une protéine à une protéase dans des conditions très contrôlées afin de ne cliver que certains morceaux de la protéine. Classiquement, cela implique de travailler avec la protéine en excès vis à vis de la protéase (par ex ratio de 1:100 (w/w)), à basse température (par ex. sur glace) et sur un temps court (par ex. 1h), puis d'arrêter brutalement la réaction par ajout d'un inhibiteur de la protéase. Les fragments de protéine qui résistent à la protéolyse sont identifiés par gel SDS-PAGE et MS (Fontana et al., 2012).

Les protéases reconnaissent des sites de clivage spécifiques dans les séquences, mais leur capacité à cliver ces sites dépend directement de leur capacité à y accéder. 3 paramètres contrôlent cela, par ordre d'importance : l'exposition au solvant du site, la flexibilité de la chaîne peptidique, et les interactions locales (par ex. liaisons hydrogènes). Concrètement, les clivages sont observés majoritairement dans les boucles flexibles et les parties désordonnées de la protéine (Kazanov et al., 2011). Pour cette raison, la protéolyse limitée peut être utilisée pour détecter les zones désordonnées (Fontana et al., 2012) et par voie de conséquence les cœurs structurés pour des protéines globulaires. La protéolyse limitée a donc été utilisée pour résoudre la structure 3D d'un grand nombre de protéines, en retirant les parties flexibles qui gênent la cristallisation (Derewenda, 2004).

Nous avons exposé la calcyanine à 2 protéases : la trypsine et l' α -chymotrypsine dans les conditions et les ratios décrits dans le Matériels & Méthodes. La trypsine clive principalement en C_{ter} des lysines et des arginines. L' α -chymotrypsine clive principalement en C_{ter} des a.a aromatiques (tryptophane, phénylalanine et tyrosine) et de certains hydrophobes comme la leucine, l'isoleucine ou la méthionine.

Ces 3 protéases avaient un grand nombre de sites de clivage prédits sur la séquence de la calcyanine (Figure 27.A). Pour la protéolyse par la trypsine, une unique bande était visible pour un ratio protéine/protéase de 100/1, à un poids moléculaire compris entre 15 et 20 kDa. Cette bande était aussi visible à un ratio protéine/protéase de 10/1, indiquant un objet protéique très résistant à la protéolyse. Pour l' α -chymotrypsine, une unique bande était visible à un ratio de protéine/protéase de 10/1 à un poids compris entre 20 et 25 kDa (Figure 25).

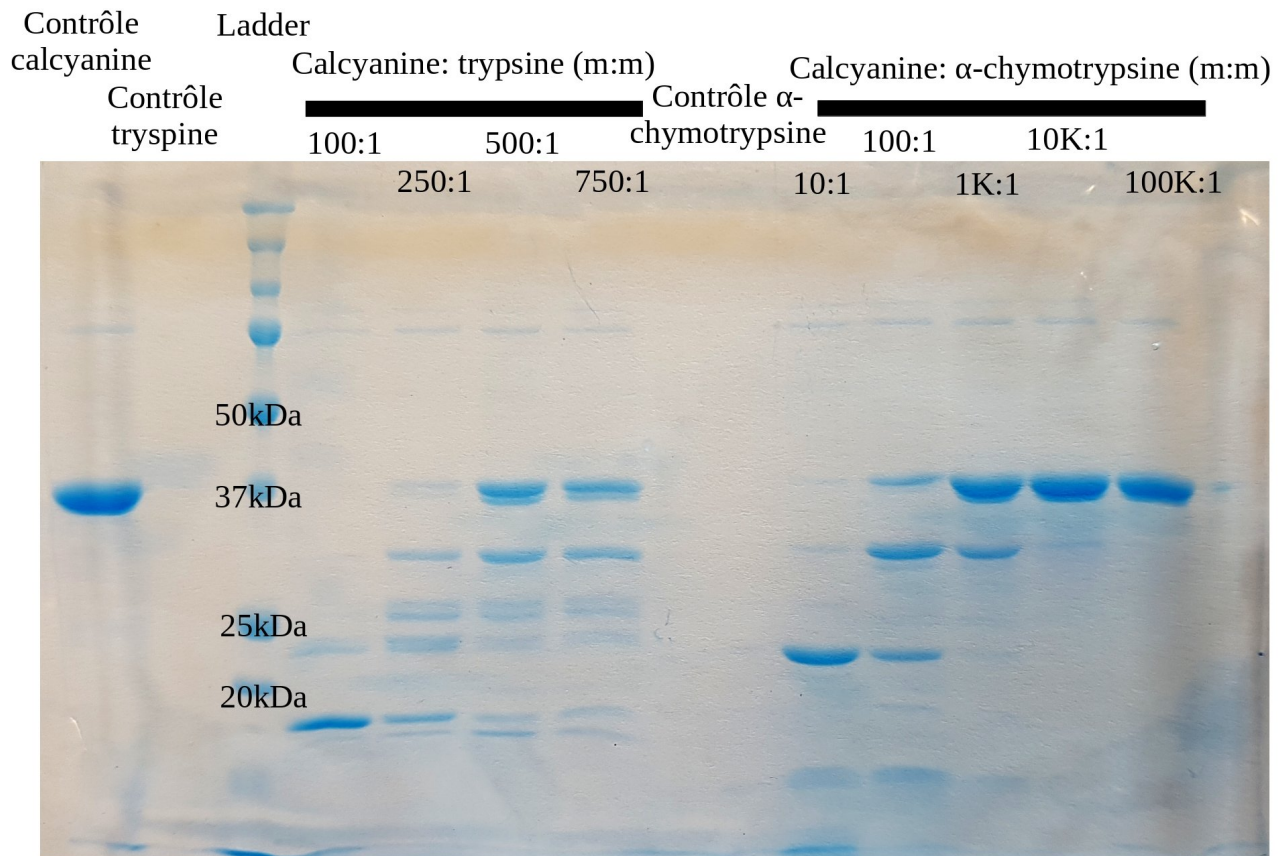
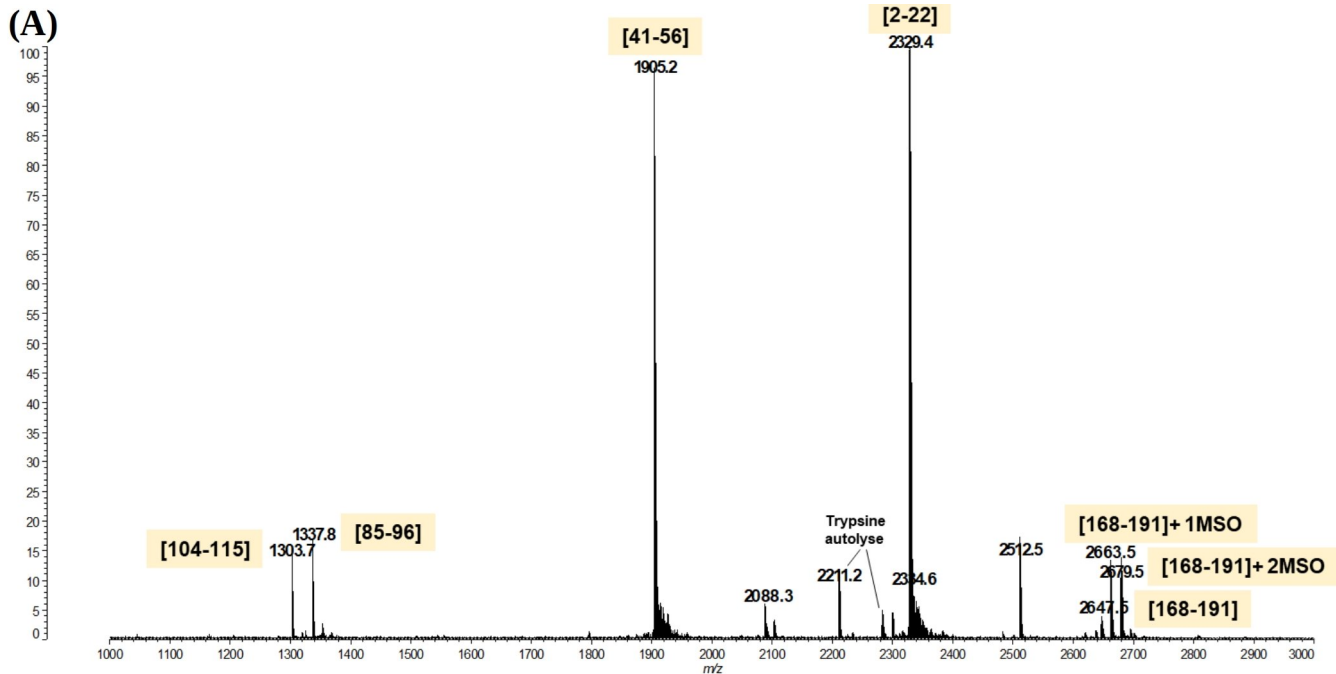


Figure 25 : Gels SDS-PAGE de la protéolyse de la calceyanine de *S. calcipolaris*.

Protéolyse par le trypsine et l' α -chymotrypsine. Les contrôles sont réalisés avec la concentration maximale de la protéine/protéase utilisée dans les différents tests.

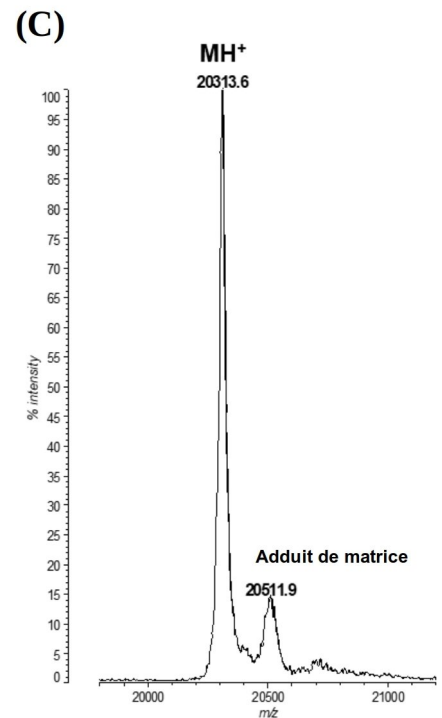
Avec l'aide de Christophe Marchand (Institut de Biologie Paris-Seine, UMR 7238 CNRS, Biologie Computationnelle et Quantitative (LCQB), Sorbonne Université), nous avons analysé les bandes du gel SDS-PAGE issus de la protéolyse par la trypsine et de l' α -chymotrypsine par MS (Spectroscopie de Masse). L'objet résistant à la trypsine est bien un fragment de la calceyanine, et correspond à la séquence P2 – K190, pour un poids moléculaire de 20,3kDa, supérieur à celui visible sur gel (Figure 26.A&B). Ce fragment recouvre donc le CoBaHMA + GlyZip1. L'objet résistant à l' α -chymotrypsine est aussi un fragment de calceyanine. Les fragments détectés en MS ne terminant pas sur un site de clivage de la protéase, il a fallu extrapoler le fragment complet. Le site de clivage le plus proche, et qui donne un fragment de la bonne taille vis à vis du gel SDS-PAGE, est la méthionine en 216. Cela correspond à la séquence P2 – M216 pour un poids moléculaire de 23,2kDa (Figure 26.D&E). Ce fragment recouvrirait aussi le CoBaHMA + GlyZip1, jusqu'au tout début du GlyZip2.

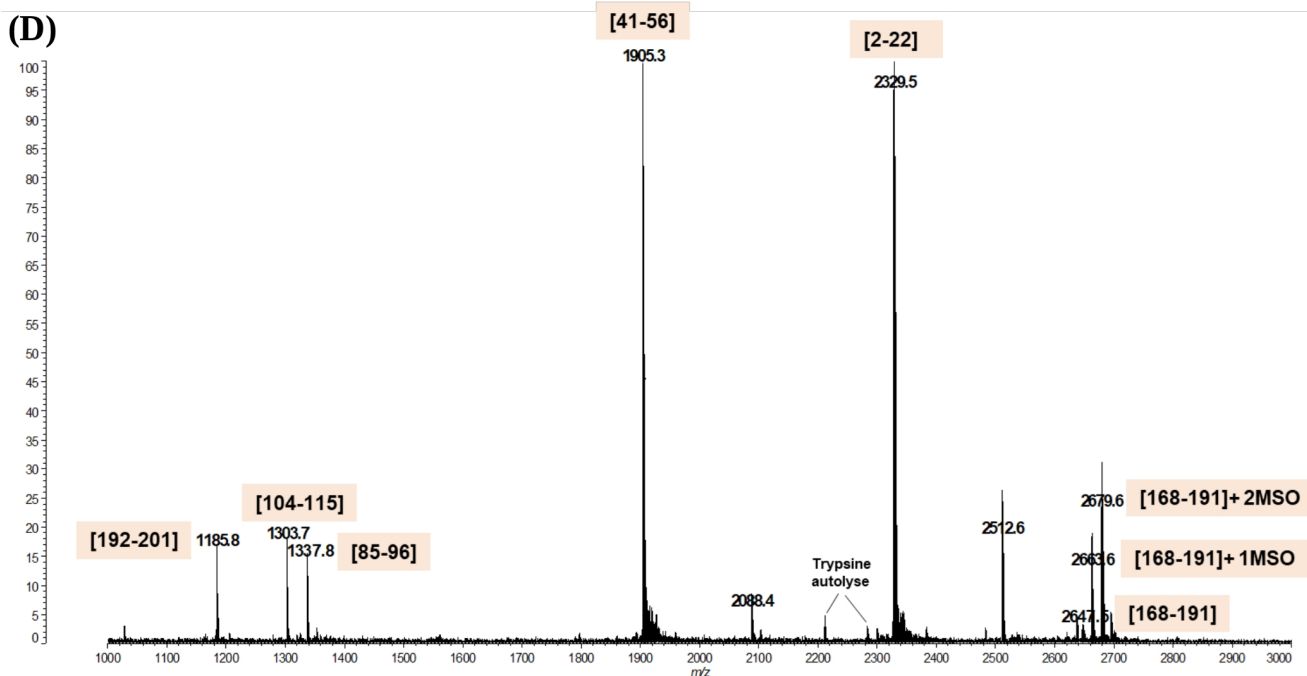
Les analyses MS faites sur gels SDS-PAGE ne permettent de détecter que certains fragments des protéines analysées. Afin d'être sûr d'avoir les bonnes bornes, et ne pas manquer un fragment en C_{ter} , nous avons complété cette analyse par une détermination de la masse complète. Comme les bornes pour le fragment issu de la protéolyse par l' α -chymotrypsine étaient moins sûres que celles obtenues lors de la protéolyse par la trypsine, nous nous sommes concentrés uniquement sur cette dernière.



(B)

| | | | | |
|--|-------------------|-----------------------|-------------------|-------------------|
| | 10 | 20 | 30 | 40 |
| | <u>MPKPSDDEES</u> | <u>LPPVAELVHL</u> | <u>TRDRLRLRLP</u> | <u>LLKKDPDYGR</u> |
| | 50 | 60 | 70 | 80 |
| | <u>YLQDYLKPIP</u> | <u>GITEVRLNLQ</u> | <u>AASLSIHYAL</u> | <u>DLITPLQILA</u> |
| | 90 | 100 | 110 | 120 |
| | <u>LIERWGDVQI</u> | <u>IGQGHKGLN</u> | <u>LTRAFELEPO</u> | <u>EVGGLKQMG</u> |
| | 130 | 140 | 150 | 160 |
| | <u>GFVVGGQIGD</u> | <u>VVGGVVGTA</u> | <u>GGVFMGPAGM</u> | <u>LMGAQVGTfV</u> |
| | 170 | 180 | 190 | 200 |
| | <u>GGVIGGR</u> | <u>LGI EAMEQISQLT</u> | <u>FQDMQAPGP</u> | <u>KTALTPEEIR</u> |
| | 210 | 220 | 230 | 240 |
| | <u>REAEIAKALE</u> | <u>IRSGAKMGEV</u> | <u>VGELAGGIAG</u> | <u>QTVLGPPGEA</u> |
| | 250 | 260 | 270 | 280 |
| | <u>VGRVLGEMLG</u> | <u>GQIGEDVSRQ</u> | <u>VAEKSEAAIP</u> | <u>TDLSVNIVLE</u> |
| | 290 | 300 | 310 | 320 |
| | <u>WWMKTSRAfV</u> | <u>GETALATLGG</u> | <u>LLSRVILGPQ</u> | <u>AESVGLRAGT</u> |
| | 330 | 340 | 350 | 360 |
| | <u>RVGRLVDWNG</u> | <u>QDQQKTKEMA</u> | <u>ETPPTETEIT</u> | <u>TSPESNREGK</u> |
| | 370 | 380 | | |
| | <u>SVGSSENLYF</u> | <u>QSASWSHPQF</u> | <u>EKGA</u> | |





(E)

```

10          20          30          40          50          60          70
MPKPSDDEES LPPVAELVHL TRDRLRLRLP LLKKDPDYGR YLQDYLKPIP GITEVRLNLQ AASLSIHIAL
80          90          100         110         120         130         140
DLITPLQILA LIERWGDVQI IGQGHKGLEN LTRAFELEPQ EVGGKLGKMG GFVVGQIGD VVGGVVGTA
150         160         170         180         190         200         210
GGVFMGPAGM LMGAQVGFV GGVIGGRLGI EAMEQISQLT FQDMQDAPGP KTALTPEEIR REAEIAKALE
220         230         240         250         260         270         280
IRSGAKMGEV VGELAGGIAG QTVLGPPGEA VGRVLGEMLG GQIGEDVSRQ VAEKSEAAIP TDLSVNIVLE
290         300         310         320         330         340         350
WWMKTSRAFV GETALATLGG LLSRVILGPQ AESVGLRAGT RVGRLVDWNG QDGQTKEMA ETPPTETEIT
360         370         380
TSPESNREGK SVGSSENLYF QSASWSHPQF EKGA

```

Figure 26 : Analyses MS de la protéolyse limitée de la calcyanine de *S. calcipolaris*.

Protéolyse limitée faite par la trypsine (A)&(B)&(C) et l' α -chymotrypsine (D)&(E). Spectre MALDI-TOFF de la bande SDS-PAGE du fragment de digestion de la calcyanine par (A) la trypsine et (E) l' α -chymotrypsine. Les poids moléculaires et les bornes qui correspondent sur la séquence sont écrits au dessus de chaque pic. Ces fragments sont illustrés en gras sur la séquence de la calcyanine en (B) & (E). La séquence totale du fragment protéolytique après digestion par la trypsine est soulignée en trait plein en (B). La séquence putative totale du fragment protéolytique après digestion par l' α -chymotrypsine est soulignée en pointillé en (E). (C) Spectre de la masse totale du fragment protéolytique digéré par la trypsine. Sa masse expérimentale est : $MW_{\text{expérimentale}} = 20313,6$ Da (MH+ average). La masse théorique de [2-191] est : $MW_{\text{théorique}} = 20319,5$ Da (MH+ average).

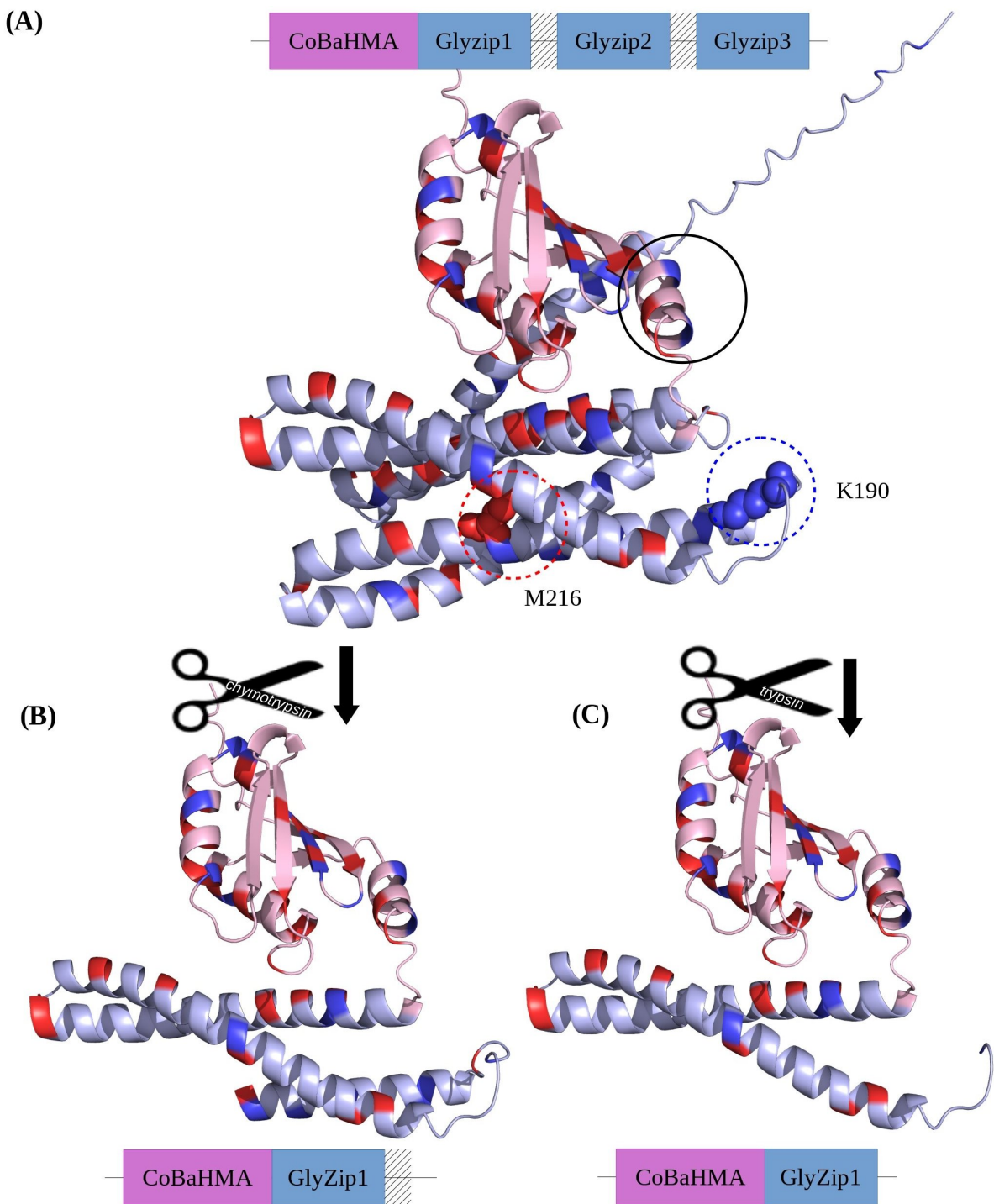


Figure 27 : Illustrations des résultats des protéolyses limitées par l' α -chymotrypsine et la trypsine. Le domaine CoBaHMA est visible en rose pâle. Le domaine (GlyZip)₃ en bleu pâle. En rouge sont illustrés les sites prédit de clivage par l' α -chymotrypsine, et en bleu ceux de la trypsine. Les sites clivés identifiés ou supposés par MS, sont illustrés avec des sphères. (A) Modèle de la calcyanine complète conçu par ESMFold, avec les sites de clivage illustrés. L'hélice α qui relie le CoBaHMA et le (GlyZip)₃ est entouré d'un cercle noir. (B)&(C) Modèle restreint au fragment résistant à (B) l' α -chymotrypsine (tel que supposé) (C) la trypsine (tel qu'identifié par MS).

Nous avons donc protéolysé la calcyanine par la trypsine dans un ratio en masse de 50/1 protéine/protéase, puis nous avons isolé le fragment par SEC sur Superdex 200 5/150 GL. L'analyse MS sur ce fragment purifié a confirmé le poids exact du fragment (Figure 26. C), et donc les bornes identifiées à partir du gel SDS-PAGE.

Sur le modèle de la structure 3D de la calcyanine, le site de clivage qui a été protéolysé par la trypsine est localisé sur la boucle désordonnée qui relie le GlyZip1 au GlyZip2 (Figure 27. A&C). Les protéases attaquent principalement les sites qui sont accessibles au solvant, donc majoritairement des boucles et des zones désordonnées (Kazanov et al., 2011). La protéolyse de la trypsine sur cette boucle et la résistance du CoBaHMA + GlyZip1, vient donc supporter la répartition des zones structurées/désordonnées proposées par le modèle, pour cette région de la calcyanine.

Par contre la protéolyse par l' α -chymotrypsine est plus complexe à interpréter. En effet, selon le modèle ESMFold, la coupure a eu lieu sur une hélice α à la limite N_{ter} du GlyZip2 (Figure 27. A&B). Bien que plus rares que les clivages sur parties désordonnées, les protéases peuvent attaquer les hélices α (Kazanov et al., 2011). Cependant, comme illustré sur la Figure 27. A, l' α -chymotrypsine avait d'autres sites de clivage *a priori* bien plus accessibles sur la séquence. Nous pouvons formuler plusieurs explications au fait que ce site là ait été privilégié.

Premièrement il est possible que nous n'ayons pas les bonnes bornes pour ce fragment. En effet, nous avons fait une analyse MS de ce fragment seulement sur gel SDS-PAGE, sans détermination en masse entière. Dans ces conditions, il est possible que nous ayons oublié un fragment de la calcyanine en C_{ter}. Peut être que l' α -chymotrypsine protéolyse la boucle désordonnée qui se situe après le GlyZip2, mais que les fragments C-terminaux ne sont pas visibles en MS. Le fragment ainsi formé aurait une taille de 28 kDa, bien supérieure au ~23 kDa constaté sur gel SDS-PAGE. Toutefois, en l'absence de détermination en masse entière pour vérifier le poids exact du fragment, nous ne pouvons pas exclure cette possibilité.

Deuxièmement, il est possible que nous ayons les bonnes bornes pour ce fragment, mais que le GlyZip2 (et éventuellement le GlyZip3) soit mal structuré. De fait, l'absence d'ordre permettrait le clivage par la protéase.

Enfin, il est aussi possible que le modèle ESMFold soit erroné sur cette section-là de la calcyanine, et que la zone en N_{ter} du GlyZip2 ne soit pas structurée.

La résistance de ce fragment aux protéases indiquait qu'il avait donc de bonnes chances d'être une zone structurée, avec peu de zones flexibles ou désordonnées, et donc d'être adapté pour une détermination structurale.

L'isolement du fragment par protéolyse limitée n'était pas souhaitable, car cela posait le risque d'obtenir un mélange complexe composé de la calcyanine dans différents états d'avancement de protéolyse. Nous avons préféré exprimer le fragment CoBaHMA-GlyZip1 sous forme de protéine recombinante dans *E. coli* (DE3) C43.

En plus de l'étiquette 6His, nous avons fusionné le fragment à la MBP (Maltose-Binding Protein), qui a souvent été rapportée comme favorisant un bon repliement des protéines auxquelles elle est associée (Kapust & Waugh, 1999).

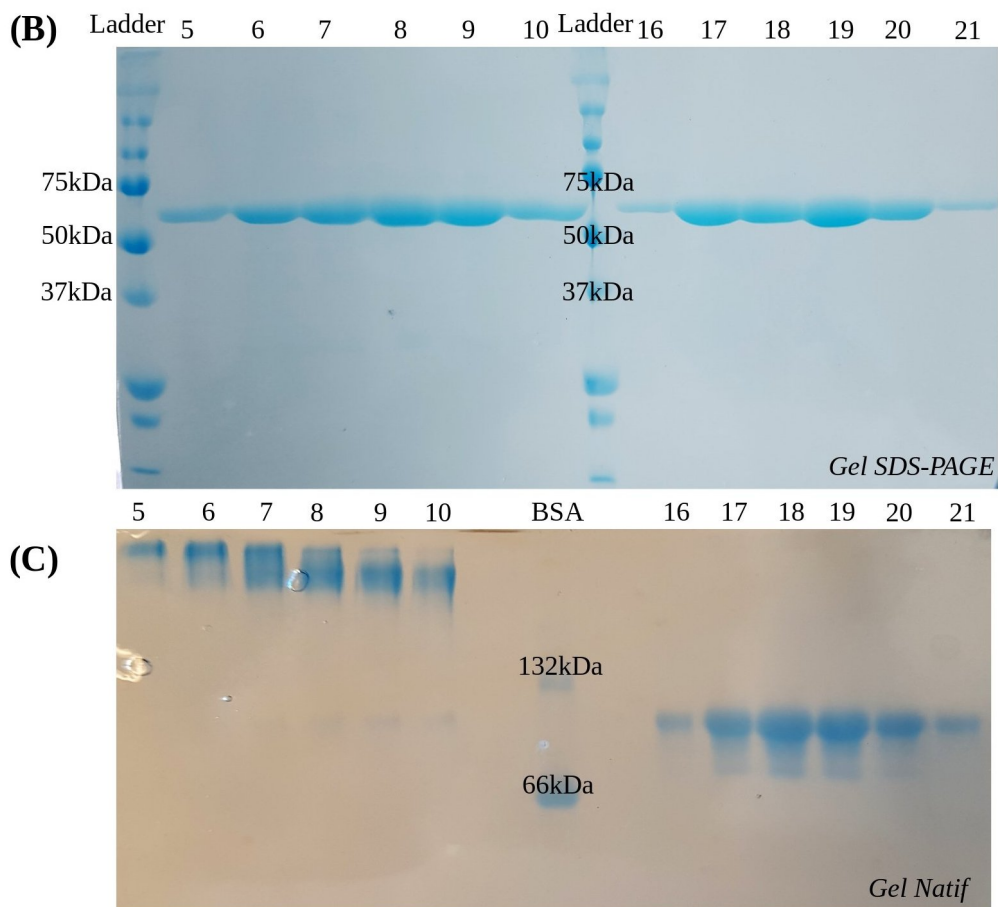
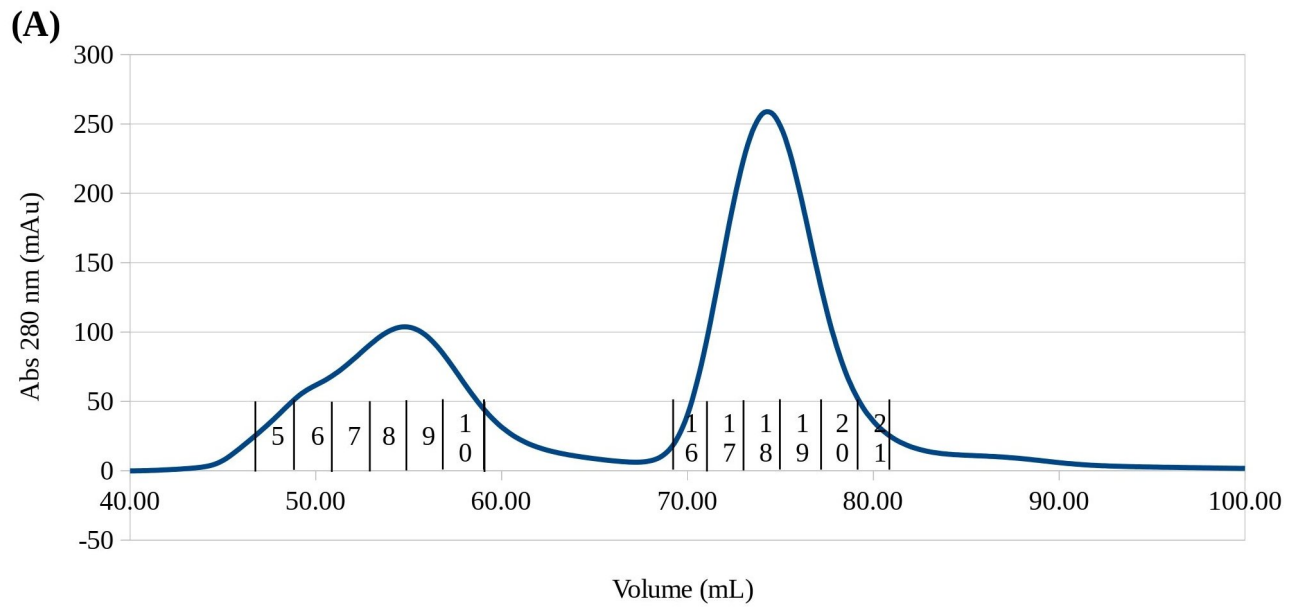


Figure 28: Purification de la construction MBP-CoBaHMA-GlyZip1.

(A) Profil d'éluion en sortie de SEC sur Superdex 200 16/600 PG. (B) Gel SDS-PAGE des fractions d'éluion. (C) Gel natif des fractions présentes sur le gel SDS-PAGE (B).

Les numéros au dessus des puits correspondent aux fractions d'éluion de la SEC.

La construction MBP-CoBaHMA-GlyZip1 se comporte comme une protéine cytosolique dans *E.coli*. Nous l'avons purifiée comme décrit dans le Matériels & Méthodes. Sur le chromatogramme de sortie de SEC, nous avons pu voir 2 pics bien séparés. Le 1^{er} de ces pics avait un léger épaulement au niveau des hauts poids moléculaires (figure 28.A). Un gel SDS-PAGE des fractions correspondant à ces 2 pics nous a confirmé que ces 2 pics correspondaient bien à la construction MBP-CoBaHMA-GlyZip1. Sur gel natif, le 1^{er} pic semblait être constitué d'au moins 2 populations de taille > 132 kDa, ce qui pourrait expliquer l'épaulement visible sur le chromatogramme pour ce pic. Le 2nd pic était monodisperse sur gel natif, avec une taille comprise entre 66 et 132 kDa, et qui pourrait correspondre à un dimère de la protéine fusion (~130 kDa) (Figure 28).

Les rendements étaient spectaculairement plus élevés que ceux obtenus pour la protéine entière, de l'ordre de 20mg de protéines/L de culture bactérienne contre 500µg/L de culture bactérienne pour la protéine entière. De plus il était très aisé de concentrer cette protéine fusion : nous l'avons ainsi concentré à plus de 50 mg/mL sans voir de précipité apparaître.

Ces résultats étaient de très loin les meilleurs que nous ayons jamais eus dans la partie expérimentale de ce projet. Pour la première fois nous avons réussi à obtenir des protéines pures, probablement homogènes, dans des quantités suffisantes pour faire des analyses biophysiques. Exprimer le fragment issu de la protéolyse, fusionné à la MBP semblait donc être une bonne stratégie pour obtenir au moins une partie de la structure de la calcanine.

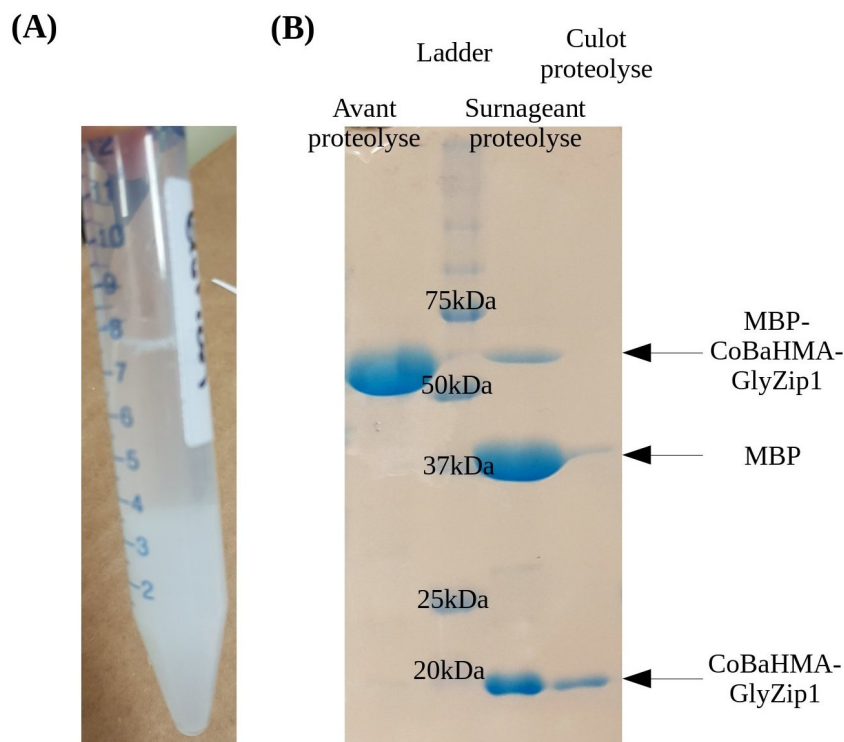


Figure 29 : Protéolyse de la construction MBP-CoBaHMA-GlyZip1 par la protéase TEV.

(A) Photo du précipité blanc apparu après la protéolyse. (B) Gel SDS-PAGE de la protéolyse.

Comme nous étions intéressés par la structure 3D du fragment CoBaHMA-GlyZip1, et non par celle de la MBP qui est déjà connue (Spurlino et al., 1991), nous avons isolé ce fragment. Pour cela, après la chromatographie d'affinité, à la place de faire une SEC, nous avons protégé la protéine fusion avec la protéase TEV. A notre surprise, alors que la construction fusion pouvait être fortement concentrée (>50 mg/mL), le clivage à la TEV s'est accompagné de la formation d'un abondant précipité blanc (Figure 29.A). Le précipité était difficile à resuspendre dans une solution de Laemmli, de fait les résultats du gel SDS peuvent être imprécis, mais il semblerait que ce précipité soit enrichi en fragment CoBaHMA-GlyZip1, par rapport à la MBP (Figure 29.B).

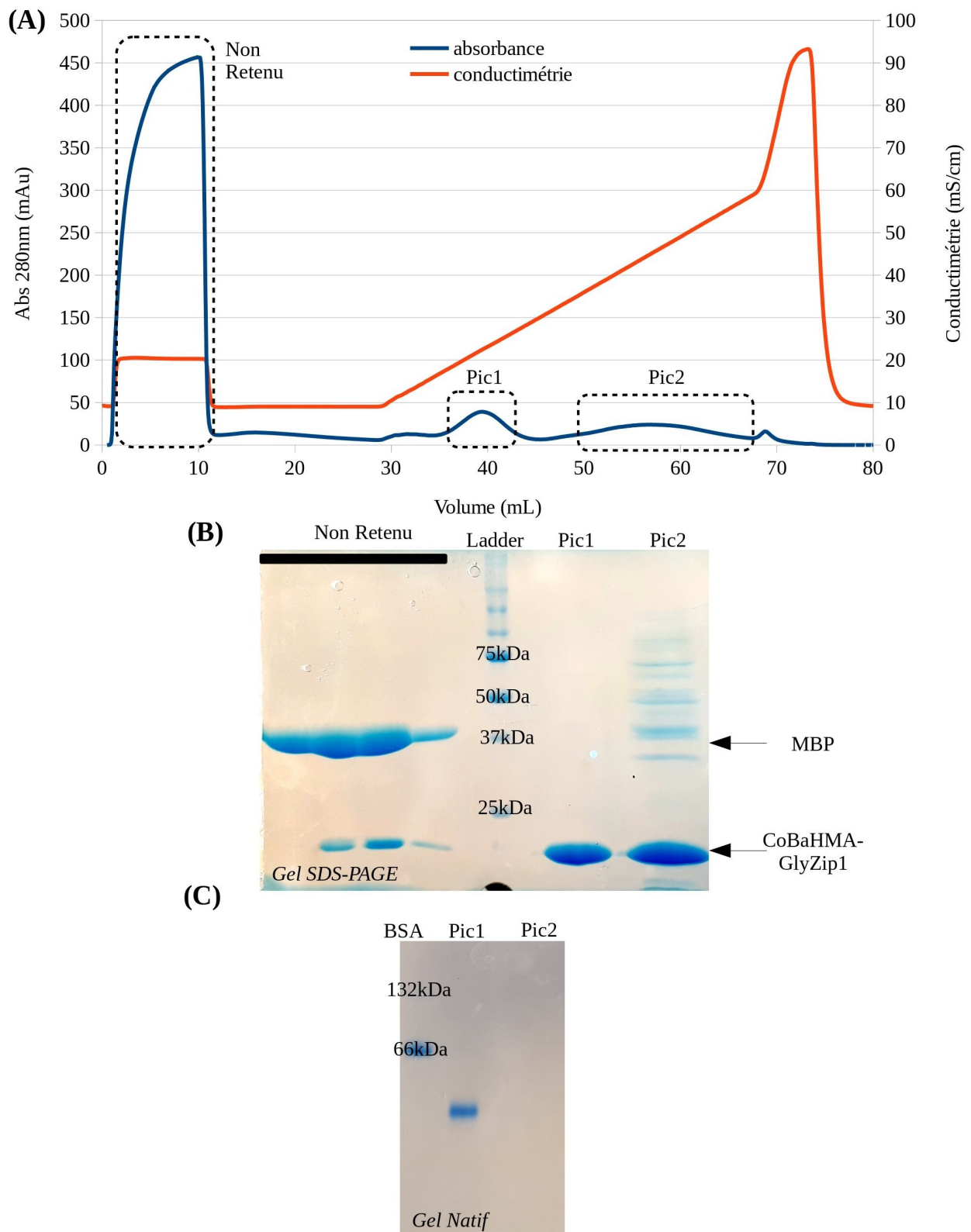


Figure 30 : Purification de la construction CoBaHMA-GlyZip1.

(A) Profil d'éluion en sortie de chromatographie échangeuse d'ions. Le gradient de conductivité utilisé pour l'éluion est indiqué. (B) Gel SDS-PAGE des pics. (C) Gel natif des pics.

Cette précipitation a entraîné une perte importante de protéine, mais une fraction restait en solution. Ce fragment encore en solution précipitait à nouveau si nous le concentrons à plus de $\sim 5\text{mg/mL}$.

Pour séparer le fragment purifié de la MBP, de la TEV et de la protéine fusion non clivée nous avons injecté le surnageant sur une colonne échangeuse d'anions qui sépare les protéines selon leur charge de surface. Les protéines se décrochent quand leurs charges sont écrantées, au moins partiellement. Ceci présente un intérêt pour la cristallogénèse : une fois les charges écrantées, ces protéines se retrouvent dans un régime attractif, ce qui favorise la nucléation de cristaux (McPherson, 2017).

Lors de l'éluion, 2 populations se sont distinguées (Figure 30.A). La première formait un pic relativement fin et était constituée du fragment pur et mono-disperse. La taille de cette population sur gel natif était inférieure à celle du monomère de la BSA ($\sim 66\text{ kDa}$), ce qui pourrait correspondre à un trimère (66 kDa), un dimère (44 kDa) ou un monomère (22 kDa). La seconde formait un large pic, et était constituée du fragment légèrement contaminé et ne formait pas de bande visible sur gel natif (Figure 30.B&C). Cette absence pourrait être due au fait que les objets formés seraient trop gros pour rentrer dans le gel. Potentiellement ce 2nd pic pourrait être un début de précipité.

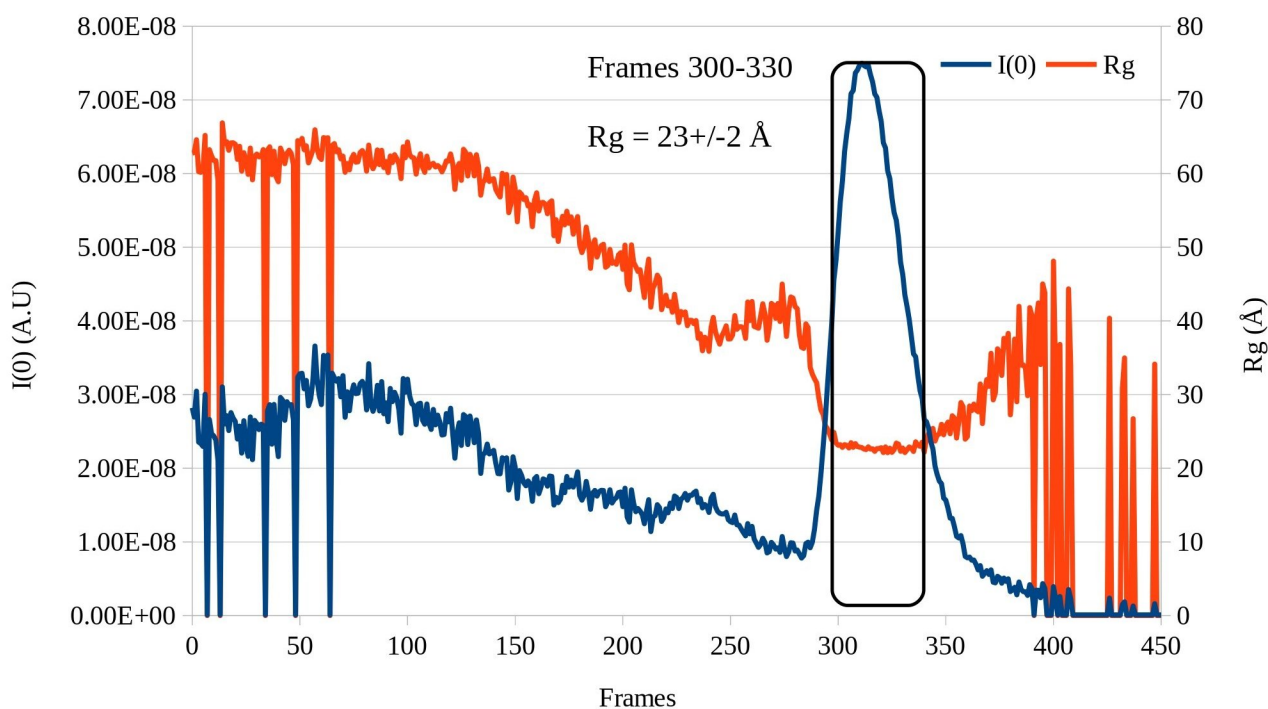


Figure 31 : Analyse SAXS du fragment CobGly1.

Courbes du R_g et $I(0)$. La partie des courbes correspondant à la population avec un R_g stable est encadrée en noir. Le calcul du R_g s'est fait en sélectionnant la partie $q \in [0.03; 0.05]\text{Å}^{-1}$ des frames du cadre noir.

La première population sur ce chromatogramme nous paraissait donc beaucoup plus intéressante à étudier, puisque pure et homogène en taille sur le gel natif. Nous l'avons isolée pour faire des analyses SAXS et des essais de cristallisation.

Les analyses SAXS sur cette population ont confirmé son homogénéité en taille (Figure 31). En effet, elle a un R_g stable à $23 \pm 2\text{Å}$. Pour une protéine globulaire, cela correspondrait à un objet de $\sim 40\text{kDa}$

(Smilgies & Folta-Stogniew, 2015), soit proche d'un dimère de fragment CoBaHMA-GlyZip1. Bien que le fragment CoBaHMA-GlyZip1 ne soit pas nécessairement globulaire, ces résultats de SAXS, combinés aux résultats des gels natifs sur le fragment et la protéine fusion, indiquent clairement que le fragment CoBaHMA-GlyZip1 peut oligomériser sous forme de dimère. La présence d'autres populations à plus haut poids moléculaires sur le chromatogramme et le gel natif de la protéine fusion semble indiquer que de plus haut degrés d'oligomérisation sont accessibles pour ce fragment.

L'absence du Glyzip2 et GlyZip3 est une différence majeure avec la calcyanine complète, mais devant ces résultats sur le fragment CoBaHMA-GlyZip1 ainsi que ce que nous avons observé sur la protéine complète, nous pouvons considérer que la calcyanine complète est une protéine qui peut oligomériser.

La précipitation du fragment CoBaHMA-GlyZip1 lors de la protéolyse à la TEV, quant à elle, est un phénomène qui est pour l'heure difficile d'expliquer.

Tout d'abord, l'enrichissement du fragment dans le précipité issu de la protéolyse, ainsi que la précipitation du fragment pur au delà de ~ 5 mg/mL indique que ce phénomène n'est très certainement pas dû à une interaction avec d'autres protéines.

Nous avons essayé de comprendre si ce comportement de précipitation pouvait être relié à la séquence du fragment. Pour cela nous avons utilisé Protein-Sol, un outil qui évalue la propension d'une protéine à rester en solution à partir de sa seule séquence. Cet outil se base sur le travail de Niwa *et al.* qui ont quantifié la proportion d'un grand nombre de protéines d'*E. coli* à rester en solution après expression dans un système cell-free dépourvu de chaperonnes, suivie d'une centrifugation à 21600g (Niwa et al., 2009). Cet outil calcule un Predicted Scale Solubility, qui représente la proportion de protéine qui resterait en solution lors d'un protocole similaire à celui décrit par Niwa *et al.*. Pour les protéines d'*E. coli* la moyenne est de 0,45 (Hebditch et al., 2017).

Toutefois, dans notre cas, les scores obtenus avec ce logiciel sont surprenants. Dans le cas de la protéine fusion ce score est de 0,343; pour la MBP et le linker qui la relie au fragment, ce score est de 0,501; et pour le fragment CoBaHMA-GlyZip1 ce score est de 0,568. Pour comparaison, la calcyanine complète étiquetée 6His à un score de 0,558. Il paraît très surprenant que la protéine fusion ait un score plus bas que le fragment individuel quand il a été démontré que la MBP augmente fortement la capacité des protéines auxquelles elle est fusionnée à rester en solution (Kapust & Waugh, 1999). Devant ces résultats il y a 2 possibilités : soit ce logiciel n'est pas capable d'étudier correctement le fragment de la calcyanine, soit la précipitation du fragment individuel n'est pas liée à sa séquence.

Cette précipitation pourrait avoir une origine plus structurale. La protéine fusion MBP-CoBaHMA-GlyZip1 se partage en 2 grandes populations : un probable dimère, et une population de plus haut poids moléculaire. Lors du clivage par la protéase TEV ces 2 populations sont clivées. Après ce clivage, nous retrouvons le dimère probable de fragment, par contre la population de plus haut poids moléculaire n'est plus visible sur gel natif.

Il est possible que le précipité formé lors de la protéolyse par la protéase TEV soit cette population de haut poids moléculaire, qui précipiterait une fois séparée de la MBP.

La précipitation du fragment au delà de ~ 5 mg/mL, quant à elle, pourrait alors s'expliquer de 2 manières. Soit comme un changement d'état d'oligomérique du fragment avec la concentration, comme le font certaines protéines comme LptA (Merten et al., 2012), avec un passage du dimère soluble à

l'oligomère qui précipite, soit un seuil de saturation en solution qu'atteindrait le fragment à cette concentration (Trevino et al., 2008).

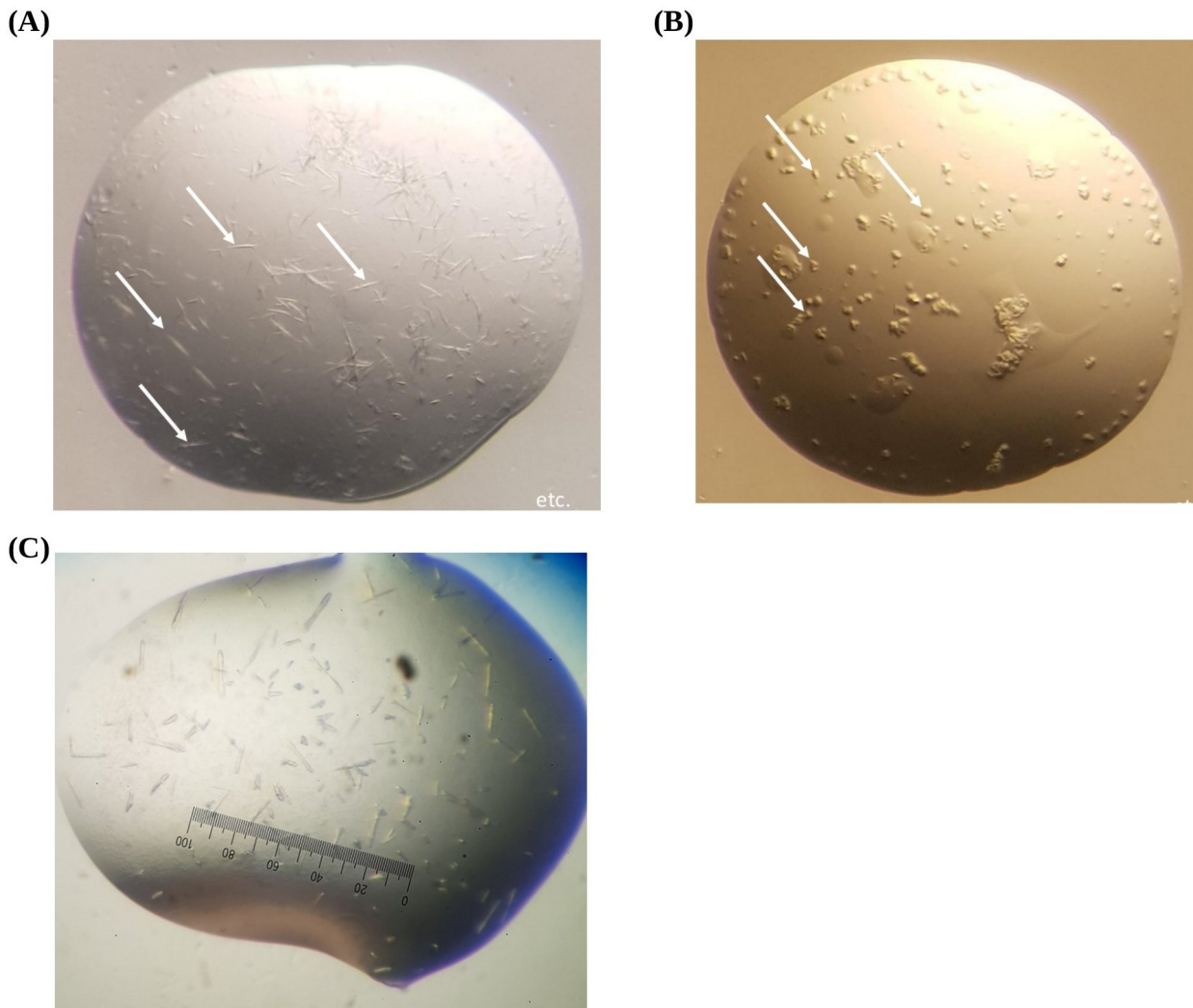


Figure 32 : Essais de cristallogénèse sur le fragment CoBaHMA-GlyZip1.

(A) Conditions : 0,2 M Potassium Sodium tartrate; 20% (w/v) PEG 3350. (B) 0,1 M HEPES pH 6,5; 5% (w/v) PEG 6000. (C) Réplicat de la condition en (A).

Bien que nous eûmes perdu une quantité non négligeable de protéine à l'issue de la précipitation après la protéolyse à la protéase TEV, nous en avons tout de même suffisamment en solution pour mener des essais de cristallisation.

Nous avons donc une protéine pure, mono dispersée, concentrée à un niveau proche de sa précipitation, dans une solution où le contenu en sel écranait au moins partiellement ses charges, et la plaçait dans un régime attractif. C'était donc un excellent candidat pour des essais de cristallogénèse.

Une première série d'essais de cristallisation, menée avec Julien Henri (Institut de Biologie Paris-Seine, UMR 7238 CNRS, Biologie Computationnelle et Quantitative (LCQB), Sorbonne Université),

sur une large gamme de conditions ont montré la présence d'objets qui pourraient être des cristaux dans 2 conditions :

_ 0,2M Potassium Sodium Tartrate et 20% (w/v) PEG (PolyEthylène Glycol) 3350 (JCSG I 22) (Figure 32.A);

_ 0,1M HEPES pH 6,5 et 5% (w/v) PEG 6000 (JCSG III 44) (Figure 32.B)

Dans le premier cas, les objets formés étaient des aiguilles (Figure 32.A). Dans le second les objets formés semblaient majoritairement amorphes, avec des facettes possiblement cristallines (Figure 32.B).

Seul les objets de la première condition de cristallisation ont pu être répliqués dans un plus grand volume (Figure 32.C). Ils ont été analysés par diffraction des rayons X au synchrotron SOLEIL.

Ces objets étaient bien des cristaux de protéine, et non des cristaux de sel. Les premières mesures permettaient d'estimer une maille en groupe d'espace P1. Cependant il n'a pas été possible d'obtenir des tâches de diffraction à moins de 30 Å sur ces cristaux, ce qui ne permettait pas d'obtenir une structure 3D pour le fragment.

Ces résultats, s'ils ne sont pas concluants, sont néanmoins très encourageants puisqu'ils ouvrent la voie à une détermination de la structure de ce fragment. Répéter cette cristallisation pour augmenter la taille des cristaux ou affiner les conditions de précipitations pourrait permettre, à terme, d'avoir des cristaux plus parfaits et plus gros, et donc plus exploitables (McPherson, 2017).

Chapitre 3 : Les protéines à domaine CoBaHMA.

Le domaine CoBaHMA est présent sur d'autres architectures protéiques, principalement associé à des domaines membranaires.

Comme décrit précédemment, le domaine CoBaHMA a pour particularité d'être le seul domaine des calcyanines qui présente des similarités de séquence et de structure avec des domaines déjà décrits (HMA, YAM...) de la superfamille HMA (Cf Chapitre 1 & Benzerara et al., 2022). Les domaines HMA sont présents sur différentes architectures, 799 selon l'entrée Interpro (Paysan-Lafosse et al., 2023) IPR006121 (HMA). Pour ne citer que quelques exemples, le domaine HMA peut se trouver seul (Hearnshaw et al., 2009), répété plusieurs fois (UniProt ID : Q59HD1), associé à une P1B-type ATPase (Bitter et al., 2022), associé au domaine de transport du mercure MerT (UniProt ID : A0A076HZU1)... Cette apparente diversité peut en fait se restreindre à un contexte fonctionnel d'interaction avec des cations métalliques divalents (Bull & Cox, 1994). Les domaines HMA peuvent ainsi jouer le rôle de chaperonnes (Hearnshaw et al., 2009) ou de domaines auxiliaires de régulation de transporteurs de métaux comme les P1B-type ATPases (Smith et al., 2014) ou d'enzymes de détoxification (comme par exemple l'oxido-reductases MerA (Ledwidge et al., 2005)).

L'analyse des architectures protéiques associées à un domaine, permet, de manière plus générale, de préciser le contexte fonctionnel de ce domaine, comme cela a, par exemple, pu être fait pour le domaine LOTUS, une famille de domaines associée au métabolisme des ARN (Callebaut & Mornon, 2010).

Dans le but de mieux comprendre le domaine CoBaHMA, nous avons cherché à savoir si ce domaine est également présent dans d'autres architectures protéiques que les calcyanines, afin d'en déduire des caractéristiques ainsi que des hypothèses fonctionnelles.

