



**HAL**  
open science

**Exploration de la diversité des protéines à solénoïdes alpha, régulatrices de l'expression des gènes des organites dans les lignées eucaryotes photosynthétiques et étude de la dynamique conformationnelle des protéines à "Pentatricopeptide Repeats"**

Céline Cattelin

► **To cite this version:**

Céline Cattelin. Exploration de la diversité des protéines à solénoïdes alpha, régulatrices de l'expression des gènes des organites dans les lignées eucaryotes photosynthétiques et étude de la dynamique conformationnelle des protéines à "Pentatricopeptide Repeats". Biochimie, Biologie Moléculaire. Sorbonne Université, 2023. Français. NNT : 2023SORUS158 . tel-04164696

**HAL Id: tel-04164696**

**<https://theses.hal.science/tel-04164696v1>**

Submitted on 18 Jul 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.

THÈSE DE DOCTORAT  
SPÉCIALITÉ : BIOINFORMATIQUE

SORBONNE UNIVERSITÉ  
Ecole Doctorale Complexité du Vivant (ED 515)

Laboratoire de Biologie du Chloroplaste et Perception de la Lumière chez les  
Micro-algues, UMR7141

Laboratoire de Biochimie Théorique, UPR9080

INSTITUT DE BIOLOGIE PHYSICO-CHIMIQUE

**Exploration de la diversité des protéines à solénoïdes  
alpha, régulatrices de l'expression des gènes des  
organites dans les lignées eucaryotes  
photosynthétiques et étude de la dynamique  
conformationnelle des protéines à  
"Pentatricopeptide Repeats".**

par Céline CATTELIN

Dirigée par Ingrid LAFONTAINE et Charles H. ROBERT

Présentée et soutenue publiquement le 4 avril 2023

Devant un jury composé de :

Hédi SOULA	Professeur, Sorbonne Université	Président du jury
Catherine ETCHEBEST	Professeure, Université Paris Cité	Rapportrice
Kamel HAMMANI	Chargé de recherche, IBMP Strasbourg	Rapporteur
Julie MENETREY	Directrice de recherche, I2BC Gif-sur-Yvette	Examinatrice
Benoît CASTANDET	Maître de conférence, Université Paris Cité	Examinateur
Ingrid LAFONTAINE	Professeure, Sorbonne Université	Directrice de thèse
Charles H. ROBERT	Directeur de recherche, LBT Paris	Co-directeur de thèse







## Remerciements

J'aimerais tout d'abord remercier ma directrice de thèse Ingrid Lafontaine pour sa bienveillance, ses conseils, son honnêteté franche et sa patience. J'ai beaucoup apprécié ces années en ta compagnie dans ce grand bureau au sous-sol de l'institut ainsi que tes idées et ton enthousiasme durant nos discussions devant des résultats qui parfois me laissaient (peut être un peu trop) dubitative.

Je remercie évidemment Charles Robert, co-directeur de ma thèse avec qui j'ai eu le plaisir d'approfondir mes connaissances et mes compétences, tant en informatique en apprenant à me servir des clusters de calculs qu'en dynamique moléculaire, domaine dans lequel je n'étais pas spécialiste mais que j'ai pu largement découvrir grâce à tes explications précises, ta patience et ta gentillesse.

Merci à tous les deux de m'avoir accordé votre confiance pour ce projet. Bien que nous ayons rencontré des difficultés conséquentes durant notre travail, j'espère vous avoir donné satisfaction.

Je souhaite particulièrement remercier Angela Falciatore pour m'avoir accueillie dans son laboratoire ainsi que Francis-André Wollman pour les discussions que nous avons eu, son dynamisme et pour la relecture de ce manuscrit. Je suis heureuse d'avoir fait un bout de chemin au sein de ce laboratoire historique. De la même façon, j'aimerais remercier Marc Baaden notamment pour avoir soutenu mon utilisation des ressources du cluster de calculs du Laboratoire de Biochimie Théorique à l'IBPC.

Je remercie également les membres de mon jury : Catherine Etchebest et Kamel Hammani qui ont tous les deux acceptés d'être les rapporteurs de ce travail ainsi que Julie Ménétreay, Benoît Castandet et Hédi Soula d'avoir bien voulu faire partie de mon jury.

Merci aussi à Yves Choquet et Olivier Vallon pour les riches discussions que nous avons eu au début du projet à propos des protéines OPR.

Je dois aussi bien sûr remercier Rebecca Goulancourt et Alexis Astourian pour leur motivation, leur gentillesse et surtout leur excellent travail durant les stages de Master 2 qu'ils ont effectués avec Ingrid et moi. J'espère qu'ils en ont été satisfaits.

J'aimerais aussi remercier les membres de mon comité de suivi de thèse qui ont toujours été d'une bienveillance sans faille durant nos réunions : Etienne Delannoy, Sophie Pasek, Hédi Soula et Ludovic Sanguet.

Enfin, un grand merci à tous les membres du laboratoire avec lesquels j'ai passé un peu plus de trois années mémorables et notamment Clotilde, Matthieu, Amel, Sandrine, Raphaël, Eliora, Dany, Gwenaëlle, Alessandro, Carole, Marcio, Marcello, Erik, Katia, Alix, Benjamin, Domitille et tous les autres.

La présence et les encouragements de mes amis Ulysse et Julie ont aussi été des piliers sur lesquels j'ai toujours pu m'appuyer et je tiens à le faire remarquer. Je veux maintenant remercier Claire, mon amie et colocataire depuis maintenant cinq ans qui cuisine toujours de bons plats et qui a partagé tous les moments difficiles que j'ai pu rencontrer durant ces années, qui a relu ma thèse et qui m'a bien aidée à corriger les petites imprécisions que j'avais faites parfois.

Pour finir, je veux remercier ma famille et notamment mes parents, Marie-Hélène et Gilles ainsi que mon frère Benoît pour leur soutien indéfectible, mais aussi mon oncle et ma grand-mère pour toujours me sourire, croire en moi et m'encourager.

Je suis profondément reconnaissante des échanges (scientifiques ou non) que j'ai pu avoir avec vous tous durant cette période et qui m'ont apporté beaucoup de joie !

Merci encore à tous !





# Table des matières

## Remerciements

Avant-propos 1

## I Introduction 8

### 1 De la chimie organique à la biologie : l'organisation du premier "être vivant" 9

1.1 Première hypothèse : la "soupe" prébiotique ou le monde à ARN . . . 11

1.2 Seconde hypothèse : le monde "fer-souffre" de surface, ou l'être vivant à deux dimensions . . . . . 11

1.3 Autres hypothèses à propos de l'émergence de la vie . . . . . 12

1.3.1 La co-évolution hasardeuse . . . . . 12

1.3.2 La vie extraterrestre . . . . . 13

1.4 L'émergence de la vie : que conclure? . . . . . 13

### 2 La vie primitive à l'aube de l'évolution 14

### 3 LUCA : le dernier ancêtre commun universel et la dichotomie fondamentale de la biologie 15

3.1 Les organismes procaryotes . . . . . 16

3.1.1 Le règne des bactéries . . . . . 16

3.1.2 Le règne des Archaea . . . . . 18

3.2 Les organismes eucaryotes . . . . . 19

### 4 La variété des interactions biologiques entre espèces 20

### 5 Former des espèces en assemblant d'autres espèces : les endosymbioses 22

### 6 La théorie endosymbiotique : vers la formation des lignées eucaryotes 23

6.1 Une première endosymbiose primaire : l'origine de la mitochondrie . 24

6.2 La seconde endosymbiose primaire : l'apparition du chloroplaste . . 25

6.3 Les endosymbioses secondaires à l'origine d'autres algues . . . . . 27

6.4 D'autres endosymbioses à l'origine de la variété des plastes . . . . . 29

<b>7</b>	<b>Les endosymbioses apportent de nouvelles compétences</b>	<b>31</b>
7.1	La respiration . . . . .	31
7.2	La photosynthèse . . . . .	31
<b>8</b>	<b>Les impacts de l'évolution sur l'endosymbiose et les innovations post-endosymbiotiques</b>	<b>33</b>
8.1	Transferts et pertes de gènes . . . . .	33
8.2	La mosaïque génétique des complexes respiratoires et photosynthétiques . . . . .	33
8.3	La machinerie d'import des protéines . . . . .	34
8.4	Une ploïdie variable au sein des cellules . . . . .	35
8.5	Le contrôle par épistasie de la synthèse . . . . .	35
8.6	Les acteurs nucléaires de la régulation des génomes des organites . . . . .	37
<b>9</b>	<b>À la découverte de grandes familles d'OTAFs</b>	<b>37</b>
9.1	PPR : PentatrigoPeptide Repeat . . . . .	38
9.1.1	Les deux principales classes de protéines PPR . . . . .	39
9.1.2	La variété des motifs composant les protéines PPR . . . . .	41
9.2	OPR : OctatrigoPeptide Repeat . . . . .	43
<b>10</b>	<b>Structure type des protéines à solénoïde alpha</b>	<b>45</b>
<b>11</b>	<b>La diversité des protéines à solénoïde alpha</b>	<b>47</b>
<b>12</b>	<b>État de l'art des approches précédemment développées pour identifier les familles de protéines à solénoïde alpha</b>	<b>52</b>
12.1	Identifier des protéines à solénoïde alpha via leur répétitions et leur structure . . . . .	52
12.2	Identifier des protéines PPR et TPR . . . . .	58
12.3	Identifier des protéines OPR . . . . .	60
<b>13</b>	<b>Objectifs</b>	<b>60</b>
13.1	Évolution et diversité des protéines OTAF à solénoïde alpha : et en dehors des espèces modèles? . . . . .	60
13.2	Saisir la complexité de l'interaction protéine-ARNm : quelle spécificité pour la liaison à l'ARN? . . . . .	62
<b>II</b>	<b>Exploration de la diversité des protéines à solénoïde alpha dans les organismes photosynthétiques</b>	<b>65</b>
<b>1</b>	<b>Introduction</b>	<b>66</b>

1.1	L'évolution des protéines OTAF . . . . .	66
1.2	Méthode : identification des protéines à solénoïde alpha . . . . .	68
<b>2</b>	<b>Article</b>	<b>70</b>
2.1	A propos de l'approche développée . . . . .	70
2.2	Qualité des protéomes utilisés . . . . .	71
2.3	Article 1 . . . . .	72
<b>3</b>	<b>Discussion</b>	<b>115</b>
3.1	Améliorer les capacités de détection de notre approche . . . . .	115
3.2	Appliquer nos méthodes à d'autres espèces photosynthétiques : le cas des diatomées . . . . .	116
<b>4</b>	<b>Conclusion et perspectives</b>	<b>121</b>

### III Étude de la dynamique conformationnelle des protéines à "Pentatricopeptide Repeat" 123

<b>1</b>	<b>Premier chapitre : présentation des simulations de dynamique moléculaire et du principe des analyses réalisées</b>	<b>124</b>
1.1	Les simulations de dynamique moléculaire : quelques grands principes	124
1.2	Les analyses des résultats issus des simulations de dynamique moléculaire . . . . .	130
1.2.1	Les distances entre acides aminés . . . . .	130
1.2.2	La RMSD . . . . .	131
1.2.3	La fraction de contacts natifs . . . . .	133
1.2.4	La RMSF . . . . .	133
1.2.5	La surface accessible au solvant . . . . .	133
1.2.6	La surface enfouie . . . . .	134
1.2.7	Les rotamères . . . . .	135
1.2.8	Les paramètres hélicoïdaux . . . . .	135
1.3	Discussion et conclusion . . . . .	137
<b>2</b>	<b>Second chapitre : la dynamique conformationnelle des protéines à "Pentatricopeptide repeat"</b>	<b>138</b>
2.1	Introduction . . . . .	138
2.2	Article 2 . . . . .	144
2.3	Discussion de l'article 2 . . . . .	163
2.3.1	Interaction ARN-PPR . . . . .	163
2.3.2	Comparaison avec la protéine apo . . . . .	164

2.4 Conclusion et perspectives de l'article 2 . . . . .	165
<b>IV Conclusion et perspectives</b>	<b>166</b>
Références	171
Résumé	188





## Table des figures

1	Disque protoplanétaire autour d'une étoile (le Soleil) à l'origine de la formation des planètes telluriques, gazeuses et de planétésimaux selon la distance à l'étoile et la matière disponible (figure issue de l'Observatoire de Paris). . . . .	2
2	Vue simplifiée de l'intérieur du Soleil. Les photons sont produits au coeur de l'étoile puis traversent les couches supérieures vers la surface où ils sont émis (d'après une figure de l'Encyclopédie de l'énergie). . . . .	3
3	L'atmosphère est une couche gazeuse entourant une planète et composée d'atomes et de molécules non chargés. Une planète possédant un champ magnétique baigne dans une magnétosphère qui est une région où les phénomènes physiques sont dominés par son champ magnétique. La magnétosphère est peuplée de particules chargées et neutres originaires de l'atmosphère (figure modifiée issue de l'Université de Picardie). . . . .	4
4	Morphologie de la magnétosphère de la Terre qui dévie le vent solaire autour de la planète (figure modifiée issue du CNES). . . . .	5
5	Vue d'artiste de l'intérieur de la lune Europe, satellite de la planète Jupiter. De gauche à droite : la croûte de glace, l'océan d'eau salée liquide, le manteau rocheux et le coeur ferreux ainsi qu'une vue schématique des stries de surface. Une configuration similaire est proposée pour Encelade (image modifiée tirée d'une vidéo de l'ESA). . . . .	6
6	Une cellule de la bactérie <i>Bdellovibrio bacteriovorus</i> imagée par ECT (Electron CryoTomography) (OIKONOMOU, CHANG et JENSEN 2016). . . . .	17
7	Représentation d'une cellule eucaryote non photosynthétique avec les nombreux compartiments et certains des acteurs essentiels à la vie de la cellule (issue de e-biologie.fr). Les cellules eucaryotes photosynthétiques contiennent en plus un compartiment où se passe la photosynthèse (cf. 5, 7.2). . . . .	19
8	Représentation de l'arbre du vivant réalisée avec 16 séquences d'ARN ribosomiques (ARNr) de 92 phyla bactériens, 26 phyla d'Archaea et les 5 super-groupes des eucaryotes (HUG et al. 2016). . . . .	21
9	Photos de spécimens d' <i>Hydra viridissima</i> par Peter Schuchert le 25 mai 2009 (gauche) et par Frank Fox le 4 mars 2012 (droite). Sources : marinespecies.org (gauche), mikro-foto.de (droite) . . . . .	24

10	Les trois endosymbioses primaires connues à ce jour qui ont amené à l'apparition des mitochondries et des organites responsables de la photosynthèse chez les eucaryotes (chloroplastes et chromatophores). Les Archaeplastida (Glaucophytes, Rhodophytes, algues vertes ou chlorophytes et plantes terrestres) contiennent des mitochondries et un ou plusieurs chloroplastes issus des endosymbioses primaires successives de deux bactéries (alphaprotéobactérie et cyanobactérie respectivement) (adapté de KEELING 2004). . . . .	26
11	Les lignées ayant subi des endosymbioses secondaires et tertiaires et dont le chloroplaste provient initialement d'une Rhodophyte. Certains plastes acquis ont subi de fortes réductions de leur taille, de leur génome et de leur activité comme c'est le cas chez certaines Alveolates (notamment les Ciliés) (adapté de KEELING 2004). . . . .	28
12	Arrangement de spécimens de Diatomées à but artistique ("Diatom arrangement" par Klaus Kemp). . . . .	29
13	Schéma de trois endosymbioses secondaires distinctes impliquant une algue verte et un eucaryote (adapté de KEELING 2004). . . . .	30
14	Les chaînes des transport des électrons dans les mitochondries et les chloroplastes. A : Représentation de la chaîne de transport des électrons et des acteurs de la chaîne photosynthétique dans le chloroplaste. B : Représentation de la chaîne de transport des électrons dans le processus de respiration ayant lieu dans les mitochondries (CHADEE et al. 2021). . . . .	32
15	Le mécanisme d'adressage des protéines aux organites (adapté du manuscrit de thèse de Clotilde Garrido (GARRIDO 2021)). . . . .	34
16	Provenance des sous-unités des complexes protéiques impliqués dans la photosynthèse dans le chloroplaste (adapté du manuscrit de thèse de Domitille Jarrige (JARRIGE 2019)). . . . .	36
17	Nombre de protéines PPR de type PLS et P selon l'espèce. Le cercle extérieur correspond au nombre de protéines de type PLS et le cercle intérieur à celui des protéines de type P (GUTMANN et al. 2020). . . . .	40
18	Illustration représentant les différentes classes de PPR et leur composition en motifs types (MANNA 2015). . . . .	42
19	Logo des types S, P et L des motifs PPR. Les rectangles gris correspondent à l'emplacement des hélices alpha constituant la paire d'hélice dans le motif consensus de type P (BARKAN et SMALL 2014). . . . .	42

20	Les protéines OTAF (facteurs M et T) jouent un rôle de protection contre la dégradation des transcrits mono-cistroniques par les exonucléases mais aussi un rôle de correction des erreurs sur les séquences ARNm des organites (MACEDO-OSORIO, MARTÍNEZ-ANTONIO et BADILLO-CORONA 2021). . . . .	44
21	Trois portions de la structure PDB 5i9f disponible sur la base de données RCSB via le logiciel VMD. A : une répétition d'un motif caractéristique d'une protéine à solénoïde alpha avec une première hélice alpha, un coude aussi appelé "linker" et une seconde hélice alpha. B : assemblage de deux répétitions du même motif que dans A. C : assemblage de quatre répétitions du même motif que A. On peut remarquer ici un début de torsion de cette chaîne de motifs répétés. . . . .	45
22	Différentes vues d'une protéine PPR (identifiant sur la base de données RCSB de la structure utilisée : 5i9f) (SHEN et al. 2016) (réalisé avec VMD). . . . .	46
23	Photo d'un spécimen de la plante terrestre à fleurs (Angiospermae) <i>Arabidopsis thaliana</i> (photo par Benjamin Zwiitnig). . . . .	46
24	Photo d'un spécimen de la microalgue verte <i>Chlamydomonas reinhardtii</i> par Sandrine Bujaldon (données du laboratoire). . . . .	48
25	Nombre de témoins vrais positifs contre nombre de témoins faux positifs pour 5 méthodes parmi celles présentées sur un même jeu de données (VO, NGUYEN et HUANG 2010). Le jeu de données provient de MARSELLA et al. 2009 : les témoins positifs correspondent à 105 domaines se structurant en solénoïde et les témoins négatifs correspondent à 247 domaines protéiques ne formant pas de solénoïde (d'après leur structure obtenue aux rayons X). . . . .	57
26	Arbre phylogénétique des eucaryotes d'après les consensus jusqu'à 2020 (BURKI et al. 2020). Chaque couleur représente un supergroupe et les flèches rouges pointent vers les groupes d'intérêt sur lesquels mon travail d'identification des TAF a porté. . . . .	62
27	Structure d'une protéine PPR vue par le dessus. A l'intérieur du sillon, l'ARN se lie de façon spécifique, un nucléotide par répétition du motif PPR (SHEN et al. 2016). . . . .	64

28	Arbres phylogénétiques des 22 espèces de Chlorophytes : à gauche la famille des protéines OPR, à droite la famille des protéines LHCa. A : En noir est indiqué le nombre de groupes d'orthologues, en bleu le nombre de protéines OPR connues, en rouge le nombre de clusters de protéines OPR connus. B : Les évènements de pertes (noir), de contractions (orange), de gains (bleu) et d'expansions (rouge) constatés au sein de la phylogénie. La longueur des barres de couleur représente le nombre de groupes d'orthologues ayant subi l'évènement en question (adapté du rapport de stage de master de Shogofa Mortaza (MORTAZA 2018)). . . . .	67
29	Schéma du taux d'hydrophobicité le long de la séquence protéique illustrant l'hypothèse de travail de début de stage d'Alexis Astatourian (ASTATOURIAN 2021). . . . .	116
30	Taux d'hydrophobicité le long de la séquence de la protéine AT1G28020.1 qui est une protéine PPR connue d' <i>Arabidopsis thaliana</i> . Le taux est calculé sur une fenêtre glissante de 19 acides aminés via l'échelle d'hydrophobicité des acides aminés proposée dans FAUCHÈRE et PLISKA 1983. Les points rouges représentent les linkers de paires d'hélices alpha détectés par le logiciel ard2 (FOURNIER et al. 2013 ; ASTATOURIAN 2021). . . . .	117
31	Trois vues d'une même protéine PPR sous différents angles. En bleu les acides aminés apolaires, en vert les acides aminés polaires, les autres acides aminés gris ne font pas partie d'hélice amphiphile. Figure réalisée sur une protéine PPR de la diatomée <i>Phaeodactylum tricornutum</i> avec le logiciel de visualisation VMD. . . . .	118
32	Distribution du nombre de protéines OPR candidates dans 57 protéomes issus du MMETSP (ASTATOURIAN 2021). . . . .	119
33	Distribution du nombre de protéines PPR candidates dans 57 protéomes issus du MMETSP (ASTATOURIAN 2021). . . . .	120
34	Distribution du nombre de protéines à solénoïde alpha candidates dans 57 protéomes issus du MMETSP (ASTATOURIAN 2021). . . . .	120
35	Schéma simplifié illustrant les différents éléments pouvant varier au cours de la simulation qui sont retrouvés dans l'équation à quatre termes du champs de force (figure réalisée à partir d'un modèle issu de LEACH 2001). . . . .	127
36	Le potentiel de Lennard-Jones caractérise l'énergie potentielle de l'interaction entre deux atomes en fonction de leur proximité. Deux atomes sont à l'équilibre lorsqu'ils sont à une distance à laquelle l'énergie potentielle de leur interaction est minimale. . . . .	128

37	Le système de seuil (ou "cutoff" en anglais) est utilisé pour optimiser les temps de calculs en ne prenant que les forces provenant des atomes à l'intérieur d'une sphère de rayon de la taille du seuil et ayant pour centre l'atome duquel on calcule l'accélération et la position. On schématise ici deux seuils : "cutoff" est la limite à laquelle les atomes situés plus loin ne sont plus utilisés pour calculer la somme des forces appliquées sur l'atome et cutnb est une zone tampon (ou "buffer" en anglais) qui permet de voir entrer ou sortir un atome de la zone "cutoff". . . . .	128
38	Schéma simplifié du principe des boîtes de simulation strictement identiques entre lesquelles les particules peuvent circuler. Entourée en rouge, il s'agit de la boîte centrale que l'on observe lorsqu'on visualise la simulation par la suite. . . . .	129
39	Schéma d'une liaison hydrogène (ligne pointillée) entre un atome d'hydrogène (H) lié à un azote d'azote (N) par une liaison covalente (ligne pleine) et un atome d'oxygène (O). . . . .	130
40	Illustrations de deux représentations types de la RMSD que nous avons utilisé pour ce travail. A gauche : exemple d'une figure sur laquelle est tracée la RMSD en fonction du temps par rapport à la structure initiale. A droite : exemple d'une figure sur laquelle est tracée la RMSD de la structure à chaque pas de temps contre les structures à tous les autres pas de temps. . . . .	132
41	Exemple de la RMSF dans un motif PPR. . . . .	134
42	Illustration du principe de calcul de la surface accessible au solvant d'une molécule (adapté de DABERDAKU 2018). . . . .	134
43	Illustration du calcul de la surface enfouie entre deux motifs PPR. La surface de van der Waals des atomes composant les motifs PPR est affichée, les pointillés représentent donc la surface accessible au solvant sur un unique plan. . . . .	135
44	Atomes utilisés pour définir les angles $\chi_1$ (atomes en bleu) et $\chi_2$ (atomes en vert) dans un acide aminé schématisé. . . . .	136
45	Quelques uns des paramètres hélicoïdaux que l'on peut mesurer au cours de la simulation pour des éléments disposés régulièrement autour d'un seul axe hélicoïdal (adapté d'une illustration de Rajarshi Rit). . . . .	136
46	Structures d'une protéine PPR synthétique liée (vert) et non liée (gris) à un ARNsb SHEN et al. 2016. . . . .	139
47	Logo des motifs des protéines PPR de type P de l'espèce de plante terrestre <i>Arabidopsis thaliana</i> . Les résidus du code PPR sont indiqués par des flèches noires (adapté de SHEN et al. 2016). . . . .	140

48	Les quatre constructions réalisées par Marion Sisqueillas lors de son stage de master 2 en 2018 pour son étude de la dynamique des protéines PPR qui a ouvert la voie au travail présenté ici (SISQUEILLAS 2018). . . . .	141
49	Arbre regroupant différentes lignées eucaryotes photosynthétiques ayant subi une ou plusieurs endosymbioses (cf. FALCIATORE et al. 2022). . . . .	168
50	Les quatre structures des protéines synthétiques se liant à des ARNs différents (cf. SHEN et al. 2016). La structure 5I9F correspond à celle utilisée dans ce travail mais il est possible d'utiliser la dynamique moléculaire sur les trois autres. En rouge sont entourés les acides nucléiques qui varient d'une structure à l'autre : 5i9d contient un ARN composé de 4 uraciles, 2 adénines et 4 uraciles ; 5i9f contient un ARN composé de 10 uraciles ; 5i9g contient un ARN composé de 4 uraciles, 2 cytosines et 4 uraciles ; 5i9h contient un ARN composé de 4 uraciles, 2 guanines et 4 uraciles. . . . .	170





## Liste des tableaux

- 1 Les principales familles de protéines formant des solénoïdes alpha. Certaines se lient à des protéines, d'autres à des ARNs et d'autres encore à de l'ADN, la longueur des motifs répétés pouvant varier grandement. . . . . 51
- 2 Principales méthodes développées depuis 1999 permettant d'identifier des protéines à motifs répétés dont des protéines à solénoïde alpha. . . . . 55
- 3 Les 15 systèmes de la protéine PPR avec et sans ARN simulé puis analysés par dynamique moléculaire . . . . . 143







## Avant-propos

La genèse d'une planète apte à développer la vie : du  
macro au micro

"[...] the cosmos is our environment, with all human beings who have ever lived we share the same view of the stars and ultimately we ourselves are stardust. Like every thing on Earth we consist of the vestiges of long extinguished celestial bodies. [...] all the elements were formed in the stars out of hydrogen and helium through nuclear fusion. If you are less romantically inclined you can call human beings stellar nuclear waste."

Conversation entre Martin Rees et Stefan Klein en référence à Carl Sagan,  
*We are all stardust*, Stefan Klein (2015)

En 1969 au Mexique quelques mois avant Apollo 11, 250 tonnes de roche se sont écrasées : la météorite d'Allende (GARGAUD et al. 2009). Il s'agit d'une chondrite carbonée (météorite indifférenciée, c'est-à-dire non formée d'une superposition de couches de compositions variées, et contenant une grande quantité de carbone) datant des prémices de la formation du système solaire. Elle est donc composée de poussières et de gaz provenant du nuage de matière à l'origine du système solaire. Ainsi, son étude a apporté de nombreux indices quant à l'âge et à la genèse des corps du système solaire.

Le Soleil est une sphère pleine constituée presque exclusivement d'un mélange de gaz (hydrogène et hélium) (NASA 2022) qui correspond à 99,86% de la masse totale du système solaire (WOOLFSON 2000). Initialement, il s'est formé et a commencé à briller il y a 4,5 à 4,6 milliards d'années (NASA 2022; CEA 2012) grâce à l'effondrement gravitationnel de la majeure partie d'un immense nuage moléculaire. La cause de cet effondrement est encore débattue et les hypothèses actuellement les plus soutenues sont l'effet de l'onde de choc provenant d'une supernova (explosion d'une étoile) voisine ou du passage d'une étoile inconnue à proximité. Quelle

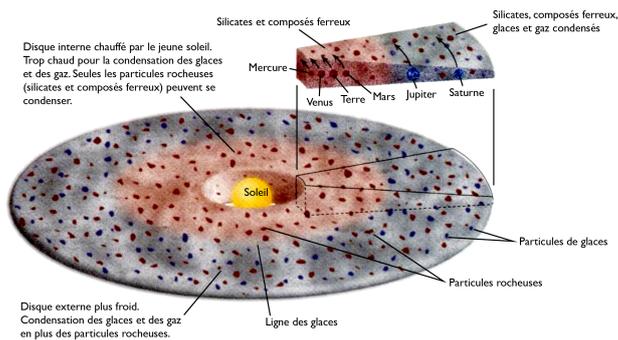
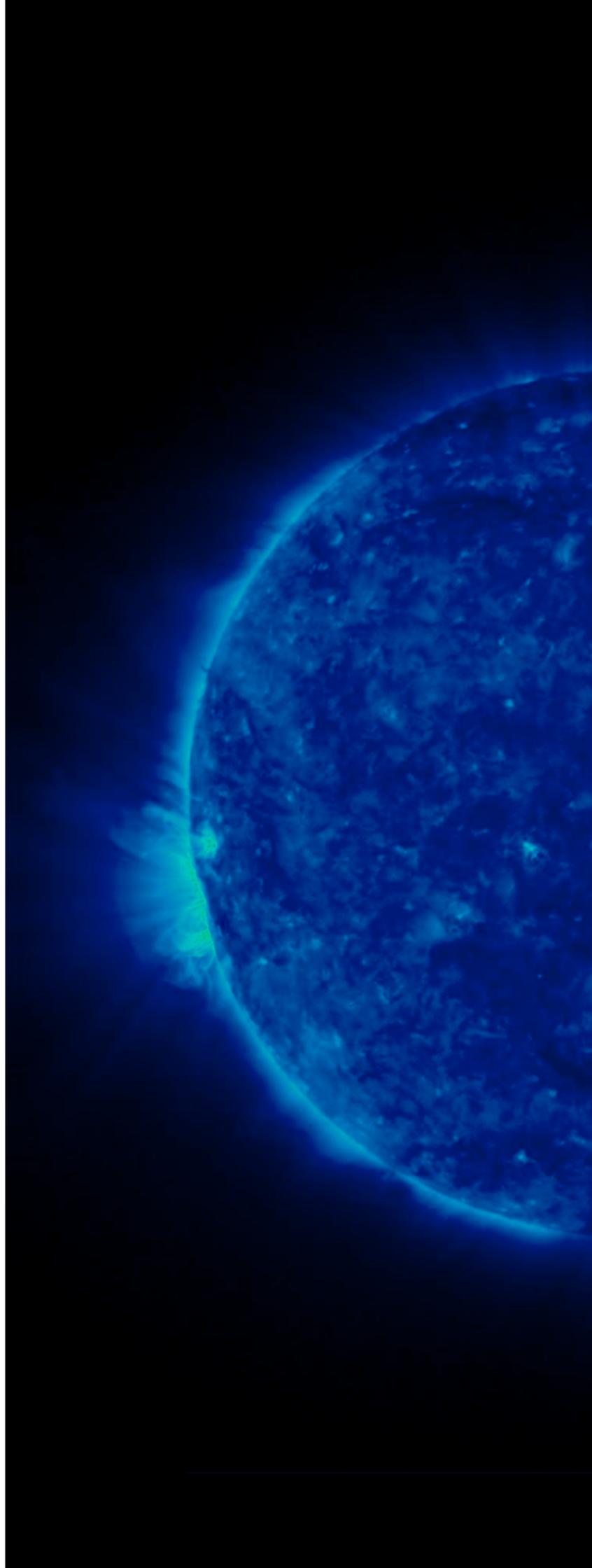
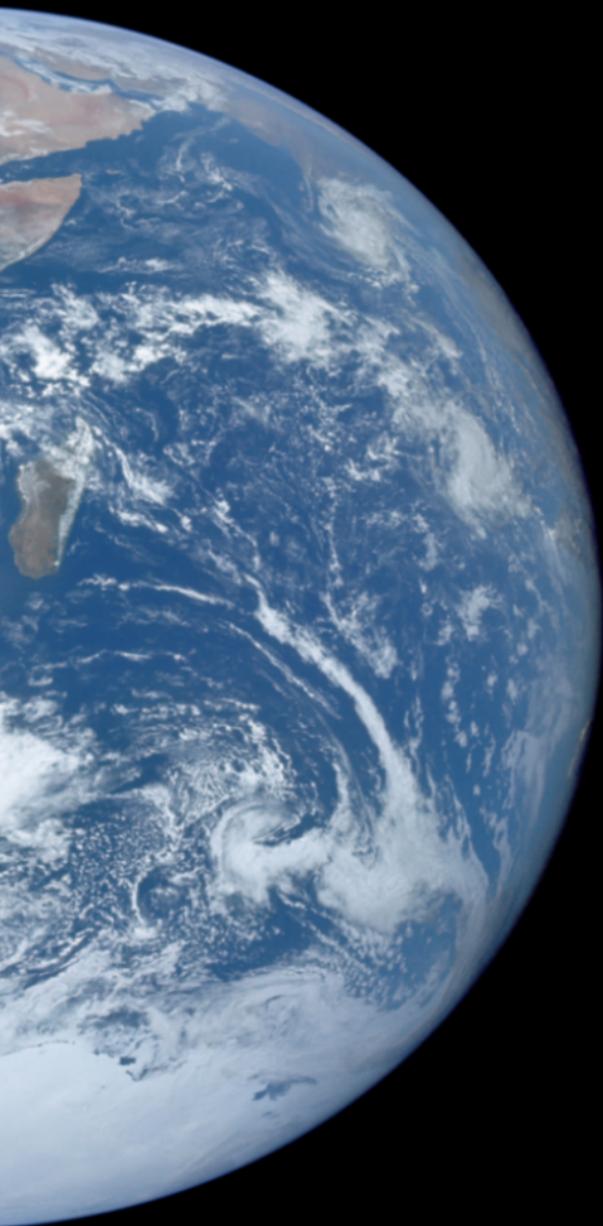


FIGURE 1 – Disque protoplanétaire autour d'une étoile (le Soleil) à l'origine de la formation des planètes telluriques, gazeuses et de planétésimaux selon la distance à l'étoile et la matière disponible (figure issue de l'Observatoire de Paris).

que soit l'hypothèse retenue, le reste de ce nuage initial s'est ensuite agrégé en quelques dizaines à centaines de millions d'années pour former les quatre planètes gazeuses (Jupiter, Saturne, Uranus et Neptune en s'éloignant du Soleil) et une multitude de planétésimaux (petits corps rocheux formant un disque pouvant être qualifié de protoplanétaire) (cf. Figure 1).

Photo d'illustration : "Sun", NASA, STEREO's SECCHI/Extreme Ultraviolet Imaging Telescope le 4 décembre 2006. Image en fausses couleurs de l'atmosphère du Soleil à 1 million °C (image modifiée).





Les collisions et agrégations successives entre ces planétésimaux ont finalement donné naissance aux quatre planètes rocheuses (Mercure, Vénus, Terre et Mars en s'éloignant du Soleil) et à certaines lunes. À ce jour, notre système solaire est donc constitué de huit planètes et de très nombreux astéroïdes, planètes naines et autres petits corps (NASA 2022).

Dans les profondeurs de notre étoile, en son coeur, la pression est 200 milliards de fois plus grande que la pression atmosphérique terrestre et en conséquence, la température du noyau du Soleil est particulièrement élevée (environ 15 millions °C) (CEA 2012). Le noyau du Soleil est le lieu où des atomes d'hydrogène, des électrons et d'autres particules se percutent et fusionnent pour finalement engendrer un atome d'hélium 4 au terme de plusieurs étapes de fusion (CEA 2012). Durant ce processus, d'immenses quantités d'énergie sont émises et se dégagent sous forme de rayonnement électromagnétique (CEA 2012), aussi appelé flux de photons (quantums d'énergie qui composent la lumière, ils sont décrits par la théorie quantique (Le Robert en ligne, définition au 16 septembre 2022)) (cf. Figure 2). La dis-

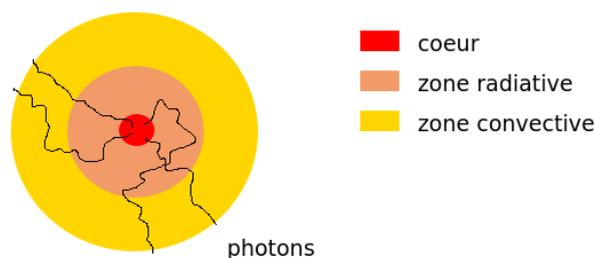


FIGURE 2 – Vue simplifiée de l'intérieur du Soleil. Les photons sont produits au coeur de l'étoile puis traversent les couches supérieures vers la surface où ils sont émis (d'après une figure de l'Encyclopédie de l'énergie).

tance moyenne de la Terre au Soleil est d'environ 150 millions de kilomètres, la lumière met donc en moyenne 8 minutes à faire le trajet (NASA 2022 ; CEA 2012). Cependant, l'âge de la lumière provenant du Soleil est bien supérieur et se compte en milliers voire en centaines de milliers d'années. Le Soleil est en effet très dense et même si les photons sont de petites particules, après leur naissance dans les profondeurs du Soleil, ils sont régulièrement absorbés et ré-émis par les atomes environnants durant leur trajet vers la surface, celui-ci est donc en réalité assimilable à une marche aléatoire (cf. Figure 2).

Photo d'illustration : "The Blue Marble" (ou "La Bille Bleue"), NASA/équipage Apollo 17 le 7 décembre 1972 à environ 29 000 km de la Terre (image modifiée).

Une certaine variabilité intervient aussi dans l'estimation de l'âge de la lumière provenant du Soleil car celui-ci n'est pas uniformément dense, ce qui implique qu'à leur sortie du Soleil, des photons traversant des zones denses seront plus vieux que ceux qui ne les traversent pas (ODENWALD s. d.) (cf. Figure 2).

La quantité d'énergie solaire reçue sur Terre est très faible par rapport à la quantité émise par le Soleil (CEA 2012). En effet, l'atmosphère terrestre filtre la plupart des rayonnements autres que la lumière visible grâce à différents mécanismes de diffusion, de réflexion et d'absorption qui ont notamment pour effet de réchauffer la surface de notre planète (DANIEL 2003; DEMIRDJIAN 2007). Les rayonnements les plus énergétiques (ou ionisants), eux, sont repoussés bien avant d'arriver au niveau de la surface de la Terre grâce à la magnétosphère (région de l'espace dominée par le champ magnétique de l'objet, ici la Terre) formant un bouclier face au Soleil (cf. Figures 3 et 4). Ce champ magnétique est généré par un effet de dynamo provoqué par des mouvements de convection dans le noyau externe ferreux terrestre (IPGP 2018).

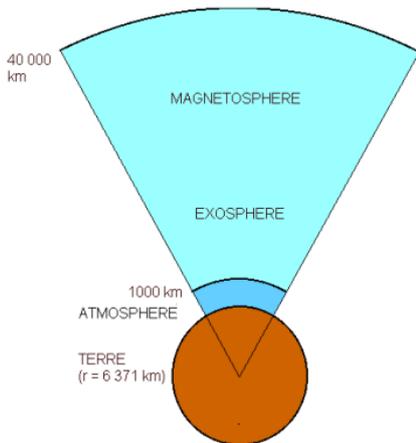
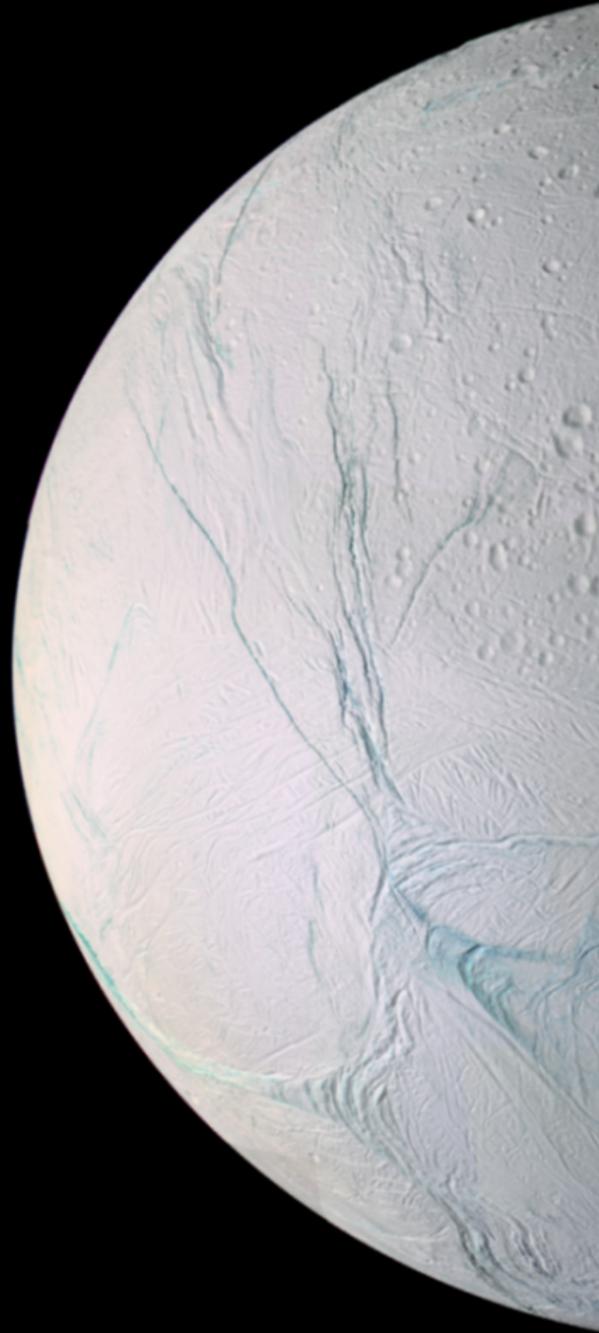


FIGURE 3 – L'atmosphère est une couche gazeuse entourant une planète et composée d'atomes et de molécules non chargés. Une planète possédant un champ magnétique baigne dans une magnétosphère qui est une région où les phénomènes physiques sont dominés par son champ magnétique. La magnétosphère est peuplée de particules chargées et neutres originaires de l'atmosphère (figure modifiée issue de l'Université de Picardie).

Photo d'illustration : Encelade, une lune de Saturne le 14 juillet 2005. Image en fausses couleurs composée de 21 clichés pris par la sonde Cassini alors qu'elle passait au pôle sud de la lune. Les motifs complexes observés sont dus à des fissures dans la couche de glace recouvrant la lune et la variété des couleurs peut s'expliquer par l'absence de la matière blanche poudreuse à leurs abords (NASA/JPL/Space Science Institute, image modifiée).



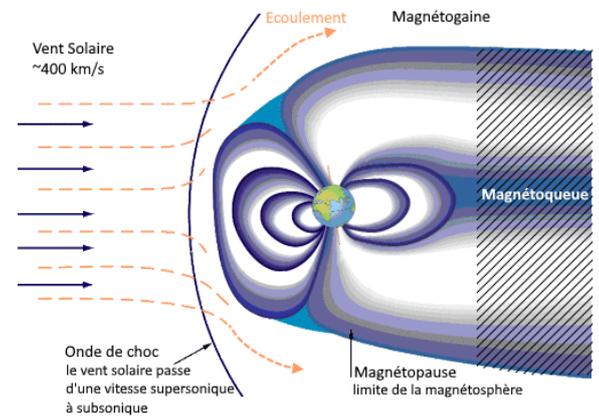


FIGURE 4 – Morphologie de la magnétosphère de la Terre qui dévie le vent solaire autour de la planète (figure modifiée issue du CNES).

La surface de la Terre est couverte à plus de 70% d'eau liquide et l'origine de cette eau est une question toujours débattue. Une hypothèse propose que l'eau à la surface d'un corps planétaire soit issue du dégazage de la planète lors de sa formation (GREENWOOD et al. 2018) tandis qu'une autre avance qu'elle aurait été apportée via des collisions avec de nombreux corps riches en eau (sous forme de glace) tels que des comètes ou des astéroïdes (ALBAREDE 2009). Une combinaison de ces deux hypothèses est également possible et la question porte alors sur l'apport de chaque source à la quantité d'eau actuellement présente à la surface de la Terre. Notons que l'eau existe à la surface d'autres corps du système solaire tels qu'Encelade, une lune de Saturne, Europe, une lune de Jupiter ou encore aux pôles de Mars et de la Lune mais elle ne s'y trouve qu'à l'état de glace contrairement à la Terre (NIMMO et PAPPALARDO 2016) (bien que l'eau coulait à la surface de Mars il y a plusieurs millions d'années, les nombreux deltas et lits de rivières asséchés visibles depuis l'espace s'en faisant la preuve). Cependant, de l'eau à l'état liquide a récemment été détectée dans ce qui est proposé comme un réseau de lacs d'eau salée (saumure) sous la surface du pôle sud de Mars (LAURO et al. 2022 ; STILLMAN et al. 2022) et deux sondes orbitales de la NASA (Cassini et Galileo) ont permis de mettre en évidence la présence d'un océan d'eau salée liquide en sous-surface pouvant atteindre plusieurs dizaines de kilomètres de profondeur sur Encelade (SPENCER et NIMMO 2013) et les lunes dites "glacées" de Jupiter (Europe, Callisto (ZIMMER, KHURANA et KIVELSON 2000) et Ganymède (KIVELSON, KHURANA et VOLWERK 2002)).

Photo d'illustration : L'astéroïde Bennu le 2 décembre 2018. Image composée à partir de 12 clichés PolyCam pris par la sonde OSIRIS-Rex à 24km de l'objet (NASA/Goddard/University of Arizona, image modifiée).

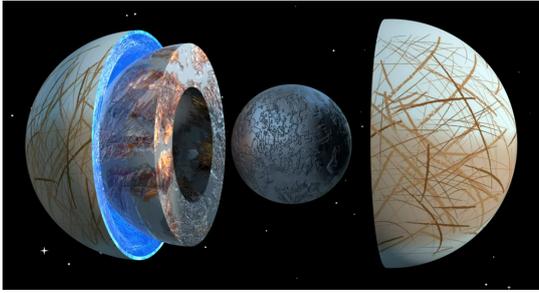
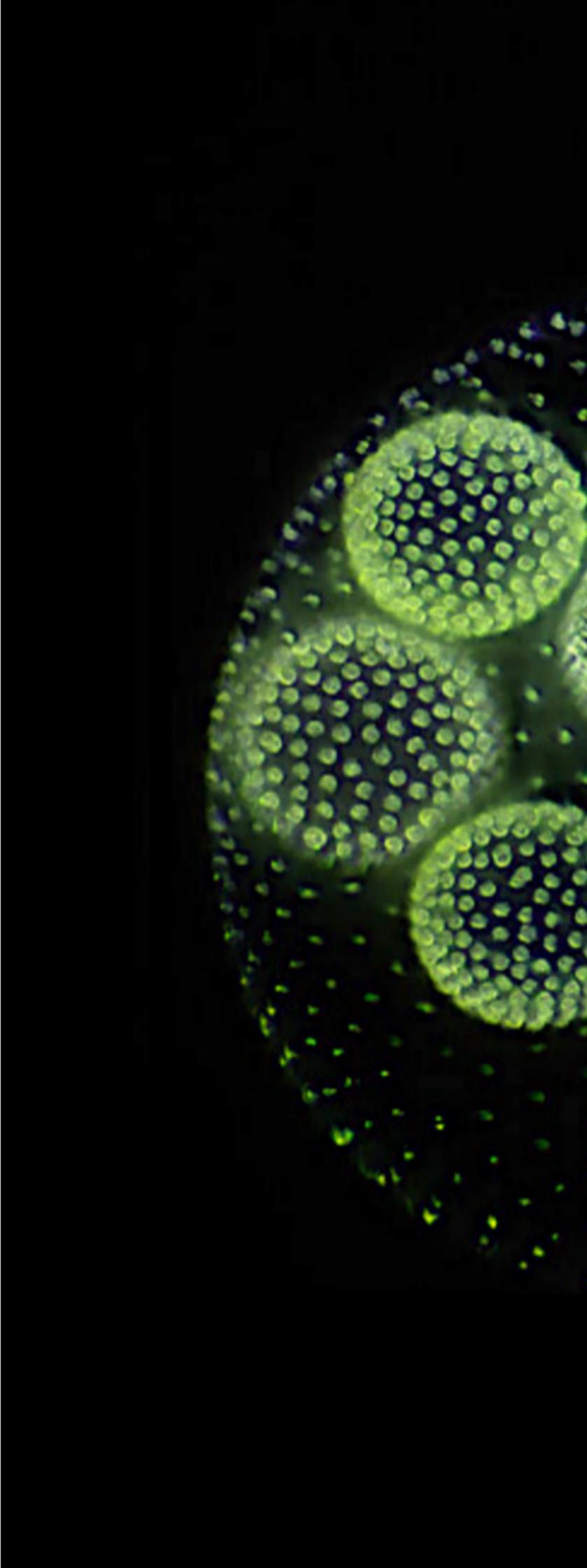


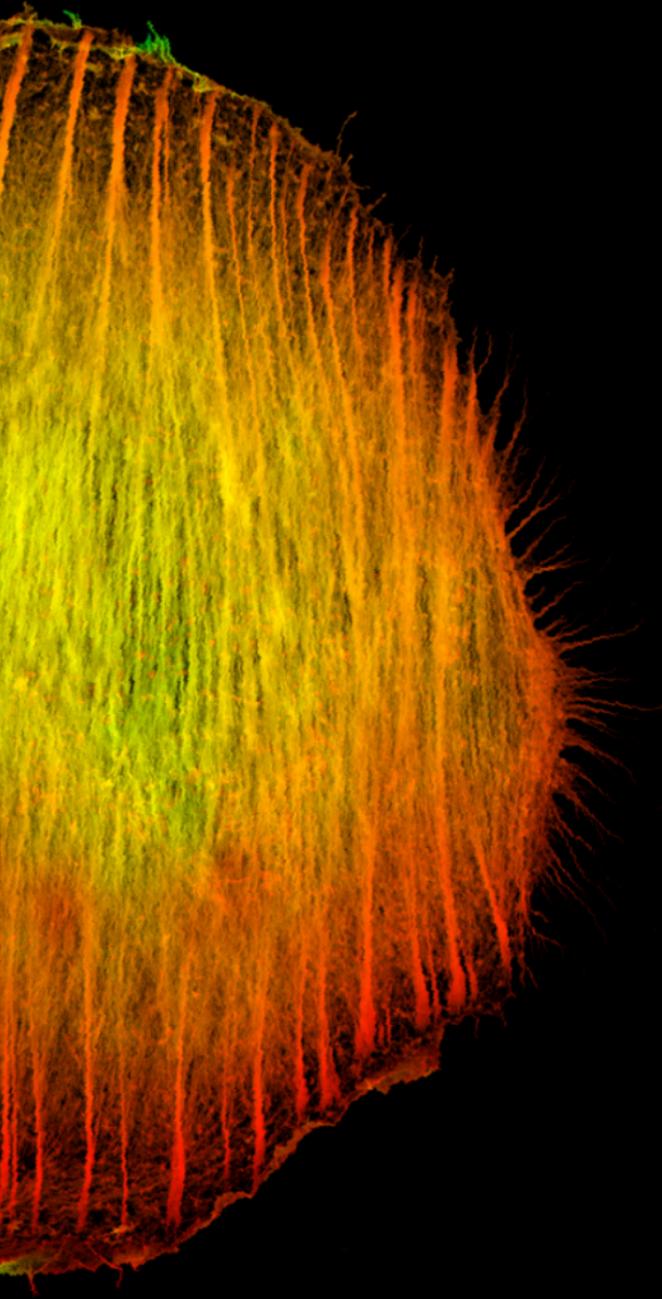
FIGURE 5 – Vue d'artiste de l'intérieur de la lune Europe, satellite de la planète Jupiter. De gauche à droite : la croûte de glace, l'océan d'eau salée liquide, le manteau rocheux et le coeur ferreux ainsi qu'une vue schématique des stries de surface. Une configuration similaire est proposée pour Encelade (image modifiée tirée d'une vidéo de l'ESA).

Ces océans sous-glaciaires seraient dus aux effets de marée provoqués par les planètes gazeuses Jupiter et Saturne, réchauffant ainsi l'intérieur de leurs lunes et amenant à la fonte d'une partie de la couche de glace. L'eau liquide nécessaire à la vie semble donc rare dans notre système solaire, et ce, surtout à la surface des corps.