Ci-après, je décris un résumé succinct de l'article, dont le manuscrit est présenté en Annexe 2, et qui sera prochainement soumis pour publication. Toutes les figures qui sont citées dans ce chapitre sont issues de cet article. Pour plus de détail, les lecteurs sont invités à s'y référer.

Notre processus pour rechercher des domaines CoBaHMA est schématisé sur la Figure 2 de l'article. Nous avons utilisé les 15 séquences de domaines CoBaHMA de calcyanine, présentées dans Benzerara *et al.* 2022, comme sondes que nous avons fournies au programme HHblits (Remmert et al., 2012). HHblits est un outil qui permet de rechercher des similarités de séquences entre une sonde et une banque de séquences par comparaison de HMMs (Hidden Markov Model). Les HMMSs sont des versions condensées d'alignements de séquences, qui spécifient pour chaque position la probabilité d'observer chacun des 20 a.a dans des séquences apparentées. Les méthodes qui se basent sur des alignements HMM-HMM (ou profil-profil) sont les plus performantes pour la recherche de parentés éloignées. Avec HHblits, nous avons fouillé de manière itérative (8 itérations) la base de données UniClust30 (Mirdita et al., 2017), une version clusterisée à 30% d'identité de séquence de la base de séquences de protéines UniProt (The UniProt Consortium et al., 2023). Le recours à une base de données clusterisée permet de faciliter l'analyse et sonder de manière exhaustive la diversité des architectures existantes dans la banque d'origine, tout en réduisant les biais dus à la redondance présente dans cette dernière.

De cette manière, nous avons identifié 38444 protéines ayant au moins un segment similaire à un des 15 domaines CoBaHMA des calcyanines. Malheureusement, après inspection de la structure de quelques unes de ces séquences, nous avons constaté qu'une grande majorité d'entre elles étaient des domaines HMA. Sur la base des séquences, il était donc très difficile de distinguer ces 2 familles de domaines. Cependant, comme nous l'avons montré dans le Chapitre 1 de cette thèse, ces 2 domaines ont des structures 3D proches, mais différentes : le domaine CoBaHMA a un brin β supplémentaire par rapport au domaine HMA. Nous avons alors tiré parti du récent développement d'AlphaFold2, qui a donné naissance à la base de données AlphaFoldDB (Varadi et al., 2022). Cette base de données rassemble les modèles de structure 3D d'un très grand nombre de protéines de la base de données UniProt. Nous nous sommes servis de ces modèles pour discriminer finement et de manière automatique les domaines HMA des domaines CoBaHMA sur la base de critères structuraux. Pour chaque segment de séquence identifié par HHblits comme étant similaire en termes de séquence aux domaines CoBaHMA, nous avons extrait la structure 3D d'AlphaFoldDB. Nous avons ensuite traduit cette structure sous forme d'un enchaînement de structures secondaires *via* l'outil DSSP (Kabsch & Sander, 1983; Touw et al., 2015). Nous avons filtré toutes les structures, principalement sur la présence de l'enchaînement $\beta\beta\alpha\beta\alpha\beta(\beta)$ typique des domaines CoBaHMA. Les séquences qui passaient ce filtre structural étaient considérées comme étant des domaines CoBaHMA.

Les nouveaux domaines CoBaHMA identifiés de cette manière ont alors été utilisés à leur tour comme sondes pour répéter cette recherche de manière transitive. Ce processus a été répété 6 fois, jusqu'à ce que nous ne trouvions plus de nouveaux domaines CoBaHMA et que nous atteignions ainsi la convergence de ce processus. De cette manière, nous avons identifié 2305 domaines CoBaHMA répartis dans 2227 séquences.

Nous avons ensuite cherché à identifier les zones conservées et variables des domaines CoBaHMA. Afin d'éviter la sur-représentation d'une sous-famille du domaine CoBaHMA, nous avons clusterisé les 2305 séquences que nous avons identifiées pour ce domaine, avec l'outil mmseqs2 (Steinegger & Söding, 2017) (seuil de couverture 80%; seuil d'identité 60%). Nous avons ensuite aligné les représentants des 1432 clusters ainsi produits. Nous avons constaté dans le processus d'alignement, que les hélices α du domaine CoBaHMA sont trop variables pour être alignées. A l'inverse, (Figure 3 de l'article) les brins β , à l'exception du β_4 , ont tous des sites conservés.

Contrairement à ce que nous avons observé chez les domaines CoBaHMA des calcyanines, le motif HxxxxRxxR complet n'est pas totalement conservé, seul 27 % des séquences l'ont. Sur ce motif les 2 arginines centrales sont très conservées (88 % des séquences), et 94 % des séquences ont au moins l'une de ces 2 arginines, quand l'histidine et l'arginine aux extrémités peuvent être remplacées par d'autres a.a polaires. Par ailleurs, une sous-population minoritaire des domaines CoBaHMA n'a aucun des a.a de ce motif basique.

Les séquences incluant un domaine CoBaHMA sont quasi exclusivement des séquences bactériennes, à l'exception d'une séquence eucaryote, et de 4 séquences d'archée, mais dont l'origine semble douteuse au vu du très faible nombre de séquences concernées.

Nous avons regroupé ces séquences avec l'outil mmseqs2 (Steinegger & Söding, 2017), et l'algorithme de Louvain (Blondel et al., 2008) en communautés de séquences similaires de longueurs proches. Nous nous sommes focalisés sur les 48 communautés qui comptaient au moins 5 membres, afin d'avoir

suffisamment de séquences pour produire des MSA et ainsi être en mesure de déterminer les conservations et analyser finement les architectures. Ces communautés rassemblent au total 1410 séquences.

Nous avons analysé ces communautés pour identifier dans leurs séquences les domaines déjà répertoriés dans les banques de domaines à l'aide des outils InterProScan (P. Jones et al., 2014) et DomainMapper (Manriquez-Sandoval & Fried, 2022)).

Comme le domaine (GlyZip)₃ des calcyanines (Benzerara et al., 2022) est absent des bases de domaines, nous avons aussi scanné ces communautés avec l'outil pCALF, développé au laboratoire par Maxime Millet. Cet outil permet de reconnaître les 3 GlyZips du domaine (GlyZip)₃. Cela nous a permis d'identifier les calcyanines de notre ensemble de séquences.

De plus, nous avons recherché les segments transmembranaires dans ces communautés avec l'outil DeepTMHMM (Hallgren et al., 2022). Au delà de l'intérêt général d'identifier les protéines transmembranaires, nous avons pour hypothèse (cf Chapitre 1) que le domaine CoBaHMA puisse interagir avec des membranes. Cette recherche nous permettait d'avancer sur cette hypothèse.

Enfin, nous avons complété ces analyses de séquences par une étude des modèles complets de la structure 3D du représentant de chaque communauté.

Nous avons ainsi identifié que le domaine CoBaHMA est présent dans d'autres architectures protéiques que la calcyanine (Figure 5 de l'article). En effet le domaine CoBaHMA est observé principalement :

- _ Seul;
- _ Répété à 2 ou 3 exemplaires dans la même protéine;
- _ Associé à des domaines de type P-type ATPase;
- _ Associé à des domaines ABC transporteur de type IV;
- _ Associé à des domaines phosphatidylglycerol phosphatase de type 2 (PAP2);

A cela s'ajoutent des architectures représentées par un petit nombre de séquences, comme par exemple des domaines CoBaHMA associés à des domaines diacylglycerol kinase.

Enfin une part non négligeable des domaines CoBaHMA sont associés à des régions encore non décrites dans les bases de domaines.

Le domaine CoBaHMA se trouve donc associé à au moins 3 domaines membranaires (P-type ATPase, ABC transporteur de type IV, PAP2). Deux de ces domaines sont des transporteurs, qui ont fait l'objet d'une brève description dans l'introduction.

L'association la plus représentée dans notre ensemble de données, est CoBaHMA - P-type ATPase, avec une forte prévalence de CoBaHMA - P1B-type ATPase (Figure 7 de l'article). Les P1B-type ATPases ont été abondamment décrites comme des transporteurs de cations métalliques (Smith et al., 2014).

Elles présentent les 4 domaines canoniques des P-type ATPases :

- _ Le domaine T, qui est constitué de 6 hélices transmembranaires qui permet la translocation du substrat. Chez les P1B-type ATPase, un site CPC situé sur ce domaine permet l'interaction avec les cations métalliques divalents ;
- _ Le domaine N, pour Nucleotide-binding;
- _ Le domaine P, pour Phosphorylation, et qui permet, avec le domaine N, l'interaction avec l'ATP;

_ Le domaine A, pour Actuator qui transmettrait l'information d'interaction avec l'ATP au domaine T; (Palmgren & Nissen, 2011)

Dans le cas des P1B-type ATPases, s'ajoute au moins un domaine supplémentaire :

_ Le domaine S, qui est constitué de 2 hélices associées au domaine T. L'une de ces 2 hélices est amphiphatique et est supposée jouer un rôle dans l'entrée des ions transportés par cette protéine (Bitter et al., 2022);

_ A cela peut enfin s'ajouter en N_{ter} un ou plusieurs domaines HMA dont les rôles sont pour l'heure mal compris, mais qui pourraient avoir un rôle régulateur en interagissant avec le domaine A et/ou en amenant les cations métalliques jusqu'au domaine S (Bitter et al., 2022).

De façon inédite, les P1B-type ATPases associées à un ou plusieurs domaines CoBaHMA n'ont pas nécessairement le motif conservé CPC dans leur domaine T. Ce motif peut être remplacé par d'autres motifs conservés. Cela suggère que l'architecture CoBaHMA - P1B-type ATPase n'est pas spécifique du transport de cations métalliques divalents.

Notre analyse a aussi montré qu'au contraire des domaines HMA qui apparaissent limités au P1B-type ATPases, les domaines CoBaHMA sont également observés occasionnellement associés à des ATPases à Ca²⁺ (famille P2A, SERCA).

Comme évoqué ci-dessus, un des rôles possibles pour le HMA serait d'interagir avec le domaine A dans un mécanisme de régulation du P1B-type ATPase. De même, il est possible que les domaines CoBaHMA soient des domaines régulateurs dans cette architecture CoBaHMA - P-type ATPase, par une interaction avec un ou plusieurs des domaines de la P-type ATPase.

La 2^{ème} architecture la plus présente, est CoBaHMA - ABC type IV (Figure 8 de l'article). Les ABC type IV, anciennement type I, sont des domaines membranaires qui exportent des substrats. Ils sont constitués d'un domaine transmembranaire et d'un domaine NBD (Nucleotide Binding Domain), et forment des homo ou des hétéro-dimères (Thomas et al., 2020). Les ABC de type IV exportent une large gamme de substrats, mais ceux observés associés au domaine CoBaHMA sont proches en termes de séquences d'exporteurs de lipides comme MsbA (Mi et al., 2017). De fait, il est possible que l'association CoBaHMA - ABC type IV soit aussi spécifique du transport de lipides. Le domaine CoBaHMA pourrait avoir un rôle régulateur en interagissant directement avec les domaines du transporteur ABC, et/ou interagir avec les têtes chargées négativement des lipides de la membrane, comme supposé dans le Chapitre 1.

Enfin, les domaines PAP2 sont des domaines membranaires qui catalysent la déphosphoylation de lipides ou de carbohydrates (Figure 8 de l'article). Ceux associés aux domaines CoBaHMA appartiennent à la famille LPT (lipid phosphatase/phosphotransferase), et interagissent avec de nombreux lipides dans des bactéries Gram-negative.

La co-occurrence du domaine CoBaHMA avec des domaines membranaires, et plus spécifiquement de domaines qui interagissent directement avec des lipides (ABC type IV et PAP2) suggère une interaction entre le domaine CoBaHMA et ces derniers. Comme nous l'avons déjà développé dans le Chapitre 1, le domaine CoBaHMA pourrait interagir avec les têtes chargées négativement des lipides par ses a.a basiques conservés, d'une manière similaire au domaine eucaryote C2.

Notre analyse a permis de montrer que le domaine CoBaHMA n'était pas limité aux cyanobactéries, mais se trouvait aussi dans de nombreux phylums de bactéries (Figure 6 de l'article). Or bactéries et

cyanobactéries ont des lipidomes très différents (Wada & Murata, 2004). Il n'y a que 2 types de lipides avec une tête chargée négativement qui sont communs aux lipidomes des cyanobactéries et des bactéries (autres que cyanobactéries) : le phosphatidylglycerol (PG) et l'acide phosphatidique (PA). Si le domaine CoBaHMA interagit bien avec des lipides, l'hypothèse la plus vraisemblable est que cette interaction implique l'un ou l'autre de ces candidats.

Cependant, les séquences avec une région annotée par un autre domaine que le domaine CoBaHMA ne représentent qu'une partie de notre ensemble de séquences. Nous avons réalisé une analyse exhaustive des modèles de structure 3D des séquences où seul un domaine CoBaHMA était annoté. Dans ces modèles, nous avons souvent observé en C_{ter} du domaine CoBaHMA un court domaine constitué d'une épingle à cheveux de 2 hélices (Figure 9 de l'article), qui est proche en terme de structure du motif GlyZip de la calcyanine. Ce court domaine est parfois identifié comme membranaire par DeepTMHMM, probablement en raison du fort taux d'a.a hydrophobes. Toutefois il porte souvent également des a.a chargés en particulier basiques. Ce court domaine pourrait servir de « plateforme » sur laquelle des a.a pourraient être greffés pour interagir spécifiquement avec des ligands ou des ions. Cette épingle à cheveux pourrait par ailleurs constituer un motif structural de base pour la constitution d'architectures protéiques plus complexes telle qu'observée pour le domaine (GlyZip)₃ des calcyanines ou d'autres domaines.

En conclusion, nous avons identifié, décrit et analysé la famille de domaines CoBaHMA. Les architectures sur lesquelles ces domaines sont observés permettent de proposer des hypothèses sur le rôle du CoBaHMA, et par conséquent, de progresser dans la compréhension de la fonction de la calcyanine.

Cette analyse, délicate en raison de la proximité en termes de séquences des différentes familles au sein de la superfamille HMA, a pu être menée à bien grâce à la mise en place d'un filtre structural pour discriminer finement les domaines CoBaHMA. En intégrant l'information structurale dans notre approche, nous avons pu étudier les architectures connues (P1B-type ATPase, ABC..) mais aussi nous intéresser aux régions encore non-décrites en apportant des observations inédites pour ces dernières, qui représentent encore des régions vierges d'annotations structurales et fonctionnelles. Si nous nous étions contentés de l'approche par séquence nous n'aurions, par exemple, pas observé l'épingle à cheveux de 2 hélices qui se trouve régulièrement associée aux domaines CoBaHMA.

Toutefois ce filtre structural apporte aussi la principale limitation à notre étude. 25% des séquences identifiées par HHblits comme ayant potentiellement un domaine CoBaHMA n'ont pas de modèle dans la base AlphaFoldDB les excluant donc de notre analyse. Par ailleurs, pour limiter la présence de faux positifs, nous avons été très sélectifs sur notre filtre structural, ce qui a induit des faux négatifs. Pour ces raisons, notre analyse n'est pas exhaustive. Nous pourrions avoir manqué des architectures incluant le domaine CoBaHMA. Notre analyse pourrait être complétée dans le futur, par exemple en utilisant ESM Metagenomic Atlas (Lin et al., 2023) plutôt qu'AlphaFoldDB comme base de données de modèles, car elle rassemble 772 millions de modèle, contre 200 millions pour AlphaFoldDB.

Par ailleurs, dans cet article nous n'avons pas encore pleinement exploité le potentiel des modèles 3D. Nous avons développé, en plus du travail de l'article, un outil pour détecter automatiquement les contacts inter-domaines uniquement dans les régions des modèles d'AlphaFold2 jugées comme fiables. Pour cela nous prenons en compte les distances inter-résidus uniquement pour les a.a où les métriques (pLDDT, PAE; cf Chapitre 1 pour ces métriques) sont au dessus d'un seuil de confiance. Nous n'avons

malheureusement pas eu le temps d'intégrer cet outil à notre étude sur le domaine CoBaHMA. Un exemple de l'utilisation de cette métrique est cependant donné en Figure Supplémentaire 8 de l'article, illustrant un réseau d'interaction entre le domaine CoBaHMA et le domaine A et le domaine T d'un membre de la communauté C140 (P1B-type ATPase). Mais il serait intéressant dans le futur d'établir, de manière automatique, les contacts potentiels établis entre les domaines CoBaHMA et d'autres domaines. Cela permettrait d'aller plus en avant dans la formulation d'hypothèses fonctionnelles de ce domaine, notamment en mettant en lumière des a.a impliqués dans des interactions inter-domaines.

Conclusion et Discussion :

Dans l'introduction de cette thèse, nous avons soulevé un certain nombre de questions auxquelles nous espérons apporter des éléments de réponse. En rassemblant les résultats que nous avons obtenus ainsi que les nombreuses hypothèses que nous avons soulevées au cours de ce travail, nous pouvons discuter de certains points.

Tout d'abord, nous avons progressé dans la compréhension du domaine CoBaHMA. La concordance de nos approches de modélisation (modélisation comparative, AlphaFold2 et ESMFold) nous permet de proposer un modèle robuste de la structure 3D de ce domaine CoBaHMA, représentant une nouvelle famille dans la superfamille HMA. Grâce à ce modèle, nous avons développé une nouvelle approche pour l'identification de familles de domaines protéiques, qui repose sur l'intégration des informations séquentielles et structurales pour distinguer une famille de domaines de ses familles proches. Par cette méthode nous avons réussi à discriminer finement le domaine CoBaHMA des autres domaines de la superfamille HMA. A partir de cela, nous avons décrit la variété des organisations modulaires dans lesquelles se trouve le domaine CoBaHMA. Cette analyse a mis à jour plusieurs architectures, voir même domaines, encore non décrits dans la littérature.

Notre analyse du domaine CoBaHMA est la plus complète (et la seule) faite à ce jour, mais il est probable qu'elle ne soit pas exhaustive. En effet, notre méthodologie (choix des bases de données, filtre sur la structure 3D...) a privilégié la spécificité à la couverture pour éviter la contamination de l'ensemble de données par des faux positifs. Nous avons donc peut-être manqué des architectures possédant le domaine CoBaHMA, dont la description permettrait d'encore affiner notre analyse. Cependant, même avec cette possible extension, on peut d'ores et déjà constater que les domaines CoBaHMA ne sont présents que dans un ensemble relativement restreint de protéines au sein de différentes familles, ce qui suggère une fonction spécifique au sein de ces familles, limitée à certaines espèces ou phylums.

Grâce à l'analyse des modèles 3D, nous avons constaté que le domaine CoBaHMA est couramment associé à un motif en épingle à cheveux constituée de 2 hélices. Cette association est celle que nous avons modélisé pour la structure du fragment CoBaHMA – GlyZip1, que nous avons isolé depuis la calcyanine complète par protéolyse limitée. La récurrence de cet assemblage structurale est peut-être l'indication que nous devrions considérer la calcyanine comme l'association de 2 blocs : CoBaHMA-GlyZip1 d'un côté et GlyZip2- GlyZip3 de l'autre. Cela pourrait expliquer pourquoi nos essais pour exprimer les domaines CoBaHMA et (GlyZip)₃ de la calcyanine n'ont pas pu pleinement aboutir (Targowla, 2019, Rapport de M2).

Nous avons aussi observé que le domaine CoBaHMA est souvent présent dans un contexte membranaire, en particulier associé à des protéines permettant le transport de substrats au travers des membranes ou des protéines modifiant les lipides. Un certain nombre de domaines CoBaHMA sont en effet associés à des domaines transmembranaires, en particulier à des familles transporteurs (P-type ATPases, ABC transporteurs). Il est difficile de prédire quels substrats pourraient être transportés par ces architectures. Par analogie avec les P1B-type ATPases à domaine HMA, ces substrats pourraient être des ions. Mais, en se basant sur la proximité des séquences des transporteurs ABC à domaines CoBaHMA avec des transporteurs de lipides comme MsBA (Padayatti et al., 2019), il est aussi possible d'envisager des transports de lipides. Dans le cas des P-type ATPases, ce transport de lipides rappellerait alors celui des P4-type ATPases (dit « flippases ») (Lyons et al., 2020).

L'alignement des 2305 séquences de domaines CoBaHMA, couplé à l'examen du modèle de la structure 3D, nous a permis de mettre en évidence une signature fonctionnelle constituée de résidus basiques conservés sur la face exposée au solvant du feuillet β . Par analogie avec des domaines présents chez les eucaryotes, nous avons formulé l'hypothèse que le domaine CoBaHMA de la calcyanine puisse lier une membrane via une interaction avec des phospholipides chargés négativement. En nous basant sur la considération des lipidomes des cyanobactéries et des bactéries (non-cyanobactéries) (Kralj et al., 2022; Wada & Murata, 2004), nous proposons que le domaine CoBaHMA interagisse avec le phosphatidylglycérol (PG) et/ou l'acide phosphatidique (PA). Une autre possibilité, plus spécifique aux cyanobactéries, concerne le lipide majoritaire des membranes des thylakoïdes, le sulfoquinovosyldiacylglycérol (SQDG), un lipide chargé négativement.

Cette hypothèse permet de réconcilier le comportement de la calcyanine de *S. calcipolaris* avec celui de *Gloeomargarita lithophora*. En effet, il a été constaté que la calcyanine de *G. lithophora*, qui a pour architecture protéique X - (GlyZip)₃, est membranaire dans *E. coli* (Görgen, 2017, Rapport de M2). De même, la construction X - GlyZip1 exprimée récemment au laboratoire se retrouve dans la fraction membranaire de *E. coli*, contrairement à la construction CoBaHMA - GlyZip1. Enfin le domaine X semble être un domaine membranaire, probablement même transmembranaire, du moins chez *E. coli*. En effet, nous avons purifié cette calcyanine en présence de détergent, puis éliminé ce détergent avec de la cyclodextrine, ce qui a induit la précipitation de la protéine. Ce comportement est typique de protéine transmembranaire. Nous pouvons donc émettre l'hypothèse que les domaines N_{ter} de la calcyanine X et CoBaHMA, auraient en commun une capacité d'interaction avec des lipides.

Dans ce contexte, il faut aussi noter que, des essais d'expression de calcyanines à domaine CoBaHMA associées à l'eGFP dans une souche biominéralisante ont conduit à observer une localisation de la protéine aux membranes (Görgen, 2020, Thèse). Par ailleurs, des essais de purification de la calcyanine de *S. calcipolaris* entière menés par nos partenaires du BIAM (Institut de Biosciences et biotechnologies d'Aix-Marseille), en présence de β -D-maltopyranoside de décyle (un détergent), semblaient stabiliser la protéine sous forme de dimère/trimère. Ceci indiquerait qu'une partie de la protéine a besoin d'un environnement hydrophobe pour se structurer correctement.

A partir de ces résultats et observations, une hypothèse plausible est que la calcyanine de *S. calcipolaris* interagit avec des membranes.

La nature précise de cette interaction, transmembranaire ou simplement associée aux membranes, est encore sujette à discussion. En effet, les résultats expérimentaux présentent le domaine CoBaHMA, ainsi que le domaine (GlyZip)₃ comme des domaines cytosoliques, ce qui indiquerait que la calcyanine de *S. calcipolaris* n'est pas transmembranaire mais qu'elle s'associe aux membranes via son domaine CoBaHMA.

Toutefois, notre modélisation des GlyZips du domaine (GlyZip)₃, nous a amenés à formuler l'hypothèse que ces éléments pourraient être transmembranaires. En effet, nous les avons modélisés comme des épingles à cheveux formées de 2 hélices interrompues par un motif GP, avec le motif glycine zipper enfoui à l'interface entre ces 2 hélices. Ces modèles présentent des similarités structurales avec des structures 3D de protéines membranaires, en raison de leur caractère largement apolaire.

Une hypothèse qui permettrait de lever cet apparent paradoxe est que le domaine (GlyZip)₃ pourrait être, chez la cyanobactérie, cytosolique ou membranaire en fonction de sa conformation. La protéine

SmhABC (constitué de 3 composants : A, B et C) qui forme un pore dans une membrane adopte ce comportement. Avant de s'insérer dans la membrane, les composants A et B de cette protéine sont cytosoliques. Ils sont ensuite recrutés par le composant C qui est membranaire, puis opèrent un changement conformationnel pour insérer une de leur partie dans la membrane (Churchill-Angus et al., 2021). La partie du composant B qui s'insère dans la membrane est précisément celle qui a des similarités structurales avec le modèle de la structure 3D des GlyZips de la calcyanine. Nous pourrions donc envisager un comportement similaire pour le domaine (GlyZip)₃. Comme les GlyZips ont une face hydrophile et une face hydrophobe, le domaine (GlyZip)₃ pourrait permuter la face qui se trouve à la surface de la protéine pour passer de cytosolique (face polaire vers l'extérieur, en contact avec le solvant) à membranaire (face hydrophobe vers l'extérieur, en contact avec les lipides). Ce changement de conformation du domaine (GlyZip)₃, s'il existe, pourrait être déclenché par une interaction du domaine CoBaHMA avec la membrane ou par une interaction avec une autre protéine. Si cette hypothèse est vérifiée, la calcyanine de *S. calcipolaris* pourrait être transmembranaire. L'incapacité de la calcyanine de *S. calcipolaris* à s'insérer dans la membrane d'*E. coli* pourrait être imputable à un lipidome non adapté, notamment si la calcyanine interagit avec des lipides spécifiques des cyanobactéries comme le SQDG, ou à l'absence du partenaire protéique.

Enfin, les expériences menées sur le fragment CoBaHMA-GlyZip1 ainsi que sur la protéine fusion MBP-CoBaHMA-GlyZip1 ont montré que ce fragment oligomérisait. Nous n'avons pas pu réaliser de détermination précise de cet état oligomérique, et les expériences de caractérisation que nous avons mis en œuvre à ce sujet ne sont qu'indirectes (SAXS, profil d'élution, gel natif...). Cependant il semblerait que ce fragment puisse former au moins des dimères, voir des objets de plus haut poids moléculaire. Si le fragment peut s'oligomériser, il est très probable que la calcyanine complète de *S. calcipolaris* oligomériser aussi. Cela expliquerait le grand nombre d'objets constaté sur gel natif pour cette protéine quand nous avons essayé de la purifier. Les gels natifs et les profils d'élution de la calcyanine complète témoignent en effet de la présence de gros assemblages. La calcyanine semble tendre vers des objets de haut poids moléculaires (> 15 unités monomères).

Il est vraisemblable que l'interface d'oligomérisation, que ce soit pour le fragment CoBaHMA-GlyZip1 ou la protéine complète, soit localisée au niveau des motifs GlyZip, faisant intervenir la tranche du motif en épingle à cheveux, comme illustré dans un modèle préliminaire que j'ai réalisé d'un dimère de 2 motifs GlyZip1 (Figure 33). En amont de cette thèse, des essais d'expression du domaine (GlyZip)₃ de la calcyanine de *G. lithophora* avaient montré que ce domaine tend à former des objets de différentes tailles, jusqu'à ~25 unités monomériques (Targowla, 2019, Rapport de M2). Ce comportement rappelle fortement celui observé pour la calcyanine de *S. calcipolaris*. Au regard de leur modèle, les GlyZips pourraient s'assembler entre eux et former des assemblages de haut poids moléculaire.

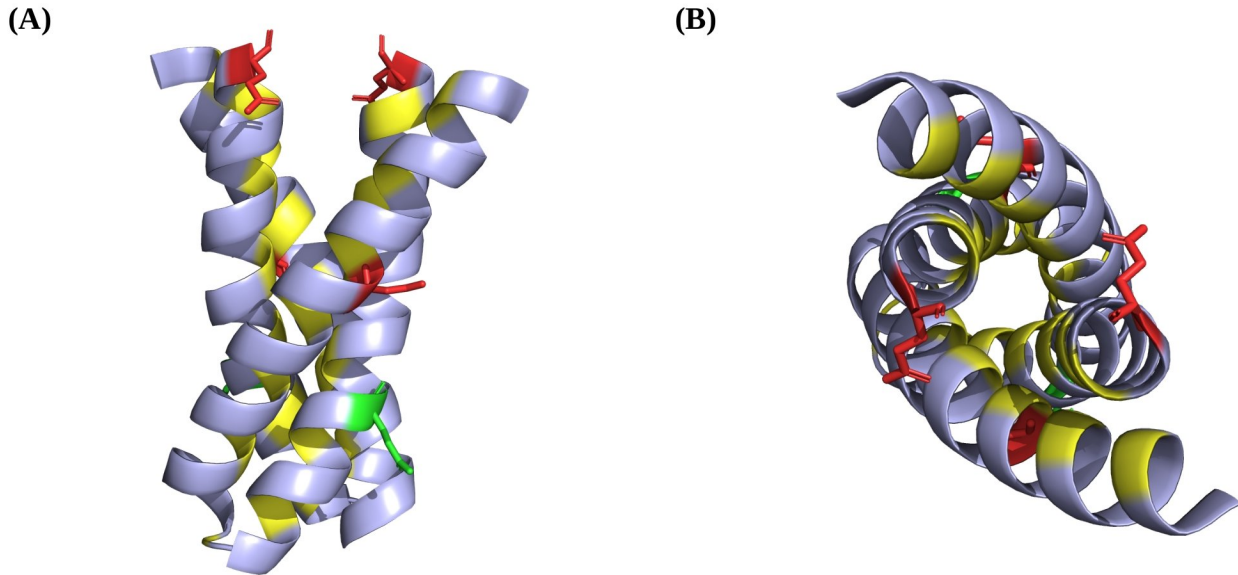


Figure 33 : Modèle d'assemblage de 2 motifs GlyZip.

Modèle d'assemblage de 2 motifs GlyZip vu (A) de côté et (B) de haut. Les glycines sont colorées en jaune. Les a.a acides et polaires conservés sont colorés en vert et en rouge.

Nous avons pour ambition au départ de cette thèse, d'avancer sur la compréhension de la fonction de la calcyanine. Si nous n'avons pas encore abouti, nos résultats nous permettent de mettre en avant 2 grandes hypothèses fonctionnelles:

1. La calcyanine forme une enveloppe protéique autour des iACC via son domaine (GlyZip)₃. Le domaine CoBaHMA lierait une membrane de la cellule, pour des raisons encore inconnues (Figure 34).
2. Le domaine (GlyZip)₃ s'insère dans une membrane lipidique (iACC ou autre), avec l'aide du domaine CoBaHMA, par exemple pour former un pore (Figure 36).

La co-occurrence du seul gène marqueur de cette biominéralisation, *ccyA*, et des iACC incite fortement à supposer que la calcyanine est directement en interaction avec ces derniers.

Si la tendance à former de gros oligomères est bien une propriété de la calcyanine de *S. calcipolaris* nous pourrions envisager que l'assemblage des GlyZips des domaines (GlyZip)₃ forme l'enveloppe qui entoure les iACC dans les cyanobactéries. En effet il a été supposé que cette enveloppe pouvait être, entre autres, de nature protéique (Blondeau et al., 2018).

Dans cette hypothèse, le domaine CoBaHMA pourrait interagir avec des lipides de thylakoïdes ou de la membrane cellulaire. Bien que les iACC ne soient pas systématiquement co-localisés à proximité de membrane (Blondeau et al., 2018), il se pourrait tout de même qu'il y ait un lien entre ces 2 éléments cellulaires. Dans de nombreuses souches, dont *S. calcipolaris* les iACC sont localisées aux pôles de la cellule. Chez ces souches il semblerait que la formation des inclusions soit concomitante de la formation du septum, lors de la division cellulaire (Benzerara et al., 2014). De même, chez des souches où les iACC sont dispersés dans la cellule, certains de ces iACC sont localisés à proximité des

thylakoïdes (Blondeau et al., 2018). Il pourrait donc y avoir un lien entre les membranes lipidiques et les iACC, par exemple au moment de leur formation.

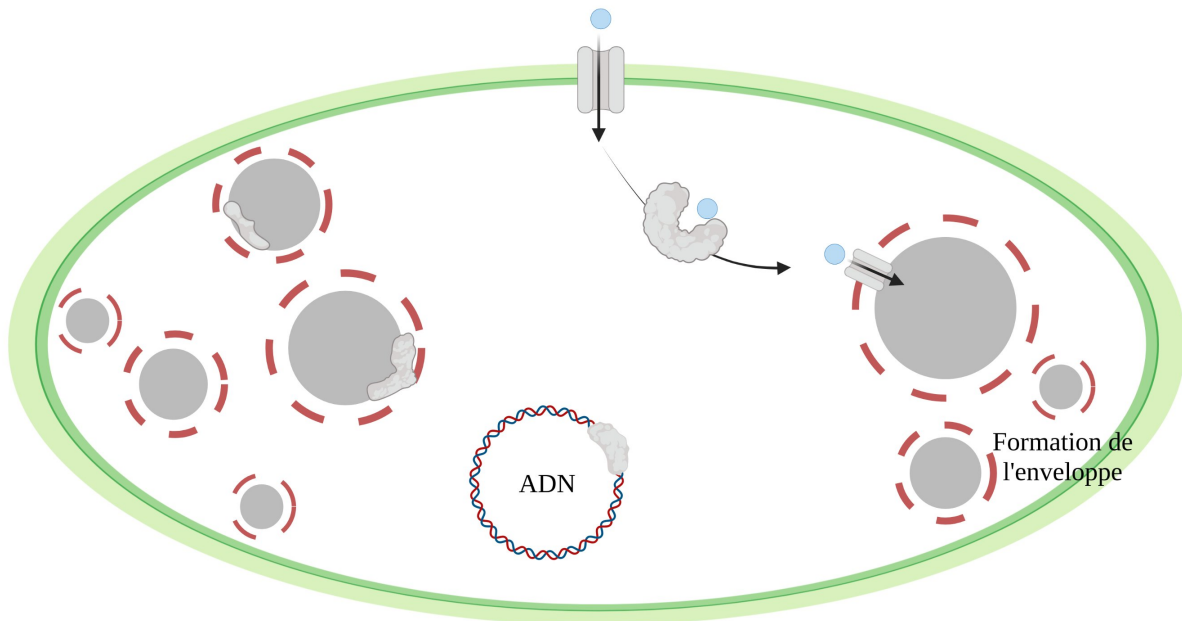


Figure 34 : La calcyanine pourrait former une enveloppe protéique autour des iACC.

Figure produite avec www.BioRender.com.

Il est intéressant de noter à ce sujet que le domaine CoBaHMA de la calcyanine est similaire d'un point de vue structural au domaine BMC (Bacterial Microcompartment Domain; pfam PF00936) qui forme, par oligomérisation, certaines parties de l'enveloppe protéique du carboxysome, le compartiment où a lieu la fixation du carbone chez les cyanobactéries. En effet, ces 2 domaines partagent le même cœur structural caractérisé par l'enchaînement de structure secondaire $\beta\alpha\beta\alpha\beta$ (Figure 35; Kinney et al., 2011).

Comme l'enveloppe autour des iACC a aussi été observée chez des cyanobactéries qui expriment des variants de la calcyanine autres que le variant à domaine CoBaHMA (Benzerara et al., 2022; Blondeau et al., 2018), le domaine CoBaHMA ne peut pas être le seul responsable de la formation de cette enveloppe. Mais il se pourrait que le domaine CoBaHMA soit un module supplémentaire dans la formation de l'enveloppe protéique. Dans ce cas la proximité structurale entre le domaine BMC et le domaine CoBaHMA pourrait fournir des indications sur la façon dont ce dernier interviendrait.

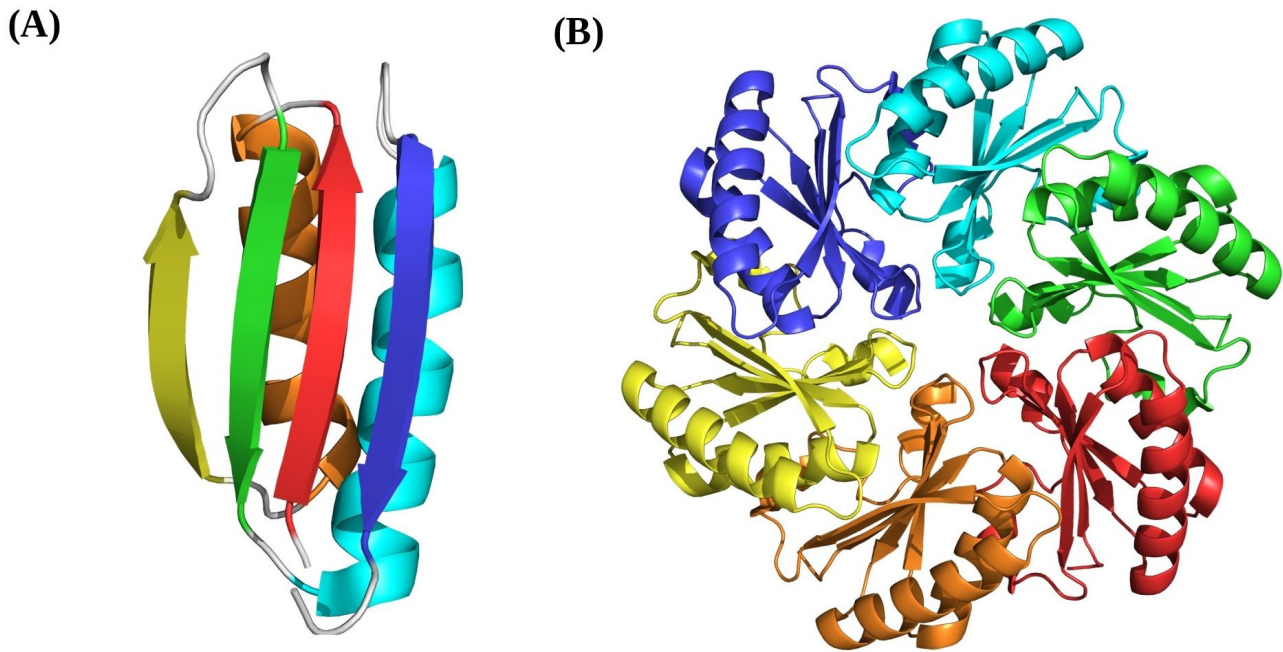


Figure 35 : Structure du domaine BMC.

ID PDB RCSB : 2A1B. (A) Structure expérimentale du domaine BMC. (B) Structure de l'hexamère de domaine BMC qui forme l'enveloppe du carboxysome.

L'autre hypothèse présentée ici, serait que le domaine $(\text{GlyZip})_3$ s'insère dans une membrane lipidique, que ce soit celle qui forme la membrane cellulaire, les thylakoïdes, ou encore la mono-couche qui pourrait entourer les iACC (Blondeau et al., 2018). Le domaine CoBaHMA pourrait alors déclencher le changement de conformation du domaine $(\text{GlyZip})_3$ en reconnaissant les membranes.

Dans ce cas, nous pourrions envisager que les GlyZips s'assemblent en pore dans la membrane, comme les protéines avec qui ils partagent des similarités structurales. Les faces hydrophiles des glycine zippers permettraient le passage de substrat à travers la membrane. Ces pores pourraient permettre le passage de molécules d'eau, impliquées dans la formation des iACCs, ainsi que dans leur structures et propriétés (Goodwin et al., 2010). Ce transport pourrait ainsi assurer les conditions nécessaires au maintien d'iACC sous forme amorphe (Du & Amstad, 2020), alors qu'il pourrait aussi permettre le passage concomitant d'ions, médiés par les a.a acides conservés qui pourraient interagir avec les ions Ca^{2+} , comme c'est le cas dans de nombreuses protéines impliquées dans la biominéralisation. Ce pore pourrait être localisé dans l'enveloppe qui entoure les iACC, ou dans les membranes cellulaires.

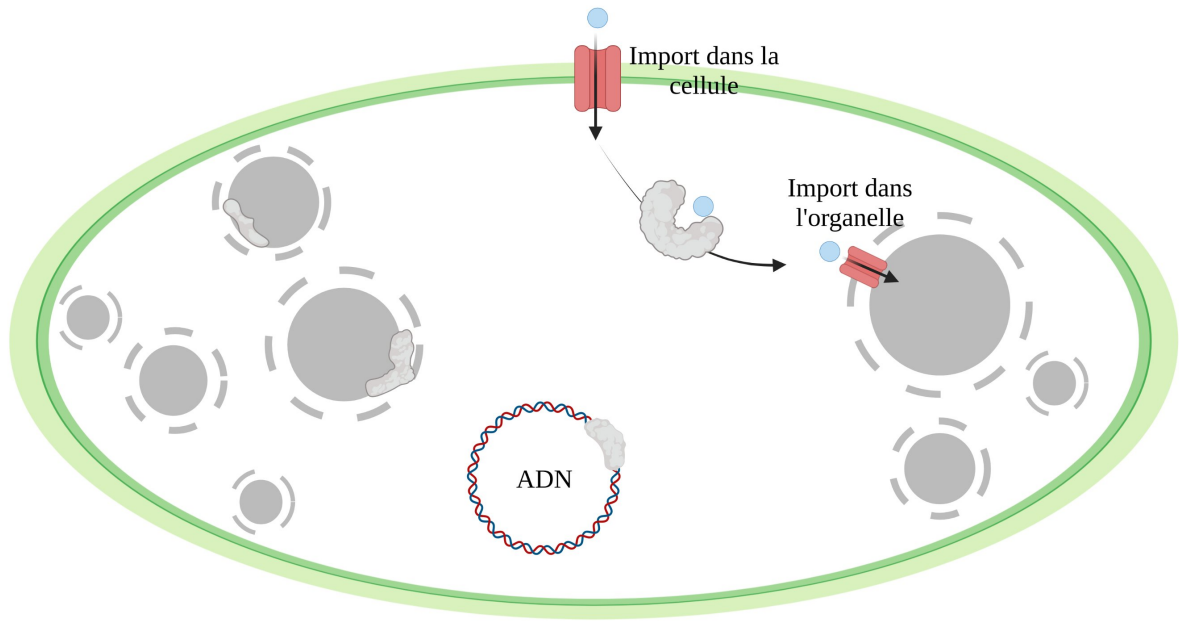


Figure 36 : La calcyanine pourrait être un pore.
Figure produite avec www.BioRender.com.

Perspectives :

Durant cette thèse, nous avons avancé sur la compréhension des propriétés structurales de la calcyanine et avons avancé des hypothèses sur sa/ses fonction(s). Nous proposons ici des expériences qui permettraient de continuer à progresser et aboutir sur ces 2 points.

Enrichissement de la modélisation de la structure 3D de la calcyanine.

Les modèles que nous proposons de la structure 3D de la calcyanine de *S. calcipolaris* sont encore incomplets. Il existe cependant des pistes pour espérer aboutir à des modèles plus aboutis, notamment sur l'assemblage des GlyZips du domaine (GlyZip)₃. Si nous avons utilisé ESMFold et AlphaFold2 durant cette thèse, d'autres logiciels de modélisation ont été publiés à la suite d'AlphaFold2, tel qu'OmegaFold (R. Wu et al., 2022) ou RoseTTAFold (Baek et al., 2021). Plusieurs études ont montré la pertinence de comparaison des résultats de plusieurs logiciels de modélisation lors de prédictions de structure 3D (Aubel et al., 2023; Azzaz et al., 2022; Téletchéa et al., 2023). Il semblerait que selon les protéines, certains logiciels soient plus adaptés que d'autres. Dans le cas de la calcyanine de *S. calcipolaris*, RoseTTAFold pourrait être un excellent candidat. En effet, au vu des hypothèses avancées pour les 2 domaines de la calcyanine, cette protéine pourrait être membranaire par l'intermédiaire de son domaine (GlyZip)₃. Or Azzaz *et al.* ont montré plusieurs exemples de protéines transmembranaires pour lesquelles Robetta modélise avec succès le domaine membranaire, là où AlphaFold2 échoue (Azzaz et al., 2022). De même, ESMFold semble être moins performant que RoseTTAFold pour modéliser les protéines membranaires (Téletchéa et al., 2023).

Par ailleurs, une autre piste pour améliorer le modèle serait de modéliser plus exhaustivement les oligomères de la calcyanine. Nous avons essayé de modéliser la calcyanine sous forme de trimère, sans succès. Cependant, les protéines présentant des similitudes de structure avec les motifs GlyZips sont des protéines qui forment des oligomères avec un haut nombre de sous unités (jusqu'à 10 (J. S. Wilson et al., 2019)). De même, comme nous l'avons supposé, la calcyanine pourrait former des oligomères de haut poids moléculaire. La calcyanine a peut être besoin d'un grand nombre d'unités pour se structurer en homo-oligomère. Cependant, réaliser de tels modèles à haut nombre d'oligomères serait extrêmement gourmand en ressources, surtout avec AlphaFold2.

Outre son intérêt intrinsèque, un modèle plus précis de la calcyanine pourrait être utilisé pour résoudre sa structure par remplacement moléculaire, si nous parvenons à obtenir une carte de diffraction des rayons X à partir de cristaux de la calcyanine ou un de ses fragments.

Structure 3D expérimentale du fragment CoBaHMA-GlyZip1.

La suite immédiate de la partie expérimentale de ce projet serait de répéter la purification du fragment CoBaHMA-GlyZip1 pour obtenir de nouveaux des cristaux. En travaillant sur les conditions et/ou l'échelle pour la cristallisation, nous pourrions aboutir à des cristaux plus gros et plus réguliers qui seraient plus propices pour obtenir une carte de diffraction.

Une telle structure devrait permettre de comprendre les potentielles interactions entre le domaine CoBaHMA et le GlyZip1. Enfin cette structure du fragment pourrait former un oligomère, ce qui devrait permettre de caractériser les surfaces impliquées dans ces mécanismes d'oligomérisation et ainsi, de mieux comprendre le comportement de la protéine complète en solution.

Structure 3D expérimentale de la calcyanine complète de S. calcipolaris.

Si l'approche par fragment fonctionne, nous pourrions exprimer l'autre fragment de la calcyanine de *S. calcipolaris*, c'est à dire la partie GlyZip2 - GlyZip3 de cette protéine pour obtenir l'ensemble de la structure 3D par fragment, et ainsi recueillir des informations utiles quant aux interactions GlyZip/GlyZip et GlyZip/CoBaHMA.

S'il s'avère que le fragment GlyZip2 - GlyZip3 ne peut être exprimé ou purifié, nous pourrions rebasculer sur la calcyanine entière avec de nouvelles stratégies.

Comme évoqué précédemment, des expériences de purification de la calcyanine entière en présence de β -D-maltopyranoside de décyle (un détergent), menées par nos partenaires au BIAM (Institut de Biosciences et biotechnologies d'Aix-Marseille), semblent indiquer que ce détergent stabilise un dimère/trimère de la calcyanine, et limite la présence d'une multitude d'objet en solution. Toutefois, la présence du détergent n'est pas une condition favorable à la cristallogénèse. De plus, un dimère/trimère de la calcyanine (~78-117kDa) est probablement trop petit pour être confortablement caractérisé par Cryo-EM, cette technique n'est pas vraiment appropriée pour des assemblages de moins de 150kDa (Benjin & Ling, 2020). Une dernière solution serait donc la RMN (Résonance Magnétique Nucléaire) mais dans ce cas, elle risque de s'avérer complexe pour des assemblages de cette taille et nécessiterait probablement la mise en œuvre de marquages des protéines (N15 et C13).

Si ce détergent s'avère être trop gênant pour la détermination structurale, nous pourrions exprimer la calcyanine complète de *S. calcipolaris* associée à la MBP, comme nous l'avons fait pour le fragment. La MBP a stabilisé le fragment CoBaHMA-GlyZip1, et elle pourrait stabiliser la protéine complète.

Une autre solution, plus exotique et donc plus difficile à mettre en place, serait d'exprimer la calcyanine de *S. calcipolaris* dans une cyanobactérie, avec une étiquette 6His. L'expression dans une cyanobactérie permettrait de garantir la présence d'un lipidome adapté à une éventuelle interaction avec la calcyanine. De cette manière, nous devrions aussi avoir une meilleure idée de la localisation de la calcyanine dans les cyanobactéries. Par ailleurs, il pourrait y avoir des partenaires protéiques chez les cyanobactéries qui pourraient stabiliser la protéine.

Des souches de cyanobactéries ont déjà été utilisées pour exprimer et purifier des protéines étiquetées 6His, tel que la protéine PsbQ du photo-système II dans *Synechococcus* sp. PCC 6803 (Roose et al., 2007) ou CikA étiquetée 6His dans *Synechococcus elongatus* PCC 7492 (Mutsuda et al., 2003). Cependant, ces systèmes sont encore peu utilisés, et donc mal maîtrisés.

Étude du GlyZip.

Nous pourrions compléter l'analyse bioinformatique des domaines de la calcyanine par l'étude ses motifs GlyZip. Ce type d'étude était très difficile à poursuivre en début de thèse, en raison du caractère original de ce type de motif, ne présentant pas d'homologues dont les structures 3D aient été étudiées expérimentalement. Cependant, le développement de l'outil Pcalf dans notre laboratoire (travaux de Maxime Millet), ainsi que l'émergence d'outils comme FoldSeek (Van Kempen et al., 2022) permettent de fouiller avec une plus grande sensibilité les bases de données.

Aussi, nous pourrions réaliser une étude similaire à celle conduites pour le domaine CoBaHMA, avec une description étendue de la diversité des protéines qui auraient ce motif. A terme, cela pourrait permettre de préciser le rôle fonctionnel de ce type de long motif glycine zipper.

Localisation de la calcyanine dans S. calcipolaris.

Pour préciser la fonction de la calcyanine, il est crucial de déterminer sa localisation cellulaire dans la cyanobactérie. Comme nous l'avons discuté, ce sujet reste très ouvert. Plusieurs techniques sont envisageables dans ce but. L'expression d'une protéine fusionnée à un fluorophore dans les cyanobactéries a déjà été testée comme cela a été décrit précédemment, et s'est avérée particulièrement complexe à mettre en œuvre pour des résultats qui n'ont pas été entièrement concluants (Görgen, 2020, Thèse).

Une autre option serait d'utiliser un anticorps contre la calcyanine, puis de l'identifier soit dans une cellule fixée, soit dans un extrait cellulaire de *S. calcipolaris*.

Dans le cas d'un fractionnement cellulaire, il serait très intéressant d'isoler 3 fractions : les membranes, le cytosol, et les iACC, bien que l'isolement de cette dernière fraction ait été testé au labo par Gabriel Brandt et Neha Mehta, est en cours d'optimisation. Néanmoins cette extraction d'iACC inclut des solvants, comme le chloroforme, qui pourraient avoir un impact sur les protéines.

Interactions de la calcyanine.

Il a été supposé dans cette thèse que la calcyanine interagirait avec des lipides et/ou le calcium. Il serait judicieux de tester ces hypothèses.

Dans le cas des lipides, des expériences de sédimentation pourraient être mises en place (Zhao & Lappalainen, 2012), en choisissant soigneusement les lipides étudiés afin de correspondre au mieux à la membrane des cyanobactéries. Le candidat principal pour ce type d'expériences est le phosphatidylglycérol. Mais il serait aussi intéressant d'étudier le sulfoquinovosyldiacylglycérol et qui est un lipide chargé négativement chez les cyanobactéries.

Pour l'interaction avec le calcium, des expériences de titrage calorimétrique isotherme (Beccia et al., 2015), de fluorescence ou de dynamique moléculaire permettraient de déterminer si interaction il y a.

Durant cette thèse, nous avons fait des premiers essais d'interaction entre la calcyanine et le calcium. Ces essais n'ont pas été entièrement concluants, notamment car ils nécessitent de partir d'une protéine dans une solution sans calcium. Or le calcium est omniprésent dans l'eau, et peut se trouver sous forme de trace sur les plastiques. Il est dès lors très difficile de s'assurer de cette absence de calcium. De plus, la calcyanine était étiquetée 6His, ce qui pouvait fausser les mesures faites. Néanmoins, ces premiers essais ont permis de préciser les paramètres expérimentaux pour une future expérience mieux maîtrisée.

De même, une éventuelle interaction avec des ions carbonates seraient intéressante à étudier.

Nous pourrions aussi chercher des partenaires protéiques à la calcyanine en immunoprécipitant la calcyanine recombinante étiquetée (6 His ou Strep II) avec un extrait cellulaire de cyanobactéries. Nous pourrions identifier les protéines interagissant avec la calcyanine par spectroscopie de masse.

Travail similaire sur d'autres variants.

Sur du plus long terme, nous pourrions répéter l'étude que nous avons mis œuvre dans cette thèse pour les autres variants de la calcyanine, ceux à domaine X, Y et Z.

Pour la calcyanine à domaine X, des essais d'expression et de purification ont déjà été effectués sur la souche de *G. lithophora*. Il semblerait que cette protéine soit propice à une détermination expérimentale. Par contre, il y a très peu de séquences de calcyanine à domaine X répertoriées, avec un domaine X ne présentant pas, de plus, d'apparentes similitudes de séquence avec d'autres séquences, ce qui limite fortement les possibilités d'analyse bioinformatique et de modélisation. Nos quelques essais de modélisation de la structure 3D du domaine X par des logiciels ne requérant, en théorie, pas de MSA comme ESMFold, n'ont pas abouti.

A l'inverse, les calcyanines à domaine Y et Z ont encore été très peu étudiées. Les quelques expériences préliminaires sur des calcyanines avec ces domaines ont montré que ce sont des protéines difficiles à exprimer dans *E. coli*, ce qui semble être un motif très récurrent avec cette famille de protéines. De même, les essais de modélisation de la structure 3D des domaines Y et Z n'ont pas abouti.

La comparaison de ces protéines devraient permettre de mieux comprendre le rôle de la calcyanine et de ses différents domaines. Il est à noter qu'une partie des calcyanine à domaines Y n'ont pas le GlyZip2 (Benzerara et al., 2022). Étudier une de ces séquences pourrait fournir des informations pertinentes sur le rôle structural et fonctionnel de ce GlyZip.

A terme, finaliser l'étude de la calcyanine permettrait de mettre en évidence des caractéristiques structurales et fonctionnelles très originales.

Matériels et Méthodes :

Le Chapitre 3 étant un résumé de l'article présent en Annexe 2, le Matériels et Méthodes associé à ce chapitre est contenu dans l'article en question.

1. Bioinformatique.

1.1. Pymol, Chimera : visualisation, hydrophobicité, RMSD.

Les structures 3D de protéines, qu'elles soient expérimentales ou modélisées, ont été observées et analysées à l'aide des logiciels PyMOL (Schrödinger) et Chimera (Pettersen et al., 2004).

Nous avons coloré chaque a.a en fonction de son hydrophobicité à l'aide de la fonction python `color_h` (https://pymolwiki.org/index.php/Color_h), sur la base de l'échelle d'hydrophobicité définie par (Eisenberg et al., 1984). Cette échelle n'a pas d'unité, les valeurs numériques d'hydrophobicité de chaque a.a sont attribuées de telle manière que 0 corresponde à un a.a ni hydrophobe, ni hydrophile.

L'écart quadratique moyen entre atomes (RMSD pour Root Mean Square Deviation) a été calculé avec la fonction *super* de pymol (<https://pymolwiki.org/index.php/Super>), qui effectue une superposition structurale sans alignement de séquences.

1.2. Modélisation.

1.2.1 Modélisation comparative.

La modélisation comparative du domaine CoBaHMA de la calcyanine de *Synechococcus* RS9917 a été faite avec Modeller 9.23 (Webb & Sali, 2016).

1.2.2 AlphaFold2.

AlphaFold2 est un logiciel de modélisation de structure 3D *de novo* développé en 2021 par Jumper *et al.*. Il repose sur des techniques d'apprentissage profond (deep learning) pour prédire la structure 3D d'une protéine, à partir d'un MSA (pour Multiple Sequence Alignment ou Alignement Multiple de Séquences) de séquences homologues. AlphaFold2 a été entraîné sur des structures expérimentales issues de la PDB (Protein Data Bank), antérieures à avril 2018. Sans rentrer dans les détails techniques, AlphaFold2 repose sur un dialogue entre 2 blocs : un bloc interprète le MSA sous forme de co-variation et de co-évolution, et l'autre positionne les paires d'acide aminés dans l'espace 3D. Les échanges entre ces 2 blocs permettent d'obtenir une représentation précise de la structure 3D de la protéine. Il faut toutefois préciser que les auteurs eux-mêmes ne sont pas certains du fonctionnement exact

d'AlphaFold2 : par exemple la délétion complète du bloc MSA qui paraît essentiel en théorie, n'affecte que très peu les performances final du logiciel (D. T. Jones & Thornton, 2022; Jumper et al., 2021). Paradoxalement, en l'absence de MSA, les performances d'AlphaFold2 chutent tout de même drastiquement (Lin et al., 2023).

Ses extraordinaires performances durant la compétition CASP14 (Critical Assessment of Techniques for Protein Structure Prediction) lui ont valu d'être considéré comme la solution au problème, vieux de 50 ans, de la prédiction de la structure 3D des protéines à partir de leur séquence seule. Cette affirmation doit être nuancée, car de nombreux cas posent encore problème à AlphaFold2, mais il n'en reste pas moins qu'AlphaFold2 a révolutionné le champ de la prédiction des structures de protéines (Bertoline et al., 2023). Depuis sa publication, AlphaFold2 a été utilisé pour créer la database AlphaFoldDB (Varadi et al., 2022) qui regroupe les prédictions de la structure de plus de 200 millions de protéines issues de UniProt_2021_04 (The UniProt Consortium et al., 2023).

Pour modéliser le monomère de calcyanine de *S. calcipolaris*, nous avons utilisé l'outil Colabfold (Mirdita et al., 2021), qui permet d'utiliser Alphafold2 en passant par un serveur plutôt que via une installation locale. Nous sommes passés par le notebook AlphaFold2_advanced (https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/beta/AlphaFold2_advanced.ipynb) qui est devenu obsolète entre temps. A partir de la séquence de la protéine, Colabfold génère un alignement multiple avec l'outil mmseq2 (Steinegger & Söding, 2017) à partir de la base de données BFD/MGnify (Mirdita et al., 2021). Sur la base de ce MSA, Alphafold2 modélise 5 structures 3D. Le meilleur de ces 5 modèles est évalué manuellement sur la base du pLDDT (cf partie 1.3.2 du Matériels et Méthodes) et du PAE (cf partie 1.3.4 du Matériels et Méthodes). Colabfold ne permettant pas de modéliser les protéines ou assemblage protéiques de plus de ~ 1000 acides aminés, les modèles de trimère de calcyanine de *S. calcipolaris* ont été prédits par une installation locale d'Alphafold2, par le Dr. Thibault Tubiana (Institute of Integrative Biology of the Cell, Université Paris Saclay).

1.2.3. ESMFold.

ESMFold est un logiciel modélisation de structure 3D qui repose sur l'utilisation de modèles de langage. Contrairement à AlphaFold2, il ne nécessite pas de MSA incluant la séquence à modéliser. ESMFold permet de modéliser une structure 3D à partir d'une séquence seule, jusqu'à 60x plus vite qu'AlphaFold2 au prix d'une précision moindre. Sur l'ensemble de structure de la compétition CASP14, AlphaFold2 a obtenu un score de modélisation (TM-score) de 0,85/1 en moyenne, contre 0,68/1 pour ESMFold (Lin et al., 2023).

ESMFold a été utilisé pour modéliser >617 millions de structure des séquences de la base de métagénomiques MGnify90 (Lin et al., 2023; Mitchell et al., 2019). Tout comme AlphaFold2, ESMFold utilise le pLDDT comme score de confiance dans ses prédictions (Lin et al., 2023).

Pour modéliser le monomère de calcyanine *S. calcipolaris* nous avons utilisé ESMFold (Lin et al., 2022), directement sur le site de ESM Metagenomic Atlas (<https://esmatlas.com/resources?action=fold>). A partir de la séquence de la protéine, ESMFold génère un modèle avec le pLDDT associé. Le site ne génère pas le PAE associé à la structure.

Le logiciel présent sur le site de l'Atlas ne permettant pas de modéliser les protéines ou assemblage protéiques de plus de 400 acides aminés, les modèles de trimère de calcyanine de *S. calcipolaris* ont été

fait sur une installation locale d'ESMFold, par le Dr. Thibault Tubiana (Institute of Integrative Biology of the Cell, Université Paris Saclay).

1.3. Outils d'évaluation des structures 3D.

1.3.1. ProSA.

ProSA est un programme qui permet d'évaluer la qualité du repliement d'une structure 3D de protéine d'un point de vue énergétique, pour repérer les mauvais repliements ou les erreurs locales. Pour une structure 3D évaluée, ProSA calcule la somme de toutes les énergies d'interaction entre chaque paire d'a.a (notée E_x). Une même somme est calculée pour des protéines de séquences similaires mais de conformations 3D aléatoires (notées E_a). Le Z-score est calculé comme la différence entre E_x et la moyenne des E_a , divisée par l'écart type de cette moyenne. Pour une protéine correctement repliée, l'énergie d'interaction totale doit être un minimum vis à vis des conformations aléatoires, donc le Z-score doit être négatif. Il est intéressant de noter que les énergies d'interactions entre a.a sont issues d'un ensemble de protéines globulaires (Sippl, 1993). Pour des protéines membranaires, à cause de l'environnement hydrophobe, ces énergies ne sont plus les mêmes (Mbaye et al., 2019), et il est possible que ProSA ne soit plus adapté.

Nous avons utilisé ProSA via l'interface web ProSA-web, en y téléchargeant les modèles de la structure 3D de la calcyanine (<https://prosa.services.came.sbg.ac.at/prosa.php>). ProSA web indique un le Z-score global et local de la protéine ainsi qu'une comparaison de ce score vis à vis de Z-score de structures expérimentales déterminées par diffraction des rayons X et RMN (Wiederstein & Sippl, 2007).

1.3.2. pLDDT.

Pour évaluer la qualité de leurs modèles AlphaFold2 et ESMFold utilisent le pLDDT.

Le pLDDT (predicted Local Difference Test) est la prédiction du score lDDT (local superposition Difference Distance Test) sur les $C\alpha$. Le lDDT mesure la similarité entre 2 structures sous la forme d'un score allant de 0 à 100, 0 correspondant à aucune similarité, et 100 à une identité parfaite. A la différence de métriques couramment utilisées, comme le RMSD, le lDDT ne nécessite pas de superposition de structures. A l'inverse pour chaque atome d'une structure, il considère les distances de cet atome à tous les atomes inclus dans une sphère d'inclusion de rayon R_0 . Cet ensemble de distances est comparé entre les 2 structures considérées. Le pourcentage de distances conservées entre les deux structures donne le lDDT d'un atome. C'est la métrique principalement employée au CASP depuis le CASP9 (Mariani et al., 2011, 2013). Jumper *et al.* ont montré que le pLDDT était un bon prédicteur du lDDT sur les $C\alpha$, avec toutefois de grandes différences dans certains cas (Jumper et al., 2021). Le pLDDT peut être vu comme un score de confiance du logiciel prédicteur dans sa prédiction. Ce score est découpé en 4 sections :

_ ≥ 90 : très haute confiance, avec orientation des chaînes latérales correctement prédite;

_ $[70; 90[$: haute confiance;

_ $[50; 70[$: basse confiance;

_<50 : non interprétable (Tunyasuvunakool et al., 2021).

Il a été montré qu'un pLDDT<50 était cependant équivoque chez AlphaFold2. En effet cela peut correspondre à 2 situations :

_ une région repliée dont la structure 3D ne peut être prédite avec confiance, par exemple en raison d'un manque de séquence homologues ou d'un repliement inédit (Bruley et al., 2022, 2023);

_ une région désordonnée (Tunyasuvunakool et al., 2021; C. J. Wilson et al., 2022).

Pour l'heure il n'y a pas d'observation similaire démontrée pour ESMFold.

1.3.3. PROCHECK.

PROCHECK est un ensemble de programmes dédiés à l'évaluation de la qualité d'une structure 3D. Il évalue notamment la stéréochimie des a.a de la structure, et les place dans un diagramme de Ramachandran. Il les classe aussi selon les différentes régions autorisées ou interdites pour la stéréochimie des a.a, identifiées par Morris et al., 1992.

Nous avons utilisé PROCHECK via le site PDBSum (<http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/>). Nous avons directement téléchargé les modèles de la calcyanine sur PDBSum.

1.3.4. PAE.

Le PAE (Predicted Alignment Error) se présente comme une matrice qui indique l'erreur de positionnement relative de toutes les paires d'a.a de la séquence, en Ångström (Jumper et al., 2021; Mirdita et al., 2021). Elle est directement calculée par AlphaFold2 lors de la génération d'un modèle.

1.4. FoldSeek.

FoldSeek est un outil de recherche de similitudes entre les structures 3D de protéines. Il ne repose pas sur une recherche de similitudes de séquences. Par rapport aux outils préexistants, comme DALI (Holm, 2020), FoldSeek est spectaculairement plus rapide (180 000x plus rapide que DALI), mais généralement moins précis (Van Kempen et al., 2022).

Nous avons utilisé FoldSeek directement sur le site associé (<https://search.foldseek.com/search>). Le modèle de la calcyanine de *S.calcipolaris* construit par ESMFold a été téléchargé sur le site. Nous avons demandé à FoldSeek de fouiller 6 bases de données : AlphaFold/Uniprot50 v4; AlphaFold/Swiss-Prot v4; AlphaFold/Proteome v4; MGnify-ESM30 v1; PDB100 2201222; GMGCL 2204 avec le mode TM-align, sans filtre taxonomique. L'alignement affiché sur la figure **[2B changed]** est celui fourni par FoldSeek.

1.5. Téléchargement de structure 3D.

Les structures 3D expérimentales ont été téléchargées depuis le site RCSB PDB (<https://www.rcsb.org/>) (Berman, 2000).

1.6. Alignements de séquences.

Les séquences et les MSA ont été étudiés avec le logiciel JalView (Waterhouse et al., 2009).

Nous avons aligné les séquences à l'aide d'une installation locale du logiciel mafft version 7 avec les options par défaut du logiciel (<https://mafft.cbrc.jp/alignment/software/>) (Katoh & Standley, 2013).

Nous avons calculé l'identité et la similarité entre deux séquences alignées à l'aide de la version web de Sequence Manipulation Suite (https://www.bioinformatics.org/sms2/ident_sim.html) (Stothard, 2000). L'identité est définie par la proportion d'a.a strictement conservés entre les 2 séquences. La similarité est définie par la proportion de positions de la séquence qui conservent la même propriété d'une séquence à l'autre. Les groupes de similarités sont définis comme suit par le logiciel: {GAVLI}; {FYW}; {CM}; {ST}; {KRH}; {DENQ}; {P}.

1.7. Prédiction de l'agrégation.

La prédiction de la propension à l'agrégation a été calculée avec l'outil Protein-Sol sur la base des séquences seules. Cet outil calcule un score de solubilité qui est le rapport de la quantité de protéines qui resterait en solution après centrifugation et de la quantité totale de protéines; les protéines éliminées par agrégation étant agrégées. Cet outil se base sur les travaux de Niwa *et al.* (Niwa et al., 2009), qui ont mesuré cette propension à l'agrégation pour un grand nombre de protéines d'*E.coli* exprimées dans un système Cell-free, sans chaperonne (Hebditch et al., 2017).

Nous avons utilisé Protein-Sol via son le site <https://protein-sol.manchester.ac.uk/>.

2. Experimental.

2.1. Différentes constructions de calcyanine étudiées.

| Souches | Constructions (Largeur des blocs non proportionnelles à la longueur des séquences) | Poids moléculaire (kDa) | $\epsilon_{1\%}$ |
|--|---|-------------------------|------------------|
| <i>Synechococcus caldipolaris</i> (séquence entière) | CoBa HMA (GlyZip) ₃ 6His | 39,1 | 7,1 |
| <i>Synechococcus caldipolaris</i> (séquence entière) | CoBa HMA (GlyZip) ₃ L1 TE Vsp Strep II | 40,8 | 8,6 |
| <i>Synechococcus caldipolaris</i> (fragment) | CoBa HMA GlyZip 1 L1 TE Vsp Strep II | 22,8 | 8,1 |
| <i>Synechococcus caldipolaris</i> (fragment) | CoBa HMA GlyZip 1 L1 TE Vsp 10His | 22,8 | 5,7 |
| <i>Synechococcus caldipolaris</i> (fragment) | MBP L2 CoBa HMA GlyZip 1 L1 10His | 63 | 12,4 |
| <i>Synechococcus caldipolaris</i> (fragment) | MBP L3 TE Vsp CoBa HMA GlyZip 1 L1 10His | 65,1 | 12,2 |
| <i>PCC 6716</i> (séquence entière) | CoBa HMA (GlyZip) ₃ L4 6His | 35 | 7,6 |
| <i>PCC 6717</i> (séquence entière) | CoBa HMA (GlyZip) ₃ L4 6His | 35 | 7,6 |
| <i>Thermosynechococcus elongatus BP-1</i> (séquence entière) | CoBa HMA (GlyZip) ₃ L4 6His | 35 | 6 |

Table 1 : Constructions de la calcyanine étudiées pendant cette thèse.

6His (6 histidines) = HHHHHH

10His (10 histidines) = HHHHHHHHHH

Strep II = ASWSHPQFEKGA

TEV sp (Site de protéolyse de la TEV) = ENLYFQ|S

MBP = Maltose Binding Proteins

L1 (Linker 1) = GSS

L2 (Linker 2) = AARAFAAA

L3 (Linker 3) = NSSSTSGSGGGGRLVPRGSMS

L4 (Linker 4) = AAAL

Les poids moléculaires et les $\epsilon_{1\%}$ ont été calculés avec l'outil ProtParam d'Expasy (<https://web.expasy.org/protparam/>) (Gasteiger et al., 2005).

Les étiquette 6His et 10His, sont couramment rassemblées sous le nom d'étiquette poly-his (poly-histidines). Les étiquette poly-his, sont parmi les plus simples et les plus utilisées dans la purification de protéines, notamment en raison de leur faible impact sur les protéines auxquelles elles sont attachées, leurs petites tailles et le faible coût associé à leur utilisation (Wood, 2014). La chaîne latérale de l'histidine est un cycle imidazole qui a une forte affinité pour les métaux de transition comme le nickel ou le cuivre. Il est donc possible d'isoler les protéines étiquetées poly-his des protéines non étiquetées, à l'aide d'une résine sur laquelle est fixée l'un de ces métaux (Porath et al., 1975). En augmentant le nombre d'histidine de l'étiquette (en passant de 6 à 10 par exemple), il est possible d'augmenter l'affinité de l'étiquette pour la résine, ce qui permet de faire des lavages plus astringents et donc d'augmenter la pureté de la protéine (Grisshammer & Tucker, 1997), au prix d'un risque accru de perturber la protéine (Bornhorst & Falke, 2000). Cette étiquette présente tout de même quelques limitations. Tout d'abord, elle est incompatible avec certains milieux couramment utilisés en biochimie, comme les milieux contenant de l'EDTA ou du DTT (Malhotra, 2009), ou les milieux avec un pH inférieur à 6 (Bornhorst & Falke, 2000). Des cas où elle interfère avec le repliement ou la fonction de la protéine à laquelle elle était associée ont aussi été reportés (Fukushima et al., 2013; Nakatani et al., 2013). Elle ne permet pas une bonne séparation avec les protéines naturellement riches en histidines ou cystéines (Bornhorst & Falke, 2000). Enfin, c'est une étiquette qui a de moins bonnes performances en terme de purification de que d'autres étiquettes couramment utilisées telle que Strep II ou FLAG (Lichty et al., 2005).

L'étiquette Strep II (T. G. M. Schmidt et al., 1996) est aussi une étiquette couramment utilisée en biochimie. C'est une amélioration de l'étiquette Strep (T. G. M. Schmidt & Skerra, 1993). L'étiquette Strep II permet d'obtenir une grande pureté des protéines auxquelles elle est fusionnée grâce à son interaction spécifique et réversible avec la biotin. Elle n'impose quasiment aucune contrainte sur les conditions de purification (Malhotra, 2009). De plus cette étiquette est sensée être biologiquement inerte (T. G. Schmidt & Skerra, 2007), bien qu'au moins un cas où l'étiquette Strep II perturbe la cristallisation d'une protéine a été répertoriée (Bucher et al., 2002). Elle est par ailleurs, plus coûteuse d'utilisation que l'étiquette poly-his (Wood, 2014). Il est possible d'isoler les protéines étiquetées Strep II des protéines non étiquetées en les fixant sur une résine Strep-Tactin (T. G. Schmidt & Skerra, 2007).

La protéase TEV (Tobacco Etched Virus) est le nom simplifié de la protéase TEV N1a (nuclear inclusion a), une protéase de 27kDa isolée du TEV, un potyvirus de plante (Carrington & Dougherty, 1987; Dougherty & Dawn Parks, 1991). La protéase TEV reconnaît le site ENLYFQ|S, clivant la séquence au niveau du | entre le Q et le S avec une haute efficacité (Yi et al., 2013). Cette protéase est l'une des plus utilisée dans la purification des protéines pour séparer les protéines des étiquettes ou des protéines qui y auraient été fusionnées pour faciliter leur détection ou leur purification (Parks et al., 1994; Raran-Kurussi et al., 2017). Elle a une grande spécificité pour son site de clivage, ainsi qu'une capacité à fonctionner dans une large gamme de pH (4-9) et de tampon (Tris, MES, acetate...) (Raran-Kurussi et al., 2017). Elle a comme défaut de laisser après clivage des acides aminés sur la protéine purifiée. Soit la séquence ENLYFQ, soit une serine.

La MBP (Maltose Binding Protein) est une protéine monomérique de 43 kDa codée par le gène malE chez *E.coli*. Comme son nom, l'indique elle peut se fixer au maltose (Kellermann & Szmelcman, 1974).

La MBP peut être utilisée comme les étiquettes poly-his et Strep II pour purifier des protéines par fusion. En effet, les protéines fusionnées à la MBP peuvent être isolées, par accrochage sur une résine d'amylose (Malhotra, 2009). Néanmoins, le principal attrait de la MBP est d'augmenter la solubilité des protéines auxquelles elle est associée. En effet, il existe des protéines qui sont en théorie solubles mais qui ne forment que des corps d'inclusions insolubles quand elles sont exprimées seules dans *E.coli*. Pour certaines d'entre elles, une fois exprimées fusionnées à la MBP, elles deviennent parfaitement solubles et ne s'agrègent plus, même après clivage et élimination de la MBP (Kapust & Waugh, 1999). Bien que cette caractéristique de la MBP ait été bien décrite dans la littérature, les mécanismes sous-jacents qui permettent cela sont encore mal compris. Parmi les hypothèses fréquemment citées on peut trouver la création d'un environnement entropiquement favorable, la formation de micelles hydrophiles par la MBP, ou encore une action chaperonne moléculaire de la MBP (Jin et al., 2017; Raran-Kurussi & Waugh, 2012). La MBP a aussi été rapportée comme facilitant la cristallisation de certaines protéines, comme le domaine extracellulaire de la GCPR (Pioszak & Xu, 2008). Cependant, il est à noter que la large taille de la MBP peut diminuer le rendement effectif de l'expression : pour une protéine de 40kDa, il faut que les cellules produisent 2mg de protéines fusionnées à la MBP pour former 1mg de protéines non fusionnées.

Parmi les différents linkers employés, L1 et L4 ont été conçus pour avoir le moins d'impact structurel et fonctionnel possible. Ils sont utilisés uniquement pour éloigner la calcyanine des différentes étiquettes utilisées et ainsi minimiser les risques d'interaction entre elles.

L2 a été conçu comme un linker rigide qui se structure en hélice α . Cela permet de bloquer la position des deux domaines qu'il relie. Cela a été utilisé pour favoriser la cristallisation de protéines fusionnées à la MBP (Jin et al., 2017).

L3 est un long linker flexible, qui permet de maintenir éloigné le bloc MBP du bloc CoBaHMA-GlycineZipper1 afin de ne pas mettre de contrainte structurale entre eux.

2.2. Gène et expression de la calcyanine.

Au cours de la thèse, nous avons testé plusieurs souches d' *E. coli* pour l'expression : *E. coli* C41 (DE3), C43 (DE3), C44 (DE3) et C45 (DE3), des dérivées de la souche d' *E. coli* BL21 (DE3). La souche BL21 (DE3) possède le gène nécessaire à l'expression de la T7 ARN polymérase. Cette polymérase reconnaît spécifiquement le promoteur T7. Sa vitesse de transcription est 5 fois plus élevée que celle de l'ARN polymérase native d' *E. coli* (Golomb & Chamberlin, 1974). L'expression de la T7 ARN polymérase est contrôlée par le promoteur lacUV5, lui-même réprimé par le gène *LacI*. Cette répression peut être levée par l'ajout d'un inhibiteur de lacI, tel que l'IPTG. BL21 (DE3) permet donc une expression inductible du gène qui lui est fourni sur un vecteur pET (plasmid for Expression by T7 RNA polymerase) (Studier & Moffatt, 1986; William Studier et al., 1990). BL21 a aussi l'avantage de ne pas posséder l'endoprotéase OmpT ce qui diminue sa capacité à dégrader les protéines produites (Grodberg & Dunn, 1988). La production de protéines hétérologues dans *E.coli* à l'aide d'un système T7 ARN polymérase inductible est un des systèmes les plus populaires dans le champ de la purification de protéines (Angius et al., 2018). Néanmoins un certain nombre de protéines sont toxiques pour BL21 (DE3), et donc ne peuvent y être exprimées. Les souches *E.coli* C41 (DE3), C43 (DE3), C44 (DE3) et C45 (DE3) ont été sélectionnées pour leur capacité à exprimer des protéines réputées difficiles d'expression (Miroux & Walker, 1996) (Angius et al., 2018). *E.coli* C41 (DE3), C43 (DE3) sont des souches dérivées de BL21 (DE3), où le promoteur *lacUV5* est muté. Ceci limite l'expression de la T7

ARN polymérase, ce qui ralentit l'expression des protéines recombinantes par rapport à BL21 (DE3) (Wagner et al., 2008). La présence d'un trop grand nombre de protéine recombinantes inutiles à *E.coli* déclenche une destruction des ribosomes et une mort cellulaire (Dong et al., 1995). De plus, dans le cas des protéines membranaires, une sur-expression de protéines recombinantes sature le système d'insertion des protéines dans la membrane. Cela empêche la cellule de maintenir son protéome membranaire, ce qui a pour effet, entre autres, d'empêcher la génération d'ATP (Wagner et al., 2008). La souche C43 (DE3) a une mutation supplémentaire sur son gène *LacI*, ce qui ralentit encore l'expression de la T7 ARN Polymerase, qui permet une plus grande tolérance aux protéines membranaires (Kwon et al., 2015). C44 (DE3) et C45 (DE3) sont des souches spécifiquement conçues pour exprimer des protéines membranaires à très faible vitesse mais sur de très longues périodes, pour atteindre de haut taux d'expression, supérieurs à ceux obtenus avec C41 (DE3) et C43 (DE3). Ces souches expriment une majorité de T7 ARN Polymérase tronquée et inactive, et seulement une minorité de T7 ARN Polymérase active (Angius et al., 2018).

Chaque construction mentionnée dans ce manuscrit correspond à un gène synthétique dont la séquence a été optimisée pour l'usage des codons de *E.coli* puis qui a été cloné dans le plasmide pet24d ou pet24b (soustraité à la société GeneCust). Ces plasmides confèrent une résistance à la Kanamycine (Umezawa et al., 1957) pour permettre la sélection des mutants correctement transformés. La séquence codante pour la protéine d'intérêt est située en aval d'un promoteur T7. L'expression de la protéine d'intérêt est induite par ajout d'IPTG dans le milieu de culture.

Pour exprimer les constructions, les cellules d'*E.coli* sont transformées avec le plasmide choisi, soit par choc thermique, selon le protocole fournit par Lucigen (https://biosearchtech.a.bigcontent.io/v1/static/manual_COMCEL-004_OverExpress-Chemicallycompetent-Cells), soit par électroporation. Les cellules sont ensuite étalées dans une boîte de pétri contenant de l'Agar-LB-Lennox (Lysogeny Broth) (Bertani, 1951) (Lennox, 1955) et de la Kanamycine à 40µg/mL, pour sélectionner les cellules correctement transformées. La boîte est incubée une nuit à 37°C. Ensuite la boîte de pétri est conservée en chambre froide jusqu'à utilisation. Les colonies ayant poussé après la nuit à 37°C sont prélevées à l'aide d'une hanse sous hotte à flux laminaire, et mises en pré-culture dans du milieu LB-Lennox contenant 40µg/mL de Kanamycine. La pré-culture est incubée sur nuit à 37°C à 200 rpm. Pour la culture, la pré-culture est ajoutée à 1L de milieu riche 2xYT (20g/L poudre de LB-Lennox + 6g/L d'Hydrolysate de Caséine + 5g/L d'Extrait de Levure) contenant 40µg/mL de Kanamycine, de manière à atteindre une DO₆₀₀ (Densité Optique à 600nm) de 0,1. La culture est placée à 37°C, 200rpm et la DO₆₀₀ est suivie au cours du temps. Lorsqu'elle atteint [0,6-0,8], l'IPTG est ajouté de manière atteindre 0,1 à 0,5 mM final selon l'expérience, pour induire l'expression du gène. Ensuite, la culture est placée à 19°C, 30°C ou 37°C selon l'essai, sous 200 rpm d'agitation. A 30°C et 37°C la culture est laissée 4h sous agitation. A 19°C, la culture est laissée une nuit sous agitation. Les cellules sont ensuite concentrées par centrifugation à ~4000g, lavées avec du tampon (50mM Tris pH7.5, 100mM NaCl) puis les culots cellulaires sont congelés à -20°C pour être utilisés ultérieurement.

2.3. Lyse cellulaire.

Quelle que soit la construction protéique considérée, la lyse cellulaire a été effectuée avec la même méthode et le même protocole.

Pour lyser les cellules d'*E.coli* transformées, le culot cellulaire est décongelé sur glace. Il est ensuite re-suspendu dans un tampon Tris-NaCl (concentrations dépendants de la protéine purifiée, cf les paragraphes sur la purification) contenant 1 pastille d'antiprotéase et 1mM PMSF (Fluorure de phénylméthylsulfonyle) pour bloquer l'action des protéases, 50u/mL de Benzonase Nucléase pour dégrader l'ADN, et 600u/mL de rLysozyme pour dégrader la paroi d'*E.coli*. La suspension est homogénéisée mécaniquement à l'aide d'un potter. La lyse cellulaire se fait ensuite par 2 passages successifs de la suspension homogène dans un CF Broyeur (CellD), à une pression de 2kBar, à 4°C. La sortie du broyeur est centrifugée à ~9500 g 30 min pour séparer les cellules non cassées du lysat. Le surnageant, qui contient le produit de lyse, est isolé.

2.4. Purification des constructions de calcyanine.

Pour l'ensemble des constructions, la première étape de purification est une centrifugation à 100 000g à 4°C pendant 1h, qui permet de séparer les membranes contenant les protéines membranaires, qui se retrouvent dans le culot, des protéines cytosoliques qui restent dans le surnageant. Comme toutes les protéines que nous avons étudiées dans cette thèse se comportaient comme des protéines cytosoliques, nous avons travaillé uniquement sur le surnageant de cette centrifugation.

2.4.1. Calcyanine de *Synechococcus calcipolaris* étiquetée 6-Histidines.

Cette construction est purifiée en 2 étapes (Figure 37.A) :

_Chromatographie d'affinité sur résine nickel (Ni Sepharose 6 Fast Flow, GE Healthcare)

_Chromatographie d'exclusion de taille sur colonne Superdex (Superdex 200 16/600 PG ou Superdex 200 10/300 increase, Cytiva)

La purification est réalisée dans un tampon Tris 50mM, NaCl 100mM à pH 8,0.

La résine nickel est lavée à l'eau distillée puis équilibrée avec le tampon Tris-NaCl contenant 10mM d'imidazole, pour limiter les interactions non spécifiques. De la même manière, le surnageant de la centrifugation à 100 000g est ajusté à 10mM d'imidazole final. Le surnageant est mis au contact de la résine pendant 20min, sur roue, en chambre froide. Une résine contenant du nickel permet de séparer les protéines étiquetées 6His des protéines non étiquetées (Schmitt et al., 1993). La résine est ensuite déposée sur colonne, puis lavée avec de l'imidazole 20mM pour éliminer les protéines contaminantes qui se seraient fixées non spécifiquement. Les protéines sont ensuite éluées à 400mM d'imidazole. L'imidazole ayant aussi une forte affinité pour le nickel, il va entrer en compétition avec les protéines fixées et les décrocher de la résine.

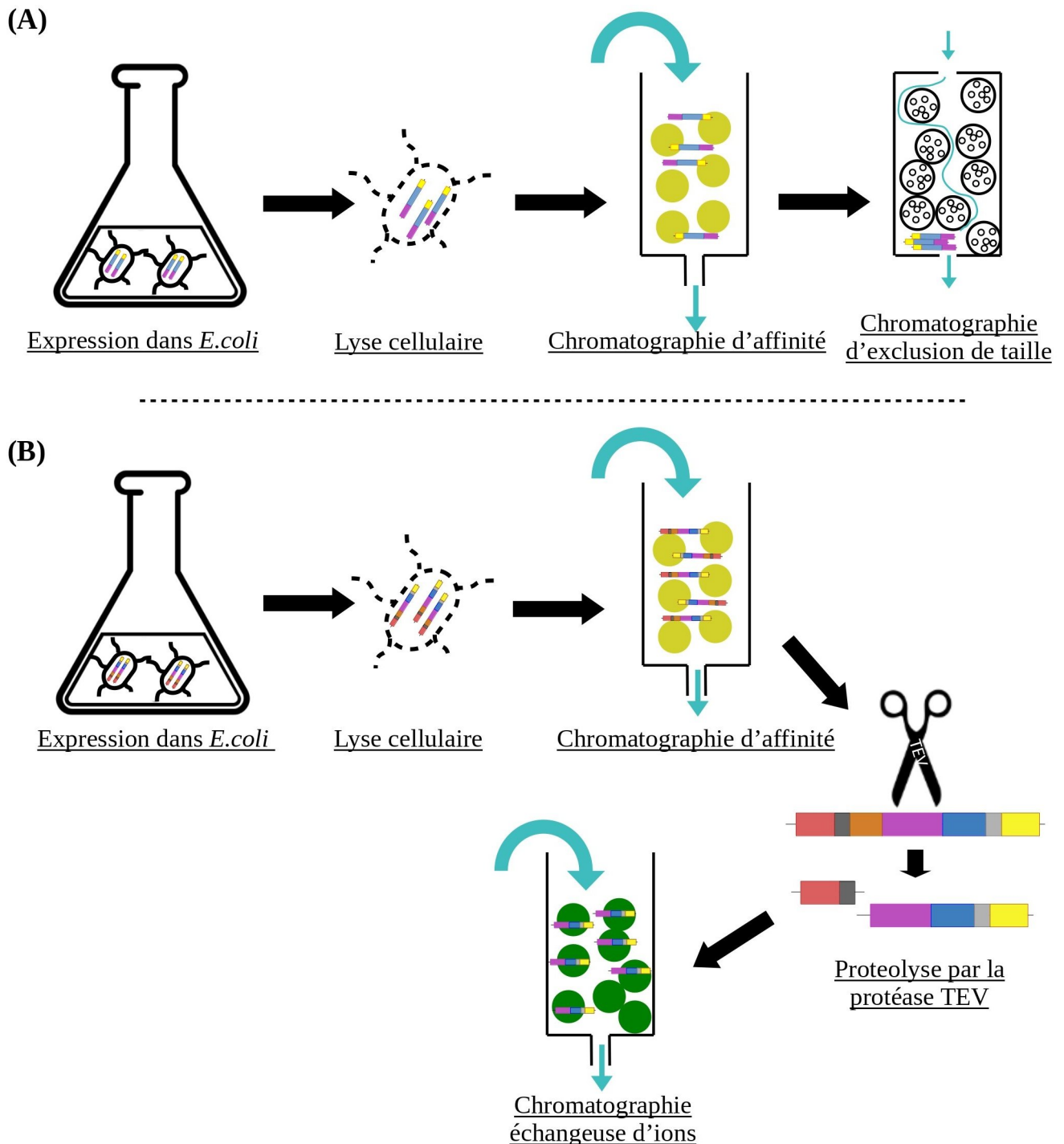


Figure 37 : Schéma des différentes étapes des 2 grands types de purification qui ont été mis en œuvre pendant la thèse.

Les fractions contenant la calcyanine sont identifiées par gel SDS-page, rassemblées puis injectées sur une colonne Superdex 200 16/600 PG ou Superdex 200 10/300 increase, pour séparer les protéines selon leur poids moléculaire apparent. La présence de protéine en sortie de colonne est suivie par absorbance à 280nm. En effet, principalement grâce à la tyrosine et au tryptophane, les protéines absorbent à

280nm (Noble & Bailey, 2009). Les fractions contenant la calcyanine sont de nouveau identifiées par gel SDS-page, avec confirmation éventuelle par Western Blot à l'aide d'un anticorps commercial polyclonal anti-étiquette 6 histidines. Ce gel et ce Western Blot permettent aussi d'évaluer la pureté de l'échantillon. L'état oligomérique de la calcyanine dans les différentes fractions est estimé par gel natif. Les fractions contenant la calcyanine pure, dans l'état oligomérique souhaité, sont rassemblées. La concentration de l'échantillon est mesurée au Nanodrop (Labtech). Selon les applications désirées, il peut être nécessaire de concentrer la calcyanine. Dans ce cas, la protéine est concentrée par centrifugation sur Protein Concentrator (cut-off ≥ 30 kDa: VIVASPIN, Sartorius). Si la calcyanine ne peut pas être utilisée immédiatement pour des analyses ou des expériences, elle est congelée en azote liquide avec 10% de glycérol final, puis stockée à -80°C .

2.4.2. *Calcyanine de Synechococcus calcipolaris étiquetée Strep II clivable à la protéase TEV (Tobacco etch virus).*

Cette construction est purifiée en 2 étapes (Figure 37.A) :

_ Chromatographie d'affinité sur résine Strep-Tactin (Strep-Tactin Superflow, Iba)

_ Chromatographie d'exclusion de taille sur colonne Superdex (Superdex 200 increase 10/300 GL, Cytiva)

La purification est réalisée dans un tampon Tris 50mM, NaCl 150mM à pH 8,0.

La résine Strep-Tactin est lavée à l'eau distillée puis équilibrée avec le tampon Tris NaCl, contenant 5mM β -mercaptoéthanol et 5% glycérol. Le surnageant de la centrifugation est mis au contact de la résine pendant 20min, sur roue, en chambre froide. La résine est déposée sur colonne puis lavée avec par le tampon qui a servi à équilibrer. Les protéines qui se sont fixées sur la résine sont éluées avec un tampon Tris NaCl, contenant 5mM β -mercaptoéthanol, 5% glycérol et 10mM de d-desthiobiotine. La d-desthiobiotine ayant aussi une forte affinité pour la Strep-Tactin, elle va entrer en compétition avec les protéines fixées et les décrocher de la résine. Les fractions contenant la protéine sont identifiées par gel SDS-page, rassemblées et concentrées à 500 μL par centrifugation sur Protein Concentrator (cut-off ≥ 30 kDa) puis injectées sur colonne Superdex 200 increase 10/300 GL, pour séparer les protéines selon leur poids moléculaire apparent.

La détermination de la pureté et de l'état oligomérique de la calcyanine, ainsi que son éventuelle concentration et congélation sont réalisées selon le protocole décrit en 3.1.

2.4.3. *Construction CoBaHMA-GlycineZipper1 étiquetée Strep clivable à la protéase TEV.*

Cette construction est purifiée comme décrit en 3.2. avec les variations suivantes :

_ L'élué de la résine Strep-Tactin est réalisée sur nuit, à 25mM d-desthiobiotine, et en présence de 250 $\mu\text{g}/\text{mL}$ de protéase TEV, pour décrocher la construction et cliver son étiquette dans le processus.

_ La chromatographie d'exclusion de taille est réalisée sur colonne Superdex 200 16/600 PG

2.4.4. *Constructions fusion MBP(Maltose Binding Protein)-CoBaHMA-GlycineZipper1-10His (linker rigide) et MBP-Site de Clivage TEV-CoBaHMA-GlycineZipper1-10His (linker souple).*

Ces 2 constructions sont purifiées en 2 étapes (Figure 37.A) :

- _Chromatographie d'affinité sur résine nickel (Ni Sepharose 6 Fast Flow, GE Healthcare)
- _Chromatographie d'exclusion de taille sur colonne Superdex (Superdex 200 16/600 PG, Cytiva)

Pour ces deux étapes, le protocole de purification est quasiment identique à celui décrit à 3.1, aux exceptions suivantes :

- _Le tampon de purification est Tris 50mM, NaCl 150mM à pH 7.6.
- _Lors de la chromatographie d'affinité, l'équilibrage se fait à 20mM imidazole, le lavage à 40mM imidazole et l'élution à 400mM imidazole.

2.4.5. Fragment CoBaHMA-GlycineZipper1-10His à partir de la protéolyse de la construction fusion MBP-Site de Clivage TEV-CoBaHMA-GlycineZipper1-10His (linker souple).

Cette construction est purifiée en 3 étapes (Figure 37.B):

- _Chromatographie d'affinité sur résine nickel (Ni Sepharose 6 Fast Flow, GE Healthcare)
 - _Clivage par la protéase TEV étiquetée 6His
 - _Chromatographie échangeuse d'ions (HiTrap ANX FF (highsub), GE Healthcare)
- La purification est réalisée dans un un tampon Tris 50mM, NaCl 150mM à pH 7,6.

Pour isoler le fragment CoBaHMA-GlycineZipper1 issu de la calcyanine de *S.calcipolaris*, la protéine fusion MBP-Site de Clivage TEV-CoBaHMA-GlycineZipper1-10His est purifiée jusqu'à la fin de l'étape de chromatographie d'affinité, comme indiqué dans la partie 3.4.

Les fractions contenant la protéine fusion issue de la chromatographie d'affinité sont rassemblées. La protéase TEV (Parks et al., 1994) étiquetée 6His produite au laboratoire est ajoutée à la protéine fusion à un rapport de 1/100e en masse. La protéolyse est réalisée sur nuit, en chambre froide. La TEV clive la construction fusion sur son site de clivage dans la séquence, ce qui sépare la MBP de la partie CoBaHMA-GlycineZipper1-10His. A l'issue de ce clivage, un précipité blanc est présent dans le tube. Le tube est alors centrifugé à ~9500g à 4°C pendant 10 min pour séparer le précipité des protéines encore en solution. Le surnageant est dilué au 2/3 par un tampon Tris 50mM pH 7.6 pour abaisser la concentration en NaCl à 100mM.

Afin de séparer le fragment CoBaHMA-GlycineZipper1-10His de la MBP et de la TEV, le surnageant est injecté sur colonne échangeuse d'ions HiTrap ANX FF (highsub). Une fois les protéines accrochées, l'élution est établie par un gradient de sel, de 50mM à 500mM NaCl pour séparer les différentes protéines selon leur charge apparente. Les fractions contenant le fragment sont identifiées par gel SDS-page, avec confirmation éventuelle par Western Blot à l'aide d'un anticorps anti-6-His. Ce gel et ce Western Blot permettent aussi d'évaluer l'état de pureté de l'échantillon. L'état oligomérique du fragment dans les différentes fractions est estimé par gel natif. Les fractions contenant le fragment pur, dans l'état oligomérique souhaité sont rassemblées. La concentration du fragment est mesurée au Nanodrop. Selon les applications désirées, il peut être nécessaire de concentrer le fragment. Dans ce cas le fragment est concentré sur par centrifugation sur Protein Concentrator (cut-off \geq 5kDa, VIVASPIN, Sartorius).

2.5. Expression et purification de la protéase TEV étiquetée 6His.

La protéase TEV étiquetée 6His est exprimée et purifiée selon un protocole similaire à ceux utilisés pour les constructions de calcyanine, avec les changements suivants :

_ La protéase TEV est exprimée comme une protéine recombinante dans un plasmide PRK793 (résistance à l'ampicilline (Acred et al., 1962)), dans la souche d'*E.coli* BL21(DE3)-RIL (résistance au chloramphénicol (Ehrlich et al., 1947)). La sélection des cellules correctement transformées est réalisée à 34µg/mL chloramphénicol et 50µg/mL d'ampicilline.

_ L'expression dure 24h à 4°C après l'induction à l'IPTG.

_ La purification est réalisée en une étape d'affinité sur résine nickel, dans un tampon de purification Tris 20mM, NaCl 500mM pH = 8.

_ La protéine purifiée est dialysée en 2 bains successifs, avec une membrane ≥ 8000 kDa, dans un tampon Tris 20mM, NaCl 200mM, Glycerol 10%, EDTA (Ethylenediaminetetraacetic acid ou acide éthylène diamine tétra-acétique) 2mM, pH = 8.

_ Les éventuels précipités formés avant ou pendant la dialyse sont éliminés par centrifugation à $\sim 4000g$ pendant 10min.

2.6. Protéolyse limitée.

La calcyanine de *S.calcipolaris* étiquetée 6-Histidines purifiée est exposée à 2 protéases :

_ La trypsine (de pancréas de bovin, Sigma-Aldrich) (Kühne, 1877) (Gutfreund et al., 1976).

_ L' α -chymotrypsine (de pancréas de bovin, Sigma-Aldrich) (Bender et al., 1967).

La trypsine clive principalement en C_{ter} des lysines et des arginines. L' α -chymotrypsine clive principalement en C_{ter} des acides aminés aromatiques (tryptophane, phenylalanine et tyrosine) et de certains a.a hydrophobes comme la leucine et l'isoleucine (Fontana et al., 2012).

Avant utilisation, les protéases sont conservées en alicquot à -80°C. La trypsine est congelée dans un tampon Tris 50mM pH=7,5, NaCl 150mM, CaCl₂ 5mM. L' α -chymotrypsine est congelée en HCl 1mM, CaCl₂ 2mM.

Pour la protéolyse limitée, les protéases sont décongelées sur glace. La calcyanine est mise en contact avec chaque protéase à un ratio masse/masse connue, dans un tampon Tris 50mM pH=7,5, NaCl 150mM, CaCl₂ 5mM. Typiquement ce ratio va de 10:1 à 100 000:1 masse de calcyanine par masse de protéase. Pour la trypsine et l' α -chymotrypsine, la digestion dure 1h sur glace.

La réaction de protéolyse est arrêtée par l'ajout de 5mM final de PMSF, qui inhibe les protéases à sérine dont font parties la trypsine et l' α -chymotrypsine (Garcia-Carreno, 1992). Le résultat est analysé par gel SDS-page et les bandes les plus intéressantes sont découpées puis analysées en spectroscopie de masse. La séquence exacte des fragments de protéolyse est déterminée par spectroscopie de masse en phase liquide.

Dans le but d'isoler le fragment et en analyser la séquence, la digestion est suivie par une chromatographie d'exclusion de taille (Superdex 200 5/150Gl ou 16/600PG selon la quantité de protéine disponible).

La prédiction des potentiels sites de clivage de la trypsine et de l' α -chymotrypsine sur la calcyanine est réalisée grâce au site internet PeptideCutter d'Expasy (https://web.expasy.org/peptide_cutter/) (Gasteiger et al., 2005).

2.7. Spectroscopie de masse.

Les expériences de spectroscopie de masse (MS) ont été réalisées par le Dr. Christophe Marchand (Institut de Biologie Paris-Seine, UMR 7238 CNRS, Biologie Computationnelle et Quantitative (LCQB), Sorbonne Université).

2.7.1. Préparation des bandes des gels SDS-PAGE.

Les bandes d'intérêt ont été excisées du gel de manière précise à l'aide d'un scalpel sur une surface en verre préalablement nettoyée à l'éthanol. Les bandes sont alors redécoupées en petits cubes d'environ 1-2 mm³ afin d'optimiser la pénétration des différents réactifs dans les mailles du gel de polyacrylamide. Afin d'éliminer le bleu de Coomassie fixé aux protéines ainsi que les différents contaminants possibles (SDS, sels, ...), les cubes sont ensuite placés à 37°C en présence de 150 μ L d'une solution de bicarbonate d'ammonium (AMBIC) 50 mM. Après 15 min d'incubation au thermomixer (Eppendorf) à 37°C sous agitation (1000 rpm), le surnageant est éliminé et 150 μ L d'acétonitrile (CH₃CN ; ACN) sont ajoutés. Après 5 min d'incubation à 37°C au thermomixer, le surnageant est prélevé et éliminé. Ces lavages successifs sont répétés autant de fois que nécessaire pour obtenir la décoloration complète des morceaux de gels et ces derniers sont alors déshydratés sous vide à l'aide d'un speedvac (Concentrator plus, Eppendorf) pendant 20 min.

En l'absence de cystéines dans la calcyanine, les étapes de réduction des ponts disulfure, et d'alkylation des cystéines n'ont pas été mise en œuvre. La faible acidité (pH~5) constatée est suffisante pour inhiber l'activité de la trypsine sans l'inactiver et ainsi éviter son autolyse lors de la réhydratation.

Les morceaux de gels sont réhydratés sur glace en présence de 10 à 15 μ L d'une solution de trypsine (Trypsine Gold, Proméga) à 12,5 ng/ μ L en HCl 1 mM.

Après une incubation d'au moins 30 min, 60 μ L d'une solution d'AMBIC 50 mM sont alors ajoutés sur les morceaux de gels réhydratés pour corriger le pH (pH~8,5) et initier la digestion par la trypsine. Les tubes sont alors placés à 37°C sous agitation (800 rpm) sur nuit.

A l'issue de la digestion, les surnageants sont prélevés et conservés et les morceaux de gels sont lavés pendant 15 min à 30°C par 100 μ L d'une solution d'acide trifluoroacétique 1% (TFA ; volume/volume). Les surnageants sont prélevés et ajoutés aux précédents. Les morceaux de gels sont lavés une seconde fois pendant 15 min à 30°C par 100 μ L d'une solution aqueuse d'ACN/TFA 60/1 (volume/volume). Les surnageants sont prélevés et ajoutés aux précédents.

Les surnageants sont alors concentrés sous vide à l'aide d'un speedvac (Concentrator plus, Eppendorf) jusqu'à évaporation presque complète (5-10 μ L) et conservés à -20°C jusqu'à l'analyse par spectrométrie de masse.

Pour l'analyse MS proprement dite, 1 μ L de surnageant concentré par évaporation est mélangé rapidement avec 1,5 μ L d'une solution aqueuse d'acide \square -cyano-4-hydroxycinnamique (matrice)

préparée à demi-saturation (poids/volume) dans de l'ACN/TFA 50/0,3 (volume/volume). 2 μ L de ce mélange sont alors déposés sur la plaque porte échantillon et laissés sous un léger courant d'air pour faciliter l'évaporation des solvants et la co-cristallisation des peptides avec la matrice.

2.7.2. Préparation du fragment de la calcyanine pour une masse entière.

Le fragment est purifié après protéolyse limitée, comme décrit dans la partie 2.6. Le pic contenant la protéine est concentré puis diluée avec un tampon Tris 50mM pH 7,5, puis concentrée à nouveau pour se placer dans des conditions salines compatibles avec la spectroscopie de masse, à la concentration la plus élevée possible. Une fois concentré jusqu'à 4 μ L, 2 μ L d'une solution aqueuse d'ACN/TFA 30/0,3 (volume/volume) sont alors ajoutés pour acidifier l'échantillon et 1,1 μ L de cet échantillon final est mélangé à 1,8 μ L d'une solution aqueuse d'acide sinapinique (Sigma) préparée à saturation (poids/volume) dans de l'ACN/TFA 30/0,3 (volume/volume). 2 μ L de ce mélange sont alors déposés sur la plaque porte échantillon et laissés sous un léger courant d'air pour faciliter l'évaporation des solvants et la co-cristallisation des peptides avec la matrice.

2.7.3. MALDI-TOF (pour Spectrométrie de masse Désorption-Ionisation Laser Assistée par Matrice-Temps de Vol).

Les analyses ont été réalisées sur un spectromètre de masse MALDI-TOF (Axima Performance, Shimadzu, Manchester, Angleterre) en mode positif linéaire (masse intacte) ou bien en mode positif réflectron (empreintes peptidiques massiques) après calibration du spectromètre de masse sur les pics de masse de protéines et de peptides références respectivement. Pour l'analyse de masse intacte du fragment, le pulse extraction a été fixé à 25000 tandis que pour l'analyse des mélanges peptidiques il a été fixé à 2800.

2.8. Small Angle X-rays Scattering : expériences et analyses.

Le SAXS (pour Small Angle X-Ray Scattering) consiste à exposer des particules en solution, telles que des protéines, à un faisceau de rayons X monochromatique et collimaté. Au contact de ces particules, les rayons X vont être diffusés, ce qui va aboutir à une figure de diffusion qui peut être enregistrée à l'aide d'un détecteur. Comme les particules en solution sont orientées aléatoirement, la figure de diffusion obtenue après le passage d'un grand nombre de particules est isotrope. La figure de diffusion peut alors être considérée comme la moyenne sphérique de la transformée de Fourier de la différence de densité électronique entre celles des particules et celle du solvant. L'intensité de diffusion 2D, $I(q)$ avec $q = 4\pi \sin(\theta)/\lambda$ où 2θ est l'angle de diffusion et λ la longueur d'onde des rayons X, est isotrope, elle peut donc être ramenée à une courbe 1D par intégration azimutale.

De cette courbe, il est possible d'extrapoler le rayon de giration, R_g , et l'intensité à l'origine à partir de l'approximation de Guinier, qui stipule que pour $q < 1/R_g$, $I(q) \propto I(0) \cdot \exp(-(q \cdot R_g)^2/3)$. R_g peut être vu comme une détermination de la taille globale de la particule (Guinier, 1939).

En condition diluée, tel qu'en sortie de SEC, les différentes populations de la particule sont séparées, et l'on se rapproche des conditions mono-disperse. Dans les conditions de mono-dispersité, le poids moléculaire peut être extrapolé à partir de $I(0)$. Le tracé de Kratky, $q^2I(q) = f(q)$ permet de déterminer la globularité de la protéine. Enfin, l'enveloppe, $\log(I(q)) = f(q)$, peut servir à construire une structure basse résolution (de l'ordre du nm) pour la particule considérée (Kikhney & Svergun, 2015) (Mertens & Svergun, 2010).

Toutes les expériences et analyses de SAXS ont été réalisées avec la Dr. Stéphanie Finet (Institut de minéralogie, de physique des matériaux et de cosmochimie, UMR 7590, Sorbonne Université, Muséum National d'Histoire Naturelle). Les expériences de SEC-SAXS ont été réalisées sur la ligne SWING du Synchrotron SOLEIL (St-Aubin, France) (Thureau et al., 2021). Sur cette ligne, le faisceau de rayon-X a une longueur d'onde de $\lambda = 1.03\text{\AA}$, la distance échantillon-détecteur est de 2m et la gamme de fréquences spatiales accessibles, q , est de $0.045\text{-}0.546\text{ \AA}^{-1}$. La protéine analysée est injectée sur colonne HPLC (High Performance Liquid Chromatography) Bio Sec-3 300 \AA (Agilent), ou sur colonne Superdex 5/150 GL (Cytiva) équilibrée avec le tampon de purification de la protéine considérée. La colonne est branchée sur le système Agilent HPLC de la ligne, lui-même relié à la cellule SAXS. La quantité de protéine en sortie de colonne est évaluée par absorbance à 280nm. L'absorbance à 260nm est utilisée pour évaluer la contamination par des acides nucléiques. La durée des frames lors de l'acquisition SAXS est de 1s.

Les données issues du SAXS sont analysées par le logiciel Foxtrot (<https://www.synchrotron-soleil.fr/fr/lignes-de-lumiere/swing>) proposé par la ligne SWING. Les spectres de diffusions 2D sont ramenés à une courbe de diffusion 1D par intégration, à q constant, des intensités transmises. L'intensité diffusée à l'angle nul $I(0)$ et le rayon de giration (R_g) de la molécule sont déterminés en se basant sur l'approximation de Guinier (Guinier, 1939).

2.9. Essais de cristallogénèse.

Les essais de cristallisation ont été réalisés avec le Dr. Julien Henri (Institut de Biologie Paris-Seine, UMR 7238 CNRS, Biologie Computationnelle et Quantitative (LCQB), Sorbonne Université) à l'aide d'un robot Mosquito TTP Labtech, qui permet de pipeter des nanovolumes de solutions liquides. La technique utilisée est la diffusion de vapeur en goutte assise.

Pour identifier des conditions de cristallogénèse, la protéine concentrée à environ 10 mg/mL est mélangée en volume équivalent dans une goutte de ~ 100 nL avec 384 conditions standardisées (crible JCSG (Lesley & Wilson, 2005) ou complémentaires (NeXtal Biotechnologies 6201 Trust Drive Holland, OH, Etats-Unis)). Ces kits permettent de couvrir une large gamme de pH, de force ionique et d'agents précipitants. Une fois la protéine mélangée aux conditions, la plaque d'essais de cristallisation est scellée et maintenue à 20°C. Les plaques sont régulièrement observées jusqu'à 12 mois après le début de l'essai. En cas d'apparition de cristaux, les conditions de précipitation sont répliquées et optimisées dans une goutte de 2 μL . Pour cela, la condition initiale est modifiée en faisant varier la concentration des différents composants du milieu de précipitation. Les cristaux sont cryo-protégés en éthylène glycol ou glycérol, congelés en azote liquide puis exposés au rayon X sur les lignes PROXIMA-1 ou PROXIMA-2 du synchrotron SOLEIL (Saint-Aubin, France) afin de collecter les informations de diffraction.