Plusieurs milieux sont proposés quant à l'émergence des premières molécules organiques sur Terre : les sources hydrothermales, les surfaces ferreuses, les météorites ou encore l'atmosphère. Les modèles de chimie prébiotique (c'est-à-dire précédant l'apparition de la vie) montrent aussi qu'une source d'énergie pour activer certaines réactions est nécessaire (la chaleur, la lumière ou la foudre par exemple). Certains modèles sont plus convaincants que d'autres quant à l'organisation des molécules organiques en systèmes ordonnés précurseurs des molécules du vivant. En effet, le milieu aqueux permettrait la rencontre plus aisée des molécules pour qu'elles s'assemblent et forment les précurseurs des molécules organiques caractéristiques du vivant comme les acides nucléiques et les acides aminés (les interactions hydrophobes et l'hydrolyse peuvent aussi avoir joué un rôle). Notons cependant que la présence de matière organique n'est pas nécessairement synonyme de vie car des molécules organiques (notamment des acides aminés) sont aussi assemblées dans des milieux qui ne semblent pas favorables à la vie, comme dans l'espace sur des météorites ou des comètes (GARGAUD et al. 2009).

Photo d'illustration : Microalgues du genre Volvox (mikrophoto.de, image modifiée).





Sachant cela, on pourrait s'attendre à trouver des traces de vie (ancienne ou présente) sur les lunes glacées de Jupiter et Saturne ou sur Mars. L'hypothèse d'une vie souterraine sur ces autres planètes, notamment sous forme de micro-organismes ne pouvant être exclue, de futures missions spatiales de la NASA et de l'ESA telles que respectivement Europa Clipper et JUPITER ICy moons Explorer (JUICE) se pencheront sur cette question.

Pour tenter de savoir quand et où la vie est apparue sur Terre, on recherche des traces comme des macro- ou des micro-fossiles. Les traces de vie incontestables (fossiles conservés et identifiables) datent de 2,7 milliards d'années. Toutes les traces de vie plus vieilles sont sujettes à controverse, notamment à cause des phénomènes volcaniques basiques ayant lieu alors et du métamorphisme des roches qui sont plus âgées (le métamorphisme est la transformation d'une roche à l'état solide en fonction des conditions de chaleur et de pression et de l'environnement chimique). Les plus anciennes potentielles traces de vie sont trouvées dans des micro-fossiles datant d'environ 3,5 à 3,8 milliards d'années (GARGAUD et al. 2009). Cependant, il a récemment été suggéré que certaines traces chimiques âgées de plus de 4 milliards d'années extraites dans les montagnes de Jack Hills dans l'Ouest de l'Australie peuvent être interprétées comme des traces de vie (proposition de la biogenèse d'atomes de carbone préservés dans des zircons (minéral du groupe des silicates)) (BELL et al. 2015). Certains avancent donc que la vie aurait pu commencer à se développer rapidement après la formation de la Terre tandis que d'autres proposent que la vie ne soit apparue que plus d'un milliard d'années après.

Récemment, plusieurs "briques élémentaires du vivant" telles que les nucléotides composant l'ARN et l'ADN, les acides aminés composant les protéines, des acides gras et des sucres ont été synthétisés artificiellement à partir d'éléments présents sur la jeune Terre dans des milieux ressemblant à ceux qui pouvaient y être rencontrés il y a plus de 4 milliards d'années (BECKER et al. 2019; HUD et FIALHO 2019). Cependant la synthèse d'éléments du vivant n'est pas suffisante pour créer la vie, il faut en effet combiner ces fragments en des ensembles cohérents.

Notons enfin que nous connaissons aujourd'hui la vie à ADN mais il ne peut pas être exclu que d'autres formes de vie dont nous n'avons pas de traces ou dont nous ne savons pas reconnaître les traces l'aient précédée. La vie à ADN aurait alors simplement été la plus efficace et adaptée et elle aurait ainsi mieux subsisté tandis que les autres se seraient éteintes (HAWKING 2018).

Photo : "Depth Coded Phalloidin Stained Actin Filaments Cancer Cell", Howard Vindin. Une cellule d'un ostéosarcome à la microscopie confocale, coloration à la phalloïdine pour l'observation des filaments d'actine F qui sont impliqués dans la structure interne des cellules (image modifiée).







## Première partie

# Introduction

Courte rétrospective : du non-vivant à nos jours

"Quelles sont nos chances de trouver d'autres formes de vie en explorant la galaxie ? Si l'échelle temporelle d'apparition de la vie sur Terre n'est pas exceptionnelle, il devrait y avoir de nombreuses étoiles entourées de planètes où la vie est apparue. Certains de ces systèmes stellaires ont pu se former 5 milliards d'années avant la Terre, pourquoi alors notre galaxie ne fourmille-t-elle pas de formes de vie biologiques ou robotisées ? Pourquoi la Terre n'a-t-elle pas encore été visitée ou colonisée ? [...] Il se peut que la probabilité d'apparition de la vie soit si faible que la Terre soit la seule planète où cela s'est produit. Il se peut aussi que la vie soit apparue ailleurs et ait évolué jusqu'à former des cellules mais pas une vie intelligente."

*Brèves réponses aux grandes questions*, Stephen Hawking  
(Edition Odile Jacob, 2018)



# 1 De la chimie organique à la biologie : l'organisation du premier "être vivant"

Dater l'apparition de la vie est une question difficile mais à laquelle certains indices peuvent permettre de proposer des réponses. Il y a 4,4 milliards d'années, la Terre est progressivement devenue compatible avec l'apparition de la vie à l'aube de ses 200 millions d'années. Les conditions physico-chimiques et la géologie étaient réunies pour permettre la formation de molécules organiques comme les acides aminés ou les acides nucléiques et leurs rencontres mutuelles. La vie, par le biais de petits peptides ou de chaînes d'acides nucléiques auraient ainsi pu émerger.

Comme évoqué précédemment, certains cristaux vieux de plus de 4,1 milliards d'années retrouvés sur le site actuellement connu comme portant les plus anciens vestiges de la Terre pourraient en avoir gardé la trace (BELL et al. 2015). Cependant si la vie avait commencé à émerger il y a plus de 4 milliards d'années, le grand bombardement tardif de la Terre (un événement pouvant être qualifié de cataclysmique) a pu tout anéantir. Il s'agissait d'une intense pluie de météorites qui a duré plusieurs dizaines de millions d'années et dont les traces sont encore visibles sur la Lune, Mars, Mercure et Venus (BOTTKÉ et NORMAN 2017) car leurs surfaces n'ont pas été renouvelées depuis (notamment par manque d'activité géologique et de volcanisme suffisant en surface) à l'inverse de la surface de la Terre. Cette hypothèse se base en partie sur les prélèvements de roches dans les cratères lunaires qui ont été rapportés par les missions Apollo. Cet épisode aurait eu lieu il y a environ 3,9 milliards d'années et si la vie était apparue avant, il a pu l'annihiler à cause de l'échauffement de la surface de la planète dû aux réguliers impacts de météorites. On peut alors supposer que le processus d'émergence de la vie a dû recommencer bien que certains proposent que la vie aurait pu en partie résister (GARGAUD et al. 2009).

Toujours est-il que des témoins plus récents dont la provenance biologique est presque certaine de nos jours, datant au plus de 3,5 milliards d'années (selon les estimations) sont trouvés en grands nombres notamment en Australie : les stromatolites. Ce sont des amas de carbonates fabriqués par des cyanobactéries photosynthétiques (bactéries de couleur bleue ou verte utilisant l'énergie lumineuse pour fabriquer des molécules organiques essentielles à leur survie) (ALLWOOD et al. 2006 ; MOYEN et THOMAS 2007). Pour en arriver à l'apparition de bactéries, la vie et surtout les cellules ont du émerger bien avant.

Si la question de la temporalité de l'émergence de la vie divise, c'est aussi le cas des questions "Comment la vie s'est-elle formée au départ ?" et "Quel objet est

apparu en premier : les molécules catalytiques ou bien l'ancêtre du compartiment cellulaire?".

Tout d'abord, plusieurs définitions quant à ce qu'est la vie sont proposées mais on peut essayer de les accorder pour trouver trois conditions qui semblent essentielles pour différencier le vivant du non-vivant : la capacité à reconnaître ce qui est à soi de ce qui ne l'est pas (la compartimentation), le maintien dans un environnement (le métabolisme), l'aptitude à se reproduire et à évoluer (le système génétique) (GARGAUD et al. 2009). Cette définition semble bien convenir à l'unité du vivant : la cellule.

La description de la cellule comme unité capable de se diviser et de générer ainsi deux cellules a été proposée par Rudolf Virchow en 1859 (VIRCHOW 1859). Pasteur a prouvé par la suite que la vie ne se génère effectivement pas spontanément sur des temps courts (contrairement à ce que d'autres avançaient alors) en utilisant les principes de stérilisation et d'isolement des milieux : il faut déjà avoir une cellule pour en former une autre, ce qui est le principe de la division cellulaire).

Ceci étant dit, il reste maintenant à définir ce qui est apparu en premier et deux choix s'offrent à nous : le métabolisme ou la capacité à reproduire une molécule (ANET 2004). Ce clivage est toujours d'actualité bien que des propositions de réconciliation des deux thèses aient été avancées dernièrement, arguant que des molécules de même type pourraient jouer les deux rôles (SALADINO et al. 2012). C'est entre autre le cas des molécules d'ARN dont certaines ont des activités catalytiques. Elles sont appelées ribozymes et il en existe toujours de nos jours dans les cellules. On peut notamment citer la ribonucléase P dont le site actif est uniquement composé d'ARN mais qui porte plus ou moins de protéines accessoires selon les espèces. Cette enzyme joue un rôle dans la maturation des ARN de transferts (ARNt). De la même façon, le site actif du ribosome est lui aussi composé uniquement d'ARN et catalyse la synthèse des liaisons peptidiques entre les acides aminés pour former les chaînes peptidiques à la base des protéines. Certains ARN sont aussi capables de s'autorépliquer (SIEVERS et VON KIEDROWSKI 1994) ou de s'autocliver (couper sa propre chaîne de nucléotides), c'est notamment le cas d'introns que l'on appelle auto-épissables (c'est-à-dire qui peuvent se séparer de l'ARN codant de façon entièrement autonome) (ROBERTSON, ALTMAN et J. D. SMITH 1972; WEINER 1993; GARGAUD et al. 2009).

Une fois cette étape franchie, on peut alors se demander comment isoler ces machineries du milieu environnant pour former les ancêtres des cellules. Plusieurs hypothèses cherchant à répondre à la formation d'un métabolisme et d'une

proto-cellule sont proposées et ici nous en discutons rapidement quelque unes des plus soutenues.

## **1.1 Première hypothèse : la "soupe" prébiotique ou le monde à ARN**

L'apparition du vivant à la suite de la formation de molécules organiques élémentaires est sujette à débat mais c'est le modèle du "monde à ARN" qui est actuellement le plus étudié. Il propose ainsi que les premières molécules caractéristiques du vivant aient été des chaînes d'ARN (ALBERTS et al. 2015). Celles-ci auraient été des ARN catalytiques capables de s'auto-répliquer dans un milieu aqueux. Par la suite, un mécanisme de traduction pour former des protéines émerge et le milieu est alors composé d'ARN et de protéines (ALBERTS et al. 2015). Enfin, les prémices des membranes cellulaires que l'on connaît aujourd'hui seraient apparus à partir de la formation de micelles, de vésicules ou bien de coacervats. Il s'agit de différents types d'assemblages hydrophobes ou amphiphiles (molécules qui sont à moitié hydrophobe et à moitié hydrophile) plus ou moins sphériques permettant de différencier un milieu intérieur d'un milieu extérieur.

On peut ensuite imaginer que l'association de tous ces éléments à des composants minéraux de l'environnement a entraîné l'apparition du premier ancêtre cellulaire. L'évolution de ce système au fil des générations a ensuite donné naissance à des cellules se complexifiant et s'adaptant à différents environnements (GARGAUD et al. 2009).

Notons que l'apparition de la membrane cellulaire est ici bien plus tardive que l'émergence des ARN se répliquant et des protéines, mais sans barrière, les molécules complémentaires d'ARN alors de plus en plus nombreuses dans le milieu auraient été plus difficiles à séparer alors qu'il s'agit d'une étape nécessaire pour leur réplication (d'après les travaux de SIEVERS et VON KIEDROWSKI 1994). Il s'agit de l'une des limites à l'apparition tardive de la membrane cellulaire.

## **1.2 Seconde hypothèse : le monde "fer-souffre" de surface, ou l'être vivant à deux dimensions**

Des surfaces minérales notamment composées de fer et de soufre au sein de sources hydrothermales peuvent avoir joué un rôle critique dans l'apparition du premier être vivant. En effet, un proto-métabolisme basé sur des proces-

sus d'oxydo-réduction (notamment du fer) permettrait à un proto-organisme d'obtenir de l'énergie pour produire des molécules organiques ou se reproduire. Cet organisme primitif aurait ainsi été en deux dimensions car nécessitant en permanence la surface ferreuse d'une source hydrothermale (d'après les travaux de WÄCHTERSCHÄUSER 1992). Par la suite un système génétique suivi d'une machinerie de traduction ainsi qu'une membrane cellulaire ont pu apparaître ou co-évoluer pour former un organisme à trois dimensions au lieu de deux.

Cependant, cette proposition échoue à expliquer certains points essentiels à la vie, notamment à montrer comment l'énergie générée grâce aux réactions d'oxydo-réduction serait redirigée et utilisée par le proto-organisme. De nos jours nous connaissons plusieurs molécules permettant de stocker et de transporter l'énergie (l'ATP produite par l'ATPase par exemple) mais il s'agit probablement de mécanismes trop complexes pour apparaître lors des balbutiements de la vie (GARGAUD et al. 2009).

## **1.3 Autres hypothèses à propos de l'émergence de la vie**

### **1.3.1 La co-évolution hasardeuse**

La co-évolution hasardeuse de vésicules ou de micelles ainsi que d'un métabolisme primitif pouvant impliquer des acides aminés (faciles à synthétiser dans les conditions environnementales d'une jeune Terre) ou des ARN catalytiques peut être envisagée comme une autre solution à l'émergence du premier être vivant. Cette hypothèse est parfois proposée en ajoutant une composante génétique pouvant être héritée. Il s'agirait alors d'une proto-cellule contenant trois sous-mécanismes auto-catalytiques qui co-évoluent : une membrane pour garder la cohésion du système global (la membrane cellulaire primitive), un système génétique contenant les informations transmises à la descendance et un métabolisme permettant de capturer des ressources et de générer de l'énergie notamment pour reproduire la proto-cellule (GÁNTI 2003). La complexification et l'évolution darwinienne interviennent ensuite au fil des générations. On note cependant que cette hypothèse semble décrire une proto-cellule bien plus avancée que les autres hypothèses déjà évoquées et ce, même si les proto-métabolismes et systèmes génétiques ne sont pas complexes.

### 1.3.2 La vie extraterrestre

Une autre hypothèse propose que la vie soit apparue sur un autre corps du système solaire ou bien dans l'espace sans support près d'une autre étoile (COMTE et al. 2023) et qu'elle ait ensuite été apportée sur Terre par des collisions météoritiques ou cométaires. En effet, certains corps du système solaire semblent recouverts d'une couche de molécules organiques et leur impact sur Terre aurait contribué à l'apparition de la vie en apportant des molécules alors impossibles à synthétiser sur notre planète (REUELL 2019). Une autre proposition parle de l'apparition de la vie également sur un autre corps mais cette fois jusqu'à former des entités vivantes et non seulement des molécules prébiotiques qui auraient ensuite été apportées sur Terre via des météorites. Cette dernière hypothèse ne fait cependant que déplacer la question de l'origine de la vie sur une autre planète autour du Soleil ou même sur une planète dans un autre système solaire.

## 1.4 L'émergence de la vie : que conclure ?

Quelle que soit l'hypothèse envisagée pour expliquer le début de la vie sur Terre, elle a mené à l'apparition de la cellule et celle-ci s'est complexifiée avec l'ajout de machineries de traduction puis de transcription lors du passage de l'ARN à l'ADN comme méthode de stockage de l'information génétique.

Notons qu'on ne peut pas exclure le fait que certaines des hypothèses citées ont pu concourir ensemble à l'apparition du vivant et elles sont pour la plupart basées ou vérifiées grâce à des expériences en laboratoire. En effet, des conditions s'approchant de celles qui sont proposées comme celles de la Terre primitive (il y a 3 à 4 milliards d'année) ont été reproduites plusieurs fois en milieu contrôlé en laboratoire afin de tester si la formation de certaines molécules organiques comme les acides aminés, les peptides ou encore les vésicules était possible. Certaines de ces expériences ont fonctionné, notamment celle portant sur la génération d'acides aminés simples à partir d'ammoniac, de méthane et d'eau (MILLER 1953).

Il n'est finalement toujours pas possible de savoir avec certitude quel a été le scénario d'émergence de la vie car bien qu'il soit relativement facile de créer les "briques du vivant", comment peut-on s'assurer de la véracité de tels propos sans moyen direct de voir dans le passé ?

## 2 La vie primitive à l'aube de l'évolution

Dans cette partie, nous allons parler de la théorie darwinienne de l'évolution. La cellule est l'unité qui se reproduit et se multiplie et à cela s'ajoute la capacité à évoluer. Celle-ci est due à des erreurs (que l'on appelle "mutations") arrivant au hasard durant la réplication du matériel génétique de la cellule mère. L'une des cellules filles hérite alors du génome sans la nouvelle mutation et l'autre hérite de la version du génome la comportant. On note une exception : certaines mutations peuvent avoir lieu lors de la réparation de l'ADN dans une cellule après un dommage. Les deux cellules filles portent alors ces mutations une fois la division cellulaire effectuée après la réplication du génome.

Une mutation dans le génome peut avoir une incidence sur la viabilité de la cellule selon la position et le type de mutation. Par exemple, dans un centromère, la mutation peut entraîner la mort de la cellule ou des cellules filles à cause d'un mauvais appariement des chromatides soeurs lors de la division cellulaire. A l'origine de réplication du chromosome bactérien (ne contenant qu'une seule origine de réplication contrairement aux chromosome d'Archaea ou d'Eucaryotes), une mutation peut entraîner un échec immédiat de la réplication et de la division cellulaire. Il peut aussi y avoir directement un gène impacté si la mutation est située dans le gène lui-même, dans un exon (séquence codante) comme dans un intron (séquence non codante), ou encore dans son promoteur par exemple, et le produit du gène peut alors être non fonctionnel ou bien fonctionner différemment. Ainsi, la majorité des mutations sont délétères pour la cellule et sont donc perdues car les cellules les possédant ne survivent pas ou moins bien que les autres et/ou ne se reproduisent pas ou moins.

Il arrive cependant que certaines mutations aient un impact positif sur le fonctionnement de la cellule, par exemple en apportant une nouvelle fonctionnalité ou bien en permettant de mieux réagir à l'environnement. La mutation a alors tendance à être conservée au cours des générations et au fur et à mesure des mutations, la population de cellules s'adapte ainsi à son environnement. Enfin, on note que certaines mutations peuvent être fixées dans le génome d'une population via le phénomène de dérive génétique. Il s'agit de la fixation d'une version d'un gène (un allèle) par hasard dans une population. Une mutation peut donc être conservée sans qu'elle n'ait aucun impact avantageux pour l'organisme.

L'accumulation de mutations dans une population de cellules ou d'individus est l'un des phénomènes responsables de l'émergence de nouvelles espèces. On peut définir une espèce comme un groupe d'êtres vivants pouvant se reproduire entre eux et avoir une descendance viable et fertile. Cependant cette définition n'est pas

adaptée aux espèces qui ne font pas de reproduction sexuée comme la majorité de la diversité du vivant que représentent les bactéries et les Archaea. D'autres exemples comme certains parasites comme les *Taenia* ou certaines levures comme *Schizosaccharomyces pombe* font de la reproduction asexuée via des mécanismes de strobilation (segmentation du corps de l'individu en plusieurs parties qui vont chacune donner naissance à un ou plusieurs nouveaux individus) ou de scissiparité (séparation d'un individu pour donner deux clones).

Il est plus facile d'adopter le concept d'espèce phylogénétique, inférée à partir de la reconstruction des liens de parentés entre les organismes sur la base de la comparaison de leurs caractères homologues. Ces caractères peuvent être morphologiques, physiologiques ou moléculaires (séquences nucléiques ou protéiques). On peut alors utiliser un arbre phylogénétique pour représenter les relations de parenté entre les espèces étudiées : la racine modélise l'ancêtre commun à tous les organismes de l'arbre tandis que chaque feuille de l'arbre représente un organisme et chaque noeud interne (ou jointure de branches) est un organisme ou un groupe ancestral d'organismes.

L'utilisation des caractères moléculaires permet d'obtenir un fort pouvoir résolutif du fait de leur nombre et a l'avantage de pouvoir être applicable à l'ensemble des organismes vivants connus car ils descendent tous d'un ancêtre commun et contiennent chacun un jeu de séquences moléculaires utilisables pour les comparer. Dans ce cas, la phylogénie représente l'histoire évolutive des séquences moléculaires utilisées (phylogénie de gènes), qui sera utilisée pour inférer la phylogénie des espèces.

### **3 LUCA : le dernier ancêtre commun universel et la dichotomie fondamentale de la biologie**

LUCA pour "Last Universal Common Ancestor", aussi appelé le cenancêtre, est le dernier ancêtre commun à toutes les espèces sur Terre. Beaucoup d'hypothèses sont proposées quant à quoi ressemblait le cenancêtre. On pourrait le définir comme l'organisme ou l'ensemble d'organismes (ou de proto-cellules) duquel l'information génétique aurait été transmise jusqu'à nos jours au gré des mutations au fil des générations. Cet organisme est à l'origine de toutes les lignées desquelles sont issues toutes les espèces passées ou présentes qui ont été décrites par les humains. En effet, si l'on remonte dans le passé, on constate que toutes les espèces sont reliées les unes aux autres au fil des générations.

L'idée d'une séparation initiale entre deux groupes de proto-cellules pour former la lignée virale et la lignée à l'origine des proto-organismes microbiens semble séduisante. On aurait ainsi une lignée virale (considérée comme non vivante car incapable de se reproduire sans recourir à la machinerie de réplication d'un autre organisme) lointaine soeur de toutes les espèces issues de la lignée proto-microbienne.

Par la suite, la lignée bactérienne s'est écartée et enfin, les lignées Archaea et eucaryotes se sont formées (cf. figure 8). Notons que les lignées bactériennes et Archaea sont procaryotes, c'est-à-dire ne contiennent pas de noyau tandis que la lignée eucaryote contient un noyau stockant le matériel génétique et des compartiments spécialisés. Il s'agit là d'une différence fondamentale dans l'étude des espèces. La première mention de ces termes provient de CHATTON 1925.

Il ne peut pas être exclu que d'autres lignées d'organismes vivants aient émergées en plus des trois qui ont survécu jusqu'à nos jours. Nous n'avons cependant actuellement pas observé de trace de ces lignées car nous ne savons peut être pas les reconnaître mais il est probable que ces lignées n'aient simplement pas survécu suffisamment longtemps pour laisser des traces en quantité suffisante. L'arbre du vivant qui permet d'organiser le monde biologique est donc composé de trois branches principales qui se subdivisent ensuite de plus en plus pour obtenir une feuille par espèce actuellement connue.

## **3.1 Les organismes procaryotes**

### **3.1.1 Le règne des bactéries**

Les bactéries sont des êtres vivants unicellulaires (bien qu'il existe des bactéries filamenteuses s'associant les unes avec les autres) qui n'ont pas de compartiment intracellulaire (il existe ici aussi quelques exceptions à cette affirmation comme les thylakoïdes pour la photosynthèse des cyanobactéries (cf. 7.2)). Tous les composants de la cellule bactérienne comme le matériel génétique (l'ADN), les protéines et les acides gras baignent donc dans le même milieu intracellulaire (cf. figure 6).

Elles ont été observées pour la première fois en 1676 par Antoni van Leeuwenhoek au milieu d'autres espèces non bactériennes comme des ciliés (ou protozoaires) et des microalgues vertes (cf. 3.2) via des études réalisées grâce à des microscopes de sa fabrication (PORTER 1976) sur des échantillons d'eau récoltés

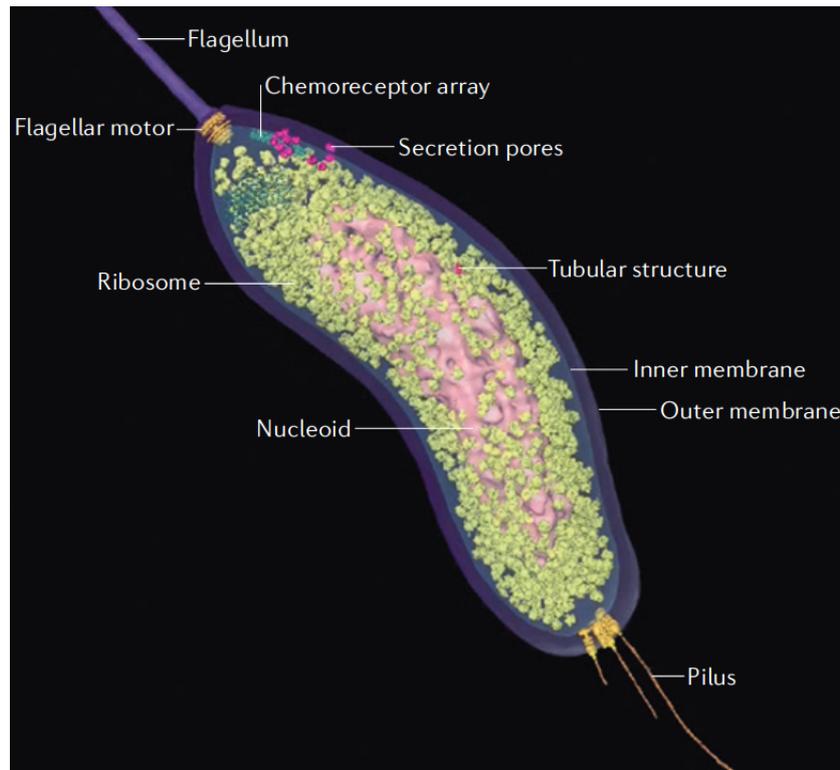


FIGURE 6 – Une cellule de la bactérie *Bdellovibrio bacteriovorus* imagée par ECT (Electron CryoTomography) (OIKONOMOU, CHANG et JENSEN 2016).

autour de son domicile. Par la suite, Louis Pasteur démontre à la fin du 19ème siècle que les bactéries jouent un rôle dans les infections et qu'elles ne proviennent pas d'une génération spontanée (van Leeuwenhoek à son époque n'était déjà pas partisan de la thèse de la génération spontanée d'après ses écrits).

Les plus anciens témoins de la présence de bactéries sur lesquels aucun doute n'est soulevé (contrairement à ceux datant de 3,5 milliards d'années évoqués précédemment en 1) remontent à 2,7 milliards d'années (fossiles de cyanobactéries, cf. 7.2) : la ou les lignées à l'origine des bactéries observées de nos jours sont apparues tôt dans l'histoire de la Terre (GARGAUD et al. 2009). On trouve des bactéries dans presque tous les milieux et environnements sur Terre et elles jouent des rôles essentiels comme notamment la décomposition de la matière organique ou la fixation du carbone. Leur vitesse de croissance est la plus rapide des trois règnes que nous évoquons dans cette partie, certaines espèces sont capables de doubler leur masse en l'espace de quelques dizaines de minutes seulement dans un environnement riche en matière organique. Elles n'ont pour la plupart qu'un seul

chromosome circulaire bien qu'on observe parfois aussi la présence d'autres petits éléments génétiques appelés plasmides.

Les cellules bactériennes sont entourées d'une paroi dont la composition varie entre les groupes bactériens, indépendamment des relations phylogénétiques entre les espèces (PARKS et al. 2018). On distingue les bactéries à Gram positif de celles à Gram négatif, distinction due à la coloration de la cellule bactérienne lors d'une coloration de Gram (technique mise au point en 1884 par Hans Christian Gram) révélant la paroi bactérienne (cf. figure 6). Les parois bactériennes se colorant en violet sont caractéristiques des bactéries à Gram positif dont la paroi est composée en majorité de peptidoglycanes et n'ont qu'une seule membrane (elles sont dites monodermes) et les bactéries se colorant en rose possèdent une paroi à faible teneur en peptidoglycanes et deux membranes (elles sont dites didermes) et font partie des bactéries à Gram négatif (TAIB et al. 2020).

### 3.1.2 Le règne des Archaea

Longtemps les espèces d'Archaea ont été considérées comme des bactéries extrémophiles à cause de la ressemblance physiologique entre elles. En effet les Archaea ne possèdent pas de noyau ni de compartimentation, tout comme les bactéries. Il s'agit cependant d'une lignée à part entière adaptée aux milieux dont la température, le taux de salinité ou encore l'acidité sont élevés. On les trouve notamment dans les sources chaudes ou les saumures.

Pour la première fois en 1968, une espèce est isolée depuis une source d'eau chaude acide par Thomas Brock. Cependant, leur découverte comme branche de l'arbre du vivant remonte à 1977 avec l'article de Woese et Fox (BALCH et al. 1977). Cet article présente la construction d'un arbre du vivant des procaryotes en utilisant des séquences de gènes codant pour des ARN ribosomiques (ARNr). Cet arbre a ainsi montré deux groupes distincts toujours d'actualité (cf. figure 8) : les bactéries d'un côté et les Archaea de l'autre (BALCH et al. 1977) mais il reste difficile d'estimer l'ancienneté de cette séparation (GARGAUD et al. 2009). Certaines hypothèses proposent que les Archaea soient à l'origine des eucaryotes (cf. 3.2). En effet, une étude très récente renforce l'idée que le groupe des Archaea Asgard possédait un proto-cytosquelette (les eucaryotes possèdent un cytosquelette) avant l'émergence des eucaryotes (RODRIGUES-OLIVEIRA et al. 2023), ce qui peut ainsi permettre de penser que c'est du groupe des Asgard que les eucaryotes ont divergés initialement après la première endosymbiose primaire (cf. 6.1).

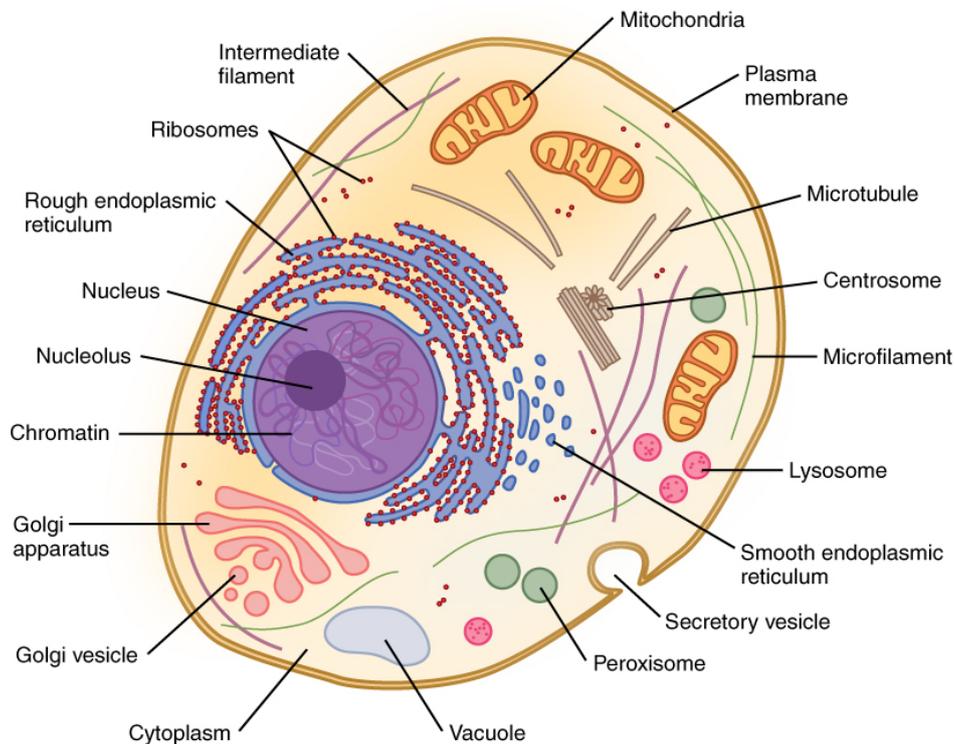


FIGURE 7 – Représentation d’une cellule eucaryote non photosynthétique avec les nombreux compartiments et certains des acteurs essentiels à la vie de la cellule (issue de e-biologie.fr). Les cellules eucaryotes photosynthétiques contiennent en plus un compartiment où se passe la photosynthèse (cf. 5, 7.2).

### 3.2 Les organismes eucaryotes

Tous les animaux, champignons et plantes en font partie, il s’agit d’un règne dont l’origine la plus acceptée est l’endosymbiose d’une bactérie par une autre cellule procaryote (cf. 6). On y trouve ainsi des organismes pluricellulaires mais aussi unicellulaires.

L’étymologie du terme eucaryote signifie vrai noyau (du grec ancien *eu* : bien, *karyon* : noyau). Il s’agit de la différence fondamentale entre les cellules eucaryotes et les cellules procaryotes (dont l’étymologie *pro* : avant, *karyon* : noyau du grec ancien signifie qu’il s’agit d’organismes sans noyau). Les cellules eucaryotes ont donc toutes un noyau qui stocke leur matériel génétique sous forme d’ADN mais elles ont aussi d’autres compartiments appelés organites qui sont délimités par

une membrane et qui ont différentes fonctions et dont voici quelques exemples (cf. figure 7) :

- le reticulum endoplasmique (RE) est composé d'une membrane reliée à la membrane du noyau. On distingue le RE lisse qui est impliqué dans la synthèse des acides gras et des phospholipides que l'on retrouve dans la bicouche lipidique formant la membrane des cellules. Le RE rugueux est quant à lui le siège de la synthèse des protéines via les ribosomes qu'il contient mais aussi du contrôle qualité et de l'envoi de ces protéines vers d'autres régions de la cellule.
- l'appareil de Golgi est le lieu dans lequel les protéines vont subir des modifications ou finir leur repliement. Elles sont ensuite envoyées vers leur destination finale dans la cellule ou bien sécrétées.
- les lysosomes sont de petites vésicules à l'intérieur desquelles le pH est très acide pour permettre l'activité catalytique de certaines enzymes qui lysent la plupart des macromolécules de la cellule.
- les mitochondries qui contiennent la chaîne respiratoire (cf. 7.1).
- les chloroplastes qui sont les organites responsables de la couleur verte des cellules photosynthétiques et qui contiennent la chaîne photosynthétique (cf. 7.2).

## 4 La variété des interactions biologiques entre espèces

Les règnes des bactéries, des Archaea et des eucaryotes contiennent chacun d'innombrables espèces, les estimations allant jusqu'à en proposer plusieurs dizaines de millions vivant actuellement sur Terre. On sait que de nouvelles espèces (découvertes ou non) peuvent apparaître ou disparaître chaque jour, à un rythme variable qu'il est difficile d'estimer de nos jours et ce, en partie à cause des changements climatiques majeurs actuels (CEBALLOS, EHRLICH, BARNOSKY et al. 2015 ; CEBALLOS, EHRLICH et DIRZO 2017). Selon les milieux et les écosystèmes, certaines de ces espèces sont en interaction entre elles et ce, malgré leur appartenance à différents règnes. Ainsi, les espèces peuvent mutuellement influencer sur la vie et la survie les unes des autres grâce à des mécanismes d'interactions bénéfiques ou délétères. Il existe plusieurs types d'interactions biologiques desquelles on peut citer par exemple :

- le mutualisme qui est une interaction mutuellement profitable entre orga-

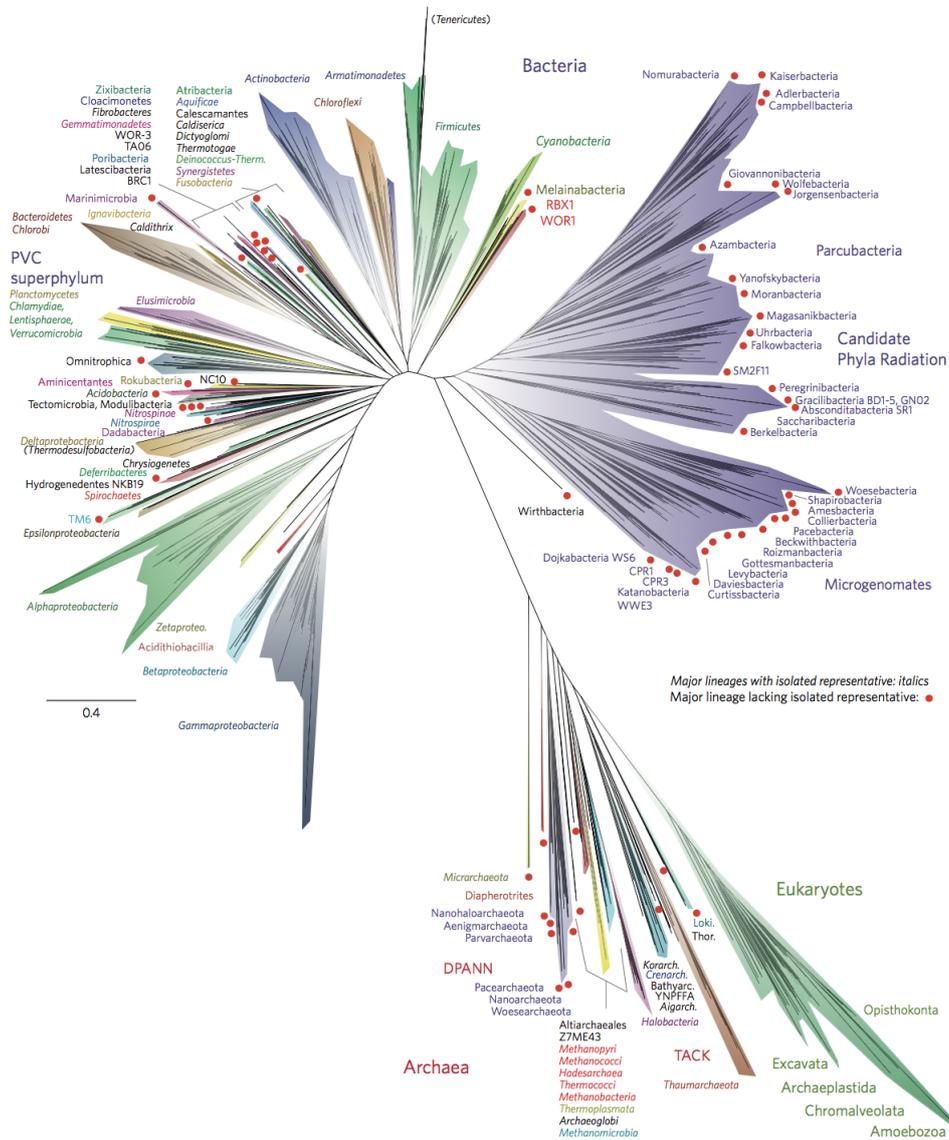


FIGURE 8 – Représentation de l’arbre du vivant réalisée avec 16 séquences d’ARN ribosomiques (ARNr) de 92 phyla bactériens, 26 phyla d’Archaea et les 5 super-groupes des eucaryotes (HUG et al. 2016).

nismes d’espèces différentes, c’est-à-dire que toutes les espèces impliquées dans l’interaction retirent un bénéfice. L’interaction n’est cependant pas obligatoire et les espèces peuvent donc survivre sans. Il peut s’agir par exemple de l’interaction entre un pollinisateur et une plante à fleur (comme les abeilles et les fleurs) ou encore les oiseaux comme les merles ou les étour-

neaux dispersant les graines d'une plante produisant des fruits comme le merisier.

- le commensalisme est une interaction entre différentes espèces dans laquelle l'un des organismes (appelé l'organisme commensal) bénéficie de l'interaction et l'autre ne subit aucun impact. On peut par exemple citer les petits crabes pinnothères commensaux des moules.
- la symbiose est une interaction obligatoire et/ou permanente où les deux organismes vivent et évoluent ensemble. Il s'agit généralement d'une interaction mutualiste mais il arrive qu'elle évolue vers le parasitisme. Il peut par exemple s'agir de la mycorhize, symbiose entre un champignon et les racines d'une plante ou encore de la symbiose entre une espèce animale et des bactéries dans son tube digestif. Un cas particulier notable est l'endosymbiose, où une espèce vit à l'intérieur d'une autre (cf. 5).
- le parasitisme est l'interaction dans laquelle l'une des espèces bénéficie de la relation tandis que l'autre en souffre. C'est par exemple le cas de la relation commensale citée précédemment entre les pinnothères et les moules où les crabes se nourrissent de la moule lorsque la nourriture vient à manquer.

Il existe d'autres catégories d'interactions comme la compétition ou la prédation. On note que dernièrement, les recherches montrent que les interactions entre espèces ne sont pas statiques et évoluent au cours du temps en fonction de l'environnement et du cycle de vie de certaines espèces.

## 5 Former des espèces en assemblant d'autres espèces : les endosymbioses

Le phénomène d'endosymbiose est une association (ou co-habitation) entre deux espèces qui bénéficient chacune de la relation (une forme de mutualisme, cf. 4), l'une des espèces se trouvant dans le corps de l'autre (le corps est en fait la cellule si l'organisme est unicellulaire). Lorsque l'espèce phagocytée vit à l'intérieur de son hôte, il s'agit d'un endocytobionte et on parle alors du phénomène d'endocytobiose. C'est par exemple le cas de nombreuses bactéries qui vivent à l'intérieur d'eucaryotes. On peut aussi citer l'association de coraux avec des bactéries (MAIRE, BLACKALL et OPPEN 2021) ou d'insectes avec d'autres bactéries (MCCUTCHEON, BOYD et DALE 2019). D'autres exemples, cette fois impliquant des eucaryotes comme endosymbiotes peuvent être évoqués :

- Les espèces *Hydra viridissima* et *Hydra vulgaris* sont des Cnidaires qui vivent en eau douce. Leur couleur verte (cf. figure 9) provient, pour chacune des deux espèces, d'algues vertes de la famille des Chlorelles qu'elles ont absorbées lors d'une endosymbiose (ou endocytobiose dans ce cas). Ces espèces sont particulièrement étudiées pour mieux comprendre les premières étapes d'une endosymbiose car il a été déterminé que l'endosymbiose de l'espèce *Hydra vulgaris* était bien plus récente que celle de l'espèce *Hydra viridissima* (ISHIKAWA, YUYAMA et al. 2016). En effet, toutes les souches de *Hydra viridissima* maintenues au "National Institute of Genetics" à Mishima au Japon montrent une endosymbiose tandis que seules 2 des 25 souches de *Hydra vulgaris* en comportent une (12 des 23 souches non symbiotiques sont cependant capables d'héberger des algues jusqu'à deux générations par reproduction asexuée (bourgeonnement) après l'ingestion) (ISHIKAWA, Hiroshi SHIMIZU et al. 2016).
- Certains coraux (colonies de polypes, une forme physiologique des Cnidaires, dans un exosquelette calcaire) peuvent former une endocytobiose avec des cellules d'algues de la famille des Xanthelles (faisant partie des Dinoflagellés) capables de faire la photosynthèse (STANLEY et SWART 1995; MEYER et WEIS 2012). Les échanges entre les deux espèces sont alors les suivants : les algues se nourrissent des déchets organiques et azotés du polype et fournissent à ce dernier des sucres issus de la photosynthèse.

## 6 La théorie endosymbiotique : vers la formation des lignées eucaryotes

La théorie endosymbiotique a pour vocation de proposer une explication à la présence des mitochondries dans les cellules eucaryotes et des chloroplastes dans les lignées eucaryotes photosynthétiques. Chloroplastes et mitochondries dériveraient d'endosymbiotes s'étant progressivement intégrés à la cellule hôte pour en devenir un compartiment intracellulaire (ARCHIBALD 2015). Ce compartiment est délimité par une membrane (pouvant être formée de plusieurs couches), il contient quelques gènes rémanents de la cellule initialement ingérée et a une fonction spécialisée. L'endosymbiote est appelé ainsi car il ne peut dès lors plus survivre seul à l'extérieur de son hôte.

En effet, il a été remarqué que ces deux organites possèdent des gènes et ressemblent étrangement à certaines bactéries, aussi bien du point de vue de leur physiologie que de leur contenu en gènes (tout en sachant que le nombre de



FIGURE 9 – Photos de spécimens d'*Hydra viridissima* par Peter Schuchert le 25 mai 2009 (gauche) et par Frank Fox le 4 mars 2012 (droite). Sources : marinespecies.org (gauche), mikro-foto.de (droite)

gènes dans les organites est faible par rapport à celui des bactéries auxquelles elles ressemblent, voir 1.6.3). C'est Constantin Sergueïevitch Merejkovski qui propose le premier dans deux articles en 1905 puis en 1910 que les chloroplastes dérivent de bactéries anciennement ingérées par les cellules de plantes grâce à plusieurs observations de la similitude de la structure des thylakoïdes de plantes et des thylakoïdes de cyanobactéries. Cette théorie est popularisée au sein de la communauté scientifique seulement plusieurs dizaines d'années plus tard grâce à l'apport de preuves visuelles de la présence d'ADN dans le chloroplaste de microalgue du genre *Chlamydomonas* (RIS et PLAUT 1962) puis par une revue de la théorie depuis sa proposition (SAGAN 1967).

## 6.1 Une première endosymbiose primaire : l'origine de la mitochondrie

Les endosymbioses que l'on suppose être à l'origine de la mitochondrie des eucaryotes et du chloroplaste de certaines lignées photosynthétiques eucaryotes

(cf. figure 10) sont nommées endosymbioses primaires. Un évènement d'endosymbiose primaire se joue entre une bactérie (le futur symbiote) et une cellule hôte (aucun des deux organismes n'ayant jamais subi d'endosymbiose auparavant). Il peut par exemple s'agir d'une bactérie ingérée par une Archaea ou d'une bactérie ingérée par un protiste (eucaryote unicellulaire ou pluricellulaire à tissu non spécialisé). Par la suite, la bactérie ingérée est maintenue à l'intérieur de l'hôte et une forme de coopération puis de contrôle du symbiote par l'hôte s'éveille. Le symbiote évolue ainsi en parallèle de l'hôte, des échanges de ressources et des mouvements de matériel génétique du symbiote vers l'hôte se mettent aussi en place. Le symbiote finit par devenir une partie intégrante et nécessaire à la survie de la cellule hôte (tout comme le symbiote ne peut plus vivre hors de son hôte).

La lignée eucaryote que nous avons évoquée en 1.3.3 provient d'un évènement d'endosymbiose primaire il y a plus d'1,7 milliards d'années (DACKS et al. 2016). Il est communément admis que l'endosymbiose a eu lieu entre une bactérie ancestrale proche des alphaproteobactéries et une cellule hôte proche des Archaea. Les positions exactes de ces cellules ancestrales par rapport aux alphaprotéobactéries et aux Archaea restent aujourd'hui toujours sujettes à débat (ARCHIBALD 2015 ; ROGER, MUÑOZ-GÓMEZ et KAMIKAWA 2017). Cet évènement a pu se produire si on fait l'hypothèse que les alphaprotéobactéries étaient des proies de la cellule hôte proto-eucaryote. Il suffit alors que l'alphaprotéobactérie ne soit pas détruite et ce par exemple, grâce à un mécanisme de défense mis en place contre l'hôte. La bactérie se maintient donc dans la cellule hôte et au fil des générations, elle évolue au sein et avec son hôte via des pertes et des transferts de gènes et de molécules organiques pour devenir les mitochondries (cf. 8.1).

## **6.2 La seconde endosymbiose primaire : l'apparition du chloroplaste**

Les plantes terrestres à fleurs (Angiospermae), les arbres (Streptophyta) et les algues vertes (Chlorophyta) formant la lignée des Viridiplantae ainsi que les algues rouges (Rhodophyta) et les algues bleues-vertes (Glaucophyta) proviennent d'une seconde endosymbiose primaire il y a environ 1,5 milliards d'années (YOON, HACKETT, CINIGLIA et al. 2004). Cette fois-ci, il s'agit d'une endosymbiose entre un eucaryote hétérotrophe (une cellule contenant donc un noyau et des mitochondries) et une cyanobactérie (bactérie de couleur verte capable de générer de l'ATP grâce à la photosynthèse) (PONCE-TOLEDO et al. 2017 ; SIBBALD et ARCHIBALD 2020). Au fil du temps et des générations, la cyanobactérie ingérée évolue alors au sein de la cellule eucaryote pour devenir le chloroplaste. Cette

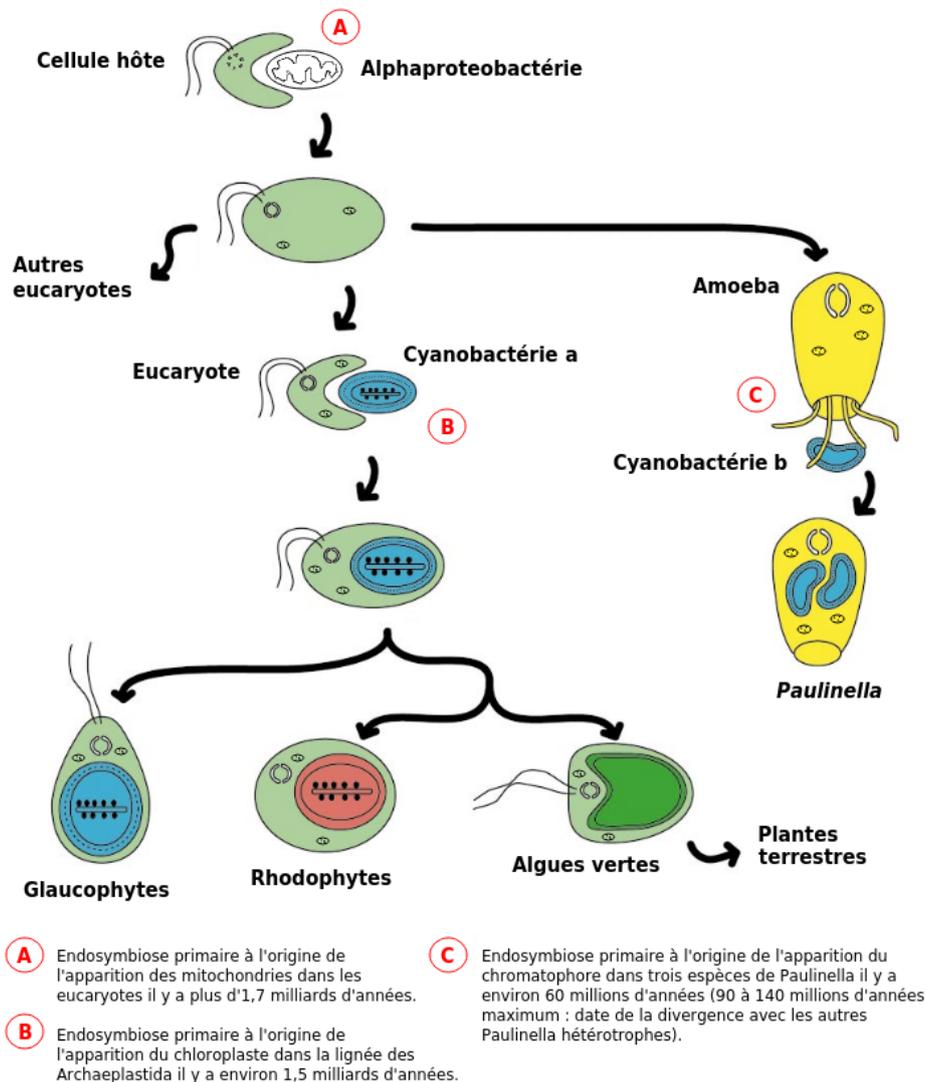


FIGURE 10 – Les trois endosymbioses primaires connues à ce jour qui ont amené à l'apparition des mitochondries et des organites responsables de la photosynthèse chez les eucaryotes (chloroplastes et chromatophores). Les Archaeplastida (Glaucophytes, Rhodophytes, algues vertes ou chlorophytes et plantes terrestres) contiennent des mitochondries et un ou plusieurs chloroplastes issus des endosymbioses primaires successives de deux bactéries (alphaprotéobactérie et cyanobactérie respectivement) (adapté de KEELING 2004).

branche de l'arbre du vivant est appelée Archaeplastida et c'est ainsi que nous la nommerons tout au long de ce manuscrit (cf. figure 10).