Références :

- Acred, P., Brown, D. M., Turner, D. H., & Wilson, M. J. (1962). PHARMACOLOGY AND CHEMOTHERAPY OF AMPICILLIN-A NEW BROAD-SPECTRUM PENICILLIN. *British Journal of Pharmacology and Chemotherapy*, 18(2), 356–369. <https://doi.org/10.1111/j.1476-5381.1962.tb01416.x>
- Aizenberg, J., Addadi, L., Weiner, S., & Lambert, G. (1996). Stabilization of amorphous calcium carbonate by specialized macromolecules in biological and synthetic precipitates. *Advanced Materials*, 8(3), 222–226. <https://doi.org/10.1002/adma.19960080307>
- Ali, J., Yu, M., Sung, L.-K., Cheung, Y.-W., & Lai, E.-M. (2022). A glycine zipper motif governs translocation of type VI secretion toxic effectors across the cytoplasmic membrane of target cells [Preprint]. *Microbiology*. <https://doi.org/10.1101/2022.07.12.499750>
- Angius, F., Ilioiaia, O., Amrani, A., Suisse, A., Rosset, L., Legrand, A., Abou-Hamdan, A., Uzan, M., Zito, F., & Miroux, B. (2018). A novel regulation mechanism of the T7 RNA polymerase based expression system improves overproduction and folding of membrane proteins. *Scientific Reports*, 8(1), 8572. <https://doi.org/10.1038/s41598-018-26668-y>
- Arp, G., Reimer, A., & Reitner, J. (1999). Calcification in cyanobacterial biofilms of alkaline salt lakes. *European Journal of Phycology*, 34(4), 393–403. <https://doi.org/10.1080/09670269910001736452>
- Aubel, M., Eicholt, L., & Bornberg-Bauer, E. (2023). Assessing structure and disorder prediction tools for de novo emerged proteins in the age of machine learning. *F1000Research*, 12, 347. <https://doi.org/10.12688/f1000research.130443.1>
- Azzaz, F., Yahi, N., Chahinian, H., & Fantini, J. (2022). The Epigenetic Dimension of Protein Structure Is an Intrinsic Weakness of the AlphaFold Program. *Biomolecules*, 12(10), 1527. <https://doi.org/10.3390/biom12101527>
- Bacchetta, T., López-García, P., Gutiérrez-Preciado, A., Mehta, N., Skouri-Panet, F., Benzerara, K., Ciobanu, M., Yubuki, N., Tavera, R., & Moreira, D. (2022). *Description of Gloeomargarita*

- ahousahtiae sp. Nov., a thermophilic member of the order Gloeomargaritales with intracellular carbonate inclusions [Preprint]. *Microbiology*. <https://doi.org/10.1101/2022.11.03.515036>
- Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., ... Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557), Article 6557. <https://doi.org/10.1126/science.abj8754>
- Bakshi, M. S., Kaur, H., Banipal, T. S., Singh, N., & Kaur, G. (2010). Biomineralization of Gold Nanoparticles by Lysozyme and Cytochrome c and Their Applications in Protein Film Formation. *Langmuir*, 26(16), 13535–13544. <https://doi.org/10.1021/la101701f>
- Banci, L., Bertini, I., Ciofi-Baffoni, S., Gonnelli, L., & Su, X.-C. (2003). Structural Basis for the Function of the N-terminal Domain of the ATPase CopA from *Bacillus subtilis*. *Journal of Biological Chemistry*, 278(50), Article 50. <https://doi.org/10.1074/jbc.M307389200>
- Barrán-Berdón, A. L., Rodea-Palomares, I., Leganés, F., & Fernández-Piñas, F. (2011). Free Ca²⁺ as an early intracellular biomarker of exposure of cyanobacteria to environmental pollution. *Analytical and Bioanalytical Chemistry*, 400(4), 1015–1029. <https://doi.org/10.1007/s00216-010-4209-3>
- Barton, L. L., & Tomei, F. A. (1995). Characteristics and Activities of Sulfate-Reducing Bacteria. In L. Barton (Ed.), *Sulfate-Reducing Bacteria* (pp. 1–32). Springer US. https://doi.org/10.1007/978-1-4899-1582-5_1
- Bauer, P., Elbaum, R., & Weiss, I. M. (2011). Calcium and silicon mineralization in land plants: Transport, structure and function. *Plant Science*, 180(6), 746–756. <https://doi.org/10.1016/j.plantsci.2011.01.019>
- Beccia, M. R., Sauge-Merle, S., Lemaire, D., Brémond, N., Pardoux, R., Blangy, S., Guilbaud, P., & Berthomieu, C. (2015). Thermodynamics of Calcium binding to the Calmodulin N-terminal domain to evaluate site-specific affinity constants and cooperativity. *JBIC Journal of Biological Inorganic Chemistry*, 20(5), Article 5. <https://doi.org/10.1007/s00775-015-1275-1>

- Bender, M. L., Kezdy, F. J., & Wedler, F. C. (1967). alpha-Chymotrypsin: Enzyme concentration and kinetics. *Journal of Chemical Education*, 44(2), 84. <https://doi.org/10.1021/ed044p84>
- Benjin, X., & Ling, L. (2020). Developments, applications, and prospects of cryo-electron microscopy. *Protein Science*, 29(4), 872–882. <https://doi.org/10.1002/pro.3805>
- Benzerara, K., Bolzoni, R., Monteil, C., Beyssac, O., Forni, O., Alonso, B., Asta, M. P., & Lefevre, C. (2021). The gammaproteobacterium *Achromatium* forms intracellular amorphous calcium carbonate and not (crystalline) calcite. *Geobiology*, 19(2), 199–213. <https://doi.org/10.1111/gbi.12424>
- Benzerara, K., Duprat, E., Bitard-Feildel, T., Caumes, G., Cassier-Chauvat, C., Chauvat, F., Dezi, M., Diop, S. I., Gaschignard, G., Görden, S., Gugger, M., López-García, P., Millet, M., Skouri-Panet, F., Moreira, D., & Callebaut, I. (2022). A New Gene Family Diagnostic for Intracellular Biomineralization of Amorphous Ca Carbonates by Cyanobacteria. *Genome Biology and Evolution*, 14(3), evac026. <https://doi.org/10.1093/gbe/evac026>
- Benzerara, K., Skouri-Panet, F., Li, J., Ferard, C., Gugger, M., Laurent, T., Couradeau, E., Ragon, M., Cosmidis, J., Menguy, N., Margaret-Oliver, I., Tavera, R., Lopez-Garcia, P., & Moreira, D. (2014). Intracellular Ca-carbonate biomineralization is widespread in cyanobacteria. *Proceedings of the National Academy of Sciences*, 111(30), Article 30. <https://doi.org/10.1073/pnas.1403510111>
- Berman, H. M. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1), 235–242. <https://doi.org/10.1093/nar/28.1.235>
- Bertani, G. (1951). STUDIES ON LYSOGENESIS I: The Mode of Phage Liberation by Lysogenic *Escherichia coli*. *Journal of Bacteriology*, 62(3), 293–300. <https://doi.org/10.1128/jb.62.3.293-300.1951>
- Bertoline, L. M. F., Lima, A. N., Krieger, J. E., & Teixeira, S. K. (2023). Before and after AlphaFold2: An overview of protein structure prediction. *Frontiers in Bioinformatics*, 3, 1120370. <https://doi.org/10.3389/fbinf.2023.1120370>

- Bitard-Feildel, T., Lamiable, A., Mornon, J.-P., & Callebaut, I. (2018). Order in Disorder as Observed by the “Hydrophobic Cluster Analysis” of Protein Sequences. *PROTEOMICS*, *18*(21–22), Article 21–22. <https://doi.org/10.1002/pmic.201800054>
- Bitter, R. M., Oh, S., Deng, Z., Rahman, S., Hite, R. K., & Yuan, P. (2022). Structure of the Wilson disease copper transporter ATP7B. *Science Advances*, *8*(9), eabl5508. <https://doi.org/10.1126/sciadv.abl5508>
- Blakemore, R. (1975). Magnetotactic Bacteria. *Science*, *190*(4212), 377–379. <https://doi.org/10.1126/science.170679>
- Blondeau, M., Sachse, M., Boulogne, C., Gillet, C., Guigner, J.-M., Skouri-Panet, F., Poinot, M., Ferard, C., Miot, J., & Benzerara, K. (2018). Amorphous Calcium Carbonate Granules Form Within an Intracellular Compartment in Calcifying Cyanobacteria. *Frontiers in Microbiology*, *9*, 1768. <https://doi.org/10.3389/fmicb.2018.01768>
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. <https://doi.org/10.1088/1742-5468/2008/10/P10008>
- Bornhorst, J. A., & Falke, J. J. (2000). [16] Purification of proteins using polyhistidine affinity tags. In *Methods in Enzymology* (Vol. 326, pp. 245–254). Elsevier. [https://doi.org/10.1016/S0076-6879\(00\)26058-8](https://doi.org/10.1016/S0076-6879(00)26058-8)
- Boudière, L., Michaud, M., Petroustos, D., Rébeillé, F., Falconet, D., Bastien, O., Roy, S., Finazzi, G., Rolland, N., Jouhet, J., Block, M. A., & Maréchal, E. (2014). Glycerolipids in photosynthesis: Composition, synthesis and trafficking. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, *1837*(4), Article 4. <https://doi.org/10.1016/j.bbabbio.2013.09.007>
- Bruley, A., Bitard-Feildel, T., Callebaut, I., & Duprat, E. (2023). A sequence-based foldability score combined with ALPHAFOLD2 predictions to disentangle the protein order/disorder continuum. *Proteins: Structure, Function, and Bioinformatics*, *91*(4), 466–484. <https://doi.org/10.1002/prot.26441>

- Bruley, A., Mornon, J.-P., Duprat, E., & Callebaut, I. (2022). Digging into the 3D Structure Predictions of AlphaFold2 with Low Confidence: Disorder and Beyond. *Biomolecules*, *12*(10), 1467. <https://doi.org/10.3390/biom12101467>
- Bucher, M. H., Evdokimov, A. G., & Waugh, D. S. (2002). Differential effects of short affinity tags on the crystallization of *Pyrococcus furiosus* maltodextrin-binding protein. *Acta Crystallographica Section D Biological Crystallography*, *58*(3), 392–397. <https://doi.org/10.1107/S0907444901021187>
- Bull, P. C., & Cox, D. W. (1994). Wilson disease and Menkes disease: New handles on heavy-metal transport. *Trends in Genetics*, *10*(7), 246–252. [https://doi.org/10.1016/0168-9525\(94\)90172-4](https://doi.org/10.1016/0168-9525(94)90172-4)
- Callebaut, I., Labesse, G., Durand, P., Poupon, A., Canard, L., Chomilier, J., Henrissat, B., & Mornon, J. P. (1997). Deciphering protein sequence information through hydrophobic cluster analysis (HCA): Current status and perspectives. *Cellular and Molecular Life Sciences (CMLS)*, *53*(8), Article 8. <https://doi.org/10.1007/s000180050082>
- Callebaut, I., & Mornon, J.-P. (2010). LOTUS, a new domain associated with small RNA pathways in the germline. *Bioinformatics*, *26*(9), 1140–1144. <https://doi.org/10.1093/bioinformatics/btq122>
- Cam, N., Benzerara, K., Georgelin, T., Jaber, M., Lambert, J.-F., Poinot, M., Skouri-Panet, F., & Cordier, L. (2016). Selective Uptake of Alkaline Earth Metals by Cyanobacteria Forming Intracellular Carbonates. *Environmental Science & Technology*, *50*(21), Article 21. <https://doi.org/10.1021/acs.est.6b02872>
- Cam, N., Benzerara, K., Georgelin, T., Jaber, M., Lambert, J.-F., Poinot, M., Skouri-Panet, F., Moreira, D., López-García, P., Raimbault, E., Cordier, L., & Jézéquel, D. (2018). Cyanobacterial formation of intracellular Ca-carbonates in undersaturated solutions. *Geobiology*, *16*(1), Article 1. <https://doi.org/10.1111/gbi.12261>
- Cam, N., Georgelin, T., Jaber, M., Lambert, J.-F., & Benzerara, K. (2015). In vitro synthesis of amorphous Mg-, Ca-, Sr- and Ba-carbonates: What do we learn about intracellular calcification by cyanobacteria? *Geochimica et Cosmochimica Acta*, *161*, 36–49. <https://doi.org/10.1016/j.gca.2015.04.003>

- Carino, A., Testino, A., Andalibi, M. R., Pilger, F., Bowen, P., & Ludwig, C. (2017). Thermodynamic-Kinetic Precipitation Modeling. A Case Study: The Amorphous Calcium Carbonate (ACC) Precipitation Pathway Unravelling. *Crystal Growth & Design*, 17(4), 2006–2015. <https://doi.org/10.1021/acs.cgd.7b00006>
- Carrington, J. C., & Dougherty, W. G. (1987). Small Nuclear Inclusion Protein Encoded by a Plant Potyvirus Genome Is a Protease. *Journal of Virology*, 61(8), 2540–2548. <https://doi.org/10.1128/jvi.61.8.2540-2548.1987>
- Cesari, S., Thilliez, G., Ribot, C., Chalvon, V., Michel, C., Jauneau, A., Rivas, S., Alaux, L., Kanzaki, H., Okuyama, Y., Morel, J.-B., Fournier, E., Tharreau, D., Terauchi, R., & Kroj, T. (2013). The Rice Resistance Protein Pair RGA4/RGA5 Recognizes the *Magnaporthe oryzae* Effectors AVR-Pia and AVR1-CO39 by Direct Binding. *The Plant Cell*, 25(4), Article 4. <https://doi.org/10.1105/tpc.112.107201>
- Chowdhury, R., Bouatta, N., Biswas, S., Floristean, C., Kharkar, A., Roy, K., Rochereau, C., Ahdriz, G., Zhang, J., Church, G. M., Sorger, P. K., & AlQuraishi, M. (2022). Single-sequence protein structure prediction using a language model and deep learning. *Nature Biotechnology*, 40(11), 1617–1623. <https://doi.org/10.1038/s41587-022-01432-w>
- Churchill-Angus, A. M., Schofield, T. H. B., Marlow, T. R., Sedelnikova, S. E., Wilson, J. S., Rafferty, J. B., & Baker, P. J. (2021). Characterisation of a tripartite α -pore forming toxin from *Serratia marcescens*. *Scientific Reports*, 11(1), 6447. <https://doi.org/10.1038/s41598-021-85726-0>
- Cosmidis, J., & Benzerara, K. (2022). Why do microbes make minerals? *Comptes Rendus. Géoscience*, 354(1), 1–39. <https://doi.org/10.5802/crgeos.107>
- Couradeau, E., Benzerara, K., Gerard, E., Moreira, D., Bernard, S., Brown, G. E., & Lopez-Garcia, P. (2012). An Early-Branching Microbialite Cyanobacterium Forms Intracellular Carbonates. *Science*, 336(6080), Article 6080. <https://doi.org/10.1126/science.1216171>
- Dana, J. D., Dana, E. S., & Gaines, R. V. (1997). *Dana's new mineralogy: The system of mineralogy of James Dwight Dana and Edward Salisbury Dana*. (8th edition). Wiley.

- De la Concepcion, J. C., Franceschetti, M., Maqbool, A., Saitoh, H., Terauchi, R., Kamoun, S., & Banfield, M. J. (2018). Polymorphic residues in rice NLRs expand binding and response to effectors of the blast pathogen. *Nature Plants*, 4(8), Article 8. <https://doi.org/10.1038/s41477-018-0194-x>
- De Wever, A., Benzerara, K., Coutaud, M., Caumes, G., Poinot, M., Skouri-Panet, F., Laurent, T., Duprat, E., & Gugger, M. (2019). Evidence of high Ca uptake by cyanobacteria forming intracellular Ca CO₃ and impact on their growth. *Geobiology*, 17(6), Article 6. <https://doi.org/10.1111/gbi.12358>
- Derewenda, Z. (2004). The use of recombinant methods and molecular engineering in protein crystallization. *Methods*, 34(3), 354–363. <https://doi.org/10.1016/j.ymeth.2004.03.024>
- Diop, S. I. (2016). *Caractérisation des Domaines Orphelins au Sein des Protéomes. Application à l'étude de mécanismes de biominéralisation intracellulaire chez les cyanobactéries.* (p. 23).
- Dong, H., Nilsson, L., & Kurland, C. G. (1995). Gratuitous overexpression of genes in Escherichia coli leads to growth inhibition and ribosome destruction. *Journal of Bacteriology*, 177(6), 1497–1504. <https://doi.org/10.1128/jb.177.6.1497-1504.1995>
- Dougherty, W. G., & Dawn Parks, T. (1991). Post-translational processing of the tobacco etch virus 49-kDa small nuclear inclusion polyprotein: Identification of an internal cleavage site and delimitation of VPg and proteinase domains. *Virology*, 183(2), 449–456. [https://doi.org/10.1016/0042-6822\(91\)90974-G](https://doi.org/10.1016/0042-6822(91)90974-G)
- Du, H., & Amstad, E. (2020). Water: How Does It Influence the CaCO₃ Formation? *Angewandte Chemie International Edition*, 59(5), 1798–1816. <https://doi.org/10.1002/anie.201903662>
- Dupraz, C., Reid, R. P., Braissant, O., Decho, A. W., Norman, R. S., & Visscher, P. T. (2009). Processes of carbonate precipitation in modern microbial mats. *Earth-Science Reviews*, 96(3), 141–162. <https://doi.org/10.1016/j.earscirev.2008.10.005>
- Durak, G. M., Brownlee, C., & Wheeler, G. L. (2017). The role of the cytoskeleton in biomineralisation in haptophyte algae. *Scientific Reports*, 7(1), 15409. <https://doi.org/10.1038/s41598-017-15562-8>

- Ehrlich, J., Bartz, Q. R., Smith, R. M., Joslyn, D. A., & Burkholder, P. R. (1947). Chloromycetin, a New Antibiotic From a Soil Actinomycete. *Science*, *106*(2757), 417–417.
<https://doi.org/10.1126/science.106.2757.417>
- Eisenberg, D., Schwarz, E., Komaromy, M., & Wall, R. (1984). Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *Journal of Molecular Biology*, *179*(1), 125–142. [https://doi.org/10.1016/0022-2836\(84\)90309-7](https://doi.org/10.1016/0022-2836(84)90309-7)
- Faivre, D., & Schüler, D. (2008). Magnetotactic Bacteria and Magnetosomes. *Chemical Reviews*, *108*(11), 4875–4898. <https://doi.org/10.1021/cr078258w>
- Fontana, A., de Laureto, P. P., Spolaore, B., & Frare, E. (2012). Identifying Disordered Regions in Proteins by Limited Proteolysis. In V. N. Uversky & A. K. Dunker (Eds.), *Intrinsically Disordered Protein Analysis* (Vol. 896, pp. 297–318). Springer New York.
https://doi.org/10.1007/978-1-4614-3704-8_20
- Fukushima, M., Iiyama, K., Yamashita, J., Furue, M., Tsuji, G., Imanishi, S., Mon, H., Lee, J. M., & Kusakabe, T. (2013). PRODUCTION OF SMALL ANTIBACTERIAL PEPTIDES USING SILKWORM–BACULOVIRUS PROTEIN EXPRESSION SYSTEM. *Preparative Biochemistry and Biotechnology*, *43*(6), 565–576.
<https://doi.org/10.1080/10826068.2012.762717>
- Gadd, G. M., & Pan, X. (2016). Biomineralization, Bioremediation and Biorecovery of Toxic Metals and Radionuclides. *Geomicrobiology Journal*, *33*(3–4), 175–178.
<https://doi.org/10.1080/01490451.2015.1087603>
- Gaëtan, J., Halary, S., Millet, M., Bernard, C., Duval, C., Hamlaoui, S., Hecquet, A., Gugger, M., Marie, B., Mehta, N., Moreira, D., Skouri-Panet, F., Travert, C., Duprat, E., Leloup, J., & Benzerara, K. (2023). Widespread formation of intracellular calcium carbonates by the bloom-forming cyanobacterium *MICROCYSTIS*. *Environmental Microbiology*, *25*(3), 751–765.
<https://doi.org/10.1111/1462-2920.16322>

- García-Bayona, L., Guo, M. S., & Laub, M. T. (2017). Contact-dependent killing by *Caulobacter crescentus* via cell surface-associated, glycine zipper proteins. *eLife*, 6, e24869.
<https://doi.org/10.7554/eLife.24869>
- Garcia-Carreno, F. L. (1992). Protease inhibition in theory and practice. *Biotechnology Education*.
- Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., & Bairoch, A. (2005). Protein Identification and Analysis Tools on the ExPASy Server. In J. M. Walker (Ed.), *The Proteomics Protocols Handbook* (pp. 571–607). Humana Press. <https://doi.org/10.1385/1-59259-890-0:571>
- Gautron, J., Stapane, L., Le Roy, N., Nys, Y., Rodriguez-Navarro, A. B., & Hincke, M. T. (2021). Avian eggshell biomineralization: An update on its structure, mineralogy and protein tool kit. *BMC Molecular and Cell Biology*, 22(1), 11. <https://doi.org/10.1186/s12860-021-00350-0>
- Gilbert, P. U. P. A., Bergmann, K. D., Boekelheide, N., Tambutté, S., Mass, T., Marin, F., Adkins, J. F., Erez, J., Gilbert, B., Knutson, V., Cantine, M., Ortega Hernandez, J., & Knoll, A. H. (2022). Biomineralization: Integrating mechanism and evolutionary history. *Science Advances*, 8(10), 17.
- Golomb, M., & Chamberlin, M. (1974). Characterization of T7-specific Ribonucleic Acid Polymerase. *Journal of Biological Chemistry*, 249(9), 2858–2863. [https://doi.org/10.1016/S0021-9258\(19\)42709-9](https://doi.org/10.1016/S0021-9258(19)42709-9)
- Goodwin, A. L., Michel, F. M., Phillips, B. L., Keen, D. A., Dove, M. T., & Reeder, R. J. (2010). Nanoporous Structure and Medium-Range Order in Synthetic Amorphous Calcium Carbonate. *Chemistry of Materials*, 22(10), 3197–3205. <https://doi.org/10.1021/cm100294d>
- Görge, S. (2017). *Processus moléculaires de la formation de carbonates de calcium par les cyanobactéries: Étude d'un gène candidat* (p. 36).
- Görge, S. (2020). *Les mécanismes moléculaires de la biominéralisation intracellulaire de CaCO₃ chez les cyanobactéries*. Sorbonne University.

- Görge, S., Benzerara, K., Skouri-Panet, F., Gugger, M., Chauvat, F., & Cassier-Chauvat, C. (2021). The diversity of molecular mechanisms of carbonate biomineralization by bacteria. *Discover Materials*, 1(1), 2. <https://doi.org/10.1007/s43939-020-00001-9>
- Gourdon, P., Liu, X.-Y., Skjørringe, T., Morth, J. P., Møller, L. B., Pedersen, B. P., & Nissen, P. (2011). Crystal structure of a copper-transporting PIB-type ATPase. *Nature*, 475(7354), 59–64. <https://doi.org/10.1038/nature10191>
- Grishin, N. V. (2001). Fold Change in Evolution of Protein Structures. *Journal of Structural Biology*, 134(2–3), 167–185. <https://doi.org/10.1006/jsbi.2001.4335>
- Grisshammer, R., & Tucker, J. (1997). Quantitative Evaluation of Neurotensin Receptor Purification by Immobilized Metal Affinity Chromatography. *Protein Expression and Purification*, 11(1), 53–60. <https://doi.org/10.1006/prev.1997.0766>
- Grodberg, J., & Dunn, J. J. (1988). ompT encodes the Escherichia coli outer membrane protease that cleaves T7 RNA polymerase during purification. *Journal of Bacteriology*, 170(3), 1245–1253. <https://doi.org/10.1128/jb.170.3.1245-1253.1988>
- Guerrero-Valero, M., Ferrer-Orta, C., Querol-Audí, J., Marin-Vicente, C., Fita, I., Gómez-Fernández, J. C., Verdaguer, N., & Corbalán-García, S. (2009). Structural and mechanistic insights into the association of PKC α -C2 domain to PtdIns(4,5)P₂. *Proceedings of the National Academy of Sciences*, 106(16), 6603–6607. <https://doi.org/10.1073/pnas.0813099106>
- Guillén, J., Ferrer-Orta, C., Buxaderas, M., Pérez-Sánchez, D., Guerrero-Valero, M., Luengo-Gil, G., Pous, J., Guerra, P., Gómez-Fernández, J. C., Verdaguer, N., & Corbalán-García, S. (2013). Structural insights into the Ca²⁺ and PI(4,5)P₂ binding modes of the C2 domains of rabphilin 3A and synaptotagmin 1. *Proceedings of the National Academy of Sciences*, 110(51), 20503–20508. <https://doi.org/10.1073/pnas.1316179110>
- Guinier, A. (1939). La diffraction des rayons X aux très petits angles: Application à l'étude de phénomènes ultramicroscopiques. *Annales de Physique*, 11(12), 161–237. <https://doi.org/10.1051/anphys/193911120161>

- Gutfreund, H., Kihne, W., & Kihne, F. (1976). WILHELM FRIEDRICH KOHNE; AN APPRECIATION. *FEBS Letters*, 62, E1–E12.
- Gwenzi, W. (2019). Carbon Sequestration via Biomineralization: Processes, Applications and Future Directions. In Inamuddin, A. M. Asiri, & E. Lichtfouse (Eds.), *Sustainable Agriculture Reviews* 37 (Vol. 37, pp. 93–106). Springer International Publishing. https://doi.org/10.1007/978-3-030-29298-0_5
- Hallgren, J., Tsigos, K. D., Pedersen, M. D., Almagro Armenteros, J. J., Marcatili, P., Nielsen, H., Krogh, A., & Winther, O. (2022). *DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks* [Preprint]. *Bioinformatics*. <https://doi.org/10.1101/2022.04.08.487609>
- Hazen, R. M., Papineau, D., Bleeker, W., Downs, R. T., Ferry, J. M., McCoy, T. J., Sverjensky, D. A., & Yang, H. (2008). Mineral evolution. *American Mineralogist*, 93(11–12), 1693–1720. <https://doi.org/10.2138/am.2008.2955>
- Hearnshaw, S., West, C., Singleton, C., Zhou, L., Kihlken, M. A., Strange, R. W., Le Brun, N. E., & Hemmings, A. M. (2009). A Tetranuclear Cu(I) Cluster in the Metallochaperone Protein CopZ. *Biochemistry*, 48(40), Article 40. <https://doi.org/10.1021/bi9011995>
- Hebditch, M., Carballo-Amador, M. A., Charonis, S., Curtis, R., & Warwicker, J. (2017). Protein–Sol: A web tool for predicting protein solubility from sequence. *Bioinformatics*, 33(19), 3098–3100. <https://doi.org/10.1093/bioinformatics/btx345>
- Holm, L. (2020). Using Dali for Protein Structure Comparison. In Z. Gáspári (Ed.), *Structural Bioinformatics* (Vol. 2112, pp. 29–42). Springer US. https://doi.org/10.1007/978-1-0716-0270-6_3
- Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., & Rives, A. (2022). *Learning inverse folding from millions of predicted structures* [Preprint]. *Systems Biology*. <https://doi.org/10.1101/2022.04.10.487779>
- Hu, M. Y., Petersen, I., Chang, W. W., Blurton, C., & Stumpp, M. (2020). Cellular bicarbonate accumulation and vesicular proton transport promote calcification in the sea urchin larva.

Proceedings of the Royal Society B: Biological Sciences, 287(1934), 20201506.

<https://doi.org/10.1098/rspb.2020.1506>

- Innocenti Malini, R., Finney, A. R., Hall, S. A., Freeman, C. L., & Harding, J. H. (2017). The Water–Amorphous Calcium Carbonate Interface and Its Interactions with Amino Acids. *Crystal Growth & Design*, 17(11), 5811–5822. <https://doi.org/10.1021/acs.cgd.7b00874>
- Jacques, D. A., Langley, D. B., Hynson, R. M. G., Whitten, A. E., Kwan, A., Guss, J. M., & Trehwella, J. (2011). A Novel Structure of an Antikinase and its Inhibitor. *Journal of Molecular Biology*, 405(1), 214–226. <https://doi.org/10.1016/j.jmb.2010.10.047>
- Javaux, E. J. (2019). Challenges in evidencing the earliest traces of life. *Nature*, 572(7770), 451–460. <https://doi.org/10.1038/s41586-019-1436-4>
- Jiang, D., Zhao, Y., Fan, J., Liu, X., Wu, Y., Feng, W., & Zhang, X. C. (2014). Atomic resolution structure of the E. coli YajR transporter YAM domain. *Biochemical and Biophysical Research Communications*, 450(2), Article 2. <https://doi.org/10.1016/j.bbrc.2014.06.053>
- Jiang, D., Zhao, Y., Wang, X., Fan, J., Heng, J., Liu, X., Feng, W., Kang, X., Huang, B., Liu, J., & Zhang, X. C. (2013). Structure of the YajR transporter suggests a transport mechanism based on the conserved motif A. *Proceedings of the National Academy of Sciences*, 110(36), Article 36. <https://doi.org/10.1073/pnas.1308127110>
- Jiang, L.-Q., Carter, B. R., Feely, R. A., Lauvset, S. K., & Olsen, A. (2019). Surface ocean pH and buffer capacity: Past, present and future. *Scientific Reports*, 9(1), 18624. <https://doi.org/10.1038/s41598-019-55039-4>
- Jin, T., Chuenchor, W., Jiang, J., Cheng, J., Li, Y., Fang, K., Huang, M., Smith, P., & Xiao, T. S. (2017). Design of an expression system to enhance MBP-mediated crystallization. *Scientific Reports*, 7(1), 40991. <https://doi.org/10.1038/srep40991>
- Jones, D. T., & Thornton, J. M. (2022). The impact of AlphaFold2 one year on. *Nature Methods*, 19(1), 15–20. <https://doi.org/10.1038/s41592-021-01365-3>
- Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong,

- S.-Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236–1240. <https://doi.org/10.1093/bioinformatics/btu031>
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*. <https://doi.org/10.1038/s41586-021-03819-2>
- Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, *22*(12), 2577–2637. <https://doi.org/10.1002/bip.360221211>
- Kalmar, L., Homola, D., Varga, G., & Tompa, P. (2012). Structural disorder in proteins brings order to crystal growth in biomineralization. *Bone*, *51*(3), 528–534. <https://doi.org/10.1016/j.bone.2012.05.009>
- Kapłon, T. M., Michnik, A., Drzazga, Z., Richter, K., Kochman, M., & Ożyhar, A. (2009). The rod-shaped conformation of Starmaker. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, *1794*(11), 1616–1624. <https://doi.org/10.1016/j.bbapap.2009.07.010>
- Kapust, R. B., & Waugh, D. S. (1999). *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Science*, *8*(8), 1668–1674. <https://doi.org/10.1110/ps.8.8.1668>
- Karathanassis, D., Stahelin, R. V., Bravo, J., Perisic, O., Pacold, C. M., Cho, W., & Williams, R. L. (2002). Binding of the PX domain of p47phox to phosphatidylinositol 3,4-bisphosphate and phosphatidic acid is masked by an intramolecular interaction. *The EMBO Journal*, *21*(19), 5057–5068. <https://doi.org/10.1093/emboj/cdf519>
- Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. <https://doi.org/10.1093/molbev/mst010>

- Kazanov, M. D., Igarashi, Y., Eroshkin, A. M., Cieplak, P., Ratnikov, B., Zhang, Y., Li, Z., Godzik, A., Osterman, A. L., & Smith, J. W. (2011). Structural Determinants of Limited Proteolysis. *Journal of Proteome Research*, *10*(8), 3642–3651. <https://doi.org/10.1021/pr200271w>
- Kellermann, O., & Szmelcman, S. (1974). Active Transport of Maltose in Escherichia coli K12. Involvement of a “Periplasmic” Maltose Binding Protein. *European Journal of Biochemistry*, *47*(1), 139–149. <https://doi.org/10.1111/j.1432-1033.1974.tb03677.x>
- Kikhney, A. G., & Svergun, D. I. (2015). A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Letters*, *589*(19PartA), 2570–2577. <https://doi.org/10.1016/j.febslet.2015.08.027>
- Kim, S., Jeon, T.-J., Oberai, A., Yang, D., Schmidt, J. J., & Bowie, J. U. (2005). Transmembrane glycine zippers: Physiological and pathological roles in membrane proteins. *Proceedings of the National Academy of Sciences*, *102*(40), Article 40. <https://doi.org/10.1073/pnas.0501234102>
- Kinney, J. N., Axen, S. D., & Kerfeld, C. A. (2011). Comparative analysis of carboxysome shell proteins. *Photosynthesis Research*, *109*(1–3), 21–32. <https://doi.org/10.1007/s11120-011-9624-6>
- Koga, N., Nakagoe, Y., & Tanaka, H. (1998). Crystallization of amorphous calcium carbonate. *Thermochimica Acta*, *318*(1–2), 239–244. [https://doi.org/10.1016/S0040-6031\(98\)00348-7](https://doi.org/10.1016/S0040-6031(98)00348-7)
- Kralj, T., Nuske, M., Hofferek, V., Sani, M.-A., Lee, T.-H., Separovic, F., Aguilar, M.-I., & Reid, G. E. (2022). Multi-Omic Analysis to Characterize Metabolic Adaptation of the E. coli Lipidome in Response to Environmental Stress. *Metabolites*, *12*(2), 171. <https://doi.org/10.3390/metabo12020171>
- Kühne, W. F. (1877). Ueber das Tyrsin (Enzym des Pankreas). *Verhandlungen des Heidelb. Naturhist.-Med.*, *1*, 194–198.
- Kumar, A., Mohanram, H., Li, J., Le Ferrand, H., Verma, C. S., & Miserez, A. (2020). Disorder–Order Interplay of a Barnacle Cement Protein Triggered by Interactions with Calcium and Carbonate Ions: A Molecular Dynamics Study. *Chemistry of Materials*, *32*(20), 8845–8859. <https://doi.org/10.1021/acs.chemmater.0c02319>

- Kutateladze, T. G. (2006). Phosphatidylinositol 3-phosphate recognition and membrane docking by the FYVE domain. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, 1761(8), 868–877. <https://doi.org/10.1016/j.bbalip.2006.03.011>
- Kwon, S.-K., Kim, S. K., Lee, D.-H., & Kim, J. F. (2015). Comparative genomics and experimental evolution of Escherichia coli BL21(DE3) strains reveal the landscape of toxicity escape from membrane protein overproduction. *Scientific Reports*, 5(1), 16076. <https://doi.org/10.1038/srep16076>
- Laipnik, R., Bissi, V., Sun, C.-Y., Falini, G., Gilbert, P. U. P. A., & Mass, T. (2020). Coral acid rich protein selects vaterite polymorph in vitro. *Journal of Structural Biology*, 209(2), 107431. <https://doi.org/10.1016/j.jsb.2019.107431>
- Lamiable, A., Bitard-Feildel, T., Rebehmed, J., Quintus, F., Schoentgen, F., Mornon, J.-P., & Callebaut, I. (2019). A topology-based investigation of protein interaction sites using Hydrophobic Cluster Analysis. *Biochimie*, 167, 68–80. <https://doi.org/10.1016/j.biochi.2019.09.009>
- Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993). PROCHECK: A program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, 26(2), Article 2. <https://doi.org/10.1107/S0021889892009944>
- Le Roy, N., Jackson, D. J., Marie, B., Ramos-Silva, P., & Marin, F. (2014). The evolution of metazoan α -carbonic anhydrases and their roles in calcium carbonate biomineralization. *Frontiers in Zoology*, 11(1), 75. <https://doi.org/10.1186/s12983-014-0075-8>
- Ledwidge, R., Patel, B., Dong, A., Fiedler, D., Falkowski, M., Zelikova, J., Summers, A. O., Pai, E. F., & Miller, S. M. (2005). NmerA, the Metal Binding Domain of Mercuric Ion Reductase, Removes Hg²⁺ from Proteins, Delivers It to the Catalytic Core, and Protects Cells under Glutathione-Depleted Conditions. *Biochemistry*, 44(34), 11402–11416. <https://doi.org/10.1021/bi050519d>
- Lem, N. W., & Stumpf, P. K. (1984). In Vitro Fatty Acid Synthesis and Complex Lipid Metabolism in the Cyanobacterium *Anabaena variabilis*. *Plant Physiology*, 74(1), 134–138.

- Lemmon, M. A. (2008). Membrane recognition by phospholipid-binding domains. *Nature Reviews Molecular Cell Biology*, 9(2), 99–111. <https://doi.org/10.1038/nrm2328>
- Lennox, E. S. (1955). Transduction of linked genetic characters of the host by bacteriophage P1. *Virology*, 1(2), 190–206. [https://doi.org/10.1016/0042-6822\(55\)90016-7](https://doi.org/10.1016/0042-6822(55)90016-7)
- Lesley, S. A., & Wilson, I. A. (2005). Protein Production and Crystallization at the Joint Center for Structural Genomics. *Journal of Structural and Functional Genomics*, 6(2–3), 71–79. <https://doi.org/10.1007/s10969-005-2897-2>
- Li, H., Sun, C.-Y., Fang, Y., Carlson, C. M., Xu, H., Ješovnik, A., Sosa-Calvo, J., Zarnowski, R., Bechtel, H. A., Fournelle, J. H., Andes, D. R., Schultz, T. R., Gilbert, P. U. P. A., & Currie, C. R. (2020). Biomineral armor in leaf-cutter ants. *Nature Communications*, 11(1), 5792. <https://doi.org/10.1038/s41467-020-19566-3>
- Li, J., Benzerara, K., Bernard, S., & Beyssac, O. (2013). The link between biomineralization and fossilization of bacteria: Insights from field and experimental studies. *Chemical Geology*, 359, 49–69. <https://doi.org/10.1016/j.chemgeo.2013.09.013>
- Li, J., Margaret Oliver, I., Cam, N., Boudier, T., Blondeau, M., Leroy, E., Cosmidis, J., Skouri-Panet, F., Guigner, J.-M., Férard, C., Poinot, M., Moreira, D., Lopez-Garcia, P., Cassier-Chauvat, C., Chauvat, F., & Benzerara, K. (2016). Biomineralization Patterns of Intracellular Carbonatogenesis in Cyanobacteria: Molecular Hypotheses. *Minerals*, 6(1), Article 1. <https://doi.org/10.3390/min6010010>
- Lichty, J. J., Malecki, J. L., Agnew, H. D., Michelson-Horowitz, D. J., & Tan, S. (2005). Comparison of affinity tags for protein purification. *Protein Expression and Purification*, 41(1), 98–105. <https://doi.org/10.1016/j.pep.2005.01.019>
- Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2022). *Evolutionary-scale prediction of atomic level protein structure with a language model* [Preprint]. *Synthetic Biology*. <https://doi.org/10.1101/2022.07.20.500902>

- Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2023). *Evolutionary-scale prediction of atomic-level protein structure with a language model*.
- Lyons, J. A., Timcenko, M., Dieudonné, T., Lenoir, G., & Nissen, P. (2020). P4-ATPases: How an old dog learnt new tricks — structure and mechanism of lipid flippases. *Current Opinion in Structural Biology*, 63, 65–73. <https://doi.org/10.1016/j.sbi.2020.04.001>
- Mackinder, L., Wheeler, G., Schroeder, D., Von Dassow, P., Riebesell, U., & Brownlee, C. (2011). Expression of biomineralization-related ion transport genes in *Emiliana huxleyi*: Biomineralization in *Emiliana huxleyi*. *Environmental Microbiology*, 13(12), 3250–3265. <https://doi.org/10.1111/j.1462-2920.2011.02561.x>
- Malhotra, A. (2009). Chapter 16 Tagging for Protein Expression. In *Methods in Enzymology* (Vol. 463, pp. 239–258). Elsevier. [https://doi.org/10.1016/S0076-6879\(09\)63016-0](https://doi.org/10.1016/S0076-6879(09)63016-0)
- Mann, K., Siedler, F., Treccani, L., Heinemann, F., & Fritz, M. (2007). Perlinhibin, a Cysteine-, Histidine-, and Arginine-Rich Miniprotein from Abalone (*Haliotis laevigata*) Nacre, Inhibits In Vitro Calcium Carbonate Crystallization. *Biophysical Journal*, 93(4), 1246–1254. <https://doi.org/10.1529/biophysj.106.100636>
- Manriquez-Sandoval, E., & Fried, S. D. (2022). DOMAINMAPPER: Accurate domain structure annotation including those with non-contiguous topologies. *Protein Science*, 31(11). <https://doi.org/10.1002/pro.4465>
- Mariani, V., Biasini, M., Barbato, A., & Schwede, T. (2013). IDDT: A local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21), 2722–2728. <https://doi.org/10.1093/bioinformatics/btt473>
- Mariani, V., Kiefer, F., Schmidt, T., Haas, J., & Schwede, T. (2011). Assessment of template based protein structure predictions in CASP9: Assessment of Template-Based Protein Structure Predictions. *Proteins: Structure, Function, and Bioinformatics*, 79(S10), 37–58. <https://doi.org/10.1002/prot.23177>

- Marin, F. (2020). Mollusc shellomes: Past, present and future. *Journal of Structural Biology*, 212(1), 107583. <https://doi.org/10.1016/j.jsb.2020.107583>
- Marin, F., & Luquet, G. (2007). Unusually Acidic Proteins in Biomineralization. In E. Buerlein (Ed.), *Handbook of Biomineralization* (pp. 273–290). Wiley-VCH Verlag GmbH. <https://doi.org/10.1002/9783527619443.ch16>
- Martignier, A., Pacton, M., Filella, M., Jaquet, J.-M., Barja, F., Pollok, K., Langenhorst, F., Lavigne, S., Guagliardo, P., Kilburn, M. R., Thomas, C., Martini, R., & Ariztegui, D. (2017). Intracellular amorphous carbonates uncover a new biomineralization process in eukaryotes. *Geobiology*, 15(2), 240–253. <https://doi.org/10.1111/gbi.12213>
- Martin, R., Gupta, K., Ninan, N. S., Perry, K., & Van Duyne, G. D. (2012). The Survival Motor Neuron Protein Forms Soluble Glycine Zipper Oligomers. *Structure*, 20(11), Article 11. <https://doi.org/10.1016/j.str.2012.08.024>
- Matz, J. M., Drepper, B., Blum, T. B., Van Genderen, E., Burrell, A., Martin, P., Stach, T., Collinson, L. M., Abrahams, J. P., Matuschewski, K., & Blackman, M. J. (2020). A lipocalin mediates unidirectional heme biomineralization in malaria parasites. *Proceedings of the National Academy of Sciences*, 117(28), 16546–16556. <https://doi.org/10.1073/pnas.2001153117>
- Mbaye, M. N., Hou, Q., Basu, S., Teheux, F., Pucci, F., & Rooman, M. (2019). A comprehensive computational study of amino acid interactions in membrane proteins. *Scientific Reports*, 9(1), 12043. <https://doi.org/10.1038/s41598-019-48541-2>
- McPherson, A. (2017). Protein Crystallization. In A. Wlodawer, Z. Dauter, & M. Jaskolski (Eds.), *Protein Crystallography* (Vol. 1607, pp. 17–50). Springer New York. https://doi.org/10.1007/978-1-4939-7000-1_2
- Mehta, N., Benzerara, K., Kocar, B. D., & Chapon, V. (2019). Sequestration of Radionuclides Radium-226 and Strontium-90 by Cyanobacteria Forming Intracellular Calcium Carbonates. *Environmental Science & Technology*, 53(21), 12639–12647. <https://doi.org/10.1021/acs.est.9b03982>

- Mehta, N., Gaëtan, J., Giura, P., Azais, T., & Benzerara, K. (2022). Detection of biogenic amorphous calcium carbonate (ACC) formed by bacteria using FTIR spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 278, 121262.
<https://doi.org/10.1016/j.saa.2022.121262>
- Merten, J. A., Schultz, K. M., & Klug, C. S. (2012). Concentration-dependent oligomerization and oligomeric arrangement of LptA: Oligomerization of LptA Is Concentration Dependent. *Protein Science*, 21(2), 211–218. <https://doi.org/10.1002/pro.2004>
- Mertens, H. D. T., & Svergun, D. I. (2010). Structural characterization of proteins and complexes using small-angle X-ray solution scattering. *Journal of Structural Biology*, 172(1), 128–141.
<https://doi.org/10.1016/j.jsb.2010.06.012>
- Meyer, J., & Riebesell, U. (2015). Reviews and Syntheses: Responses of coccolithophores to ocean acidification: a meta-analysis. *Biogeosciences*, 12(6), 1671–1682. <https://doi.org/10.5194/bg-12-1671-2015>
- Mi, W., Li, Y., Yoon, S. H., Ernst, R. K., Walz, T., & Liao, M. (2017). Structural basis of MsbA-mediated lipopolysaccharide transport. *Nature*, 549(7671), 233–237.
<https://doi.org/10.1038/nature23649>
- Mirdita, M., Ovchinnikov, S., & Steinegger, M. (2021). *ColabFold—Making protein folding accessible to all* [Preprint]. Bioinformatics. <https://doi.org/10.1101/2021.08.15.456425>
- Mirdita, M., von den Driesch, L., Galiez, C., Martin, M. J., Söding, J., & Steinegger, M. (2017). Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research*, 45(D1), D170–D176. <https://doi.org/10.1093/nar/gkw1081>
- Miroux, B., & Walker, J. E. (1996). Over-production of Proteins in *Escherichia coli*: Mutant Hosts that Allow Synthesis of some Membrane Proteins and Globular Proteins at High Levels. *Journal of Molecular Biology*, 260(3), 289–298. <https://doi.org/10.1006/jmbi.1996.0399>
- Mitchell, A. L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M. R., Kale, V., Potter, S. C., Richardson, L. J., Sakharova, E., Scheremetjew, M., Korobeynikov, A., Shlemov, A., Kunyavskaya, O., Lapidus, A., & Finn, R. D. (2019). MGnify: The microbiome

analysis resource in 2020. *Nucleic Acids Research*, gkz1035.

<https://doi.org/10.1093/nar/gkz1035>

Monteil, C. L., Benzerara, K., Menguy, N., Bidaud, C. C., Michot-Achdjian, E., Bolzoni, R., Mathon, F. P., Coutaud, M., Alonso, B., Garau, C., Jézéquel, D., Viollier, E., Ginet, N., Floriani, M., Swaraj, S., Sachse, M., Busigny, V., Duprat, E., Guyot, F., & Lefevre, C. T. (2021). Intracellular amorphous Ca-carbonate and magnetite biomineralization by a magnetotactic bacterium affiliated to the Alphaproteobacteria. *The ISME Journal*, 15(1), 1–18.

<https://doi.org/10.1038/s41396-020-00747-3>

Moreira, D., Tavera, R., Benzerara, K., Skouri-Panet, F., Couradeau, E., Gérard, E., Fonta, C. L., Novelo, E., Zivanovic, Y., & López-García, P. (2017). Description of *Gloeomargarita lithophora* gen. Nov., sp. Nov., a thylakoid-bearing, basal-branching cyanobacterium with intracellular carbonates, and proposal for *Gloeomargaritales* ord. Nov. *International Journal of Systematic and Evolutionary Microbiology*, 67(3), 653–658. <https://doi.org/10.1099/ijsem.0.001679>

Morris, A. L., MacArthur, M. W., Hutchinson, E. G., & Thornton, J. M. (1992). Stereochemical quality of protein structure coordinates. *Proteins: Structure, Function, and Genetics*, 12(4), 345–364.

<https://doi.org/10.1002/prot.340120407>

Morse, J. W., Arvidson, R. S., & Lüttge, A. (2007). Calcium Carbonate Formation and Dissolution.

Chemical Reviews, 107(2), 342–381. <https://doi.org/10.1021/cr050358j>

Murata, N., & Nishida, I. (1987). Lipids of Blue-Green Algae (Cyanobacteria). In *Lipids: Structure and Function* (pp. 315–347). Elsevier. <https://doi.org/10.1016/B978-0-12-675409-4.50018-6>

Mutsuda, M., Michel, K.-P., Zhang, X., Montgomery, B. L., & Golden, S. S. (2003). Biochemical Properties of CikA, an Unusual Phytochrome-like Histidine Protein Kinase That Resets the Circadian Clock in *Synechococcus elongatus* PCC 7942. *Journal of Biological Chemistry*, 278(21), 19102–19110. <https://doi.org/10.1074/jbc.M213255200>

Nakamura, Y. (2002). Complete Genome Structure of the Thermophilic Cyanobacterium *Thermosynechococcus elongatus* BP-1. *DNA Research*, 9(4), 123–130.

<https://doi.org/10.1093/dnares/9.4.123>

- Nakatani, K., Ishikawa, H., Aono, S., & Mizutani, Y. (2013). Heme-binding properties of heme detoxification protein from *Plasmodium falciparum*. *Biochemical and Biophysical Research Communications*, 439(4), 477–480. <https://doi.org/10.1016/j.bbrc.2013.08.100>
- Nickel, E. H. (1995). International Mineralogical Association, Commission on New Minerals and Mineral Names: Definition of a mineral. *Mineralogy and Petrology*, 55(4), 323–326. <https://doi.org/10.1007/BF01165125>
- Nims, C., Cron, B., Wetherington, M., Macalady, J., & Cosmidis, J. (2019). Low frequency Raman Spectroscopy for micron-scale and in vivo characterization of elemental sulfur in microbial samples. *Scientific Reports*, 9(1), 7971. <https://doi.org/10.1038/s41598-019-44353-6>
- Niwa, T., Ying, B.-W., Saito, K., Jin, W., Takada, S., Ueda, T., & Taguchi, H. (2009). Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of *Escherichia coli* proteins. *Proceedings of the National Academy of Sciences*, 106(11), 4201–4206. <https://doi.org/10.1073/pnas.0811922106>
- Noble, J. E., & Bailey, M. J. A. (2009). Chapter 8 Quantitation of Protein. In *Methods in Enzymology* (Vol. 463, pp. 73–95). Elsevier. [https://doi.org/10.1016/S0076-6879\(09\)63008-1](https://doi.org/10.1016/S0076-6879(09)63008-1)
- Nys, Y., & Le Roy, N. (2018). Calcium Homeostasis and Eggshell Biomineralization in Female Chicken. In *Vitamin D* (pp. 361–382). Elsevier. <https://doi.org/10.1016/B978-0-12-809965-0.00022-7>
- Olejarz, J., Iwasa, Y., Knoll, A. H., & Nowak, M. A. (2021). The Great Oxygenation Event as a consequence of ecological dynamics modulated by planetary change. *Nature Communications*, 12(1), 3985. <https://doi.org/10.1038/s41467-021-23286-7>
- Padayatti, P. S., Lee, S. C., Stanfield, R. L., Wen, P.-C., Tajkhorshid, E., Wilson, I. A., & Zhang, Q. (2019). Structural Insights into the Lipid A Transport Pathway in MsbA. *Structure*, 27(7), 1114–1123.e3. <https://doi.org/10.1016/j.str.2019.04.007>
- Palmgren, M. G., & Nissen, P. (2011). P-Type ATPases. *Annual Review of Biophysics*, 40(1), 243–266. <https://doi.org/10.1146/annurev.biophys.093008.131331>

- Parks, T. D., Leuther, K. K., Howard, E. D., Johnston, S. A., & Dougherty, W. G. (1994). Release of Proteins and Peptides from Fusion Proteins Using a Recombinant Plant Virus Proteinase. *Analytical Biochemistry*, 216(2), 413–417. <https://doi.org/10.1006/abio.1994.1060>
- Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B. L., Salazar, G. A., Bileschi, M. L., Bork, P., Bridge, A., Colwell, L., Gough, J., Haft, D. H., Letunić, I., Marchler-Bauer, A., Mi, H., Natale, D. A., Orengo, C. A., Pandurangan, A. P., Rivoire, C., ... Bateman, A. (2023). InterPro in 2022. *Nucleic Acids Research*, 51(D1), D418–D427. <https://doi.org/10.1093/nar/gkac993>
- Petroutsos, D., Amiar, S., Abida, H., Dolch, L.-J., Bastien, O., Rébeillé, F., Jouhet, J., Falconet, D., Block, M. A., McFadden, G. I., Bowler, C., Botté, C., & Maréchal, E. (2014). Evolution of galactoglycerolipid biosynthetic pathways – From cyanobacteria to primary plastids and from primary to secondary plastids. *Progress in Lipid Research*, 54, 68–85. <https://doi.org/10.1016/j.plipres.2014.02.001>
- Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera?A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13), 1605–1612. <https://doi.org/10.1002/jcc.20084>
- Pioszak, A. A., & Xu, H. E. (2008). Molecular recognition of parathyroid hormone by its G protein-coupled receptor. *Proceedings of the National Academy of Sciences*, 105(13), 5034–5039. <https://doi.org/10.1073/pnas.0801027105>
- Pizzagalli, M. D., Bensimon, A., & Superti-Furga, G. (2021). A guide to plasma membrane solute carrier proteins. *The FEBS Journal*, 288(9), 2784–2835. <https://doi.org/10.1111/febs.15531>
- Polowczyk, I., Bastrzyk, A., & Fiedot, M. (2016). Protein-Mediated Precipitation of Calcium Carbonate. *Materials*, 9(11), 944. <https://doi.org/10.3390/ma9110944>
- Ponce-Toledo, R. I., Deschamps, P., López-García, P., Zivanovic, Y., Benzerara, K., & Moreira, D. (2017). An Early-Branching Freshwater Cyanobacterium at the Origin of Plastids. *Current Biology*, 27(3), 386–391. <https://doi.org/10.1016/j.cub.2016.11.056>

- Porath, J., Carlsson, J., Olsson, I., & Belfrage, G. (1975). Metal chelate affinity chromatography, a new approach to protein fractionation. *Nature*, 258(5536), 598–599.
<https://doi.org/10.1038/258598a0>
- Price, G. D., & Howitt, S. M. (2011). The cyanobacterial bicarbonate transporter BicA: Its physiological role and the implications of structural similarities with human SLC26 transporters This paper is one of a selection of papers published in a Special Issue entitled CSBMCB 53rd Annual Meeting — Membrane Proteins in Health and Disease, and has undergone the Journal's usual peer review process. *Biochemistry and Cell Biology*, 89(2), 178–188. <https://doi.org/10.1139/O10-136>
- Rani, R. S., & Saharay, M. (2019). Molecular dynamics simulation of protein-mediated biomineralization of amorphous calcium carbonate. *RSC Advances*, 9(3), 1653–1663.
<https://doi.org/10.1039/C8RA08459A>
- Rao, A., Seto, J., Berg, J. K., Kreft, S. G., Scheffner, M., & Cölfen, H. (2013). Roles of larval sea urchin spicule SM50 domains in organic matrix self-assembly and calcium carbonate mineralization. *Journal of Structural Biology*, 183(2), 205–215.
<https://doi.org/10.1016/j.jsb.2013.06.004>
- Raran-Kurussi, S., Cherry, S., Zhang, D., & Waugh, D. S. (2017). Removal of Affinity Tags with TEV Protease. In N. A. Burgess-Brown (Ed.), *Heterologous Gene Expression in E.coli* (Vol. 1586, pp. 221–230). Springer New York. https://doi.org/10.1007/978-1-4939-6887-9_14
- Raran-Kurussi, S., & Waugh, D. S. (2012). The Ability to Enhance the Solubility of Its Fusion Partners Is an Intrinsic Property of Maltose-Binding Protein but Their Folding Is Either Spontaneous or Chaperone-Mediated. *PLoS ONE*, 7(11), e49589. <https://doi.org/10.1371/journal.pone.0049589>
- Raven, J. A., & Allen, J. F. (2003). Genomics and chloroplast evolution: What did cyanobacteria do for plants? *Genome Biology*.
- Remick, K. A., & Helmann, J. D. (2023). The elements of life: A biocentric tour of the periodic table. In *Advances in Microbial Physiology* (p. S0065291122000339). Elsevier.
<https://doi.org/10.1016/bs.ampbs.2022.11.001>

- Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2012). HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, 9(2), 173–175.
<https://doi.org/10.1038/nmeth.1818>
- Richardson, J. S. (1981). The Anatomy and Taxonomy of Protein Structure. In *Advances in Protein Chemistry* (Vol. 34, pp. 167–339). Elsevier. [https://doi.org/10.1016/S0065-3233\(08\)60520-3](https://doi.org/10.1016/S0065-3233(08)60520-3)
- Riebesell, U., Körtzinger, A., & Oeschlies, A. (2009). Sensitivities of marine carbon fluxes to ocean change. *Proceedings of the National Academy of Sciences*, 106(49), 20602–20609.
<https://doi.org/10.1073/pnas.0813291106>
- Rodriguez-Navarro, C., Cizer, Ö., Kudłacz, K., Ibañez-Velasco, A., Ruiz-Agudo, C., Elert, K., Burgos-Cara, A., & Ruiz-Agudo, E. (2019). The multiple roles of carbonic anhydrase in calcium carbonate mineralization. *CrystEngComm*, 21(48), 7407–7423.
<https://doi.org/10.1039/C9CE01544B>
- Romero, M. F., Chen, A.-P., Parker, M. D., & Boron, W. F. (2013). The SLC4 family of bicarbonate transporters. *Molecular Aspects of Medicine*, 34(2–3), 159–182.
<https://doi.org/10.1016/j.mam.2012.10.008>
- Roose, J. L., Kashino, Y., & Pakrasi, H. B. (2007). The PsbQ protein defines cyanobacterial Photosystem II complexes with highest activity and stability. *Proceedings of the National Academy of Sciences*, 104(7), 2548–2553. <https://doi.org/10.1073/pnas.0609337104>
- Rosenzweig, R., Nillegoda, N. B., Mayer, M. P., & Bukau, B. (2019). The Hsp70 chaperone network. *Nature Reviews Molecular Cell Biology*, 20(11), 665–680. <https://doi.org/10.1038/s41580-019-0133-3>
- Saghäi, A., Zivanovic, Y., Moreira, D., Benzerara, K., Bertolino, P., Ragon, M., Tavera, R., López-Archilla, A. I., & López-García, P. (2016). Comparative metagenomics unveils functions and genome features of microbialite-associated communities along a depth gradient: Comparative metagenomics of microbialites from Lake Alchichica. *Environmental Microbiology*, 18(12), 4990–5004. <https://doi.org/10.1111/1462-2920.13456>

- Sah, N., Kuehu, D. L., Khadka, V. S., Deng, Y., Peplowska, K., Jha, R., & Mishra, B. (2018). RNA sequencing-based analysis of the laying hen uterus revealed the novel genes and biological pathways involved in the eggshell biomineralization. *Scientific Reports*, 8(1), 16853. <https://doi.org/10.1038/s41598-018-35203-y>
- Saito, A., Kagi, H., Marugata, S., Komatsu, K., Enomoto, D., Maruyama, K., & Kawano, J. (2020). Incorporation of Incompatible Strontium and Barium Ions into Calcite (CaCO₃) through Amorphous Calcium Carbonate. *Minerals*, 10(3), 270. <https://doi.org/10.3390/min10030270>
- Saito, H., & Uchida, H. (1977). Initiation of the DNA Replication of Bacteriophage Lambd Escherichia coli K12. *Journal of Molecular Biology*, 113(1), 1–25. [https://doi.org/10.1016/0022-2836\(77\)90038-9](https://doi.org/10.1016/0022-2836(77)90038-9)
- Sansone, V., Maiorano, E., Galluzzo, A., & Pascale, V. (2018). Calcific tendinopathy of the shoulder: Clinical perspectives into the mechanisms, pathogenesis, and treatment. *Orthopedic Research and Reviews, Volume 10*, 63–72. <https://doi.org/10.2147/ORR.S138225>
- Sato, N. (2021). Are Cyanobacteria an Ancestor of Chloroplasts or Just One of the Gene Donors for Plants and Algae? *Genes*, 12(6), 823. <https://doi.org/10.3390/genes12060823>
- Schewiakoff, W. (1893). *Über einen neuen bakteriennähnlichen Organismus des Süßwassers (Habilitationsschrift)*. [Habilitationsschrift]. University of Heidelberg.
- Schmidt, C. A., Stifler, C. A., Luffey, E. L., Fordyce, B. I., Ahmed, A., Barreiro Pujol, G., Breit, C. P., Davison, S. S., Klaus, C. N., Koehler, I. J., LeCloux, I. M., Matute Diaz, C., Nguyen, C. M., Quach, V., Sengkhammee, J. S., Walch, E. J., Xiong, M. M., Tambutté, E., Tambutté, S., ... Gilbert, P. U. P. A. (2022). Faster Crystallization during Coral Skeleton Formation Correlates with Resilience to Ocean Acidification. *Journal of the American Chemical Society*, 144(3), 1332–1341. <https://doi.org/10.1021/jacs.1c11434>
- Schmidt, T. G. M., Koepke, J., Frank, R., & Skerra, A. (1996). Molecular Interaction Between the Strep-tag Affinity Peptide and its Cognate Target, Streptavidin. *Journal of Molecular Biology*, 255(5), 753–766. <https://doi.org/10.1006/jmbi.1996.0061>

- Schmidt, T. G. M., & Skerra, A. (1993). The random peptide library-assisted engineering of a C-terminal affinity peptide, useful for the detection and purification of a functional Ig Fv fragment. *Protein Engineering, Design and Selection*, 6(1), 109–122.
<https://doi.org/10.1093/protein/6.1.109>
- Schmidt, T. G., & Skerra, A. (2007). The Strep-tag system for one-step purification and high-affinity detection or capturing of proteins. *Nature Protocols*, 2(6), 1528–1535.
<https://doi.org/10.1038/nprot.2007.209>
- Schmitt, J., Hess, H., & Stunnenberg, H. G. (1993). Affinity purification of histidine-tagged proteins. *Molecular Biology Reports*, 18(3), 223–230. <https://doi.org/10.1007/BF01674434>
- Sekine, M., Yoshida, A., Kishi, M., Furuya, K., & Toda, T. (2023). Free ammonia tolerance of cyanobacteria depends on intracellular pH. *Biocatalysis and Agricultural Biotechnology*, 47, 102562. <https://doi.org/10.1016/j.bcab.2022.102562>
- Sekkal, W., & Zaoui, A. (2013). Nanoscale analysis of the morphology and surface stability of calcium carbonate polymorphs. *Scientific Reports*, 3(1), 1587. <https://doi.org/10.1038/srep01587>
- Senes, A., Gerstein, M., & Engelman, D. M. (2000). Statistical analysis of amino acid patterns in transmembrane helices: The GxxxG motif occurs frequently and in association with β -branched residues at neighboring positions. *Journal of Molecular Biology*, 296(3), 921–936.
<https://doi.org/10.1006/jmbi.1999.3488>
- Sharma, V., Srinivasan, A., Nikolajeff, F., & Kumar, S. (2021). Biomineralization process in hard tissues: The interaction complexity within protein and inorganic counterparts. *Acta Biomaterialia*, 120, 20–37. <https://doi.org/10.1016/j.actbio.2020.04.049>
- Sippl, M. J. (1993). Recognition of errors in three-dimensional structures of proteins. *Proteins: Structure, Function, and Bioinformatics*, 17(4), Article 4.
<https://doi.org/10.1002/prot.340170404>
- Sivaguru, M., Saw, J. J., Wilson, E. M., Lieske, J. C., Krambeck, A. E., Williams, J. C., Romero, M. F., Fouke, K. W., Curtis, M. W., Kear-Scott, J. L., Chia, N., & Fouke, B. W. (2021). Human kidney

- stones: A natural record of universal biomineralization. *Nature Reviews Urology*, 18(7), 404–432. <https://doi.org/10.1038/s41585-021-00469-x>
- Skeffington, A. W., & Scheffel, A. (2018). Exploiting algal mineralization for nanotechnology: Bringing coccoliths to the fore. *Current Opinion in Biotechnology*, 49, 57–63. <https://doi.org/10.1016/j.copbio.2017.07.013>
- Skinner, H. C. W. (2005). Biominerals. *Mineralogical Magazine*, 69(5), 621–641. <https://doi.org/10.1180/0026461056950275>
- Smilgies, D.-M., & Foltá-Stogniew, E. (2015). Molecular weight–gyration radius relation of globular proteins: A comparison of light scattering, small-angle X-ray scattering and structure-based data. *Journal of Applied Crystallography*, 48(Pt 5), 1604–1606. <https://doi.org/10.1107/S1600576715015551>
- Smith, A. T., Smith, K. P., & Rosenzweig, A. C. (2014). Diversity of the metal-transporting P1B-type ATPases. *JBIC Journal of Biological Inorganic Chemistry*, 19(6), 947–960. <https://doi.org/10.1007/s00775-014-1129-2>
- Soo, R. M., Hemp, J., Parks, D. H., Fischer, W. W., & Hugenholtz, P. (2017). On the origins of oxygenic photosynthesis and aerobic respiration in Cyanobacteria. *Science*, 355(6332), Article 6332. <https://doi.org/10.1126/science.aal3794>
- Soo, R. M., Skennerton, C. T., Sekiguchi, Y., Imelfort, M., Paech, S. J., Dennis, P. G., Steen, J. A., Parks, D. H., Tyson, G. W., & Hugenholtz, P. (2014). An Expanded Genomic Representation of the Phylum Cyanobacteria. *Genome Biology and Evolution*, 6(5), 1031–1045. <https://doi.org/10.1093/gbe/evu073>
- Spurlino, J. C., Lu, G. Y., & Quiocho, F. A. (1991). The 2.3-Å resolution structure of the maltose- or maltodextrin-binding protein, a primary receptor of bacterial active transport and chemotaxis. *Journal of Biological Chemistry*, 266(8), 5202–5219. [https://doi.org/10.1016/S0021-9258\(19\)67774-4](https://doi.org/10.1016/S0021-9258(19)67774-4)