### 6.3 Les endosymbioses secondaires à l'origine d'autres algues

Les endosymbioses secondaires cette fois ont lieu entre deux eucaryotes unicellulaires, l'un hétérotrophe et l'autre déjà muni d'un chloroplaste. Le même phénomène de maintien du symbiote que pour les endosymbioses primaires a lieu, tout comme la co-évolution des deux partenaires. Ces endosymbioses proviennent :

- de l'ingestion d'une algue rouge (Rhodophyte) par un eucaryote non photosynthétique il y a environ 1,3 milliards d'années (YOON, HACKETT, CINIGLIA et al. 2004). Cet évènement est à l'origine des familles d'algues comme les Hétérocontes (ou Straménopiles). Il s'agit de la lignée dont font partie les Diatomées (Bacillariophytes) et les algues brunes (Phaeophyceae) (cf. figure 12). Ces espèces contiennent un ou plusieurs plastes qui sont entourés de plusieurs membranes (dont deux issues de l'endosymbiose à l'origine des Archaeplastida (cf. 6.2), l'autre provenant de l'endosymbiose décrite ici (cf. figure 11) (PETERSEN et al. 2014)).
- de l'ingestion d'une algue verte unicellulaire par un eucaryote non photosynthétique menant à l'apparition des Euglenides (TOMECKOVA et al. 2020) et des Chlorarachniophytes (HIRAKAWA et al. 2011) lors de deux endosymbioses indépendantes). Pour les Euglenides, les indices montrent qu'il s'agirait de l'ingestion d'une Pyramimonadale par une Euglenozoa (cf. 6.4).

Une différence entre les endosymbioses primaires et secondaires que l'on peut observer facilement est le nombre de membranes lipidiques autour de l'endosymbiote. Lors d'une endosymbiose primaire, on observe généralement deux membranes autour du symbiote tandis que lors d'une endosymbiose secondaire, on observe une membrane de plus que précédemment autour du symbiote et chaque endosymbiose supplémentaire a le potentiel d'en ajouter une nouvelle. Il arrive aussi d'observer qu'une ou plusieurs membranes soient perdues au cours de l'évolution (WETHERBEE et al. 2019).

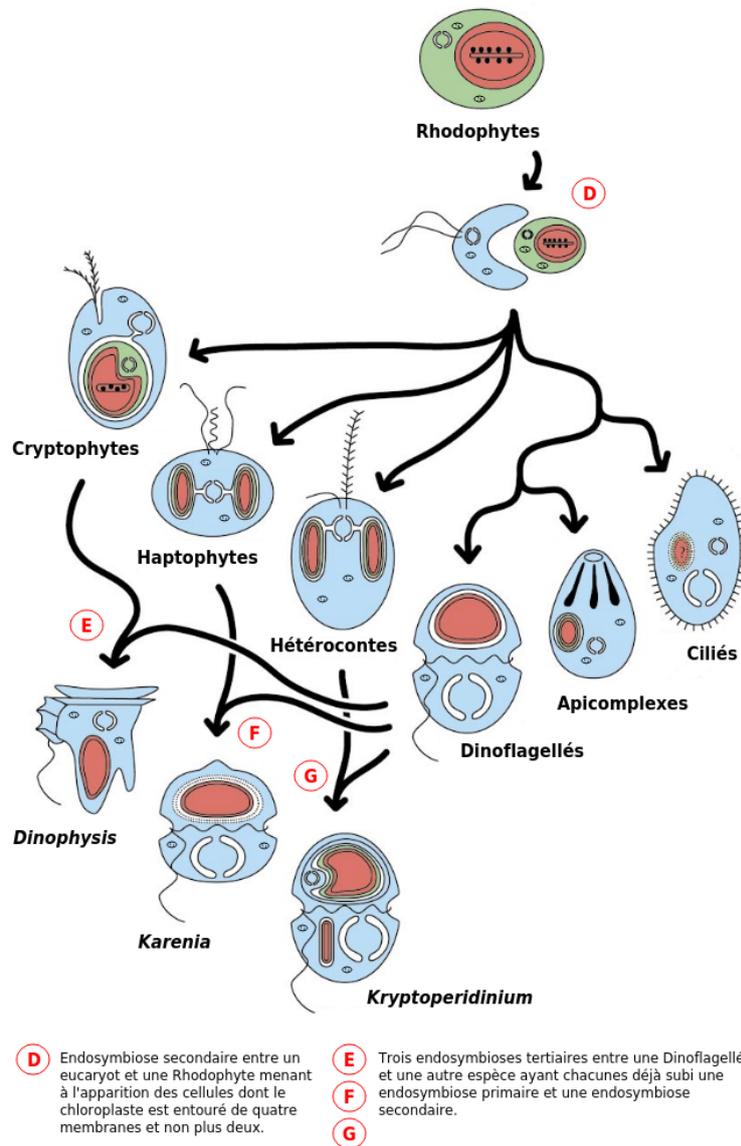


FIGURE 11 – Les lignées ayant subi des endosymbioses secondaires et tertiaires et dont le chloroplaste provient initialement d'une Rhodophyte. Certains plastes acquis ont subi de fortes réductions de leur taille, de leur génome et de leur activité comme c'est le cas chez certaines Alveolates (notamment les Ciliés) (adapté de KEELING 2004).

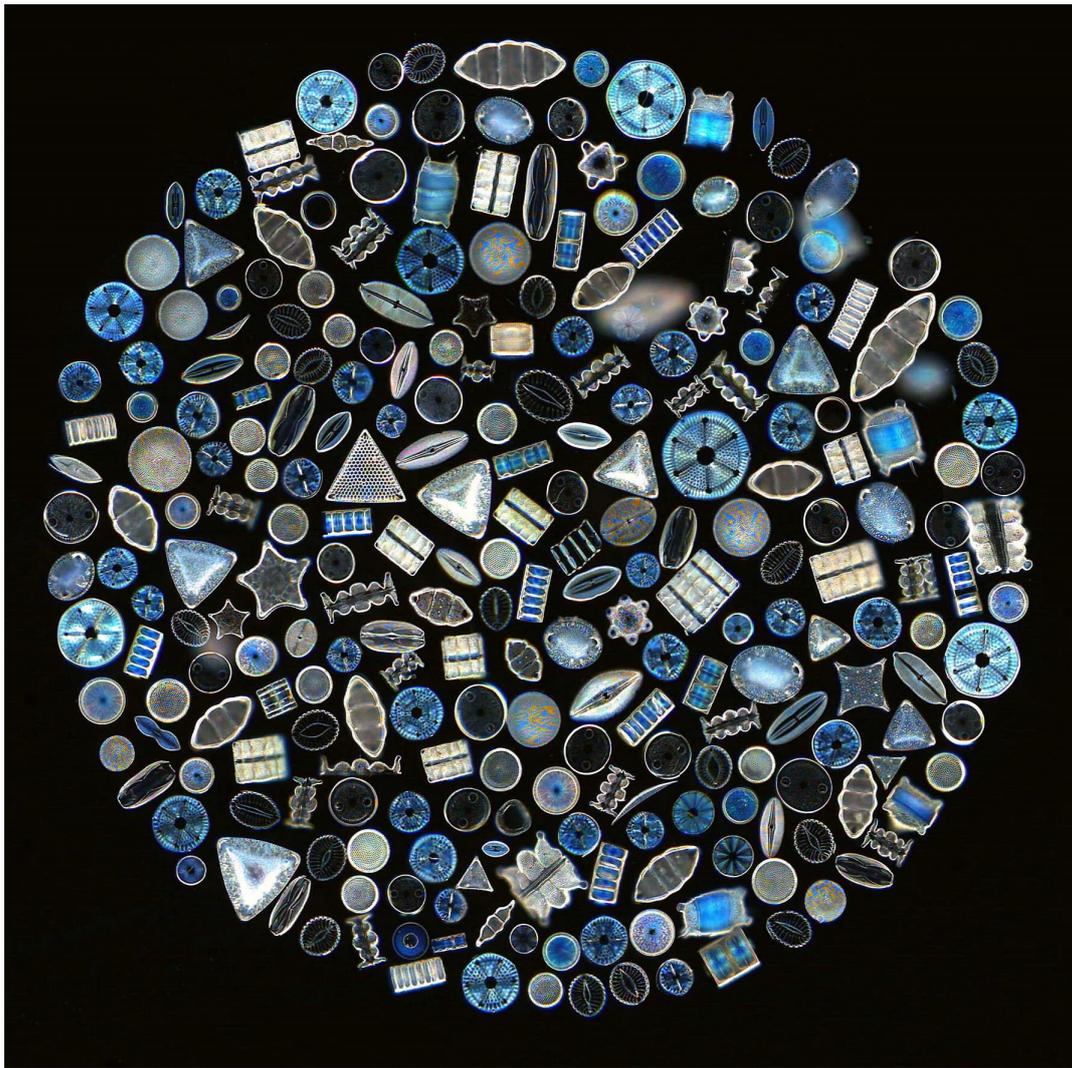


FIGURE 12 – Arrangement de spécimens de Diatomées à but artistique ("Diatom arrangement" par Klaus Kemp).

#### 6.4 D'autres endosymbioses à l'origine de la variété des plastes

Il existe encore beaucoup d'espèces en plus de celles dont nous avons déjà évoqué l'origine dans les paragraphes précédents qui possèdent des plastes. C'est notamment le cas de trois espèces issues d'une endosymbiose primaire récente à l'échelle du vivant (quelques dizaines de millions d'années) d'une cyanobactérie par un eucaryote non photosynthétique : les espèces photosynthétiques *Paulinella chromatophora*, *Paulinella micropora* et *Paulinella longichromatophora* (MARIN,

NOWACK et MELKONIAN 2005 ; DELAYE, VALADEZ-CANO et PÉREZ-ZAMORANO 2016). Ces espèces sont étudiées notamment dans l'optique de mieux comprendre l'établissement et le maintien d'un endosymbiote à l'intérieur de l'hôte, au même titre que les hydres évoquées en 5 (cf. figure 10).

Pour finir, trois autres endosymbioses indépendantes menant à des lignées d'organismes hébergeant des plastes ont aussi eu lieu indépendamment les unes des autres (cf. figure 11). Il s'agit d'endosymbioses tertiaires pour les lignées des *Karenia* (BERCEL et KRANZ 2019), des *Kryptoperidinium* (FIGUEROA et al. 2009) et des *Dinophysis* (BHATTACHARYA, YOON et HACKETT 2004 ; YOON, HACKETT, VAN DOLAH et al. 2005), à chaque fois avec une cellule de Dinoflagellé (cf. 6.4). Certains proposent aussi que certaines de ces endosymbioses soient des endosymbioses quaternaires (BHATTACHARYA, YOON et HACKETT 2004).

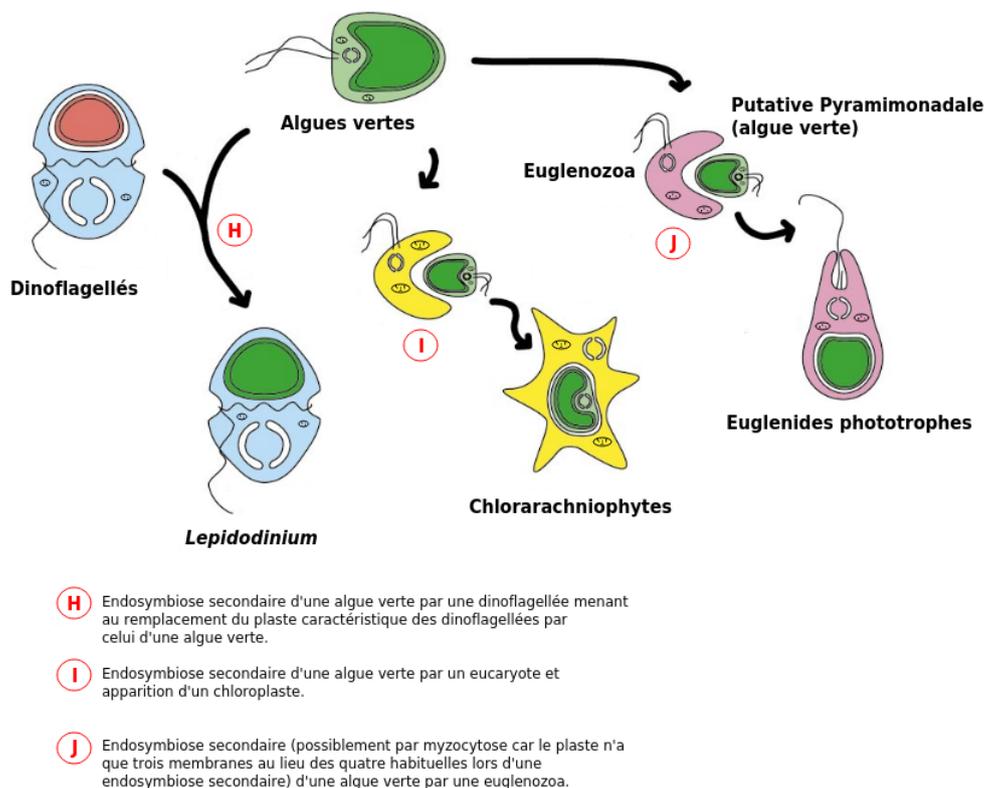


FIGURE 13 – Schéma de trois endosymbioses secondaires distinctes impliquant une algue verte et un eucaryote (adapté de KEELING 2004).

## 7 Les endosymbioses apportent de nouvelles compétences

### 7.1 La respiration

Au sein des mitochondries (cf. 6.1) a lieu la respiration, processus permettant d'alimenter la cellule en énergie. Il s'agit d'une suite de réactions chimiques aussi appelée "phosphorylation oxydative" qui utilise des composés organiques sources de carbone réduit comme le succinate mais également le NADH comme réducteurs (qui sont donc des donneurs d'électrons). L'oxydation de cette source de carbone réduit produit un transfert d'électrons entre trois complexes protéiques principaux qui sont insérés dans la membrane interne de la mitochondrie (cf. figure 14.B) et à chaque transfert d'un électron, un proton est pompé depuis la matrice de la mitochondrie vers l'espace intermembranaire. Il se crée ainsi un gradient électrochimique de protons entre ces deux zones. Par la suite, un autre complexe protéique transmembranaire, l'ATP synthase (cf. figure 14.B) utilise l'énergie de ce gradient électrochimique de protons pour produire de l'ATP (Adenosine TriPhosphate) à partir d'ADP (Adenosine DiPhosphate) et de phosphate. L'ATP est la source d'énergie de la cellule car la liaison du troisième phosphate libère de l'énergie lorsqu'elle est rompue. Ces molécules sont ensuite transportées vers les autres compartiments intracellulaires (comme le noyau ou le cytosol par exemple).

La respiration peut avoir lieu en permanence, bien que son cycle ne soit pas réellement connu. Elle est dépendante de l'environnement pour les nutriments desquels sont extraits le carbone réduit et de l'oxygène comme accepteur terminal d'électrons. Elle a donc lieu dans des environnements aérobies contenant une source de carbone réduit.

### 7.2 La photosynthèse

La photosynthèse est un processus permettant d'alimenter la cellule en molécules organiques comme les sucres. Elle fonctionne en utilisant l'énergie de photons issus d'une source lumineuse, le Soleil ou une ampoule par exemple. C'est donc un processus qui n'a lieu que le jour en conditions naturelles contrairement à la respiration qui peut être permanente (cf. 7.1) : le cycle de la photosynthèse est diurne et ne dépend pas d'autres éléments. C'est dans le chloroplaste (cf. 6.2), et plus précisément au niveau des thylakoïdes (sortes de sacs aplatis contenant les pigments assimilateurs ou pigments photosynthétiques comme les chlorophylles a et b, pigments verts ou bleus ou encore le carotène, pigment orange ou jaune

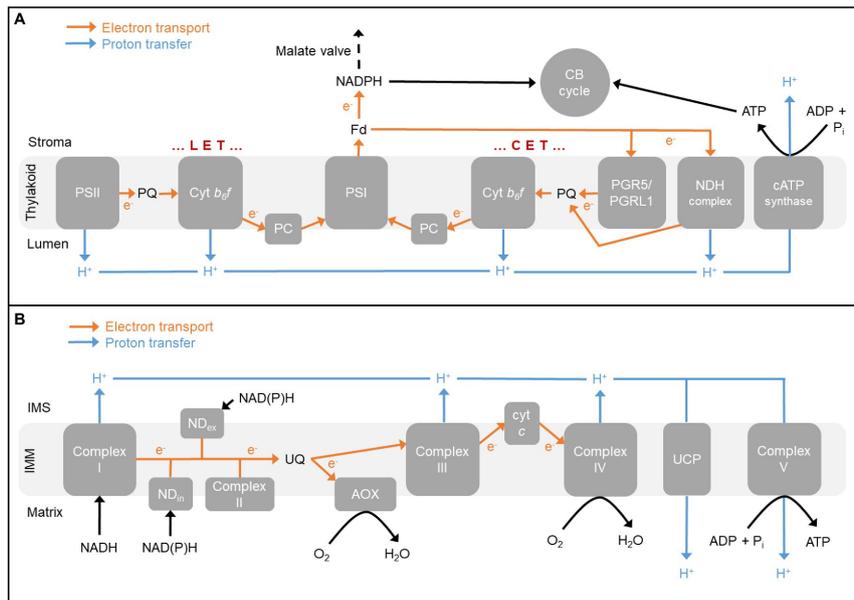


FIGURE 14 – Les chaînes des transport des électrons dans les mitochondries et les chloroplastes. A : Représentation de la chaîne de transport des électrons et des acteurs de la chaîne photosynthétique dans le chloroplaste. B : Représentation de la chaîne de transport des électrons dans le processus de respiration ayant lieu dans les mitochondries (CHADEE et al. 2021).

permettant de capter les photons nécessaires à la photosynthèse) qu'ont lieu ces échanges chimiques (cf. figure 14.A). L'absorption de photons par les pigments permet de déclencher une chaîne de transferts d'électrons entre des complexes protéiques enchâssés dans la membrane du thylakoïde. Lors de ces transferts, des protons sont pompés depuis le stroma vers le lumen des thylakoïdes, créant un gradient électrochimique de protons. Celui-ci est consommé par une ATP synthétase (complexe protéique également transmembranaire) pour produire des molécules d'ATP, sur le même principe que dans le processus de la respiration. Cependant ici, l'étape terminale du transfert photosynthétique d'électrons (la réduction du NADP en NADPH) permet, en combinaison avec les molécules d'ATP produites par l'ATP synthétase, la synthèse de sucres (des triose-phosphates) qui sont une source de carbone réduit utilisée par l'ensemble des compartiments intracellulaires de la cellule végétale.

## 8 Les impacts de l'évolution sur l'endosymbiose et les innovations post-endosymbiotiques

### 8.1 Transferts et pertes de gènes

Lors d'une endosymbiose, un long processus d'évolution entre l'endosymbiote et l'hôte se met en place et c'est avant tout par des pertes et des transferts de gènes depuis le symbiote vers le noyau de l'hôte qu'on le remarque. En effet, le symbiote perd les gènes qui sont redondants avec ceux de l'hôte ainsi que les gènes qui sont devenus inutiles dans l'environnement intracellulaire (tels que ceux codant pour des protéines de la paroi cellulaire des bactéries). D'autres gènes sont transférés dans le génome nucléaire de l'hôte et seules une dizaine à quelques dizaines de gènes sont de nos jours conservées dans la mitochondrie tandis qu'une centaine (ou plusieurs centaines selon les espèces) est toujours conservée dans le chloroplaste. Ces gènes codent principalement des sous-unités des complexes protéiques respiratoires et photosynthétiques, des protéines impliquées dans la transcription et dans la traduction, des ARN de transfert (ARNt) et des ARN ribosomiques.

### 8.2 La mosaïque génétique des complexes respiratoires et photosynthétiques

Les complexes photosynthétiques et respiratoires sont formés de plusieurs sous-unités et en raison des transferts de gènes évoqués précédemment (cf. 8.1), les gènes codant pour ces sous-unités peuvent être situés dans le génome nucléaire ou bien dans les génomes des organites (initialement, tous ces gènes étaient contenus dans le génome de l'endosymbiote). Ainsi, les complexes des chaînes respiratoires et photosynthétiques sont donc des mosaïques génétiques composées de protéines codées par des gènes du noyau et de protéines codées dans l'organite (cf. figure 16).

L'une des hypothèses pour expliquer le maintien de gènes dans le génome de l'organite serait la forte hydrophobicité de certaines des protéines très abondantes qu'ils codent, dont la translocation à haut débit serait difficile à assurer par les translocons de l'enveloppe de l'organite (cf. 8.3 et figure 15), ce qui ouvre le risque d'une traduction cytosolique conduisant à l'accumulation d'agrégats protéiques cytotoxiques dans le cytosol (BJORKHOLM et al. 2015). S'ajoutent à cela les différences entre les codes génétiques des organites et du noyau. Certains gènes transférés du génome des organites à celui du noyau pourraient être traduits de façon erronée, et alors perdre leur fonction et entraîner la mort de la cellule (contre-sélection de cet évènement de transfert) (GREY 2005). Enfin, la nécessité

d'avoir une régulation rapide de l'expression des gènes codant pour certains composants de la chaîne de transfert d'électrons pourrait aussi expliquer leur maintien dans le génome de l'organite car il est plus rapide de réguler un gène au sein du même compartiment qu'un gène codé ailleurs (ALLEN 2015).

### 8.3 La machinerie d'import des protéines

Les protéines codées dans le noyau et traduites dans le cytosol nécessitant un import dans le chloroplaste ou la mitochondrie chez les Archaeplastida sont dotées d'un peptide d'adressage à leur extrémité 5' (petite portion d'environ 60-80 acides aminés). Ce peptide est reconnu par un appareil de translocation différent selon l'organite visée (ARCHER et KEEGSTRA 1990).

La machinerie de translocation (un complexe protéique) permettant d'importer une protéine depuis l'extérieur de la mitochondrie vers la matrice s'appelle TOM/TIM pour "Translocase of Outer Membrane" et "Translocase of Inner Membrane". De même, dans le chloroplaste, un complexe protéique de translocation appelé TOC/TIC pour "Translocase of Outer Chloroplast membrane" et "Translocase of Inner Chloroplast membrane" joue le même rôle (CHOTEWUTMONTRI, HOLBROOK et BRUCE 2017; WIEDEMANN et PFANNER 2017).

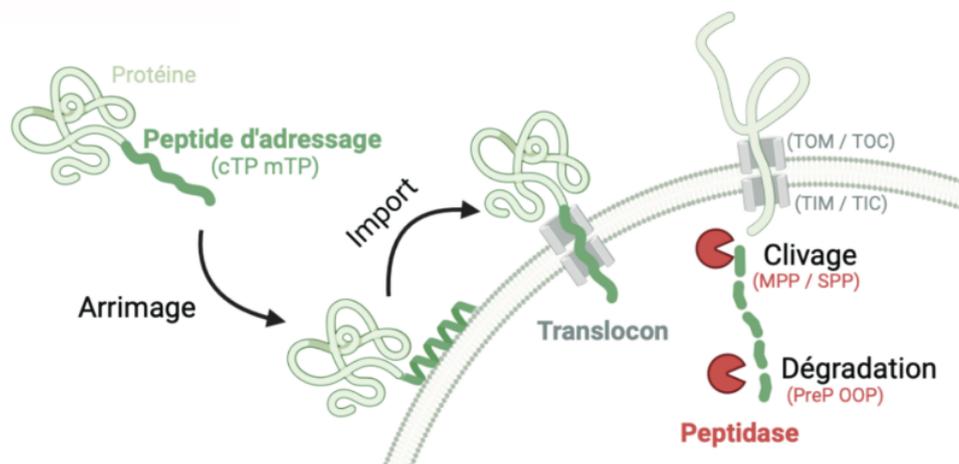


FIGURE 15 – Le mécanisme d'adressage des protéines aux organites (adapté du manuscrit de thèse de Clotilde Garrido (GARRIDO 2021)).

Le peptide d'adressage vient interagir avec la membrane des organites de la

façon suivante : c'est un peptide dont le repliement prend la forme d'une hélice amphiphile, c'est-à-dire qu'une face de l'hélice est hydrophobe (de par sa composition en acides aminés hydrophobes) tandis que l'autre face est hydrophile et chargée positivement et peut donc interagir préférentiellement avec les phosphates chargés négativement à la surface de la membrane d'un organite. Par la suite, lorsque le peptide entre en contact avec la machinerie de translocation adaptée, la protéine est importée puis le peptide d'adressage est clivé et dégradé et enfin la protéine est libérée dans l'organite (cf. figure 15). Ce mécanisme utilise de l'énergie et dépend donc d'ATP. Il a récemment été proposé que ce mécanisme d'import des protéines dans les organites provienne du détournement des mécanismes de résistance aux peptides antimicrobiens (WOLLMAN 2016 ; GARRIDO et al. 2020).

## 8.4 Une ploïdie variable au sein des cellules

On note qu'en général la ploïdie des organites est beaucoup plus importante que celle du noyau (jusqu'à un facteur 100) : certains eucaryotes ont un génome nucléaire haploïde, la plupart sont diploïdes mais certains ont une polyploïdie plus élevée comme le maïs qui est hexaploïde. Par contraste, les génomes des organites sont extrêmement polyploïdes, de l'ordre de 100 copies de leur génome dans chaque organite. Ce constat implique que les gènes codés dans les organites sont potentiellement transcrits en quantités plus importantes que ceux codés dans le noyau et qu'il faut un mécanisme d'ajustement de l'expression protéique entre les deux compartiments pour aboutir à la stoechiométrie appropriée pour l'assemblage des différentes sous-unités des complexes protéiques respiratoires et photosynthétiques (Y. CHOQUET et WOLLMAN 2002 ; WOODSON et CHORY 2008).

## 8.5 Le contrôle par épistasie de la synthèse

L'interaction entre deux ou plusieurs gènes est appelée l'épistasie. Il s'agit d'un mécanisme dans lequel un ou plusieurs gènes masquent ou empêchent l'expression d'un ou plusieurs autres gènes.

Le Contrôle par Épistasie de la Synthèse (ou CES), quant à lui, est le phénomène impliquant la modulation de la traduction dans l'organite d'un ARNm de sous-unité d'un complexe protéique en fonction de l'état d'assemblage de ce même complexe (Y. CHOQUET et WOLLMAN 2009). Il a été particulièrement étudié chez la microalgue *Chlamydomonas reinhardtii*.

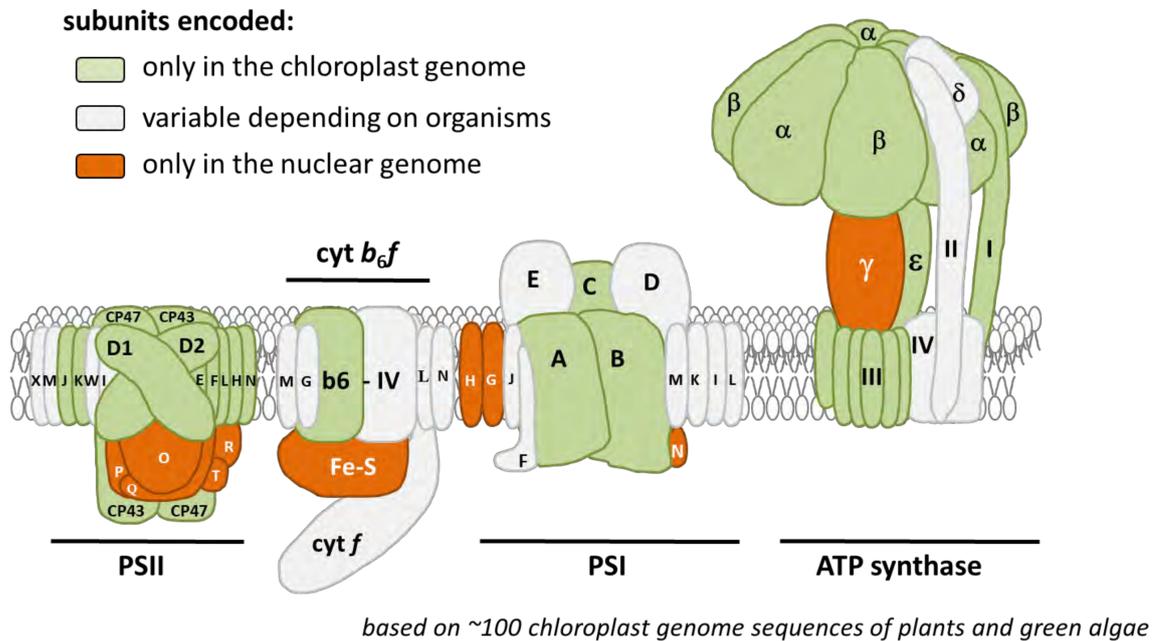


FIGURE 16 – Provenance des sous-unités des complexes protéiques impliqués dans la photosynthèse dans le chloroplaste (adapté du manuscrit de thèse de Domitille Jarrige (JARRIGE 2019)).

La plupart des complexes protéiques de la chaîne respiratoire et de la chaîne photosynthétique présente ce genre de CES. La forme la plus observée de CES est la régulation négative, c'est-à-dire que lors de l'expression d'une protéine (ici une sous-unité d'un complexe), sa forme non assemblée empêche sa propre production. Lorsque le complexe est assemblé et qu'il y a donc moins de sous-unités non assemblées, la production de nouvelles sous-unités recommence (Y. CHOQUET, STERN et al. 1998 ; WOSTRIKOFF et al. 2004 ; DRAPIER, RIMBAULT et al. 2007 ; BOULOUIS, RAYNAUD et al. 2011). Le cas d'une régulation positive est rarement observé (DRAPIER, RIMBAULT et al. 2007).

Ce mécanisme de régulation est observé aussi bien dans les mitochondries de levure que dans les chloroplastes des plantes terrestres et des microalgues (CALDER et McEWEN 1991 ; ZAMBRANO et al. 2007).

## 8.6 Les acteurs nucléaires de la régulation des génomes des organites

Les cellules gouvernent l'expression des gènes restant dans les organites et donc l'activité de l'organite en elle-même grâce à l'apparition au cours de l'évolution de protéines capables de reconnaître et de se fixer spécifiquement aux ARNm des organites. C'est dans le génome nucléaire de la cellule que sont codés les facteurs de régulation post-transcriptionnelle de l'expression génétique des organites. Ces gènes sont transcrits dans le noyau, traduits dans le cytosol, puis importés à l'intérieur des organites via la machinerie de translocation adéquate (cf. 8.3).

Dans ce manuscrit, ces facteurs nucléaires seront appelés OTAF pour "Organelar Trans-Acting Factor" mais il est possible de trouver d'autres dénominations dans la littérature comme simplement TAF pour "Trans-Acting Factor" ou encore ROGE pour "Regulator of Organelle Genome Expression".

## 9 À la découverte de grandes familles d'OTAFs

L'une des premières mentions de protéines codées dans le génome nucléaire agissant sur l'expression de protéines mitochondriales ("nuclear-encoded factors") remonte à 1986 dans une étude portant sur l'expression du cytochrome C dans la levure *Saccharomyces cerevisiae* (MCEWEN et al. 1986). Les auteurs montrent en effet que certaines sous-unités protéiques sont absentes de la cytochrome C oxydase de la levure dans une partie de leurs mutants complémentés et ils proposent d'utiliser ces mutants pour comprendre le phénomène derrière.

Un an plus tard en 1987, il est rapporté que plusieurs facteurs nucléaires sont nécessaires pour l'expression des sous-unités I, II ou III de la cytochrome oxydase de *Saccharomyces cerevisiae* (KLOECKENER-GRUISSEM, MCEWEN et POYTON 1987) mais c'est en 1995 qu'il est démontré que le noyau produit au moins 18 protéines qui interviennent dans le contrôle de l'expression du locus de COX1, la sous-unité I de la cytochrome oxydase codée dans le génome mitochondrial (MANTHEY et MCEWEN 1995). L'une de ces protéines est caractérisée, Pet309p, et son rôle dans la transcription et la stabilisation de COX1 est mis en évidence (MANTHEY et MCEWEN 1995). Les années suivantes, des protéines homologues à celle-ci sont identifiées dans divers organismes, allant de la microalgue à l'humain (COFFIN et al. 1997; FISK, WALKER et BARKAN 1999; LOWN, WATSON et PURTON 2001; MOOTHA et al. 2003).

## 9.1 PPR : PentatrigoPeptide Repeat

Les premières protéines PPR (alors appelées AtPCMP pour "*Arabidopsis thaliana* Plant Combinatorial and Modular Protein") ont finalement été identifiées début 2000 (AUBOURG et al. 2000). Il s'agissait à ce moment là d'une famille de 200 gènes dits orphelins, c'est-à-dire n'appartenant à aucune autre famille de gènes connue et qui avaient la particularité de n'être trouvées que chez les plantes. Ces protéines étaient alors particulièrement étonnantes du fait de leurs répétitions et de leur abondance dans le génome de l'espèce *Arabidopsis thaliana*. Alors même que leur fonction n'était encore pas connue, un mois plus tard, une autre équipe les identifiait également et les nommait PentatrigoPeptide Repeat en référence aux protéines TetratrigoPeptide Repeat (TPR) avec lesquelles elles partagent un repliement 3D et des répétitions proches en séquence, suggérant l'évolution de l'une des familles depuis l'autre (cf. 9.1.2) (SMALL et PEETERS 2000).

Par la suite, quelques analyses fonctionnelles ont été produites (HASHIMOTO et al. 2003; MEIERHOFF et al. 2003; YAMAZAKI, TASAKA et SHIKANAI 2004), permettant de proposer un rôle dans la régulation de l'expression des ARNm mitochondriaux et chloroplastiques via la liaison spécifique de la protéine à un ARNm à l'intérieur de l'organite. Environ 450 protéines PPR sont identifiées en 2004 via des analyses bioinformatiques et de laboratoire chez *Arabidopsis thaliana* et des données à propos de la fonction putative à l'échelle de la famille entière sont produites mais leur cible physiologique reste encore incertaine (LURIN et al. 2004).

Les arguments en faveur de la liaison spécifique des protéines PPR à des ARNm cibles sont de plus en plus nombreux à partir de 2007 et les débuts du déchiffrement du code de reconnaissance spécifique d'un nucléotide pour chaque motif PPR dégénéré répété en 2012 terminent d'asseoir l'idée que les PPR régulent de façon spécifique l'expression des ARNm des organites chez les plantes en s'y fixant (DELANNOY et al. 2007; BARKAN, ROJAS et al. 2012; BARKAN et SMALL 2014; SHEN et al. 2016; ROVIRA et SMITH 2019).

Les protéines PPR peuvent avoir des fonctions variées comme l'édition de l'ARN (remplacer/modifier une base par une autre), le "splicing" (coupure des ARN polycistroniques), la stabilisation (empêcher la dégradation de l'ARN par des exonucléases), ou encore l'épissage (suite de coupures et de ligatures visant à éliminer certaines portions d'un ARN) (BARKAN et SMALL 2014; KOTERA, TASAKA et SHIKANAI 2005; MACEDO-OSORIO, MARTÍNEZ-ANTONIO et BADILLO-CORONA 2021) (cf. figure 20). Ces fonctions sont associées à une classification plus fine en deux catégories (cf. 9.1.1). Chacune de ces catégories est composée d'une succession différente d'une variété de types de motifs (CHENG

et al. 2016 ; GUTMANN et al. 2020) (les différents motifs sont décrits dans la partie 9.1.2).

### 9.1.1 Les deux principales classes de protéines PPR

De nos jours, deux classes de protéines PPR ont été identifiées : les classes P et PLS. Les protéines de la classe P sont composées de motifs du même type : les motifs de type P qui sont décrits en 9.1.2. Il s'agit des motifs les plus couramment trouvés dans les protéines PPR (GUTMANN et al. 2020). Les protéines de la classe PLS, elles, sont composées d'une alternance de plusieurs type de motifs : les motifs P, L et S qui sont également décrits en 9.1.2.

Les protéines de la classe P sont parfois complétées d'un court domaine dont la fonction est encore mal comprise mais leur fonction reste la stabilisation d'un ARN via une liaison spécifique de type "une répétition PPR lie un nucléotide". La liaison peut avoir lieu à l'extrémité 5' comme à l'extrémité 3'. Ce mécanisme de liaison permet notamment de protéger l'ARN de la dégradation par les exoribonucléases.

Chez les protéines de la classe PLS, on retrouve des domaines avec des activités liées à l'editing de l'ARNm à côté du domaine contenant la suite de motifs P, L et S. Ces domaines sont appelés E et DYW. Ils sont souvent tous les deux présents dans les protéines de la classe PLS mais on les retrouve parfois seuls. On peut notamment observer ceci en regardant la composition en motifs des protéines PPR de l'espèce *Arabidopsis thaliana* sur la base de données "ppr.plantenergy.uwa.edu.au" (CHENG et al. 2016).

On note que ces classes ne sont pas distribuées de façon identique au sein des espèces possédant des protéines PPR. En effet, récemment, une recherche des PPR au sein de 1000 transcriptomes de plantes et d'algues a montré de fortes expansions d'une classe de PPR au sein de certaines lignées, comme par exemple chez les Monilophytes (assimilables aux fougères). Par ailleurs, la classe PLS des protéines PPR (GUTMANN et al. 2020) (cf. figure 17) semble absente des algues vertes et rouges, ce qui semble indiquer une apparition de cette classe chez l'ancêtre des plantes terrestres. Il s'agit de résultats se basant sur des travaux plus anciens qui proposaient déjà l'hypothèse de l'apparition tardive de la classe PLS des protéines PPR au sein des plantes terrestres, les exemples de protéines PPR de cette classe au sein d'autres espèces eucaryotes comme des protistes provenant ainsi de transferts horizontaux tandis que les protéines PPR de la classe P sont retrouvées chez toutes les lignées eucaryotes (SCHALLENBERG-RÜDINGER et al.

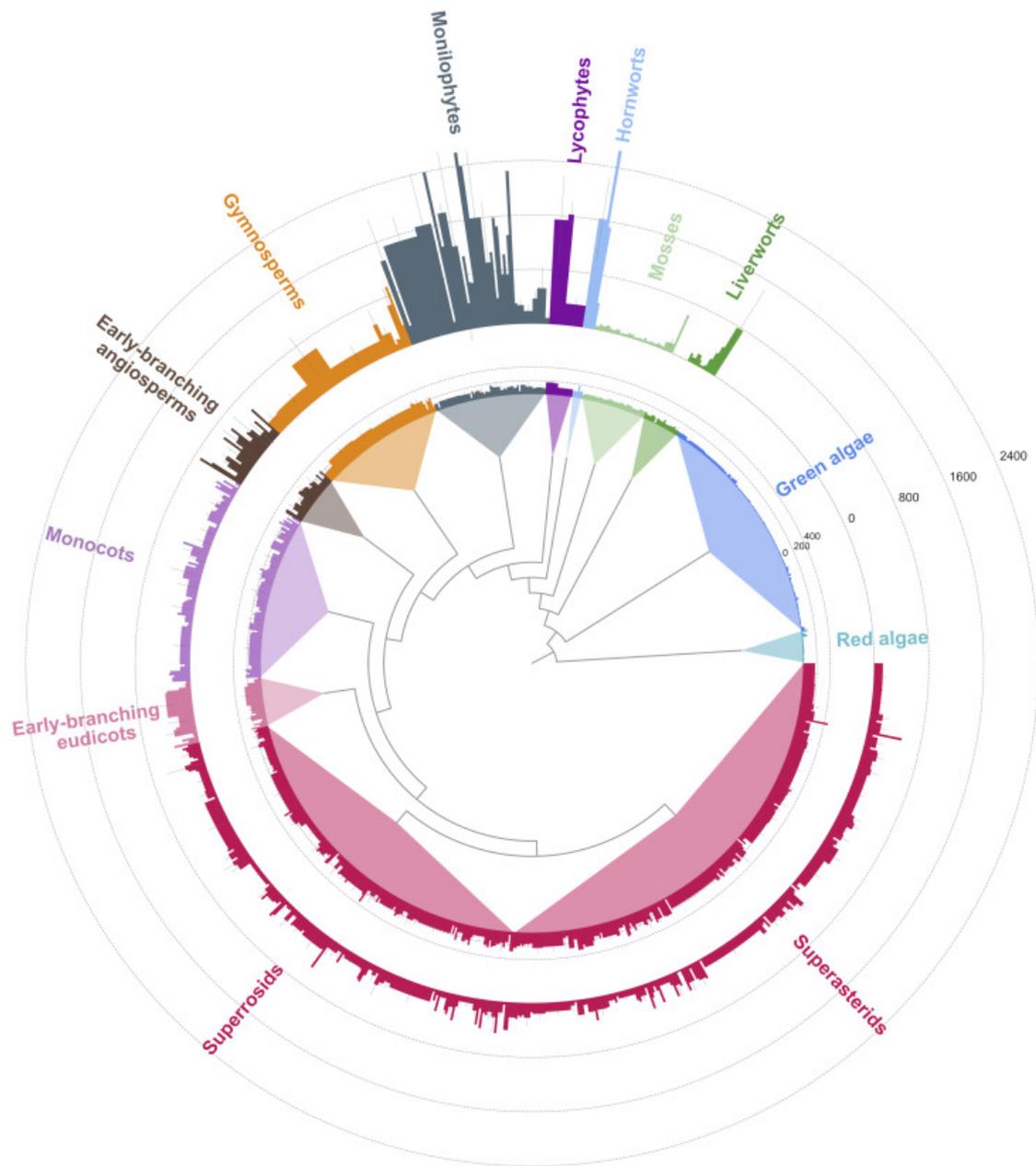


FIGURE 17 – Nombre de protéines PPR de type PLS et P selon l'espèce. Le cercle extérieur correspond au nombre de protéines de type PLS et le cercle intérieur à celui des protéines de type P (GUTMANN et al. 2020).

2013). La capacité d'editing caractéristique aux plantes terrestres semble être une

bonne piste pour expliquer l'expansion de la classe PLS. En effet, le nombre de sites d'editing retrouvés semble souvent proche de celui du nombre de protéines PPR de la classe PLS identifiées dans un génome de plante (KOTERA, TASAKA et SHIKANAI 2005). On note enfin que le protiste *Naegleria gruberi* possède des protéines PPR de type PLS et que cet organisme non photosynthétique présente lui aussi des capacités d'editing de l'ARN dans ses mitochondries. L'origine de ces protéines pourrait être un transfert horizontal (RÜDINGER et al. 2011).

Tout ceci concoure à l'idée que les protéines OTAF évoluent rapidement pour s'adapter aux besoins de l'espèce. Une explication à la variabilité du nombre de PPR a été proposée en 2014 par BARKAN et SMALL 2014 : lors d'une mutation sur un gène codant pour une protéine d'un organite, l'ARN qui en provient présente aussi ce changement et lorsqu'une PPR (dont le gène peut aussi muter et ainsi apporter des changements dans la séquence en acides aminés) est capable de s'y fixer à nouveau, l'ARN est à nouveau ou nouvellement régulé. Il s'agit ici du phénomène de co-évolution et ainsi, la protéine PPR est fixée dans l'espèce au même titre que le nouvel ARN avec lequel elle interagit. Au fil des générations et des millions d'années, de nombreuses PPR apparaissent alors et sont fixées dans l'espèce avec des rôles qui diffèrent parfois (cf. figure 18) (BARKAN et SMALL 2014; MANNA 2015).

### 9.1.2 La variété des motifs composant les protéines PPR

L'origine du motif répété des protéines PPR actuellement admise est l'évolution du motif TPR ancestral vers le motif consensus PPR que l'on connaît (le logo du motif PPR est visible dans la figure 19). Il a effectivement été observé une similarité de séquence faible mais statistiquement significative entre les motifs répétés PPR et TPR. Les protéines TPR (protéines à tetratricopeptide repeat) sont aussi formées de répétitions agencées en un solénoïde alpha mais elles se lient à des protéines et non à des ARNs (cf. 11) (bien qu'il existe des exceptions). On note également que les PPR sont trouvées chez tous les eucaryotes tandis que les protéines TPR sont trouvées chez les organismes eucaryotes mais également chez les organismes procaryotes. Il semble ainsi que les protéines PPR dérivent de la famille des protéines TPR (BARKAN et SMALL 2014).

Plusieurs types de motifs PPR ont été proposés : les motifs P, L et S (cf. figures 18 et 19). Les motifs PPR de type P et L font 35 acides aminés de long tandis que les motifs de type S font environ 31 acides aminés. Le motif retrouvé en majorité au sein de la famille des protéines PPR est le motif de type P qui est notamment le seul motif présent répété dans les protéines de la classe P (cf.

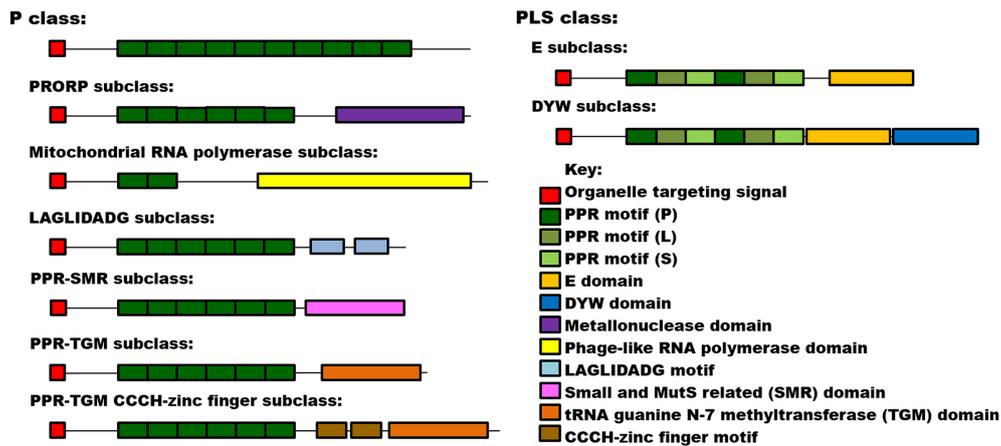


FIGURE 18 – Illustration représentant les différentes classes de PPR et leur composition en motifs types (MANNA 2015).

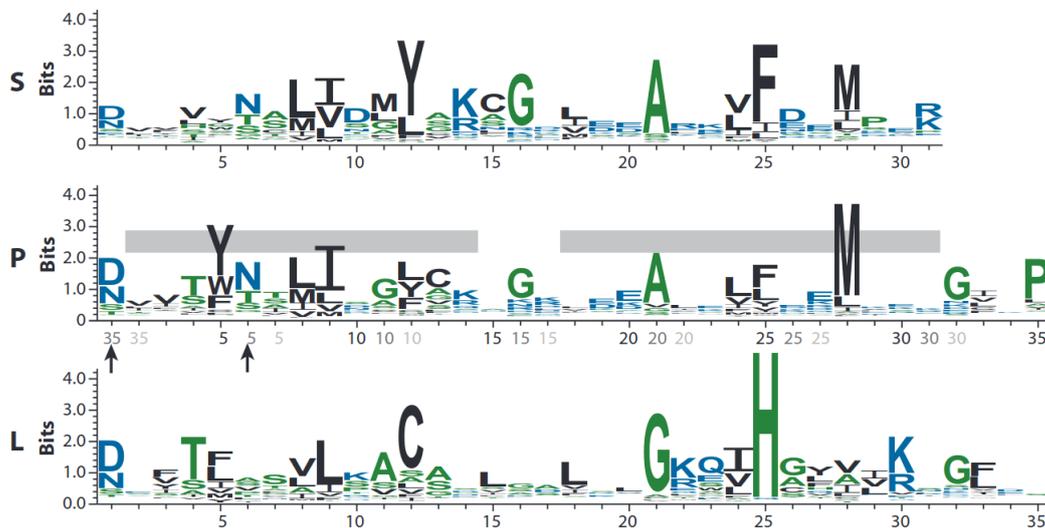


FIGURE 19 – Logo des types S, P et L des motifs PPR. Les rectangles gris correspondent à l'emplacement des hélices alpha constituant la paire d'hélice dans le motif consensus de type P (BARKAN et SMALL 2014).

9.1.1). Les motifs P et S sont très proches dans leur séquence mais le motif S est plus court de 4 acides aminés tandis que les motifs P (et S) et L semblent plus éloignés l'un de l'autre quand bien même ils sont de même taille. En effet, il semble que la séquence de la deuxième hélice des motifs P et L est différente (BARKAN et SMALL 2014; GUTMANN et al. 2020).

## 9.2 OPR : OctatricoPeptide Repeat

Les protéines OPR ont été découvertes notamment chez la microalgue *Chlamydomonas reinhardtii*. De façon générale, ces protéines comportent entre 2 et 10 répétitions du motif OPR dégénéré. Tout comme les PPR, les OPR se lient à des ARNm cibles dans les organites et elles sont aussi classées en plusieurs catégories selon leur fonction (cf. figure 20) : facteur de maturation/stabilisation ou facteur de transcription (BOULOUIS, RAYNAUD et al. 2011 ; EBERHARD et al. 2011 ; BOULOUIS, DRAPIER et al. 2015 ; CAVAIUOLO et al. 2017 ; VIOLA et al. 2019, pour une revue récente voir MACEDO-OSORIO, MARTÍNEZ-ANTONIO et BADILLO-CORONA 2021).

En 1992, des preuves de la régulation de l'activité du chloroplaste de *Chlamydomonas reinhardtii* par des gènes codés dans le génome nucléaire viennent à nouveau confirmer les précédentes propositions de l'existence d'un tel contrôle nucléaire grâce à l'étude de trois mutants dont la production des sous-unités atpA et atpB de l'ATP synthase du chloroplaste était altérée (JENSEN et al. 1986 ; CHOQUET et al. 1988 ; KUCHKA, MAYFIELD et ROCHAIX 1988 ; KUCHKA, GOLDSCHMIDT-CLERMONT et al. 1989 ; ROCHAIX et al. 1989 ; GOLDSCHMIDT-CLERMONT et al. 1990 ; SIEBURTH et al. 1991 ; DRAPIER, GIRARD-BASCOU et WOLLMAN 1992). Auparavant en effet, en 1986, une équipe montre que la biogénèse du photosystème II (complexe protéique impliqué dans la chaîne photosynthétique) est régulée entre autre de façon post-transcriptionnelle (JENSEN et al. 1986) puis en 1988 et 1990, l'action de protéines issues du génome nucléaire sur l'ARNm d'une sous-unité du photosystème I (autre complexe protéique impliqué dans cette même chaîne photosynthétique, cf. figure 14) est montrée mais le mécanisme d'action n'est pas connu (CHOQUET et al. 1988 ; GOLDSCHMIDT-CLERMONT et al. 1990).

Par la suite, bien qu'il ait été majoritairement proposé une action gouvernante des protéines OTAF en 3' des ARNm comme chez les bactéries, la transformation génétique du chloroplaste de *Chlamydomonas reinhardtii* avec des gènes chimères démontre une régulation à l'extrémité 5' de l'ARNm cible. L'idée d'une régulation spécifique de l'expression de certains gènes des organites par des protéines issues du noyau est déjà présente depuis plusieurs années (CHOQUET et al. 1988 ; ZERGES et ROCHAIX 1994) et le terme "trans-acting factor" est utilisé (ROCHAIX 1996). La concentration et la durée de vie de ces protéines seraient la clé de la capacité du noyau les produisant à réguler finement l'activité des organites (WOLLMAN, MINAI et NECHUSHTAI 1999).

Il a ensuite été montré qu'il s'agit de protéines contenant des répétitions

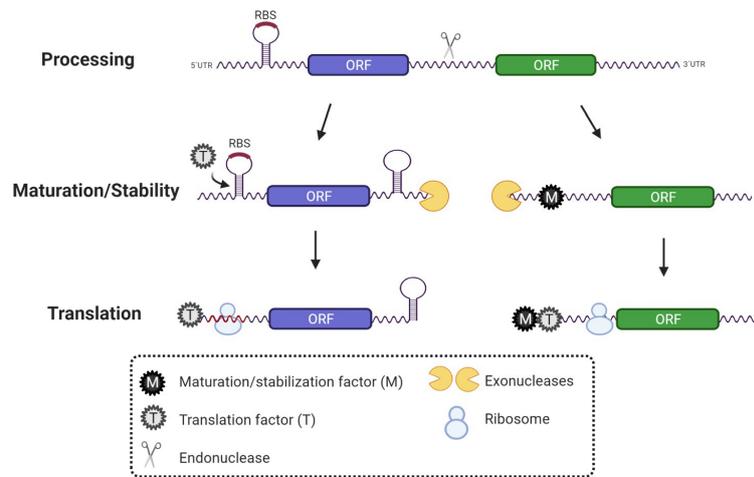


FIGURE 20 – Les protéines OTAF (facteurs M et T) jouent un rôle de protection contre la dégradation des transcrits mono-cistroniques par les exonucléases mais aussi un rôle de correction des erreurs sur les séquences ARNm des organites (MACEDO-OSORIO, MARTÍNEZ-ANTONIO et BADILLO-CORONA 2021).

d'un motif dégénéré différent de celui des protéines PPR et plusieurs dizaines de protéines ont été identifiées au cours des années 2000 et 2010 (LOISELAY et al. 2008 ; RAHIRE et al. 2012 ; BOULOUIS, DRAPIER et al. 2015 ; CLINE, LAUGHBAUM et HAMEL 2017 ; VIOLA et al. 2019). Il semble cependant difficile d'établir un code de reconnaissance entre les protéines OPR et les ARNm auxquelles elles se lient spécifiquement. Ceci peut être dû au faible nombre de protéines OPR jusqu'à présent identifiées comparé au nombre de PPR connues. Toutefois, de récents travaux au sein de mon laboratoire ont apporté des éléments à propos du code de reconnaissance entre OPR et ARNm. Il s'agit du travail de doctorat de Domitille Jarrige sous la direction d'Yves Choquet (JARRIGE 2019). Le travail a consisté à construire un mutant de la cible ARN pour observer la qualité de son interaction avec l'OTAF correspondante (ici les protéines MDB1 et MTH1) puis ensuite à modifier l'OTAF en suivant le code de reconnaissance OPR-ARNm préliminaire pour voir si l'interaction était rétablie. Il s'agit de la même méthode que celle utilisée pour le décryptage du code PPR-ARNm (BARKAN, ROJAS et al. 2012).

Les résultats ont montré que l'interaction de la protéine MDB1 avec son ARNm cible n'est pas facilement altérée. En effet, de petites modifications ne suffisent pas à perdre la reconnaissance entre les deux acteurs mais des modifications plus importantes des ARNm cibles permettent de ne plus observer aucune ou très peu d'interaction. Ceci confirme que les protéines OPR reconnaissent leur cible.

Ces résultats ont cependant été surprenants car une étude plus ancienne *in vitro* avait présenté des résultats sur ces mêmes acteurs amenant à conclure à une interaction facilement altérable entre eux (ANTHONISEN, SALVADOR et KLEIN 2001). Une hypothèse soulevée pour expliquer cette apparente contradiction est qu'*in vivo* d'autres facteurs viennent stabiliser l'interaction. Afin de tester cette hypothèse et de n'observer ainsi que la liaison de l'ARNm à la protéine OTAF, Domitille Jarrige a construit des chimères ARN ne contenant que la partie 5' se fixant à la protéine OTAF. Ses résultats ont montré l'interaction entre les ARN chimères et la protéine OTAF MDB1 est plus sensible dans ce cas que lorsque les modifications sont faites sur l'ARNm endogène. Ceci permet alors de renforcer l'hypothèse du complexe multi-protéique impliqué dans la stabilisation de l'interaction OTAF-ARNm (JARRIGE 2019).

## 10 Structure type des protéines à solénoïde alpha

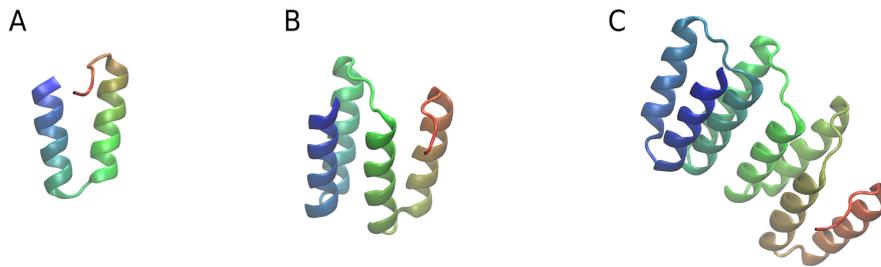


FIGURE 21 – Trois portions de la structure PDB 5i9f disponible sur la base de données RCSB via le logiciel VMD. A : une répétition d'un motif caractéristique d'une protéine à solénoïde alpha avec une première hélice alpha, un coude aussi appelé "linker" et une seconde hélice alpha. B : assemblage de deux répétitions du même motif que dans A. C : assemblage de quatre répétitions du même motif que A. On peut remarquer ici un début de torsion de cette chaîne de motifs répétés.

Les facteurs nucléaires ou TAF sont donc en majorité des protéines adoptant une structure en solénoïde alpha, c'est-à-dire une chaîne de motifs de type "hélice-coude-hélice" (cf. figure 21), dont les familles des Pentatricopeptide Repeat (PPR) (cf. 9.1 et figure 22) et des Octotricopeptide Repeat (OPR) (cf. 9.2), respectivement composées de motifs dégénérés répétés de 35 et 38-40 acides aminés. La structure en solénoïde alpha permet de reconnaître spécifiquement l'ARNm mitochondrial ou chloroplastique cible et de s'y lier plus ou moins solidement selon les types d'OTAF (cf. 11). Cette reconnaissance se fait notamment via

l'association d'un nucléotide de l'ARNm à un motif répété "hélice-coude-hélice" (cf. figure 21).

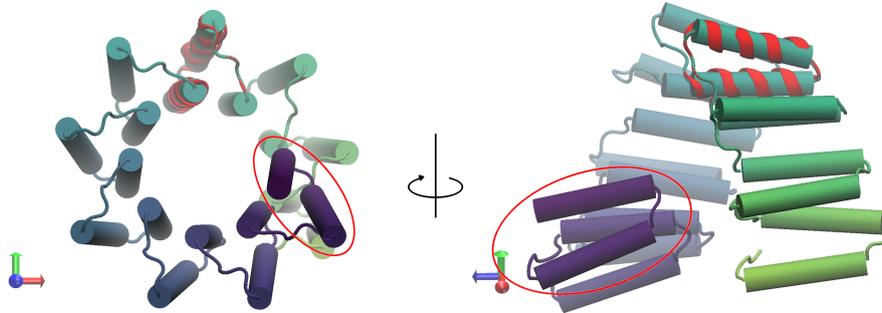


FIGURE 22 – Différentes vues d'une protéine PPR (identifiant sur la base de données RCSB de la structure utilisée : 5i9f) (SHEN et al. 2016) (réalisé avec VMD).



FIGURE 23 – Photo d'un spécimen de la plante terrestre à fleurs (Angiospermae) *Arabidopsis thaliana* (photo par Benjamin Zwittnig).

Chez les espèces modèles de microalgues et de plantes terrestres à fleurs au sein de la lignée des Archaeplastida, on observe un nombre variable de protéines

OTAF. Par exemple, la plante terrestre à fleurs *Arabidopsis thaliana*, espèce modèle du groupe des Angiospermae (cf. figure 23) compte 481 protéines PPR (CHENG et al. 2016; GUTMANN et al. 2020) et une seule OPR (KLEINKNECHT et al. 2014). À l'inverse, les microalgues vertes possèdent peu de protéines PPR (GUTMANN et al. 2020) avec une quinzaine recensées chez *Chlamydomonas reinhardtii*, mais de nombreuses protéines OPR, à raison d'environ 120 dans l'espèce modèle de micro-algue *Chlamydomonas reinhardtii* (CAVAIUOLO et al. 2017; MACEDO-OSORIO, MARTÍNEZ-ANTONIO et BADILLO-CORONA 2021). Actuellement, 58 protéines OPR ont été découvertes et publiées chez l'algue verte unicellulaire *Chlamydomonas reinhardtii* (cf. figure 24), espèce modèle des Chlorophytes (WOSTRIKOFF et al. 2004; LOISELAY et al. 2008; BOULOUIS, RAYNAUD et al. 2011; EBERHARD et al. 2011; BOULOUIS, DRAPIER et al. 2015; VIOLA et al. 2019) et environ 70 autres ont été identifiées mais ne sont pas encore publiées (MACEDO-OSORIO, MARTÍNEZ-ANTONIO et BADILLO-CORONA 2021). Cependant, d'autres espèces de Chlorophytes ne semblent pas posséder autant d'OPR que *Chlamydomonas reinhardtii* ni autant de PPR que *Arabidopsis thaliana*. De plus, il est difficile d'identifier les protéines OPR sur la base de la similarité de séquence à cause de l'évolution particulièrement rapide de leur motif répété caractéristique.

Il semble ainsi y avoir une grande variabilité dans le nombre et le type de protéines à solénoïde alpha au sein des Archaeplastida. On notera que ce mécanisme de la régulation de l'expression des génomes des organites est décrit pour la mitochondrie ainsi que pour les chloroplastes des Chlorophytes, mais aucune exploration rigoureuse n'a été entreprise dans les lignées des Rhodophytes et des Glaucophytes, ainsi que chez les autres organismes issus d'endosymbioses secondaires ou tertiaires, où la façon dont est régulée l'expression des génomes des plastes reste inconnue à ce jour (cf. 6.3).

## 11 La diversité des protéines à solénoïde alpha

Il existe une grande variété de protéines à répétition au sein du vivant (HIRSH et al. 2018) et certaines forment des solénoïdes alpha. Certaines se lient à des ARNm comme les protéines OPR et les protéines PPR qui ont été présentées en amont mais d'autres se lient à de l'ADN double brin (ADNdb) ou encore à d'autres protéines (comme la famille des protéines TPR). Pour quelques unes de ces familles de protéines, on trouve des suites de motifs à deux hélices alpha comme les protéines OPR, PPR, TPR et Ankyrin (cf. figure 21) mais d'autres contiennent des motifs à trois hélices alpha. C'est notamment le cas des protéines



FIGURE 24 – Photo d'un spécimen de la microalgue verte *Chlamydomonas reinhardtii* par Sandrine Bujaldon (données du laboratoire).

à répétitions Armadillo, mTERF et PUF.

Le tableau ci-dessous résume les connaissances actuelles sur la diversité des familles de protéines à solénoïde alpha, quelque soit leur cible de fixation.

On note quelques particularités pour certaines des protéines présentées dans le tableau :

- Les protéines HPR qui ont été identifiées chez *Plasmodium falciparum* (l'un des parasites responsables du paludisme) sont en fait assimilables à des protéines OPR ayant divergées. Les motifs OPR et HPR montrent en effet de fortes similitudes car via des motifs OPR on peut identifier des protéines HPR et *vice versa* (HILLEBRAND et al. 2018).
- De la même façon, les motifs PPR, TPR et HAT (ou "Half-a-TPR") sont supposés proches les uns des autres et il est aussi possible d'identifier des protéines d'une famille via des motifs d'une autre (SIKORSKI et al. 1990; PREKER et KELLER 1998; SMALL et PEETERS 2000).
- D'autres protéines ressemblant aux protéines à solénoïde alpha existent, comme les protéines à répétitions riches en leucine (LRR pour "Leucine-rich repeat") mais elles ne sont pas formées de répétitions de paires d'hélices alpha. Pour les LRR il s'agit de l'association d'une hélice alpha et d'un brin bêta.

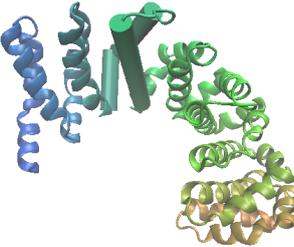
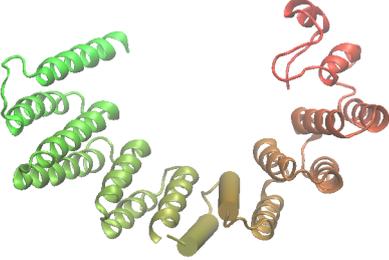
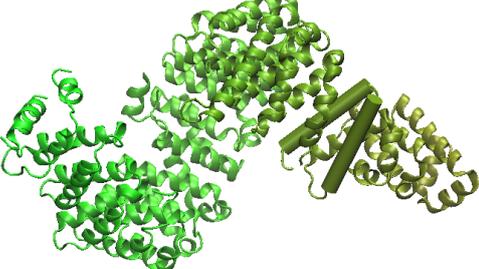
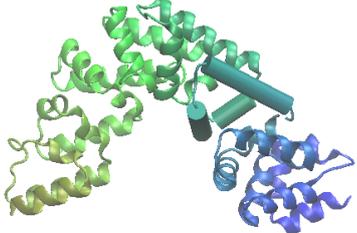
Un exemple de structure de chaque famille de protéines à solénoïde alpha citée

ici est visible dans le tableau 1, de même que des informations sur le ligand et le motif répété les composant.

Les protéines HAT peuvent se lier à des protéines (BAI et al. 2007) ou des ARN (HAMMANI, COOK et BARKAN 2012) mais leur fonction est en grande majorité liée au métabolisme de l'ARN, c'est-à-dire qu'elles peuvent jouer un rôle dans la maturation ou le splicing des ARNm par exemple. On en retrouve couramment une quinzaine chez tous les eucaryotes, dans les organites ou bien dans le cytosol. La liaison des protéines HAT à l'ARN semble proche de la liaison des protéines PPR à leur ARN cible et on peut alors imaginer qu'elles se lient également de la façon suivante : une répétition reconnaît un nucléotide (HAMMANI, BONNARD et al. 2014).

Les protéines mTERF possèdent, quant à elles, une structure 3D du motif répété différente de celle qui est observée chez les protéines PPR, OPR, TPR et HAT. En effet, ce sont trois hélices et non deux qui forment le motif répété. Deux hélices sont des hélices alpha tandis que la dernière est une hélice  $3_{10}$ , c'est-à-dire que cette dernière hélice est composée de 3 résidus par tour d'hélice (et de 10 atomes impliqués dans le cycle formé par le squelette carboné contenant une liaison hydrogène) au lieu des 3,6 acides aminés par tour habituels des hélices alpha et pour lesquelles le cycle fermé par la liaison hydrogène fait 13 atomes (les hélices alpha sont aussi appelées hélices  $3,6_{13}$  (HAMMANI, BONNARD et al. 2014). On retrouve ces protéines chez les plantes et les microalgues ainsi que chez les animaux. Elles jouent un rôle dans la terminaison de la transcription de l'ADN des organites en se liant à l'ADN (PERALTA, WANG et MORAES 2012) mais il a aussi été montré que certaines d'entre elles peuvent se lier à des ARN (CÁMARA et al. 2011).

Les protéines PUF (Pumilio and FBF (fem-3 binding factor)) sont présentes chez la plupart des espèces eucaryotes et les répétitions dont elles sont formées sont bien conservées. Il s'agit aussi d'une famille de protéines qui se lient à des ARN mais elles se n'y lient pas de façon spécifique : la cible de ces protéines est un motif consensus sur l'ARNsb, appelé le Pumilio Response Element (PRE) (BOHN et al. 2017).

Nom	Liaison	Longueur du motif	Exemple de structure cristalline RCSB
Ankyrin	Protéine	33	 <p>MICHAELY et al. 2002</p>
Armadillo	Protéine	40	 <p>Hikaru SHIMIZU et al. 2017</p>
HAT ("Half-A-Tetratricopeptide repeat")	ARNsb Protéine	34	 <p>BAI et al. 2007</p>
HEAT ("Huntingtin, elongation factor 3 (EF3), protein phosphatase 2A (PP2A), yeast kinase TOR1")	Protéine	47	 <p>CANSIZOGLU et CHOOK 2007</p>
mTERF ("Mitochondrial transcription termination factor")	ADNdb ARNsb	30	 <p>BYRNES et al. 2016</p>

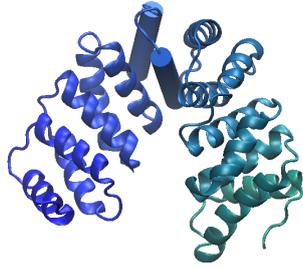
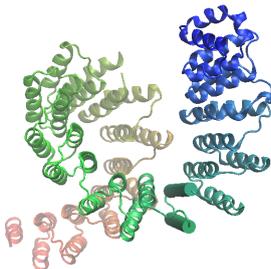
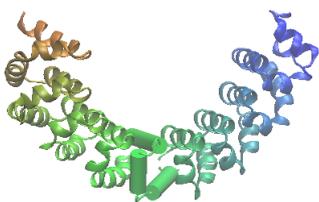
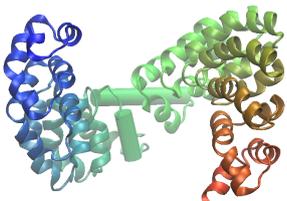
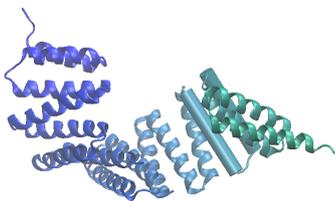
Nom	Liaison	Longueur du motif	Exemple de structure cristalline RCSB
OPR ("Octatricopeptide repeat")	ARNsb	38-40	 <p>MURPHY et al. 2019</p>
PPR ("Pentatricopeptide repeat")	ARNsb	35	 <p>YIN et al. 2013</p>
PUF ("Pumilio and FBF (fem-3 binding factor)")	ARNsb	36	 <p>JENKINS, BAKER-WILDING et EDWARDS 2009</p>
TALE ("transcription activator-like effector")	ADN	34	 <p>GAO et al. 2012</p>
TPR ("Tetratricopeptide repeat")	Protéine	34	 <p>YUZAWA et al. 2011</p>

TABLE 1 – Les principales familles de protéines formant des solénoïdes alpha. Certaines se lient à des protéines, d'autres à des ARNsb et d'autres encore à de l'ADNdb, la longueur des motifs répétés pouvant varier grandement.

## 12 État de l'art des approches précédemment développées pour identifier les familles de protéines à solénoïde alpha

### 12.1 Identifier des protéines à solénoïde alpha via leur répétitions et leur structure

Dans la littérature, on peut trouver plus d'une dizaine de méthodes proposant différentes approches pour détecter les protéines contenant des répétitions. Notons qu'il ne s'agit pas toujours de détecter les protéines à solénoïde alpha puisqu'il existe des protéines possédant des séquences répétées qui adoptent d'autres structures basées sur des brins et des feuillets bêta par exemple. Nous allons présenter ces quelques méthodes dans la 2 ainsi que la suite de cette partie.

Nom	Publication	Principe	Disponibilité	Commentaires
REP <sub>uter</sub>	KURTZ et SCHLEIERMACHER. 1999	Recherche les répétitions exactes et les palindromes	Oui	Il ne s'agit pas d'un logiciel suffisamment souple dans notre cas car les répétitions OPR et PPR sont très variables
REP	ANDRADE et al. 2000	Recherche itérative de répétitions	Non	Capable d'identifier de nouveaux candidats alors inconnus chez 11 familles de protéines contenant des répétitions parmi lesquelles 5 contiennent des protéines à solénoïde alpha comme celles qui nous intéressent dans notre travail : les familles ANK (protéines à répétitions Ankyrin), ARM (protéines à répétitions Armadillo), HAT (protéines à répétitions Half-A-Tetratricopeptide), HEAT (huntingtine (H), facteur d'élongation eucaryotique 2 (E), protéine phosphatase 2A (A) et kinase Tor1 (T)) et TPR (tetratricopeptide repeat)
RADAR	HEGER et HOLM 2000	Cherche des répétitions de séquence en autorisant les gaps et sans parti pris sur la longueur des répétitions ni sur leur nombre	Oui	Le but de cette méthode est de palier à un problème récurrent : celui de ne pas pouvoir facilement trouver des répétitions de tailles variées au sein d'une même séquence. Son optimisation principale est qu'il ne cherche les répétitions que des résidus qui sont alignés.
REPRO	GEORGE et HERINGA 2000	Cherche des répétitions à partir d'alignements locaux non chevauchant avec un algorithme Smith-Waterman adapté puis une recherche de profils issus des répétitions déjà trouvées	Non	Capable de retrouver des répétitions divergées grâce à l'étape de clustering des répétitions trouvées et de recherche de profils

Nom	Publication	Principe	Disponibilité	Commentaires
TRUST	SZKLARCZYK et HERINGA 2004	Augmenter la sensibilité de la recherche de répétitions classique utilisée dans les logiciels des lignes au dessus	Non	Après avoir aligné comme le font les autres méthodes, il utilise le principe de la transitivité des alignements (le principe de la transitivité au sein d'un alignement est très bien expliqué et illustré dans MALDE et FURMANEK 2013). De plus, il estime la significativité statistique de l'alignement analysé, ce que ne fait pas RADAR
DAVROS	MURRAY, TAYLOR et THORNTON 2004	Utilise des alignements structuraux provenant d'un algorithme basé sur les techniques de traitement du signal et les propriétés statistiques des alignements	Non	Il n'est pas comparé à d'autres méthodes et évaluer sa performance semble ainsi difficile
REPPER	GRUBER, SÖDING et LUPAS 2005	Optimisé pour détecter les répétitions de taille inférieure à 15 acides aminés, ce que RADAR et REPRO ne sont pas capables de faire de par les objectifs à l'origine leur construction	Non	Il s'agit ainsi surtout d'un logiciel qui est plus adapté pour les protéines fibreuses, contenant de courtes répétitions sans ou avec peu de gaps dans l'alignement, ce qui n'est pas notre cas ici
HHrep	SODING, REMMERT et BIEGERT 2006	Utilise des comparaisons de modèle de Markov cachés ou HMM (Hidden Markov Model) ainsi que des informations de transitivité des alignements (MALDE et FURMANEK 2013) et des informations évolutives	Oui	Après avoir trouvé des répétitions, de la même façon que REPRO, un profil des répétitions est construit et cherché dans la séquence pour trouver des répétitions plus lointaines

Nom	Publication	Principe	Disponibilité	Commentaires
HHrepID	BIEGERT et SODING 2008	Améliorer l'alignement des répétitions et identifier plusieurs types de répétitions	Non	Capable d'évaluer la significativité statistique des alignements pour améliorer la sensibilité
REPETITA	MARSELLA et al. 2009	S'affranchir de la séquence seule et utiliser des propriétés physico-chimiques et structurelles	Non	Ce logiciel est ainsi capable de détecter les protéines contenant un solénoïde (alpha ou d'un autre genre)
WAVELET	VO, NGUYEN et HUANG 2010	Reconnaître les protéines à solénoïde via la transformée en ondelettes discrète (technique notamment utilisée dans le traitement du signal)	Non	Plusieurs avantages à l'utilisation de cette technique sont avancés par les auteurs : la capacité à représenter facilement les propriétés de la structure des protéines, la possibilité d'extraire des caractéristiques jusque là restées inconnues pour mieux distinguer les protéines à solénoïde des autres et enfin obtenir des statistiques qui permettent de mieux représenter les motifs répétés des protéines à solénoïde.

TABLE 2 – Principales méthodes développées depuis 1999 permettant d'identifier des protéines à motifs répétés dont des protéines à solénoïde alpha.

Les performances d'une partie de ces logiciels ont été comparées dans la littérature au fil des années. Ces comparaisons sont résumées ici.

Pour la publication de TRUST (SZKLARCZYK et HERINGA 2004), ses performances ont été comparées à celles de RADAR (HEGER et HOLM 2000) sur un jeu de données contenant 530 protéines avec des répétitions (BALiBASE Benchmark Alignment Database 2.0 (BAHR et al. 2001)). Cette comparaison a montré que TRUST est meilleur que RADAR notamment sur la justesse de l'estimation des résidus impliqués dans les répétitions avec une "accuracy" (précision) de 82% contre 67% pour RADAR. Il semble que RADAR soit particulièrement sensible : sur 100 séquences de 1000 acides aminés générées de façon aléatoire, RADAR détecte en moyenne 5 répétitions par séquence tandis que TRUST n'en détecte aucune qui soit réellement significative (SZKLARCZYK et HERINGA 2004).

Les performances de HHrep (SODING, REMMERT et BIEGERT 2006) ont été comparées à celles de TRUST sur un jeu de protéines contenant des protéines à répétitions issues de la base de données SCOP et elles montrent que ce logiciel est 2 à 3 fois plus sensible que TRUST. De la même façon, les performances de HHrepID (BIEGERT et SODING 2008) comparées à celles de RADAR et à TRUST montrent qu'il est plus sensible.

Les performances du logiciel REPETITA (MARSELLA et al. 2009) sont comparables à celles de RADAR. Cependant, TRUST semble meilleur car les résultats de TRUST contiennent moins de faux positifs lorsqu'on demande au logiciel de trouver plus de protéines à solénoïde candidates dans un jeu de protéines (c'est-à-dire en relâchant le critère estimant la justesse de la prédiction).

Enfin, la méthode WAVELET semble largement supérieure à celles précédemment présentées et contre lesquelles elle a été comparée (cf. Figure 25). La technique des ondelettes avait déjà été utilisée par MURRAY, GORSE et THORNTON 2002 dans une méthode cherchant également à distinguer les répétitions au sein d'une séquence et ce, avec succès, mais la méthode avait des difficultés à reconnaître des répétitions lorsque celles-ci contenaient des séquences avec insertions ou délétions.

Les principales différences entre tous ces logiciels cherchant des répétitions sont la façon de déterminer la longueur et les bornes des répétitions, comment les statistiques des comparaisons de séquences et des alignements sont utilisées et enfin si les logiciels utilisent les informations de transitivité des alignements ou non (MALDE et FURMANEK 2013).

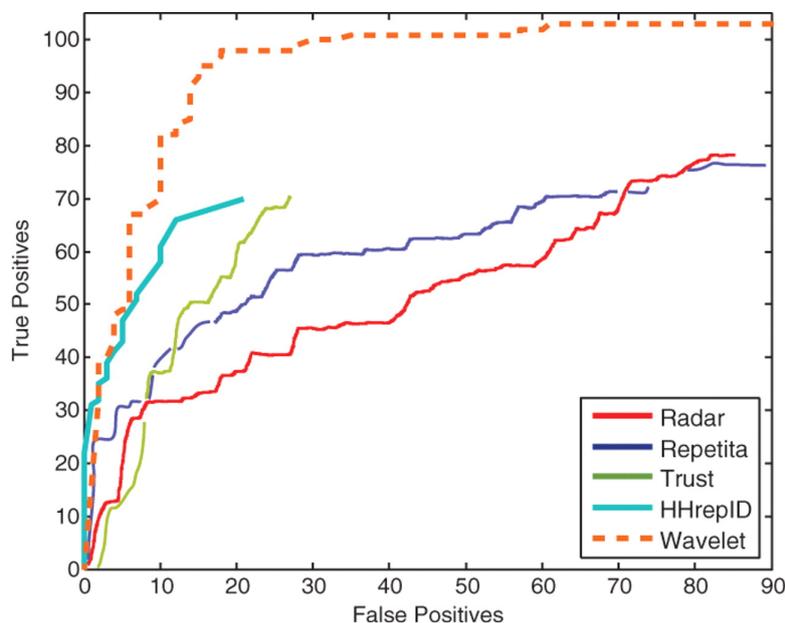


FIGURE 25 – Nombre de témoins vrais positifs contre nombre de témoins faux positifs pour 5 méthodes parmi celles présentées sur un même jeu de données (VO, NGUYEN et HUANG 2010). Le jeu de données provient de MARSELLA et al. 2009 : les témoins positifs correspondent à 105 domaines se structurant en solénoïde et les témoins négatifs correspondent à 247 domaines protéiques ne formant pas de solénoïde (d'après leur structure obtenue aux rayons X).

ard2 est un logiciel un peu différent des autres jusqu'à présent cités (FOURNIER et al. 2013). En effet, il ne cherche pas à proprement parler les répétitions mais tente de retrouver le coude entre les deux hélices de la paire d'hélices antiparallèles caractéristiques des motifs répétés formant les protéines à solénoïde alpha. Il s'agit d'un logiciel utilisant un réseau de neurones entraînés sur des jeux de protéines contenant des coudes et des paires d'hélices. C'est donc la portion d'acide aminés faisant le coude qui est recherchée ici et non la répétition du motif et on s'affranchit ainsi des limites inhérentes aux méthodes citées au dessus qui pourraient être causées par une perte de la significativité des alignements lorsque deux répétitions de séquences divergent trop. Les performances de ce logiciel semblent cependant faibles avec des valeurs de "recall" (rappel) de moins de 0,5 (FOURNIER et al. 2013).

Il faut désormais conclure l'histoire des logiciels permettant de détecter des répétitions en notant que seuls RADAR et ard2 sont actuellement disponibles au

téléchargement et utilisables et nous verrons plus loin dans ce manuscrit que nous les utilisons en les combinant pour identifier des OTAFs.

## 12.2 Identifier des protéines PPR et TPR

Il existe plusieurs méthodes développées depuis une quinzaine d'années qui permettent d'identifier les protéines PPR et TPR.

En 2007, la méthode TPRpred se basant sur une recherche par similarité de séquence est proposée (ZIMMERMANN et al. 2018). Le logiciel permet d'identifier les protéines TPR, PPR et SEL1-like lointaines (protéines proches des TPR qui ont un rôle dans la régulation de la traduction et de l'assemblage de certains complexes protéiques (GRANT et GREENWALD 1996)) via des profils de séquences. Cette approche donnait de bon résultats en comparaison avec des recherches Pfam ou SMART et une nouvelle implémentation a été produite en 2018 pour remplacer la première en améliorant l'efficacité des calculs. Celle-ci est disponible dans le "MPI Bioinformatics Toolkit" (ZIMMERMANN et al. 2018) tandis que la première n'est plus accessible.

D'autres méthodes ont été développées depuis TPRpred :

- SCIPHER est une méthode itérative proposée en 2011 (LIPINSKI et al. 2011) qui utilise des profils HMM construits à partir de motifs PPR identifiés dans des protéines PPR déjà connues au sein des espèces d'intérêt via TPRpred. Cette méthode a été appliquée à 14 espèces de levures et a donné des résultats probant en identifiant 12 protéines candidates qui ne sont pas détectées par TPRpred. Cette méthode a également permis de montrer l'évolution rapide des protéines PPR au sein des espèces de levures étudiées.
- MixedPPR est une méthode utilisant des propriétés variées pour classer les protéines via le principe de l'apprentissage machine avec trois type "classifier" (trois types d'algorithmes conduisant au choix de classer une protéine comme candidate PPR ou non) qui sont "random forest" (forêt aléatoire d'arbres de décision), "J48" (arbre de décision) et "naïve Bayes" (modèle basé sur le théorème de Bayes avec naïveté (indépendance) des hypothèses). Les propriétés choisies par les auteurs (QU et al. 2019) sont des propriétés physiques, chimiques et de séquences comme les auto-cross covariance (qui mesurent les relations entre les propriétés physico-chimiques d'une paire d'acides aminés de la séquence protéique) ou encore les fréquences des 20 types d'acides aminés. Cette méthode est disponible en ligne via le serveur suivant : <http://server.malab.cn/MixedPPR/index.jsp> mais elle n'a

- cependant pas suffisamment été appliquée au sein de protéomes bruts pour pouvoir estimer la qualité de ses prédictions en conditions réelles.
- PPRfinder : construite en 2020, il s'agit d'une méthode qui permet d'identifier les motifs PPR au sein de protéines candidates via une recherche de la meilleure suite de motifs PPR après identification de motifs PPR via des recherche de profils HMM de motifs PPR. C'est notamment la méthode utilisée pour identifier et quantifier la famille de protéines PPR au sein de 1000 transcriptomes de plantes (GUTMANN et al. 2020). Le code est disponible sur Github au lien suivant : <https://github.com/ian-small/OneKP>.
  - deux méthodes récemment publiées (FENG, ZOU et WANG 2021 ; ZHAO, WANG et al. 2021) utilisent l'apprentissage machine ainsi que le même jeu de données que celui déjà utilisé par la méthode MixedPPR précédemment décrite (487 protéines PPR et 9590 protéines non PPR). Il n'est ici aussi pas fait état d'une quelconque application à des données au niveau de l'espèce ou du groupe d'espèces. Ces méthodes utilisent toutes les deux des algorithmes d'apprentissage automatique :
    - chez ZHAO, WANG et al. 2021, c'est un algorithme précédemment implémenté par les mêmes auteurs qui est utilisé (ZHAO, JIAO et al. 2020). Il s'agit d'un algorithme combinant plusieurs approches d'apprentissage automatique pour tenter d'optimiser la justesse des résultats issus des "classifier" classiques et qui montre des résultats probants. Il semble y avoir encore plusieurs améliorations proposées par les auteurs afin d'optimiser leurs résultats comme utiliser des méthodes de clustering sur les résultats pour s'émanciper des protéines ne groupant pas avec d'autres. Les auteurs remarquent finalement que le contenu en méthionine des séquences protéiques semble particulièrement important pour identifier des protéines PPR.
    - chez FENG, ZOU et WANG 2021, c'est une variante de la méthode MixedPPR qui est proposée. Les auteurs utilisent les mêmes propriétés que chez QU et al. 2019 mais font ensuite des corrélations et de la réduction de dimensions (via PCA notamment) pour réduire le nombre de propriétés utilisées. Par la suite, les trois mêmes algorithmes que ceux de la méthode MixedPPR auxquels s'ajoute l'algorithme SVM (Support Vector Machine) sont utilisés. La réduction de dimensions permet de passer de 188 dimensions à 10 et il apparaît en conclusion que les performances des 4 "classifiers" sont très similaires lorsqu'on compare les résultats utilisant 188 ou 10 dimensions (et même meilleures avec 10 dimensions au lieu de 188 pour l'algorithme "naïve Bayes").

Pour finir, la méthode aPPRove développée en 2016 ne détecte pas de protéine

PPR à proprement parler mais elle est cependant capable de prédire l'interaction d'une protéine PPR donnée avec des candidats ARNsb en utilisant des travaux plus anciens proposant un code de reconnaissance PPR-ARN préliminaire (HARRISON et al. 2016; BARKAN, ROJAS et al. 2012). Il ne s'agit cependant pas ici de l'un des objectifs de cette thèse.

## 12.3 Identifier des protéines OPR

Actuellement, il n'existe pas de méthode publiée dédiée permettant d'identifier précisément des protéines OPR autre que la recherche de motifs OPR via similarité de séquence, par BLAST ou par construction de profils HMM depuis des séquences déjà connues.

# 13 Objectifs

## 13.1 Évolution et diversité des protéines OTAF à solénoïde alpha : et en dehors des espèces modèles ?

Chez les espèces non photosynthétiques comme les levures ou les animaux, seules quelques protéines PPR ont été identifiées. Il y en a par exemple sept chez l'*Homo sapiens* (l'humain) et une quinzaine chez la levure *Saccharomyces cerevisiae* (LIGHTOWLERS et CHRZANOWSKA-LIGHTOWLERS 2013; LIPINSKI et al. 2011). Chez les Embryophytes (plantes terrestres) ce sont plusieurs centaines de ces protéines qui y sont identifiées : 491 protéines PPR chez le riz *Oriza sativa* et 105 chez la mousse *Physcomitrella patens* (CHEN et al. 2018; SUGITA et al. 2013). Cependant, les détails de la distribution des protéines OTAF chez les eucaryotes ne sont pas connus car comme évoqué précédemment (cf. 9.2), la présence de protéines OPR, PPR et d'autres protéines à solénoïde alpha se liant à l'ARN n'a pas été examinée en détails, notamment chez les lignées d'algues rouges (Rhodophytes) et chez les Glaucophytes au sein des Archaeplastida qui sont susceptibles d'en posséder beaucoup (car proches des Embryophytes et des Chlorophytes) (cf. figure 26). Il en est de même chez les espèces provenant d'événements d'endosymbioses secondaires comme les Diatomées (cf. 5).

Ceci étant dit, si peu ou aucune protéine OPR ou PPR n'est identifiée dans une espèce eucaryote, on peut se demander comment ces espèces contrôlent l'expression des gènes de leurs organites. Il apparaît alors raisonnable de commencer par chercher à identifier de nouveaux facteurs nucléaires adoptant aussi une

structure en solénoïde alpha tels que les familles de protéines citées en 11.

Partant de ces constatations, la première idée pour étoffer le catalogue des protéines OPR et PPR est de chercher des protéines ayant une séquence similaire à celles qu'on connaît via un outil de comparaison de séquence (comme BLAST, ALTSCHUL et al. 1990) dans les protéomes des espèces d'intérêt. Cependant cette méthode, bien que permettant de trouver un certain nombre de protéines de la même famille, ne permet pas d'identifier toutes celles qui sont déjà connues. C'est notamment le cas quand on regarde dans la famille des protéines OPR. On se rend ainsi compte que la variabilité au sein du motif et que le nombre et la position des motifs que comportent les protéines influent sur notre capacité à les identifier.

L'un des objectifs de ma thèse est donc de construire une méthode qui permette :

- de retrouver l'exhaustivité des protéines d'intérêt telles que les protéines PPR et OPR dans une espèce.
- d'explorer la diversité des protéines à solénoïde alpha susceptibles de se lier à de l'ARNsb en se demandant s'il existe d'autres familles d'OTAF inconnues ou spécifiques à certains clades et susceptibles de jouer un rôle dans la régulation de l'expression des génomes des organites.

Pour cela, j'ai construit une méthode permettant d'identifier des protéines à solénoïde alpha candidates grâce à la similarité de séquence avec des motifs connus. Cette méthode, dont je démontre l'efficacité en deuxième partie de ce manuscrit (cf. II), a permis d'établir un catalogue d'OPR et de PPR dans l'ensemble des Archaeplastida, y compris de nouvelles candidates OPR et PPR dans les organismes modèles *Arabidopsis thaliana* et *Chlamydomonas reinhardtii*.

J'ai également construit une seconde méthode qui, cette fois, utilise notamment la structure 3D caractéristique des protéines à solénoïde alpha pour identifier de nouvelles familles de protéines à solénoïde alpha. Cette méthode, également discutée en deuxième partie de ce manuscrit (cf. II) a montré sa capacité à identifier toute une variété de familles de protéines à solénoïde alpha.

Durant le premier semestre 2021 j'ai co-encadré Alexis Astatourian pour son stage de master 2. Ensemble nous avons cherché à améliorer la seconde méthode. Ce travail est décrit dans la discussion de la deuxième partie (cf. II).

Par la suite j'ai co-encadré Rebecca Goulancourt durant le premier semestre

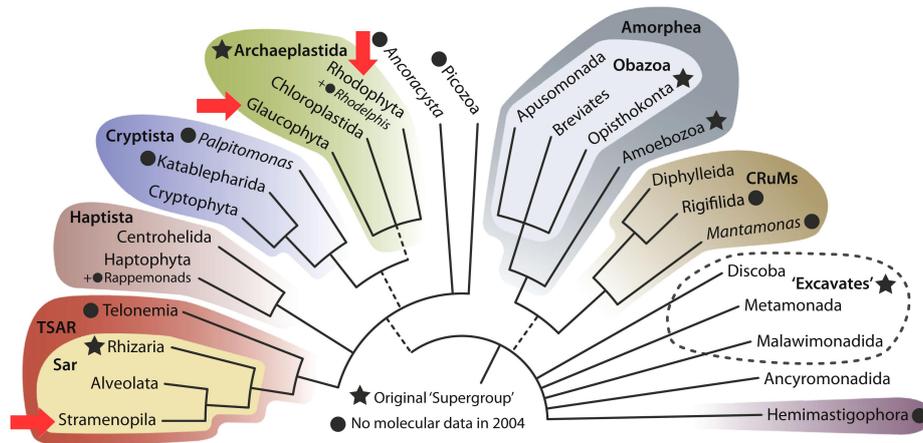


FIGURE 26 – Arbre phylogénétique des eucaryotes d’après les consensus jusqu’à 2020 (BURKI et al. 2020). Chaque couleur représente un supergroupe et les flèches rouges pointent vers les groupes d’intérêt sur lesquels mon travail d’identification des TAF a porté.

2022 pour son stage de master 2. Nous avons construit une méthode de "machine learning" (ou apprentissage automatique) se basant sur une partie du travail effectué durant ma thèse et durant le stage d’Alexis Astatourian pour mieux identifier la diversité des protéines à solénoïde alpha. Cette méthode et les résultats qui en proviennent sont décrits et discutés en deuxième partie de ce manuscrit (cf. II).

### 13.2 Saisir la complexité de l’interaction protéine-ARNm : quelle spécificité pour la liaison à l’ARN ?

S’il est désormais acquis que les protéines OPR et PPR se lient à de l’ARNm dans les organites (cf. 9.2, 9.1), il reste cependant à comprendre comment la liaison s’effectue.

Il a été montré que la liaison entre les protéines PPR et l’ARNm est séquence-spécifique, c’est-à-dire qu’un nucléotide de l’ARNm est lié par un motif dégénéré répété de la protéine et un code préliminaire de reconnaissance entre bases et motifs PPR a été proposé (BARKAN et SMALL 2014). Cependant, ce code ne permet pas de prédire avec une grande certitude quelle est la séquence de nucléotides liée par une protéine PPR étudiée bien qu’il soit exploité dans ce que nous

pouvons qualifier de prémices à l'élaboration de facteurs nucléaires synthétiques (MCDOWELL, SMALL et BOND 2022). Il semble donc qu'il reste à découvrir des facteurs impliqués dans la spécificité de la liaison.

Pour ce qui est de la famille des protéines OPR, un code préliminaire de reconnaissance entre la protéine et l'ARNm a été produit dans mon laboratoire (données non publiées) mais il ne suffit pas encore pour prédire l'ARNm lié par une protéine OPR donnée.

Ainsi, le second aspect de ma thèse porte sur l'étude de la structure et de la liaison à l'ARN des protéines à solénoïde alpha et c'est grâce à des études théoriques *in silico* que nous avons observé le détail de la structure en solénoïde alpha et d'une liaison protéine-ARN à l'échelle atomique. Nous avons produit des simulations de dynamique moléculaire d'une protéine PPR synthétique liée à un ARN simple brin (ARNsb). L'objectif est ici d'obtenir des données à propos des mouvements de la protéine avec et sans ARN lié pour :

- valider l'implication des acides aminés sur lesquels le code PPR préliminaire repose.
- identifier d'autres éléments impliqués dans la reconnaissance protéine ARN, par exemple au niveau des motifs eux-mêmes ou bien entre paires de motifs.

Notons que nous discutons seulement de la liaison protéine PPR-ARNsb dans ce manuscrit. Nous utilisons une seule structure car il y a peu de structures de protéine PPR avec ARN lié et aucune structure de protéine OPR avec ARN n'est disponible. Il semble en effet difficile d'obtenir des structures cristallines de protéines OPR et PPR avec ARNm : moins de 10 structures de protéines PPR avec ARN existent dans la banque de données RCSB ou PDB (ou "Protein Data Bank") dont quatre structures de PPR de bonne résolution, qui sont liées à un brin d'ARN et dans lesquelles il manque peu d'atomes (SHEN et al. 2016) (quatre protéines PPR synthétiques presque identiques liées à un brin d'ARNsb issues des mêmes auteurs et dont la meilleure résolution est 2,19 Ångström) et une structure de résolution plus faible (2,46 Ångström) (YIN et al. 2013). Or, sans données structurales fiables sur les OPR nous ne pouvons pas construire de simulation de dynamique moléculaire desquelles tirer une conclusion sur la liaison protéine OPR-ARN.

Nous avons choisi la structure la mieux résolue des quatre évoquées. Il s'agit d'une protéine synthétique contenant dix motifs PPR identiques se liant à un brin d'ARN poly-U (cf. figure 27). Le motif PPR répété est issu du motif consensus

des motifs P (cf. 9.1, 18) des PPR d'*Arabidopsis thaliana* (SHEN et al. 2016).

Ce travail sera détaillé en troisième partie de ce manuscrit (cf. III).

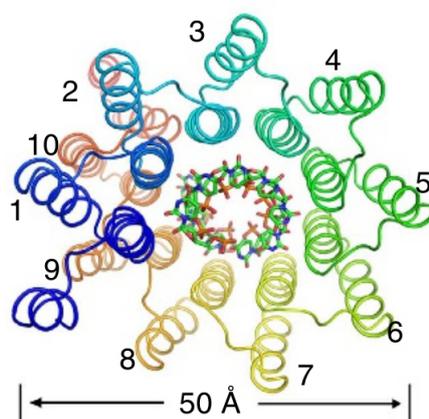


FIGURE 27 – Structure d'une protéine PPR vue par le dessus. A l'intérieur du sillon, l'ARN se lie de façon spécifique, un nucléotide par répétition du motif PPR (SHEN et al. 2016).









Deuxième partie

## Exploration de la diversité des protéines à solénoïde alpha dans les organismes photosynthétiques

"It is the tension between creativity and skepticism that has produced the stunning and unexpected findings of science."

*Broca's Brain : Reflections on the Romance of Science*, Carl Sagan  
(Ballantine Books, 1986)

"Our passion for learning ... is our tool for survival."

*Cosmos*, Carl Sagan (Random House, 1985)



# 1 Introduction

## 1.1 L'évolution des protéines OTAF

Lors de son stage de Master 2 en 2018, Shogofa Mortaza a commencé à explorer et à retracer l'histoire évolutive, en terme de gains et de pertes, des protéines OPR dans 22 espèces de Chlorophytes le long de leur phylogénie. Elle a également retracé l'histoire évolutive de la famille des protéines LHCA (Light-Harvesting Complex A) en parallèle pour comparaison. Ces protéines sont aussi codées dans le génome nucléaire et importées dans le chloroplaste car il s'agit de sous-unités des complexes de la chaîne photosynthétique mais elles ne sont pas soumises aux mêmes pressions de sélection que les protéines OPR. Shogofa a donc supposé dans son travail que la dynamique évolutive de ces deux familles de protéines était différente.

Les résultats de la comparaison de la dynamique évolutive de ces deux familles de protéines ont ainsi montré que le nombre d'OPR varie plus que celui des LHCA dans la phylogénie des 22 espèces étudiées (cf. figure 28). D'importants gains de gènes ont eu lieu dans certaines espèces comme *Monoraphidium neglectum* et *Chlamydomonas eustigma* (barres bleues), mais on observe également de fortes pertes (en noir) dans l'entièreté de l'arbre phylogénétique des 22 espèces. Les événements d'expansions et de contractions semblent plus modérés quant à eux (MORTAZA 2018).

Ainsi, il semble que les protéines OPR évoluent rapidement. Cependant, ces résultats restent préliminaires car une simple recherche de similarité de séquence a été réalisée, alors qu'il est difficile d'identifier toutes les protéines d'une même famille (OPR ou PPR) de cette façon. En effet, le nombre, l'ordre et la position des motifs caractéristiques de ces protéines varient et entraînent donc des alignements de séquence moins bons que pour d'autres familles de protéines qui ne sont pas composées de répétitions modulables. De plus, parmi les espèces étudiées, la qualité de l'assemblage des génomes est variable et pour certaines espèces, seules des données de transcriptomique étaient disponibles. Pour pallier à ces problèmes, durant ma thèse j'ai construit et expérimenté plusieurs méthodes pour détecter des OTAF (connus ou non). Ces méthodes se complètent les unes avec les autres, et je n'ai sélectionnant que des espèces dont les génomes étaient entiers et de bonne qualité.

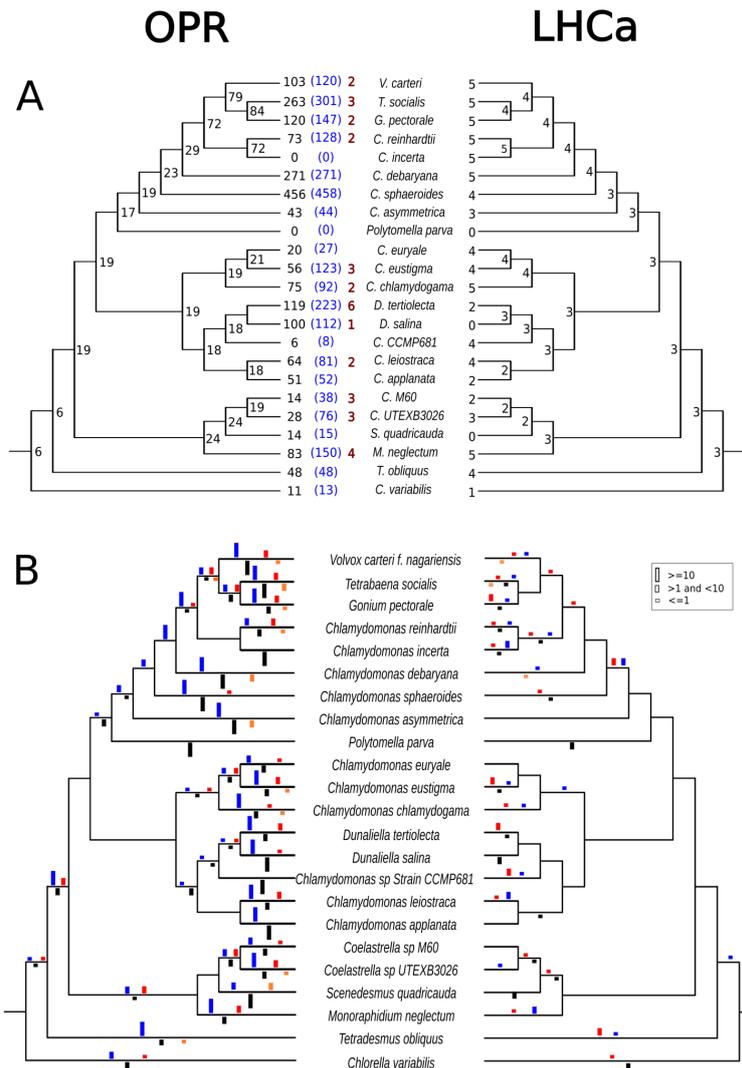


FIGURE 28 – Arbres phylogénétiques des 22 espèces de Chlorophytes : à gauche la famille des protéines OPR, à droite la famille des protéines LHCa. A : En noir est indiqué le nombre de groupes d’orthologues, en bleu le nombre de protéines OPR connues, en rouge le nombre de clusters de protéines OPR connus. B : Les évènements de pertes (noir), de contractions (orange), de gains (bleu) et d’expansions (rouge) constatés au sein de la phylogénie. La longueur des barres de couleur représente le nombre de groupes d’orthologues ayant subi l’évènement en question (adapté du rapport de stage de master de Shogofa Mortaza (MORTAZA 2018)).

## 1.2 Méthode : identification des protéines à solénoïde alpha

Nous avons construit trois méthodes utilisant différentes propriétés et différents concepts, pour dresser les catalogues complets des familles de protéines OPR et PPR dans les espèces photosynthétiques eucaryotes, et identifier de nouvelles familles de protéines à solénoïde alpha potentiellement OTAF. Pour construire ces méthodes, nous avons considéré en premier lieu les espèces *Chlamydomonas reinhardtii* et *Arabidopsis thaliana*, pour lesquelles les OPR et les PPR sont les mieux décrites, ainsi que 46 autres espèces d'Archaeplastida.

La première méthode (IPB, pour Iterative Profile-Based), se base sur le principe de la similarité de séquence pour retrouver les homologues lointains, surtout chez les protéines OPR car peu de travaux sur leur catalogue étendu ont été réalisés. Les protéines PPR, elles, ont bénéficié très tôt d'une large recherche d'homologues et un catalogue complet chez les espèces modèles de plantes terrestres comme *Arabidopsis thaliana* a été fourni grâce aux travaux de Ian Small et Alice Barkan durant les deux dernières décennies CHENG et al. 2016 ; GUTMANN et al. 2020. On peut alors se servir de ce catalogue de protéines PPR pour estimer l'efficacité et la robustesse de notre méthode.

Pour construire notre méthode, nous n'avons pas utilisé les séquences entières des protéines mais les séquences des motifs OPR (38 acides aminés) et PPR (35 acides aminés) les composant. En comparant ces séquences de motifs entre elles puis en les regroupant en clusters, il est possible de construire des profils statistiques (ici des profils HMM) permettant ensuite d'identifier des motifs de plus en plus distants de ceux utilisés initialement. Cette volonté de chercher des motifs OPR et PPR lointains au lieu de protéines entières provient des constats déjà évoqués en introduction que sont la dégénérescence du motif et la composition variable en nombre et en type de motifs au sein des protéines OTAF, qui sont un frein à la recherche d'homologues par similarité de séquence via les séquences complètes de nos protéines d'intérêt.

Cette méthode, dont les résultats sont détaillés dans l'article présenté dans cette partie permet d'identifier efficacement la majorité des protéines OPR et PPR connues ainsi que de nouvelles protéines candidates.

Nous avons également voulu construire une méthode d'identification des protéines à solénoïde alpha s'appuyant sur d'autres propriétés des protéines à solénoïde alpha que seule la similarité de séquence : DT (pour Decision Tree). Nous espérons ainsi distinguer de nouvelles familles de protéines adoptant une

structure 3D similaire à celle des protéines OPR et PPR (cf. 10) dans l'éventualité de mettre en lumière de nouvelles familles de protéines interagissant avec l'ARNm dans les organites.

Cette méthode utilise une succession de filtres sur des propriétés physico-chimiques et des propriétés de structure qui ont été calibrés sur les protéines OPR et PPR. Ceci permet par exemple de rejeter les protéines transmembranaires (protéines dont au moins une partie traverse une membrane de la cellule) ou les protéines ne possédant pas de paire d'hélices alpha (propriété structurale essentielle des protéines à solénoïde alpha). Cette méthode permet notamment d'identifier d'autres familles de protéines à solénoïde alpha. On note cependant que la méthode, bien qu'efficace, est très spécifique et ne permet pas d'identifier la totalité des protéines à solénoïde alpha d'une famille au sein d'une espèce.