- Stace, C., & Ktistakis, N. (2006). Phosphatidic acid- and phosphatidylserine-binding proteins. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, 1761(8), 913–926. <https://doi.org/10.1016/j.bbalip.2006.03.006>
- Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11), 1026–1028. <https://doi.org/10.1038/nbt.3988>
- Stothard, P. (2000). The Sequence Manipulation Suite: JavaScript Programs for Analyzing and Formatting Protein and DNA Sequences. *BioTechniques*, 28(6), 1102–1104. <https://doi.org/10.2144/00286ir01>
- Studier, F. W., & Moffatt, B. A. (1986). Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *Journal of Molecular Biology*, 189(1), 113–130. [https://doi.org/10.1016/0022-2836\(86\)90385-2](https://doi.org/10.1016/0022-2836(86)90385-2)
- Takeuchi, T., Sarashina, I., Iijima, M., & Endo, K. (2008). *In vitro* regulation of CaCO₃ crystal polymorphism by the highly acidic molluscan shell protein Aspein. *FEBS Letters*, 582(5), 591–596. <https://doi.org/10.1016/j.febslet.2008.01.026>
- Tarczewska, A., Bielak, K., Zoglowek, A., Sołtys, K., Dobryczycki, P., Ożyhar, A., & Różycka, M. (2022). The Role of Intrinsically Disordered Proteins in Liquid–Liquid Phase Separation during Calcium Carbonate Biomineralization. *Biomolecules*, 12(9), 1266. <https://doi.org/10.3390/biom12091266>
- Targowla, S. (2019). *Structural and functional study of Calcyanin, a protein involved in the intracellular biomineralization of calcium carbonates by cyanobacteria, using Gloeomargarita lithophora as a case-study* (p. 20) [M2 Internship].
- Taylor, A. R., Chrachri, A., Wheeler, G., Goddard, H., & Brownlee, C. (2011). A Voltage-Gated H⁺ Channel Underlying pH Homeostasis in Calcifying Coccolithophores. *PLoS Biology*, 9(6), e1001085. <https://doi.org/10.1371/journal.pbio.1001085>
- Teese, M. G., & Langosch, D. (2015). Role of GxxxG Motifs in Transmembrane Domain Interactions. *Biochemistry*, 54(33), Article 33. <https://doi.org/10.1021/acs.biochem.5b00495>

- Téletchéa, S., Esque, J., Urbain, A., Etchebest, C., & De Brevern, A. G. (2023). Evaluation of Transmembrane Protein Structural Models Using HPMScore. *BioMedInformatics*, 3(2), 306–326. <https://doi.org/10.3390/biomedinformatics3020021>
- Thangakani, A. M., Kumar, S., Velmurugan, D., & Gromiha, M. S. M. (2012). How do thermophilic proteins resist aggregation? *Proteins: Structure, Function, and Bioinformatics*, 80(4), 1003–1015. <https://doi.org/10.1002/prot.24002>
- The UniProt Consortium, Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bye-A-Jee, H., Cukura, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Garmiri, P., da Costa Gonzales, L. J., Hatton-Ellis, E., Hussein, A., ... Zhang, J. (2023). UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1), D523–D531. <https://doi.org/10.1093/nar/gkac1052>
- Thomas, C., Aller, S. G., Beis, K., Carpenter, E. P., Chang, G., Chen, L., Dassa, E., Dean, M., Duong Van Hoa, F., Ekiert, D., Ford, R., Gaudet, R., Gong, X., Holland, I. B., Huang, Y., Kahne, D. K., Kato, H., Koronakis, V., Koth, C. M., ... Tampé, R. (2020). Structural and functional diversity calls for a new classification of ABC transporters. *FEBS Letters*, 594(23), 3767–3775. <https://doi.org/10.1002/1873-3468.13935>
- Thureau, A., Roblin, P., & Pérez, J. (2021). BioSAXS on the SWING beamline at Synchrotron SOLEIL. *Journal of Applied Crystallography*, 54(6), 1698–1710. <https://doi.org/10.1107/S1600576721008736>
- Torrecilla, I., Leganés, F., Bonilla, I., & Fernández-Piñas, F. (2000). Use of Recombinant Aequorin to Study Calcium Homeostasis and Monitor Calcium Transients in Response to Heat and Cold Shock in Cyanobacteria. *Plant Physiology*, 123(1), 161–176. <https://doi.org/10.1104/pp.123.1.161>
- Touw, W. G., Baakman, C., Black, J., te Beek, T. A. H., Krieger, E., Joosten, R. P., & Vriend, G. (2015). A series of PDB-related databanks for everyday needs. *Nucleic Acids Research*, 43(D1), D364–D368. <https://doi.org/10.1093/nar/gku1028>

- Toyofuku, T., Matsuo, M. Y., De Nooijer, L. J., Nagai, Y., Kawada, S., Fujita, K., Reichart, G.-J., Nomaki, H., Tsuchiya, M., Sakaguchi, H., & Kitazato, H. (2017). Proton pumping accompanies calcification in foraminifera. *Nature Communications*, *8*(1), 14145.
<https://doi.org/10.1038/ncomms14145>
- Treccani, L., Mann, K., Heinemann, F., & Fritz, M. (2006). Perlwapin, an Abalone Nacre Protein with Three Four-Disulfide Core (Whey Acidic Protein) Domains, Inhibits the Growth of Calcium Carbonate Crystals. *Biophysical Journal*, *91*(7), 2601–2608.
<https://doi.org/10.1529/biophysj.106.086108>
- Trevino, S. R., Scholtz, J. M., & Pace, C. N. (2008). Measuring and Increasing Protein Solubility. *Journal of Pharmaceutical Sciences*, *97*(10), 4155–4166. <https://doi.org/10.1002/jps.21327>
- Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., ... Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature*, *596*(7873), 590–596.
<https://doi.org/10.1038/s41586-021-03828-1>
- Tyszka, J., Bickmeyer, U., Raitzsch, M., Bijma, J., Kaczmarek, K., Mewes, A., Topa, P., & Janse, M. (2019). Form and function of F-actin during biomineralization revealed from live experiments on foraminifera. *Proceedings of the National Academy of Sciences*, *116*(10), 4111–4116.
<https://doi.org/10.1073/pnas.1810394116>
- Umezawa, H., Ueda, M., Maeda, K., Yagishita, K., Kondo, S., Okami, Y., Utahara, R., Osato, Y., Nitta, K., & Takeuchi, T. (1957). Production and Isolation of a New Antibiotic, Kanamycin. *The Journal of Antibiotics, Series A*, *10*(5), 181–188. https://doi.org/10.11554/antibioticsa.10.5_181
- Van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., & Steinegger, M. (2022). *Fast and accurate protein structure search with Foldseek* [Preprint]. *Bioinformatics*. <https://doi.org/10.1101/2022.02.07.479398>
- Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Žídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J.,

- Clancy, E., Green, R., Vora, A., Lutfi, M., ... Velankar, S. (2022). AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, *50*(D1), D439–D444.
<https://doi.org/10.1093/nar/gkab1061>
- Wacey, D. (2012). Earliest evidence for life on Earth: An Australian perspective. *Australian Journal of Earth Sciences*, *59*(2), 153–166. <https://doi.org/10.1080/08120099.2011.592989>
- Wada, H., & Murata, N. (2004). Membrane Lipids in Cyanobacteria. In S. Paul-André & M. Norio (Eds.), *Lipids in Photosynthesis: Structure, Function and Genetics* (Vol. 6, pp. 65–81). Kluwer Academic Publishers. https://doi.org/10.1007/0-306-48087-5_4
- Wagner, S., Klepsch, M. M., Schlegel, S., Appel, A., Draheim, R., Tarry, M., Högbom, M., van Wijk, K. J., Slotboom, D. J., Persson, J. O., & de Gier, J.-W. (2008). Tuning *Escherichia coli* for membrane protein overexpression. *Proceedings of the National Academy of Sciences*, *105*(38), 14371–14376. <https://doi.org/10.1073/pnas.0804090105>
- Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., & Barton, G. J. (2009). Jalview Version 2—A multiple sequence alignment editor and analysis workbench. *Bioinformatics*, *25*(9), 1189–1191. <https://doi.org/10.1093/bioinformatics/btp033>
- Webb, B., & Sali, A. (2016). Comparative Protein Structure Modeling Using MODELLER. *Current Protocols in Bioinformatics*, *54*(1), Article 1. <https://doi.org/10.1002/cpbi.3>
- Weiner, S. (2003). An Overview of Biomineralization Processes and the Problem of the Vital Effect. *Reviews in Mineralogy and Geochemistry*, *54*(1), 1–29. <https://doi.org/10.2113/0540001>
- Weiss, I. M., & Schönitzer, V. (2006). The distribution of chitin in larval shells of the bivalve mollusk *Mytilus galloprovincialis*. *Journal of Structural Biology*, *153*(3), 264–277.
<https://doi.org/10.1016/j.jsb.2005.11.006>
- Weiss, I. M., Schönitzer, V., Eichner, N., & Sumper, M. (2006). The chitin synthase involved in marine bivalve mollusk shell formation contains a myosin domain. *FEBS Letters*, *580*(7), 1846–1852.
<https://doi.org/10.1016/j.febslet.2006.02.044>

- Whitton, B. A. (Ed.). (2012). *Ecology of Cyanobacteria II: Their Diversity in Space and Time*. Springer Netherlands. <https://doi.org/10.1007/978-94-007-3855-3>
- Wiederstein, M., & Sippl, M. J. (2007). ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research*, 35(Web Server), Article Web Server. <https://doi.org/10.1093/nar/gkm290>
- William Studier, F., Rosenberg, A. H., Dunn, J. J., & Dubendorff, J. W. (1990). [6] Use of T7 RNA polymerase to direct expression of cloned genes. In *Methods in Enzymology* (Vol. 185, pp. 60–89). Elsevier. [https://doi.org/10.1016/0076-6879\(90\)85008-C](https://doi.org/10.1016/0076-6879(90)85008-C)
- Wilson, C. J., Choy, W.-Y., & Karttunen, M. (2022). AlphaFold2: A Role for Disordered Protein/Region Prediction? *International Journal of Molecular Sciences*, 23(9), 4591. <https://doi.org/10.3390/ijms23094591>
- Wilson, J. S., Churchill-Angus, A. M., Davies, S. P., Sedelnikova, S. E., Tzokov, S. B., Rafferty, J. B., Bullough, P. A., Bisson, C., & Baker, P. J. (2019). Identification and structural analysis of the tripartite α -pore forming toxin of *Aeromonas hydrophila*. *Nature Communications*, 10(1), 2900. <https://doi.org/10.1038/s41467-019-10777-x>
- Wood, D. W. (2014). New trends and affinity tag designs for recombinant protein purification. *Current Opinion in Structural Biology*, 26, 54–61. <https://doi.org/10.1016/j.sbi.2014.04.006>
- Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X., Luo, S., Su, C., Wu, Z., Xie, Q., Berger, B., Ma, J., & Peng, J. (2022). *High-resolution de novo structure prediction from primary sequence* [Preprint]. Bioinformatics. <https://doi.org/10.1101/2022.07.21.500999>
- Wu, Y., Tam, T. V., Hur, S. H., Rao, P., & Yoo, I.-K. (2023). Biomineralization of titanium dioxide with enhanced photocatalytic activity induced by lysozyme–polystyrene template. *Materials Chemistry and Physics*, 293, 126935. <https://doi.org/10.1016/j.matchemphys.2022.126935>
- Yarra, T., Blaxter, M., & Clark, M. S. (2021). A Bivalve Biomineralization Toolbox. *Molecular Biology and Evolution*, 38(9), 4043–4055. <https://doi.org/10.1093/molbev/msab153>
- Yi, L., Gebhard, M. C., Li, Q., Taft, J. M., Georgiou, G., & Iverson, B. L. (2013). Engineering of TEV protease variants by yeast ER sequestration screening (YESS) of combinatorial libraries.

Proceedings of the National Academy of Sciences, 110(18), 7229–7234.

<https://doi.org/10.1073/pnas.1215994110>

Yoshino, T., Maruyama, K., Kagi, H., Nara, M., & Kim, J. C. (2012). Pressure-Induced Crystallization from Amorphous Calcium Carbonate. *Crystal Growth & Design*, 12(7), 3357–3361.

<https://doi.org/10.1021/cg2017159>

Zavarzin, G. A. (2002). *Microbial Geochemical Calcium Cycle*. 71(1).

Zexer, N., & Elbaum, R. (2022). Hydrogen peroxide modulates silica deposits in sorghum roots.

Journal of Experimental Botany, 73(5), 1450–1463. <https://doi.org/10.1093/jxb/erab497>

Zhao, H., & Lappalainen, P. (2012). A simple guide to biochemical approaches for analyzing protein–lipid interactions. *Molecular Biology of the Cell*, 23(15), 2823–2830.

<https://doi.org/10.1091/mbc.e11-07-0645>

Zhou, X.-X., Wang, Y.-B., Pan, Y.-J., & Li, W.-F. (2008). Differences in amino acids composition and coupling patterns between mesophilic and thermophilic proteins. *Amino Acids*, 34(1), 25–33.

<https://doi.org/10.1007/s00726-007-0589-x>

Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A. N., & Alva, V. (2018). A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *Journal of Molecular Biology*, 430(15), 2237–2243.

<https://doi.org/10.1016/j.jmb.2017.12.007>

Zoccola, D., Ganot, P., Bertucci, A., Caminiti-Segonds, N., Techer, N., Voolstra, C. R., Aranda, M., Tambutté, E., Allemand, D., Casey, J. R., & Tambutté, S. (2015). Bicarbonate transporters in corals point towards a key step in the evolution of cnidarian calcification. *Scientific Reports*,

5(1), 9983. <https://doi.org/10.1038/srep09983>


Annexes :

Annexe 1 : Benzerara, K., Duprat, E., Bitard-Feildel, T., Caumes, G., Cassier-Chauvat, C., Chauvat, F., Dezi, M., Diop, S. I., Gaschignard, G., Görgen, S., Gugger, M., López-García, P., Millet, M., Skouri-Panet, F., Moreira, D., & Callebaut, I. (2022). A New Gene Family Diagnostic for Intracellular Biomineralization of Amorphous Ca Carbonates by Cyanobacteria. *Genome Biology and Evolution*, 14(3), evac026. <https://doi.org/10.1093/gbe/evac026>

Annexe 2 : Gaschignard, G., Millet, M., Bruley, A., Benzerara, K., Dezi, M., Skouri-Panet, F., Duprat, E., & Callebaut, I. AlphaFold2-aided description of CoBaHMA, a novel family of bacterial domains within the Heavy-Metal Associated superfamily. *En cours de soumission*

Les images associées à cet article sont situées à la suite du texte.

A New Gene Family Diagnostic for Intracellular Biomineralization of Amorphous Ca Carbonates by Cyanobacteria

Karim Benzerara ^{1,*,+,#}, Elodie Duprat^{1,+}, Tristan Bitard-Feildel¹, Géraldine Caumes¹, Corinne Cassier-Chauvat², Franck Chauvat², Manuela Dezi¹, Seydina Issa Diop^{1,S}, Geoffroy Gaschignard¹, Sigrid Görden^{1,2}, Muriel Gugger³, Purificación López-García⁴, Maxime Millet¹, Fériel Skouri-Panet¹, David Moreira^{4,+,#} and Isabelle Callebaut^{1,*,+}

¹Sorbonne Université, Muséum National d'Histoire Naturelle, UMR CNRS 7590. Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie (IMPMC), Paris, France

²Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), Gif-sur-Yvette, France

³Institut Pasteur, Université de Paris, Collection of Cyanobacteria, Paris, France

⁴Unité d'Ecologie Systématique et Evolution, CNRS, Université Paris-Saclay, AgroParisTech, Orsay, France

[†]These authors contributed equally to this work.

^{*}Lead contact.

^SPresent address: Department of Systematic and Evolutionary Botany & Zurich-asel Plant Science Center, University of Zurich, Zollikerstrasse 107, Zurich, Switzerland.

*Corresponding authors: E-mails: karim.benzerara@sorbonne-universite.fr; isabelle.callebaut@sorbonne-universite.fr.

Accepted: January 29, 2022

Abstract

Cyanobacteria have massively contributed to carbonate deposition over the geological history. They are traditionally thought to biomineralize CaCO₃ extracellularly as an indirect byproduct of photosynthesis. However, the recent discovery of freshwater cyanobacteria-forming intracellular amorphous calcium carbonates (iACC) challenges this view. Despite the geochemical interest of such a biomineralization process, its molecular mechanisms and evolutionary history remain elusive. Here, using comparative genomics, we identify a new gene (*ccyA*) and protein family (calcyanin) possibly associated with cyanobacterial iACC biomineralization. Proteins of the calcyanin family are composed of a conserved C-terminal domain, which likely adopts an original fold, and a variable N-terminal domain whose structure allows differentiating four major types among the 35 known calcyanin homologs. Calcyanin lacks detectable full-length homologs with known function. The overexpression of *ccyA* in iACC-lacking cyanobacteria resulted in an increased intracellular Ca content. Moreover, *ccyA* presence was correlated and/or colocalized with genes involved in Ca or HCO₃⁻ transport and homeostasis, supporting the hypothesis of a functional role of calcyanin in iACC biomineralization. Whatever its function, *ccyA* appears as diagnostic of intracellular calcification in cyanobacteria. By searching for *ccyA* in publicly available genomes, we identified 13 additional cyanobacterial strains forming iACC, as confirmed by microscopy. This extends our knowledge about the phylogenetic and environmental distribution of cyanobacterial iACC biomineralization, especially with the detection of multicellular genera as well as a marine species. Moreover, *ccyA* was probably present in ancient cyanobacteria, with independent losses in various lineages that resulted in a broad but patchy distribution across modern cyanobacteria.

Key words: biomineralization, amorphous calcium carbonates, cyanobacteria, protein structure prediction, phylogeny, glycine zipper motifs, comparative genomics.

© The Author(s) 2022. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Significance

Few freshwater species of Cyanobacteria have been known to mineralize amorphous CaCO_3 (ACC) intracellularly. Despite the geochemical interest of this biomineralization, its evolutionary history and molecular mechanism remain poorly known. Here, we report the discovery of a new gene family that has no homolog with known function, which proves to be a good diagnostic marker of this process. Using this marker gene, we find new cyanobacteria forming ACC in several genera and environments such as seawater, where ACC biomineralization had not been reported before. Moreover, this gene is ancient and was independently lost in various lineages, resulting in a broad and patchy phylogenetic distribution in modern cyanobacteria.

Introduction

The formation of mineral phases by living organisms is widespread in both eukaryotes and prokaryotes (Weiner and Dove 2003). Although many cases of biomineralization in eukaryotes involve specific genes (Marron et al. 2016; Wang et al. 2021; Yarra et al. 2021), there is presently only one documented case of genetically controlled biomineralization in bacteria: the intracellular magnetite formation by magnetotactic bacteria (Lefevre and Bazylinski 2013). The formation of Ca carbonates by cyanobacteria has been studied for several decades and cyanobacteria are thought to have been among the main calcifiers at the Earth surface since their appearance several billion years ago (Altermann et al. 2006). However, it is only recently that a genetic control of intracellular amorphous calcium carbonates (iACC) biomineralization by some species of cyanobacteria has been hypothesized (Benzerara et al. 2014), but not yet proven. Interestingly, the involvement of ACC has been widely documented and studied in the formation of eukaryotic skeletons (Blue et al. 2017). By contrast and although a growing number of bacterial occurrences are described (Monteil et al. 2021), the determinants of iACC formation in prokaryotes remain poorly understood.

The iACC-biomineralizing cyanobacteria are geographically widespread in freshwater, hot spring, or karstic terrestrial systems (Ragon et al. 2014) and sometimes locally abundant (Bradley et al. 2017). They received particular attention because they challenge the usual paradigm that cyanobacteria biomineralize CaCO_3 extracellularly as an indirect byproduct of photosynthesis only (Altermann et al. 2006). Moreover, the geological history of iACC biomineralization remains mysterious because the fossilization potential of these bacteria appears uncertain (Couradeau et al. 2012; Riding 2012). They can form iACC even under thermodynamically unfavorable conditions, indicating that they consume energy to perform this process, possibly in relation with active sequestration of alkaline earth elements (Cam et al. 2018). An envelope of undetermined composition, either a lipid monolayer and/or proteins, surrounds the iACC granules (Blondeau, Sachse, et al. 2018) and it has been suggested that compartmentation is instrumental for the achievement of local Ca concentrations that are high enough for the formation of iACC (Cam et al. 2018). Furthermore, some iACC-forming species require

higher Ca amounts for optimal growth than iACC-lacking ones, indicating that they possess an unusual Ca homeostasis (De Wever et al. 2019). Interestingly, by forming iACC granules, these cyanobacteria accumulate very high Ca amounts, as well as other alkaline earth elements such as strontium (Sr) and barium (Ba) (Cam et al. 2016; Blondeau, Benzerara, et al. 2018) and may impact the geochemical cycles of these trace elements (Blondeau, Benzerara, et al. 2018). Indeed, by normalizing the uptake to their cell mass, they are among the highest Sr and Ba-scavenging organisms known (Cam et al. 2016). Moreover, they can efficiently sequester radioisotopes such as ^{90}Sr or radium (Ra) isotopes, a capability that may be used for bioremediation (Cam et al. 2016; Blondeau, Benzerara, et al. 2018; Mehta et al. 2019).

All the members of some clades of cyanobacteria, such as the *Cyanothece-Synechococcus-Thermosynechococcus* clade, share this capability to form iACC, suggesting the genetic heritability of this trait in this specific group (Benzerara et al. 2014). Yet, despite the geochemical relevance of this process, the genetic control of iACC formation has not been identified. Moreover, whether the presently known iACC-forming cyanobacteria share ancestral genetic traits related to this biomineralization process or they convergently developed this capability to form iACC during cyanobacterial evolution remains unknown. In the absence of a fossil record, investigating the genetic basis of this biomineralization process appears as the only way to track its geological history.

Results and Discussion

Detection of a Gene Family Diagnostic of iACC Biomineralization

We applied comparative genomics to search for putative genes exclusively shared by iACC-forming cyanobacteria, and therefore absent in iACC-lacking species. We analyzed the genomes of 56 cyanobacterial strains (supplementary table 1, Supplementary Material online), in which the presence or absence of iACC was previously determined by electron microscopy (EM) (Benzerara et al. 2014). Fifty strains were lacking iACC and six were shown to form iACC: *Synechococcus* sp. PCC 6312, *Synechococcus calcipolaris* PCC 11701, *Thermosynechococcus elongatus* BP-1,

Cyanothece sp. PCC 7425, *Chroococciopsis thermalis* PCC 7203, and *Gloeomargarita lithophora* D10. Among the 523,680 translated coding sequences (CDSs) contained in the 56 genomes, only one group of orthologous sequences (among the 27,230 groups comprising at least two sequences) was shared by all six iACC-forming strains and absent in all 50 iACC-lacking strains. The corresponding gene was named *ccyA*. Its predicted protein product was named calcyanin (CcyA as a protein symbol). Conversely, we found no group of orthologous sequences shared by all 50 iACC-lacking strains and absent in all 6 iACC-forming strains. No functional annotation of calcyanin could be achieved using profiles of known protein domain families.

We first investigated the architecture of calcyanin by hydrophobic cluster analysis (HCA), an approach that has already been largely applied to the detection of novel domain families (Callebaut et al. 1997; Bitard-Feildel et al. 2018). The two-dimensional HCA representation of the protein sequence provides structural information based on the distribution of strong hydrophobic amino acids in clusters (representative of regular secondary structures) and their relative arrangement. This last feature allows to appreciate the segmentation of the protein into domains and their intrinsic nature (e.g., folded, disordered), as well as to detect repeated motifs and their overall conservation between sequences. The HCA approach revealed that calcyanin is composed of two domains (fig. 1). The C-terminal domain is composed of three long repeats of a periodic pattern (called GlyZip), including glycine (or small amino acids—indicated in yellow in fig. 1) and hydrophobic amino acids (green) every four residues (long, horizontal clusters). The pattern was clearly distinct for the N-terminal domains, possessing smaller hydrophobic clusters, usually encountered in current globular domains. Although the C-terminal domain was highly conserved in the six different calcyanin sequences, the N-terminal domain appeared to be conserved in five sequences only, and exhibited significant differences in *G. lithophora*. Therefore, we used the conserved C-terminal domain to search for additional homologs in a comprehensive set of 594 cyanobacterial genomes available in public databases. We found additional *ccyA* homologs in 27 strains (supplementary table 2 and fig. 1, Supplementary Material online). Among them, we inspected 17 strains available to us by EM coupled with energy-dispersive X-ray spectrometry (EDXS), which allowed submicrometer-scale chemical mapping of several elements, including Ca and P. As shown by Benzerara et al. (2014) and Li et al. (2016), iACC can be recognized by the fact that they contain Ca but little to no P, in contrast with polyphosphate inclusions, which show a major P EDXS peak with Mg and K and, sometimes, Ca. We detected iACC in 13 of the 17 inspected strains (fig. 2 and supplementary fig. 2, Supplementary Material online), thereby increasing the number of known iACC-forming cyanobacterial species from 6 to 19. Moreover, we detected *ccyA* in the two recently sequenced genomes of *Synechococcus* sp.

PCC 6716 and PCC 6717 that were previously shown to form iACC (Benzerara et al. 2014) (supplementary table 2, Supplementary Material online).

In some strains (e.g., *Fischerella* sp. NIES-4106, *Neosynechococcus sphagnicola* sy1), most of the cells exhibited abundant iACC granules. By contrast, for strains such as *Microcystis aeruginosa* PCC 7806, cells contained none or only few iACC granules. In other strains (e.g., *Chlorogloeopsis fritschii* PCC 9212), the cells contained few iACC granules and many Ca-rich polyphosphate inclusions that could be morphologically confused with iACC by EM alone but not chemically, hence requiring the use of EDXS (fig. 2 and supplementary fig. 2, Supplementary Material online). The four strains possessing *ccyA* but lacking iACC (*C. fritschii* PCC 6912; *Fischerella* sp. NIES-3754; *M. aeruginosa* PCC 9432 and PCC 9717; fig. 3) were phylogenetically very close to iACC-forming relatives. For example, *C. fritschii* PCC 9212 (iACC-forming) and PCC 6912 (no observed iACC) had only few differences in their gene repertoires (supplementary fig. 3 and table 3, Supplementary Material online) and the nucleotide sequences of the genomic regions containing *ccyA* in these two strains (corresponding to contigs of 97,542 and 97,528 bp in length, respectively) shared 100% identity over 97,528 bp. However, 57 genes of *C. fritschii* PCC 9212 had no homolog in *C. fritschii* PCC 6912. Their functional categories were annotated using the NCBI-curated clusters of orthologous groups (COG) protein classification resource. They mostly corresponded to unknown functions (46 without COG hit, 2 genes with COG category X indicating an unknown function) or inorganic ion transport (4 genes, COG category P; supplementary table 3, Supplementary Material online). Moreover, although we did not observe iACC in *C. fritschii* PCC 6912 and *Fischerella* sp. NIES-3754 cells, they both showed Ca- and P-rich inclusions morphologically similar to iACC, suggesting that they may have some but not all the capabilities required to produce iACC (fig. 3). Benzerara et al. (2014) and Cam et al. (2018) previously concluded that iACC-forming strains tend to show iACC inclusions when cultured in different growth media and/or sampled at different stages of their growth. Moreover, we conducted observations on multiple cultures sampled at different times for the four strains, supporting the idea that iACC do not appear transiently in these cultures. However, whether these strains are genetically unable to form iACC or this capability may depend on specific conditions will need to be assessed by future studies. At any rate, the search for *ccyA* in available cyanobacterial genome sequences allowed the detection of 13 additional iACC-forming strains among the 17 strains whose genomes contained *ccyA*, largely extending and optimizing the initial detection of 8 iACC-forming strains (i.e., 6 strains whose genomes were used for comparative genomics plus *Synechococcus* sp. PCC 6716 and PCC 6717 whose genomes were recently sequenced) among 58 randomly selected, phylogenetically diverse cyanobacteria

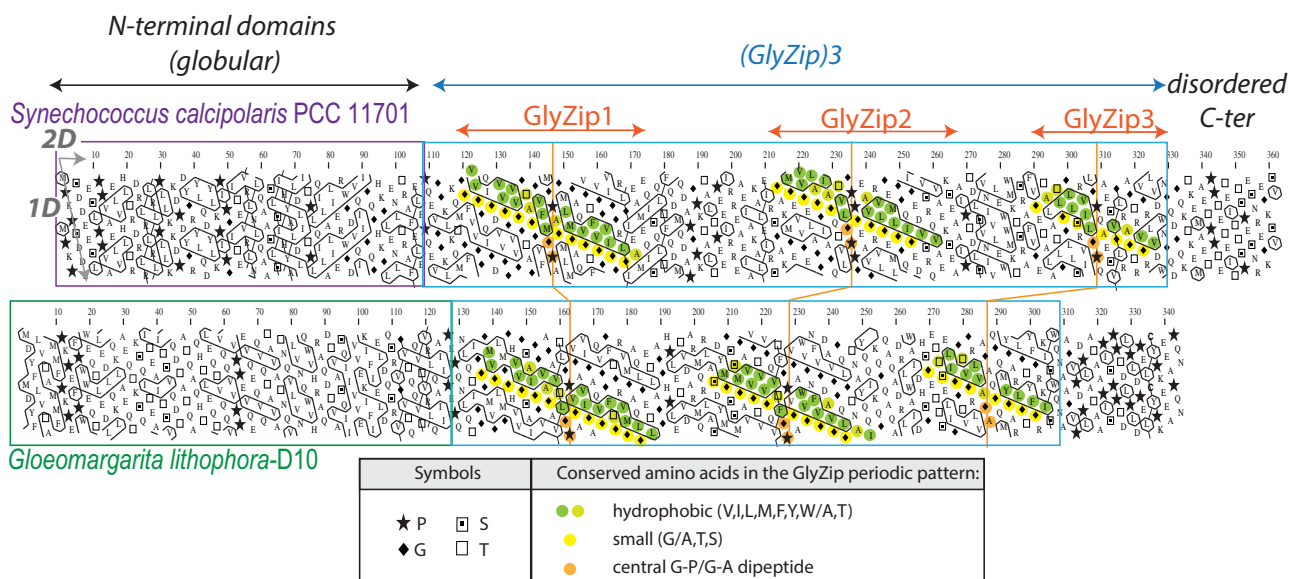


FIG. 1.—Domain architecture of calcyanins, as viewed by HCA. HCA plots of the calcyanin sequences of *Synechococcus calcipolaris* PCC 11701 and *Gloeomargarita lithophora* D10. The protein amino acid sequences (one-letter code) are displayed on a duplicated alpha-helical net, on which the strong hydrophobic amino acids (V, I, L, F, M, Y, and W) are contoured. The latter form clusters, which mainly correspond to the internal faces of regular secondary structures (α -helices and β -strands). The way to read the primary (1D) and secondary (2D) structures is shown with arrows (one amino acid or one hydrophobic cluster after another, respectively), whereas special symbols used for four amino acids with specific structural properties (P, G, S, and T) are described in the inset, together with the color code used to highlight conserved amino acids within the periodic patterns of the two calcyanin sequences. The two distinct CcyA folded domains ($\sim 1/3$ strong hydrophobic amino acids) are boxed.

(Benzerara et al. 2014). Therefore, the search for *ccyA* occurrence significantly increased the probability of success in finding iACC-forming strains (binomial exact test, $P = 9.0e-09$), indicating that *ccyA* can be used as diagnostic marker of intracellular biomineralization.

Thanks to this approach, we expanded considerably the phylogenetic diversity of known iACC-forming cyanobacteria (fig. 4A). So far, iACC biomineralization had been reported in unicellular cyanobacteria only (Benzerara et al. 2014). Here, we find iACC in several multicellular genera belonging to the most complex morphotypes of the cyanobacterial phylum with cellular differentiation and ramifications (*Chlorogloeopsis* and *Fischerella*). Moreover, we also discovered iACC in *M. aeruginosa*, one of the most common, worldwide-distributed bloom-forming cyanobacteria (Humbert et al. 2013). *Microcystis* shows a life cycle with a benthic phase in winter and a planktonic phase in warmer seasons when cells produce gas vesicles to float in the water column (Reynolds and Rogers 1976; Latour et al. 2007). Considering the high density of ACC relative to cells, a controlled production of dense iACC granules might favor a shift to benthic life. Interestingly, *ccyA* is present in the genome of some closely related *M. aeruginosa* strains but absent from others. This finding may be consistent with the high genome plasticity detected in this species, reflecting frequent horizontal gene transfers (HGTs) (Frangeul et al. 2008; Humbert et al. 2013).

The *ccyA* gene and iACC biomineralization were also found in four *Synechococcus*-like strains previously not known to produce iACC (*Neosynechococcus sphagnicola* sy1, *Synechococcus* sp. RS9917, *Synechococcus lividus* PCC 6715, and *Thermosynechococcus* sp. NK55a). *Synechococcus* is a polyphyletic genus, grouping strains isolated from very different environments (Komarek et al. 2020). We previously reported thermophilic and mesophilic freshwater iACC-biomineralizing *Synechococcus* representatives (Benzerara et al. 2014). Here, we significantly expanded this environmental distribution especially with the inclusion of the first marine (*Synechococcus* sp. RS9917) iACC-forming strain.

Sequence-Based Analysis of the Calcyanin Structure

With the exception of the *Thermosynechococcus* sp. NK55a calcyanin, fused with a polypeptide containing a PIN-TRAM domain, the other 34 calcyanin family homologs contained 264–375 amino acids (average 336 ± 25 ; supplementary table 2, Supplementary Material online). All showed the already mentioned two-domain modularity: a variable N-terminal domain and a conserved C-terminal domain.

The N-terminal domain was composed of hydrophobic clusters with lengths and shapes typical of regular secondary structures found in globular domains (Lamiable et al. 2019). According to their N-terminal domains, we classified the 35 calcyanin homologs into four groups: W, X, Y, and Z (fig. 4B). There was only one calcyanin in the X group. Amino acid

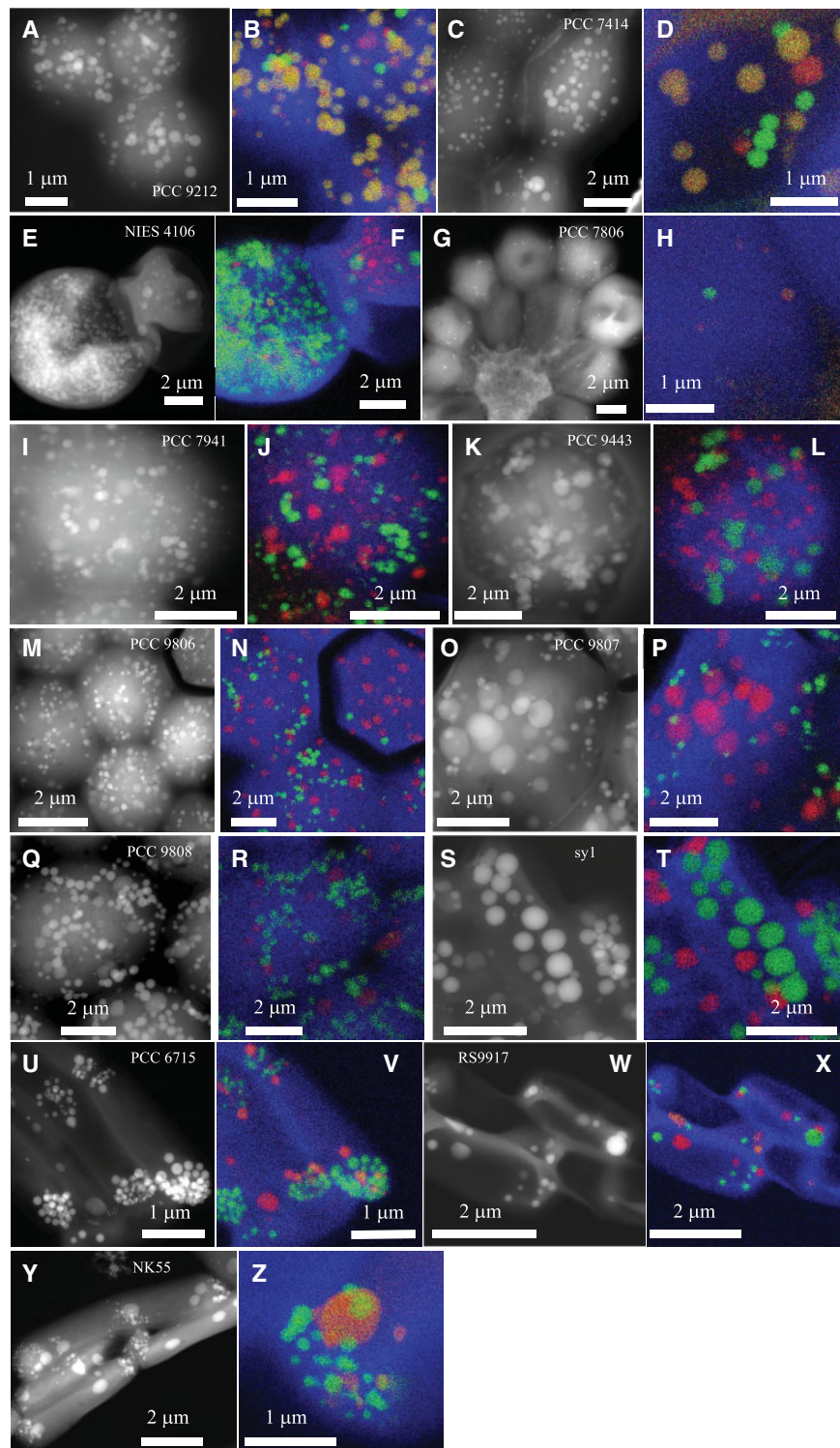


Fig. 2.—EM detection of iACC in 13 calcanin-bearing cyanobacterial strains not previously known to biomineralize carbonates. STEM-HAADF images of the 13 newly identified iACC-forming strains and overlays of C (blue), Ca (green), and P (red) chemical maps as obtained by EDXS. The name of the strains is provided on the STEM-HAADF image. Numbers in parenthesis correspond to replicate numbers of SEM-EDXS, STEM-EDXS, or both analyses. (A and B) *Chlorogloeopsis fritschii* PCC 9212 (13); (C and D) *Fischerella muscicola* PCC 7414 (4); (E and F) *Fischerella* sp. NIES-4106 (5); (G and H) *Microcystis aeruginosa* PCC 7806 (9); (I and J) *M. aeruginosa* PCC 7941 (7); (K and L) *M. aeruginosa* PCC 9443 (3); (M and N) *M. aeruginosa* PCC 9806 (4); (O and P) *M. aeruginosa* PCC 9807 (3); (Q and R) *M. aeruginosa* PCC 9808 (4); (S and T) *Neosynechococcus sphagnicola* sy1 (4); (U and V) *Synechococcus lividus* PCC 6715 (3); (W and X) *Synechococcus* sp. RS9917 (4); (Y and Z) *Thermosynechococcus* sp. NK55 (6).

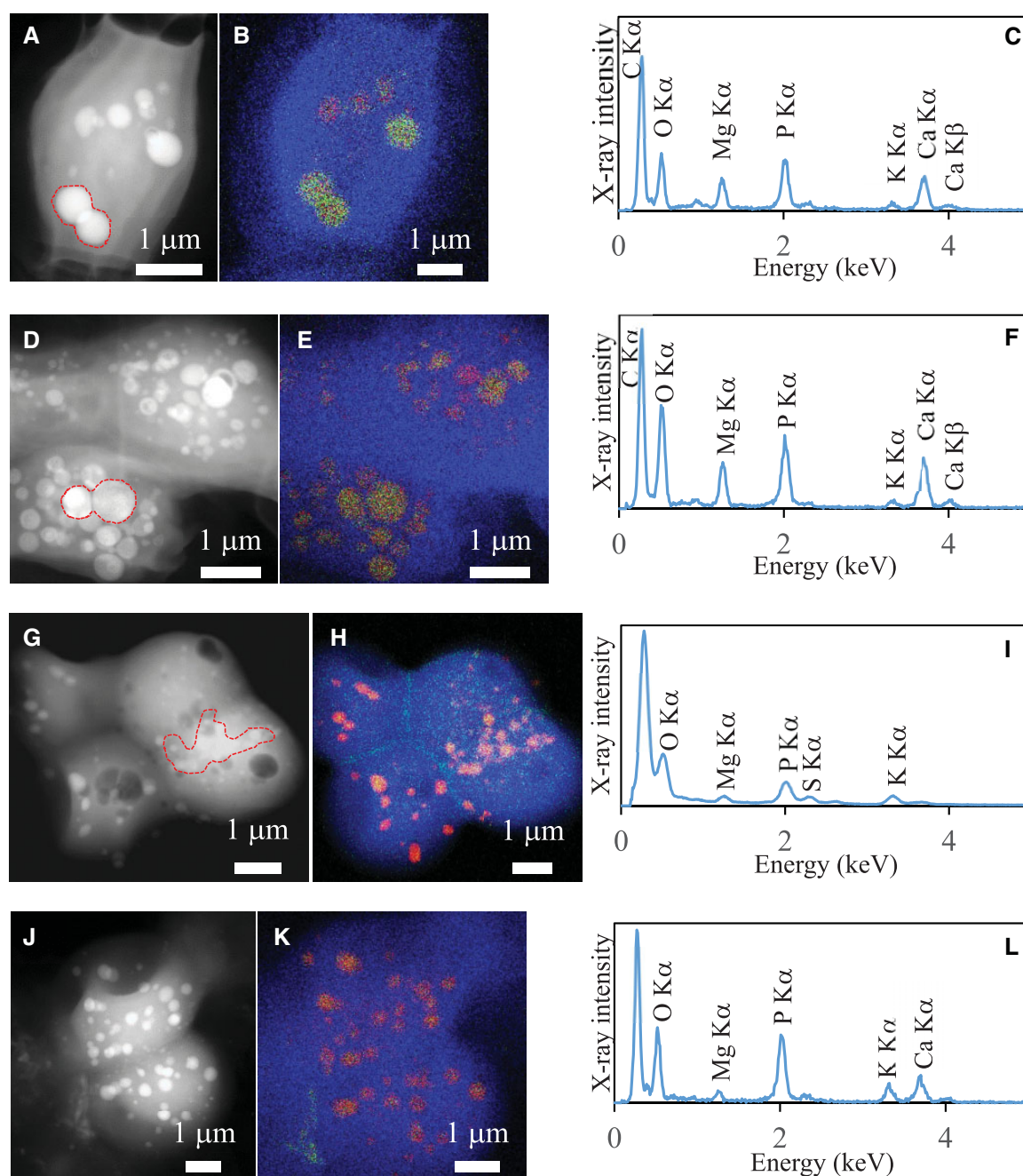


Fig. 3.—TEM analyses of the four *cyjA*-harboring strains not forming iACC. Each row corresponds to one strain. The first column shows STEM-HAADF images. The second column shows overlays of C, Ca, and P EDXS maps. The third column shows EDXS spectra of inclusions detected in the cells. (A, B, and C) *Fischerella* sp. NIES-3754. EDXS spectrum is extracted from the area indicated in (A) by a dashed line; (D, E, and F) *Chlorogloeopsis fritschii* PCC 6912. (G, H, and I) *Microcystis aeruginosa* PCC 9432; (J, K, and L) *M. aeruginosa* PCC 9717.

identities between the N-terminal domains of calycanin homologs were higher than 18%, 84%, and 82% within the W, Y, and Z groups, respectively. The Y-type N-terminal domain consisted of a duplicated small domain (measuring 66 amino acids in length, with a mean identity between the repeated domains in a same protein of 35.6%; [supplementary fig. 4, Supplementary Material online](#)), which was predicted

to contain five regular secondary structures (labeled a–e in [fig. 4](#)). As for X- and Z-type N-terminal domains, they were distinct from known protein domains, as inferred from the absence of significant similarities when searching sequence and domain databases. By contrast, significant sequence similarities were detected between the W-type N-terminal domain and three known domain families ([fig. 5](#)): 1) YAM

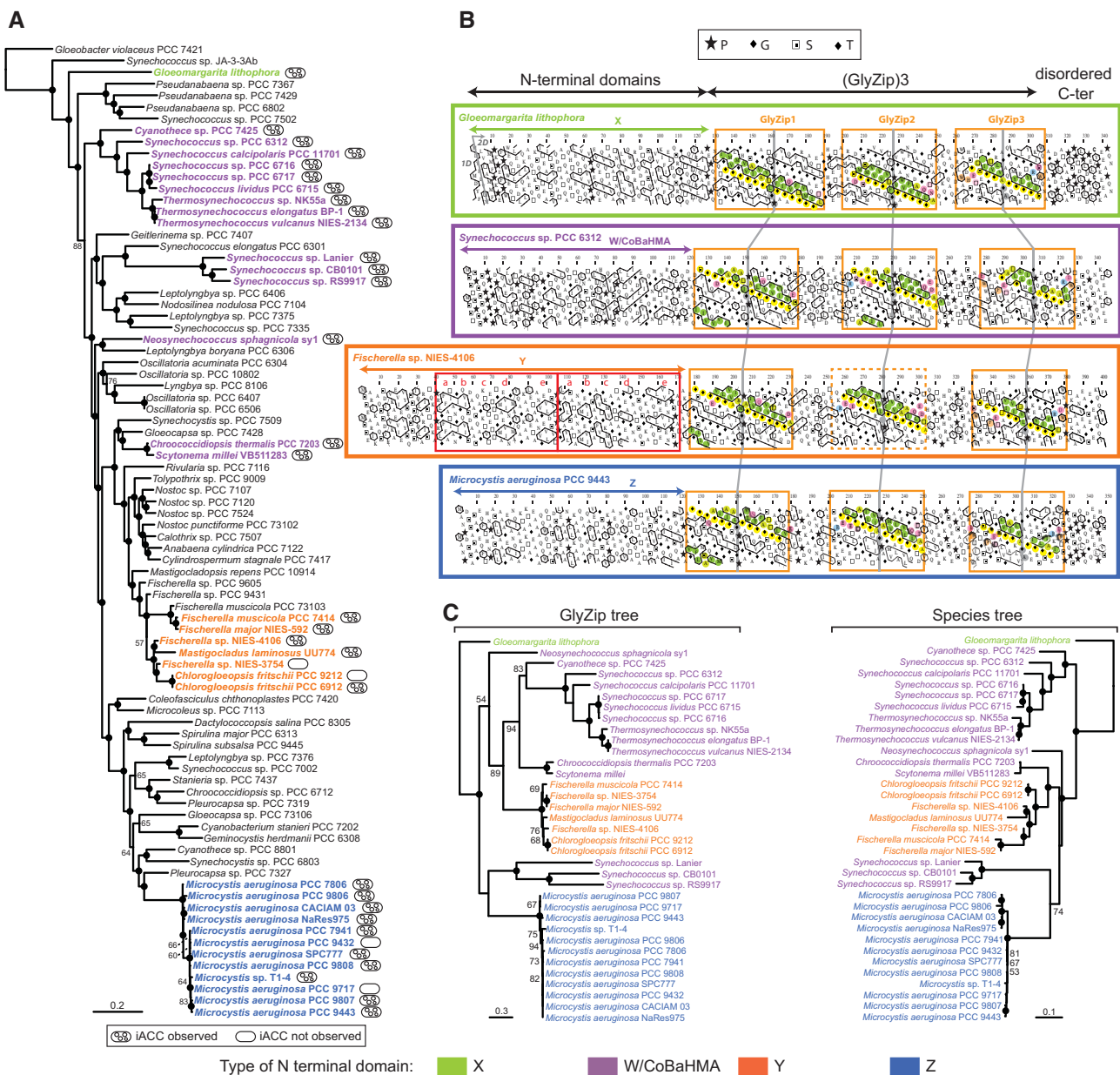


Fig. 4.—Phylogenetic analysis and domain architecture of the calcanin protein family. (A) Maximum-likelihood phylogenetic tree of Cyanobacteria based on 58 conserved proteins; the strains containing the *cyA* gene are highlighted in bold and color. (B) HCA plots of representative calcanin sequences (see fig. 1 for details of the HCA representation). The positions of the domains are indicated, with red boxes corresponding to the duplicated subdomain composing domain Y (labels a–e refer to equivalent hydrophobic clusters). The periodic patterns, made of glycine (or small amino acids—yellow) and hydrophobic amino acids (green) are highlighted for each GlyZip, with conserved signatures specific of each GlyZip shown with other colors. GlyZip2, which is present in only one species in the Y family, is indicated with a dotted box. (C) Unrooted maximum-likelihood phylogenetic tree of the GlyZip domain of calcanin (left) compared with the species tree based on 58 conserved proteins (right). Numbers on branches indicate bootstrap support (BS, only values >50% are shown), BS of 100% are indicated by black circles. The species names and HCA profiles are color-coded according to the type of N-terminal domain of calcanins (the code is shown at the bottom of the figure).

domains, found in the cytosolic C-terminus of *Escherichia coli* Major Facilitator Superfamily transporter YajR (Jiang et al. 2013, 2014); 2) heavy-metal associated (HMA) domains (also called metal binding domains) present in various proteins (e.g., P-type ATPases and metallochaperones), generally

involved in metal transport and detoxification pathways (Bull and Cox 1994); and 3) integrated HMA (iHMA) domains detected in plant immune receptors, where they are involved in fungal effector recognition (De la Concepcion et al. 2018). Similar to these three domains, the W-type N-terminal

Downloaded from https://academic.oup.com/gbe/article/14/3/evac026/6526398 by BIU Jussieu user on 22 March 2022

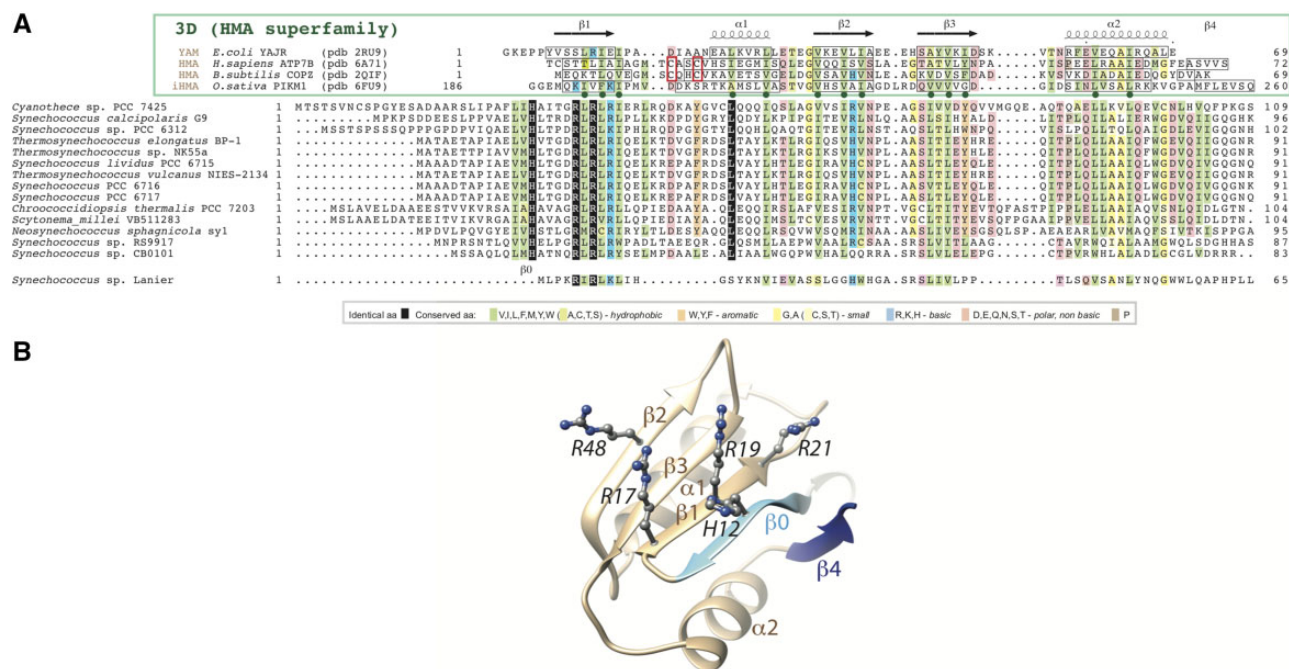


FIG. 5.—The CoBaHMA domain. (A) Multiple sequence alignment of calcyanins and members of the HMA superfamily with known 3D structures. Identical amino acids are shown in white on a black background, similarities are colored according to amino acid properties (inset). Sequences of proteins of the HMA superfamily, whose 3D structures are known and with which the CoBaHMA sequences can be aligned, are shown on top. PDB identifiers are provided. Observed 2D structures are boxed. The two cysteines of the CXXC motif specific of the HMA family are boxed in red. Green dots highlight the positions in which the hydrophobic character is strongly conserved, corresponding to amino acids participating in the hydrophobic core of the ferredoxin fold. An additional β -strand, named β_0 , is predicted in the CoBaHMA sequences, including a strictly conserved histidine. (B) Model of the CoBaHMA 3D structure, illustrated here with the *Synechococcus* sp. RS9917 sequence. The HMA common core is colored in beige, whereas specific secondary structures of the CoBaHMA family are in blue. The four highly conserved basic amino acids are shown with atomic details.

domain showed a repeated $\beta - \alpha - \beta$ motif corresponding to a ferredoxin-like fold, characteristic of the HMA superfamily (fig. 5). However, although most HMA domains possess two conserved cysteine residues directly involved in binding heavy metals, YAM, iHMA, and W-type calcyanin N-terminal domains do not conserve these amino acids (fig. 5). Moreover, the W-type domain showed a specific signature consisting of several basic amino acids distributed in strands β_1 and β_2 and a histidine located upstream of strand β_1 , in a region appearing as a calcyanin-specific extension of the HMA core (strand β_0 in fig. 5). Therefore, we named this novel domain family CoBaHMA, after *domain with Conserved Basic residues in the HMA superfamily*. A model of the CoBaHMA 3D structure was built using the experimental 3D structures of the HMA, iHMA, and YAM as templates in Modeller (Webb and Sali 2016). The position of strand β_0 was moreover putatively assigned with reference to the 3D structure of Kipl (pdb 2KWA), based on the results of HH-PRED searches and subsequent superimposition of the 3D corresponding 3D structures (pdb 2RU9 and 2KWA, root mean square value of 2.1 Å on 55 C α superimposed positions). The AlphaFold2 model (pLDDT scores above 85 from aa 7 to 81, with most of the values above 90) agreed with the first proposed model, in particular on the position of strand β_0

relative to the $\beta_1 - \beta_3$ core, but also led to propose a model for strand β_4 as well as to refine the position of amino acids within strand β_0 (fig. 5B). Although the calcyanin sequence of *Synechococcus* sp. Lanier also contained the specific signature of W-type domains with several basic amino acids, it clearly differed from the rest of the W-type N-terminal domains (fig. 5A), suggesting that calcyanin has deeply diverged in this species. Future studies should assess whether these different N-terminal domains can be found in other cyanobacterial proteins.

The C-terminal domain of the different calcyanin types consisted in three repetitions of a ~50 amino acid motif, which was largely apolar and displayed a constant periodicity in hydrophobic and small (glycine/alanine) amino acids (supplementary fig. 5, Supplementary Material online). We called this motif “GlyZip” in reference to the name proposed by Kim et al. (2005) to describe recurrent, short Gly-X(3)-Gly-X(3)-Gly motifs allowing tight packing of transmembrane helices (Senes et al. 2004). However, the calcyanin GlyZip motifs were much longer (12 basic Gly-X(3)-Gly units, interrupted in their middle by a central, highly conserved Gly-Pro dipeptide) than those already known at the 3D level, which generally contained no more than three such units (Leonov and Arkin 2005). Moreover, they did not share any obvious

Downloaded from https://academic.oup.com/gbe/article/14/3/evac026/6526398 by BIU Jussieu user on 22 March 2022

sequence similarity with known domains, suggesting that these repeated motifs form a novel architecture. The repeated presence of glycine and hydrophobic amino acids every four amino acid residues over a large sequence length, with an unusual persistence of this periodic motif across the different cyanobacterial lineages (especially for the first repeat) suggests that it may form compact and highly constrained assemblages of helices compatible with a membrane-embedded structure. These assemblages might resemble homo-oligomeric structures formed by short subunits, such as the c-rings of sodium-translocating ATP synthases (Kuehlbrandt 2019), which share similar, albeit smaller, glycine zippers. Analysis of multiple sequence alignments (supplementary fig. 5, Supplementary Material online) allowed discriminating each of the three GlyZip calcyanin motifs based on specific signatures, including the presence of aromatic and polar amino acids, outside the repeated patterns. In particular, a tryptophan and a glutamic acid were strictly conserved in the third GlyZip motif in all calcyanin sequences. The second GlyZip motif of several calcyanin sequences matched part of a family model called PdsO (sortase-associated OmpA-like protein), found in, for example, *Shewanella oneidensis* (see NCBI Conserved Domain Database [CDD] annotations in supplementary table 2b, Supplementary Material online). The matching region, located before the OmpA-like C terminal domain, shows the typical features of a GlyZip unit (supplementary fig. 6, Supplementary Material online) and is present as a single copy in PdsO, suggesting that this basic unit evolved within calcyanin by triplication and enrichment in polar amino acids (see below). Last, among Y-type calcyanins, only that of *Fischerella* sp. NIES-4106 possessed all three GlyZip motifs. By contrast, all other Y-type calcyanins, including those found in iACC-forming strains, contained only the first and third GlyZip motifs. This suggests that calcyanins with only two GlyZip motifs remain functional (supplementary fig. 5, Supplementary Material online). Interestingly, although it did not match the characteristic GlyZip profile, the duplicated domain found in the N-terminal region of Y-type calcyanins was also largely apolar and rich in small amino acids so as in GlyZip motifs.