Avec Rebecca Goulancourt, lors de son stage de Master 2 de janvier à juin 2022 dont j'ai assuré l'encadrement quotidien, nous avons désigné une méthode d'apprentissage automatique que nous appelons RF (pour Random Forest) utilisant l'algorithme d'apprentissage automatique de la forêt aléatoire. Les approches d'apprentissage automatique n'ont pas besoin de définir des seuils de filtrage, ce qui peut représenter un avantage. Nous avons sélectionné des critères identiques à ceux de la méthode DT, ainsi que des nouveaux pour mieux décrire nos protéines d'intérêt et pouvoir classer toutes les protéines d'une espèce en deux catégories : solénoïde alpha et autre. Cette volonté de construire une méthode parallèle à la précédente provient de plusieurs tentatives pour palier à la faible sensibilité que nous avons constatée lors des tests et de l'application de la méthode DT évoquée précédemment. Les résultats d'Alexis Astatourian (autre stagiaire de Master 2 que j'ai co-encadré en 2021) à propos de ces essais font partie intégrante de ce travail.

Les performances et les résultats de ces trois méthodes sont présentés dans l'article 1 et dans la discussion (cf. 2.3, (cf. 3)). Dans cet article, nous présentons notamment de nouvelles protéines à solénoïde alpha candidates à la régulation post-transcriptionnelle des ARNm mitochondriaux et chloroplastiques. Nous faisons aussi le catalogue des protéines OPR et PPR dans 48 espèces d'Archaeplastida.

Le travail présenté dans l'article qui suit est actuellement en cours de finalisation et l'article sera soumis prochainement.

## 2 Article

### 2.1 A propos de l'approche développée

Dans cet article, nous présentons et discutons les trois méthodes développées : IPB, DT et RF. La réflexion et le cheminement vers leur construction et les versions proposées dans l'article sont les fruits d'essais, de tests et de recherche avec le concours de plusieurs étudiants durant leurs stages de licence et de master. Nous présentons ainsi ici une partie de ces réflexions.

L'implémentation d'IPB, se basant sur les motifs OPR et PPR a été pensée pour explorer le contenu en protéines OPR des espèces possédant un chloroplaste et capables de faire la photosynthèse et c'est pour cela que nous avons appliqué la méthode sur 48 espèces d'Archaeplastida. L'application aux protéines PPR dans ces mêmes espèces avait principalement pour but de mieux évaluer les capacités de la méthode ; les protéines PPR étant en effet mieux connues et leur diversité mieux explorée.

Les résultats de la recherche avec IPB ont montré que seules quelques espèces possèdent suffisamment de protéines OPR ou PPR pour réguler finement l'expression de leurs génomes chloroplastiques. Certaines espèces d'Archaeplastida comme dans le groupe des Rhodophytes possèdent quelques PPR et aucune OPR à l'exception de *Porphyridium purpureum*. On se demande donc comment l'expression des gènes dans leurs génomes chloroplastiques sont régulés et notre première hypothèse a été de proposer l'existence d'autres familles de protéines à solénoïde alpha dans ces espèces. Or, l'inconvénient principal de la méthode est qu'il est nécessaire de connaître les motifs composants la protéine à solénoïde alpha dont on souhaite identifier des homologues dans la même espèce ou dans d'autres espèces.

Ainsi, pour explorer la diversité des protéines à solénoïdes alpha nous avons construit une seconde méthode, DT, utilisant d'autres propriétés que la similarité de séquence. Ces propriétés sont par exemple la présence et le nombre de coudes (ou "linkers") de paires d'hélices alpha, l'absence d'hélice transmembranaire ou encore le type de structure secondaire prédite. Cette méthode reprend le principe de l'arbre de décision dans lequel chaque suite de choix (ou décisions) mène à une issue (une feuille de l'arbre de décision).

Ici, chaque protéine suit un arbre de décision où toutes les feuilles sauf une amènent au classement de la protéine comme "protéine ne possédant pas de

solénoïde alpha" tandis que la dernière feuille amène au classement de la protéine comme "protéine à solénoïde alpha candidate". Autrement dit, à chaque étape, les protéines ne possédant pas la propriété considérée dans les témoins positifs (voir ci-dessous) sont éliminées. On utilise des seuils fixes sur chacune des propriétés pour décider si la protéine s'approche des propriétés caractéristiques des protéines à solénoïde alpha. Par ailleurs, le paramétrage de ces seuils a été réalisé avec un jeu de données contenant des protéines "témoins positifs" à solénoïde alpha comme des protéines OPR de *Chlamydomonas reinhardtii*, des protéines PPR de *Arabidopsis thaliana* ainsi que d'autres types de protéines à solénoïde alpha comme des protéines à répétitions ankyrin. Un second jeu de données "témoins négatifs".

Les paramètres que nous avons déterminés grâce à nos jeux de données sont très stricts et impliquent ainsi une faible sensibilité de notre méthode lorsqu'elle est évaluée sur les jeux de témoins (de l'ordre de 0,2). Cependant, on note que la spécificité évaluée de DT est alors de 0,99 et on peut ainsi avoir une bonne confiance dans les résultats de cette méthode. Pour tenter d'augmenter la sensibilité de notre approche, nous avons mis en place une troisième méthode, RF, qui utilise des propriétés équivalentes à celles de la méthode de l'arbre de décision en plus de nouvelles propriétés comme les fréquences des acides aminés. Ce travail a en partie été effectué par Rebecca Goulancourt durant son stage de Master 2.

RF utilise l'apprentissage automatique pour s'affranchir des seuils et des décisions binaires prises dans l'arbre de décision de la méthode DT. L'entraînement du modèle a été fait sur les mêmes jeux de données témoins positifs et négatifs que DT et on peut alors comparer les performances de ces deux méthodes. La sensibilité de la méthode avec apprentissage automatique est de l'ordre de 0.94 et la spécificité de l'ordre de 0.92. En effet, on obtient plus de protéines à solénoïde alpha candidates avec cette méthode qu'avec la méthode DT.

## 2.2 Qualité des protéomes utilisés

Nous avons estimé la qualité des protéomes d'Archaeplastida sur lesquels nous avons travaillé grâce au logiciel BUSCO. Il s'agit d'une méthode qui compare le protéome à un jeu de marqueurs spécifiques du clade avec lequel on choisit de comparer le protéome. Par exemple pour les 48 protéomes d'Archaeplastida avec lesquels j'ai travaillé, j'ai fait la comparaison avec le jeu de marqueurs des eucaryotes et le jeu de marqueurs des Viridiplantae. En observant les résultats, bien que certains de nos protéomes n'aient pas de bons scores de complétude notamment chez les Rhodophytes, la plupart ont un score supérieur à 90% et nous

avons donc estimé que la qualité des protéomes était satisfaisante pour continuer notre exploration des protéines à solénoïde alpha.

## **2.3 Article 1**

## Searching $\alpha$ -solenoid proteins involved in organellar gene expression

Céline Cattelin<sup>1;2</sup>; Rebecca Goulancourt<sup>1</sup>; Emmanuel Chatelet<sup>1</sup>; Alexis Astatourian<sup>1</sup>; Francis-André Wollman<sup>1</sup>; Charles Robert<sup>2</sup>; Ingrid Lafontaine<sup>1\*</sup>

<sup>1</sup>UMR7141, Institut de Biologie Physico-Chimique (CNRS/Sorbonne Université), 13 Rue Pierre et Marie Curie, 75005 Paris, France

<sup>2</sup>UMR9080, Institut de Biologie Physico-Chimique (CNRS/Université Paris-Cité/PSL), 13 rue Pierre et Marie Curie, 75005 PARIS, France

Correspondance: [ingrid.lafontaine@sorbonne-universite.fr](mailto:ingrid.lafontaine@sorbonne-universite.fr)

### Abstract

In photosynthetic eukaryotes of the green lineage, the expression of the chloroplast genome is mainly regulated post-transcriptionally, by RNA-binding proteins encoded in the nuclear genome (OTAF for organelle trans-acting factor). Most of those identified to date belong to two families of alpha-solenoid proteins (PPR and OPR) and interact with specific sequences on their target mRNAs, allowing their maturation, splicing, editing, stabilization and translation activation. Here we present three novel approaches for annotating  $\alpha$ -solenoid proteins targeted to the chloroplast or the mitochondria by i) profile-based searches and approaches that do not rely on sequence similarity: ii) decision tree and iii) machine learning approaches, to identify distant homologs of existing OTAF families and new OTAF families. Applied to a set of 48 proteomes of Archaeplastida, the combined approaches efficiently retrieve those OPR and PPR proteins that were previously annotated (with more than 85% accuracy). In total it identifies 280 OPR and 8880 PPR in the 48 proteomes studied. Finally, our approach allowed us to identify 4505 other  $\alpha$ -solenoid proteins families, 39 in *C. reinhardtii* and 5 *A. thaliana*, which are likely to participate as new regulators of organelle gene expression. We thus provide valuable new tools to decipher the repertoire of OTAF and new candidates for experimental characterization.

## Introduction

Eukaryotic photosynthesis is ensured by plastids, organelles originally acquired ca 1.5 billion years ago from a primary endosymbiosis involving a protist host and a cyanobacterial ancestor, which gave rise to the extant green algae and land plants (together Viridiplantae), red algae and glaucophyte algae (Archibald 2015). These endosymbiotic events were followed by massive gene transfers from the plastid progenitors to the nucleus of the host cell, as it had already happened during mitochondrial endosymbiosis, which involved an Archaeal host and an  $\alpha$ -proteobacterial ancestor, ca. 1.8 billion years ago (Archibald 2015). Most of the organelle proteome - either mitochondrion or plastid- proteome, is now nucleus-encoded, translated in the cytosol and imported in the organelle. However, these energy-providing organelles have retained a tiny genome, therefore several major protein complexes in organelles are genetic mosaics with subunits encoded in two different genomes. To ensure cell viability and acclimation of the organelle activity (energy production and metabolic activities), expression of organelle genomes became, during evolution, closely interconnected with that of the host cell. While the expression of nuclear genes is regulated at multiple levels (via epigenetic marks, transcriptionally and post-transcriptionally), it is known, based on pioneering studies in Viridiplantae and in yeast, that plastidial and mitochondrial genomes are mainly regulated post-transcriptionally, by nucleus-encoded RNA-binding proteins (Choquet and Wollman 2002; Woodson and Chory 2008). These proteins hereafter named OTAF (for organellar trans-acting factor), belong mostly to a large class of proteins containing repeated motifs forming short  $\alpha$ -helices that confer a rod-shape like structure with concave surface where ligands can bind, like HEAT, Armadillo and Pumilio repeats (Andrade and Bork 1995; Riggelman et al. 1989; Spassov and Jurecic 2003). Most OTAFs contain a succession of either PPR motifs of 35 residues (pentatricopeptide repeat) or OPR motif of 38 residues (octatricopeptide repeat) (Small and Peeters 2000; Eberhard et al. 2011; Lurin et al. 2004; Macedo-Osorio et al. 2021; Loiselay et al. 2008; Delannoy et al. 2007; Barkan and Small 2014). PPR and OPR motifs are degenerated and form a pair of antiparallel  $\alpha$ -helices. The variable number of motif repeats form an  $\alpha$ -solenoid shape with a positively charged surface that specifically binds to the mRNA (Barkan and Small 2014; Cheng et al. 2016). PPR repeats are related to TPR repeats (Sikorski et al. 1990; Small and Peeters 2000) that mainly mediate protein-protein interactions

and are involved in a variety of cell processes (D'Andrea and Regan 2003). Sel1-like repeats are themselves linked to TPR repeats, while HAT repeats are “Half-A-TPR” repeats, some of which, like PPR, bind to RNA (Preker and Keller 1998). As expected from their evolutionary role, the PPR and OPR repertoires are remarkably diverse between organisms, with land plants containing several hundreds of PPR (Gutmann et al. 2020) and few OPR (Kleinknecht et al. 2014), whereas the green algae *Chlamydomonas reinhardtii* encodes only a dozen PPRs (Tourasse et al. 2013; Gutmann et al. 2020) but hundreds of OPR (Eberhard et al. 2011; Boulouis et al. 2015). PPR and OPR proteins also contain additional domains, like the DYW domain in PLS-type PPR, involved in editing (Barkan and Small 2014; Hayes and Santibanez 2020) and the RAP domain (RNA Binding Abundant in Apicomplexa), (Lee and Hong 2004) at the C terminus of NCL OPR in *Chlamydomonas reinhardtii* (Boulouis et al. 2015).

Profile-based methods have been developed to identify PPR, like TPRpred that identify TPR motifs and related PPR and Sel1-like motifs (Karpenahalli et al. 2007); PPRFinder based on plant homologs, (Gutmann et al. 2020; Karpenahalli et al. 2007), with an extensive catalogue of PPR in plants provided at [\(Lee and Hong 2004\)](#); and SCIPHER based on yeast homologs (Lipinski et al. 2011). The motifs degeneracy and the highly biased taxonomical distribution of PPR and OPR repeats make the homology searches difficult. Thus, it is likely that those families are still incomplete, especially for OPR.

It is of note that the regulation of plastid gene expression outside Viridiplantae remains largely unexplored, especially in organisms bearing complex plastids resulting from secondary and tertiary endosymbiosis that occurred between a eukaryotic alga and a heterotrophic protist (Archibald 2015). Genes and processes implicated in the biogenesis and function of these complex plastids await further characterization and paradigms established for primary plastids in Viridiplantae must be challenged and most likely should be revisited.

Here, we present three procedures to complete the catalogue of OTAF and discover new OTAF families. The first is a profile-based similarity procedure to retrieve distant OTAFs homologs predicted to be targeted to organelles (*pto*), in order to fully describe their distribution. Applied to OPR and PPR, we show that OPR expansions were restricted within Chlorophytes and that outside of green algae and land plants, PPRs and OPRs were few in number, suggesting that other players in organelle regulation remain to be discovered. This likely reflects their genetic adaptation to different lifestyles or ecological niches. We also present a procedure to retrieve new families of nuclear-encoded candidates likely to be involved in organelle genome expression, *i.e.* *pto* proteins adopting an  $\alpha$ -solenoid shape. We thus identified several dozens of organelle-addressed  $\alpha$ -solenoid candidates, including a family of  $\alpha$ -solenoid containing ankyrin

like repeats in *Chlamydomonas*, whose experimental characterization would be relevant to understand their possible contribution to chloroplast gene expression.

## Results

### Sequence similarity procedure

In order to retrieve distant homologs of known OTAFs, we developed an iterative profile-based procedure, IPB (Figure 1). Profiles built from existing motifs are searched against the proteomes of interest and regions with significant similarity (new motifs) are retrieved. The newly identified motifs and the already existing ones are clustered together to build new profiles, avoiding the formation of very large and non-specific profiles. The clustering is validated based on the quality of the alignment between motifs within a cluster (See Methods). The proteins containing the new motifs are selected and their subcellular localization is estimated with 4 prediction algorithms (see Methods). The procedure is iterated a defined number of times, or until no new *pto* protein is identified.

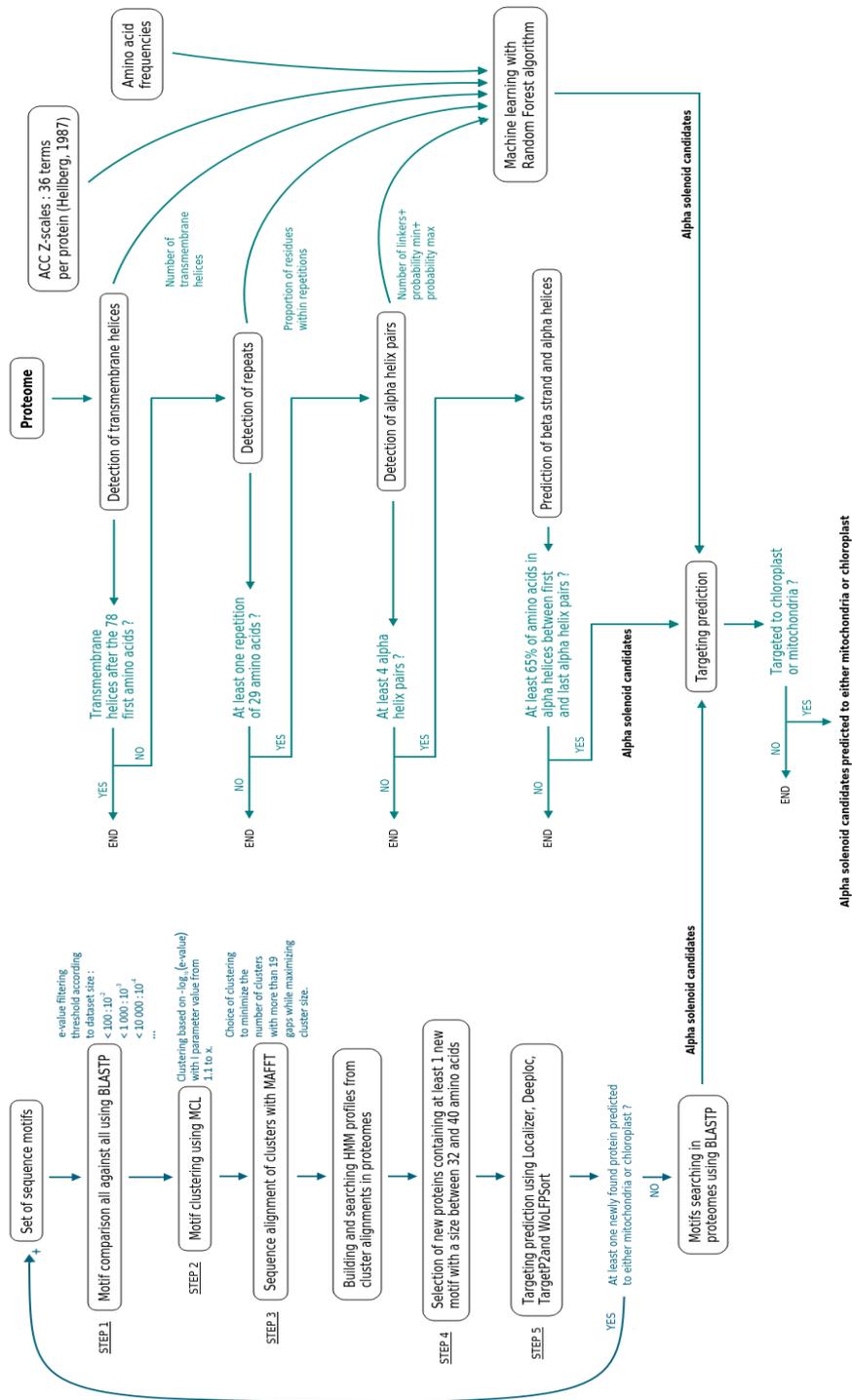
We built profiles from 107 motifs from 12 known OPR proteins and 155 PPR motifs from 11 published PPR proteins (Table 1) to retrieve homologs in a representative set of 48 proteomes from Archaeplastida (Table S1): the proteome of *Cyanophora paradoxa*, the sole representative of Glaucophytes, the 9 proteomes available in Rhodophyta, the 13 (including *C. reinhardtii*) available proteomes in Chloroplastida, respectively, and a set of 25 proteomes of Streptophyta, including 17 Angiosperms, among which *A. thaliana*).

To retrieve proteins containing PPR motifs, two iteration steps were performed without reaching convergence, due to nearly thousands of *pto* candidates in 3 Tracheophyta species: *Selaginella moellendorffii* (Lycopodiophyta) and two Euphyllophyta, *Ceratopteris richardii* and *Thuja plicata*. This is in agreement with previous results, where more than 1500 PPR proteins are found in Lycopodiophyta and Euphyllophyta (Gutmann et al. 2020). As we did not reach convergence, the number of motifs obtained after the final BLASTP search to retrieve additional candidates that could share similarities with motifs singletons that were not used during the iterative procedure

is significantly higher, more than 3 folds higher than after the 2 iterative steps and the number of clusters is more than 4 folds higher (Table 2).

With the OPR profiles, 9 iterative steps were performed to reach convergence, *i.e.* until no new *pto* protein was retrieved. Only 242 new motifs were retrieved after the final BLASTP search (see explanation above for PPR motif). This number strengthens the convergence of the iterative procedure. Note that the lower number of obtained clusters is just due to adjustment of the *l* parameter of MCL at each step according to our defined optimality criterion (maximum number of clusters with the lowest number of gapped positions, see Methods) (Table 2).

Only 846 PPR clusters contain more than 20 sequences. Most of them (771) are composed of Streptophyta motifs that are globally homogeneously distributed among the different Streptophyta studied proteomes. Only 84 OPR clusters contain more than 5 sequences. One cluster is biased towards Streptophyta motifs (186/216 motifs) and two are biased towards Angiospermae motifs (93/98 and 34/43 motifs), but homogeneously distributed among the different Angiospermes. The remaining 81 clusters are biased towards Chlorophyta motifs, 69 of them biased also towards Chlorophyceae (85% bias on average). The biased composition of the clusters of OPR motifs confirms the ability of IPB to retrieve diverged motifs and suggest that OPR motifs are more degenerate than PPR motifs.



**Figure 1:** Diagram of three developed procedures. IPB, left; DT, middle and RF, right.

OTAF	Locus Id	Gene	Type	#motifs	Source
OPR, <i>Chlamydomonas reinhardtii</i>	Cre15.g638950	NCC1	NCL	9	(Boulouis et al. 2015)
	Cre15.g640400	NCC2	NCL	9	(Boulouis et al. 2015)
	Cre07.g348800	TAB1	Translation factor	10	(Rahire et al. 2012)
	Cre08.g358350	TDA1	Translation factor	8	(Eberhard et al. 2011)
	Cre06.g262650	TAA1	Translation factor	7	(Lefebvre-Legendre et al. 2015)
	Cre09.g394150	RAA1	Splicing factor	4	(Eberhard et al. 2011)
	Cre09.g388356	TBC2	Translation factor	10	(Eberhard et al. 2011)
	Cre06.g272450	MBI1	Maturation factor	17	(Wang et al. 2015)
	Cre10.g429400	MCG1	Maturation factor	12	(Wang et al. 2015)
	Cre16.g680850	MDA1	Maturation factor	13	(Viola et al. 2019)
	Cre10.g440000	RAA8	Splicing factor	6	(Marx et al. 2015)
	Cre09.g388372	RAT2	Maturation factor	2	(Marx et al. 2015)
PPR, <i>Arabidopsis thaliana</i>	AT2G18940	-	P-class	18	(Pfalz et al. 2009)
	AT2G35130	-	P-class	12	(Small and Peeters 2000)
	AT2G41720	EMB2654	P-class	19	(Lee et al. 2019)
	AT3G06430	EMB2750	P-class	11	(Williams and Barkan 2003)
	AT3G09650	HCF152	P-class	12	(Meierhoff et al. 2003)
	AT3G53170	-	P-class	10	PPR database
	AT3G59040	-	P-class	10	PPR database
	AT4G31850	PGR3	P-class	27	(Yamazaki et al. 2004)
	AT4G39620	PPR5,EMB2453	P-class	9	(Beick et al. 2008)
	AT5G02860	-	P-class	18	PPR database
	AT5G48730	-	P-class	9	(Williams and Barkan 2003)

**Table 1.** OPR and PPR locus from which motifs were used to build the initial profiles for the IPB procedure, and associated references. Type refers of the OTAF function. P-class PPR contains an organelle targeting peptide followed by PPR motifs of 35 residues. They are involved in stabilization, processing, splicing and translation; PPR database: <https://ppr.plantenergy.uwa.edu.au/>.

		<i>OPR motifs</i>	<i>PPR motifs</i>
after iterations	#clusters	533 (3320)	5691 (50823)
after BLAST	#clusters	480 (3562)	23572 (167102)
	#Big clusters	84 (1877)	846 (3455)

**Table 2.** Number of clusters obtained after 9 and 2 iterations of IPB for OPR and PPR, respectively and after the final BLASTPsearch. The total number of motifs into clusters are given in parenthesis. Big clusters contain more than 5 and 20 motifs for OPR and PPR clusters, respectively.

We first compared our results to the set of annotated PPR in *A. thaliana*, as these are the best described OTAFs in land plants and particularly in *A. thaliana* (Delannoy et al. 2007). We retrieved 451 over the 476 annotated PPR proteins in the *Arabidopsis thaliana* proteome, with at least 1 recognized PPR motifs and only 398 with 2 PPR motifs (Fig 2). 4 additional PPR were found by the DT procedure, and another 8 were found by the RF procedure (see below).

11 of the 15 PPR annotated in *C. reinhardtii* were retrieved by IPB with the PPR motifs, in line with the previous observations that we could find more PPR candidates in the proteomes we have studied by performing more iterations.

IPB with the OPR motifs retrieved the sole OPR protein (ATRAP, AT2G31890) in *A. thaliana* and 122 *pto* candidates in *C. reinhardtii*. Among the *pto* candidates in *C. reinhardtii*, 56 published OPR were retrieved. Only Raa3 and NCL18 were not retrieved (Boulouis et al. 2011; Macedo-Osorio et al. 2021). Since Raa3 was retrieved by the RF procedure, NCL18 remains the sole published OPR that was not retrieved by any of the three developed procedures. We also retrieved the 47 additional candidates annotated as “OctotricoPeptide Repeat Protein” in the annotation v5.6 version available at JGI, Phytozome13.

IPB with OPR motifs also retrieved 11 additional candidates that are not annotated as OPR (Table 3). Four of them have a gene name in the v5.6 annotation available at Phytozome12. The first one is Raa7, the *psaA* mRNA trans-splicing factor (Lefebvre-Legendre et al. 2016). Although other trans-splicing factors Raa1, Raa8 and Rat2 contain OPR motifs (Kück and Schmitt 2021), Raa7 was not described as an OPR. However, its predicted 3D structure, although of very low quality, contains pairs of  $\alpha$ -helices and could likely fold into an  $\alpha$ -solenoid conformation. The Uniprot automatic annotation pipeline at Uniprot, using Google’s ProtNLM (protein natural language model) indicates that Raa7 contains a RAP domain (IPR013584), like other known OPR proteins in *C. reinhardtii* (Boulouis et al. 2015). However, the RAP domain is not found in Raa7 by classical similarity search via InterProScan. The second one is Cri1, annotated as a carotenoid isomerase that adopt an  $\alpha$ -solenoid shape according to its AlphaFold2 prediction (hereafter named AF2pred.), also automatically annotated as containing a RAP domain. Finally, IPB detects Mme1

and Mme6 that could likely correspond to false positive, with  $\alpha$ -helices predicted with low quality in AF2pred. that does not reveal any  $\alpha$ -solenoid shape. The 7 other candidates are uncharacterized, with only automatic domain annotation at Uniprot, domains that are not retrieved by an InterProScan search (Table 3). Among them, 2 adopt an  $\alpha$ -solenoid shape in AF2pred. (Cre09.g395880 and Cre02.g109200) and one (Cre03.g145867) is part of the large subunit of the mitoribosome (PDB entry 7pkt, chain M) that forms an  $\alpha$  solenoid according to 7pkt.

Note that whereas there are 131 proteins in *C. reinhardtii* automatically annotated by Uniprot as RAP-domain containing proteins, we retrieved only 4 of them with IPB but 116 more with DT or RF, confirming that these annotations are not based on strict similarity criterion, and also that RAP-domain are tightly related to OTAFs as previously described for NCL OPR (Boulouis et al. 2015). The IPB procedure has 95% precision and 93% sensitivity (Table S3.2).

In an attempt to find new OTAF candidates without relying on sequence similarity from already known OTAF families, we developed two other procedures, described below, that we first applied to the proteomes of *C. reinhardtii* and *A. thaliana* to assess their performance.

#### Decision tree on predicted protein properties

We first built a decision-tree procedure (referred to as DT procedure, see Methods) based on properties deduced from the primary sequences to retrieve  $\alpha$ -solenoid proteins composed of repeated motifs and targeted to an endosymbiotic organelle (Figure 1, central panel). First, we eliminated candidates with a predicted transmembrane helix (see Method), as most of the published works describe OTAFs as soluble proteins. Next, candidates must contain at least one repeat motif of 29 amino acids, at least 4 pairs of  $\alpha$ -helices (3 linkers) with a majority of residues predicted to fold into an  $\alpha$ -helix in the region defined by the N-terminal most and the C-terminal most  $\alpha$ -helix pair. The threshold values used at each step have been determined to maximize the precision (see Methods). Applied to *A. thaliana*, it retrieved only 104 PPR, 4 of which were not retrieved by IPB, and 30 new *pto* candidates. Applied to *C. reinhardtii*, it retrieved only 23/6 of

the published/unpublished OPRs, all but one (Cre09.g392579) already retrieved by IPB; it also retrieved 2 PPR and 93 new *pto* candidates. Therefore, the DT procedure is very specific (0.99) at the expense of sensibility (0.18) (Table S3.2).

<i>locus ID</i>	<i>annotation v5.6</i>	<i><math>\alpha</math>-solenoid shape</i>	<i>domain Uniprot</i>
Cre03.g145867	-	yes, 7pkt	RAP
Cre03.g201103	Raa7, psaA mRNA trans-splicing factor	yes, low quality (AF2)	Protein kinase
Cre06.g251400	MME6, NADP-dependent malic enzyme	no	malic enzyme
Cre06.g268750	MME1, NADP-dependent malic enzyme	no	malic enzyme
Cre16.g651900	Cril, carotenoid isomerase	yes (AF2)	RAP
Cre02.g109200	-	yes (AF2)	Uncharacterized
Cre04.g217909	-	no	Protein kinase
Cre09.g395880	-	yes (AF2)	RAP domain
Cre10.g431100	-	no	TPR_REGION
Cre13.g605850	-	no	RAP
Cre13.g605900	-	no	MYND-type

**Table 3.** 11 *pto* candidates retrieved by IPB with at least 2 OPR motifs. The last column indicates the domain found by the automatic annotation at Uniprot, but not retrieved by InterProScan.

### Machine learning based on predicted protein properties

We also developed a random forest (RF) procedure to avoid the definition of threshold values to select candidates. In the RF procedure, each protein is characterized by the properties used in the DT procedure, plus the 20 amino-acid frequencies, and a description of the protein physico-chemical properties by 36 Auto-Cross Correlation terms between the Z-scales amino-acid descriptors (Hellberg et al. 1987) as described in (Garrido et al. 2020). The RF classifier was trained on the same training set as for the DT procedure (see Methods). Over 1000 iterations of the model on our validation test (those 10% proteins from the positive and negative sets that were never used to train the model), on average the precision is 0.94, sensitivity 0.94 and specificity 0.92. Applied to the *A. thaliana* proteome, the RF procedure retrieved 350 PPR, 6 of which were not retrieved neither by IPB nor by DT; as well as 83 new *pto* candidates. In *C. reinhardtii*, it retrieved 53/56 of published/unpublished OPR, including NCL18, the sole published

OPR that was not detected by IPB and 12 others unpublished OPR that were not retrieved by IPB nor DT (Table S4). RF also identified 857 new *pto* candidates.

Analysis of the model showed that the most important properties (importance score > 0.02) are the sequence repeats and the linkers between  $\alpha$ -helices, the probability of being a linker, frequencies of 6 hydrophobic amino acids (leucine, methionine, isoleucine, proline, alanine and tryptophane), 3 polar residues (cysteine, lysine, and threonine) and 4 ACC values. Only Z<sub>1</sub> scale reflecting hydrophobicity and Z<sub>2</sub> scale reflecting the steric properties of the amino acids are involved in those ACC (Table Supp\_Zscale). We recently showed that these 4 ACC values reflect the amphipathic properties of an  $\alpha$ -helix (Garrido et al. 2020).

All retrieved candidates by IPB, DT and RF are listed in Table S4.

#### Comparison of the results obtained with IPB, DT and RF

In *A. thaliana*, the use of IPB with the PPR motif retrieved only PPR proteins, while IPB used with the OPR motif retrieved the only OPR previously identified in this organism. In an attempt to identify new candidate OTAFs for further analysis, we considered only *pto* candidates identified by DT (30) and by RF (83). Most of the 20 candidates identified only by DT have predicted structures containing  $\alpha$ -solenoid domains, except 2 that have no AF2pred. and 3 that have poor AF2pred., but with  $\alpha$ -helices and coiled regions. Among those 3 with poor 3D structure prediction, one hypothetical protein is worth mentioning because it has been inferred as being involved in RNA metabolic process based on co-expression networks analysis (Depuydt and Vandepoele 2021). Among the 10 candidates identified by both DT and RF (Table 4), one interesting OTAF candidate is AT3G29190, automatically annotated as a member of the Terpenoid cyclases/Protein prenyltransferases superfamily with an  $\alpha$ -solenoid shape. Two candidates contain TPR repeats and 5 contains ARM repeats. The remaining two are predicted to be localized in the nucleus according to TAIR annotations, despite the fact that they were predicted as *pto* by our procedure (at least 2 predictions among 4 for mitochondrial or

chloroplast localization, see Methods) and to be involved in DNA repair and chromosome condensation, which makes them unlikely to be actual *pto*, unless they are also involved in mitochondrial DNA repair.

locus ID	Domain Uniprot (automatic annotation)
AT1G59850	ARM repeat superfamily protein
AT2G44900	ARABIDILLO-1
AT2G45720	ARM repeat superfamily protein
AT3G02840	ARM repeat superfamily protein
AT3G14950	tetratricopeptide-repeat thioredoxin-like 2
AT3G29190	Terpenoid cyclases/Protein prenyltransferases superfamily protein
AT3G48190	Serine/Threonine-kinase ATM-like protein
AT4G15890	binding protein
AT4G31890	ARM repeat superfamily protein
AT5G37590	Tetratricopeptide repeat (TPR)-like superfamily protein

**Table 4.** 10 *pto* candidates retrieved only by DT and RF in *A. thaliana*.

In *C. reinhardtii*, 39 candidates were not retrieved by IPB, but by both DT and RF, including the OTAF Mac1, member of the HAT family that stabilizes the *psaC* mRNA, a subunit of photosystem I (Douchi et al. 2016). Note that Mbb1, the paralog of Mac1, is found only by the RF procedure. There are also 3 TPR repeats containing proteins, and the PPR Tcb1. Among the most interesting uncharacterized candidates are two proteins automatically annotated as containing a RAP domain (Cre03.g202450 and Cre14.g622900) and two proteins with an  $\alpha$ -solenoid shape in AF2pred: Cre01.g025500 containing a TPR domain and Cre12.g530750, containing a domain of unknown function (DUF4042) (Table 5).

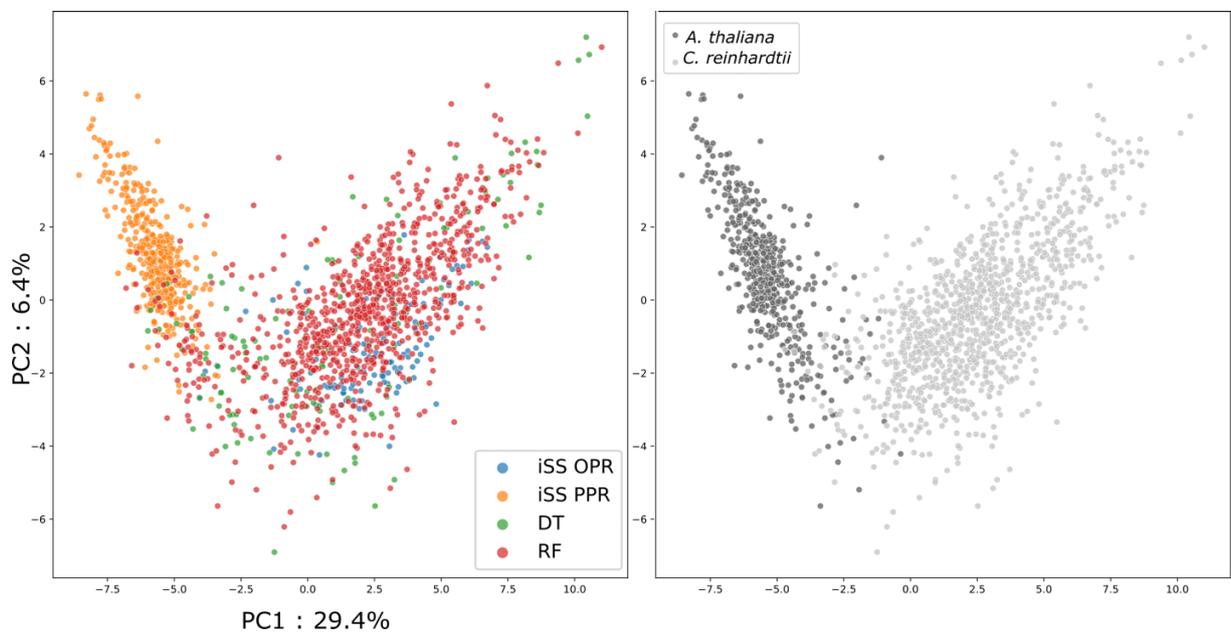
locus	v5.6	$\alpha$ -solenoid	uniprot (Phytozome)	# pto paralogs
Cre01.g007100	-	-	EXS domain	0
Cre01.g015300	-	yes	ANK_REP	12
<b>Cre01.g025500</b>	-	<b>yes</b>	<b>TPR</b>	<b>0</b>
Cre01.g034650	-	yes	APC5 Anaphase-promoting complex subunit 5*	0
Cre01.g050500	TCB1, PPR1	yes	PPR_long domain	0
Cre03.g154800	-	yes	ANK_REP	6
Cre03.g156700	FAP185, Nphp3	-	TPR_REGION domain	1
Cre03.g158700	-	yes	ANK_REP	12
<b>Cre03.g202450</b>	-	-	<b>RAP domain</b>	<b>0</b>
Cre04.g231516	-	yes	BTB domain*	0
Cre05.g230700	-	yes	ANK_REP	6
Cre05.g236350	-	yes	ANK_REP	8
Cre05.g236550	-	yes	ANK_REP	8
Cre06.g255100	-	no	GRIP domain	0
Cre06.g283450	-	yes	ANK_REP	7
Cre06.g298200	-	yes	ANK_REP	8
Cre06.g306150	-	+/-	DUF262 domain*	0
Cre07.g341925	-	-	DUF4456	2
Cre07.g345650	-	-	PEROXIDASE_4 domain	0
Cre08.g372716	-	yes	ANK_REP	7
Cre09.g389615	MAC1	yes	TPR	0
Cre09.g404500	SFI1	-	Sfi1 domain	0
Cre09.g409000	-	yes	ANK_REP	4
Cre10.g420537	-	yes	ANK_REP	7
Cre10.g428692	-	yes	ANK_REP	12
Cre10.g432950	-	yes	ANK_REP	7
Cre10.g464950	-	-	TPR	0
Cre11.g467712	-	no	CBM20 domain	0
Cre12.g493550	THK3	yes	TPR domain containing	-
Cre12.g521800	-	-	Protein kinase domain	1
Cre12.g525000	-	yes	ANK_REP	12
<b>Cre12.g530750</b>	-	<b>yes</b>	<b>DUF4042 domain</b>	<b>0</b>
Cre12.g539700	-	yes	ANK_REP	8
Cre13.g569300	-	yes	ANK_REP	8
<b>Cre14.g622900</b>	-	-	<b>RAP domain</b>	<b>0</b>
Cre15.g637249	-	-	MYND-type domain	0
Cre16.g688190	-	yes	Importin N-terminal domain	0
Cre17.g725900	-	no	uncharacterized	0
Cre18.g749247	-	yes	TOG domain	0

**Table 5.** 39 *pto* candidates retrieved only by DT and RF in *C. reinhardtii*. \*: domain retrieved by sequence search with InterProScan. Potential interesting candidates mentioned in the text are in bold.

The principal component analysis (PCA) of all *pto* candidates described by the properties used in the RF procedure (Figure 2) shows that *pto* candidates from *A. thaliana* and *C. reinhardtii* are strikingly distinct. This separation is not solely due to amino acid frequencies and is not observed between non  $\alpha$ -solenoid proteins from these two species (Figure S1). It is also found that RF and DT candidates globally overlap with IPB ones, indicating that these two similarity independent procedures succeed at identifying the defined  $\alpha$ -solenoid properties. Interestingly, there is an area (upper left for *C. reinhardtii* candidates and lower center for *A. thaliana* candidates) where RT and DT candidates co-localize, at the exclusion of IPB candidates, that could correspond to potential new families. Of note, there is also no area devoted to only RF candidates, suggesting that DT and RF do not differ that much, although they have contrasting profiles in terms of sensitivity and sensibility.

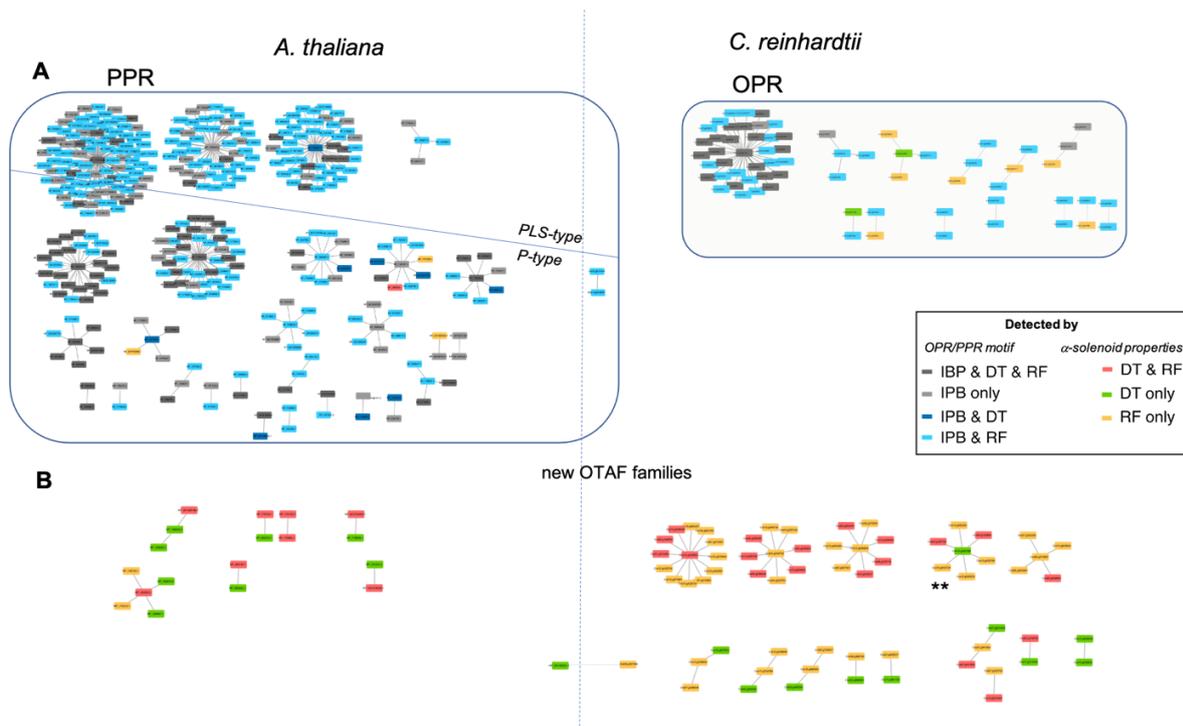
Candidates retrieved by IPB share OPR or PPR motifs while candidates retrieved by DT and RF have alpha-solenoid properties. In order to determine the extent at which those candidates share sequence similarity over their entire length, indicative of a common origin, and a possible conserved function, *pto* candidates in *C. reinhardtii* and *A. thaliana* were clustered together based on the E-value of the similarity search ( $10e^{-6}$ ) and a high hit coverage ( $> 0.7$ ) of the two compared sequences (Figure 3). The complete statistics of the clustering is given in Table S6, including the number of singletons (candidates with no paralog). Clusters of paralogs are exclusively species-specific (to the notable exception of two clusters), indicative of several expansions by independent duplications in Chlorophytes and Streptophytes. The average similarity percentage shared by homologs in a given cluster, with similarity covering more than 70% of each sequence is low on average (between 25 and 50%), suggesting that these paralogs arose, on average, from ancient duplications.

The several clusters in each species reflect the modular composition of OTAF proteins, composed of an  $\alpha$ -solenoid region provided by different combinations of degenerated motifs and additional domains. Indeed, several domains are annotated. In PPR (Figure 3A), the PLS-type proteins contain DYW and E domains (Barkan and Small 2014). In *Chlamydomonas* new pto candidates (Figure 3B) are annotated with Mynd-type, protein kinase, Reverse transcriptase, Elmo and RING domains by ProtNLM at Uniprot, where InterProScan annotations more frequently correspond to disordered regions. The clustering shows that DT and RF procedure succeed at identifying OPR and PPR homologs, also identified by IPB (colored light and dark blue, respectively). In *C. reinhardtii*, 11 proteins cluster with known OPR and could likely represent remote OPR homologs. Two of which, Cre13.g578950 and Cre12.g549900, clustering with a described OPR (Cre07.g336500) that was retrieved only by RF. Among DT and RF candidates, two contain RAP domains: Cre06.g278247 and the low-complexity Amc1 protein (Cre16.g688900), with compositional proximities to OPR and PPR, required for production of mitochondrially-encoded complex I subunit ND4 (Subrahmanian et al. 2020), therefore very likely to be an OTAF. Only 4 pto candidates retrieved by DT or RF clustered with PPR in *A. thaliana*, being annotated as part of the TPR-like or PPR-like superfamily in TAIR.



**Figure 2.** Principal component analysis of pto candidates in *C. reinhardtii* and *A. thaliana*

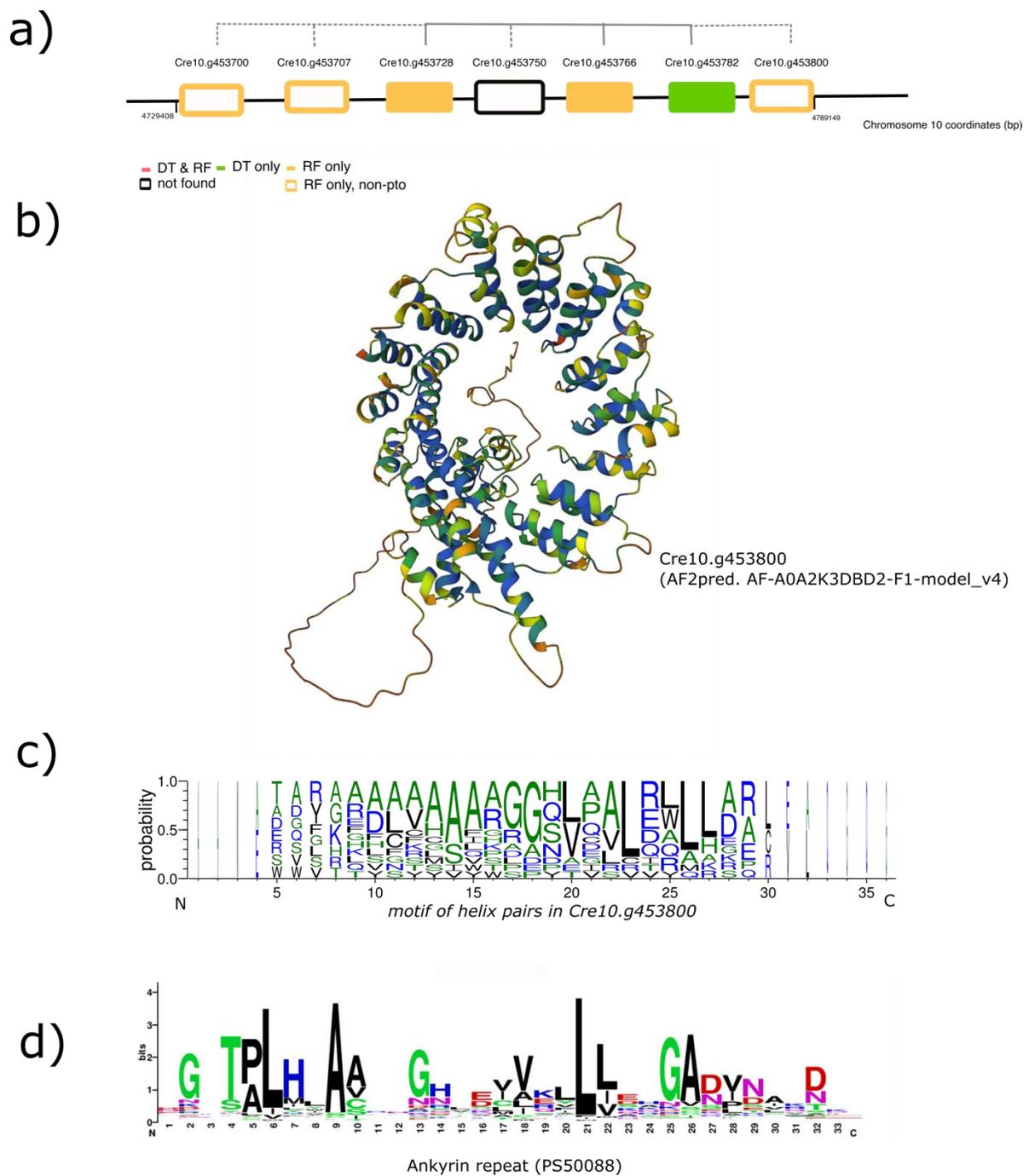
described by properties used by the RF procedure. In the left panel, candidates retrieved by IPB are colored yellow (PPR motif), blue (OPR motif); candidates not retrieved by IBP but by DT are colored green; candidates retrieved only by RF are colored red. The same PCA is represented in the right panel, with candidates colored light grey (*C. reinhardtii*) or dark grey (*A. thaliana*). Explained variance on PC1 and PC2 are given on the left panel.



**Figure 3.** Clusters of pto candidates found in *A. thaliana* (left panel) and *C. reinhardtii* (right panel). Pto candidates linked by grey edges are in the same cluster and share significant similarity (E-value < 10e-6) over at least 70% of their sequences (see Methods). Pto candidates are colored depending on the procedure that retrieved them, according to the color legend in the rectangle inset on the right center of the figure. **A.** Families of known OTAFs. Pto candidates with grey and dark dots are related to known OTAF, mainly PPR in *A. thaliana* and OPR in *C. reinhardtii*. In the PPR rectangle, PLS-type are above the oblique line, while P-type PPR are below. Some pto candidates found by DT and RF cluster with OPR and PPR. **B.** Families of new OTAF candidates. Clusters with yellow, red and red dots are new pto families. 3 pto candidates from the cluster indicated by a double star (\*\*) form a tandemly duplicated region on chromosome 10 (see Figure 4). The dotted blue line separates candidates of *A. thaliana* (left) and *C. reinhardtii* (right).

### Other species-specific families of $\alpha$ -solenoid proteins

In addition to the PPR and OPR clusters, 47 candidates are grouped in 6 clusters of more than 4 paralogs, including at least one candidate found by both by DT and RF. Only one such cluster is found in *A. thaliana*, composed of 5 uncharacterized homologs including the above mentioned AT3G29190. The 5 other clusters are specific to *Chlamydomonas reinhardtii*, of unknown function. Three of such *pto* candidates are tandemly duplicated on chromosome 10 with 3 other non-*pto* paralogs (*i.e.* identified by our approach, but not predicted to be addressed) and another paralog that was not retrieved by any of our procedures (Figure 4a). All the 7 paralogs in that duplicated region share more than 60% identity over at least 70% of their sequence and adopt an  $\alpha$ -solenoid fold (Figure 4b). The sequences of the 14 helix pairs from the predicted structure of Figure 4b (see Methods) correspond to a new motif (Figure 4c). It must be noticed that these proteins are annotated as containing an ankyrin repeat, but the determined motif diverges from the ankyrin motif (Figure 5.4d). It has been shown in *A. thaliana*, that an ankyrin repeat containing protein targeted to the chloroplast (AKRP\_ARATH) blocks chloroplast differentiation (Zhang et al. 1992). This cluster might thus represent a reservoir of candidates for further experimental characterizations in search for a possible photosynthetic phenotype of the corresponding mutants.