Calcyanin May Be Involved in Ca Homeostasis

In *C. fritschii* PCC 9212 and PCC 6912, the genes located directly upstream and downstream of *ccyA* were annotated as encoding a Ca(2+)/H(+) antiporter and a Na(+)-dependent bicarbonate transporter BicA, respectively (supplementary table 4, Supplementary Material online). This is particularly interesting because bicarbonate and calcium are obvious crucial ingredients for the synthesis of CaCO₃. Moreover, these two transporter genes are located on the same DNA strand as *ccyA* and may therefore be transcribed simultaneously with *ccyA* in a single mRNA, although this will have to be tested by future studies. By searching homologs of

these two transporters in our complete data set of 602 cyanobacterial genomes (i.e., the genomes of the 8 iACC-forming strains described by Benzerara et al. 2014 plus the 594 genomes in which we searched for new *ccyA* homologs), we observed that their combined presence was significantly associated with that of *ccyA* (χ^2 test, *P* value = 1.4e-08; supplementary table 5, Supplementary Material online). Indeed, all 35 genomes harboring *ccyA* had at least one copy of both genes, except *Synechococcus* sp. Lanier, which lacked BicA. The latter strain was also deviant from other *ccyA*-harboring strains based on very atypical N-terminal and C-terminal calcyanin sequences. Because this strain was not available for EM analysis, we could not test if it contained iACC or not. By contrast, among the 567 genomes lacking *ccyA*, only 293 contained both transporter genes. Interestingly, in *Fischerella* sp. NIES-4106 megaplasmid, *ccyA* was located downstream a calcium/proton exchanger (sharing 92.9% identity with the above-mentioned antiporter of *C. fritschii* PCC 9212 and PCC 6912), in a region containing several additional genes potentially involved in biomineralization, such as two cation-transporting ATPases and a carbonic anhydrase (supplementary fig. 7, Supplementary Material online). Overall, the correlation and/or colocalization of *ccyA* and genes involved in Ca or HCO₃⁻ transport and homeostasis supports the hypothesis of a functional role of calcyanin in Ca-carbonate biomineralization.

Attempts to obtain *ccyA* deletion mutants in the iACC-forming strains *Synechococcus* sp. PCC 6312 were unsuccessful but there is no certainty at this point that the employed technique can generate deletion mutants in this strain. Some possibilities to be further explored in the future are that *ccyA* deletion is lethal and/or increase the sensitivity to toxicity by calcium, suggesting that this gene may carry out an essential function in these cyanobacteria. In the absence of a direct loss-of-function genetic analysis, we overexpressed the *ccyA* genes of the two evolutionary distant cyanobacteria *Synechococcus* sp. PCC 6312 and *G. lithophora* in the non-iACC-forming, but genetically manipulable host, *Synechococcus elongatus* PCC 7942, which does not originally contain *ccyA*. Investigation by EM-associated elemental chemical analyses of *S. elongatus* PCC 7942 cells overexpressing these *ccyA* genes did not show the presence of typical iACC (i.e., inclusions with Ca only and little to no P), whereas polyphosphate inclusions were found in cells of all mutants (fig. 6 and supplementary fig. 8, Supplementary Material online). However, the comparison of Ca chemical maps of *S. elongatus* PCC 7942 mutants harboring the empty plasmid (pC) and mutants harboring its derivative expressing the *ccyA* genes (pC-*ccyA*_{G10e0} and pC-*ccyA*_{S6312}) showed differences. No Ca hotspot was observed in cells with the empty plasmid (pC) sampled at two different growth stages, over a total of 135 counted polyphosphate inclusions. By contrast, 23 (pC-*ccyA*_{G10e0}) and 10 (pC-*ccyA*_{S6312}) Ca hotspots were

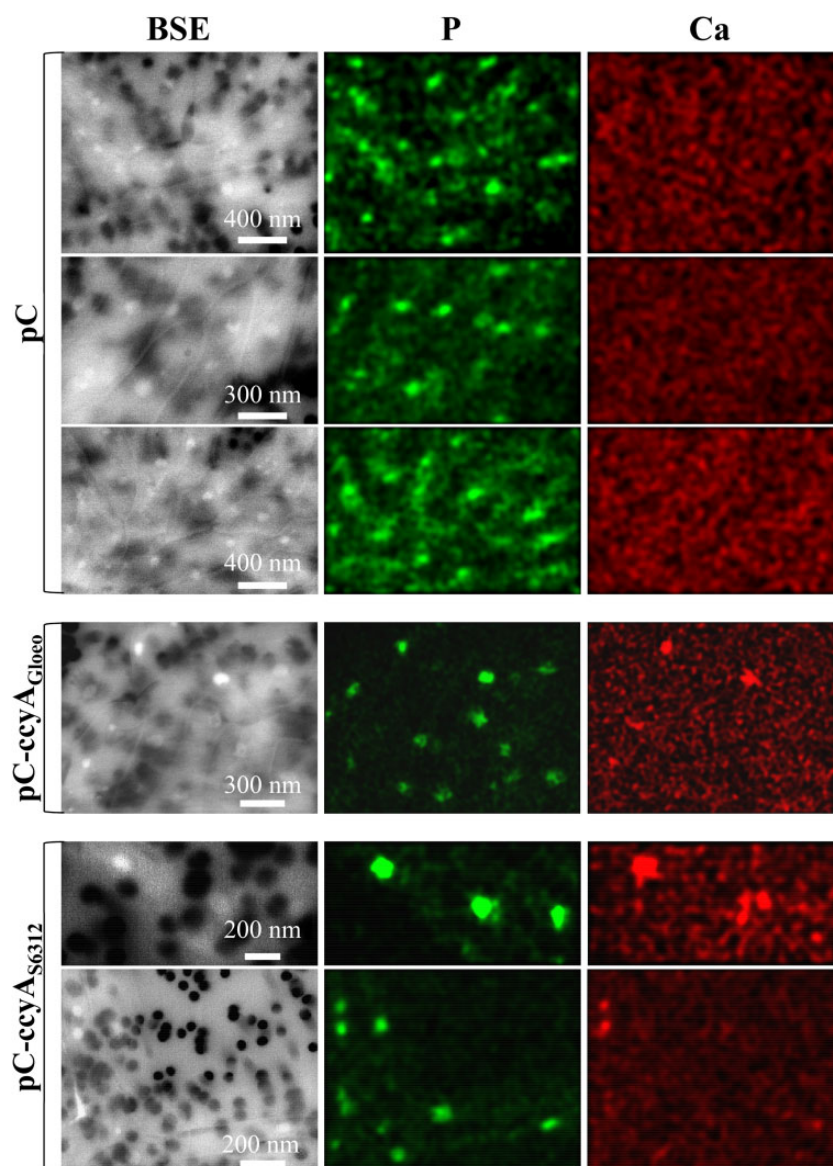


Fig. 6.—SEM analyses of mutants overexpressing *ccyA*. SEM-EDXS images (in BSE mode), P (green), and Ca (red) maps of *Synechococcus elongatus* PCC 7942 mutants. The scale bar provided on the BSE images is the same for the corresponding P and Ca maps on each row. The 0.2 μm pores of the filters appear as dark disks in the BSE images. At the accelerating voltage used for these analyses, *S. elongatus* cells appear as relatively transparent, packed rods. Polyphosphate inclusions appear as brighter dots. The first three rows show cells of a *S. elongatus* PCC 7942 mutant harboring the empty pC plasmid. No Ca-rich inclusions are observed in these cells as shown by the homogeneous background in the Ca maps. In contrast, Ca-rich inclusions (polyphosphates) are observed in cells of *S. elongatus* PCC 7942 mutants harboring the plasmids pC-*ccyA*_{Gloeo} (fourth row) or pC-*ccyA*_{S6312} (fifth and sixth rows), appearing as hotspots in Ca maps. See [supplementary data 3, Supplementary Material online](#) for details concerning the plasmid and strains.

detected over a total of 117 and 90 polyphosphate inclusions observed in the pC-*ccyA*_{Gloeo} and pC-*ccyA*_{S6312} mutants, respectively (fig. 6). The Ca detection limit of scanning electron microscopy (SEM)-EDXS is not precisely known and we likely overlook some Ca. Future studies using more sensitive spatially resolved techniques will be required to have more quantitative assessment of the Ca enrichment in these cells. However, these results suggest that higher amounts of Ca were sequestered within

polyphosphate inclusions when *ccyA* was present and that this gene may be functionally involved in Ca homeostasis, via a molecular process that remains to be fully elucidated.

Phylogenetic Distribution and Evolution of Calcyanin

Whatever the function of this diagnostic gene family, constructing its phylogeny allows to infer the possible

evolutionary history of iACC biomineralization. We placed the species containing the *ccyA* gene on a general phylogeny of cyanobacteria constructed using 58 conserved proteins (supplementary table 6, Supplementary Material online). The four calcanin types were found in various lineages widely dispersed across this cyanobacterial tree (fig. 4A). Whereas the X, Y, and Z types showed a distribution restricted to some particular clades (*Gloeomargarita*, *Fischerella* and closely related genera, and *Microcystis*, respectively), the CoBaHMA domain (i.e., W-type) was found in several distantly related branches (fig. 4A). Similarly, *ccyA* was detected in all the species of some clades (e.g., the *Cyanothece-Synechococcus-Thermosynechococcus* clade), suggesting that it already existed in the genome of their last common ancestor, whereas it is missing in some species of other clades such as the *Chlorogloeopsis-Fischerella* one, suggesting several independent losses and/or HGT events. To better characterize these evolutionary processes, we reconstructed the phylogeny of calcanin using the conserved GlyZip domain sequences and compared it with the corresponding cyanobacterial species tree (fig. 4C). Despite a weaker resolution of the deep branches, reflecting the higher sequence variability of calcanin, we retrieved the monophyly of most of the groups as found in the species tree (fig. 4C), supporting the idea that *ccyA* was ancestral in these groups, and that the *ccyA*-lacking species most likely lost it secondarily. To further compare the two trees, we carried out an approximately unbiased (AU) test (Shimodaira 2002). Whereas the species tree topology was not rejected by the GlyZip data set (P value = 0.64), the GlyZip topology was strongly rejected by the data set of conserved proteins used to build the species tree (P value = 0.00198). This strongly suggests that the differences between both trees were due to the smaller amount of phylogenetic signal contained in the GlyZip sequences compared with the set of conserved proteins and that the GlyZip sequences have evolved following the species evolution.

The overall congruence between the two trees, both retrieving the monophyly of several large cyanobacterial clades (fig. 4C), supports a very ancient origin of *ccyA* in cyanobacteria, with independent losses in various lineages. The alternative scenario of a more recent origin of *ccyA* in one group followed by its transfer to the rest by HGT was unlikely given the congruence of both trees and the extreme divergence of the N-terminal domains among the different types of calcanin (fig. 4B). Because of its larger phylogenetic distribution, the CoBaHMA-type seemed to be the most ancient calcanin version, whereas the Y- and Z-types have likely evolved in cyanobacterial groups that diverged more recently. The situation is less clear for the X-type due to its exclusive presence in *G. lithophora*, the so far single representative species of the poorly known *Gloeomargaritales*. As mentioned above, the N-terminal domains of these four types of calcanin did not share any apparent sequence similarity (fig. 4B). This could reflect either an extreme divergence from a common

ancestral domain, potentially following the adaptation of the species to their habitat needs, or the independent recruitment of nonhomologous domains generating the different calcanins by their fusion to the conserved GlyZip C-terminal domain.

To investigate if calcanin might have originated before the diversification of cyanobacteria, we used our HMM profile to search for the GlyZip domain in other sequences present in the NCBI nonredundant database. We found homologs with a complete C-terminal domain in only five noncyanobacterial species: an uncultured candidate phyla radiation *Gracilibacteria* genome and four gammaproteobacteria of the *Methylococcales* order. We included these new sequences in a phylogenetic analysis of the GlyZip domain and found that they did not form a monophyletic group but branched intermixed with the cyanobacterial sequences (supplementary fig. 9, Supplementary Material online). On the one hand, the *Gracilibacteria* sequence was very close to the *Fischerella-Chlorogloeopsis* group and, in agreement with this similarity of the GlyZip domain, it also contained the typical Y-type N-terminal domain found in these cyanobacteria. On the other hand, the *Methylococcales* sequences branched close to the *Microcystis* group and, consistently, their N-terminal domains showed some similarity with the Z-type domains of the *Microcystis* sequences. This phylogeny and the extremely sparse distribution of *ccyA* outside the Cyanobacteria phylum suggest that these few noncyanobacterial species acquired their *ccyA* genes by HGT from *Fischerella*- and *Microcystis*-like donors, respectively. It will be interesting in future work to investigate the possible presence of iACC inclusions in these bacteria.

Conclusions

Here we show that the newly identified *ccyA* gene family, belonging to the genomic “dark matter” (i.e., unclassified or poorly understood genetic material) of cyanobacteria, can be used as a diagnostic iACC biomineralization marker. The *ccyA*-encoded calcanin protein has a unique architecture composed of highly divergent N-terminal domains fused with a novel, much more conserved GlyZip-containing C-terminal domain, which may adopt an original, not yet described fold. Among the diverse N-terminal domains of calcanin that we have identified here, the domain family that we named CoBaHMA is found in the most widespread, and likely most ancient, calcanin version. This domain family likely supports an as-yet undisclosed function within the HMA superfamily, associated with a patch of conserved basic amino acids. By tracking this gene in available genome databases, we uncovered a diversity of *ccyA*-bearing cyanobacteria capable of iACC biomineralization that is phylogenetically and environmentally much broader than previously thought, supporting a potential environmental significance. Moreover, the distribution and phylogeny of *ccyA* suggest that iACC

biomineralization is ancient, with independent losses in various lineages. Additional genes are likely involved in iACC formation but, unlike *ccyA*, they may not be specific to this function and/or they are not shared by all iACC-forming cyanobacteria. The specific distribution of *ccyA* in iACC-forming cyanobacteria, its correlated presence with bicarbonate and calcium transporters, and genetic analyses, all support a pivotal role of *ccyA* in iACC biomineralization. Further investigations are required to determine whether this function may involve the conserved glutamic acid residues of the C-terminal domain, reminding Glu-rich proteins involved in ACC biomineralization (Aizenberg et al. 2002), or the basic amino acids in the N-terminal domain, which may stabilize dense liquid phases of CaCO_3 and delay the formation of ACC (Finney et al. 2020). Alternatively, calcyanin may have a more indirect role in iACC biomineralization serving as a cation transporter or a signaling molecule. In any case, iACC biomineralization clearly appears as an original case of controlled biomineralization in bacteria.

Materials and Methods

Identification of Candidate iACC-Specific Orthologous Groups

In a first step, the 56 genomic assemblies used to identify groups of orthologous genes specific to iACC-forming cyanobacteria (supplementary table 1, Supplementary Material online) were retrieved from the NCBI database. The 523,680 translated CDSs derived from these genomes were processed using OrthoMCL with default settings (Li et al. 2003). This analysis included an all-versus-all BLASTp routine (E -value $< 1e-05$) and a clustering procedure into orthologous groups using the MCL algorithm.

Iterative Search for Homologs of Calcyanin in Cyanobacterial Genomes

Homologs of calcyanin were searched based on similarities of the conserved C-terminal domain in 594 available genomes of cyanobacteria using an iterative process. This search data set corresponded to the NCBI genome assemblies assigned to Cyanobacteria, published online before December 1, 2017 (except the six identified in the first step, see above). For each genome assembly, we iteratively searched for homologs of calcyanin in the first set of amino acid sequences available in the following ordered list (supplementary data 2, Supplementary Material online): 1) translated CDS or 2) proteins in RefSeq annotation records, 3) translated CDS or 4) proteins in GenBank annotation records.

A multiple sequence alignment of the conserved C-terminal domain was built for the six calcyanin sequences identified in the first step (see above), using MAFFT (Katoh and Standley 2013). A HMM profile was generated based on this alignment with the program hmmbuild from the HMMER package

(version 3.3) (Eddy 2011). The options *wblossum* with *wid* 0.62 were used to downweight closely related sequences and *upweight* distantly related ones. To avoid biases toward glycine-rich unrelated proteins, we artificially reduced glycine weight by 20% in the profiles. The profile versus sequence similarity search was done with the program *hmmsearch* (E -value $< 1.0e-70$). The hits matching 100% of the profile length and corresponding to newly identified sequences were added to the new calcyanin data set. The multiple sequence alignment and the HMM profile of this data set were then updated. These steps (alignment, building of HMM-profile, similarity search) were repeated until no new sequence was detected. In order to detect remote homologs of calcyanin, seven iterations of the entire process were done as described in supplementary table 7, Supplementary Material online, with a progressive decrease of the stringency of the similarity search. In the beginning, we set a very low E -value and high cover to the profile. As the iterations proceeded, we increased the E -value and decreased the cover to the profile down to 70%. This cover threshold higher than 66% was designed to avoid (Gly)₂ (instead of (Gly)₃) to be matched. At the end of the whole process, we used the final HMM profile (as provided in supplementary data 3, Supplementary Material online) to search for similarities in the GenBank records of the processed genomic assemblies.

Last, *ccyA* was searched in the newly sequenced genomes of *Synechococcus* sp. PCC 6716 and PCC 6717 using tBLASTn with all previously identified *ccyA* sequences as queries. The CDS boundaries of the best BLAST hits were further assessed using Prodigal (Hyatt et al. 2010).

Comparative Genomics of *C. fritschii* PCC 6912 (No iACC Observed) and *C. fritschii* PCC 9212 (with iACC)

The search of homologous genes shared by *C. fritschii* PCC 6912 and PCC 9212 genomes was achieved based on unidirectional BLASTp best hits as implemented in the PATRIC proteome comparison tool (Gillespie et al. 2011) (E -value $< 1.0e-05$, sequence coverage $> 30\%$). For each genome assembly, we used the set of translated CDS as provided in the RefSeq annotation record. Gene functional categories were searched in COG database (v1) using CD-search (Marchler-Bauer and Bryant 2004) (E -value $< 1.0e-05$). The nucleotide sequences of the *ccyA*-containing contigs of *C. fritschii* PCC 6912 and PCC 9212 (NCBI accessions NZ_AJLN01000033.1 and NZ_AJLM01000017.1, 97,542 and 97,528 bp, respectively) were compared using BLASTn.

Search for Homologs of the $\text{Ca}(2+)/\text{H}(+)$ Antiporter and $\text{Na}(+)$ -Dependent Bicarbonate Transporter in Cyanobacterial Genomes

Homologs of the $\text{Ca}(2+)/\text{H}(+)$ antiporter and the $\text{Na}(+)$ -dependent bicarbonate transporter *BicA*, encoded in *C. fritschii* PCC 6912 and PCC 9212 by the genes located upstream and

downstream of *ccyA*, respectively, were searched using BLASTp (E -value $< 1.0e-10$) in our complete data set of 602 cyanobacterial genomes (composed of the 8 iACC-forming strains described by Benzerara et al. [2014] in which we initially detected *ccyA* and by the 594 genomes in which we iteratively searched for new *ccyA* homologs in a second step). Owing to the incompleteness of the *ccyA*-upstream gene in the genomic sequence of these two strains, we used the most similar full-length sequence as Ca(2+)/H(+) antiporter query (96% identity; accession WP_016868870.1, *Fischerella muscicola* PCC 7414). BicA homologs were identified using the protein sequence from *C. fritschii* PCC 6912 and PCC 9212 as query (accession WP_016872894.1).

Calcyanin Functional Annotation and Structure Prediction

The structural features of calcyanin were explored based on the information provided by amino acid sequences using HCA (Callebaut et al. 1997; Bitard-Feildel et al. 2018). HCA provides a global view of the protein texture, with insights into the structural features of foldable regions (Bitard-Feildel et al. 2018). Similarities between domains composing calcyanin and known domains/3D structures were searched against different databases (NCBI nr sequence database, NCBI Conserved Domain Database [CDD] [Yang et al. 2020], and the Protein Data Bank [PDB]) using tools for profile-sequence and profile-profile comparison such as PSI-BLAST (Altschul et al. 1997) and HH-PRED (Zimmermann et al. 2018), respectively.

Three-dimensional structure modeling was performed using Modeller 9.23 (Webb and Sali 2016). This modeling was refined afterwards using AlphaFold (Jumper et al. 2021), through the notebook AlphaFold2_advanced from Colabfold (Mirdita et al. 2021) (<https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb>). The full sequence of *Synechococcus* sp. RS9917 CcyA and the multiple sequence alignment of the 15 reported CoBaHMA-bearing CcyA were used as input. Three-dimensional structures were visualized using UCSF Chimera (Pettersen et al. 2004). Multiple sequence alignment handling and rendering were made using SeaView (Gouy et al. 2010) and EsPript (Robert and Gouet 2014), respectively.

Molecular Phylogenetic Analyses

Phylogeny of cyanobacteria using different sets of species was reconstructed using 58 conserved proteins (Moreira et al. 2017; supplementary table 6, Supplementary Material online). Each individual protein was aligned using MAFFT with the accurate L-INS-I option (Katoh and Standley 2013) and poorly aligned regions were removed with trimAl -automated1 (Capella-Gutiérrez et al. 2009). Trimmed alignments were concatenated to produce a supermatrix and maximum-likelihood phylogenetic trees were reconstructed with the program IQ-Tree using the mixture model LG+C60+F+G

(Nguyen et al. 2015). Statistical support was estimated using 1,000 bootstrap replicates. The phylogeny of calcyanin was studied using the manually curated alignment of the conserved GlyZip C-terminal domain. A maximum-likelihood tree was constructed with the program IQ-Tree using the mixture model LG+C20+F+G (Nguyen et al. 2015). Statistical support was estimated using 1,000 bootstrap replicates.

Tree topologies based in the set of 58 conserved proteins (species tree) and in the GlyZip C-terminal domain of calcyanin were compared using the AU test (Shimodaira 2002) implemented in IQ-TREE with the options -n 0 -zb 10,000 -au -zw (Nguyen et al. 2015). The sequence evolution models used were, as before, LG+C60+F+G for the conserved protein data set and LG+C20+F+G for the GlyZip data set.

Electron Microscopy Analyses of iACC

Strains recovered from culture collections were analyzed by scanning transmission electron microscopy (STEM) for iACC search. As previously shown by Benzerara et al (2014), Li et al. (2016), and De Wever et al. (2019), iACC can be recognized based on the fact that they mostly contain Ca with little to no P, whereas polyphosphate inclusions show a major P peak with Mg, K, and/or Ca. For that purpose, we used a field emission gun JEOL-2100F microscope operating at 200 kV, equipped with a JEOL detector with an ultrathin window allowing detection of light elements. STEM allowed Z-contrast imaging in the high angle annular dark field (HAADF) mode. EDXS analyses rely on the detection of X-rays emitted by samples excited by the electron beam. Their energy is characteristic of the atoms and their intensity depends on the atomic content. Compositional maps of Ca, P, and C were acquired by performing EDXS analysis in the STEM HAADF mode. These EDXS analyses provide hyperspectral data, that is, an image with EDXS spectra for each pixel of the image. For these analyses, a total of 0.5 mL of cultures aged between 5 and 30 days was centrifuged at $8,000 \times g$ for 10 min. Pellets were rinsed three times in Milli-Q (mQ) water (Millipore). After the final centrifugation, pellets were suspended in 200 μ L of mQ water. A drop of 5 μ L was deposited on a glow discharged carbon-coated 200-mesh copper grid and let dry at ambient temperature.

For iACC-forming strains, we systematically measured several replicates by STEM and/or EDXS associated with SEM. Even more effort was invested in the analysis of strains harboring *ccyA* but not showing iACC. Indeed, although showing the presence of iACC in a strain only requires one single positive observation, concluding about the absence of iACC is difficult, if not impossible. For *Fischerella* sp. NIES-3754, we performed seven different SEM or STEM sessions over four different cultures, including two on the same culture with a 15 days interval and three on a second culture with 3 and 9 days interval. For *C. fritschii* PCC 6912, we performed eight different SEM or STEM sessions over five different cultures,

including three on the same culture with a 4 and 6 days interval and two on another culture with a 15 days interval. For *M. aeruginosa* PCC 9432, we performed seven different SEM or STEM sessions over four different cultures, including two on the same culture with a 3 days interval and three on another culture with 3 and 8 days interval. For *M. aeruginosa* PCC 9717, we performed six different SEM or STEM sessions over four different cultures, including two on the same culture with a 3 days interval and two on another culture with 25 days interval.

Mutant strains of *S. elongatus* PCC 7942 harboring the pC, pC-ccyA_{Gloe}, or ccyA_{S6312} were analyzed by SEM in the back-scattered electron (BSE) mode, coupled with EDXS analyses to search for Ca enrichment. Analyses were replicated twice on at least three and up to six areas. Ca hotspots were identified each time and the signal in the Ca energy range was higher than the background by 1 σ . One example of EDXS spectrum is provided per type of mutant in [supplementary figure 8](#), [Supplementary Material online](#).

Genetics

The pC-ccyA_{Gloe} and pC-ccyA_{S6312} plasmids were derivatives of the RSF1010-derived pC vector (Veaudor et al. 2018) replicating in *E. coli* ([supplementary table 8 and fig. 10](#), [Supplementary Material online](#)). Chenebault et al. (2020) showed that this expression plasmid allowed strong gene expression in cyanobacteria. The pC-ccyA_{Gloe} and pC-ccyA_{S6312} plasmids were transferred to *S. elongatus* PCC 7942 by trans-conjugation (Mermet-Bouvier and Chauvat 1994), using the improved triparental-mating protocol that follows. Overnight-grown cultures of the *E. coli* strains CM404, which propagates the self-transferable mobilization vector pRK2013, and TOP10, which propagates either pC, pC-ccyA_{Gloe}, or pC-ccyA_{S6312}, were washed twice and resuspended in LB medium (1×10^9 cells.mL⁻¹). Meanwhile, *S. elongatus* PCC 7942 mid-log phase cultures grown in mineral growth medium (MM, a version of BG-11 supplemented with 3.78 mM Na₂CO₃) were centrifuged and concentrated five times (about 1×10^8 cells/mL) in fresh MM. Then, 100 μ L of *S. elongatus* PCC 7942 cells were mixed with 30 μ L of CM404 cells and 30 μ L of TOP10 cells harboring either pC, pC-ccyA_{Gloe}, or pC-ccyA_{S6312}. A total of 30 μ L aliquots of this mixture were spotted onto MM solidified with 1% agar (Difco), and incubated for 48 h under standard temperature (30 °C) and light (2,500 lux, i.e., 31 μ E.m⁻².s⁻¹) conditions. Then, cells were collected from each plate and resuspended into 50 μ L of liquid MM, prior to plating onto MM containing 5 μ g.mL⁻¹ of each the streptomycin (Sm) and spectinomycin (Sp) selective antibiotics. After about 10 days of incubation under standard light and temperature conditions, Sm^RSp^R-resistant conjugant clones were collected and restreaked onto selective plates, prior to analyzing their plasmid content by PCR and DNA sequencing (Eurofins Genomics) using specific primers

([supplementary data 1c and 1d and fig. 10](#), [Supplementary Material online](#)).

Supplementary Material

[Supplementary data](#) are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank two anonymous reviewers for their constructive comments which improved the overall quality of the manuscript. We thank Alexis De Wever and Marine Blondeau for helping acquiring some TEM data. We thank Mélanie Poinset for helping in the preparation of some samples for transmission electron microscopy analyses. We thank Eva Jahodarova for shipping a culture of *Neosynechococcus sphagnicola*. This work was supported by the Agence Nationale de la Recherche (ANR Harley, ANR-19-CE44-0017-01; ANR PHOSTORE, ANR-19-CE01-0005) and the European Research Council under the European Union's Seven Framework Program: ERC grants Calcyan (PI: K. Benzerara, grant agreement no. 307110) and PlastEvol (PI: D. Moreira, grant agreement no. 787904). Sigrid Görden PhD grant was funded by the Sorbonne Université doctoral program Interfaces pour le Vivant.

Author Contributions

K.B.E., E.D.U., C.C.C., F.C.H., D.M.O., P.L.G., and I.C.A. conceived and designed the work. K.B.E., E.D.U., T.B.F., G.C.A., C.C.C., M.D.E., I.D.I., G.G.A., M.G.U., S.G.O., F.S.P., D.M.O., and I.C.A. acquired, analyzed, and/or interpreted data. K.B.E., E.D.U., C.C.C., F.C.H., M.G.U., P.L.G., D.M.O., and I.C.A. drafted the work or substantively revised it.

Data Availability

Further information and requests for resources, codes and reagents should be directed to and will be fulfilled by the lead contact, Karim Benzerara (karim.benzerara@upmc.fr).

- Plasmids and mutant strains generated in this study are available upon request to the lead contact.
- The genomic assemblies are available at GenBank as follows:

Synechococcus calcipolaris PCC 11701—BioProject PRJNA800269

Synechococcus sp. PCC 6716—BioProject PRJNA801107

Synechococcus sp. PCC 6717—BioProject PRJNA801158

Accession numbers are listed in the key resources table.

- TEM-EDXS and SEM-EDXS data and the structure of the CoBaHMA domain have been uploaded to Zenodo: 10.5281/zenodo.5964253. DOIs will be listed in the key resources table.

Literature Cited

- Aizenberg J, Lambert G, Weiner S, Addadi L. 2002. Factors involved in the formation of amorphous and crystalline calcium carbonate: a study of an ascidian skeleton. *J Am Chem Soc.* 124(1):32–39.
- Altermann W, Kazmierczak J, Oren A, Wright DT. 2006. Cyanobacterial calcification and its rock-building potential during 3.5 billion years of Earth history. *Geobiology* 4(3):147–166.
- Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25(17):3389–3402.
- Benzerara K, et al. 2014. Intracellular Ca-carbonate biomineralization is widespread in cyanobacteria. *Proc Natl Acad Sci U S A.* 111(30):10933–10938.
- Bitard-Feildel T, Lamiabie A, Mornon J-P, Callebaut I. 2018. Order in disorder as observed by the “hydrophobic cluster analysis” of protein sequences. *Proteomics* 18(21–22):e1800054.
- Blondeau M, Benzerara K, et al. 2018. Impact of the cyanobacterium *Gloeomargarita lithophora* on the geochemical cycles of Sr and Ba. *Chem Geol.* 483:88–97.
- Blondeau M, Sachse M, et al. 2018. Amorphous calcium carbonate granules form within an intracellular compartment in calcifying cyanobacteria. *Front Microbiol.* 9:1768.
- Blue CR, et al. 2017. Chemical and physical controls on the transformation of amorphous calcium carbonate into crystalline CaCO₃ polymorphs. *Geochim Cosmochim Acta.* 196:179–196.
- Bradley JA, et al. 2017. Carbonate-rich dendrolitic cones: insights into a modern analog for incipient microbialite formation, Little Hot Creek, Long Valley Caldera, California. *NPJ Biofilms Microbiomes.* 3:32.
- Bull PC, Cox DW. 1994. Wilson disease and Menkes disease: new handles on heavy-metal transport. *Trends Genet.* 10(7):246–252.
- Callebaut I, et al. 1997. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cell Mol Life Sci.* 53(8):621–645.
- Cam N, et al. 2016. Selective uptake of alkaline earth metals by cyanobacteria forming intracellular carbonates. *Environ Sci Technol.* 50(21):11654–11662.
- Cam N, et al. 2018. Cyanobacterial formation of intracellular Ca-carbonates in undersaturated solutions. *Geobiology* 16(1):49–61.
- Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15):1972–1973.
- Couradeau E, Benzerara K, Gerard E, Moreira D, Bernard S, Brown GE, Lopez-Garcia P. 2012. An early-branching microbialite cyanobacterium forms intracellular carbonates. *Science* 336:459–462.
- De la Concepcion JC, et al. 2018. Polymorphic residues in rice NLRs expand binding and response to effectors of the blast pathogen. *Nat Plants.* 4(8):576–585.
- De Wever A, et al. 2019. Evidence of high Ca uptake by cyanobacteria forming intracellular CaCO₃ and impact on their growth. *Geobiology* 17(6):676–690.
- Eddy SR. 2011. Accelerated profile HMM searches. *PLoS Comput Biol.* 7(10):e1002195.
- Finney AR, Innocenti Malini R, Freeman CL, Harding JH. 2020. Amino acid and oligopeptide effects on calcium carbonate solutions. *Cryst Growth Des.* 20(5):3077–3092.
- Frangeul L, et al. 2008. Highly plastic genome of *Microcystis aeruginosa* PCC 7806, a ubiquitous toxic freshwater cyanobacterium. *BMC Genomics.* 9(274):274.
- Gillespie JJ, et al. 2011. PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infect Immun.* 79(11):4286–4298.
- Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol.* 27(2):221–224.
- Humbert J-F, et al. 2013. A tribute to disorder in the genome of the bloom-forming freshwater cyanobacterium *Microcystis aeruginosa*. *PLoS One.* 8(8):e70747.
- Hyatt D, et al. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics.* 11:119.
- Jiang D, et al. 2013. Structure of the YajR transporter suggests a transport mechanism based on the conserved motif A. *Proc Natl Acad Sci U S A.* 110(36):14664–14669.
- Jiang D, et al. 2014. Atomic resolution structure of the *E. coli* YajR transporter YAM domain. *Biochem Biophys Res Commun.* 450(2):929–935.
- Jumper J, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596(7873):583–589.
- Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30(4):772–780.
- Kim S, et al. 2005. Transmembrane glycine zippers: physiological and pathological roles in membrane proteins. *Proc Natl Acad Sci U S A.* 102(40):14278–14283.
- Komarek J, Johansen JR, Smarda J, Strunecky O. 2020. Phylogeny and taxonomy of *Synechococcus*-like cyanobacteria. *Fottea* 20(2):171–191.
- Kuehlbrandt W. 2019. Structure and mechanisms of F-type ATP synthases. *Annu Rev Biochem.* 88: 515–549.
- Lamiabie A, et al. 2019. A topology-based investigation of protein interaction sites using hydrophobic cluster analysis. *Biochimie* 167:68–80.
- Latour D, Salençon M-J, Reyss J-L, Giraudet H. 2007. Sedimentary imprint of *Microcystis aeruginosa* (cyanobacteria) blooms in grangent reservoir (Loire, France). *J Phycol.* 43(3):417–425.
- Lefevre CT, Bazylinski DA. 2013. Ecology, diversity, and evolution of magnetotactic bacteria. *Microbiol Mol Biol Rev.* 77(3):497–526.
- Leonov H, Arkin IT. 2005. A periodicity analysis of transmembrane helices. *Bioinformatics* 21(11):2604–2610.
- Li J, et al. 2016. Biomineralization patterns of intracellular carbonatogenesis in cyanobacteria: molecular hypotheses. *Minerals* 6(1):10.
- Li L, Stoekert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13(9):2178–2189.
- Marchler-Bauer A, Bryant SH. 2004. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.* 32(Web Server Issue):W327–W331.
- Marron AO, et al. 2016. The evolution of silicon transport in eukaryotes. *Mol Biol Evol.* 33(12):3226–3248.
- Mehta N, Benzerara K, Kocar BD, Chapon V. 2019. Sequestration of radionuclides radium-226 and strontium-90 by cyanobacteria forming intracellular calcium carbonates. *Environ Sci Technol.* 53(21):12639–12647.
- Mermet-Bouvier P, Chauvat F. 1994. A conditional expression vector for the cyanobacteria *Synechocystis* sp. strains PCC6803 and PCC6714 or *Synechococcus* sp. strains PCC7942 and PCC6301. *Curr Microbiol.* 28(3):145–148.
- Mirdita M, et al. 2021. ColabFold—making protein folding accessible to all. *bioRxiv.* doi:10.1101/2021.08.15.456425.
- Monteil CL, et al. 2021. Intracellular amorphous Ca-carbonate and magnetite biomineralization by a magnetotactic bacterium affiliated to the Alphaproteobacteria. *ISME J.* 15(1):1–18.
- Moreira D, et al. 2017. Description of *Gloeomargarita lithophora* gen. nov., sp. nov., a thylakoid-bearing, basal-branching cyanobacterium with intracellular carbonates, and proposal for *Gloeomargaritales* ord. nov. *Int J Syst Evol Microbiol.* 67(3):653–658.
- Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* 32(1):268–274.
- Pettersen EF, et al. 2004. UCSF chimera—a visualization system for exploratory research and analysis. *J Comput Chem.* 25(13):1605–1612.

- Ragon M, Benzerara K, Moreira D, Tavera R, López-García P. 2014. 16S rDNA-based analysis reveals cosmopolitan occurrence but limited diversity of two cyanobacterial lineages with contrasted patterns of intracellular carbonate mineralization. *Front Microbiol.* 5:331.
- Reynolds CS, Rogers DA. 1976. Seasonal variations in the vertical distribution and buoyancy of *Microcystis aeruginosa* Kütz. emend. Elenkin in Rostherne Mere, England. *Hydrobiologia* 48(1):17–23.
- Riding R. 2012. A hard life for cyanobacteria. *Science* 336(6080):427–428.
- Robert X, Gouet P. 2014. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* 42(Web Server Issue):W320–W324.
- Senes A, Engel DE, DeGrado WF. 2004. Folding of helical membrane proteins: the role of polar, GxxxG-like and proline motifs. *Curr Opin Struct Biol.* 14(4):465–479.
- Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. *Syst Biol.* 51(3):492–508.
- Veaudor T, et al. 2018. Overproduction of the cyanobacterial hydrogenase and selection of a mutant thriving on urea, as a possible step towards the future production of hydrogen coupled with water treatment. *PLoS One.* 13(6):e0198836.
- Wang X, et al. 2021. The evolution of calcification in reef-building corals. *Mol Biol Evol.* 38:3543–3555.
- Webb B, Sali A. 2016. Comparative protein structure modeling using MODELLER. *Curr Protoc Protein Sci.* 86:2.9.1–2.9.37.
- Weiner S, Dove PM. 2003. An overview of biomineralization processes and the problem of the vital effect. *Rev Mineral Biochem.* 54: 1–29.
- Yang M, Derbyshire MK, Yamashita RA, Marchler-Bauer A. 2020. NCBI's conserved domain database and tools for protein domain analysis. *Curr Protoc Bioinformatics.* 69(1):e90.
- Yarra T, Blaxter M, Clark MS. 2021. A Bivalve Biomineralization Toolbox. *Mol Biol Evol.* 38(9):4043–4055.
- Zimmermann L, et al. 2018. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J Mol Biol.* 430(15):2237–2243.

Associate editor: Tal Dagan

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

AlphaFold2-aided description of CoBaHMA, a novel family of bacterial domains within the Heavy-Metal Associated superfamily

Geoffroy Gaschignard^a, Maxime Millet^a, Apolline Bruley, Karim Benzerara, Manuela Dezi, Ferial Skouri-Panet, Elodie Duprat^b, Isabelle Callebaut^b

Sorbonne Université, Muséum National d’Histoire Naturelle, UMR CNRS 7590, Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, IMPMC, 75005 Paris, France

- a. Geoffroy Gaschignard and Maxime Millet should be considered joint first authors.
- b. Elodie Duprat and Isabelle Callebaut should be considered joint senior authors

Correspondence :

Elodie Duprat and Isabelle Callebaut, Sorbonne Université, Muséum National d'Histoire Naturelle, UMR CNRS 7590, Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, IMPMC, 75005 Paris, France. Email: elodie.duprat@sorbonne-universite.fr (E.D.) and isabelle.callebaut@sorbonne-universite. fr (I.C.).

23 **ABSTRACT**

24 Three-dimensional structure information, now available at the proteome scale, may facilitate
25 the detection of remote evolutionary relationships in protein superfamilies. Here, we illustrate
26 this with the identification of a novel family of protein domains related to the ferredoxin-like
27 superfold, by combining (i) transitive sequence similarity searches, (ii) clustering approaches
28 and (iii) the use of AlphaFold2 3D structure models. Domains of this family called CoBaHMA,
29 were initially identified in relation with the intracellular biomineralization of calcium
30 carbonates by Cyanobacteria. They are part of the large Heavy-Metal Associated (HMA)
31 superfamily, departing from the latter by specific sequence and structural features. In
32 particular, most CoBaHMA domains share conserved basic amino acids, forming a positively
33 charged surface which is likely to interact with anionic partners. CoBaHMA domains are found
34 in diverse modular organizations in bacteria, existing in the form of monodomain proteins or
35 as part of larger proteins, some of which are made up of membrane domains involved in
36 transport or lipid metabolism. This suggests that the CoBaHMA domains may exert a
37 regulatory function, involving interactions with anionic lipids. This hypothesis might have a
38 particular resonance in the context of the compartmentalization observed for Cyanobacteria
39 intracellular calcium carbonates.

40

41 **KEYWORDS:** Heavy-Metal Associated, AlphaFold2, sequence similarity search, graph
42 clustering, modular organization, functional annotation, P1B-type ATPase, ABC transporter,
43 PAP2, biomineralization, CoBaHMA

44

45 **ABBREVIATIONS**

46 **aa:** amino acids, **A:** Actuator, **ABC:** ATP-Binding Cassette, **AF2:** AlphaFold2, **AFDB:** AlphaFold
47 Protein Structure DataBase, **ATPDB:** ATP-Binding Domain, **ccya:** calcyanin, **CoBaHMA:**
48 Conserved Basic residues HMA, **DAG:** diacylglycerol, **GlyZip:** glycine-zipper, **HCA:** Hydrophobic
49 Cluster Analysis, **HMA:** Heavy Metal-Associated, **MSA:** Multiple Sequence Alignment, **N:**
50 Nucleotide, **NBD:** Nucleotide-Binding Domain, **P:** phosphorylation, **PAE:** Predicted Aligned
51 Error, **PAP2:** type 2 phosphatidylglycerol phosphatase, **PDB:** Protein Data Bank, **PG:**
52 phosphatidylglycerol, **pLDDT:** predicted Local Distance Difference Test, **R:** regulatory, **RSSs:**
53 Regular Secondary Structures, **S:** auxiliary membrane, **SCOP:** Structural Classification of
54 Proteins, **TM:** Transmembrane, **TMD:** Transmembrane Domain

55

56 **FUNDING INFORMATION**

57 This work was supported by the Agence Nationale de la Recherche (ANR Harley, ANR-19-
58 CE44-0017-01; ANR PHOSTORE, ANR- 19-CE01-0005).

59

60

61 **1. Introduction**

62 Superfolds are folds observed in a large number of evolutionary unrelated protein domain
63 superfamilies¹ and are characterized by compact super-secondary structure patterns^{2,3}. One
64 of these superfolds is the ferredoxin-like fold, found in 62 superfamilies according to the
65 SCOPe classification (d.58,⁴) and present in many domains with various functions^{5,6}. It is made
66 of a repeated β - α - β super-secondary structure, forming a four-stranded β -sheet, with the two
67 α -helices packed into an α - β sandwich. As for other superfolds, the ferredoxin-like fold is
68 subject to circular permutations, a mechanism which allows adaptation and the emergence of
69 new functions. This can be visualized by the ligation of the amino- and carboxyl-termini and
70 subsequent cleavage at another site^{7,8}. Among the 62 superfamilies comprising the
71 ferredoxin-like fold, the Heavy-Metal Associated (HMA) superfamily (SCOP d58.17) contains
72 the eponymous HMA family (**Figure 1A**). Domains of this HMA family contain two conserved
73 cysteine residues involved in heavy metal binding. They are found in a variety of metal-
74 trafficking proteins, which play essential roles in transport and homeostasis^{9,10}. While they
75 can be found alone, HMA domains are also part of diverse domain architectures, especially
76 associated with P_B-type ATPases, which are integral membrane proteins allowing transport of
77 metals across cell membranes¹¹.

78 We recently identified a novel family of domains belonging to the HMA superfamily, called
79 CoBaHMA (after Conserved Basic residues HMA). This discovery originates from the
80 characterization of a novel family of two-domain proteins, named calcyanin, which is
81 associated with intracellular biomineralization of calcium carbonates by cyanobacteria¹².
82 Calcyanins share a common architecture consisting in a conserved C-terminal domain, made
83 of a three-fold repeated, unusually very long glycine zipper motif ((GlyZip)₃). Glycine zippers
84 themselves consist of repeated GXXXG motifs, commonly found in transmembrane domains
85 and bacterial toxins^{13,14}. The calcyanin (GlyZip)₃ domain is preceded by a variable N-terminal
86 domain, specific to the 4 distinct calcyanin subgroups identified to this date, which are found
87 in distinct clades of cyanobacteria. The N-terminal domain of one of these calcyanin subgroups
88 is a CoBaHMA domain. It is represented by 15 sequences described in¹². It shares significant
89 sequence similarities with HMA domains, as well as plant integrated HMA domains¹⁵ and YAM
90 domains found in the C-terminal part of the bacterial YajR, an integral membrane protein

91 which belongs to the Major Facilitator Superfamily (MFS, ^{16,17}). These three families of
92 domains share the same 3D structure, but only the first one (HMA) is referenced in the SCOPe
93 classification (d.58.17.1, ⁴). Sequence analysis and molecular modeling ¹² have revealed
94 specific features of the CoBaHMA family, including the presence of an additional strand β_0 at
95 the N-terminus of the domain, which replaces the C-terminal strand β_4 , thus conforming to
96 the circular permutation scenario described before (**Figure 1B**). Calcyanins' CoBaHMA
97 domains are also characterized by the presence of conserved basic amino acids in β_0 (H) and
98 β_1 (R/K) strands, forming a charged patch on one side of the β -sheet. Despite the clear
99 identification of all these specific sequence and structural features, the function(s) of
100 CoBaHMA domains remain(s) unknown.

101 Here, we searched for the presence of CoBaHMA domains in proteins distinct from calcyanins.
102 We posit that the identity of co-occurring domains or combinations of domains in other
103 proteins may inform about their functional context ¹⁸⁻²⁰. For this purpose, we propose a novel
104 methodological framework, taking advantage of the structural information provided by the
105 Alphafold2 3D structure models ²¹ of the retrieved sequences, now widely available in the
106 Alphafold DataBase ²². This allows a both sensitive and specific detection of CoBaHMA
107 domains within the HMA superfamily. As a result, we describe a wide diversity of modular
108 organizations in which CoBaHMA domains are included, with some yet-to-be-characterized
109 regions. Moreover, the CoBaHMA domain appears to be specific to bacteria and frequently
110 found associated with transmembrane domains involved in transport of substrates and lipid
111 metabolism, suggesting a regulatory role with regard to these functions, probably via an
112 interaction with charged lipids.

113

114

115 2. MATERIAL AND METHODS

116

117 The workflow followed by this study is shown in **Figure 2**.

118

119 2.1. Structure-guided detection of CoBaHMA domain sequences (Figure 2-A)

120 We searched the UniClust30/UniRef30 databases (version 2022_02, downloaded from
121 https://gwdu111.gwdg.de/~compbiol/uniclust/2022_02/)²³ using the standalone version of
122 HHBlits²⁴. The 15 sequences of CoBaHMA domains identified by Benzerara *et al.*¹² were used
123 as individual probes. The sequence similarity search was made of 8 iterations with 50% hit
124 coverage and a E-value threshold of 1e-3 for each sequence probe. For each iteration and each
125 probe, we gathered the sequences extracted from the output a3m multiple sequence
126 alignment. We removed trivial redundancy using an in-house python3 script, by
127 merging entries with identical identifiers and whose sequences were identical on at least 50%
128 of their respective length. In order to find which entry had a 3D model available in AlphaFold
129 DB (<https://alphafold.ebi.ac.uk/>)²², we downloaded the accession_ids.csv file from
130 <http://ftp.ebi.ac.uk/pub/databases/alphafold/>. Based on this file, we downloaded the 3D
131 models corresponding to our entries from AlphaFold DB V4. We cropped each 3D model based
132 on the boundaries of the CoBaHMA domain candidates identified with the HHBlits search.
133 From the 3D coordinates of these sub-3D models, we computed the secondary
134 structures associated with each amino acid using DSSP^{25,26}. The 8 states output of DSSP was
135 converted to 3 states using the EVA convention²⁷. Following this convention, α -helix, 3_{10} helix
136 and π -helix are converted to helices (H), extended and isolated β -bridge to extended (E) and
137 turn, bend and other to coils (C). To validate the presence of a secondary structure, we set a
138 threshold of at least 3 consecutive amino acids with the same secondary structure assignment.
139 As a result, we could infer the string of regular secondary structures of the sub-model. We set
140 4 criteria to identify a CoBaHMA domain: First, it should contain the secondary pattern
141 EEHEEH(E) typical of CoBaHMA domains¹². Second, it should contain neither the Cys-X-X-Cys
142 pattern, typical of HMA domains¹⁰, nor the Cys-X-X-X-Cys pattern typical of TRASH domains
143²⁸. Third, the secondary pattern should not start with EH, as this was characteristic of some
144 false positives. Fourth, the first two beta strands had to be separated by a coil of 2 amino-
145 acids or less, to avoid confusing a β_0 strand of a CoBaHMA domain with the β_4 strand of an

146 HMA domain, which is followed by a second HMA domain starting with a β 1 strand). A final
147 manual check based on the alignment of sequences was conducted to remove the remaining
148 false positives. Entries meeting these 4 criteria and passing the manual check were considered
149 as CoBaHMA domains.

150 We performed transitive searches with newly identified CoBaHMA domains. This process was
151 iterated 6 times. We removed the redundancies out of the 6 iteration outputs by merging the
152 entries that had the same identifiers and shared at least 50% of their sequences.

153 Finally, since HHblits alignment only gave the portions of the sequences that match the
154 probes, we downloaded from UniProt (<https://www.uniprot.org/>)²⁹ the full sequences of
155 each identifier that had at least one CoBaHMA domain in its sequence.

156

157 **2.2. Annotation of the full-length sequences (Figure 2-B)**

158 The TaxID and organism name associated with the sequences were retrieved using the
159 UniProt REST API by downloading uniprot ttl files and serializing sequences information.
160 Taxonkit³⁰ was used to convert TaxID into taxonomic lineages based on the NCBI taxonomy.
161 Sequence functional features were assessed using several annotation tools. First, sequences
162 were annotated with InterProscan³¹ and InterPro³² (interproscan-5.59-91.0). Additionally,
163 domainMapper³³ [version 3.0.2] and the ECOD classification of known 3D structures of
164 domains were used to complete the sequence annotation. This HMM parsing algorithm
165 provides the detection of non-contiguous, insertional, and circularly permuted domains as
166 well. Results from both tools were filtered with an E-value threshold of 1e-10. Transmembrane
167 regions were annotated using deepTMHMM³⁴ [web version 1.0.20]. Finally, calcyanin
168 sequences were detected using *pCALF* (see below and **Supplementary Data 1**). Hydrophobic
169 Cluster Analysis (HCA^{35,36}) was also used to assess the foldability of the analyzed sequences
170^{37,38}.

171

172 **2.3. Calcyanin detection and classification (Supplementary Data 1)**

173 Sequences of calcyanins were detected in our data set using a dedicated tool called
174 *pCALF* (standing for python CALcyanin Finder). *pCALF* used four Hidden Markov Model (HMM)
175 profiles describing the glycine zipper triplication specific of calcyanin, called the (GlyZip)₃
176 motif, as described in¹². The (GlyZip)₃ HMM profile describes the whole triplication while Gly1,
177 Gly2 and Gly3 HMMs describe each glycine zipper individually. These HMMs profiles were

178 searched against amino acid sequences using pyHMMER^{39,40}. Additionally, a set of domains
179 (Y-type, X-type, Z-type and CoBaHMA), known to be found together with the glycine zipper
180 triplication, was used to annotate the N-terminal extremity of sequences. A sequence is
181 classified as calcyanin if it has a significant hit against the (GlyZip)₃ HMM profile (coverage
182 threshold: 0.6 ; e-value threshold: 1e-20) and significant hits against Gly1, Gly2 and Gly3
183 zippers in this particular order (coverage threshold: 0.7, e-value threshold: 1e-10). A sequence
184 is also classified as calcyanin if the second glycine zipper is missing and the N-terminal domain
185 is of Y-type.

186

187 **2.4. Classification of the full-length sequences (Figure 2-C)**

188 Full length proteins were clustered using mmseqs2 and a coverage threshold of 0.97⁴¹
189 (coverage mode: 0, cluster mode: 0, identity threshold: 0%).

190 This threshold was the lowest that kept the length difference between the shortest and the
191 longest sequences in a cluster below 50 residues and ensured that sequences belonging to the
192 same cluster had the same number of domains capabilities. Sequences representative of each
193 cluster, including singletons, were extracted and a self-versus-self search was performed using
194 mmseqs2 (alignment mode: 0). Alignment results of the clustering and search were combined
195 and filtered using a reciprocal coverage threshold of >0.7, an e-value threshold of <1e-10 and
196 a sequence length amplitude lower than 50 residues. The bitscore was normalized by the
197 length of the shortest sequences involved in the alignment⁴² and an additional filter was
198 applied using the median of normalized bitscore values (0.44).

199

200 A network was built using networkX⁴³ with sequences as nodes and alignments as edges,
201 weighted by the normalized bitscore. The network was refined iteratively in two steps:
202 community detection and edges removal. The best partition was found using the Louvain
203 Community Detection Algorithm⁴⁴. For each community, edges were removed if one or both
204 of its nodes has a length (i.e. number of residues) outside the interval $[X_c - SEM_c + N; X_c +$
205 $SEM_c - N]$, where X_c is the mean sequence length of a community, SEM_c is the standard error
206 of the mean sequence length of a community and N is a fixed number of residues $N=35$.
207 Louvain community detection and edges removal were repeated until there was no more edge
208 to remove and when the maximum length difference within all communities was less than 50
209 residues. Finally, the partitioning quality of the final network was assessed. Network layout

210 (edge-rendering and weighted spring embedded layout) and rendering were produced with
211 cytoscape ⁴⁵.

212

213 **2.5. Multiple alignment of representative sequences of the CoBaHMA family**

214 In order to build an alignment that is representative of CoBaHMA's diversity, and avoid
215 the over-representation of a subgroup of CoBaHMA domains, the 2305 CoBaHMA sequences
216 were clustered with mmseq2 (coverage mode: 0, cluster mode: 0, identity threshold: 60%,
217 coverage: 80%) ⁴¹. The representatives of the clusters and singletons were gathered,
218 amounting to a total of 1434 CoBaHMA sequences. To facilitate their alignment, the
219 sequences were split in subgroups of ~ 200 sequences. The sequences of these subgroups
220 were aligned using mafft v7.487 ⁴⁶ (maxiterate 1000; localpair), with some manual correction,
221 before merging them (mafft option: merge). The multiple sequence alignment (MSA) was
222 viewed and analyzed with Jalview 2.11.2.6 ⁴⁷.

223 WebLogo v2.8.2 ⁴⁸ was used on the Berkeley server (<https://weblogo.berkeley.edu/logo.cgi>)
224 to build the logo of the MSA, restricted to the CoBaHMA β strands. For the sake of clarity, we
225 removed the indels from alignment before making the logo.

226

227 **2.6. Analysis of the 3D structure models**

228 Manipulation and visualization of AF2 3D structure models were achieved using Chimera
229 ⁴⁹. Based on ⁵⁰, the pLDDT were split in 4 categories: pLDDT ≥ 90 (very high confidence); pLDDT
230 $\in [70, 90[$ (high confidence); pLDDT $\in [50, 70[$ (low confidence); and pLDDT <50 (non-
231 interpretable). The associated predicted aligned error (PAE), extracted from AFDB ²³ was also
232 considered in order to evaluate interdomain contacts. Each position of the PAE is the
233 incertitude in Å for the relative positions of 2 amino acids, whose positions in the sequence
234 are given by the x and y coordinates.

235 Structural similarities were searched using Foldseek ⁵¹. Three dimensional structure models
236 were analyzed in light of the multiple sequence alignments of the sequences of each
237 communities, built using mafft v7.487 ⁴⁶ and rendered using ESPrpt3 ⁵².

238

239

240

241 **3. RESULTS**

242 **3.1. Identification of CoBaHMA domains not contained in calcyanins**

243 We used the sequences of CoBaHMA domains from the 15 calcyanins reported in ¹² as
244 queries and performed iterative, profile-based sequence similarity searches, combined
245 with the consideration of structural features provided by AlphaFold2 models to specifically
246 identify CoBaHMA domains in the HMA superfamily (see Material and Methods; **Figure 2**). The
247 structural features specifically improved the discrimination between HMA and CoBaHMA
248 domains, which are otherwise close sequence-wise. However, this specificity was achieved at
249 the cost of a lower recovery of CoBaHMA domains, since not every sequence in the UniClust30
250 database had a model available in the AlphaFoldDB at the time of our study. We increased the
251 sensitivity by considering transitive searches (6 iterations), and using newly detected
252 CoBaHMA domains as probes for additional searches. A total of 38444 distinct sequences
253 identifiers were recorded by these searches. Among these identifiers, 28918 had AF2 models.
254 The remaining ones mostly corresponded to UniParc sequences. CoBaHMA-specific features
255 were considered using the AF2 models, restricting the set to a total of 2358 domains
256 corresponding to 2280 sequences, the manual inspection of which then identified 68 false
257 positive domains. Within this whole set, we identified a total number of 2305 (2290 domains
258 from Uniclust30 + 15 probes) CoBaHMA domains within 2227 different proteins (2212
259 sequences from Uniclust30 + 15 probes).

260

261 Most of the sequences were identified during the two first iterations (see **Supplementary**
262 **Data 2** for details). Only 8 phyla are represented by more than 20 sequences within the
263 UniClust30 database. All are affiliated to Bacteria. Among them, the Proteobacteria,
264 Cyanobacteria and Bacillota phyla are represented by 1025 (861, 153), 445 (349,78) and 365
265 (274,89) sequences, respectively. Forty-three, 87 and 40 sequences affiliated to the
266 Actomycetota phylum were identified during the three first iterations. During the whole
267 iterative similarity search process, only one sequence from Eukaryotes was detected during
268 the first transitive search (Chordata, Chondryctyes class, UniRef100_A0A401TJW7).

269

270 **Figure 3** illustrates the conserved sequence patterns of the CoBaHMA domains, derived
271 from the alignment of 1432 sequences representative of the whole domain diversity (see

272 Material and Methods). The β sheet displays several conserved features, spread over all β
273 strands, except strand β_4 , which is not always present. By contrast, the two α -helices are
274 highly variable and could not be aligned. First, strand β_1 possesses two arginines that are
275 highly conserved: 1266, i.e.88% of the sequences had the two arginines and 1353 (94% of the
276 sequences) had at least one of them. These two arginines are accompanied by another basic
277 residue (arginine or lysine) on the C-terminus of strand β_1 as well as a fairly conserved
278 histidine (925 (65%) sequences) in strand β_0 , and together, form a basic patch at the surface
279 of the β sheet. The full motif HxxxxRxxR that was originally identified in calcyanins ¹² was
280 present in only 27% of the CoBaHMA sequences. A continuum can thus be highlighted in the
281 CoBaHMA family, from sequences that do not have the basic patch to sequences that have
282 the full HxxxRxxR motif, hinting at the possible existence of sub-families with specific
283 features. Besides this basic patch, strand β_1 has a conserved PG motif on its N-terminus.
284 Strand β_3 has an array of conserved small amino acids (G/A/T/S) on its N-terminus, as well as
285 an aromatic (Y/F/H) position on its C-terminus, which is oriented toward the hydrophobic core
286 of the CoBaHMA in the 3D structure model. Considering their nature (small or apolar), and/or
287 their position (loops or orientations towards the hydrophobic core), all these conserved amino
288 acids are likely of structural importance. Finally, two polar positions occupied by an asparagine
289 (C-terminus of strand β_2) and an acidic residue (C-terminus of strand β_3) are also worth
290 mentioning. Interestingly, strand β_2 , which is the farthest from the basic patch located on
291 strands β_0 and β_1 appears to have less sequence conservation, except a central position
292 sometimes occupied by a basic residue. This further strengthens the hypothesis that the
293 functional feature of the CoBaHMA is yielded by strands β_0 and β_1 .

294

295 **3.2. Analysis of the full-length sequence communities**

296 Communities of full-length sequences have been defined from a similarity network and
297 the robustness of the affiliation of full length sequences to a given community was achieved
298 by using a refinement method described in the Material and methods (section 2.6).

299 Communities can be divided into several categories, depending on whether the CoBaHMA
300 domain alone constitutes the entire protein (what we call a “single-CoBaHMA domain
301 protein”), or the protein containing it is longer. Moreover, several scenarios can be
302 distinguished in the latter case, depending on whether or not the other regions of the
303 proteins, apart from the CoBaHMA domain, are already annotated by reference to profiles

304 contained in domain databases (InterPro (IPR), Pfam) or related to the calcyanin GlyZip motifs
305 (pCALF, see Material and Methods). An additional processing was added, by merging
306 communities sharing identical annotations and/or distinct annotations but corresponding to
307 the same major functional families. We describe the communities using this analysis
308 workflow, adding information about their 3D structures from AlphaFold2 models, as well as
309 conserved motifs identified from multiple sequence alignments (MSA, see **Supplementary**
310 **Data 3**).

311

312 We focused our analysis on the 48 most populated communities, containing at least 5
313 sequences and amounting to a total of 1410 sequences. Sparsely populated communities are
314 only analyzed on the basis of annotations also found within these large communities.

315 Most of these 48 large communities are composed of sequences affiliated with different
316 phyla. However, 15 of them (233 sequences) are specific to one bacterial phylum (**Figure 4**):
317 6, 5 and 4 communities comprised sequences affiliated to Bacillota (C675, 23 sequences; C521,
318 20 sequences; C4, 18 sequences; C413, 11 sequences; C397, 8 sequences; C369, 7 sequences),
319 Proteobacteria (C201, 53 sequences; C320, 13 sequences; C562, 10 sequences; C584, 8
320 sequences; C344, 5 sequences), or Cyanobacteria (C283, 31 sequences; C192, 14 sequences;
321 C20, 7 sequences; C654, 5 sequences) only, respectively.

322 While most of the large communities do not have any significant functional annotation (non-
323 annotated communities), 20 different domain families from InterPro (IPR) are detected within
324 the full-length sequences of 10 large communities (**Supplementary Table 1-A**). These families
325 are related to P-type ATPases (10 IPR families), ABC exporters (6 IPR families), type 2
326 phosphatidylglycerol phosphatases (2 IPR domain families) or HMA with no overlap with the
327 CoBaHMA domains (2 IPR).

328 Overall, the 48 sequence communities correspond to proteins with different modular
329 organization (**Figure 5** and **Supplementary Data 4** for the details in each community). These
330 communities may be grouped into categories as follows: small sequences (~100 aa length)
331 containing one CoBaHMA domain only, larger sequences combining a CoBaHMA domain with
332 functionally annotated regions (from IPR or other sources), non-annotated regions, or
333 additional CoBaHMA domains (either 2 or 3). The position of the communities relative to the
334 functional families/phyla and within the network are described in **Supplementary Data 5** and
335 **6**, respectively.

336

337 **3.2.1. Single-CoBaHMA domain proteins**

338 Eight of the 48 communities (C437, C201, C583, C675, C399, C397, C71, C685) fall into this
339 category, scattered over several phyla (**Figure 6**). They include a total of 402 sequences, and
340 represent more than 85% of the sequences in communities C437 (181 sequences), C685 (63
341 sequences), C201 (53 sequences), C583 (48 sequences). Superimposition of the AF2 3D models
342 of their representative members (in which all the core α helices and β strands are predicted
343 with very high/high pLDDT values) indicate that these single-CoBaHMA domain proteins
344 possess extra-regular secondary structures (mostly α helices) at the N- and/or C-terminus of
345 the domain, which pack against the core (**Figure 6**). Some variations are observed in the β -
346 sheet conserved motifs, with the particular case of community C397 (**Figure 6-F**), in which the
347 conserved basic amino acids are not present and the extra-N-terminal helix takes the place of
348 helix $\alpha 2$.

349

350 **3.2.2. CoBaHMA in multidomain proteins**

351 **• IPR-annotated communities**

352 Only a few large communities show IPR annotations. They exclusively correspond to
353 membrane proteins. We grouped together the IPR categories relating to the same protein
354 families (**Supplementary Table 1-A**), and also estimated what proportion of the less populated
355 communities ($n < 5$, number of communities and number of sequences) can be affiliated to
356 these large protein families. Apart from these major functional families, a limited number of
357 IPR annotations are found in small communities, essentially in singletons (**Supplementary**
358 **Table 1-B**)

359

360 **1) P-type ATPases.** P-type ATPases represent the bulk of the set, amounting to 216 sequences,
361 out of a total of 516 sequences over the whole set of communities (**Supplementary Table 1**).
362 P-type ATPases are composed of a common core of three conserved domains: (i) a
363 discontinuous transport-domain (**T-domain**) made of six membrane-spanning helices (M1 to
364 M6) providing the substrate translocation pathway, (ii) an ATP-binding domain (**ATPBD** -
365 between M4 and M5), which includes the nucleotide-binding domain (**N-domain**) and the
366 phosphorylation domain (**P-domain**), and (iii) an actuator domain (**A-domain**, between M2
367 and M3), which is believed to transmit changes in the ATPBD to the T-domain and to drive

368 dephosphorylation ¹¹. Two additional domains can complete this common core, depending on
369 the considered P-type ATPase subset: (i) the **S-domain**, which is an auxiliary membrane unit
370 providing support to the T domain and is located at various positions in the sequence (N- or
371 C-terminal relative to the T domain); (ii) **regulatory (R) domains**, which are located at the N-
372 terminus and/or C-terminus and act as intramolecular inhibitors, sensors for transported
373 cations and/or regulators for cation affinities ¹¹. Transport is accomplished via a so-called Post-
374 Albers cycle in which phosphorylation of a conserved aspartate residue in the ATPBD causes
375 the protein to cycle between high (E1)- and low (E2)-affinity ion-binding states. InterPro
376 entries (IPR, **Supplementary Table 1-A**) are available to annotate the P-type ATPases over
377 their full-length common core (IPR001757) or domains (IPR023298 : T-domain; IPR008250 : A-
378 domain; IPR023299 : N-domain; IPR044492, IP036412, IPR023214 : HAD/HAD-like), while
379 other IPRs provide annotations for specific P-type subsets (*e.g.* IPR027256 for P_{1B}-type,
380 IPR004014 and IPR006068 for cation-transporting P-type ATPases N-terminal and C-terminal,
381 respectively). IPR entries specific of HMA domains (IPR036163 and IPR006121) are also found,
382 outside the limits of the CoBaHMA domains, as accompanying some P_{1B}-type ATPases
383 Seven communities with at least 5 sequences and scattered over several phyla (**Figure 9**) are
384 annotated as P-type ATPases (**Supplementary Table 1-A**). Five out of the seven communities
385 (**C366, C429, C525, C140, C20**) belong to the P_{1B}-**type**. Indeed, the AlphaFold2 models of their
386 representative sequences (**Figure 7**) include a S-domain specific to this subset, comprising two
387 transmembrane helices, a long and curved MA helix and a kinked MB helix with an
388 amphipathic MB' segment at the cytoplasmic membrane interface, lining the ion entry point
389 ^{53,54}. Four out of these 5 communities (**C366, C429, C525, C140**) match the IPR027256 (P_{1B}-
390 type) profile. The five communities differ by the architecture of their whole proteins: the
391 biggest community, **C366** (115 sequences, *representative member: A0A3P1Y6T0, Figure 7-A*),
392 as well as **C429** (7 sequences, *representative member: A0A7I7MP94*, not represented in **Figure**
393 **7**) have a N-terminal CoBaHMA domain. Community **C525** (36 sequences, *representative*
394 *member: A0A5C7KBI4 - Figure 7-B*) is characterized by a N terminal CoBaHMA + HMA couple.
395 Community C140 (18 sequences, *representative member: A0A1Y6CVZ1, Figure 7-C*) contains a
396 tandem of CoBaHMA domains forming a continuous beta-sheet, with two additional strands
397 in the sequence linking them. However, only the C-terminal CoBaHMA domain was detected
398 by our search, while the N-terminal one lacked most of the basic residues. Last, the **C20**
399 community (7 sequences, *representative member: A0A3S1CNK6, Figure 7-D*), specific to

400 cyanobacteria, lacks the P-domain and has degenerated consensus motifs in domains A and
401 N, in contrast to the other communities which preserve these critical sequences (domains A
402 ([TS]-G-[DE]), P (DKTGT) and N (HP)) (**Supplementary Data 3**).

403 In addition, among the 7 communities annotated as P-type ATPases, the **C413** community (11
404 sequences, *representative member: R1CS5*, **Figure 7-E**) possesses a single N-terminal
405 CoBaHMA domain but has an atypical MA-MB segment, which does not match the usual
406 topology encountered in P_{1B}-type ATPases and could not be modelled accurately.

407 Finally, the community **C556** (22 sequences, *representative member: A0A7C4R520*, **Figure 7-**
408 **F**) has also a CoBaHMA domain in the N-terminus but do not belong to the P_{1B}-type. Instead,
409 it is annotated by cation N-terminal (IPR004014) and cation C-terminal (IPR006068) profiles
410 (**Supplementary Table 1-A**), which are found in several cation-transporting ATPases (Na⁺, K⁺,
411 Ca²⁺). Inspection of the AlphaFold2 model indicated a conserved calcium binding site in the T-
412 domain^{55,56}.

413

414 Members of the P_{1B}-type subsets were described heretofore as specific to the translocation of
415 heavy metal ions. They are divided into several groups based on conserved sequence motifs
416 (in the unwound part of M4, but also in M5 and M6) and the selectivity of the transported
417 metal ion⁵⁷. The six communities (**C366**, **C429**, **C525**, **C140**, **C20**, **C413**) of P_{1B}-type ATPases
418 highlighted here possess conserved motifs in the unwound part of M4, which differ from one
419 community to the other. The biggest community **C366** can be divided into two sub-
420 communities, according to these motifs (**Figure 7-A**): (i) Part of the sequences, such as *D8F5K4*,
421 has a conserved C-P-C motif (dotted box in **Figure 7-A**), typical of heavy metal binding sites;
422 (ii) other sequences, such as the representative sequence *A0A3P1Y6T0*, contain a
423 characteristic conserved motif including an aspartate as well as a basic residue (D-[YF]-x-[TC]-
424 x(2)-[KRH]). Both sub-communities have the same basic signature in their CoBaHMA domain,
425 including a histidine in stand β0. Community **C429** has a S-[SC]-P-C motif, similar to the **C366**
426 *D8F5K4* to which it is linked. A C-P-C motif is also found conserved in community **C525**, while
427 community C170 possesses a conserved M4 motif also including acidic and basic residues (N-
428 X-D-X-G-T-G-I-R). Last, the **C413** community has no strictly conserved residue in the unwound
429 part of M4, except from a central proline.

430 In conclusion, our results indicate that CoBaHMA domains form a novel family found
431 frequently in association with P_{1B}-type ATPases, similarly to HMA domains. However,

432 sequence signatures of heavy metal binding in M4 are not systematically found in these P_{1B}-
433 type ATPases, giving way to other signatures, including conserved acidic and basic residues
434 and suggesting that P_{1B}-type ATPases are not exclusively transporting heavy metals.

435

436 **2) ABC exporters.** CoBaHMA domains are also found in the N-terminus of **type I ABC exporters**
437 (17 out of 31 sequences over the whole set of communities). They are present in two large
438 communities: **C538** (*representative member UniProt A0A2W6BRH6*, 11 sequences) and **C520**
439 (*representative member UniProt E3FPA8*, 6 sequences).

440 Type I ABC exporters, formerly known as type IV exporters⁵⁸, transport a wide variety of
441 substrates across membranes. They consist in a TMD with six transmembrane (TM) helices
442 and a Nucleotide Binding Domain (NBD), which form homo- or hetero-dimers, with a swapped
443 arrangement of two TMs. Three InterPro IPR are associated with these two communities:
444 IPR036640, an ABC transporter type I of the transmembrane domain superfamily; IPR027417,
445 a P-loop containing nucleoside triphosphate hydrolase, and IPR039421, a type 1 protein
446 exporter, encompassing both the transmembrane domain (TMD) and the nucleotide binding
447 domain (NBD). The experimental 3D structures closest to the AF2 models of the C538 and
448 C520 representative sequences (FoldSeek searches) are those of bacterial ABC exporters
449 involved in the transport of various substrates, including lipid A (MsbA, pdb 7PH4), peptides
450 (TmrAB, pdb 6RAI) or multiple drugs (Sav1866, pdb 2HYD). The CoBaHMA-containing ABC
451 exporter sequences exhibit canonical ABC conserved motifs in the NBD (Walker A, Walker B,
452 ABC signature), suggesting that they are active transporters (**Supplementary Data 3**). In
453 contrast to other communities with CoBaHMA domains, the ABC exporter communities do
454 not have the conserved histidine in strand β_0 , while strands β_0 and β_1 include several basic
455 amino acids (**Figure 8-A**). Linkers of variable length separate the N-terminal CoBaHMA domain
456 from the TMD. These are predicted with lower pLDDT values as random coil or most often, as
457 a TMD hairpin (e.g. A0A2W6BRH6, community C538) or both (e.g. A0A6G4WXH4, community
458 **C455** or A0A7V8NHJ8, community **C499**), depending on the ABC exporter sequence. It is
459 precisely the length of the linker that makes the difference between the two most populated
460 communities, **C538** and **C520**, **Figure 8-A**).

461

462 Worth noting, a group of small communities, united under the common denominator **EcsC**
463 **proteins** (IPR024787: **C663**: 4 sequences, *representative member: A0A552EVV8*; **C684**: 1

464 sequence, *A0A098TP70*; **C83**: 1 sequence, *A0A6H2NMM5*, C52: 1 sequence, *Q606U9*) is
465 related to ABC transport systems. Indeed, in *Bacillus subtilis*, EcsC is found in an operon with
466 EcsA and EcsB, which are components of an ABC transport system⁵⁹. The AF2 model of the
467 EcsC domain folds as an *a priori* soluble bundle of TM helices, with most of the amino acids
468 characterized by low pLDDT values (**Figure 8-B**). A FoldSeek search did not highlight any
469 significant similarity with known 3D structures, suggesting that this domain adopts an as yet
470 uncharacterized fold. The EcsC CoBaHMA domains possess the characteristic His/Arg
471 signature in the first two β strands.

472

473 **C) Type 2 phosphatidylglycerol phosphatases (PAP2)**. CoBaHMA domains are present in the
474 N-terminus of some type 2 phosphatidylglycerol phosphatases (PAP2) in Cyanobacteria.
475 Community **C419** (representative member UniProt A0A3S1IWR7) contains 22 sequences,
476 whereas single members are found in 8 additional communities (**C16, C311, C418, C542, C617,**
477 **C629, C642, C643**), i.e., amounting in total to 30 sequences. PAP2 sequences are described by
478 two InterPro entries: IPR036938 and IPR000326, both entries being defined as phosphatidic
479 acid phosphatase type 2/haloperoxidase. Integral membrane proteins from the PAP2 family
480 dephosphorylate a variety of compounds, including lipids and carbohydrates⁶⁰. They consist
481 in a core TMD, with six tightly packed TM helices connected by extramembrane loops, two of
482 which interacting together to form the catalytic site (**Figure 8-C**). The sequences of the
483 CoBaHMA-containing PAP2 belong to the lipid phosphatase/phosphotransferase (LPT) family,
484 as they all contain the conserved tripartite active site motif (KX₆RP---PSGH---SRX₅HX₃D)⁶¹
485 (**Supplementary Data 3**). Members of this family modify several types of lipids in Gram-
486 negative bacteria, *e.g.* phosphatidylglycero-phosphate (PGP) for PgpB⁶², or lipid A for LpxE⁶³.
487 The AF2 model of the representative sequence resembles the 3D structure of *B. subtilis*
488 bsPgpB (pdb 6FMX(A)-5JKI(A), El Ghachi et al. 2017; FoldSeek TM-score 0.87, 26.7 % identity),
489 which contains eight TM α -helices, six of them ($\alpha 1, \alpha 4$ – $\alpha 8$) being tightly packed, while the $\alpha 2$
490 helix is amphiphilic, lying at the surface of the lipid bilayer on the active site side (**Figure 8-C**).
491 The linker separating the CoBaHMA domain, located in the intracellular milieu, from the
492 terminal TM $\alpha 1$ -helix is variable. It is predicted as a random coil or even as integrating
493 additional TM α -helices (*e.g.* A0A433NH73, community **C542**), always with very low pLDDT
494 values. Again, the PAP2 CoBaHMA domains possess the characteristic His/Arg signature in the
495 first two β strands.

496

497 **D) Diacylglycerol kinases**

498 Although not included in large communities, two small communities of cyanobacterial
499 proteins, **C515** (1 sequence, A0A0S3UCN3) and **C553** (2 sequences, *representative member:*
500 *A0A841V906*), caught our attention. They match several IPR profiles (IPR045540; IPR017438;
501 IPR016064; IPR005218; IPR001206; IPR004363) related with bacterial **diacylglycerol (DAG)**
502 **kinases**. These enzymes convert DAG, formed by the turnover of membrane phospholipids, to
503 phosphatidic acid ⁶⁴. In these proteins, the CoBaHMA domain is located at an unusual C-
504 terminal position (**Figure 8-D**), while the N-terminal domain corresponds to the catalytic DAK
505 kinase (FoldSeek matches with the putative DAG kinase from *Bacillus anthracis* (pdb 3T5P(B),
506 TM-score 0.83, 25 % identity, Hou et al. *unpublished*) and the DAG kinase DgkB from
507 *Staphylococcus aureus* (pdb 2QVL(A), TM-score 0.81, 22 % identity ⁶⁴). The 3D structure of
508 DgkB has a two-domain architecture, similar to that found in *E. coli* YegS, which
509 phosphorylates *in vitro* phosphatidyl glycerol ⁶⁵. Members of the **C515** and **C553** families share
510 the conserved P-loop (ϕ -x-x-G-G-D-G-T- ϕ , where ϕ represents a hydrophobic amino acid), but
511 exhibit slight difference in the two other conserved motifs, as described in ⁶⁴ ($(\phi$ - ϕ -x-N-P-x-
512 S/A-G instead of ϕ - ϕ -x-N-P-x-S-G and ϕ - ϕ -P-x-G-T-x-N-A- ϕ -x-N instead of ϕ - ϕ -P-x-G-T-x-N-D-
513 ϕ -x-R; side and top of the nucleotide-binding site, respectively). All three sequences share a
514 common domain in between the DAG kinase and the CoBaHMA domains, modeled with very
515 low pLDDT values as a long helix (A0A841V906) or a two-helix hairpin (A0A0S3UNC3). The
516 **C515** single member community (A0A0S3UCN3) differs from the two other sequences of the
517 **C553** community by an additional C-terminal domain, which is related to the GlyZip motif
518 described below. The DAGK CoBaHMA domains possess the characteristic His/Arg signature
519 in the first two β strands.

520

521 **• Other annotated, non-IPR, communities**

522 **A) HMA-2**

523 Eight hundred and thirty-two CoBaHMA domains overlap the Pfam profile HMA-2 (PF19991).
524 This profile is described as distantly related to HMA domains in its N-terminal part, containing
525 in particular the highly conserved histidine we also highlighted in this study. Not all the
526 CoBaHMA domains match the N-terminal part of the HMA-2 profile, indicating that proteins
527 matching this profile constitute a subset of the CoBaHMA family. However, the HMA-2 profile

528 is larger than CoBaHMA domains, with a total length of 180-190 amino acids, and includes at
529 its C-terminal part a conserved region generally predicted as two contiguous helices. Matches
530 to this C-terminal region are observed for sequences in 11 communities (**C75**, 81 sequences,
531 *representative member: A0A564ZMZ6*; **C393**, 62 sequences, *representative member: A0A1Z4FYZ0*;
532 **C141**, 46 sequences, *representative member: A0A1J1CTN7*; **C586**, 24 sequences, *representative*
533 *member: C9KNI5*; **C329**, 23 sequences, *representative member: A0A7Y6UHT*; **C521**, 20 sequences,
534 *representative member: A0A366XQZ5*; **C30**, 17 sequences, *representative member: A0A2U3KUJ9*;
535 **C562**, 10 sequences, *representative member: A0A4R3M3I3*; **C294**, 7 sequences, *representative*
536 *member: A0A2V7B4L6*, **C344**, 5 sequences, *representative member: A0A4P2PVK5*; **C5**, 5 sequences,
537 *representative member: A0A662ZLY5*). For communities having a C-terminal region associated
538 with high AF2 pLDDT values (≥ 70), such as **C393**, the two helices, often predicted as
539 transmembrane segments, pack together to form a hairpin (**Figure 9-A**). However, no obvious
540 similarity with any known 3D structure could be detected by FoldSeek for this case, outside
541 helix hairpins belonging to larger assemblies.

542 Matches with the HMA-2 profile were also detected in other communities, however limited
543 to the N-terminal CoBaHMA domain. We analyzed the regions C-terminal to the HMA-2 N-
544 ter/CoBaHMA domain, especially for detecting possible distant relationships to the HMA2-
545 Cter profile. Three communities **C202** (56 sequences, *representative member: A0A5C7T941*),
546 **C320** (13 sequences, *representative member: A0A0X8JQ54*) and **C584** (8 sequences,
547 *representative member: A0A3B9QA35*) also possess a hairpin of two helices, often predicted
548 as transmembrane segments, with conserved amino acids. However, they differ from those
549 defining the HMA2-Cter profile (illustrated with the representative sequence of the **C202**
550 community on **Figure 9-B**). In particular, the C202, C320 and C584 C-terminal regions are
551 characterized by highly conserved histidine residues, together with basic (arginine, lysine)
552 residues. Moreover, a FoldSeek search against PDB detected significant similarities (TM-score
553 0.82) between the C202 representative sequence and the long alpha-hairpin domain of a
554 manganese/iron superoxide dismutase (pdb 4BR6), which forms a four-helix bundle through
555 protein dimerization and provides histidine residues to the ion-binding site at the interface
556 with the preceding domain⁶⁶. In community **C331** (16 sequences, *representative member:*
557 *Q8YVH2*), in which proteins possess two CoBaHMA domains, the first HMA-2 N-ter/CoBaHMA
558 is also followed by a helical hairpin, predicted with low pLDDT values. This helical hairpin
559 contains conserved charged and aromatic amino acids. The second CoBaHMA domain is also

560 followed by a helical region, although less defined (**Figure 9-C**). The representative sequence
561 (A0A6MOJ010) of community **C341** (19 sequences) has a C-terminal region with three non-
562 packed alpha-helices, also predicted with very low pLDDT values and containing conserved
563 basic and acidic residues (**Figure 9-D**), while the representative sequence (A0A522V264) of
564 community **C444** (100 sequences) has two non-assembled alpha-helices, with less amino acid
565 conservation. Representative sequences from communities **C725** (A0A6F9WUC5, 12
566 sequences) and **C352** (A0A7Y4FMF1, 31 sequences) are apparently disordered (**Figure 9-E**),
567 but contain conserved charged residues accompanying hydrophobic clusters, suggesting a
568 hidden fold (**Supplementary Data 3**). Interestingly, a cyanobacterial community (**C283**, 31
569 sequences, *representative sequence: B1WT30*) has the HMA2-C-ter region at its N-terminus,
570 preceding the HMA2-Nter/CoBaHMA domain. This N-terminal HMA2-C-ter region is also
571 folded as a helical hairpin, with conserved charged/polar residues (**Figure 9-F**).

572

573 Finally, CoBaHMA domain-containing sequences belonging to the community **C678** (12
574 sequences, *representative member: A0A7W8FDY5*), which do not match the HMA2 profile,
575 also contain a helical hairpin with two conserved histidine residues in the second helix (**Figure**
576 **9-H**). These features are similar to the ones observed in the C202, C320 and C584
577 communities. A helical hairpin, with highly conserved charged amino acids, is also present in
578 three other non-HMA2 communities: **C106** (9 sequences, *representative member:*
579 *A0A7C2VAZ9*; **Figure 9-I**), **C578** (7 sequences, *representative member: A0A1H4BM11*; **Figure**
580 **9-J**) and **C487** (50 sequences, *representative member: H8GQW7*; **Figure 9-K**).

581

582 **B) Calcyanins: a (GlyZip)³ motif detected by a dedicated tool (pCALF)**

583 **Community C192** (14 sequences) corresponds to calcyanins, in which the CoBaHMA domain
584 was first identified. Some of the CoBaHMA domains of this community match the HMA2 N-ter
585 profile described above. Calcyanins contain a C-terminal domain, consisting in a three-fold
586 repeat of a large glycine-zipper (GlyZip) motif. This large GlyZip motif itself corresponds to a
587 duplicated smaller glycine zipper motif, interrupted in its middle part by a conserved Gly-Pro
588 dipeptide (**Supplementary Data 7**). AF2 modeled this GlyZip motif as a hairpin of tightly
589 packed helices, consistent with the presence of conserved glycine residues repeated every
590 four residues (**Figure 9-G**)^{14,67}. The low/very low pLDDT values may be due to the very large
591 sequence distance between this GlyZip motif and known hairpins of this type, present in

592 different architectures (as explored with FoldSeek). However, AF2 fails to assemble these
593 GlyZip motifs in a consistent way.

594

595 • **Non-annotated communities**

596 Last, very few communities with $n \geq 5$ other than those described above have been detected.
597 Protein segments associated with CoBaHMA domains correspond to disordered or ordered
598 regions. The order (foldability) in these regions was assessed by examining AF2 predictions, in
599 particular the segments associated with high/very high pLDDT values (≥ 0.7), as done for the
600 HMA-2 N-ter containing proteins. By this way, we also retrieved helical hairpins in other than
601 HMA-2 CoBaHMA communities (**Figure 9-H to -K**). In contrast, low values of pLDDT are
602 sometimes indicative of disorder, although in some cases, low pLDDT values are associated
603 with genuinely well folded regions (predicted as folded or in random coil), but for which the
604 prediction cannot be supported. For example, this is the case of new folds and sequences
605 lacking homologs^{37,38}. One way to evidence these "hidden" folded domains is to assess
606 foldability using Hydrophobic Cluster Analysis (HCA)^{37,38}. Therefore, we investigated
607 unannotated sequences by combining AF2 structure predictions and, when needed, HCA
608 analyses of the protein sequences.

609

610 The most populated community (**C233**; 39 sequences; *representative sequence: U2SKL7*) is
611 predicted by AlphaFold2 as a CoBaHMA domain with a C-terminal extension comprising three
612 strands completing the core beta-sheet (**Figure 10-A**). A four-helix bundle predicted with high
613 pLDDT values is inserted in between these the CoBaHMA domain and this C-terminal
614 extension. Interestingly, a FoldSeek search of this four-helix bundle, which contains strictly
615 conserved histidine, arginine and aromatic residues, indicated a possible structural
616 relationship with the MA-MB-M1-M2 block of P1B-type ATPases (pdb 4UMWTM-score 0.54,
617 11.3 identity, **Figure 10-A**). The MB kink (resulting in MB') is not present in the **C233** four-helix
618 bundle, whereas the basic residues (histidine and arginine) are located at the entry site of the
619 P1B-type ATPase.

620 The **C4** community (18 sequences, *representative sequence: A0A1HOCOW7*) is also predicted
621 by AlphaFold2 as composed of a N-terminal CoBaHMA domain and a four-helix bundle (**Figure**
622 **10-B**), also with conserved amino acids (especially arginine and histidine residues) located at
623 the tip of the bundle. However, in this case, no significant structural similarity was found

624 between the representative member of the community and any experimental 3D structures
625 using FoldSeek.

626 A third community (**C93**, 6 sequences, *representative sequence: A0A552LCK1*) possesses a C-
627 terminal CoBaHMA domain, preceded by a globular domain, with conserved residues, but
628 which does not share obvious similarity with any available experimental 3D structure
629 (FoldSeek search) (**Figure 10-C**).

630 Finally, we also examined the **C654** community (5 sequences, *representative sequence:*
631 *A0A0M0SSI8*). It is characterized by a C-terminal HMA-2 N-ter/CoBaHMA domain, preceded
632 by a tandem of CoBaHMA-like domains (devoid of the conserved basic signature). However, it
633 did not contain any other folded domain based on AlphaFold2 modelling and HCA analysis
634 (**Figure 10-D**).

635

636

637 4. Discussion

638 The large-scale predictions of 3D structures now enabled by AI-based approaches allow to
639 functionally annotate large sets of proteomes at the amino acid level, and identify new folds
640 (e.g. ⁶⁸). The 3D models provided by AlphaFold2 via a dedicated database ²², connected to
641 UniProt (²⁹), offer an unprecedented tool for extending the exploration of the universe of
642 protein domains and studying their evolutionary trajectory. Here, we have taken advantage
643 of this large-scale structural information, combined with sequence similarity search and
644 clustering tools, to identify a new family of domains called CoBaHMA. This domain family
645 shares a common evolutionary origin with HMA domains, as evidenced by the signature of a
646 common hydrophobic core. However, CoBaHMA can be discriminated from HMA based on an
647 additional, external β strand (called $\beta 0$) and, in many cases, a specific sequence signature,
648 conferring a positive electrostatic charge on one side of the beta sheet, which is likely to be
649 associated with a specific function. The AlphaFold2 models provided a structural constraint
650 throughout the workflow we built, enabling a fine discrimination between CoBaHMA domains
651 and the rest of the very large HMA superfamily. Besides providing information about a novel
652 family of domains, our study illustrates how evolution may operate within a superfold to
653 provide broad functional diversity.

654 This study may have some limitations. Indeed, the structural features we have automatically
655 defined lightly suffer from the definition of cutoffs, which does not rule out the possibility that
656 highly divergent members of the CoBaHMA with large loops between the two first β strands
657 may be overlooked. The methodology developed is also dependent on the availability and
658 accuracy of AF2 models. Therefore, the CoBaHMA family described in this work constitutes a
659 minimum set as we have considered a non-redundant bank of sequences (UniRef30).
660 However, even with this possible extension, CoBaHMA domains are only present in a
661 restricted set of proteins, suggesting a specific function within different protein families
662 limited to certain species or phyla.

663

664 Further prediction of the CoBaHMA-specific function, or at least the biological environment in
665 which it is performed, can be aided by analyzing the domain architecture of the proteins
666 containing the CoBaHMA domain. Here, deciphering this architecture was again aided by the
667 AlphaFold2 predictions, combined with domain database (InterPro) annotations. A large part
668 of the CoBaHMA communities corresponds to single CoBaHMA domain proteins. This is

669 reminiscent of the single-HMA domain proteins, behaving as chaperones ⁶⁹. CoBaHMA
670 domains are also found in association with a limited number of protein families, which mostly
671 correspond to membrane proteins, at least for those annotated through InterPro profiles and
672 predictors of transmembrane segments. Moreover, CoBaHMA domains appear to be specific
673 to bacteria, in contrast to HMA domains which are found in bacteria, archaea and eukaryote.
674 Overall, These observations suggest a membrane-related function specific to bacteria for
675 CoBaHMA..

676

677 Like HMA domains, CoBaHMA domains are particularly abundant in P_{1B}-type ATPases. The
678 latter are commonly defined as integral membrane proteins that couple ATP hydrolysis to the
679 transport of metal cations, such as copper, zinc and cobalt ⁵⁷. The specificity of P_{1B}-type
680 ATPases towards heavy metals is linked to conserved motifs in the middle of the fourth
681 transmembrane helix (TM4). These motifs directly coordinate the ion through
682 cysteine/histidine side chains. Besides, soluble N- and C-terminal metal-binding extensions
683 (known under the generic term Heavy Metal Binding Domains (HMBDs)), also rich in
684 cysteine/histidine and including HMA domains, seemingly play a regulatory role ^{53,54,70}. In
685 particular, HMBDs interact with the amphipathic helix MB', lying at the membrane-cytosol
686 interface at the end of a P_{1B}-specific MA and MB membrane hairpin ⁷¹. This amphipathic MB'
687 helix is connected to the high affinity ion-binding site through a conserved electronegative
688 funnel. HMBDs also interact with the cytosolic domains (A and P domains), thus playing a
689 potential regulatory role by interfering with conformational changes coupling ATP hydrolysis
690 with ion transport across the membrane ⁷¹. Here, we show that P_{1B}-type ATPases are not
691 restricted to heavy metals. Indeed, the proteins identified here share a P_{1B}-specific MA and
692 MB membrane hairpin but contain CoBaHMA domains instead of the usually encountered
693 HMA domains and have conserved motifs in TM4 different from those coordinating heavy
694 metals. These motifs vary depending on the considered community, often including charged
695 (acidic and/or basic) or polar (serine/threonine) residues. Interestingly, one community
696 contains the TM4 CPC motif, typical of heavy metal transport, together with a CoBaHMA
697 domain only, and no HMA domain at its N-terminus. This suggests that the coupling of HMA
698 with TM4 Cys-rich motif is not necessary as initially thought. Moreover, a large
699 sequence/structure diversity is observed at the level of the MA/MB hairpin, probably linked
700 to the substrate diversity of CoBaHMA-containing P_{1B}-type ATPases, as also evidenced by the

701 diversity of the M4 conserved motifs. Finally, it is worth noting that CoBaHMA domains are
702 also found associated with members of the P_{2A}-type ATPases (SERCA, Ca²⁺ ATPases), and are
703 thus not restricted to the P_{1B}-type subgroup.

704 The question remains as to what are the substrate specificities of these P-type proteins
705 associated with CoBaHMA domains, and what is the role of the CoBaHMA domains in this
706 specific modular organization. A regulatory role similar to that played by HMA could be
707 expected, supported by the fact that some AF2 models displayed significant interfaces with A
708 domains, involving amino acids outside the conserved basic patch (**Supplementary Data 8**).
709 The identity of the ligands for these charged amino acids remains to be explored.

710

711 Accessory domains provide an additional, often regulatory effect on the functions of ATP-
712 binding cassette (ABC) exporter cores, which are formed by transmembrane domains (TMDs)
713 and nucleotide-binding domains (NBDs) ⁷². For instance, the cytosolic Cystathionin Beta
714 Synthase (CBS) domains, at the C-terminus of the osmoregulatory OpuA, inhibits the
715 transporter activity by binding to cyclic-di-AMP ⁷³. This protein is gated by ionic strength,
716 which modulates the interaction of positively charged amino acids in the NBDs with negatively
717 charged lipids ⁷³. Although, like for the P-type ATPases, the specific function of CoBaHMA
718 domains in this ABC context remains to be discovered, a number of points can be considered.
719 First, the ABC transporters with the highest sequence identities (~30 %) include lipid
720 transporters, including MsbA, suggesting that (i) the CoBaHMA-domain-containing ABC
721 transporters might be involved in lipid transport, and (ii) the CoBaHMA domain might be
722 directly involved in the uptake of lipids. Second, the specific position of the domain, N-terminal
723 to the NBD, places it at the right location to interact with the polar heads of lipids, as observed
724 for instance with the lasso domain found in some ABCC transporters such as the Cystic Fibrosis
725 Transmembrane conductance Regulator (CFTR) protein (ABCC7) (⁷⁴ for a review). This suggests
726 a specific role in contacting the membrane via the conserved basic patch at the surface of the
727 domain. The predicted presence of additional TM helices between the CoBaHMA domain and
728 the TMD in most of the ABC communities, like in the ABCC transporters ABCC1 (MRP1) and
729 ABCC8 (SUR1) ⁵⁸, might serve an additional regulatory purpose. This “lipid hypothesis” is also
730 interesting to consider with regard to the function of CoBaHMA domain in the context of P_{1B}-
731 type ATPases, particularly as specific transport of lipids is carried out by another class of P-
732 type ATPases (P₄-type, ^{75,76}).

733

734 The hypothesis of the CoBaHMA domain serving as a binding module for positively charged
735 lipids (phospholipids) in bacteria is appealing considering the wide knowledge about
736 phospholipids-binding domains in eukaryotes. Indeed, the tray of basic amino acids offered
737 by CoBaHMA domains resembles that observed in C2 domains for instance, which interact
738 with membranes in a Ca²⁺-dependent manner through a polybasic cluster, with specificity to
739 phosphatidylinositol-4,5-bisphosphate ^{77,78}.

740 The phosphatidylinositol phosphates that are largely recognized in eukaryotic membranes are
741 also the targets of some bacterial proteins acting as effectors or toxins (*e.g.* the lipid raft
742 targeting domain of the *Bordetella* pathogens ⁷⁹, also see ⁸⁰ for a review). However, bacterial
743 membranes are distinct in lipid composition from eukaryotic membranes, and their lipid-
744 binding modules are far less well known. Phosphatidylglycerol (PG) might be a target for the
745 CoBaHMA domain. This phospholipid is present in both Gram - and Gram + bacteria, and plays
746 a central role in the synthesis of cardiolipin (CL, diphosphatidylglycerol),
747 lysophosphatidylglycerol (LPG) and oligosaccharides ⁸¹. Phosphatidic acid (PA) is another
748 potential candidate. It is linked to the activity of diacylglycerol (DAG) kinase, and serves as a
749 precursor for glycerolipids ⁸². An appealing hypothesis is that, in addition to being specifically
750 recognized by the CoBaHMA domains, these phospholipids could be transported by
751 membrane systems in which CoBaHMA is included (*e.g.* ABC exporter), a mechanism which
752 could contribute to the general lipid homeostasis. However, such hypotheses remain highly
753 speculative and need to be extensively tested.

754 Phosphoglycerolipids are far less abundant in cyanobacterial membranes, with PG being the
755 only phospholipid present) ⁸³. It is present in the thylakoid membranes in low proportion
756 (10 %) relative to the more abundant glycerolipids monogalactosyldiacylglycerol (MGDG),
757 digalactosyldiacylglycerol (DGDG) and sulfoquinovosyldiacylglycerol (SQDG) ⁸³. PG proportion
758 is regulated in response to phosphate availability ⁸⁴, but is essential not only for
759 photosynthesis ⁸⁵ but also cell division and metabolism ⁸⁶. We note that CoBaHMA domains
760 are found in cyanobacteria-restricted communities of phosphatidic acid phosphatases (PAP2)
761 and diacylglycerol (DAG) kinases. Both types of enzymes are involved in the biosynthesis of
762 lipids starting from phosphatidic acid (PA).

763

764 One of the outstanding features of our domain grammar analysis was the co-occurrence of
765 the CoBaHMA domain not only with already annotated, membrane-specific domains, but also
766 with hairpins of two consecutive helices. These helical hairpins show both a conserved
767 structural motif as revealed by AF2 models, and a wide variety of sequences: some include a
768 lot of strong hydrophobic amino acids, as in the HMA-2 C-ter profile, and are predicted as
769 forming transmembrane segments; others include small but also globally apolar residues.
770 However, most of them include conserved, charged residues, in particular histidine and basic
771 residues. This suggests that these hairpins may serve as platforms on which amino acids can
772 be grafted to interact with specific ions or ligands. From an evolutionary point of view, it is
773 tempting to speculate that these hairpins can be used as basic units for integrating more
774 complex architectures, such as ABC exporter TMDs or those present in calcyanins (three
775 repeats of a glycine-zipper helical hairpin). These calcyanin glycine-zippers are structures
776 characterized by very compact assemblies due to the presence of glycine every 4 residues, but
777 it remains yet to be specified whether they are soluble or membrane-bound. One open
778 question is to what extent the MA-MB hairpin specific to P_{1B}-type ATPases may have evolved
779 from this basic module. Finally, from a methodological point of view, it would be interesting
780 to consider this CoBaHMA-specific grammar to improve sequence similarity searches, as done
781 for instance by Terrapon et al.⁸⁷ and Faure and Callebaut¹⁸ or, more recently, by Buchan and
782 Jones⁸⁸ using natural language word embedding techniques.
783

- 785 1. Orengo CA, Jones DT, Thornton JM. Protein superfamilies and domain superfolds.
786 *Nature*. 1994;372(6507):631-634.
- 787 2. Chitturi B, Shi S, Kinch LN, Grishin NV. Compact Structure Patterns in Proteins. *J Mol*
788 *Biol*. 2016;428(21):4392-4412.
- 789 3. Kolodny R. Searching protein space for ancient sub-domain segments. *Curr Opin Struct*
790 *Biol*. 2021;68:105-112.
- 791 4. Chandonia JM, Guan L, Lin S, Yu C, Fox NK, Brenner SE. SCOPe: improvements to
792 the structural classification of proteins - extended database to facilitate variant
793 interpretation and machine learning. *Nucleic Acids Res*. 2022;50(D1):D553-d559.
- 794 5. Caetano-Anollés G, Caetano-Anollés D. An evolutionarily structured universe of
795 protein architecture. *Genome Res*. 2003;13(7):1563-1571.
- 796 6. Thornton JM, Orengo CA, Todd AE, Pearl FM. Protein folds, functions and evolution.
797 *J Mol Biol*. 1999;293(2):333-342.
- 798 7. Grishin NV. Fold change in evolution of protein structures. *J Struct Biol*. 2001;134(2-
799 3):167-185.
- 800 8. Jung J, Lee B. Circularly permuted proteins in the protein structure database. *Protein*
801 *Sci*. 2001;10(9):1881-1886.
- 802 9. Arnesano F, Banci L, Bertini I, et al. Metallochaperones and metal-transporting
803 ATPases: a comparative analysis of sequences and structures. *Genome Res*.
804 2002;12(2):255-271.
- 805 10. Bull PC, Cox DW. Wilson disease and Menkes disease: new handles on heavy-metal
806 transport. *Trends Genet*. 1994;10(7):246-252.
- 807 11. Palmgren MG, Nissen P. P-type ATPases. *Annu Rev Biophys*. 2011;40:243-266.
- 808 12. Benzerara K, Duprat E, Bitard-Feildel T, et al. A New Gene Family Diagnostic for
809 Intracellular Biomineralization of Amorphous Ca Carbonates by Cyanobacteria.
810 *Genome Biol Evol*. 2022;14(3).
- 811 13. Kim S, Chamberlain AK, Bowie JU. Membrane channel structure of *Helicobacter pylori*
812 vacuolating toxin: role of multiple GXXXG motifs in cylindrical channels. *Proc Natl*
813 *Acad Sci U S A*. 2004;101(16):5988-5991.
- 814 14. Kim S, Jeon TJ, Oberai A, Yang D, Schmidt JJ, Bowie JU. Transmembrane glycine
815 zippers: physiological and pathological roles in membrane proteins. *Proc Natl Acad Sci*
816 *U S A*. 2005;102(40):14278-14283.
- 817 15. De la Concepcion JC, Franceschetti M, Maqbool A, et al. Polymorphic residues in rice
818 NLRs expand binding and response to effectors of the blast pathogen. *Nat Plants*.
819 2018;4(8):576-585.
- 820 16. Jiang D, Zhao Y, Fan J, et al. Atomic resolution structure of the *E. coli* YajR transporter
821 YAM domain. *Biochem Biophys Res Commun*. 2014;450(2):929-935.
- 822 17. Jiang D, Zhao Y, Wang X, et al. Structure of the YajR transporter suggests a transport
823 mechanism based on the conserved motif A. *Proc Natl Acad Sci U S A*.
824 2013;110(36):14664-14669.
- 825 18. Faure G, Callebaut I. Identification of hidden relationships from the coupling of
826 hydrophobic cluster analysis and domain architecture information. *Bioinformatics*.
827 2013;29(14):1726-1733.
- 828 19. Jin J, Xie X, Chen C, et al. Eukaryotic protein domains as functional units of cellular
829 evolution. *Sci Signal*. 2009;2(98):ra76.
- 830 20. Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA. Structure, function and
831 evolution of multidomain proteins. *Curr Opin Struct Biol*. 2004;14(2):208-216.

- 832 21. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with
833 AlphaFold. *Nature*. 2021;596(7873):583-589.
- 834 22. Varadi M, Anyango S, Deshpande M, et al. AlphaFold Protein Structure Database:
835 massively expanding the structural coverage of protein-sequence space with high-
836 accuracy models. *Nucleic Acids Res*. 2022;50(D1):D439-d444.
- 837 23. Mirdita M, von den Driesch L, Galiez C, Martin MJ, Söding J, Steinegger M. Uniclust
838 databases of clustered and deeply annotated protein sequences and alignments. *Nucleic*
839 *Acids Res*. 2017;45(D1):D170-d176.
- 840 24. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. HH-suite3
841 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*.
842 2019;20(1):473.
- 843 25. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of
844 hydrogen-bonded and geometrical features. *Biopolymers*. 1983;22(12):2577-2637.
- 845 26. Touw WG, Baakman C, Black J, et al. A series of PDB-related databanks for everyday
846 needs. *Nucleic Acids Res*. 2015;43(Database issue):D364-368.
- 847 27. Rost B, Eyrich VA. EVA: large-scale analysis of secondary structure prediction.
848 *Proteins*. 2001;Suppl 5:192-199.
- 849 28. Ettema TJ, Huynen MA, de Vos WM, van der Oost J. TRASH: a novel metal-binding
850 domain predicted to be involved in heavy-metal sensing, trafficking and resistance.
851 *Trends Biochem Sci*. 2003;28(4):170-173.
- 852 29. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res*.
853 2023;51(D1):D523-d531.
- 854 30. Shen W, Ren H. TaxonKit: A practical and efficient NCBI taxonomy toolkit. *J Genet*
855 *Genomics*. 2021;48(9):844-850.
- 856 31. Jones P, Binns D, Chang HY, et al. InterProScan 5: genome-scale protein function
857 classification. *Bioinformatics*. 2014;30(9):1236-1240.
- 858 32. Blum M, Chang HY, Chuguransky S, et al. The InterPro protein families and domains
859 database: 20 years on. *Nucleic Acids Res*. 2021;49(D1):D344-d354.
- 860 33. Manriquez-Sandoval E, Fried SD. DomainMapper: Accurate domain structure
861 annotation including those with non-contiguous topologies. *Protein Sci*.
862 2022;31(11):e4465.
- 863 34. Hallgren J, Tsigirgos KD, Pedersen MD, et al. DeepTMHMM predicts alpha and beta
864 transmembrane proteins using deep neural networks. *bioRxiv*.
865 2022:2022.2004.2008.487609.
- 866 35. Bitard-Feildel T, Lamiable A, Mornon JP, Callebaut I. Order in Disorder as Observed
867 by the "Hydrophobic Cluster Analysis" of Protein Sequences. *Proteomics*. 2018;18(21-
868 22):e1800054.
- 869 36. Callebaut I, Labesse G, Durand P, et al. Deciphering protein sequence information
870 through hydrophobic cluster analysis (HCA): current status and perspectives. *Cell Mol*
871 *Life Sci*. 1997;53(8):621-645.
- 872 37. Bruley A, Bitard-Feildel T, Callebaut I, Duprat E. A sequence-based foldability score
873 combined with AlphaFold2 predictions to disentangle the protein order/disorder
874 continuum. *Proteins*. 2023;91(4):466-484.
- 875 38. Bruley A, Mornon JP, Duprat E, Callebaut I. Digging into the 3D Structure Predictions
876 of AlphaFold2 with Low Confidence: Disorder and Beyond. *Biomolecules*.
877 2022;12(10).
- 878 39. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity
879 searching. *Nucleic Acids Res*. 2011;39(Web Server issue):W29-37.
- 880 40. Larralde M, Zeller G. PyHMMER: a Python library binding to HMMER for efficient
881 sequence analysis. *Bioinformatics*. 2023;39(5).

- 882 41. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the
883 analysis of massive data sets. *Nat Biotechnol.* 2017;35(11):1026-1028.
- 884 42. Gibbons TR, Mount SM, Cooper ED, Delwiche CF. Evaluation of BLAST-based edge-
885 weighting metrics used for homology inference with the Markov Clustering algorithm.
886 *BMC Bioinformatics.* 2015;16:218.
- 887 43. Hagberg A, Swart P, S Chult D. Exploring network structure, dynamics, and function
888 using networkx. Conference: SCIPY 08 ; August 21, 2008 ; Pasadena; 2008; United
889 States.
- 890 44. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities
891 in large networks. *Journal of Statistical Mechanics: Theory and Experiment.*
892 2008;2008(10):P10008.
- 893 45. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated
894 models of biomolecular interaction networks. *Genome Res.* 2003;13(11):2498-2504.
- 895 46. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7:
896 improvements in performance and usability. *Mol Biol Evol.* 2013;30(4):772-780.
- 897 47. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview Version 2--a
898 multiple sequence alignment editor and analysis workbench. *Bioinformatics.*
899 2009;25(9):1189-1191.
- 900 48. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator.
901 *Genome Res.* 2004;14(6):1188-1190.
- 902 49. Pettersen EF, Goddard TD, Huang CC, et al. UCSF ChimeraX: Structure visualization
903 for researchers, educators, and developers. *Protein Sci.* 2021;30(1):70-82.
- 904 50. Tunyasuvunakool K, Adler J, Wu Z, et al. Highly accurate protein structure prediction
905 for the human proteome. *Nature.* 2021;596(7873):590-596.
- 906 51. van Kempen M, Kim SS, Tumescheit C, et al. Fast and accurate protein structure search
907 with Foldseek. *Nat Biotechnol.* 2023.
- 908 52. Robert X, Gouet P. Deciphering key features in protein structures with the new
909 ENDscript server. *Nucleic Acids Res.* 2014;42(Web Server issue):W320-324.
- 910 53. Andersson M, Mattle D, Sitsel O, et al. Copper-transporting P-type ATPases use a
911 unique ion-release pathway. *Nat Struct Mol Biol.* 2014;21(1):43-48.
- 912 54. Gourdon P, Liu XY, Skjørringe T, et al. Crystal structure of a copper-transporting PIB-
913 type ATPase. *Nature.* 2011;475(7354):59-64.
- 914 55. Olesen C, Picard M, Winther AM, et al. The structural basis of calcium transport by the
915 calcium pump. *Nature.* 2007;450(7172):1036-1042.
- 916 56. Toyoshima C, Nakasako M, Nomura H, Ogawa H. Crystal structure of the calcium
917 pump of sarcoplasmic reticulum at 2.6 Å resolution. *Nature.* 2000;405(6787):647-655.
- 918 57. Smith AT, Smith KP, Rosenzweig AC. Diversity of the metal-transporting P1B-type
919 ATPases. *J Biol Inorg Chem.* 2014;19(6):947-960.
- 920 58. Thomas C, Aller SG, Beis K, et al. Structural and functional diversity calls for a new
921 classification of ABC transporters. *FEBS Lett.* 2020;594(23):3767-3775.
- 922 59. Leskelä S, Kontinen VP, Sarvas M. Molecular analysis of an operon in *Bacillus subtilis*
923 encoding a novel ABC transporter with a role in exoprotein production, sporulation and
924 competence. *Microbiology (Reading).* 1996;142 (Pt 1):71-77.
- 925 60. Sigal YJ, McDermott MI, Morris AJ. Integral membrane lipid
926 phosphatases/phosphotransferases: common structure and diverse functions. *Biochem*
927 *J.* 2005;387(Pt 2):281-293.
- 928 61. Stuke J, Carman GM. Identification of a novel phosphatase sequence motif. *Protein*
929 *Sci.* 1997;6(2):469-472.

- 930 62. Dillon DA, Wu WI, Riedel B, Wissing JB, Dowhan W, Carman GM. The Escherichia
931 coli pgpB gene encodes for a diacylglycerol pyrophosphate phosphatase activity. *J Biol*
932 *Chem*. 1996;271(48):30548-30553.
- 933 63. Zhao J, An J, Hwang D, et al. The Lipid A 1-Phosphatase, LpxE, Functionally Connects
934 Multiple Layers of Bacterial Envelope Biogenesis. *mBio*. 2019;10(3).
- 935 64. Miller DJ, Jerga A, Rock CO, White SW. Analysis of the Staphylococcus aureus DgkB
936 structure reveals a common catalytic mechanism for the soluble diacylglycerol kinases.
937 *Structure*. 2008;16(7):1036-1046.
- 938 65. Bakali MA, Nordlund P, Hallberg BM. Expression, purification, crystallization and
939 preliminary diffraction studies of the mammalian DAG kinase homologue YegS from
940 Escherichia coli. *Acta Crystallogr Sect F Struct Biol Cryst Commun*. 2006;62(Pt 3):295-
941 297.
- 942 66. Borgstahl GE, Parge HE, Hickey MJ, Beyer WF, Jr., Hallewell RA, Tainer JA. The
943 structure of human mitochondrial manganese superoxide dismutase reveals a novel
944 tetrameric interface of two 4-helix bundles. *Cell*. 1992;71(1):107-118.
- 945 67. Kleiger G, Grothe R, Mallick P, Eisenberg D. GXXXG and AXXXA: common alpha-
946 helical interaction motifs in proteins, particularly in extremophiles. *Biochemistry*.
947 2002;41(19):5990-5997.
- 948 68. Koehler Leman J, Szczerbiak P, Renfrew PD, et al. Sequence-structure-function
949 relationships in the microbial protein universe. *Nat Commun*. 2023;14(1):2351.
- 950 69. Jordan IK, Natale DA, Koonin EV, Galperin MY. Independent evolution of heavy
951 metal-associated domains in copper chaperones and copper-transporting atpases. *J Mol*
952 *Evol*. 2001;53(6):622-633.
- 953 70. Wang K, Sitsel O, Meloni G, et al. Structure and mechanism of Zn²⁺-transporting P-
954 type ATPases. *Nature*. 2014;514(7523):518-522.
- 955 71. Mattle D, Sitsel O, Autzen HE, Meloni G, Gourdon P, Nissen P. On allosteric
956 modulation of P-type Cu(+)-ATPases. *J Mol Biol*. 2013;425(13):2299-2308.
- 957 72. Biemans-Oldehinkel E, Doeven MK, Poolman B. ABC transporter architecture and
958 regulatory roles of accessory domains. *FEBS Lett*. 2006;580(4):1023-1035.
- 959 73. Sikkema HR, van den Noort M, Rheinberger J, et al. Gating by ionic strength and safety
960 check by cyclic-di-AMP in the ABC transporter OpuA. *Sci Adv*. 2020;6(47).
- 961 74. Hwang TC, Braakman I, van der Sluijs P, Callebaut I. Structure basis of CFTR folding,
962 function and pharmacology. *J Cyst Fibros*. 2023;22 Suppl 1:S5-s11.
- 963 75. Coleman JA, Quazi F, Molday RS. Mammalian P4-ATPases and ABC transporters and
964 their role in phospholipid transport. *Biochim Biophys Acta*. 2013;1831(3):555-574.
- 965 76. Lyons JA, Timcenko M, Dieudonné T, Lenoir G, Nissen P. P4-ATPases: how an old
966 dog learnt new tricks - structure and mechanism of lipid flippases. *Curr Opin Struct*
967 *Biol*. 2020;63:65-73.
- 968 77. Corbalan-Garcia S, Gómez-Fernández JC. Signaling through C2 domains: more than
969 one lipid target. *Biochim Biophys Acta*. 2014;1838(6):1536-1547.
- 970 78. Lemmon MA. Membrane recognition by phospholipid-binding domains. *Nat Rev Mol*
971 *Cell Biol*. 2008;9(2):99-111.
- 972 79. Malcova I, Bumba L, Uljanic F, Kuzmenko D, Nedomova J, Kamanova J. Lipid binding
973 by the N-terminal motif mediates plasma membrane localization of Bordetella effector
974 protein BteA. *J Biol Chem*. 2021;296:100607.
- 975 80. Varela-Chavez C, Blondel A, Popoff MR. Bacterial intracellularly active toxins:
976 Membrane localisation of the active domain. *Cell Microbiol*. 2020;22(7):e13213.
- 977 81. López-Lara IM, Geiger O. Bacterial lipid diversity. *Biochim Biophys Acta Mol Cell Biol*
978 *Lipids*. 2017;1862(11):1287-1299.

- 979 82. Petroutsos D, Amiar S, Abida H, et al. Evolution of galactoglycerolipid biosynthetic
980 pathways--from cyanobacteria to primary plastids and from primary to secondary
981 plastids. *Prog Lipid Res.* 2014;54:68-85.
- 982 83. Wada H, Murata N. Membrane lipids in cyanobacteria. In: Siegenthaler P-A, Murata N,
983 eds. *Lipids in photosynthesis.* Dordrecht, The Netherlands: Kluwer Academic
984 Publishers; 1998:65-81.
- 985 84. Boudière L, Michaud M, Petroutsos D, et al. Glycerolipids in photosynthesis:
986 composition, synthesis and trafficking. *Biochim Biophys Acta.* 2014;1837(4):470-480.
- 987 85. Wada H, Murata N. The essential role of phosphatidylglycerol in photosynthesis.
988 *Photosynth Res.* 2007;92(2):205-215.
- 989 86. Kóbori TO, Uzumaki T, Kis M, et al. Phosphatidylglycerol is implicated in divisome
990 formation and metabolic processes of cyanobacteria. *J Plant Physiol.* 2018;223:96-104.
- 991 87. Terrapon N, Weiner J, Grath S, Moore AD, Bornberg-Bauer E. Rapid similarity search
992 of proteins using alignments of domain arrangements. *Bioinformatics.* 2014;30(2):274-
993 281.
- 994 88. Buchan DWA, Jones DT. Learning a functional grammar of protein domains using
995 natural language word embedding techniques. *Proteins.* 2020;88(4):616-624.