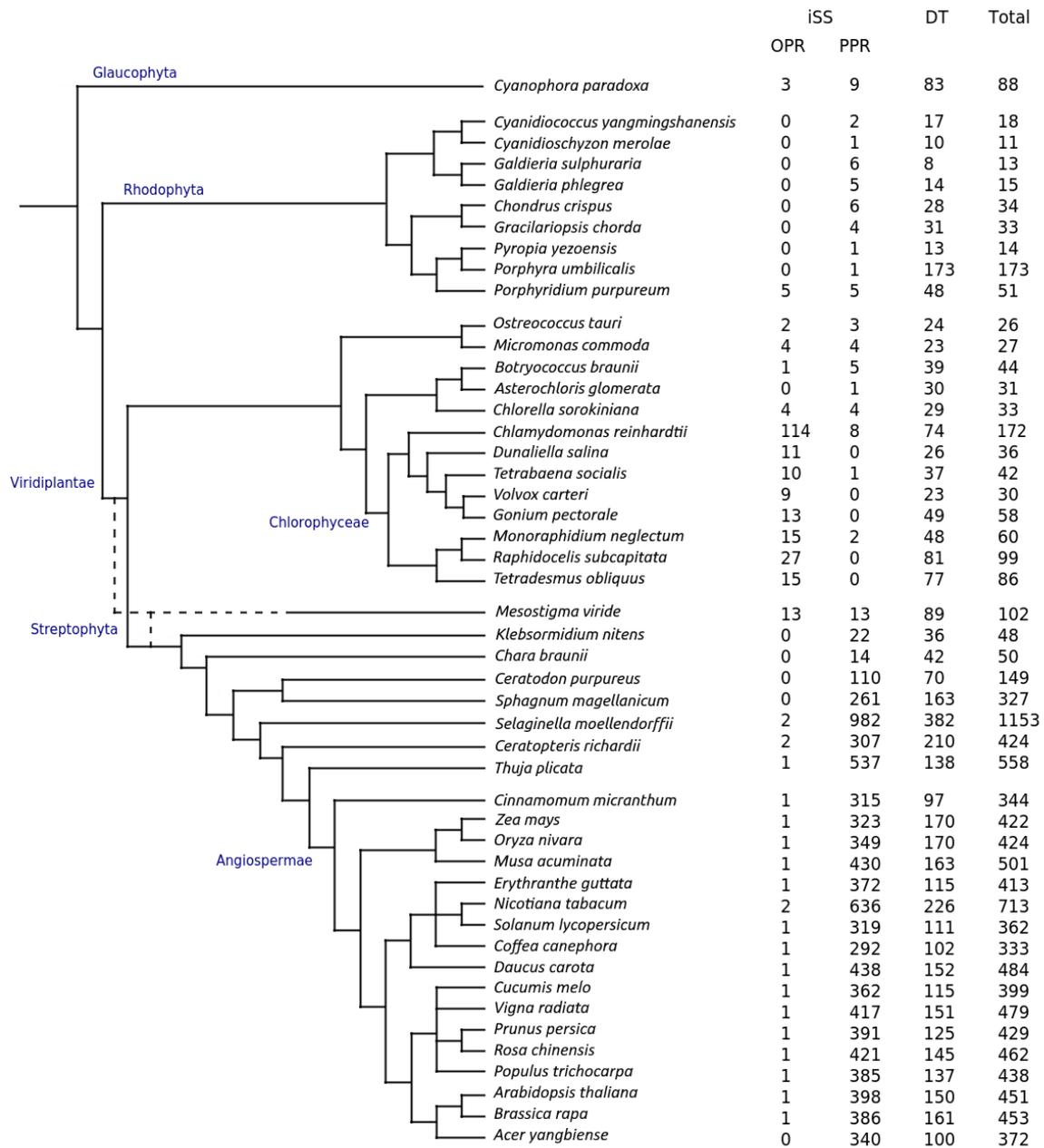


**Figure 4** **a)** Tandems of pto candidates on chromosome 10 of *C. reinhardtii*. All proteins contain an ankyrin repeat motif (found by InterProScan search) and colored similarly as in Fig. 4. Solid grey lines relate proteins in the same cluster of pto candidates (Fig. 4). Dashed grey lines relate non-ptc to pto candidates. **b)** AlphaFold2 prediction of Cre10.g453800, **c)** Sequence logo of the alignment of the sequences of 14  $\alpha$ -helix pairs retrieved from the PDB file of the AlphaFold2 model. **d)** Sequence logo of the Prosite Ankyrin repeat.

### OPR proteins have undergone major expansions in Chlorophyceae

Figure 5 shows the distribution of the 11407 *pto* candidates retrieved by IPB and DT across the 48 Archaeplastida proteomes studied. Note that they are the most confident candidates with at least 2 predictions of targeting to mitochondria or chloroplast. For *C. reinhardtii*, the higher number of *pto* candidates can be explained by the fact that 6 subcellular localization predictions are used instead of 4 for all the other species (see Methods).

PPR are present in all groups, in low copy numbers (less than 10 per species) in Rhodophyta, Glaucophyta and Chlorophyta and underwent massive expansions in Streptophyta with several hundred copies per species.



**Figure 5:** Mean numbers of detected pto candidates detected by IPB (with OPR and PPR motifs) and new pto OTAF candidates detected by DT across proteomes of 48 Archaeplastida species. The last column indicates the total number of retrieved pto proteins (IPB+DT). In *C. reinhardtii* and in *A. thaliana*, there are 39 and 10 pto candidates detected by DT that are also detected by RF. For *C. reinhardtii*, pto candidates are selected based on 6 subcellular localization predictions, while for all the other species it based only on 4 ones. See Methods for the reference topologies.

OPR are present in the Glaucophyta *Cyanophora paradoxa* but absent from Rhodophyta, except in *Porphyridium purpureum* where 5 with an  $\alpha$ -solenoid shape in AF2pred. are retrieved by IPB and DT (three are annotated as Tbc2 translation factor, and 2 as containing a RAP domain). OPR expanded in Chlorophyceae, most of them containing more than 10 OPR copies. On the contrary, in most of the Streptophyta, OPR are present in a unique copy or absent, except in *Mesostigma viride*, that contain 13 candidates OPR found by IPB. Five of them clustering together and with two paralogs in the Glaucophyta *Cyanophora paradoxa*, and two Chlorophyta homologs (in *Micromonas commoda* and *Ostreococcus tauri*) found by DT (Table Supp 5 to be added).

## Discussion

### Critical assessment of the new proposed procedures IBP, DT and RF

The proposed IPB procedure differs from PPRfinder (Gutmann et al. 2020) in several ways: first we restrict our search on complete proteomes, therefore missing all unannotated CDS, while PPRfinder search on translated genome assemblies into the six frames. Second, in IPB, the motifs are based on stringent similarity criterion applied to multiple alignments, leading to a large number of different profiles, and the procedure is iterative, while PPRfinder relies on large profiles, with supplementary criterions: PPRfinder retain only proteins with 2 neighboring motifs with a similarity score threshold. To increase the specificity of the IPB search, one can remove the last BLASTPsearch, *i.e.* not taking singletons and poorly aligned motifs into account.

For the DT procedure, that has been implemented before the release of AlphaFold2, it could be worth testing if it can be improved by using more efficient tools for single residue 2D predictions such as NetSurfP (Høie et al. 2022). However, NetSurfP already uses machine learning. If its use in a DT procedure performs better or as well as a full machine learning procedure, it will be worth using the DT procedure that will obviously have a lower computational cost.

It would also be worth testing different parametrization of the profile-based procedure to determine if this could be improved, notably by testing different criterion to estimate the quality

of the alignment and of the clustering. Our proposed list of candidates is therefore most probably incomplete but already provides targets for experimental characterization.

Some attempt to develop machine learning procedures has been already made for PPR discovery, based on protein features extracted from PPR as a positive training set and non-PPR as negative training sets has also been proposed, MixedPPR (Qu et al. 2019), but the performance of these methods have been estimated only on the training sets, and it is not known how they compare to profile PPR-based methods and how it copes with taxonomical bias. Similar methodologies, based on the same dataset with features differing in the way machine learning is performed (Feng et al. 2021; Zhao et al. 2021), mimic the above publication. There are therefore avenues to improve the machine learning procedure, by exploring different properties or different descriptors, by trying natural language processing models and neural networks models and most importantly applying them to real datasets. Our candidate list is also determined by the training set we used. It could be greatly improved with AlphaFold predictions also some are very low, especially for proteins containing disordered and low complexity regions, as it appears to be the case for OTAFs.

### Biological assessment

IPB, the iterative sequence similarity-based procedure efficiently retrieves known OTAFs in *C. reinhardtii* and *A. thaliana* and even identified potential new remote homologs in those two species. Starting from motifs found in a single species, the iterative process allows to capture specific signatures in diverse taxonomic groups. It confirms that PPR motifs are well conserved and present in all Archaeplastida lineages with a massive expansion in Angiospermae; while OPR motifs are highly divergent between groups and mainly present in green microalgae. Note that more candidates, failing to be correctly predicted to be addressed in organelles could be present.

The DT procedure is very stringent and failed to detect a number of the known OTAFs. However, the RF procedure, on the contrary to DT, is very sensitive but lacks specificity. RF showed that the amphipathic nature of the helices, allowing the formation of a hydrophilic cavity within the  $\alpha$ -solenoid structure, in which the positively charged mRNA could bind, is crucial for the

classification. Together DT and RF identified new *pto*  $\alpha$ -solenoid candidates, of yet unknown OTAFs families: 39 new candidates in *C. reinhardtii*, 5 new candidates in *A. thaliana*, and most excitingly dozens of new candidates in the Glaucophyta *Cyanophora paradoxa* and in the red alga *Porphyridium purpureum*. The distribution of OTAF is heterogeneous, and are the most numerous in land plants due to PPR. In Rhodophyta and Glaucophyta, new OTAFs candidates were found, suggesting that modes of regulation might vary. The proposed procedures, the improvement of which is currently investigated, provide complementary properties and represent a first step to decipher repertoires of OTAFs candidates in largely unexplored lineages of photosynthetic eukaryotes, in Rodophytes, Glaucophytes and all eukaryotes with complex plastid resulting from secondary endosymbiosis, paving the way for elucidating the rules of the regulation of organelle genomes in these complex plastids.

## **Materials and methods**

### Sequence data, assembly quality and software programs

The 48 proteomes from Archaeplastida species were retrieved either at NCBI, Uniprot or JGI websites. The details of the proteomes and their versions are given in Table S1.

Quality and completeness of proteomes were assessed with BUSCO 5.3.2 (Manni et al. 2021), with protein mode on the Eukaryota and Viridiplantae lineages (eukaryota\_odb10, viridiplantae\_odb10, 2020-09-10). All the details of parameters used and results are given in Table S2.

As starting points for the motif-based procedure (IPB), we used published and described repeat motifs from 12 known OPR proteins from *Chlamydomonas reinhardtii* (green microalga model species) and from 11 known PPR proteins from *Arabidopsis thaliana* (land plant model species), listed in Table 1.

### Annotations, structural predictions

Proteins annotations were retrieved at Uniprot for all proteins. For *Chlamydomonas reinhardtii*, the annotation v5.6 available at Phytozome, version 13 was also used, as well as the TAIR10 annotation for *A. thaliana*. 3D structure predictions were retrieved at the AFold Protein Structure Database or estimated locally with AlphaFold2 (Jumper et al. 2021). Domain structure predictions were performed via the InterProScan server in January 2023. Secondary structure assignment from PDB coordinates was made with STRIDE (Heinig and Frishman 2004).

### Reference phylogeny

The 48 species from Archaeplastida were selected based on proteome status as provided by the Published Plant Genomes (PPG) database (<https://www.plabipd.de/>) on June 2020. Reference phylogeny was inferred from PPG and the following articles (Hanschen and Starkenburg 2020; Sibbald and Archibald 2020; Gawryluk et al. 2019; One Thousand Plant Transcriptomes Initiative 2019; Lemieux et al. 2007; Nakamura et al. 2013; The Angiosperm Phylogeny Group et al. 2016; Leliaert et al. 2012; Cheng et al. 2019), with the help of the NCBI Taxonomy tool.

### Motifs-based similarity procedure

The IPB procedure follows the four steps described below, that we ran iteratively either on OPR or on PPR motifs.

#### *Step 1. Pairwise comparison of motifs*

First, an all-against-all pairwise comparison of all motifs is performed by BLASTP v2.6.0, Camacho et al. 2009). We kept only the pairs of motifs whose e-values were lower than a threshold  $t$  varying according to  $X$  the size of the data set, as follows: if  $10^{n-1} \leq X \leq 10^n$ , then  $t = 10^{-n}$ .

#### *Step2. Motifs clustering*

Second, motifs are clustered with the MCL v.14-137 (Enright et al. 2002) based on the  $-\log(E\text{-value})$  of the BLASTP hits. Clusters are computed with inflation parameter  $I$  values,

starting from 1.1 and by increasing it by step of 0.1 until. For each I parameter tested, a multiple alignment of the motifs in a given cluster is performed with MAFFT v7.450 (Kato and Standley 2013). We choose the clustering with the I value providing a maximum number of clusters with at most 18 and 16 gaps in their multiple alignment, for OPR and PPR motifs respectively. These values correspond to half of an OPR or a PPR motif as we wanted to detect motifs that were similar to the starting motifs over at least half of their sequence and avoid “motif slippage”.

### *Step3. HMM profiles search against proteomes of interest*

Third, profiles are built with *hmmbuild* from the HMMer suite, v3.1b2 (Mistry et al. 2013) . Each HMM profile was then searched against each of the 48 proteomes with *hmmsearch*. The length of an OPR/PPR motif being about 35/38 amino acids, only newly found sequences whose lengths are -6/+2 the length of an OPR motif, *i.e.* 32/40 amino acids long are kept. Two motifs overlapping by more than 80% are considered identical and the motif boundaries are defined by the minimum overlapping region to reduce the tendency of the motifs from slipping in one side or the other. Two motifs overlapping by less than 20% are considered different and kept without changing their boundaries. Pairs of overlapping motifs between 20% and 80% are both eliminated, being considered as no longer representative of the typical starting motif.

### *Step5. Selection of candidates on targeting properties*

The procedure is iterated restarting from step 1 after adding the new motifs to the initial data set only if at least 1 new pto candiate, *i.e.* predicted to be addressed by at least two prediction softwares is retrieved. See “prediction of localization” paragraph for details of the prediction.

### *Step6. Final sequence similarity search with all motifs found*

As described above, only clusters with a good multiple alignments, with less than dozens of gaps are used in Step 2. To consider motifs within poorly aligned clusters and singleton motifs (clusters of size one), a last BLASTP search of all motifs was performed against a meta-proteome

(proteome made up of all the proteomes of interest). Hits with E-value < 0.001 are added to the already retrieved motifs.

#### *Step7. Selection of the candidate proteins*

The final step consists in selecting only *pto* proteins. For *C. reinhardtii*, predictions from TargetP v1.1b (Emanuelsson et al. 2000) and PredAlgo (Tardif et al. 2012) as provided in the *C. reinhardtii* annotation version v5.6 are also taken into account. Therefore, a protein in *C. reinhardtii* will be considered as *pto* candidate if it has at least 2 predictions to be localized in an endosymbiotic organelle over 6 predictions, while in all other species, *pto* candidates have at least 2 good predictions over 4.

#### Similarity-independent procedures

In order to detect  $\alpha$  solenoid proteins without relying on sequence similarity from already known  $\alpha$  solenoid proteins, we draw a portrait of each protein based on some properties deduced from its primary sequence: the presence of repeats, the 2D helix structure, the presence of  $\alpha$ -helix pairs able to adopt an  $\alpha$ -solenoid shape, and the subcellular localization. We exploit these properties by two similarity-independent procedures, one based on a decision tree (DT) and one based on a random forest (RF) classification.

First of all, proteins with a predicted transmembrane helix after its N-terminal part (corresponding to the targeting peptide and that could erroneously be predicted as a transmembrane helix) were removed. TMHMM v2.0c (Krogh et al. 2001) was used to predict transmembrane helix. The end of the N-terminal part was defined to the 78<sup>th</sup> residue for the DT/RF procedure.

Repeats are predicted with RADAR v1.3 (Heger and Holm 2000). The 2D prediction at the residue level were performed with S2D v2 (Sormanni et al. 2015), the presence of  $\alpha$ -helix pairs was estimated with ard2 (Fournier et al. 2013), which uses a neural network procedure with a training

set containing X-ray resolved  $\alpha$ -helix pairs and gives a score for each amino acid position to be a linker between two  $\alpha$  helices.

#### Decision tree (DT) procedure

The decision tree used to select the good candidates by the DT procedure is given in Figure 1. Parameters for selection along the decision tree have been defined based on a set of validated  $\alpha$ -solenoid protein and a set of validated non  $\alpha$ -solenoid proteins (see the paragraph *Training set composition*). Table S3 recapitulates numbers of retrieved candidates, precision, accuracy and recall for all tested combinations of parameter values. We choose the global combination maximizing precision (TP/TP+FP).

#### *Filtering of proteins according to the number of sequence repeats*

We selected the proteins in which at least one repeat of at least 29 amino acids is detected by RADAR. 29 amino acid length gives the best precision score among a series of tested values, from 20 to 40 (Table S3).

#### *Detection of $\alpha$ helices pairs*

In order to keep only proteins with multiple pairs of  $\alpha$ -helices, we constructed and tested dozens of sets of criteria based on the distance between 2  $\alpha$ -helix pairs and the number of linkers between the helices forming a pair (Table S3). We kept only the proteins with at least 4 linkers, with the distance between two consecutive linkers comprised between 32 and 400 amino acids. A given amino acid is considered as a linker if its ard2 score is above 0.15, based on the results obtained on the training set.

#### *Protein filtering according to 2D structure predictions*

We kept only those proteins for which 65% of the amino acids were predicted in  $\alpha$ -helix by s2D, between the two most extreme linkers predicted by ard2. We tested for 50% to 100% and we chose the percentage filter having the best precision score (65%, Table S3).

### Sublocalization prediction

After these selection steps, we kept only those *pto* proteins, as in the final selection step of IPB.

### Random forest procedure

The RF procedure uses the above-described properties also used in the DT procedure: the number of repeats predicted by RADAR, the proportion of the sequence in the detected repeats, the number of linkers predicted by ard2 and the probability for an amino acid to be a linker; 4 predictions of subcellular localization. In addition, the RF model uses the 20 amino acid frequencies and 36 Auto-Cross Correlation (ACC) values of neighboring amino acids over a window of 4 residues computed for each protein based on the Z-scales amino acid descriptors (Hellberg et al. 1987) as described in (Garrido et al. 2020). The training set is the same as the one used for the DT procedure, described in the paragraph Training set composition. The training set was divided in a validation set (10%) never used for learning or testing and the remaining 90% was split in two to use 70% of the set for model training and 20% for performance testing.

The random forest classification was performed with the *RandomForestClassifier* function from the scikit-learn Python package (version 0.21.2). We tested a range of values for the number of estimators (from 100 to 800, selected value 500) and for the minimum sample split (from 5 to 50, selected value 30). The selected values give the best performance on the validation set (after 1000 trials of the model: an accuracy of 0.94, a sensitivity of 0.94 and a specificity of 0.92), all other parameters were kept by default. We used the mean decrease in impurity procedure to determine the importance of each feature in the model.

### Sublocalization prediction

Deeploc v2.0 (Almagro Armenteros et al. 2017), LOCALIZER v1.0.4 (Sperschneider et al. 2017), TargetP-2.0 (Emanuelsson et al. 2007) and WoLFPSort v0.2 (Horton et al. 2007) were used to predict the chloroplast or mitochondrial localization.

### Training set composition

The training set for DT and RF procedures comprises 1081 known  $\alpha$ -solenoids (positive set) and 1196 known non  $\alpha$ -solenoids (negative set). The positive set contains the 481 annotated PPR proteins from *A. thaliana* available at the PPR database <https://ppr.plantenergy.uwa.edu.au/>, retrieved in June 2020. To increase the diversity of the positive training set, we included the 111 candidates from *C. reinhardtii* (101 with at least one OPR motif and 9 with at least one PPR motif) retrieved by the IPB procedure and candidates retrieved by a first exploratory DT search. All IPB and DT candidates were clustered based on their sequence similarity (proteins sharing a BLASTP hit with E-value  $< 10^{-6}$ ). 18 proteins in *C. reinhardtii* and 470 proteins from the 48 other Archaeplastida species but *A. thaliana* were selected among clusters devoid of known PPR or OPR, with an  $\alpha$ -solenoid domain based on 3D structures retrieved in the literature (the selection was performed before the AFold release). The negative dataset contains 553, 391 and 252 from *C. reinhardtii*, *A. thaliana* and the other 46 Archaeplastida species, respectively. The details of the sequences used as positive and negative controls are available in the archive 3.

### Principal components analysis

The two principal components of a principal component analysis (PCA) of the proteins defined by the properties used in the RF procedure were performed. The weights of each variable in the PCA are summarized by correlation circles in Figure S1. Analyses were performed with the scikit-learn Python package version 0.21.2.

### Motifs analysis

Sequence logo were performed with the standalone version of WebLogo 3 (Crooks et al. 2004).

### Clustering

Selected pto candidates were clustered by the *easy-cluster* mode of mmseq2 v14.7e284 (Steinegger and Söding 2017), with an E-value  $< 10e^{-6}$  and a hit coverage of at least 50% of both sequences, all other parameters set to default values. Visualisation was performed with Cytoscape (v3.9.1) (Shannon et al. 2003).

Code availability: The Python scripts are available on Github

### **Supplementary data**

Supplementary Tables are available at:

<https://dropsu.sorbonne-universite.fr/s/NHMZR5Y5M5c9Xf2>

### **Acknowledgments**

We thank Hédi Soula for fruitful discussions and advice for setting up the machine learning procedure; Richard Lavery for initial discussions on molecular dynamics; Yves Choquet for fruitful discussions on OPR; Olivier Vallon for providing access to the unreleased OPRdb database and for stimulating discussions with CC.

This work was supported by funding from the Centre National de la Recherche Scientifique and Sorbonne University to UMR7141; the “LabEx Dynamo” (ANR-LABX-011), the DECRYPTOR grant (CNRS 80prime 2019) and the DecryProtARN grant (CNRS Infiniti 2018).

### **References**

- Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. 2017. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* **33**: 3387–3395. <http://academic.oup.com/bioinformatics/article/33/21/3387/3931857>.
- Andrade MA, Bork P. 1995. HEAT repeats in the Huntington’s disease protein. *Nat Genet* **11**: 115–116. <https://www.nature.com/articles/ng1095-115>.
- Archibald JM. 2015. Endosymbiosis and Eukaryotic Cell Evolution. *Current Biology* **25**: R911–R921. <http://www.sciencedirect.com/science/article/pii/S0960982215008891>
- Barkan A, Small I. 2014. Pentatricopeptide Repeat Proteins in Plants. *Annual Review of Plant Biology* **65**: 415–442. <http://dx.doi.org/10.1146/annurev-arplant-050213-040159>.

- Beick S, Schmitz-Linneweber C, Williams-Carrier R, Jensen B, Barkan A. 2008. The Pentatricopeptide Repeat Protein PPR5 Stabilizes a Specific tRNA Precursor in Maize Chloroplasts. *Molecular and Cellular Biology* **28**: 5337–5347. <https://journals.asm.org/doi/10.1128/MCB.00563-08>.
- Boulouis A, Drapier D, Razafimanantsoa H, Wostrikoff K, Tourasse NJ, Pascal K, Girard-Bascou J, Vallon O, Wollman FA, Choquet Y. 2015. Spontaneous dominant mutations in *Chlamydomonas* highlight ongoing evolution by gene diversification. *The Plant cell* **27**: 984–1001.
- Boulouis A, Raynaud C, Bujaldon S, Aznar A, Wollman F-A, Choquet Y. 2011. The Nucleus-Encoded trans-Acting Factor MCA1 Plays a Critical Role in the Regulation of Cytochrome f Synthesis in *Chlamydomonas* Chloroplasts[W]. *Plant Cell* **23**: 333–349. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3051260/>.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.
- Cheng S, Gutmann B, Zhong X, Ye Y, Fisher MF, Bai F, Castleden I, Song Y, Song B, Huang J, et al. 2016. Redefining the structural motifs that determine RNA binding and RNA editing by pentatricopeptide repeat proteins in land plants. *Plant J* **85**: 532–547. <http://onlinelibrary.wiley.com/insb.bib.cnrs.fr/doi/10.1111/tpj.13121/abstract>.
- Cheng S, Xian W, Fu Y, Marin B, Keller J, Wu T, Sun W, Li X, Xu Y, Zhang Y, et al. 2019. Genomes of Subaerial Zygnematophyceae Provide Insights into Land Plant Evolution. *Cell* **179**: 1057–1067.e14. [http://www.cell.com/cell/abstract/S0092-8674\(19\)31169-9](http://www.cell.com/cell/abstract/S0092-8674(19)31169-9).
- Choquet Y, Wollman F-A. 2002. Translational regulations as specific traits of chloroplast gene expression. *FEBS Letters* **529**: 39–42. <https://www.sciencedirect.com/science/article/pii/S001457930203260X>.
- Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res* **14**: 1188–1190.
- D’Andrea LD, Regan L. 2003. TPR proteins: the versatile helix. *Trends in Biochemical Sciences* **28**: 655–662. [http://www.cell.com/trends/biochemical-sciences/abstract/S0968-0004\(03\)00273-1](http://www.cell.com/trends/biochemical-sciences/abstract/S0968-0004(03)00273-1).
- Delannoy E, Stanley WA, Bond CS, Small ID. 2007. Pentatricopeptide repeat (PPR) proteins as sequence-specificity factors in post-transcriptional processes in organelles. *Biochemical Society Transactions* **35**: 1643–1647. <https://doi.org/10.1042/BST0351643>.

- Depuydt T, Vandepoele K. 2021. Multi-omics network-based functional annotation of unknown Arabidopsis genes. *The Plant Journal* **108**: 1193–1212. <https://onlinelibrary.wiley.com/doi/abs/10.1111/tpj.15507>.
- Douchi D, Qu Y, Longoni P, Legendre-Lefebvre L, Johnson X, Schmitz-Linneweber C, Goldschmidt-Clermont M. 2016. A Nucleus-Encoded Chloroplast Phosphoprotein Governs Expression of the Photosystem I Subunit PsaC in *Chlamydomonas reinhardtii*. *Plant Cell* **28**: 1182–1199. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4904667/>.
- Eberhard S, Loiselay C, Drapier D, Bujaldon S, Girard-Bascou J, Kuras R, Choquet Y, Wollman F-A. 2011. Dual functions of the nucleus-encoded factor TDA1 in trapping and translation activation of atpA transcripts in *Chlamydomonas reinhardtii* chloroplasts. *The Plant Journal* **67**: 1055–1066. <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-313X.2011.04657.x>.
- Emanuelsson O, Brunak S, Heijne G von, Nielsen H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocols* **2**: 953–971. <http://www-nature-com/articles/nprot.2007.131>.
- Emanuelsson O, Nielsen H, Brunak S, von Heijne G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol* **300**: 1005–1016.
- Feng C, Zou Q, Wang D. 2021. Using a low correlation high orthogonality feature set and machine learning methods to identify plant pentatricopeptide repeat coding gene/protein. *Neurocomputing* **424**: 246–254. <https://www.sciencedirect.com/science/article/pii/S0925231220302691>.
- Fournier D, Palidwor GA, Shcherbinin S, Szengel A, Schaefer MH, Perez-Iratxeta C, Andrade-Navarro MA. 2013. Functional and Genomic Analyses of Alpha-Solenoid Proteins. *PLOS ONE* **8**: e79894. <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0079894>.
- Garrido C, Caspari OD, Choquet Y, Wollman F-A, Lafontaine I. 2020. Evidence Supporting an Antimicrobial Origin of Targeting Peptides to Endosymbiotic Organelles. *Cells* **9**: 1795. <https://www.mdpi.com/2073-4409/9/8/1795>.
- Gawryluk RMR, Tikhonenkov DV, Hehenberger E, Husnik F, Mylnikov AP, Keeling PJ. 2019. Non-photosynthetic predators are sister to red algae. *Nature* **572**: 240–243. <https://www.nature.com/articles/s41586-019-1398-6>.
- Gutmann B, Royan S, Schallenberg-Rüdinger M, Lenz H, Castleden IR, McDowell R, Vacher MA, Tonti-Filippini J, Bond CS, Knopp V, et al. 2020. The Expansion and Diversification of

- Pentatricopeptide Repeat RNA-Editing Factors in Plants. *Molecular Plant* **13**: 215–230. [http://www.cell.com/molecular-plant/abstract/S1674-2052\(19\)30366-1](http://www.cell.com/molecular-plant/abstract/S1674-2052(19)30366-1).
- Hanschen ER, Starkenburg SR. 2020. The state of algal genome quality and diversity. *Algal Research* **50**: 101968.
- Heger A, Holm L. 2000. Rapid automatic detection and alignment of repeats in protein sequences. *Proteins* **41**: 224–237.
- Heinig M, Frishman D. 2004. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res* **32**: W500-502.
- Hellberg S, Sjoestroem M, Skagerberg B, Wold S. 1987. Peptide quantitative structure-activity relationships, a multivariate approach. *J Med Chem* **30**: 1126–1135. <http://dx.doi.org/10.1021/jm00390a003>.
- Høie MH, Kiehl EN, Petersen B, Nielsen M, Winther O, Nielsen H, Hallgren J, Marcatili P. 2022. NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning. *Nucleic Acids Res* **gkac439**.
- Horton P, Park K-J, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K. 2007. WoLF PSORT: protein localization predictor. *Nucleic Acids Res* **35**: W585-587.
- Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**: 583–589. <https://www.nature.com/articles/s41586-021-03819-2>.
- Karpenahalli MR, Lupas AN, Söding J. 2007. TPRpred: a tool for prediction of TPR-, PPR- and SEL1-like repeats from protein sequences. *BMC Bioinformatics* **8**: 1–8. <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-2>.
- Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**: 772–780. <https://doi.org/10.1093/molbev/mst010>.
- Kleinknecht L, Wang F, Stübe R, Philippar K, Nickelsen J, Bohne A-V. 2014. RAP, the Sole Octotricopeptide Repeat Protein in Arabidopsis, Is Required for Chloroplast 16S rRNA Maturation. *Plant Cell* **26**: 777–787. <http://www.plantcell.org/content/26/2/777>.
- Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**: 567–580.

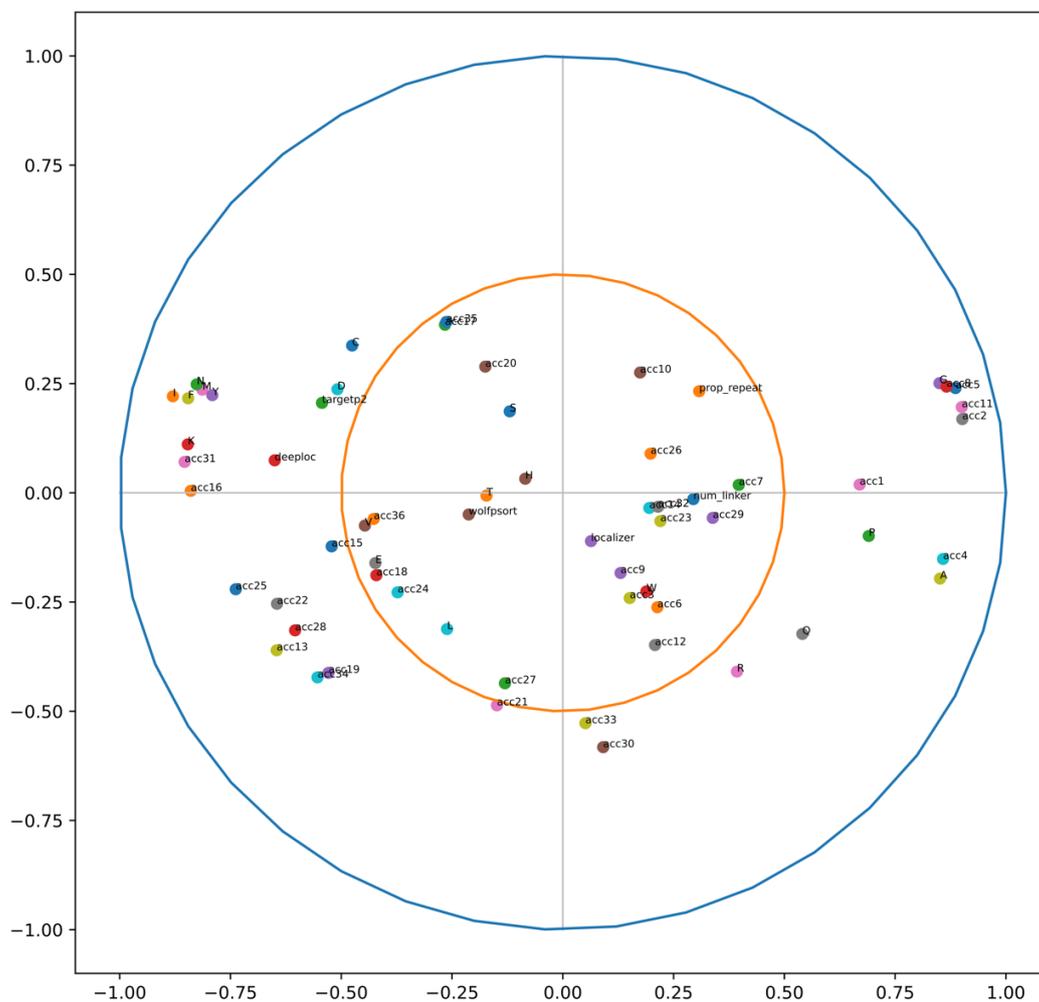
- Kück U, Schmitt O. 2021. The Chloroplast Trans-Splicing RNA–Protein Supercomplex from the Green Alga *Chlamydomonas reinhardtii*. *Cells* **10**: 290. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7912774/>.
- Lee K, Park SJ, Colas des Francs-Small C, Whitby M, Small I, Kang H. 2019. The coordinated action of PPR4 and EMB2654 on each intron half mediates trans-splicing of rps12 transcripts in plant chloroplasts. *The Plant Journal* **100**: 1193–1207. <https://onlinelibrary.wiley.com/doi/abs/10.1111/tpj.14509>.
- Lefebvre-Legendre L, Choquet Y, Kuras R, Loubéry S, Douchi D, Goldschmidt-Clermont M. 2015. A Nucleus-Encoded Chloroplast Protein Regulated by Iron Availability Governs Expression of the Photosystem I Subunit PsaA in *Chlamydomonas reinhardtii*. *Plant Physiology* **167**: 1527. <https://www.ncbi.nlm.nih.gov/insb.bib.cnrs.fr/pmc/articles/PMC4378161/>.
- Lefebvre-Legendre L, Reifschneider O, Kollipara L, Sickmann A, Wolters D, Kück U, Goldschmidt-Clermont M. 2016. A pioneer protein is part of a large complex involved in trans-splicing of a group II intron in the chloroplast of *Chlamydomonas reinhardtii*. *Plant J* **85**: 57–69.
- Leliaert F, Smith DR, Moreau H, Herron MD, Verbruggen H, Delwiche CF, Clerck OD. 2012. Phylogeny and Molecular Evolution of the Green Algae. *Critical Reviews in Plant Sciences* **31**: 1–46. <http://dx.doi.org/10.1080/07352689.2011.615705>.
- Lemieux C, Otis C, Turmel M. 2007. A clade uniting the green algae *Mesostigma viride* and *Chlorokybus atmophyticus* represents the deepest branch of the Streptophyta in chloroplast genome-based phylogenies. *BMC Biol* **5**: 1–17. <https://bmcbiol.biomedcentral.com/articles/10.1186/1741-7007-5-2>.
- Lipinski KA, Puchta O, Surendranath V, Kudla M, Golik P. 2011. Revisiting the Yeast PPR Proteins—Application of an Iterative Hidden Markov Model Algorithm Reveals New Members of the Rapidly Evolving Family. *Molecular Biology and Evolution* **28**: 2935–2948. <https://doi.org/10.1093/molbev/msr120>.
- Loiselay C, Gumpel NJ, Girard-Bascou J, Watson AT, Purton S, Wollman F-A, Choquet Y. 2008. Molecular identification and function of cis- and trans-acting determinants for petA transcript stability in *Chlamydomonas reinhardtii* chloroplasts. *Mol Cell Biol* **28**: 5529–5542.
- Lurin C, Andrés C, Aubourg S, Bellaoui M, Bitton F, Bruyère C, Caboche M, Debast C, Gualberto J, Hoffmann B, et al. 2004. Genome-Wide Analysis of Arabidopsis Pentatricopeptide Repeat Proteins Reveals Their Essential Role in Organelle Biogenesis. *The Plant Cell* **16**: 2089–2103. <http://www.plantcell.org/content/16/8/2089>.

- Macedo-Osorio KS, Martínez-Antonio A, Badillo-Corona JA. 2021. Pas de Trois: An Overview of Penta-, Tetra-, and Octo-Tricopeptide Repeat Proteins From *Chlamydomonas reinhardtii* and Their Role in Chloroplast Gene Expression. *Front Plant Sci* **12**: 775366. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8635915/>.
- Manna S. 2015. An overview of pentatricopeptide repeat proteins and their applications. *Biochimie* **113**: 93–99. <http://www.sciencedirect.com/science/article/pii/S030090841500108X>.
- Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution* **38**: 4647–4654. <https://doi.org/10.1093/molbev/msab199>.
- Marx C, Wünsch C, Kück U. 2015. The octatricopeptide repeat (OPR) protein Raa8 is required for chloroplast trans-splicing. *Eukaryotic Cell* **15**: EC.00096-15. <http://ec.asm.org/content/early/2015/07/21/EC.00096-15>.
- Meierhoff K, Felder S, Nakamura T, Bechtold N, Schuster G. 2003. HCF152, an Arabidopsis RNA Binding Pentatricopeptide Repeat Protein Involved in the Processing of Chloroplast psbB-psbT-psbH-petB-petD RNAs. *The Plant Cell* **15**: 1480–1495. <https://doi.org/10.1105/tpc.010397>.
- Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucl Acids Res* **41**: e121–e121. <http://nar.oxfordjournals.org/content/41/12/e121>.
- Nakamura Y, Sasaki N, Kobayashi M, Ojima N, Yasuike M, Shigenobu Y, Satomi M, Fukuma Y, Shiwaku K, Tsujimoto A, et al. 2013. The First Symbiont-Free Genome Sequence of Marine Red Alga, Susabi-nori (*Pyropia yezoensis*). *PLOS ONE* **8**: e57122. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0057122>.
- One Thousand Plant Transcriptomes Initiative. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**: 679–685. <http://www.nature.com/articles/s41586-019-1693-2>.
- Preker PJ, Keller W. 1998. The HAT helix, a repetitive motif implicated in RNA processing. *Trends in Biochemical Sciences* **23**: 15–16. [https://www.cell.com/trends/biochemical-sciences/abstract/S0968-0004\(97\)01156-0](https://www.cell.com/trends/biochemical-sciences/abstract/S0968-0004(97)01156-0).
- Pfalz J, Bayraktar OA, Prikryl J, Barkan A. 2009. Site-specific binding of a PPR protein defines and stabilizes 5' and 3' mRNA termini in chloroplasts. *The EMBO Journal* **28**: 2042–2052. <http://www.embopress.org/doi/full/10.1038/emboj.2009.121>.

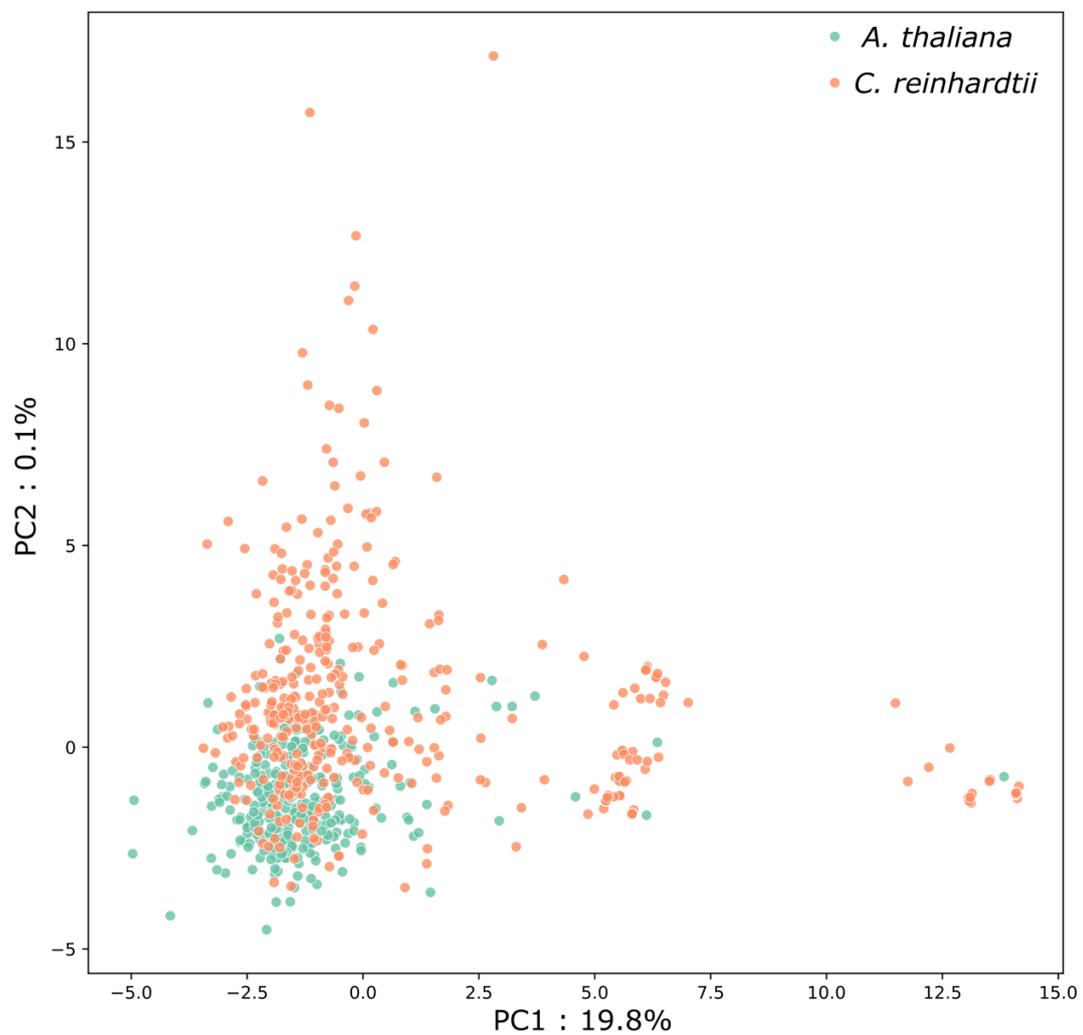
- Qu K, Wei L, Yu J, Wang C. 2019. Identifying Plant Pentatricopeptide Repeat Coding Gene/Protein Using Mixed Feature Extraction Methods. *Front Plant Sci* **9**. <https://www.frontiersin.org/articles/10.3389/fpls.2018.01961/full>.
- Rahire M, Laroche F, Cerutti L, Rochaix J-D. 2012. Identification of an OPR protein involved in the translation initiation of the PsaB subunit of photosystem I. *The Plant Journal* **72**: 652–661. <http://onlinelibrary.wiley.com.gate1.inist.fr/doi/10.1111/j.1365-3113.2012.05111.x/abstract>.
- Riggleman B, Wieschaus E, Schedl P. 1989. Molecular analysis of the armadillo locus: uniformly distributed transcripts and a protein with novel internal repeats are associated with a *Drosophila* segment polarity gene. *Genes Dev* **3**: 96–113.
- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**: 2498–2504.
- Sibbald SJ, Archibald JM. 2020. Genomic insights into plastid evolution. *Genome Biol Evol* **12**: 978–990. <https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evaa096/5836826>.
- Sikorski RS, Boguski MS, Goebel M, Hieter P. 1990. A repeating amino acid motif in CDC23 defines a family of proteins and a new relationship among genes required for mitosis and RNA synthesis. *Cell* **60**: 307–317.
- Small ID, Peeters N. 2000. The PPR motif – a TPR-related motif prevalent in plant organellar proteins. *Trends in Biochemical Sciences* **25**: 45–47. [http://www.cell.com/trends/biochemical-sciences/abstract/S0968-0004\(99\)01520-0](http://www.cell.com/trends/biochemical-sciences/abstract/S0968-0004(99)01520-0).
- Sormanni P, Camilloni C, Fariselli P, Vendruscolo M. 2015. The s2D Method: Simultaneous Sequence-Based Prediction of the Statistical Populations of Ordered and Disordered Regions in Proteins. *Journal of Molecular Biology* **427**: 982–996. <http://www.sciencedirect.com/science/article/pii/S002228361400641X>.
- Spassov DS, Jurecic R. 2003. The PUF family of RNA-binding proteins: does evolutionarily conserved structure equal conserved function? *IUBMB Life* **55**: 359–366.
- Sperschneider J, Catanzariti A-M, DeBoer K, Petre B, Gardiner DM, Singh KB, Dodds PN, Taylor JM. 2017. LOCALIZER: subcellular localization prediction of both plant and effector proteins in the plant cell. *Sci Rep* **7**: 44598. <https://www.nature.com/articles/srep44598>.
- Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**: 1026–1028. <https://www.nature.com/articles/nbt.3988>.

- Subrahmanian N, Castonguay AD, Remacle C, Hamel PP. 2020. Assembly of Mitochondrial Complex I Requires the Low-Complexity Protein AMC1 in *Chlamydomonas reinhardtii*. *Genetics* **214**: 895–911. <https://doi.org/10.1534/genetics.120.303029>.
- Tardif M, Atteia A, Specht M, Cogne G, Rolland N, Brugière S, Hippler M, Ferro M, Bruley C, Peltier G, et al. 2012. PredAlgo: a new subcellular localization prediction tool dedicated to green algae. *Mol Biol Evol* **29**: 3625–3639.
- The Angiosperm Phylogeny Group, Chase MW, Christenhusz MJM, Fay MF, Byng JW, Judd WS, Soltis DE, Mabberley DJ, Sennikov AN, Soltis PS, et al. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society* **181**: 1–20. <https://doi.org/10.1111/boj.12385>.
- Tourasse NJ, Choquet Y, Vallon O. 2013. PPR proteins of green algae. *RNA Biology* **10**: 1526–1542. <http://www-tandfonline-com.insb.bib.cnrs.fr/doi/full/10.4161/rna.26127>.
- Viola S, Cavaiuolo M, Drapier D, Eberhard S, Vallon O, Wollman F-A, Choquet Y. 2019. MDA1, a nucleus-encoded factor involved in the stabilization and processing of the atpA transcript in the chloroplast of *Chlamydomonas*. *The Plant Journal* **98**: 1033–1047. <http://onlinelibrary.wiley.com/doi/abs/10.1111/tpj.14300>.
- Wang F, Johnson X, Cavaiuolo M, Bohne A-V, Nickelsen J, Vallon O. 2015. Two *Chlamydomonas* OPR proteins stabilize chloroplast mRNAs encoding small subunits of photosystem II and cytochrome b6 f. *Plant J* **82**: 861–873.
- Williams PM, Barkan A. 2003. A chloroplast-localized PPR protein required for plastid ribosome accumulation. *The Plant Journal* **36**: 675–686. <https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-313X.2003.01915.x>.
- Woodson JD, Chory J. 2008. Coordination of gene expression between organellar and nuclear genomes. *Nature Reviews Genetics* **9**: 383. <https://www-nature-com.insb.bib.cnrs.fr/articles/nrg2348>.
- Yamazaki H, Tasaka M, Shikanai T. 2004. PPR motifs of the nucleus-encoded factor, PGR3, function in the selective and distinct steps of chloroplast gene expression in Arabidopsis. *The Plant Journal* **38**: 152–163. <http://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-313X.2004.02035.x>.
- Zhang H, Scheirer DC, Fowle WH, Goodman HM. 1992. Expression of antisense or sense RNA of an ankyrin repeat-containing gene blocks chloroplast differentiation in Arabidopsis. *The Plant Cell* **4**: 1575–1588. <https://doi.org/10.1105/tpc.4.12.1575>.

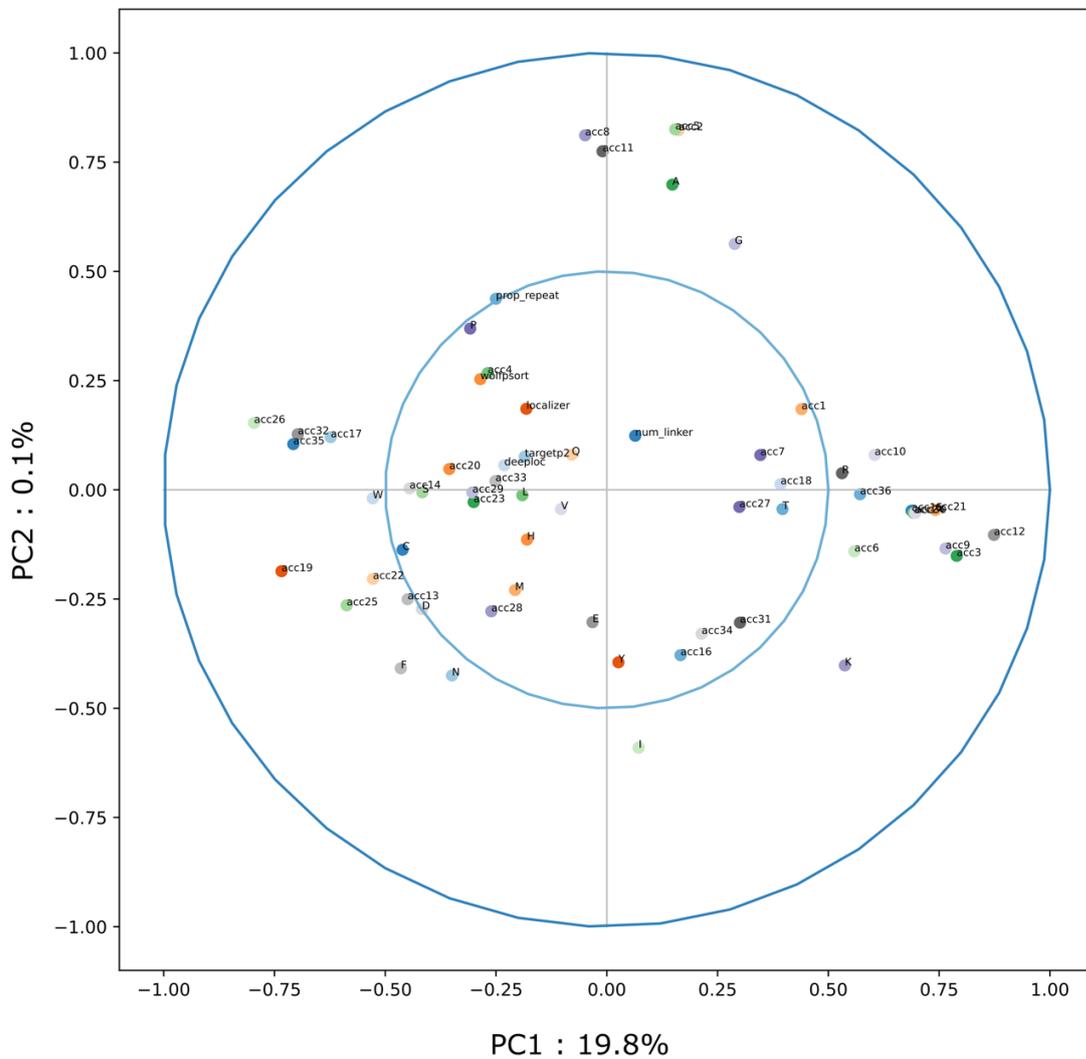
Zhao X, Wang H, Li H, Wu Y, Wang G. 2021. Identifying Plant Pentatricopeptide Repeat Proteins Using a Variable Selection Method. *Frontiers in Plant Science* **12**. <https://www.frontiersin.org/articles/10.3389/fpls.2021.506681>



**Figure S1A.** Correlation circles for the PCA presented in Figure 1. The frequencies of Isoleucine (I), Phenylalanine (F), Methionine (M), Asparagine (N), Tyrosine (Y) and Glycine (G) and Lysine (K) residues are the main contributors to PC1, together with the ACC terms (acc) 31, 16, 2, 5 and 11. Only the ACC terms 33 and 30 contributes for more than 0.5 into PC2.



**Figure S1B.** Principal component analysis of proteins from the negative set used to train DT and RF, described by properties used by the RF procedure. Proteins from *C. reinhardtii* are colored orange and those from *A. thaliana* in green. Explained variance is given for PC1 and PC2.



**Figure S1C.** Correlation circles for the PCA presented in Figure S1B. The main contributors to PC1, differs from the one observed for the PCA in Figure 3 (Figure S1A).