996

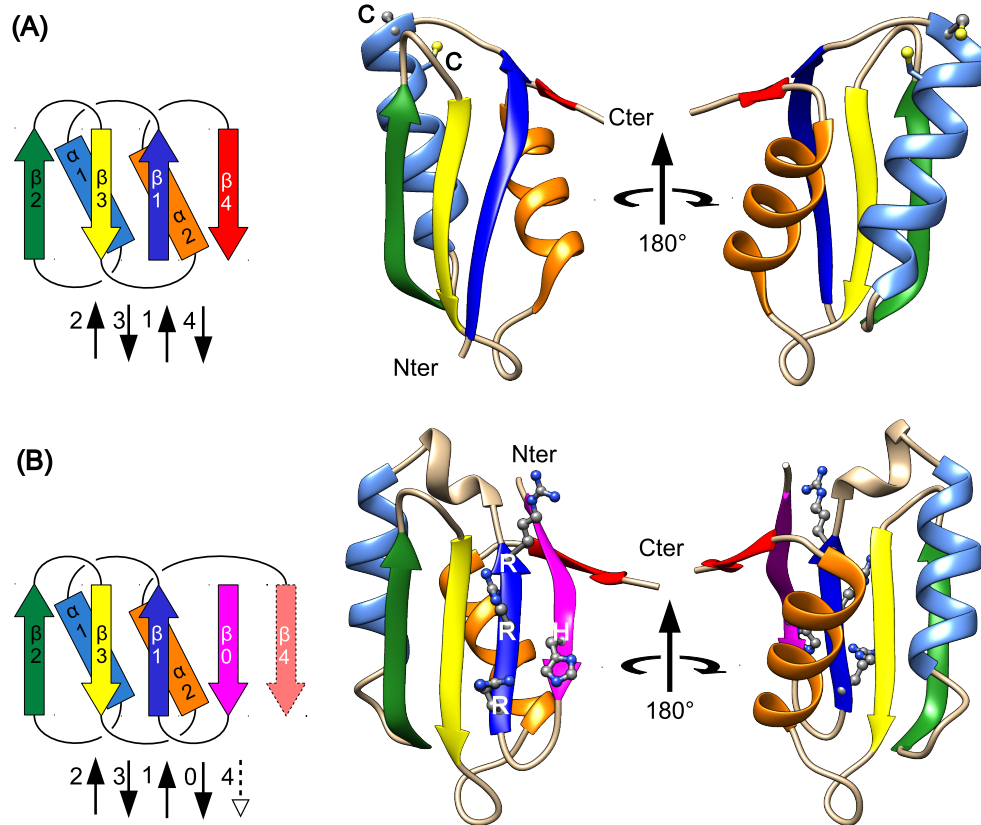


Figure 1: Topology diagrams and ribbon representations of the 3D structures of (A) an HMA domain (Human CopZ; experimental 3D structure - PDB 2QIF); (B) a CoBaHMA domain (*S. calicopolaris* calcyanin; AlphaFold2 model). Regular secondary structures are colored rainbow, from the N-terminus (blue) to the C-terminus (red), with the exception of the additional strand β_0 , specific of the CoBaHMA domain (pink). Conserved amino acids, specific to the two families, are highlighted in a ball-and-stick representation.

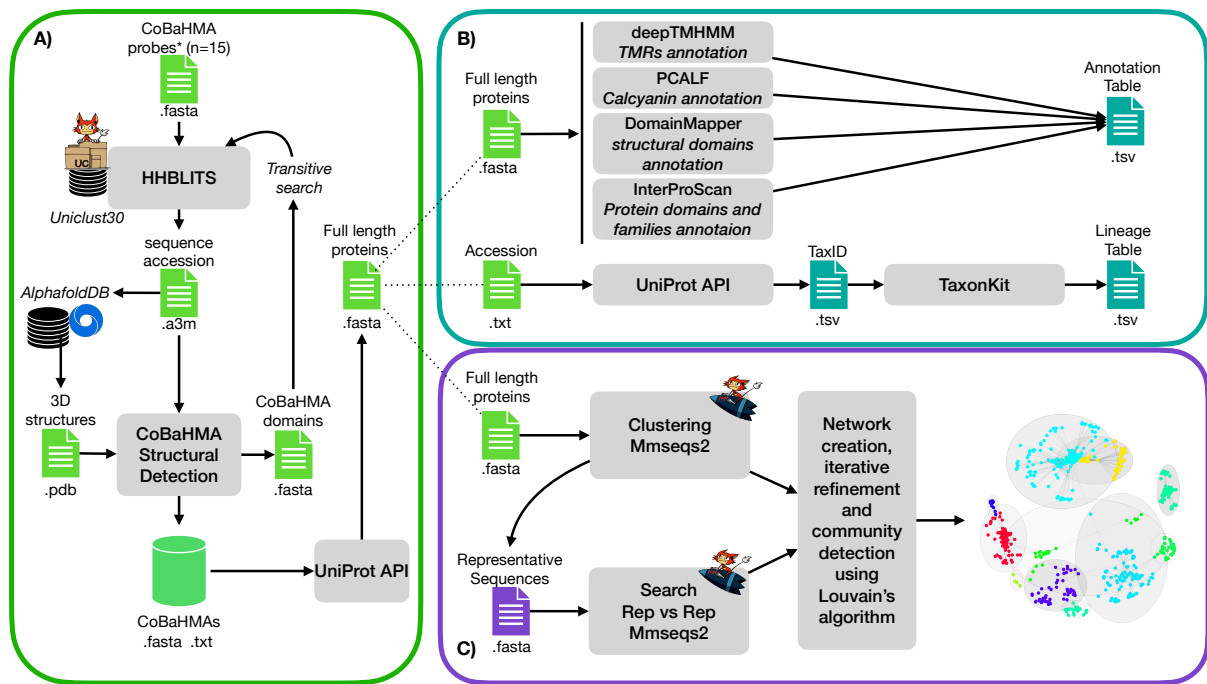


Figure 2: Workflow of the current analysis. (A) The 15 sequences of CoBaHMA domains were used as probes for sequence similarity searches with HHblits against the Uniclust30 database. The 3D structure models of the sequences that were identified by HHblits were considered in order to assess the presence of strand β_0 , a feature allowing to distinguish true CoBaHMA domains. The new oBaHMA domain sequences were then used as additional probes with HHblits to perform transitive searches. (B) Full length proteins with at least one CoBaHMA domain were annotated relative to transmembrane regions, calcyanin signature, structural domains and domains families using deepTMHMM, PCALF, DomainMapper and InterProScan, respectively. In addition, the taxonomy of all sequences is retrieved through the UniProt API. (C) Sequences were clustered using mmseqs2 and representative sequences were searched against themselves. Resulting alignment scores were used to build a similarity network. The network was refined iteratively to ensure that all sequence within a community share a similar length with a maximal amplitude of 50 residues. Finally, nodes and edges were rendered using the edge-rendering and weighted spring embedded layout from cytoscape.

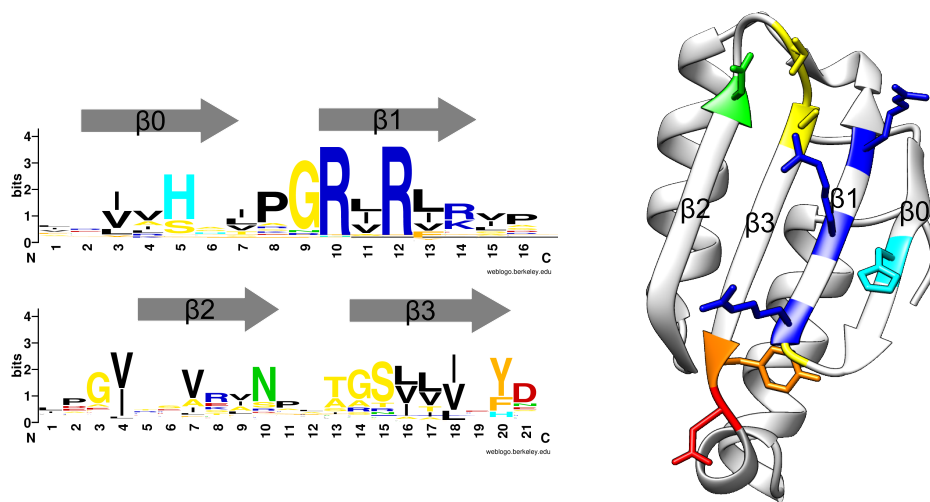


Figure 3: Logo of the amino acid conservation in the CoBaHMA family sequences. Amino acids are colored according to their properties (Black: Apolar, Green: Polar, Yellow: Small, Orange: Aromatic, Blue: Basic, Red: Acid, Cyan: Histidine). The most conserved polar amino acids are displayed on the CoBaHMA AlphaFold2 3D structure model (UniProt A0A545SE61).

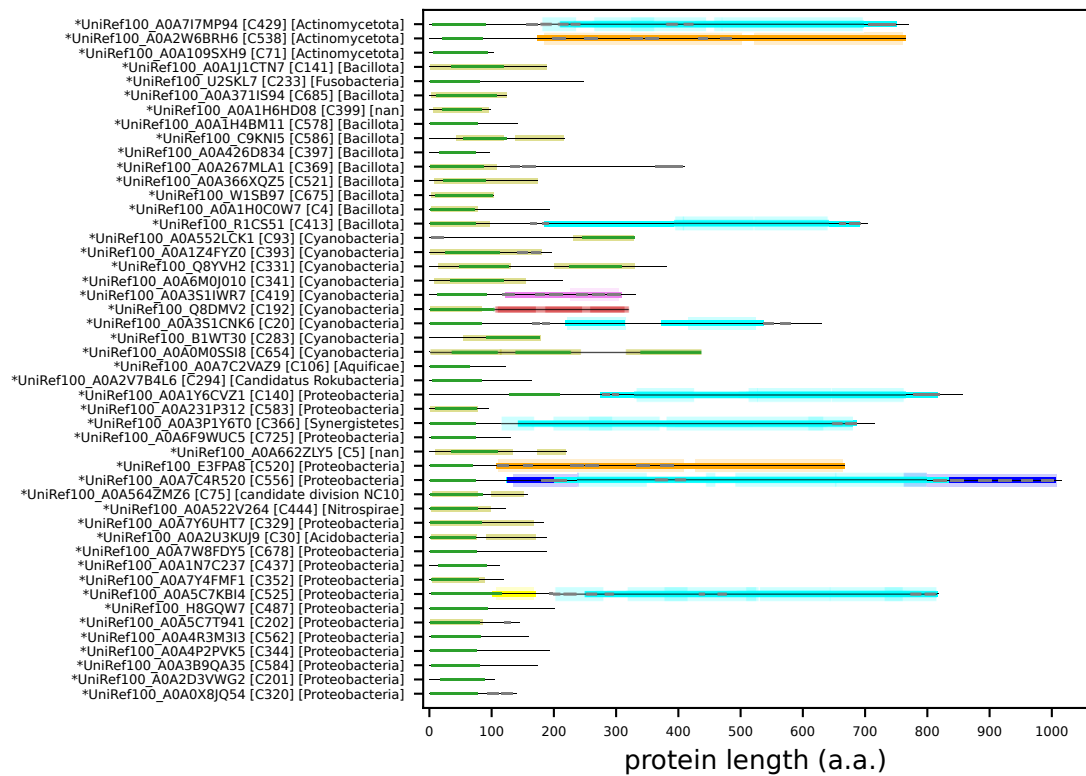


Figure 5: Modular organization of the full-length CoBaHMA domain proteins. Each line illustrates the representative sequence of a large community (the vertical order follows the horizontal order in Figure 4). Line labels represent the accession of the representative sequence, the community number and the dominant phylum. Functional annotations are indicated along the sequence by shaded areas. Each functional category of domain is highlighted by the following color code: CoBaHMA (green), HMA_2 (lime), HMA (yellow), P-type (cyan) and Serca (darkblue), ABC (orange), Calcyanin Gly-Zip (red), PAP2 (magenta). Membrane regions as identified by deepTMHMM are indicated by gray areas.

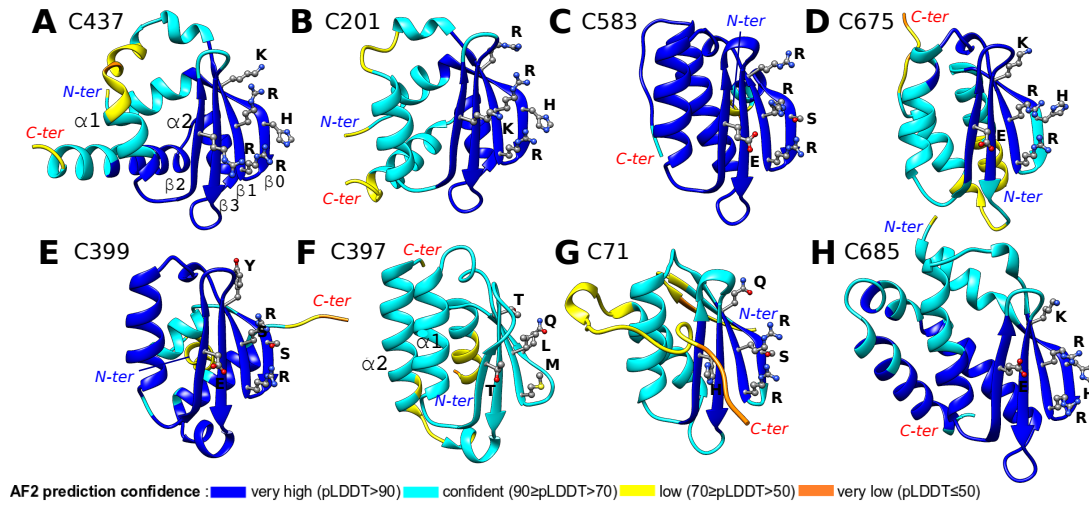


Figure 6: Single CoBaHMA proteins.

AF2 3D structure models of the representative proteins from the eight communities of single CoBaHMA domain proteins (ribbon representations), colored according to the pLDDT values. The conserved amino acids are shown as ball-and-sticks. Core α -helices and β -strands are only labeled for the first communities, with the exception of C399 (panel E), in which the extra-N-terminal helix takes the place of helix α 2. Proteins are referenced with their UniProt accession numbers: A) C437: A0A1N7C237, B) C201: A0A2D3VWG2, C) C583: A0A231P312, D) C675: W1SB97, E) C399: F5RIY2, F) C397: A0A426D834, G) C71: A0A109SXH9, H) C685: A0A371IS94. The multiple sequence alignments of the communities, together with the AF2 predicted secondary structures are provided in Supplementary Data 3.

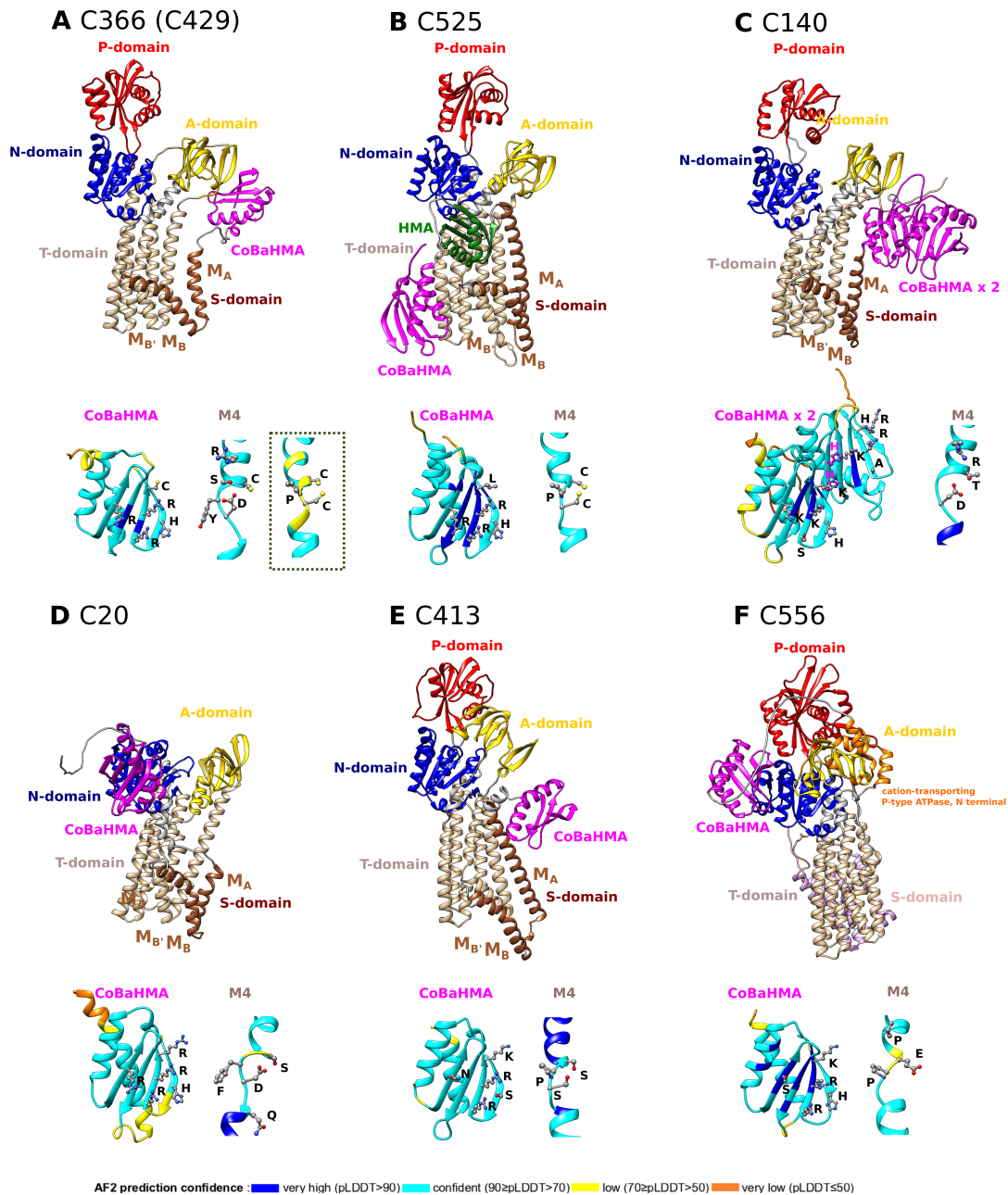


Figure 7: P-type ATPases. AF2 3D structure models of the representative proteins from six P-type ATPase communities (ribbon representations), colored according to modular organization (top). Domains are designated as in {Palmgren, 2011 #11}. At the bottom are shown the CoBaHMA domains and the M4 transmembrane helices, colored according to the pLDDT values and with the conserved amino acids shown as ball-and-sticks. Proteins are referenced with their UniProt accession numbers: A) C366: A0A3P1Y6T0 (in the dotted box is shown the M4 3D structure AF2 model of D8F5K4 (C429, full-length 3D structure not shown), B) C525: A0A5C7KBI4, C) C140: A0A1Y6CVZ1, D) C20: A0A3S1CNK6, E) C413: R1CS51, F) C556: A0A7C4R520. C429 (A0A7I7MP94) is not shown, as it shares the same domain architecture as C366 (S-P-C motif in M4). The multiple sequence alignments of the communities, together with the AF2 predicted secondary structures are provided in Supplementary Data 3.

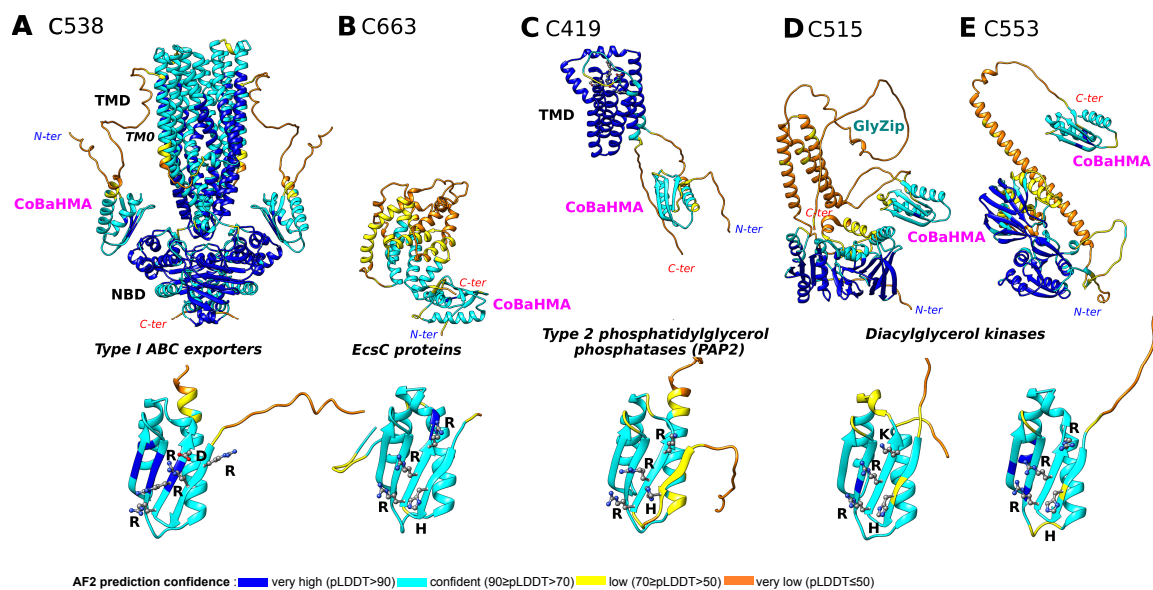


Figure 8: ABC exporters - EcsC proteins - PAP2 - DAGK. AF2 3D structure models of the representative proteins from communities including ABC exporters, EcsC proteins, type 2 phosphatidylglycerol phosphatases (PAP2) and diacylglycerolkinases (DAGK), colored according to the pLDDT values. CoBaHMA domains are shown at the bottom, with conserved amino acids shown as ball-and-sticks. Proteins are referenced with their UniProt accession numbers: A) C538: A0A2W6BRH6, B) C663: A0A552EVV8, C) C419: A0A3S1IWR7, D) C515: A0A0S3UCN3, E) C553: A0A841V906). The 3D structures of the ABC exporter dimer model (A) was built after superimposition of the AF2 model single chain on the experimental 3D structure of TM287/TM288 (best hit in a HH-PRED search, respecting the distance between the two swapped TMs and the TMD core). The multiple sequence alignments of the communities, together with the AF2 predicted secondary structures are provided in Supplementary Data 3.

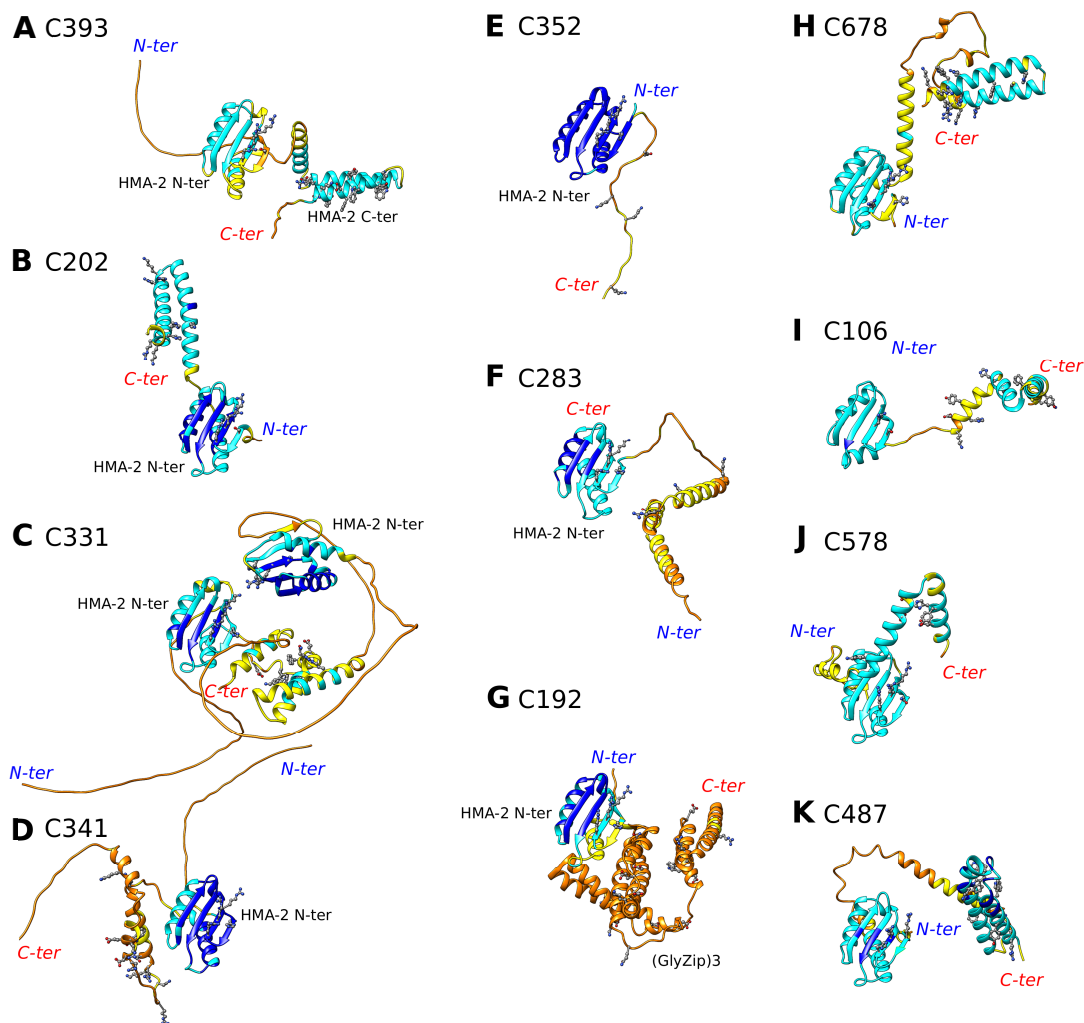


Figure 9: CoBaHMA domains associated with helical hairpins. AF2 3D structure models of the representative proteins from communities of CoBaHMA domains associated with helical hairpins (ribbon representations), colored according to the pLDDT values. CoBaHMA domains are shown in the same orientation, with those matching the HMA-2 N-ter profile (Pfam 19991) highlighted with a label. Other profile-matching domains (HMA-2 C-ter, GlyZip) are also labeled. Proteins are referenced with their UniProt accession numbers: A) C393: A0A1Z4FYZ0, B) C202: A0A5C7T941, C) C331: Q8YVH2, D) C341: A0A6M0J010, E) C352: A0A7Y4FMF1, F) C283: B1WT30, G) C192: Q8DMV2, H) C678: A0A7W8FDY5, I) C106: A0A7C2VAZ9, J) C578: A0A1H4BM11, K) C487: H8GQW7. The multiple sequence alignments of the communities, together with the AF2 predicted secondary structures is provided in Supplementary Data 3.

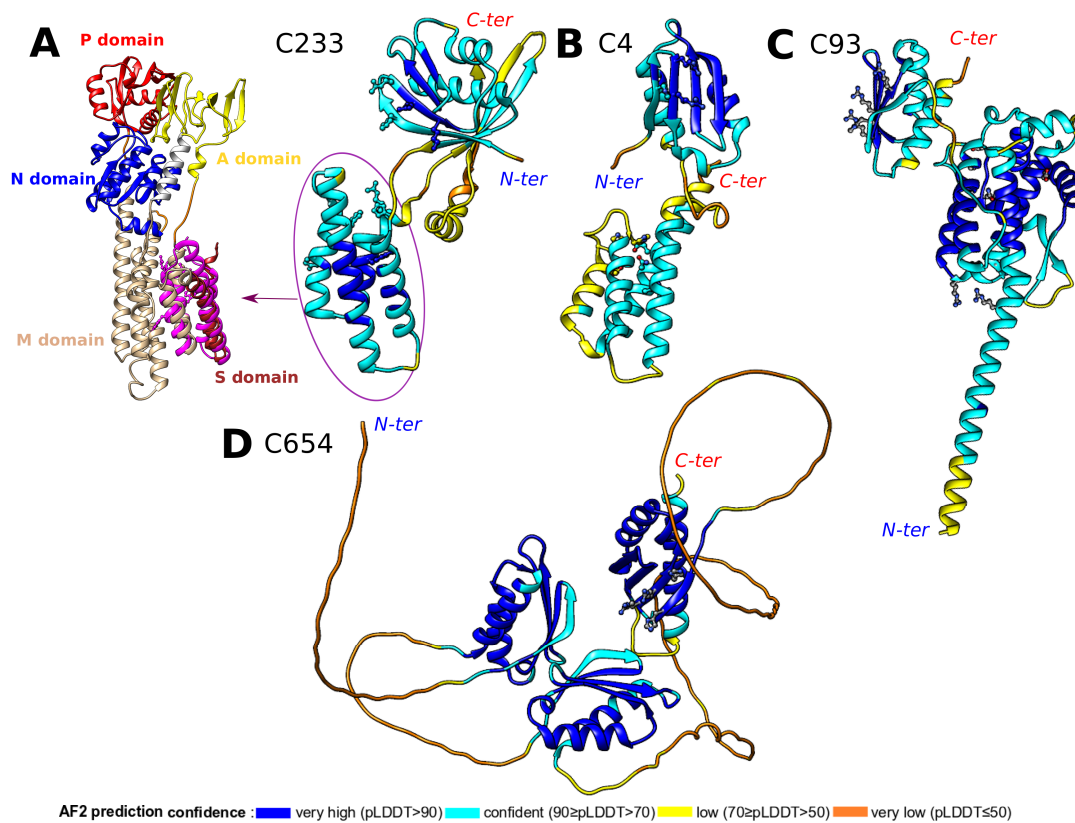


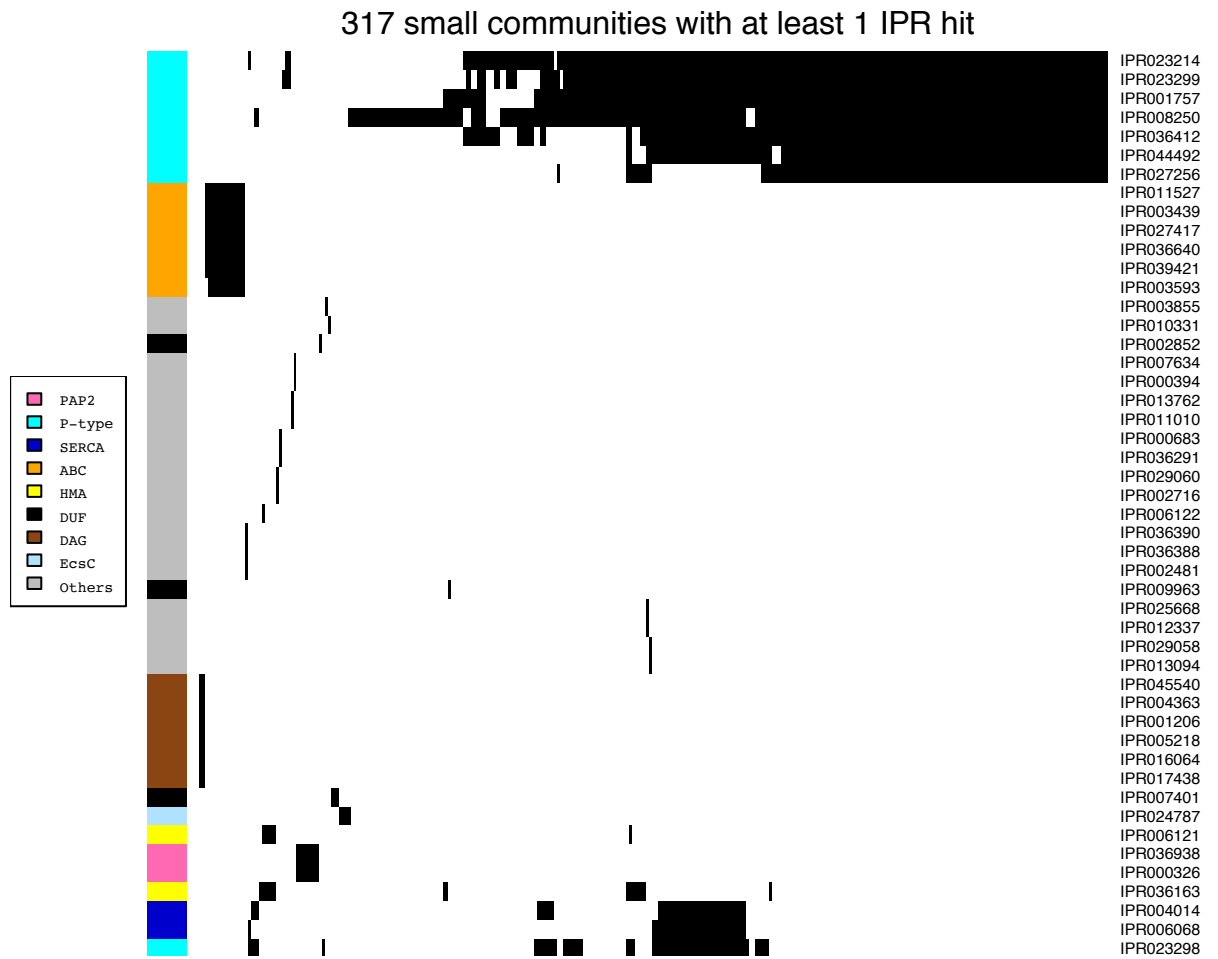
Figure 10: CoBaHMA domains with unknown regions. AF2 3D structure models of the representative proteins from communities of CoBaHMA domains associated with unknown regions (ribbon representations), colored according to the pLDDT values. Proteins are referenced with their UniProt accession numbers: A) C233: U2SKL7, the CoBaHMA domain is followed by an helical bundle, which superimposed with the MA/MB/M1/M2 bundle of P_{1B}-type ATPases (left), B) C4: A0A1H0C0W7, the CoBaHMA domain is followed by an helical bundle, with no striking similarities with any known 3D structures, C) C93: A0A552LCK1, the CoBaHMA domain is preceded by an all-alpha domain, with no striking similarities with any known 3D structures, D) C654: A0A0M0SSI8, three CoBaHMA domains are separated from each other by disordered linkers. The multiple sequence alignments of the communities, together with the AF2 predicted secondary structures is provided in Supplementary Data 3.

Supplementary Table 1:

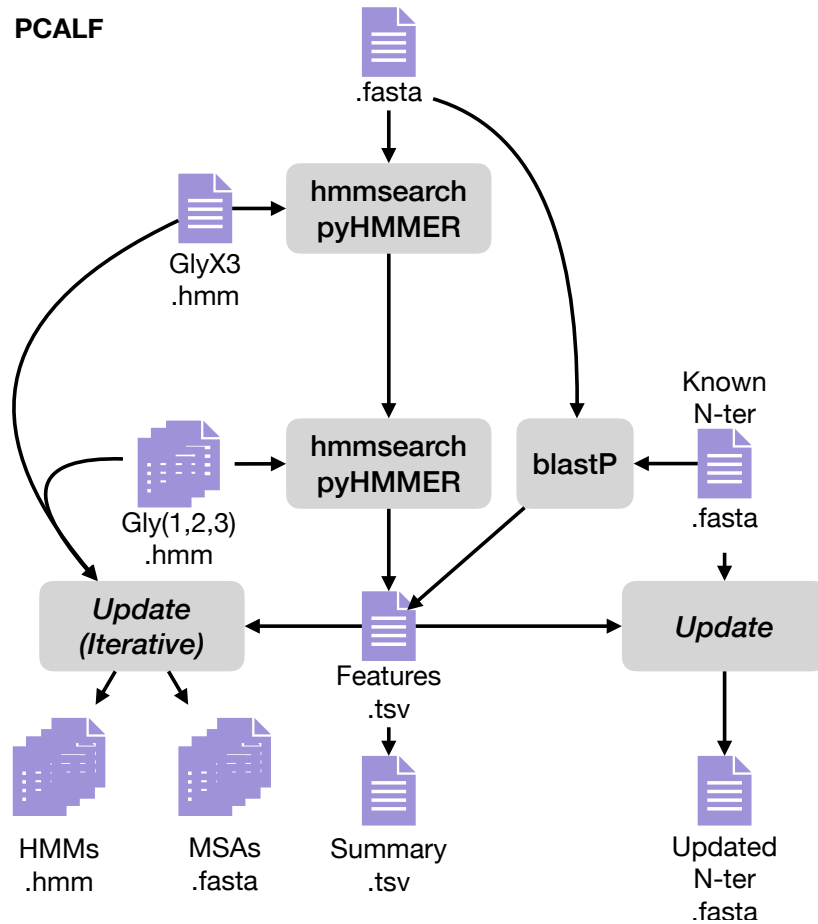
A. Annotations in the 48 most populated communities ($n \geq 5$) using domainMapper and InterProscan. The numbers of sequences which are detected are indicated, with * corresponding to an annotation of the representative sequences of the communities.

| E-value < 1.0e-10 | | | | C20 (n=7) | C556 (n=22) C655 (n=1) C430 (n=1) | C413 | C140 (n=18) | C366 (n=115) | C525 (n=36) C566 (n=1) | C444 (n=100) | C520 (n=6) | C538 (n=11) | C419 (n=22) C617 (n=1) | |
|-------------------|--------------|---------------|---|---|--|------|----------------|-----------------|---------------------------------|-----------------|---------------|----------------|---------------------------------|-----|
| DomainMapper | | 2006.1.1.5 | HAD-like | | 1 | | | | | | | | | |
| | | 5073.1.1.14 | Calcium ATPase transmembrane domain M | | 24* | | | | | | | | | |
| | | 267.1.1.3 | Metal cation-transporting ATPase, ATP-binding domain | | 24* | | | | | | | | | |
| | | 10.13.1.2 | Calcium ATPase, transduction domain A | 7* | 23* | 5 | 18* | 115* | 28* | | | | | |
| | | 2006.1.1.42 | HAD-like | 3* | 21* | 11* | 18* | 115* | 37* | | | | | |
| | | 267.1.1.5 | Metal cation-transporting ATPase, ATP-binding domain | | | 10* | 18* | 115* | 37* | | | | | |
| | | 5073.1.2.3 | Copper efflux ATPase transmembrane domain | | | | | 79* | | | | | | |
| | | 304.3.1.7 | HMA, heavy metal-associated domain | | | | | | 35* | 4 | | | | |
| | | 1075.4.1.1 | Type I ABC exporter transmembrane domain fold | | | | | | | | | 6* | 11* | |
| | | 2004.1.1.1405 | P-loop containing nucleoside triphosphate hydrolases | | | | | | | | | 6* | 11* | |
| | | 180.1.1.1 | Acid phosphatase/Vanadium-dependent haloperoxidase | | | | | | | | | | | 23* |
| | InterProscan | | IPR006068 | Cation-transporting P-type ATPase, C-terminal | | 23* | | | | | | | | |
| | | IPR004014 | Cation-transporting P-type ATPase, N-terminal | | 23* | | | | | | | | | |
| | | IPR023298 | P-type ATPase, transmembrane domain superfamily | | 24* | | | 2 | 29* | | | | | |
| | | IPR027256 | P-type ATPase, subfamily IB | | | | 18* | 113* | 37* | | | | | |
| | | IPR023299 | P-type ATPase, cytoplasmic domain N | | 24* | 11* | 18* | 115* | 37* | | | | | |
| | | IPR001757 | P-type ATPase | | 24* | 11* | 18* | 115* | 37* | | | | | |
| HAD | | | IPR023214 | HAD superfamily | 4* | 24* | 11* | 18* | 115* | 37* | | | | |
| | | | IPR036412 | HAD-like superfamily | 2 | 24* | 11* | 18* | 115* | 37* | | | | |
| | | | IPR044492 | P-type ATPase, haloacid dehalogenase domain | | 24* | 5 | 17* | 108* | 35* | | | | |
| | | IPR008250 | P-type ATPase, A domain superfamily | 7* | 24* | 4 | 18* | 115* | 37* | | | | | |
| | | IPR036163 | Heavy metal-associated domain superfamily | | | | | | 36* | | | | | |
| | | IPR006121 | Heavy metal-associated domain, HMA | | | | | | 12* | | | | | |
| | | IPR000326 | Phosphatidic acid phosphatase type 2/haloperoxidase | | | | | | | | | | | 23* |
| | | IPR036938 | Phosphatidic acid phosphatase type 2/haloperoxidase superfamily | | | | | | | | | | | 23* |
| | | IPR039421 | Type 1 protein exporter | | | | | | | | | 5* | 11* | |
| ATP-binding | | | IPR027417 | P-loop containing nucleoside triphosphate hydrolase | | | | | | | | 6* | 11* | |
| | | | IPR003439 | ABC transporter-like, ATP-binding domain | | | | | | | | 6* | 11* | |
| | | | IPR003593 | AAA+ ATPase domain | | | | | | | | 6* | 11* | |
| TM | | IPR036640 | ABC transporter type 1, transmembrane domain superfamily | | | | | | | | 6* | 11* | | |
| | | IPR011527 | ABC transporter type 1, transmembrane domain | | | | | | | | 6* | 11* | | |

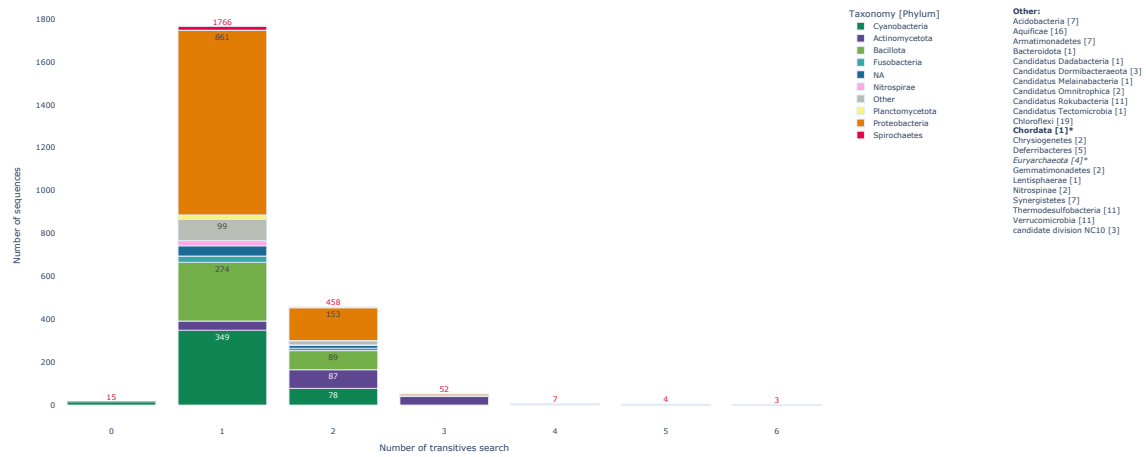
B. InterProscan annotations in the 317 small communities (n<5). DUF stands for Domain of Unknown Function.



PCALF



Supplementary Data 1: pCALF workflow, a tool to annotate calcyanins. First, HMMs profiles specific of the glycine zipper triplication were searched against a set of proteins using pyHMMER. Then, a set of domains known to be related to calcyanins (Y-type, Z-type, X-type and CoBaHMA) were searched using blastp. Proteins with a complete glycine zipper triplication were classified as calcyanins and their features were respectively re added to MSAs and HMMs.



Supplementary Data 2: Number of sequences by phylum recruited during the transitive search. Phyla represented by less than 20 sequences are labeled as others. Sequences from the Chordata and Euryarchaeota phyla were found during the first transitive search. For each search, the total number of sequences is indicated in red.

Supplementary Data 3:

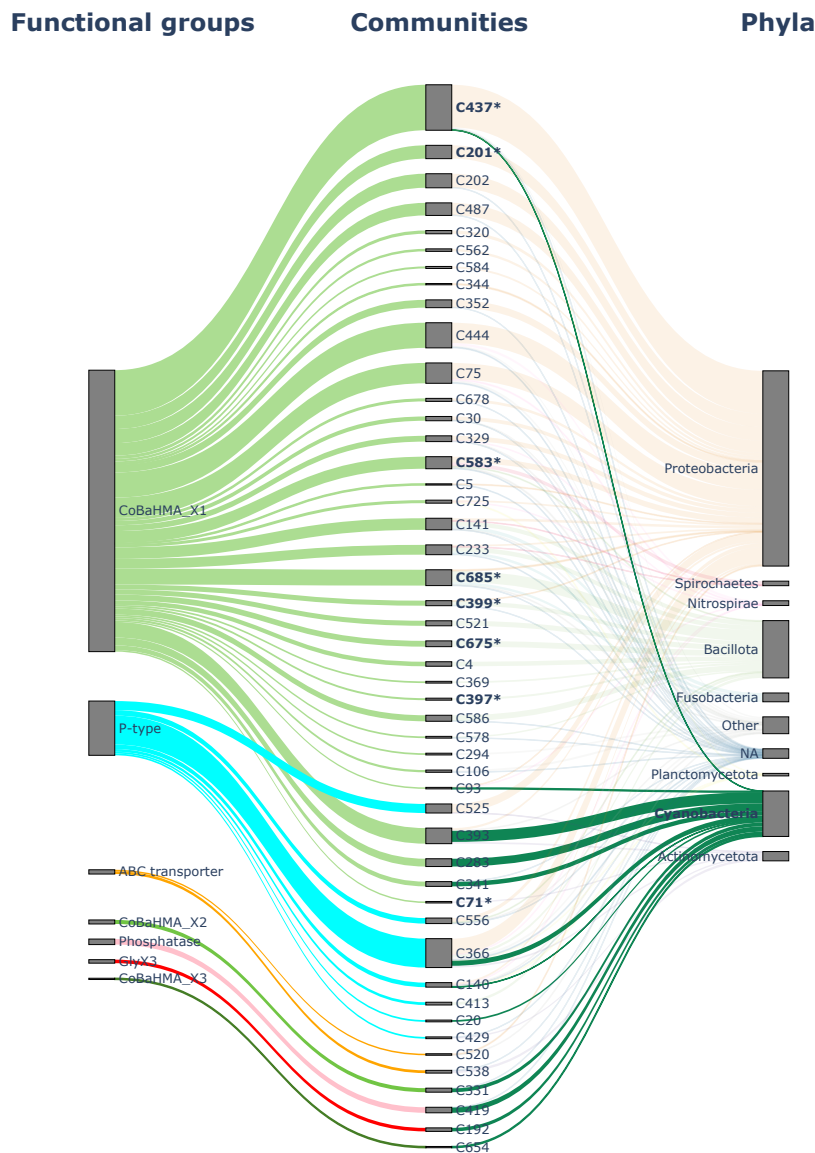
Multiple sequence alignments (MSA) of the sequences of the most populated communities, build using MAFFT and rendered using EspPript3. The first two lines report the 2D structures of the AF2 model of the representative sequence of the community, and its amino acid sequence. The initial MSA (fasta format) is also provided, together with the coordinates of the 3D structure model (pdb format) and the HCA plot (postscript file) of the representative sequence of the considered community.

<https://dropsu.sorbonne-universite.fr/s/PatmeKee7GGbkgP>

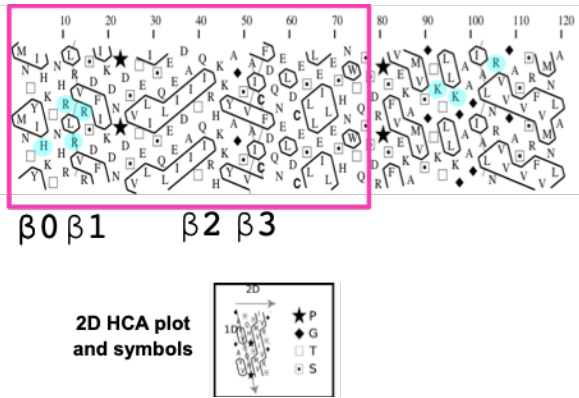
Supplementary Data 4:

Annotations by communities: Modular organization of the full-length CoBaHMA domain proteins within the 48 large community (Functional annotations are indicated along the sequence by shaded areas. Each functional category of domain is highlighted by the following color code: CoBaHMA (green), HMA_2 (lime), HMA (yellow), P-type (cyan) and Serca (darkblue), ABC (orange), Calcyanin Gly-Zip (red), PAP2 (magenta). Membrane regions as identified by deepTMHMM are indicated by gray areas.

<https://dropsu.sorbonne-universite.fr/s/J7kMXsFY44TpCBK>

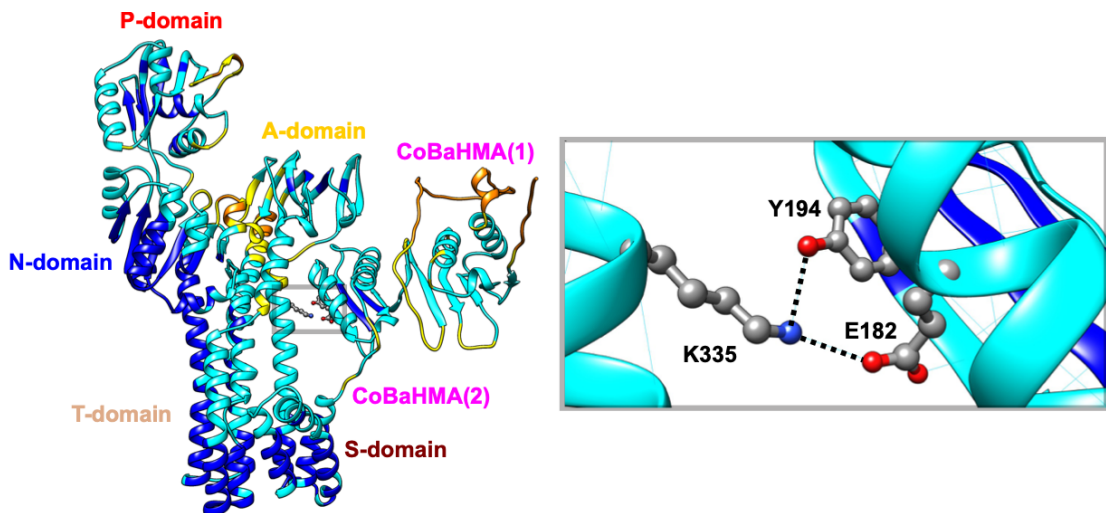


Supplementary Data 5: Alluvial plot of the distribution of sequences belonging to large communities from their great functional group to their taxonomy. There are three sets of nodes from left to right corresponding respectively to: great functional groups, communities (with at least 5 sequences) and taxonomy (at the phylum level). The height of a node represents the total number of sequences in the considered category. The width of a link represents the number of sequences belonging to the starting node and the ending node. The color of links between great functional groups and communities and links between communities and phyla correspond to the colors defined in Figure 5 and 4/6, respectively. Communities associated with Single CoBaHMA proteins are marked with an asterisk.



Supplementary Data 7:

HCA plot of the representative sequence of the C352 community (UniProt A0A7Y4FMF1). The sequence is shown on an duplicated α -helical net, on which strong hydrophobic amino acids (V,I,L,M,F,Y,W) are contoured. They form hydrophobic clusters, which have been shown to mainly correspond to regular secondary structures. Symbols and the way to read the sequence and secondary structures are indicated in the inset. Although the C-terminus is predicted as disordered on the AF2 model, it contains large hydrophobic clusters, which are indicative of the presence of regular secondary structures.



Supplementary Data 8:

Example of the interaction of a CoBaHMA domain with another domain of a PIB-type ATPase (community C140, UniProt A0A1U7CQU6). The inset illustrates the bond network involving two amino acids of the CoBaHMA domain

Abstract :

Biomineralisation is all the processes that lead to the formation of minerals by living beings. In 2012, a new biomineralization phenotype has been described in cyanobacteria, characterized by the presence of amorphous alkaline-earth carbonate inclusions inside the cells.

A comparative genomic analysis revealed that this intracellular biomineralisation phenotype is linked to the presence of one gene, unknown at the time, which has been called *ccyA*. It codes for one protein called calcyanin. Calcyanin has 4 variants, that share the same C-terminus domain ((GlyZip)₃), but which differ in their N-terminus domain (CoBaHMA, X, Y or Z). None of these 5 domains has already been described in the literature.

The goal of this PhD was to characterize the 3D structure of *Synechococcus calcipolaris*'s calcyanin, which has a CoBaHMA domain, by combining bioinformatics and experimental approaches, in order to make hypothesis regarding its role.

Through sequence analysis and 3D structure modeling, we showed that the CoBaHMA domain belongs to the “ferredoxin-like” fold, typical of the superfamily HMA (“Heavy Metal Associated”), and sets itself as a new family in it, characterized by conserved basic amino acids and an additional β strand (1). We have performed sequence similarity searches, refined with the structural information of the 3D structure models. This way, we showed that the CoBaHMA domain can be found on several different protein architectures, in various taxa. It has an independent domain, or in conjunction with other domains, especially membrane systems which, among others, allow transports of substrates through the membrane (P_{1B}-type ATPases, ABC exporters) or new families with unknown functions. These results lead us to formulate hypotheses regarding the CoBaHMA domain function (2).

We also proposed a robust model for the individual glycine zipper from which the name of the C-terminus domain (GlyZip)₃ of calcyanins comes from. These glycine zippers have a structure of a compact hairpin made of two helices, which is akin to the ones of transmembrane proteins that form pore. However, we were not able to model satisfyingly their assembly nor their possible interactions with the CoBaHMA domain, emphasizing the importance of studying the protein experimentally.

We successfully expressed calcyanin in *Escherichia coli*, and purified it. However the protein proved to be quite unstable, with a propensity to form a great diversity of objects with different sizes. A limited proteolysis experiment revealed the existence of a protease-resistant fragment of calcyanin, which encompasses the CoBaHMA domain and the first glycine zipper of the (GlyZip)₃ domain. We expressed and purified this fragment, fused to MBP (« Maltose Binding Protein »). The fragment forms only one object in solution, but is prone to precipitation once separated from MBP. Yet we have successfully obtained crystals of this fragment, which pave the way to solve its experimental 3D structure.

Calcyanin is a difficult protein to work with, both experimentally and by bioinformatics. But we managed to model and characterize several of its fragments. From that, we inferred relevant information on calcyanin. More specifically, we highlighted a new family of domains, CoBaHMA, which presence on other protein architectures opens up new hints to understand its function and evolution.

(1) Benzerara *et al.*, 2022 14(3): evav026. doi :10.1093/gbe/evac026.

(2) Gaschignard *et al.*, En préparation.

Résumé :

La biominéralisation est l'ensemble des processus qui aboutissent à la formation de minéraux par des êtres vivants. En 2012, un nouveau phénotype de biominéralisation a été décrit chez les cyanobactéries, caractérisé par la présence de précipités amorphes de carbonate d'alcalino-terreux au sein des cellules. Une analyse de génomique comparative a montré que ce phénotype de biominéralisation intracellulaire était associé à la présence d'un gène, encore non caractérisé, qui a été nommé *ccyA*, et qui code pour une protéine nommée calcyanine. La calcyanine a 4 variants qui partagent un même domaine C-terminal ((GlyZip)₃), mais qui diffèrent au niveau de leur domaine N-terminal (CoBaHMA, X, Y ou Z). Aucun de ces 5 domaines n'est encore décrit dans la littérature.

L'objectif de cette thèse a été de caractériser la structure 3D de la calcyanine de *Synechococcus calcipolaris*, variant à domaine CoBaHMA, en combinant des approches bioinformatiques et expérimentales, afin de progresser dans la connaissance de sa fonction.

En combinant un ensemble de méthodes d'analyse de séquences et de modélisation de 3D structures, nous avons montré que le domaine CoBaHMA adopte un repliement « ferredoxin-like » typique de la superfamille HMA (« Heavy-Metal Associated ») et forme au sein de celle-ci une nouvelle famille, caractérisée par la conservation d'acides aminés basiques et la présence d'un brin β additionnel (1). Nous avons ensuite réalisé des recherches avancées de similitudes de séquence, intégrant les informations dérivées des modèles de structure 3D. Nous avons ainsi montré que le domaine CoBaHMA est présent dans différentes organisations modulaires au sein de taxa variés. Il existe sous la forme de domaine unique, ou accompagné d'autres domaines, en particulier des systèmes membranaires qui permettent, entre autre, le transport de substrats au travers des membranes (P_{1B}-type ATPases, exporteurs ABC) ou qui constituent de nouvelles familles aux fonctions encore inconnues. Ces résultats permettent de proposer des hypothèses quant à la fonction moléculaire du domaine CoBaHMA et aux processus cellulaires dans lesquels il est impliqué (2).

Nous avons également proposé un modèle cohérent des trois motifs glycine zippers individuels qui donnent leur nom au domaine (GlyZip)₃ C-terminal des calcyanines. Ces glycine zippers forment une structure compacte de 2 hélices en épingle à cheveux, qui rappelle celle adoptée par des protéine transmembranaires formant des pores. Cependant, nous n'avons pu obtenir de modèle satisfaisant de l'assemblage des motifs glycine zipper (domaine (GlyZip)₃), ni de leur interaction potentielle avec le domaine CoBaHMA, renforçant ainsi l'intérêt d'une étude expérimentale.

Nous avons ainsi exprimé la calcyanine de *S.calcipolaris* chez *Escherichia coli*, et l'avons purifiée. Cependant cette protéine s'est avérée très peu stable, avec une propension à former un grand nombre d'objets de taille différentes en solution. Une expérience de protéolyse limitée a montré l'existence d'un fragment de la calcyanine résistant aux protéases, constitué du domaine CoBaHMA et du premier glycine zipper du domaine (GlyZip)₃. Nous avons donc exprimé et purifié ce fragment fusionné à la protéine MBP (« Maltose Binding Protein »), afin d'en augmenter la solubilité. Celui-ci ne forme qu'une espèce en solution, mais est très peu soluble après avoir été séparé de la MBP. Nous avons cependant réussi à obtenir des cristaux de ce fragment offrant des perspectives encourageantes pour résoudre sa structure expérimentale.

La calcyanine est donc une protéine difficile à étudier, tant par voie expérimentale que par modélisation. Néanmoins nous avons réussi à modéliser et caractériser plusieurs de ses fragments. Nous avons déduit de cette étude des informations pertinentes sur cette protéine, et plus particulièrement de son domaine CoBaHMA, dont la présence dans d'autres architectures protéiques ouvre des perspectives pour comprendre sa fonction et son évolution.

(1) Benzerara *et al.*, 2022 14(3): evav026. doi :10.1093/gbe/evac026.

(2) Gaschignard *et al.*, En préparation.