## List of Tables and Figures

Figure 1: Diagram of three developed protocols. ISS, left; DT, middle and RF, right.

Figure 2. Principal component analysis of *pto* candidates in *C. reinhardtii* and *A. thaliana*.

Figure 3. Clusters of *pto* candidates found in *A. thaliana* and *C. reinhardtii*.

Figure 4. Tandems of *pto* candidates on chromosome 10 of *C. reinhardtii*

Figure 5: Mean numbers of detected *pto* candidates detected by IPB (with OPR and PPR motifs) and new *pto* OTAF candidates detected by DT across proteomes of 48 Archaeplastida species.

Table 1: OPR and PPR motifs used to build the initial profiles for the ISS procedure

Table 2. Number of clusters obtained after 9 and 2 iterations of IPB for OPR and PPR.

Table 3. 11 *pto* candidates retrieved by IPB with the OPR motif. The last column indicates the domain found by the automatic annotation at Uniprot, but not retrieved by InterProScan.

Table 4. 10 *pto* candidates retrieved only by DT and RF in *A. thaliana*.

Table 5. 39 *pto* candidates retrieved only by DT and RF in *C. reinhardtii*.

Table S1: List and sources of the 48 proteomes studied, with link for download

Table S2: BUSCO quality assessment

Table S3: DT parameterization

Table S4: List of all candidates found by IBP and DT in 48 proteomes, and by RF in *A. thaliana* and *C. reinhardtii*.

Table S5: Clusters of *pto* candidates

Table Supp\_Zscale: correspondence between ACC numbers and composition

Figure S1A. Correlation circles for the PCA presented in Figure 1

Figure S1B. Principal component analysis of proteins from the negative set used to train DT and RF, described by properties used by the RF procedure

Figure S1C. Correlation circles for the PCA presented in Figure S1B

Supplementary Tables are available at:

<https://dropsu.sorbonne-universite.fr/s/NHMZR5Y5M5c9Xf2>



## 3 Discussion

### 3.1 Améliorer les capacités de détection de notre approche

L'analyse des résultats issus de la combinaison des approches IPB, DT et RF montre leur complémentarité notamment grâce au clustering de toutes les protéines candidates identifiées chez les espèces modèles *Chlamydomonas reinhardtii* et *Arabidopsis thaliana*. Les protéines PPR de *Arabidopsis thaliana* sont identifiées à plus de 95% grâce à la combinaison des trois méthodes et de nouvelles familles de protéines à solénoïde alpha ont également été mises en lumière.

Nous avons ensuite cherché à améliorer notre approche en testant de nouvelles propriétés pour les méthodes DT et RF. L'hypothèse de départ de ce travail, réalisé par Alexis Astatourian, durant son stage de M2 en 2021 que j'ai co-encadré, était que la structure ouverte de la protéine à solénoïde alpha permettant la fixation de l'ARNsb est une surface hydrophile car en contact avec le solvant aqueux de la cellule (ASTATOURIAN 2021). Le taux d'hydrophobicité le long de la séquence de la protéine serait donc plus faible au niveau du domaine en solénoïde alpha que dans le reste de la protéine, amenant à un profil type d'hydrophobicité tel que schématisé en (cf. figure 29).

Pour produire des figures comparables à celle illustrant son hypothèse, nous avons calculé le taux d'hydrophobicité le long de la protéine en utilisant une fenêtre glissante de 19 acides aminés (la moitié d'un motif OPR) et placé les linkers de paires d'hélices alpha détectés par le logiciel ard2 (FOURNIER et al. 2013). L'échelle d'hydrophobicité utilisée pour calculer le taux d'hydrophobicité de la fenêtre glissante provient de FAUCHÈRE et PLISKA 1983. Après essais sur les jeux de données précédemment décrits, nous nous sommes rendus à l'évidence que cette hypothèse ne donnait pas de résultats satisfaisants. Un exemple typique de ce que nous avons observé est présenté en figure 30 et ne correspond pas au profil type attendu présenté en 29. Nous n'avons donc pas ajouté de propriété basée sur le taux d'hydrophobicité à nos méthodes DT et RF.

La méthode RF nous a permis de montrer que dans les PPR, certaines hélices de protéines à solénoïde alpha ont une face hydrophile et une face hydrophobe (cf. figure 31). Ces hélices sont appelées amphipathiques, et l'utilisation des termes ACC (Auto-Cross Correlation) entre les descripteurs Z-scales des acides aminés le long de la protéine permet de les caractériser (GARRIDO et al. 2020). On a pu constater qu'il s'agit en grande majorité des hélices à l'intérieur du sillon des protéines à solénoïde alpha : la face hydrophobe est tournée vers l'autre hélice de la paire et la face hydrophile est tournée vers le sillon où se lie l'ARNm. Ce

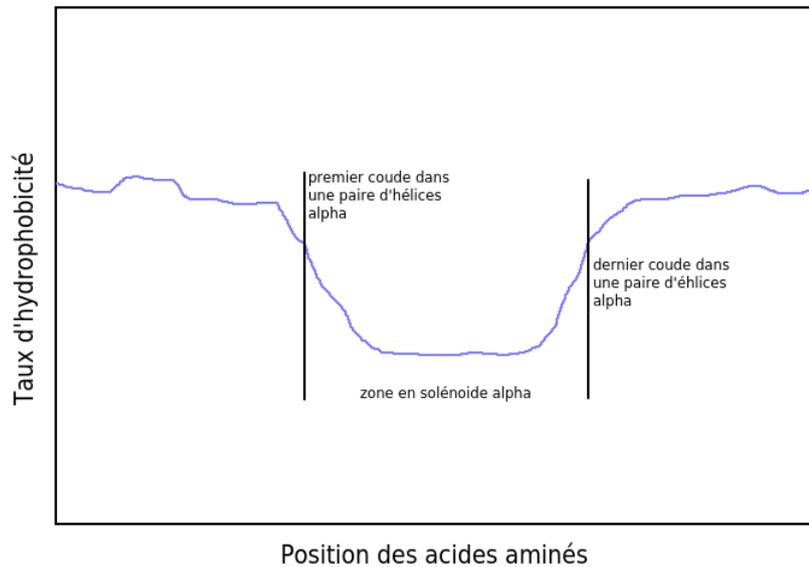


FIGURE 29 – Schéma du taux d'hydrophobicité le long de la séquence protéique illustrant l'hypothèse de travail de début de stage d'Alexis Astatourian (ASTATOURIAN 2021).

constat, s'il est montré que cette propriété est aussi présente chez les OPR et d'autres types de protéines à solénoïdes alpha, permettrait donc d'ajouter une nouvelle propriété pour affiner nos résultats.

### 3.2 Appliquer nos méthodes à d'autres espèces photosynthétiques : le cas des diatomées

Bien qu'il soit nécessaire d'affiner les méthodes DT et RF, nous avons montré les capacités de notre approche à identifier des protéines alpha solénoïdes au sein des Archaeplastida. L'objectif suivant est d'explorer la diversité des protéines à solénoïde alpha au sein d'autres groupes et espèces eucaryotes photosynthétiques. Nous avons commencé à tester et à appliquer certaines de nos approches sur un petit nombre d'espèces en dehors des Archaeplastida.

Chez les diatomées (ou Bacillariophytes), algues photosynthétiques unicellulaires (cf. illustration 12) provenant de l'endosymbiose secondaire d'une Rhodophyte par un eucaryote unicellulaire (cf. 6.3, 13), les mécanismes de régulation de l'expression du génome du plaste sont peu connus et comme plusieurs

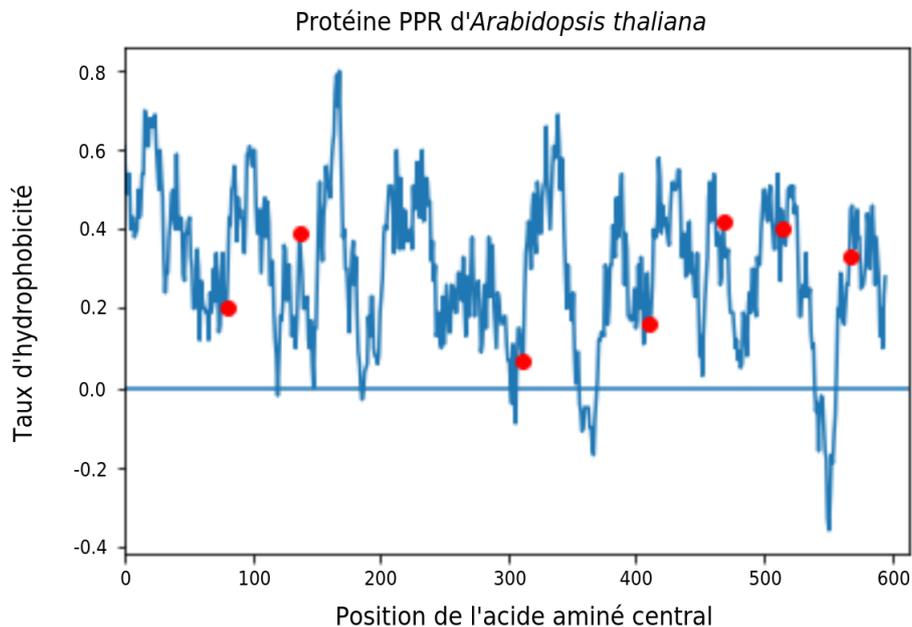


FIGURE 30 – Taux d'hydrophobicité le long de la séquence de la protéine AT1G28020.1 qui est une protéine PPR connue d'*Arabidopsis thaliana*. Le taux est calculé sur une fenêtre glissante de 19 acides aminés via l'échelle d'hydrophobicité des acides aminés proposée dans FAUCHÈRE et PLISKA 1983. Les points rouges représentent les linkers de paires d'hélices alpha détectés par le logiciel ard2 (FOURNIER et al. 2013 ; ASTATOURIAN 2021).

espèces de diatomées sont étudiées dans notre laboratoire, nous avons choisi de commencer par ces espèces.

Lors du stage de master 2 d'Alexis Astatourian nous avons choisi d'appliquer IPB et DT aux 8 protéomes de diatomées disponibles sur les banques de données et aux 57 protéomes issus des données transcriptomiques du Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) (JOHNSON, ALEXANDER et BROWN 2018). La qualité des protéomes issus des données du MMETSP a été évaluée en comparaison de la qualité des autres protéomes de diatomées et d'Archaeplastida via la méthode BUSCO. Cette analyse a montré que les protéomes sont incomplets mais restent de qualité satisfaisante pour commencer notre analyse (86% des marqueurs BUSCO sont retrouvés dans les huit protéomes complets et 44% dans les protéomes issus du MMETSP).

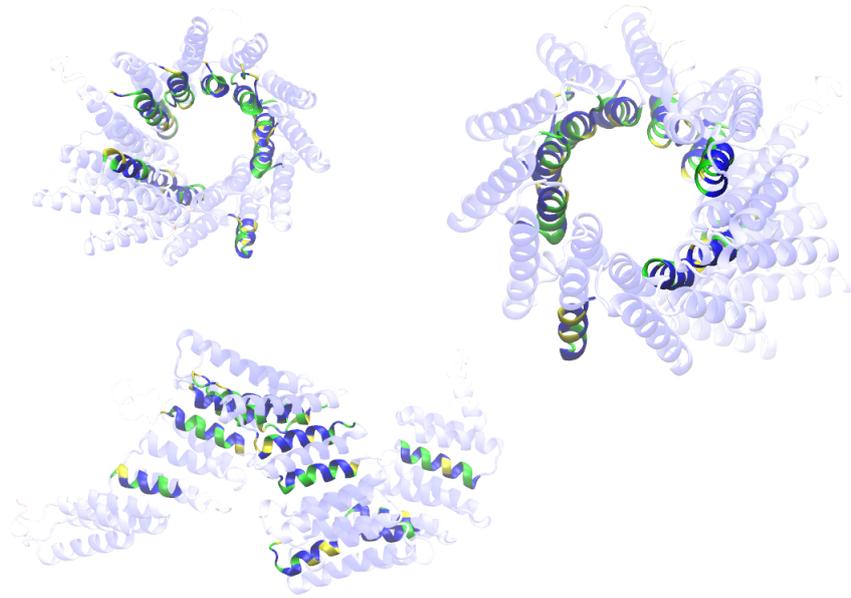


FIGURE 31 – Trois vues d’une même protéine PPR sous différents angles. En bleu les acides aminés apolaires, en vert les acides aminés polaires, les autres acides aminés gris ne font pas partie d’hélice amphiphile. Figure réalisée sur une protéine PPR de la diatomée *Phaeodactylum tricornerutum* avec le logiciel de visualisation VMD.

Lors de son stage, Alexis Astatourian n’a effectué qu’une seule itération de la méthode IPB sur les diatomées et les résultats sont donc préliminaires. Les motifs qu’il a utilisé initialement sont ceux issus de mon travail avec la méthode IPB sur les espèces d’Archaeplastida. Il a pu observer que certaines espèces de diatomées contiennent plus de 30 protéines OPR candidates, tandis que d’autres n’en ont aucune (cf. figure 32). En effet, 26 des protéomes issus du MMETSP ne contiennent aucune protéine OPR candidate mais 3 en possèdent plus de 30 et ceci montre à nouveau la grande variabilité du nombre de protéines OPR.

Pour ce qui est des protéines PPR, après également une itération de la méthode IPB dont les motifs était aussi issus de mon travail sur les PPR d’Archaeplastida, Alexis Astatourian a observé un nombre de protéines PPR candidates bien supérieur à celui des protéines OPR candidates. En effet, même chez les espèces pour lesquelles il n’avait pas trouvé de protéine OPR candidate, il a montré la présence de plusieurs dizaines de protéines PPR (en quantité toujours variable selon les organismes). Les résultats ont montré 79 protéines PPR candidates contre 5 protéines OPR en moyenne dans les protéomes de diatomées étudiés (cf. figure

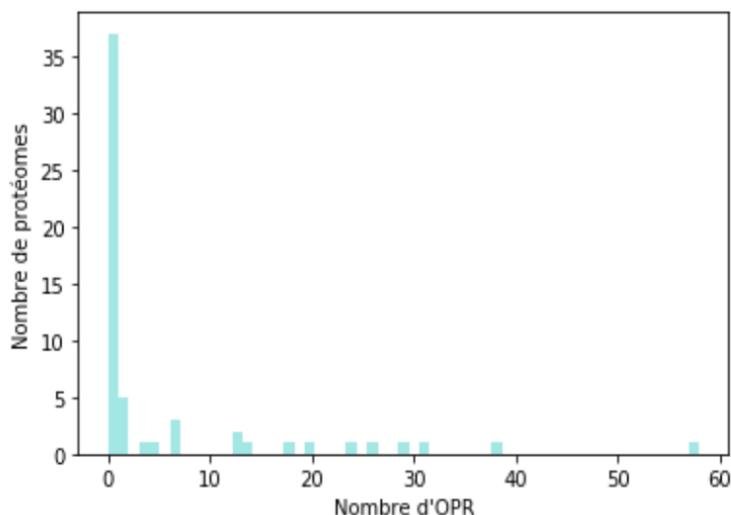


FIGURE 32 – Distribution du nombre de protéines OPR candidates dans 57 protéomes issus du MMETSP (ASTATOURIAN 2021).

33). Le biais de distribution des protéines OPR et PPR s’approche donc de celui observé chez les plantes terrestres mais il semble moins exacerbé (ASTATOURIAN 2021).

Durant la fin de son stage, Alexis Astatourian a appliqué la méthode utilisant l’arbre de décision (DT) sur les protéomes de diatomées. Par manque de temps, il a cependant tronqué l’étape filtrant sur la proportion d’acides aminés impliqués dans une hélice alpha car le logiciel à utiliser pour cette étape est particulièrement coûteux en temps de calcul. Il a ainsi mis en évidence en moyenne 101 protéines par protéome (cf. figure 34) et jusqu’à plusieurs centaines de protéines à solénoïde alpha candidates (389 chez la diatomée *Fistulifera solaris* qui ne fait pas partie des protéomes issus du MMETSP).

Il sera nécessaire de finaliser ces analyses par plusieurs itérations de la méthode IPB et d’y appliquer la méthode RF pour avoir une vue globale du paysage des OTAF candidates chez les diatomées.

Lors de son stage, Rebecca Goulancourt a pu appliquer la méthode RF au protéome de la diatomée *Phaeodactylum tricornutum*. Une différence notable avec les résultats de cette méthode sur *Arabidopsis thaliana* et *Chlamydomonas reinhardtii* présentés dans l’article est qu’il n’était pas possible d’utiliser les mêmes

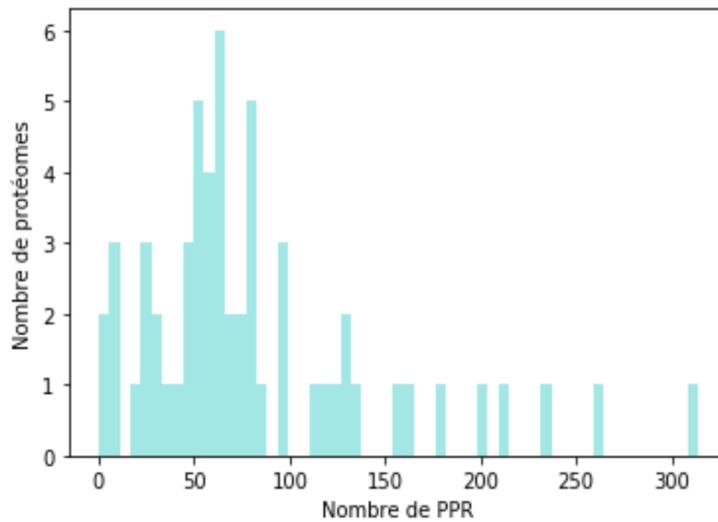


FIGURE 33 – Distribution du nombre de protéines PPR candidates dans 57 protéomes issus du MMETSP (ASTATOURIAN 2021).

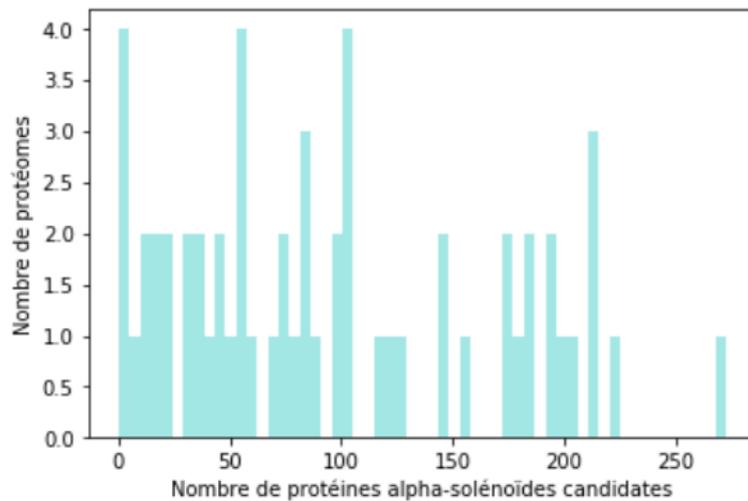


FIGURE 34 – Distribution du nombre de protéines à solénoïde alpha candidates dans 57 protéomes issus du MMETSP (ASTATOURIAN 2021).

logiciels de prédiction de l'adressage car ils ne sont pas adaptés aux diatomées (les peptides d'adressage des protéines de diatomées sont différents de ceux des Archaeplastida). Nous avons donc utilisé Hectar, qui est adapté aux diatomées

pour prédire l'adressage au chloroplaste et Deeploc pour prédire l'adressage aux mitochondries.

Les résultats obtenus par Rebecca Goulancourt sont en accord avec ceux obtenus précédemment par Alexis Astatourian : 107 protéines à solénoïdes alpha candidates sont identifiées dont 26 sont prédites comme adressées aux organites chez *Phaeodactylum tricornutum*. Alexis Astatourian trouvait 3 protéines OPR candidates et 55 protéines PPR candidates mais il n'avait pas fait de filtrage sur la prédiction de l'adressage aux organites. Il est maintenant prévu d'explorer le reste des protéomes de diatomées grâce à la méthode utilisant l'apprentissage automatique (RF).

## 4 Conclusion et perspectives

Nous avons montré la capacité de nos méthodes à identifier différentes familles de protéines à solénoïde alpha prédites comme adressées aux organites. On peut noter que la méthode DT est très spécifique et que sa sensibilité semble faible. Il s'agit du principal problème de cette méthode et il semble nécessaire de chercher à l'améliorer. D'autres logiciels sont capables de fournir une prédiction du repliement de la protéine en hélices et feuillets. Il est donc possible de remplacer certains logiciels par d'autres mais on peut aussi choisir de relâcher certains des critères éliminant trop de nos témoins positifs et ajouter des filtres sur des propriétés qui ne sont pas encore utilisées comme l'amphipathicité des hélices alpha à l'intérieur du sillon où se lie l'ARNsb.

Nous évoquons dans l'article que la méthode RF n'est pas assez spécifique car elle identifie beaucoup de protéines à solénoïde alpha candidates adressées à la mitochondrie ou au chloroplaste pour nos deux espèces modèles. Il y a ainsi bien trop de protéines OTAF candidates par rapport au nombre de gènes codés dans les génomes chloroplastiques et mitochondriaux (il n'y a que quelques dizaines de protéines dont les gènes sont codés dans les organites). Il semble donc très peu probable qu'autant de protéines jouent un rôle d'OTAF.

Actuellement nous utilisons l'algorithme du Random Forest basé sur le principe du consensus des arbres décisionnels mais il est également possible d'explorer d'autres algorithmes d'apprentissage supervisé comme les algorithmes de machines à vecteurs de support (ou SVM pour "Support-Vector Machine") qui permettent également de classer des objets dans différentes catégories.

Pour finir, l'ajout d'une étape de filtrage des résultats provenant de la méthode RF avec des valeurs seuils utilisées dans la méthode DT pourrait permettre de combiner les deux approches et d'obtenir de meilleurs résultats du point de vue des statistiques de sensibilité et de spécificité.

Depuis une autre perspective, nous avons entamé des cultures pour commencer la validation expérimentale *in vivo* chez les micro-algues de certaines des protéines à solénoïde alpha candidates non annotées que nous avons mises en évidence. En collaboration avec Sandrine Bujaldon au sein de notre laboratoire, nous avons notamment commencé à cultiver des mutants de certaines OTAF candidates non annotées chez *Chlamydomonas reinhardtii* pour observer le phénotype des cellules et leur réaction à différents environnements. Les phénotypes obtenus pour l'instant quant à l'activité photosynthétique demeurent faibles.

Enfin, bien que nous soyons désormais capables de donner une estimation préliminaire de la distribution des protéines à solénoïde alpha au sein de deux lignées eucaryotes photosynthétiques provenant de deux événements d'endosymbioses distincts, il reste encore beaucoup de lignées issues d'autres événements d'endosymbioses à explorer (cf. 5) dans lesquelles aucune quantification de la distribution des protéines à solénoïde alpha régulatrices des génomes des organites n'est actuellement disponible. Les approches DT et RF seront particulièrement utiles pour caractériser les OTAF dans ces lignées.





Troisième partie  
Étude de la dynamique  
conformationnelle des protéines à  
"PentatrigoPeptide Repeat"

"L'homme sait enfin qu'il est seul dans l'immensité indifférente de l'univers d'où il a émergé par hasard. Non plus que son destin, son devoir n'est écrit nulle part."

*Le Hasard et la Nécessité*, Jacques Monod (1970)

"Les constituants universels que sont les nucléotides d'une part, les acides aminés de l'autre, sont l'équivalent logique d'un alphabet dans lequel serait écrite la structure, donc les fonctions associatives spécifiques des protéines."

*Le Hasard et la Nécessité*, Jacques Monod (1970)



# 1 Premier chapitre : présentation des simulations de dynamique moléculaire et du principe des analyses réalisées

## 1.1 Les simulations de dynamique moléculaire : quelques grands principes

Pour étudier l'évolution d'un système d'objets tout en contrôlant le milieu dans lequel se trouve ce système, on utilise des techniques de dynamique moléculaire. Il s'agit de simulations *in silico* durant lesquelles la position des objets composant le système et leurs interactions sont évaluées au cours du temps. Ce genre de techniques peut s'appliquer avec différents niveaux d'approximations à des systèmes aux dimensions de l'ordre de la dizaine d'années lumière comme des systèmes planétaires ou des galaxies pour mieux comprendre les interactions entre planètes, étoiles et autres corps, mais également à des systèmes microscopiques comme les cellules et leurs composants comme l'ADN ou les protéines. C'est dans ce dernier cas de figure que nous étudions ici les protéines à solénoïde alpha et leur ligand ARNsb.

Dans notre cas, les objets du système sont les atomes composant la protéine et son ligand ainsi que le milieu aqueux composé de molécules d'eau et d'ions sodium ( $Na^+$ ) et calcium ( $Cl^-$ ) dans lequel ils baignent. A chaque pas de temps défini, la position de chaque atome est calculée et les forces agissant sur chaque atome sont enregistrées. A la fin de la simulation on obtient ainsi une trajectoire pour chaque atome au cours du temps.

Dans ces simulations, le temps n'est pas continu car on divise le temps en petites portions appelées pas de temps (ou "timestep" en anglais). La durée d'un pas de temps est définie par l'utilisateur selon la précision avec laquelle il souhaite simuler les mouvements des particules dans le système. En général, le pas de temps  $\Delta t$  est de l'ordre de la femtoseconde (fs) et ce, afin de prendre en considération les vibrations atomiques les plus rapides pour que les simulations ne deviennent pas instables. Les coordonnées des particules sont calculées à chaque pas de temps en résolvant l'équation provenant de la deuxième loi de Newton, issue de la mécanique classique (ou newtonienne) qui décrit l'accélération d'une particule en fonction de sa masse et de la somme des différentes forces s'appliquant dessus :  $F = ma$ .

Pour un système de N particules, il y a 3N degrés de liberté indépendants et

pour chacun on peut écrire :

$$F_i = m_i a_i = m_i \ddot{x}_i$$

La force totale est le gradient de l'énergie potentielle de l'interaction  $U$ , qui est une fonction de  $X$ , l'ensemble des  $3N$  variables indépendantes.

$$F_i = -\frac{\partial U(X)}{\partial x_i}$$

En principe on peut décrire précisément  $x_i$  après un pas de temps  $\Delta t$  par une expansion de Taylor dans laquelle  $\dot{x}_i$  est la vitesse et  $\ddot{x}_i$  représente l'accélération :

$$x_i(t + \Delta t) = x_i(t) + \dot{x}_i(t)\Delta t + \frac{1}{2}\ddot{x}_i(t)\Delta t^2 + \dots$$

En pratique, la dynamique moléculaire utilise une approximation comme l'intégration de Verlet (VERLET 1967) (notamment utilisée par NAMD (PHILLIPS et al. 2020)) dans laquelle les accélérations et les vitesses suffisent à calculer les prochaines valeurs des variables à partir de l'étape actuelle et l'étape précédente.

Ainsi, dans les simulations de dynamique moléculaire on a besoin :

- des positions de départ des atomes du système. Ces données proviennent d'une structure initiale (une structure cristalline par exemple). Cette structure peut être obtenue par cristallographie aux rayons X (aussi appelée radiocristallographie), par résonance magnétique nucléaire (RMN) ou encore par cryo-microscopie électronique (cryo-ME), une technique basée sur la microscopie électronique permettant de déterminer la structure d'une protéine congelée dans une glace ne filtrant pas les électrons. Cette dernière technique existe depuis longtemps mais de récentes avancées ont permis d'atteindre des résolutions comparables à celle de la radiocristallographie qui se base sur l'irradiation aux rayons X de cristaux de la protéine dont on veut obtenir les informations structurales et sur l'analyse des spectres de diffraction qui en résultent. La limite de cette méthode est la difficulté d'obtenir des cristaux des protéines d'intérêt, certaines d'entre elles s'y prêtant mal. La RMN, quant à elle, est une technique qui ne nécessite pas l'obtention de cristaux de protéines car le travail se fait avec des protéines en solution. Cependant, la taille des protéines qu'il est possible d'analyser avec cette méthode est limitée.

- des vitesses de départ qui sont généralement générées aléatoirement en accord avec la température choisie du système.
- du potentiel d'interaction  $U$  qui est la somme de toutes les interactions agissant sur les atomes du système et qui peut être décrit par l'équation suivante :

$$U = \sum_i^{N_{liaisons}} \frac{k_i}{2} (l_i - l_{i,0})^2 + \sum_i^{N_{angles}} \frac{k_i}{2} (\theta_i - \theta_{i,0})^2 + \sum_i^{N_{diedres}} \frac{V_i}{2} (1 + \cos(n_i - \gamma_i)) + \sum_i^N \sum_j^N (4\epsilon_{ij} [(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^6] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}})$$

D'autres termes peuvent être ajoutés à cette équation dans les champs de force plus spécifiques par exemple mais la plupart des champs de force contiennent au moins ces quatre termes :

- Le premier terme modélise les interactions entre les paires d'atomes liés par des liaisons covalentes. Ce modèle est un potentiel harmonique qui implique une énergie potentielle grandissante en fonction de la différence de la longueur de la liaison  $l_i$  par rapport à un optimal de référence  $l_{i,0}$  (cf. figure 35).
- Le second terme de l'équation correspond à la différence de tous les angles de valence  $\theta_i$  de la molécule par rapport à une valeur de référence  $\theta_{i,0}$ , c'est à dire de tous les angles formés par trois atomes (2 atomes liés à un troisième mais pas entre eux (cf. figure 35).
- Le troisième terme caractérise comment l'énergie potentielle change en fonction de la rotation des liaisons covalentes en termes d'une fonction périodique avec  $n_i$  minimal selon le cas. Il s'agit d'un potentiel de torsion (cf. figure 35).
- Le quatrième terme correspond aux autres interactions non-liées, c'est-à-dire aux interactions non covalentes comme les interactions électrostatiques ou de van der Waals (cf. figure 35). Ce dernier terme concerne toutes les paires d'atomes  $ij$  du système (faisant partie de la molécule principale ou non) dont les paires avec le solvant.

Afin d'optimiser les temps de calculs, tous les atomes situés au delà d'un rayon défini par l'utilisateur ne sont pas pris en compte dans la somme des forces appliquées à l'atome duquel part ce rayon car la force qu'ils appliquent chacun sur

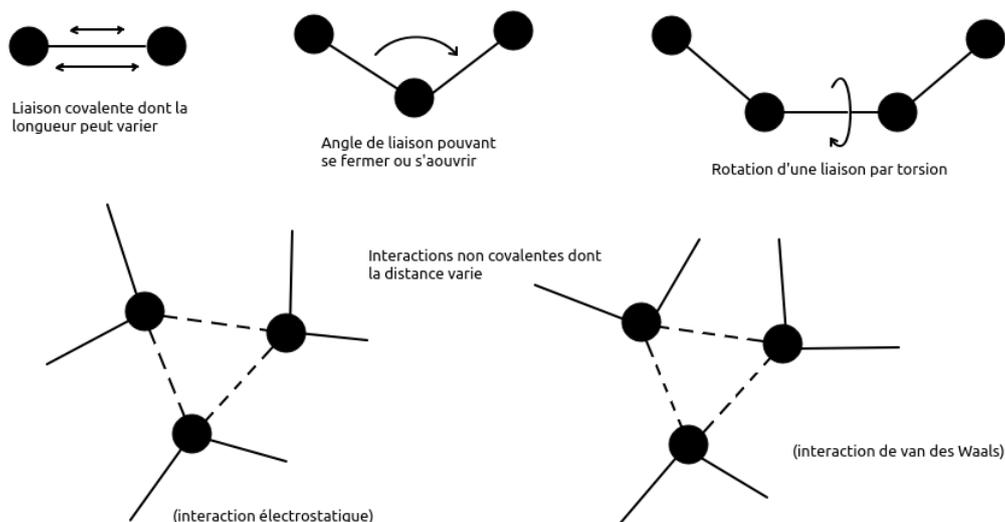


FIGURE 35 – Schéma simplifié illustrant les différents éléments pouvant varier au cours de la simulation qui sont retrouvés dans l'équation à quatre termes du champs de force (figure réalisée à partir d'un modèle issu de LEACH 2001).

l'atome dont on souhaite calculer la position est souvent très faible (cf. figures 36 et 37). Cependant, cette technique ne donne pas de résultats satisfaisants lorsque la force d'interactions diminue lentement comme c'est le cas des interactions électrostatiques par exemple. Dans le cadre des simulations dans des conditions périodiques décrites ci-dessous, nous utilisons le principe de PME (Particle Mesh Ewald) qui permet de lisser la courbe de l'énergie potentielle des liaisons électrostatiques plus rapidement vers le 0. Elle est basée sur la méthode de la sommation d'Ewald qui modélise les interactions électrostatiques comme la somme de deux termes, l'un modélisant la courte portée de l'interaction et l'autre la longue portée.

Le système de particules est en effet enfermé dans une boîte de simulation qui est répétée de façon strictement identique dans toutes les directions (cf. 38). Ainsi, lorsqu'une particule sort d'un côté de la boîte de simulation, elle passe dans la boîte d'à côté. Cela revient à faire entrer une nouvelle particule identique avec la même accélération que celle de la particule sortie, et ce, au même instant et à l'opposé du lieu de sortie. La boîte doit donc être suffisamment grande pour que le système ne puisse pas interagir fortement avec le système de la boîte d'à côté (lui-même en réalité).

La puissance que demande la simulation est très importante car il est néces-

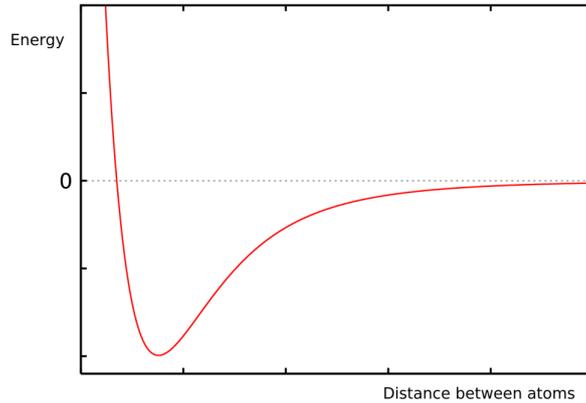


FIGURE 36 – Le potentiel de Lennard-Jones caractérise l'énergie potentielle de l'interaction entre deux atomes en fonction de leur proximité. Deux atomes sont à l'équilibre lorsqu'ils sont à une distance à laquelle l'énergie potentielle de leur interaction est minimale.

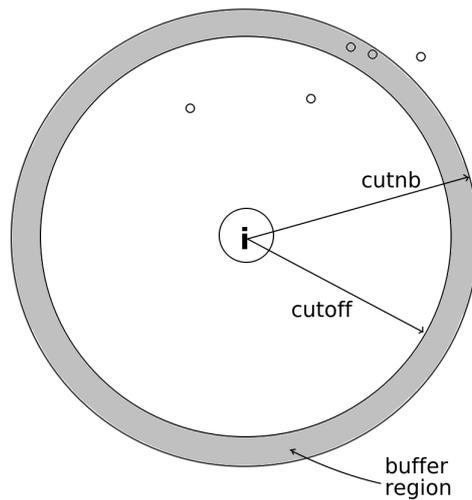


FIGURE 37 – Le système de seuil (ou "cutoff" en anglais) est utilisé pour optimiser les temps de calculs en ne prenant que les forces provenant des atomes à l'intérieur d'une sphère de rayon de la taille du seuil et ayant pour centre l'atome duquel on calcule l'accélération et la position. On schématise ici deux seuils : "cutoff" est la limite à laquelle les atomes situés plus loin ne sont plus utilisés pour calculer la somme des forces appliquées sur l'atome et cutnb est une zone tampon (ou "buffer" en anglais) qui permet de voir entrer ou sortir un atome de la zone "cutoff".

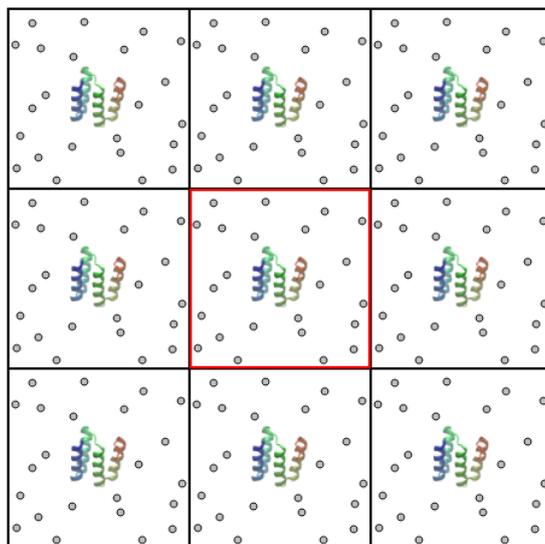


FIGURE 38 – Schéma simplifié du principe des boîtes de simulation strictement identiques entre lesquelles les particules peuvent circuler. Entourée en rouge, il s’agit de la boîte centrale que l’on observe lorsqu’on visualise la simulation par la suite.

saire d’évaluer la position de chaque atome à chaque pas de temps en fonction de tous ses voisins, et les temps de calculs donc sont considérables. En effet, il y a plusieurs dizaines voir centaines de milliers d’atomes présents dans le système (même si on ne simule en général qu’une seule protéine dans une même boîte de simulation) car le milieu dans lequel baigne la protéine est lui aussi simulé et tous les atomes en faisant partie doivent aussi voir leur position calculée à chaque pas de temps. Le pas de temps étant généralement de l’ordre de la femtoseconde, il est ainsi rare que les temps de simulation aillent au delà de l’échelle de la microseconde voir de la milliseconde même avec des ressources de calcul importantes.

Les simulations de dynamique moléculaire produisant notamment un fichier contenant la position de chaque atome à chaque pas de temps, il faut ensuite analyser les mouvements qui en découlent et on peut analyser les mouvements généraux de la protéine tout comme les détails des interactions atomiques par exemple au sein d’un groupe d’acides aminés proches dans l’espace.

## 1.2 Les analyses des résultats issus des simulations de dynamique moléculaire

### 1.2.1 Les distances entre acides aminés

L'une des analyses parmi les plus simples que l'on puisse imaginer est le calcul de l'évolution de la distance entre deux atomes au cours de la simulation. Il faut pour cela déjà avoir connaissance des positions des atomes et des acides aminés éventuellement impliqués dans des mouvements les rapprochant ou les éloignant l'un de l'autre.

Cette analyse couplée avec une mesure d'un angle peut notamment permettre d'observer la formation ou la disparition d'une liaison hydrogène entre deux acides aminés qui sont éloignés dans la séquence de la protéine mais proches lorsque celle-ci adopte sa structure 3D repliée. Une liaison hydrogène est une liaison dont la force est intermédiaire entre celle d'une liaison covalente et celle d'une force de van der Waals. Elle implique un atome d'hydrogène et un atome électronégatif (c'est-à-dire qui a la capacité d'attirer vers lui les électrons partagés dans une liaison) comme l'oxygène ou l'azote (cf. figure 39).

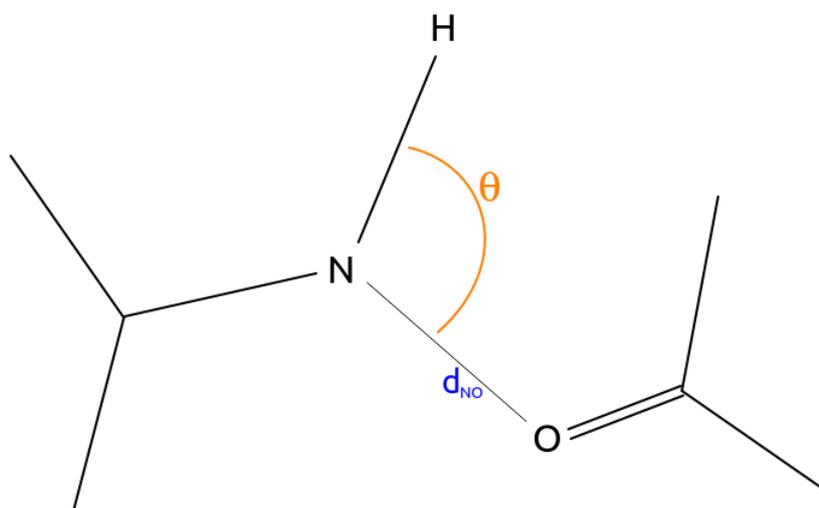


FIGURE 39 – Schéma d'une liaison hydrogène (ligne pointillée) entre un atome d'hydrogène (H) lié à un atome d'azote (N) par une liaison covalente (ligne pleine) et un atome d'oxygène (O).

### 1.2.2 La RMSD

La "Root Mean Square Deviation" (RMSD) ou la déviation de la racine de la moyenne des carrés permet de mesurer la différence moyenne entre deux jeux de données. La RMSD s'exprime dans l'unité des valeurs des jeux de données et plus la valeur est grande, plus la différence entre les données est importante. Dans le cas de la dynamique moléculaire, on se sert de cette mesure pour évaluer l'évolution de la structure de la protéine au cours du temps durant la simulation en comparant les positions des atomes d'intérêt (notamment ici les atomes  $C_\alpha$  de chaque acide aminé de la protéine) à une structure de référence au cours du temps.

Pour cela il faut d'abord calculer les différences entre les valeurs de référence ( $X_i$ ) et les valeurs que l'on souhaite évaluer ( $Y_i$ ) et appliquer la formule suivante :

$$RMSD = \sqrt{\frac{\sum_{i=1}^n (X_i - Y_i)^2}{n}}$$

En pratique on utilise la RMSD minimale, c'est-à-dire qu'on superpose la structure d'intérêt à une structure de référence (celle de départ par exemple) avant de calculer la RMSD. On cherche à produire le plus petit écart possible entre les deux structures grâce à une transformation  $T$  (rotation et translation) rigide de la structure d'intérêt  $Y'_i$  vers la structure de référence  $X$  comme suit :

$$Y_i = T_{\min}(Y'_i, X_i)$$

On peut ensuite tracer l'évolution de la valeur de la RMSD au cours du temps par rapport à la structure de la protéine sous forme cristalline ou bien essayer de caractériser différentes conformations de la structure de la protéine et les changements entre chacune.

#### Observer la dynamique de la protéine au cours du temps

Cette analyse est l'une des plus classiques pour commencer à observer et à comprendre les mouvements de la protéine durant la simulation, par exemple pour vérifier si la protéine a exploré des conformations proches de la conformation initiale ou bien si elle a été dépliée et n'a plus de réel sens biologique. Les graphiques obtenus ressemblent alors à celui présenté dans la figure 40 à gauche. On constate sur cette figure d'exemple que la protéine s'est rapidement éloigné

de sa conformation initiale pour y revenir mais s'est ensuite stabilisé autour de 1 Angström de différence avec la structure initiale puis a fait quelques excursions avant un changement important final à 3 Angström.

### Visualiser les changements de conformations

Avec les figures issues de cette analyse, on cherche à visualiser les changements de conformations de la protéine au cours de la simulation. Pour cela, on trace la RMSD à chaque instant  $t$  de la simulation contre tous les autres instants. On obtient ainsi une matrice de valeur de RMSD à deux dimensions sur laquelle on peut appliquer une échelle de couleurs. On peut ainsi graphiquement observer des blocs de couleurs sombres qui illustrent des périodes de la simulation durant lesquelles la protéine gardait une conformation similaire. On peut notamment voir un exemple de représentation sur la figure 40 à droite. Les couleurs sombres représentent les périodes durant lesquelles les structures sont proches tandis que les couleurs plus claires représentent les structures qui divergent. On peut remarquer notamment une période de la frame 500 à la frame 700 symbolisée par un carré sombre durant laquelle la structure change peu.

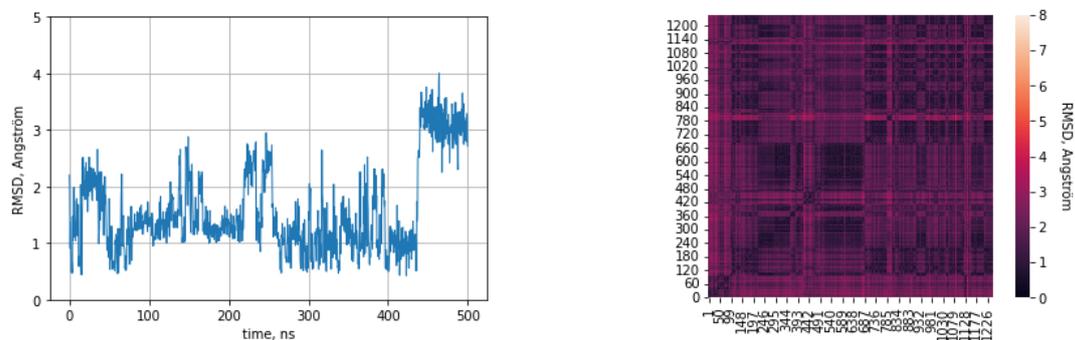


FIGURE 40 – Illustrations de deux représentations types de la RMSD que nous avons utilisé pour ce travail. A gauche : exemple d'une figure sur laquelle est tracée la RMSD en fonction du temps par rapport à la structure initiale. A droite : exemple d'une figure sur laquelle est tracée la RMSD de la structure à chaque pas de temps contre les structures à tous les autres pas de temps.

### 1.2.3 La fraction de contacts natifs

Cette mesure, souvent abrégée FNAT est complémentaire de la mesure de la RMSD en cela qu'il s'agit de mesurer le nombre de paires d'atomes pour lesquelles leur distance est inférieure à un seuil. Ce seuil est fixé de façon à ce que les paires d'atomes dont la distance est inférieure à ce seuil soient considérées comme ayant un contact. Ce seuil est fixé à 4,5Å pour les simulations que nous faisons.

La FNAT est ensuite calculée en comparant le nombre de contacts dans la structure d'intérêt à celui dans la structure initiale qui sert ici de référence comme pour le calcul de la RMSD. Plus la FNAT est proche de 1, plus la structure a gardé les contacts déjà présents dans la structure initiale.

### 1.2.4 La RMSF

La "Root Mean Square Fluctuation" (RMSF) est la RMSD de la position d'un atome donné autour de sa position moyenne sur le temps de la simulation. On obtient donc une valeur moyenne du mouvement de chaque acide aminé composant la protéine dans notre cas. Ceci permet de mieux saisir quelles sont les zones de la protéine qui sont les plus mobiles.

Par exemple, dans la figure 41 on observe que les deux acides aminés aux extrémités d'un motif PPR sont plus mobiles que les acides aminés à l'intérieur du motif. En effet, les acides aminés aux extrémités sont soumis à moins de contraintes et ont plus de liberté de mouvement au contraire des acides aminés au sein de la paire d'hélices composant le motif PPR.

### 1.2.5 La surface accessible au solvant

Dans le cas de notre étude, nous avons étudié la surface accessible au solvant (appelée ASA pour "Accessible Surface Area" ou encore SASA pour "Solvent-Accessible Surface Area" en anglais) pour mieux comprendre les impacts des mouvements de la protéine au cours du temps sur la taille et la disponibilité des zones de contacts avec une autre partie du système, par exemple ici l'ARNsb. Le calcul de cette surface équivaut à estimer la surface de la protéine sur laquelle une bille (généralement d'un diamètre équivalent à celui d'une molécule d'eau) pourrait rouler (cf. figure 42). C'est la surface parcourue par le centre de la bille que l'on appelle la surface accessible au solvant. Elle est mesurée en Angström carrés.

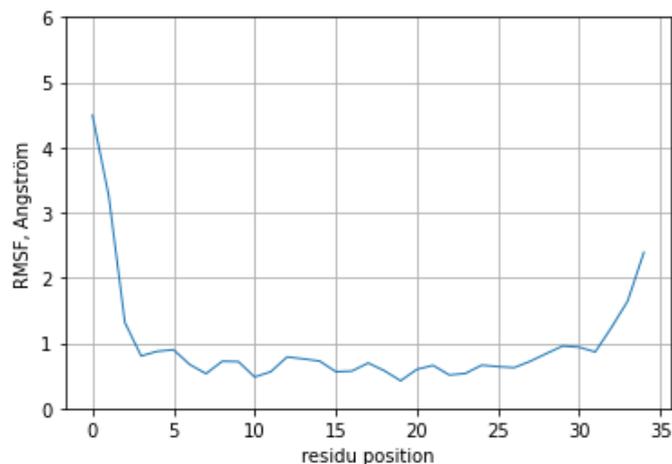


FIGURE 41 – Exemple de la RMSF dans un motif PPR.

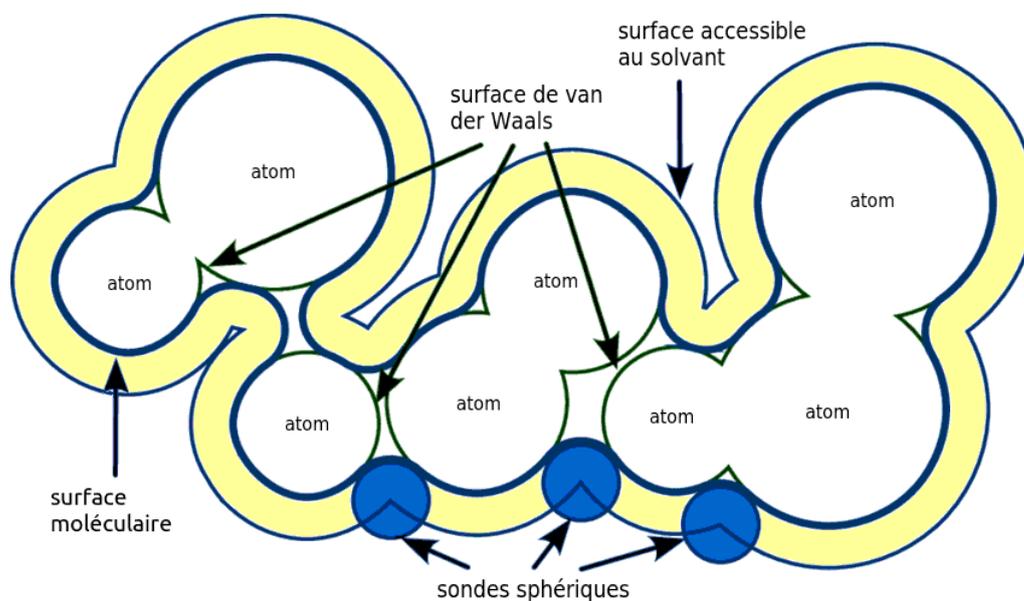


FIGURE 42 – Illustration du principe de calcul de la surface accessible au solvant d'une molécule (adapté de DABERDAKU 2018).

### 1.2.6 La surface enfouie

En parallèle de l'analyse des mouvements des acides aminés, on peut évaluer des mouvements impactant plus globalement des parties de la protéine. La mesure de la surface enfouie (ou BSA pour "Buried Surface Area" en anglais) de la

protéine au cours de la simulation permet de se rendre compte des zones de la protéine qui entrent en contact ou qui perdent contact. En effet, le principe de mesurer la surface enfouie est d'estimer les zones qui ne sont pas en contact avec le solvant au cours de la simulation (CHAKRAVARTY et al. 2013).

Dans notre cas, on cherche par exemple à calculer la surface enfouie entre deux motifs successifs de la protéine simulée. Pour cela, on calcule la surface accessible au solvant (cf. 1.2.5) des deux motifs individuellement puis on soustrait à cette valeur la surface accessible au solvant des deux motifs ensemble (cf. figure 43). La valeur de surface enfouie est exprimée en Angström carrés.

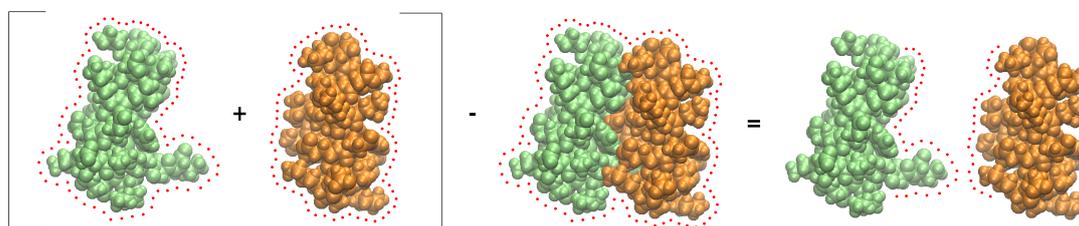


FIGURE 43 – Illustration du calcul de la surface enfouie entre deux motifs PPR. La surface de van der Waals des atomes composant les motifs PPR est affichée, les pointillés représentent donc la surface accessible au solvant sur un unique plan.

### 1.2.7 Les rotamères

L'étude des rotamères, par exemple des angles  $\chi_1$  et  $\chi_2$  des acides aminés est aussi une analyse importante (JANIN et al. 1978; GUHARROY, JANIN et ROBERT 2010). Il s'agit d'étudier le caractère mobile du squelette carboné de la chaîne latérale d'un acide aminé. Pour cela, on calcule un angle dièdre entre quatre atomes de carbone tel qu'illustré dans la figure 44. Cette information permet de mieux caractériser les suites de mouvements à l'échelle locale dans la protéine, par exemple les variations de l'orientation d'un groupe carboxyl dans un Asp (aspartate) au cours du temps.

### 1.2.8 Les paramètres hélicoïdaux

Les protéines à solénoïdes alpha sont localement assimilables à des superhélices composées d'un même élément (le motif répété) subissant une séquence de trans-

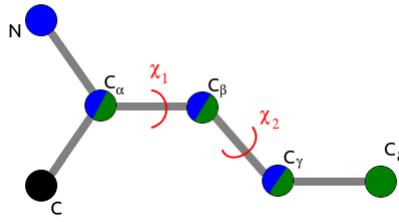


FIGURE 44 – Atomes utilisés pour définir les angles  $\chi_1$  (atomes en bleu) et  $\chi_2$  (atomes en vert) dans un acide aminé schématique.

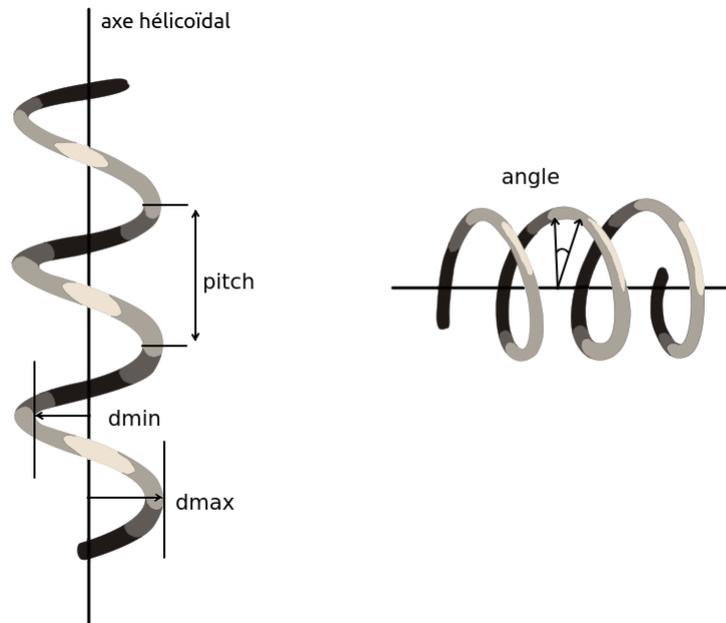


FIGURE 45 – Quelques uns des paramètres hélicoïdaux que l'on peut mesurer au cours de la simulation pour des éléments disposés régulièrement autour d'un seul axe hélicoïdal (adapté d'une illustration de Rajarshi Rit).

formations quasi-rigides autour et le long d'un axe. Il est alors possible de mesurer les paramètres de ces transformations en analysant les paramètres permettant de passer d'un motif au suivant (ANGELIDIS 2004; BOYER et al. 2015). Pour cela, nous avons mis en place des analyses entre deux motifs consécutifs, c'est à dire une paire de motifs. Les paramètres mesurés sont notamment les suivants (cf. figure 45) :

- l'axe hélicoïdal

- la distance minimum du motif PPR de l'axe hélicoïdal.
- la distance maximum du motif PPR de l'axe hélicoïdal.
- l'angle de la rotation d'un motif vers l'autre.
- le pitch d'un tour de l'hélice si l'hélice était composée uniquement de motifs provenant de la même paire de motif.

### 1.3 Discussion et conclusion

Certaines analyses sont essentielles et très classiques comme la RMSD ou les calculs de distance entre paires d'acides aminés. Il est nécessaire de produire des analyses variées pour qu'elles soient complémentaires les unes des autres afin de s'assurer que les conclusions soient justes. L'objectif étant de comprendre les mouvements globaux mais aussi locaux au sein de la protéine, c'est en accumulant plusieurs analyses, qu'on peut comparer à des points de vues différents et qu'on est capable de mieux appréhender certains mouvements ou phénomènes. Par ailleurs, il arrive très souvent que les résultats issus d'une analyse amènent à d'autres questions et à imaginer de nouvelles analyses pour tenter de répondre à ces nouvelles interrogations.

## 2 Second chapitre : la dynamique conformationnelle des protéines à "Pentatricopeptide repeat"

### 2.1 Introduction

Les protéines PPR codées dans le noyau et importées dans les organites (mitochondries et chloroplastes) sont formées de répétitions d'un même motif de 35 acides aminés. Elles ont la capacité de se lier à des séquences d'ARNsb pour agir ensuite dessus dans des rôles variés. D'autres familles de protéines à solénoïde alpha comme les protéines OPR ou PUF se fixent aussi à de l'ARNsb, dans les organites pour la famille des protéines OPR et dans le cytosol pour la famille des protéines PUF. La reconnaissance de l'ARNsb cible par la protéine est spécifique, c'est-à-dire qu'un motif répété reconnaît la séquence d'une portion de l'ARNsb cible. Dans le cas de notre étude nous nous intéressons exclusivement aux protéines à solénoïde alpha responsables de la régulation de l'expression des gènes des organites, c'est-à-dire les protéines OPR et PPR actuellement connues.

Le code de reconnaissance entre les protéines OPR et l'ARNsb est mal compris et celui entre les protéines PPR et l'ARNsb cible reste encore approximatif malgré les travaux de la dernière décennie (BARKAN, ROJAS et al. 2012; BARKAN et SMALL 2014; MCDOWELL, SMALL et BOND 2022) et c'est en partant de ce constat que nous étudions ces dernières protéines. De plus, bien que des codes de reconnaissance préliminaires plus ou moins avancés et fonctionnels soient disponibles pour les protéines OPR et PPR sur leur ARNsb cibles, seuls quelques acides aminés ont été caractérisés comme y étant impliqués. Qu'en est-il alors du reste des acides aminés des motifs répétés? Jouent-ils un rôle en plus des positions déjà identifiées dans la liaison spécifique de la protéine et de l'ARNsb?

D'autres questions ont guidé notre réflexion au fur et à mesure de l'avancement de nos travaux :

- Quel est le rôle de la dynamique de la protéine? Quel rôle pourraient jouer des changements de forme de la protéine dans la fixation de l'ARN, qui est très souple? Quelle est la stabilité de la liaison de la protéine à l'ARN? Une étude réalisée par SHEN et al. 2016 a comparé des structures cristallines de protéine PPR synthétiques liées et non liées à un ARNsb cible lui aussi synthétique. Les résultats de cette étude ont mis en évidence l'enroulement superhélicoïdal de la protéine lorsqu'elle est liée à un ARNsb (cf. figure 46).
- Les protéines PPR pouvant avoir jusqu'à 30 motifs répétés (SCHMITZ-LINNEWEBER et SMALL 2008), quel est l'effet du nombre de motifs ou

- de la longueur de la protéine sur la stabilité de la liaison ?
- Sachant que les protéines PPR peuvent avoir toute une variété de rôles comme la maturation ou l'épissage des ARN auxquels elles se fixent, y a-t-il un effet des domaines N- et C-terminaux sur la stabilité de la protéine et de la liaison ?
  - Comment caractériser la variabilité de la protéine en solution ? Quel rôle peut jouer l'association de l'ARN dans la structuration du site de fixation dans la protéine ?
  - Quels rôles jouent les interactions inter-motifs sur la dynamique de la protéine et de la liaison ? Peut-on visualiser une coordination entre les mouvements des motifs les uns par rapport aux autres ?

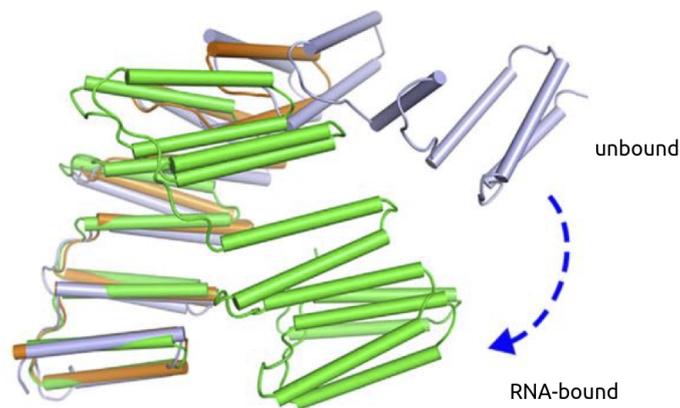


FIGURE 46 – Structures d'une protéine PPR synthétique liée (vert) et non liée (gris) à un ARNs. SHEN et al. 2016.

Pour tenter d'améliorer les connaissances sur le code de reconnaissance protéine PPR-ARNs et de répondre à ces questions, nous avons utilisé un système modèle constitué d'une protéine PPR synthétique contenant 10 motifs PPR identiques ainsi que des régions N- et C-terminales liée à un ARNs. Ce système a été cristallisé par SHEN et al. 2016. Le motif PPR synthétique est issu du consensus de tous les motifs PPR des protéines PPR de type P connues chez *Arabidopsis thaliana* (cf. figure 47) (SHEN et al. 2016). Il s'agit également de l'une des rares structures de protéine PPR liée à un ARNs dans toute sa longueur disponible publiquement (cf. figures 22 et 27). Ce système offre l'avantage de pouvoir faire des analyses systématiques qui ne seront pas influencées par des variations de séquences entre motifs comme dans une protéine PPR naturelle. De plus, il s'agit d'une structure de très bonne qualité avec une résolution de l'ordre de 2Å.

En parallèle de ce système, les auteurs de SHEN et al. 2016 ont résolu la structure de trois autres systèmes similaires. Elles ne diffèrent que par deux de leurs motifs et deux des nucléotides de l'ARNsb pour chacune sur la base du code PPR (BARKAN, ROJAS et al. 2012; BARKAN et SMALL 2014). L'objectif derrière ces constructions est l'étude du code de reconnaissance protéine-ARN, de la spécificité de l'interaction et de l'affinité de l'ARN pour la protéine. Nous n'avons utilisé que celle ayant la meilleure résolution (2,19Å) dans cette étude.

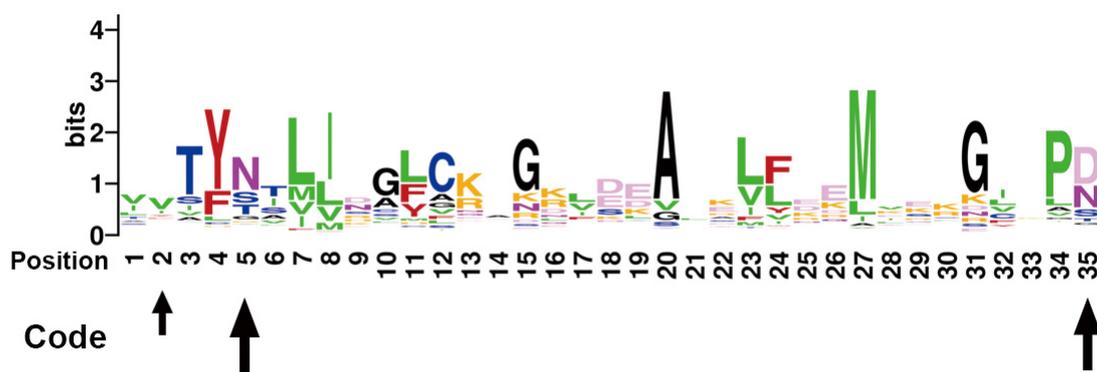


FIGURE 47 – Logo des motifs des protéines PPR de type P de l'espèce de plante terrestre *Arabidopsis thaliana*. Les résidus du code PPR sont indiqués par des flèches noires (adapté de SHEN et al. 2016).

Issues de la structure PDB portant l'identifiant 5i9f (SHEN et al. 2016), nous avons généré et utilisé plusieurs constructions qui sont composées d'un nombre variable de motifs PPR : 1, 2, 4 et 10 (cf. Figure 48). Ces constructions ont pour intérêt de permettre d'analyser la stabilité d'un motif seul et des motifs lorsqu'ils sont de plus en plus nombreux pour appréhender la stabilité intrinsèque et la plasticité du solénoïde alpha.

L'avantage d'utiliser une protéine PPR synthétique est qu'il s'agit d'un modèle simplifié et régulier qui n'est constitué que de répétitions d'un même ensemble. On peut donc examiner les données sur les 10 motifs indépendamment ou bien les comparer entre eux, sans biais de séquence. On peut ajouter également des parties Nter et Cter consécutives résolues dans la structure cristalline d'une quinzaine de résidus chacune, ce qui permet d'approximer l'impact des domaines terminaux sur la dynamique de la protéine et sa stabilité lorsqu'elle n'est pas liée à de l'ARNsb.

En 2018 durant son stage de master 2, Marion Sisquellas a réalisé les premières

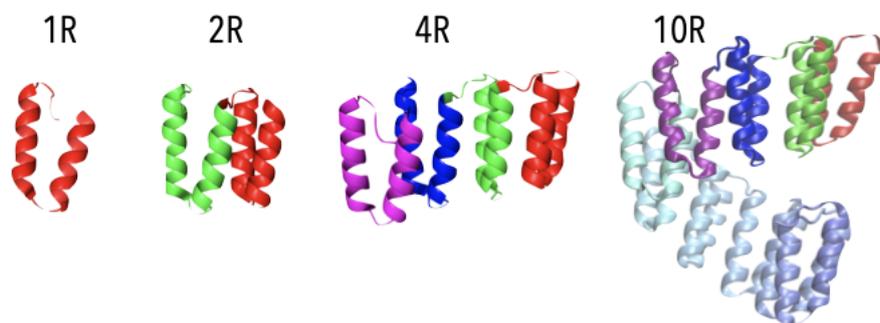


FIGURE 48 – Les quatre constructions réalisées par Marion Sisquellas lors de son stage de master 2 en 2018 pour son étude de la dynamique des protéines PPR qui a ouvert la voie au travail présenté ici (SISQUELLAS 2018).

études sur ces systèmes et le développement d'un jeu initial de constructions de protéines basées sur les données de SHEN et al. 2016. Elle a également commencé le développement préliminaire d'une stratégie de simulation de ces systèmes (SISQUELLAS 2018).

Ici, nous avons aussi repris ces constructions et nous en avons généré d'autres pour entreprendre nos travaux :

- une version des constructions qui n'est formée que des motifs PPR (il s'agit du jeu initial généré par Marion Sisquellas).
- des constructions PPR auxquels nous avons ajouté des groupements neutres aux extrémités (les groupements ACE et CT2 au N- et C-terminus respectivement).
- des constructions PPR avec des portions terminales Nter et Cter résolues dans la structure cristalline, auxquelles nous avons ajouté les mêmes groupements que ceux du point précédent.
- enfin, nous avons généré un système comprenant le complexe protéine-ARNsb auquel nous avons supprimé les nucléotides incomplets aux extrémités de l'ARNsb.

Ces constructions ont été réalisées dans l'objectif de caractériser la stabilité intrinsèque des motifs PPR ainsi que l'action stabilisatrice des groupes terminaux. Nous avons également cherché à pouvoir mieux comprendre l'évolution dynamique du site de fixation de l'ARNsb, sa pré-structuration dans le cas des constructions sans ARNsb et la reconnaissance spécifique des nucléotides. En effet, la variété de nos systèmes nous a permis de montrer la flexibilité de la protéine sans ARN et sa plus grande stabilité une fois liée à l'ARNsb ainsi que le caractère dynamique

de la liaison de la protéine à l'ARNsb.

Nous avons fait des simulations de dynamique moléculaire pour chacune des constructions décrites dans le tableau 3.

Système	motifs	ter	grp ter	ARN	$\Delta t_{prep}(fs)$	prep (ns)	pas <sub>equi</sub>	$\Delta t_{equi,prod}(fs)$	equi (ns)	pas <sub>prod</sub> ( <i>millions</i> )	prod ( $\mu s$ )
1R <sub>hmr</sub>	1	Non	Non	Non	2	3,3	2335000	4	93,4	125	0,5
1R <sub>ter+hmr</sub>	1	Oui	Non	Non	2	3,3	2335000	4	93,4	125	0,5
1R <sub>ter+ACE+CT2+hmr</sub>	1	Oui	Oui	Non	2	3,3	2335000	4	93,4	125	0,5
2R <sub>hmr</sub>	2	Non	Non	Non	2	3,3	2335000	4	93,4	125	0,5
2R <sub>ter+hmr</sub>	2	Oui	Non	Non	2	3,3	2335000	4	93,4	125	0,5
2R <sub>ter+ACE+CT2+hmr</sub>	2	Oui	Oui	Non	2	3,3	2335000	4	93,4	125	0,5
4R	4	Non	Non	Non	2	3,3	2335000	2	46,7	250	0,5
4R <sub>bis</sub>	4	Non	Non	Non	2	3,3	2335000	2	46,7	250	0,5
4R <sub>ter+hmr</sub>	4	Oui	Non	Non	2	3,3	2335000	4	93,4	250	1
4R <sub>ter+ACE+CT2+hmr</sub>	4	Oui	Oui	Non	2	3,3	2335000	4	93,4	250	1
10R	10	Non	Non	Non	2	3,3	2335000	2	46,7	250	0,5
10R <sub>ter+ACE+CT2+hmr</sub>	10	Oui	Oui	Non	2	3,3	2335000	4	93,4	1250	5
10R <sub>ter+ACE+CT2+hmr+bis</sub>	10	Oui	Oui	Non	2	3,3	2335000	4	93,4	250	1
10R <sub>ter+ACE+CT2+ARN+hmr</sub>	10	Oui	Oui	Oui	2	3,3	2335000	4	93,4	250	1
10R <sub>ter+ACE+CT2+hmr+c36</sub>	10	Oui	Oui	Non	2	3,3	2335000	4	93,4	500	2

TABLE 3 – Les 15 systèmes de la protéine PPR avec et sans ARN simulé puis analysés par dynamique moléculaire

Nous avons fait des répliques de certains systèmes qui montrent le même comportement qualitatif. Par exemple, pour le système 4R (4 motifs sans groupements terminaux), les fluctuations le long de chaque motifs sont présentées dans les figures de l'article et sont plus importantes au niveau des résidus situés aux extrémités du système, et le premier et le dernier motifs sont les plus fluctuants. Ces mouvements qui interviennent sur des échelles de temps de l'ordre de la centaine de ns sont échantillonnés dans les deux répliques. Il s'agit de mouvements de déplie-ment des extrémités en absence des régions N- et C-ter ainsi que des groupements neutres qui ne sont plus observés dans les simulations ces systèmes les comportant.

On note également qu'en général, les fluctuations sont plus importantes pour les résidus situés aux extrémités des motifs : ce sont les régions formant les linkers de 4 résidus entre motifs.

Ce travail est présenté dans un second article qui est actuellement en cours de rédaction et qui sera soumis prochainement.

## **2.2 Article 2**

# Repeat-motif protein flexibility in relation to RNA binding in a consensus PPR scaffold

Céline Cattelin<sup>1,2,3,4</sup>, Marion Sisquellas<sup>1†</sup>, Chantal Prévost<sup>1,3</sup>, Ingrid Lafontaine<sup>2,4</sup>, and Charles H. Robert<sup>1,3\*</sup>

Last updated June 26, 2023

<sup>1</sup> CNRS Laboratory for Theoretical Biochemistry UPR 9080, Institut de Biologie Physico-Chimique, Paris France

<sup>2</sup> CNRS Laboratory for Chloroplast Biology and Light Sensing in MicroAlgae UMR 7141, Institut de Biologie Physico-Chimique, Paris France

<sup>3</sup> Univ Paris Cité, Paris, France

<sup>4</sup> Sorbonne University, Paris, France

† current address: Inserm U1268, CITCOM UMR 8038 CNRS - Univ Paris Cité

\* E-mail: robert@ibpc.fr

## 1 Abstract

## 2 Introduction

Tandemly-repeated motif proteins are found throughout biology (Filipovska & Rackham, 2012, Paladin *et al.*, 2021). In eukaryotic cells, classes of proteins containing tandemly-repeated alpha-solenoid motifs are coded in the nuclear genome, expressed in the cytoplasm, and then enter mitochondria and chloroplasts as agents of nuclear control of these organelles' own gene expression. They do this by binding to specific mRNA targets in a *one motif : one nucleotide* manner. Key residues implicated in base recognition in each motif have been identified in some families, permitting the elucidation of protein-RNA binding codes (*e.g.*, Barkan *et al.*, 2012), which in turn has permitted the development of designed binders for biotechnology and medicine (Coquille *et al.*, 2014, Hall, 2016, Shen *et al.*, 2016, Zhao *et al.*, 2018). The existence of such codes also offers the possibility that sequences of natural proteins may be used to predict those of their cognate RNA. In this way an improved understanding of the physical chemical determinants of RNA base recognition in these regulation systems can not only help decipher specificity and biological function in general but also shed light on the detailed evolution of these energy-producing organelles themselves.

Proteins capable of binding sequence-specifically to single-stranded RNA intervene in many different cellular processes, including translation, splicing, and editing (Filipovska & Rackham, 2012). Several protein families in particular are relevant in this context, including:

- PPR proteins: pentatrico (35 aa) peptide repeats active in the regulation of organellar gene expression, found predominantly in plants (Barkan & Small, 2014, Nakamura *et al.*, 2012)
- OPR proteins: octatrico (38 aa) peptide repeats, having a similar function as PPR proteins, essentially in *μ*algae (*e.g.*, Eberhard *et al.*, 2011, Marx *et al.*, 2015)

- PUF-family proteins: Pumilio and FBF factors, acting in translational gene expression in the cytoplasm (Wang *et al.*, 2018).

As is frequently the case for repeat-motif proteins in general, these proteins are found to fold into non-globular, alpha-solenoid architectures.

The PPR proteins in particular are one of the largest protein families in land plants, and have been the focus of extensive experimental studies. Their findings, including several crystal structures, have played a fundamental role in elucidating the PPR code for protein-ssRNA binding. Available evidence suggests that the PPR system may also provide a model for RNA recognition by OPR proteins as well (Marx *et al.*, 2015). Nevertheless the PPR code remains incomplete (*e.g.*, Miranda *et al.*, 2017). Furthermore, although a solution structure has been proposed from X-ray scattering data (Gully *et al.*, 2015), the roles of flexibility and dynamics in PPR-RNA binding and recognition and regulation have not yet been studied in detail, and merit attention.

To better understand the physical-chemical nature of sequence-specific recognition of mRNA by PPR and related repeat-motif proteins and the possible roles of flexibility, we have chosen to analyze their dynamics at the atomic level by employing molecular dynamics simulation, leveraging crystal structure data that has been obtained to high resolution. The protein we have chosen is the result of a refined consensus design with identical motifs (Shen *et al.*, 2016), which presents several advantages for dissecting the different levels of flexibility in the protein.

## 3 Methods

### 3.1 Constructs

We constructed a series of starting structures derived from pdbid 5i9f of the 10-repeat consensus PPR in complex with its cognate RNA poly-U<sub>10</sub>— the highest-resolution structure of the 4 PPR-RNA complexes of closely related consensus-design sequences obtained in that work (Shen *et al.*, 2016). The tandemly repeated sequences in 5i9f are strictly identical for all motifs. We defined constructs containing 1, 2, 4, or all 10 repeat motifs counting from the N-terminal side of the repeat-motif domain in the structure. Repeat motifs in the constructs were sandwiched between the 14 residue N- and XX-residue C-terminal regions reported in the PDB file and capped by ACE (N-terminal) and CT2 (C-terminal) groups. The C-ter region coordinates were determined via superpositions of the final motif in each construct with that in 5i9f. Separate MD studies (not shown) showed that the absence of these terminal regions led to higher conformational instability in constructs of all sizes. Each construct is referred to by the number of repeats together with a suffix to emphasize the presence of the terminal regions, *e.g.*, **4Rtern** for a construct consisting of four contiguous motifs along with the neutral N- and C-terminal regions.

### 3.2 Molecular dynamics simulations

Each construct was solvated and ions placed in energetically favorable positions in a periodic box using Charmm-gui (Jo *et al.*, 2008, Lee *et al.*, 2016) keeping at least 20 Å to the closest image in each direction. Na and Cl ions were added to achieve a neutral solution with 150mM excess salt. Amber ff14sb (), with ff99bsc0<sub>χOL3</sub> in system 10Rtern\_rna, and TIP3P water was used in all simulations save one: a second simulation of the 10Rtern system using the charmm36 forcefield was carried out to identify possible force-field-dependent effects. Energy-minimization and heating of the starting structure was designed to gradually accommodate the absence of RNA in the protein-only simulations; for this, restraint forces operating differentially on backbone atoms ( $K = 25 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ ) and on sidechain atoms and ions ( $K = 12.5 \text{ kcal-mol}^{-1} \text{ \AA}^{-2}$ ) were relaxed to zero in a stepwise manner, with a short

cycle of heating and re-cooling applied at each reduction in order to favor exploration of low-energy regions of the potential energy surface (success of which was noted through the achievement of lower and lower potential energy averages after each cycle). For consistency the same protocol was applied in the 10Rtern-rna simulation as well. NAMD2 (Phillips *et al.*, 2020) was used for all simulations, using hydrogen mass repartitioning to permit a 4 ps integration timestep and a Langevin thermostat and barostat to define an NPT ensemble at 300 K and 1 atm pressure. Periodic boundary conditions were employed with particle mesh electrostatics. Standard amino acid protonation states were employed for all simulations after pKa calculations of the starting structures using propka3 (Olsson *et al.*, 2011).

Constructs containing 1 or 2 motifs as well as the 10Rtern\_rna system were either small enough or spherical enough to simulate using a truncated icosahedral box. For longer constructs (4R, 10R, ...), the Colvars dashboard (Hénin *et al.*, 2022), implemented in VMD (Humphrey *et al.*, 1996), permitted applying a harmonic collective-variable orientational restraint in NAMD defined in terms of the backbone CA atom coordinates ( $k_{orient} = X \text{ kcal-mol}^{-1}\text{-deg}^{-2}$ ) and acting on the instantaneous deviation of the long axis of the protein from its starting orientation, which was directed along the x-axis of the orthorhombic periodic cell.

## 4 Results

### 4.1 Conformational dynamics of PPR constructs

#### apo- versus RNA-bound PPR protein dynamics

In MD simulations of the 5i9f PPR-RNA complex, the PPR protein remained close to the starting crystal structure (Shen *et al.*, 2016), with least-rms C $\alpha$  distance (least-rmsd or lrmsd) from the superhelical starting structure varying in the range of 2-4 Å throughout the 1  $\mu$ s trajectory Fig. 1(*left*).

In contrast, in the absence of RNA, the protein unwound significantly and became more flexible overall: in a 5  $\mu$ s simulation the protein explored an extended region of conformational space in which the overall dimensions of the molecule varied continuously, ranging from about 5 Å up to 20 Å rmsd from the RNA-bound structure, and with far greater fluctuations, as can be seen from Fig. 1(*right*). Similar results were seen when using the charmm36 force field (not shown). Representative snapshots from the RNA-bound and apo-protein simulations are shown at bottom in Fig. 1.

We note that although these MD simulations revealed high flexibility of the unbound PPR scaffold compared to the protein-RNA complex, the sampled conformations are consistent with forms seen both in crystal structures of other designed PPR proteins in the unbound state (Coquille *et al.*, 2014) and in a structure of a modified PPR10 protein partially bound to its RNA ligand (Yin *et al.*, 2013), and binding-induced conformational change was highlighted by Shen *et al.* (2016).

#### Repeat motif stability

To examine the origins of the flexibility of the apo PPR protein in detail, we studied the dynamics of the repeat motifs themselves in constructs of different sizes ranging from 1 to 10 repeats, in comparison to the motif dynamics in the RNA-bound protein (Fig 7).

We note first in Fig 7 that the motif lrmsd is less than 1.5 Å in all constructs, and less than 1 Å in virtually all motifs in the 10R molecule. This is far below the variability one might expect from the 10-repeat apo-PPR molecule results discussed above (Fig. 1, *right*). This point will be addressed in a later section. Second, motif dynamics depend on the length of the repeat domain in which they are

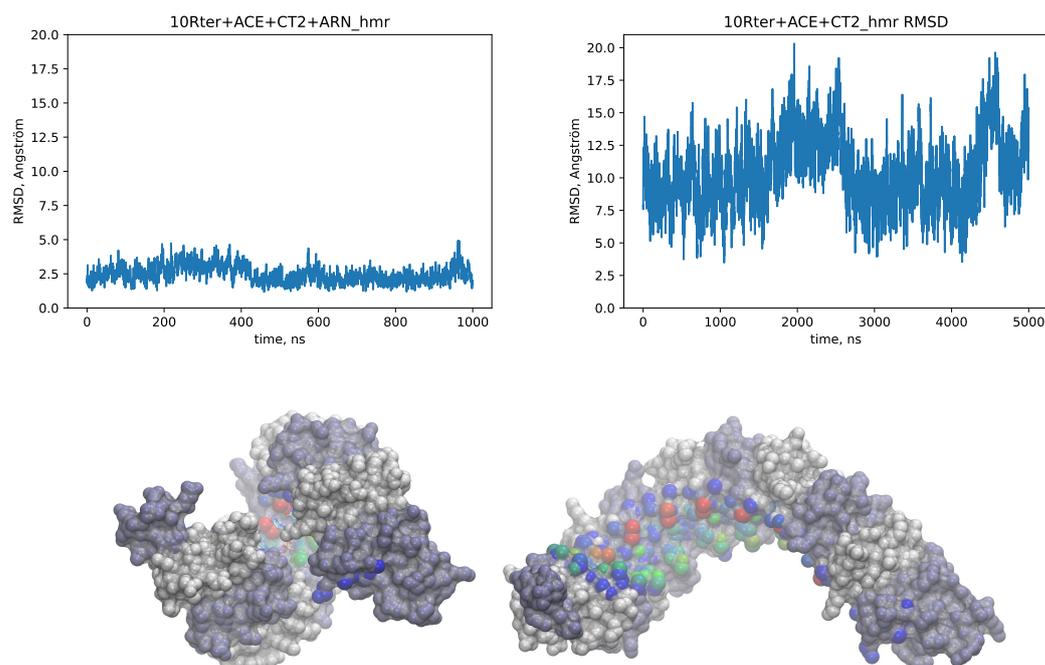


Figure 1: *Top row:* PPR protein RMS deviation after optimal superposition (*lrmsd*) during molecular dynamics simulation of the PPR-RNA complex (*left*) and of the apo-PPR molecule (*right*). *Bottom row:* Space-filling representations of the protein in the corresponding systems at representative instants during the simulations, in which the terminal regions are colored in gray-blue and the motifs colored alternately by white and gray-blue. In both protein representations, atoms contacting the RNA in the PPR-RNA simulation are colored from blue to red in proportion to their average degree of surface burial in the MD simulation. [fig-rmsd-apo]

found: motifs in constructs containing only 1 or 2 motifs tend to deviate farther from their RNA-bound conformation than motifs in the longer constructs with 4 and 10 motifs.

The overall form of the distribution of *lrmsd* values is different for different motifs in the constructs. For example, the *lrmsd* of the (single) 1Rtern motif is essentially monomodal, with an average *lrmsd* of about 1 Å with respect to the RNA-bound starting structure. Qualitatively similar monomodal distributions were found for the C-terminal repeats in the MD of each construct, including that the full-length PPR-RNA (Fig 7, (*second from bottom*)). However, while a 1 Å *lrmsd* indicates that the motif structures are very similar to those in the RNA-bound protein, the first motif in the 2Rtern construct in particular shows a dominant mode that is even closer – about 0.6 Å. Indeed, virtually all the motifs except the last show a dominant mode at this *lrmsd* distance from the RNA-bound structure. In the RNA-bound PPR simulation 10Rtern\_rna (Fig 7, (*bottom*)), the distributions of all the motifs except the last have a single mode at about 0.6 Å throughout the MD simulation, and the (slightly) more distant mode at 1 Å is all but absent.

Finally, as one tracks from left to right comparing the last two panels of Fig 7, the overall pattern of conformational deviation as a function of motif number is seen to be quite similar for the apo construct 10Rtern and the RNA-bound 10Rtern\_rna. However, the apo-form motifs have a more marked bimodal character than the motifs in the RNA-bound PPR. This suggests that small differences in the motif structures on the order of 1/2 Å play a role in the much larger conformational variability of the apo-construct 10Rtern.

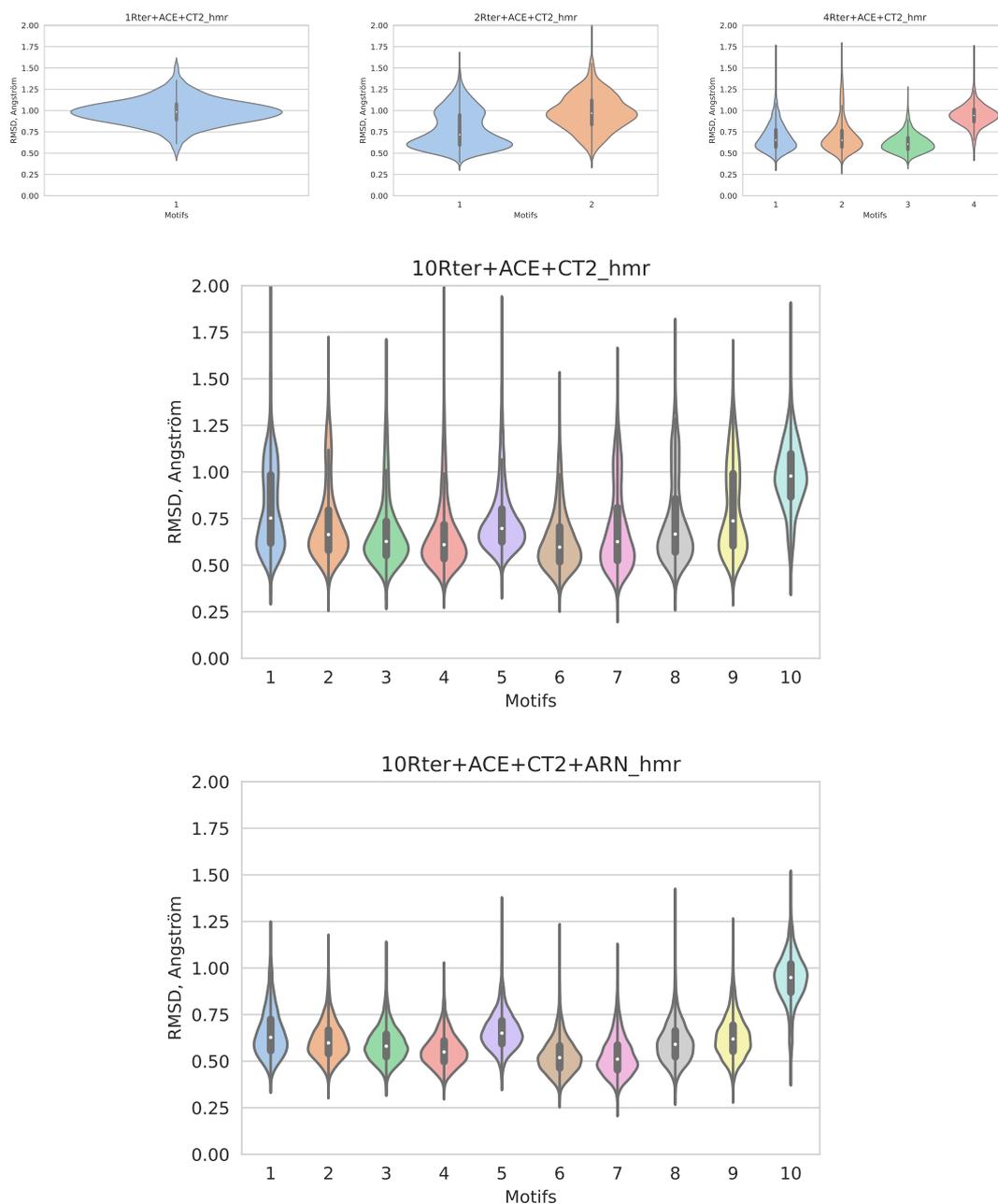


Figure 2: Least-rmsd distributions for individual motifs compared to the starting structure in apo PPR constructs containing 1, 2, 4 (*top row*), and 10 motifs (*middle row*). and in the 10-motif apo-PPR molecule(*bottom*). [fig-rmsd-distr-motif].

### Motif-pair dynamics

Although the motifs themselves are therefore remarkably rigid in the longer constructs, by examining the dynamics of neighboring motif pairs a larger degree of structural variability is seen in the apo simulations. Fig. 3 shows the distributions of lrmsd values broken down by motif-motif interface in the 10Rtern system without (left) and with (right) bound RNA. The average rmsd difference from the crystal structure form is larger in most motifpairs in the apo protein than in the RNA-bound PPR, and the fluctuations are much greater.

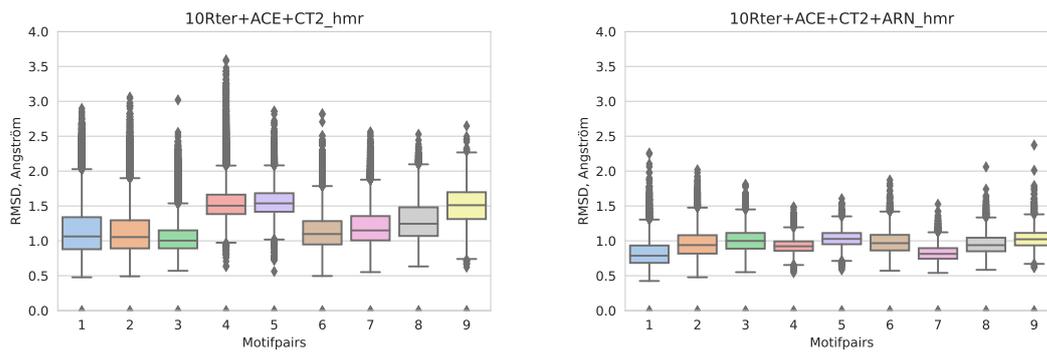


Figure 3: Least-rmsd distributions for motif pairs 1 to 9 compared to their starting structures in apo PPR 10Rtern construct (left), and 10Rtern-rna PPR-RNA complex(right). [fig-rmsd-motifpairs].

**Surface burial in motif-motif interfaces**

In order to study the motif-motif packing in detail, we examined the BSA of each residue in the motif-motif interfaces.

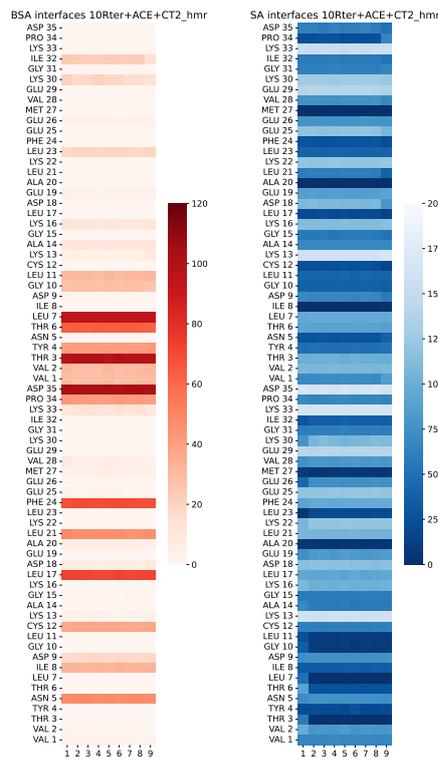


Figure 4: default [fig-heatmap-motifpair-bsa]

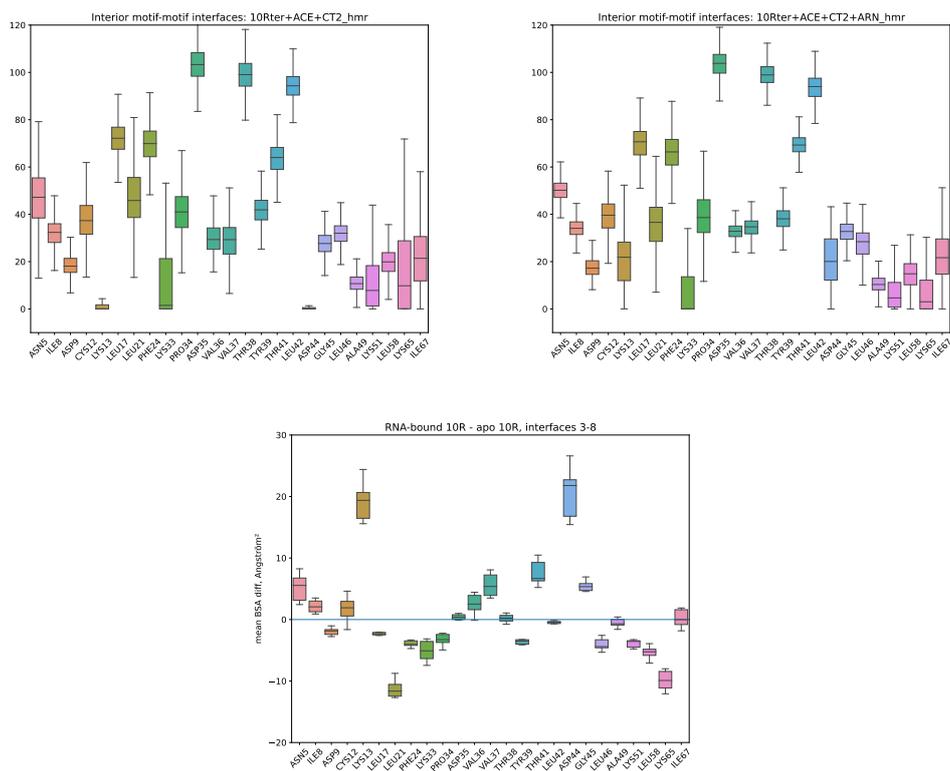


Figure 5: default [fig-box-bsa-internal-apo-rna]

### Motif-motif packing geometry

Any rigid-body movement in 3D space can be described as a twist, or screw transformation, defined by a helical axis and a rotation around that axis (Angelidis, 2004). We analyzed the instantaneous helical twist defined between two successive quasi-rigid motifs (Boyer *et al.*, 2015) throughout the MD simulations. An example of a collection of helix axes from such an analysis is shown in Fig. 6.

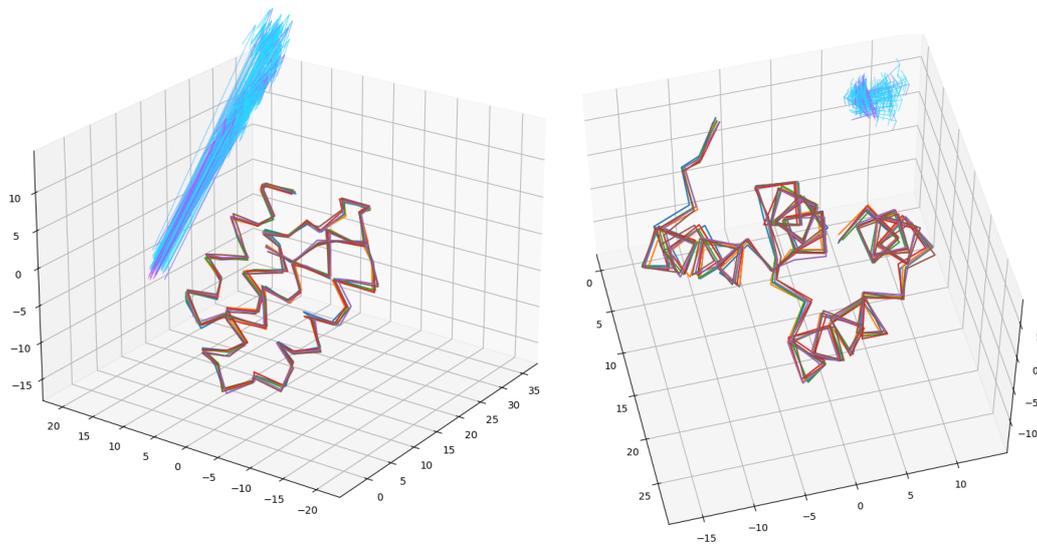


Figure 6: Two different views of a superposition of neighboring PPR motifs X and Y (in wireframe) from construct XYZ together with the collection of helix axes describing the geometrical screw transformation relating one to the other (lines colored by angular deviation from the average axis).

The screw transformation is defined in terms of a vector and a point in space which fix the helix axis, the translation along the axis, and the rotation angle. These parameters can be used to define the radius, the pitch and the number of monomers per turn of the helix (Boyer *et al.*, 2015). The distribution of the number of monomers per turn and the pitch is plotted in Fig. 7 for the PPR-RNA complex 10Rtern\_rna and the apo-protein 10Rtern. The positions of the peaks of the distributions differ by about 20 Å in pitch, indicating that the geometrical relationship between neighboring motifs is significantly different for the apo-PPR protein simulation 10Rtern with respect to the 10Rtern\_rna. Nevertheless, motif pairs in the RNA-bound protein simulation are seen to sample motif-motif packing geometries that are also seen in the most populated region of the apo-PPR simulations.

It would thus appear that thermal variations in the motif-motif packing geometry are the major contributor to the large fluctuation of the apo-PPR protein, as can be seen from the examples shown in Fig. 8 of 10-repeat regular PPR geometries generated using perfectly rigid motifs and propagating the helix parameters taken from the apo-PPR simulation to reconstruct hypothetical regular helical geometries. In the simulation themselves the structure is not regular as each interface is not rigidly coupled to its neighbor.

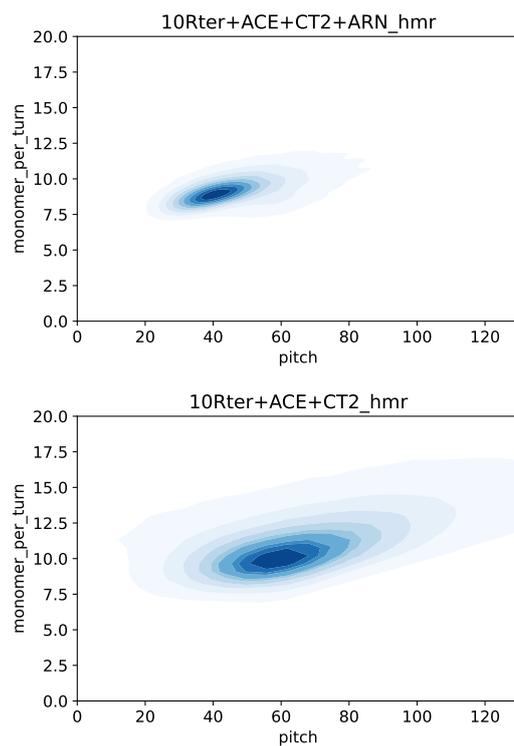


Figure 7: Distribution of the helical parameters describing motif-motif packing geometry in terms of number of monomers per superhelical turn and helix pitch, for all motifpairs in the PPR-RNA simulation 10Rtern\_rna (*top*) and apo-PPR construct 10Rtern (*bottom*). [fig-rmsd-distr-motif]

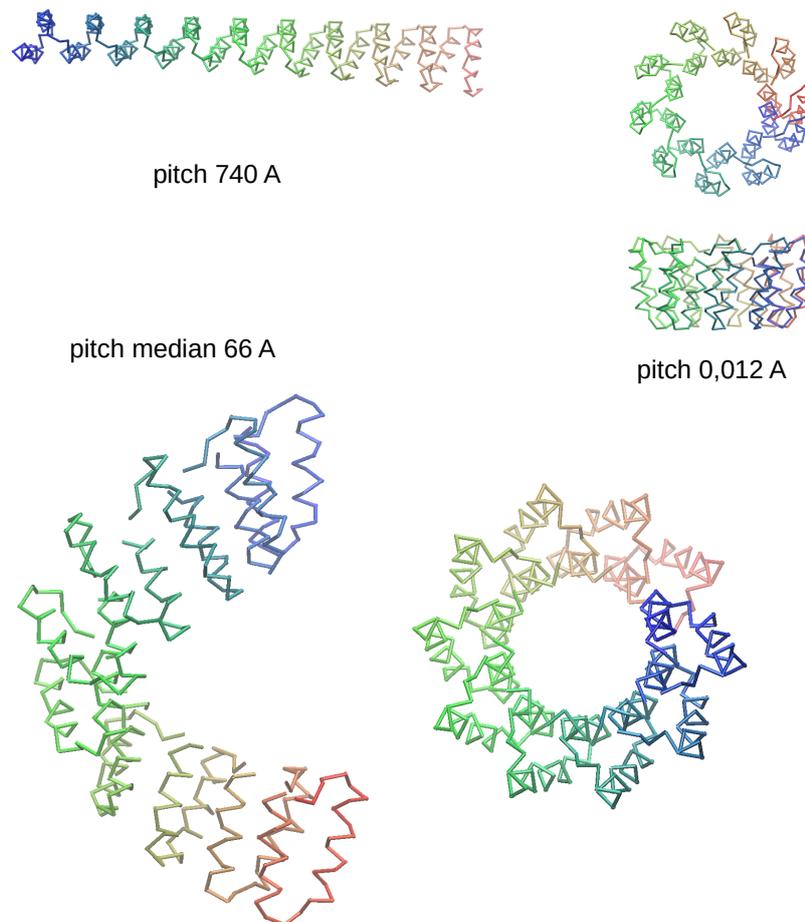


Figure 8: Reconstructed regular helical structures obtained from rigid motifs associated using different examples of motif-motif packing geometries observed during the MD simulation of the apo-PPR construct 10Rtern. [fig-regular-helical-structures]

## 4.2 PPR-RNA dynamics

We also studied the dynamics of the 10R protein in complex with its cognate RNA (10Rtern\_rna).

### Protein-RNA interface

The area of the protein-RNA interface (BSA) as a function of time is shown in Fig. 9. The BSA in these simulations was seen to fluctuate in a range of about  $1000 \text{ \AA}^2$ . The right side Fig. 9 breaks down the BSA contributions by residue and by motif. Consistent with the crystal structure results (Shen *et al.*, 2016), the contacts are essentially limited to helix a of the repeat motifs and Asp35, which line the inner cavity of the superhelix. the most extensive surface burial is contributed by Val2, which "sandwiches" the Hbond to Asn5. Near the phosphate group one finds Lys13 (from motif  $i-2$  or  $i-1$  depending on whether the phosphate 5' or 3' to the nucleotide is considered), whereas Asp9 (from motif  $i-1$  or  $i$ , see preceding) which positions itself between successive Lys13 sidechains in the vicinity of the C4' and C5' atoms. Thr6 from motif  $i+1$  contacts the region near the bond joining the uracil nucleobase to the ribose.

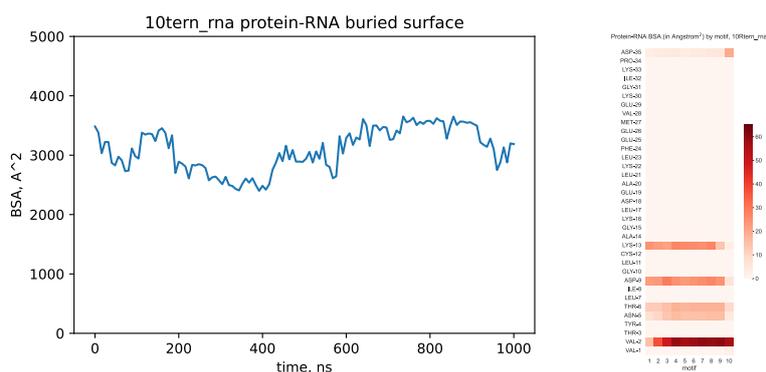


Figure 9: *Left*: Surface in  $\text{A}^2$  buried at the PPR protein-RNA binding interface as a function of simulation time. *Right*: Time average of the residue surface buried in the protein-RNA interface, by motif. [fig-bsa-protein-RNA]

Whereas the sugar phosphate backbone remained bound throughout the simulation period, intermittent unbinding of the first 3 RNA bases took place during the trajectory. This is reflected in the weaker BSA seen in particular for the Val2 residues in the first 3 motifs in Fig. 9(right). To illustrate the partial unbinding, Fig. 10 shows the protein colored by protein-RNA BSA, together with three snapshots of the RNA strand: before unbinding (X ns), at a point in which the first 3 bases were unbound (Y ns) and finally at Z ns at which the entire RNA strand was again bound to the protein. ¶

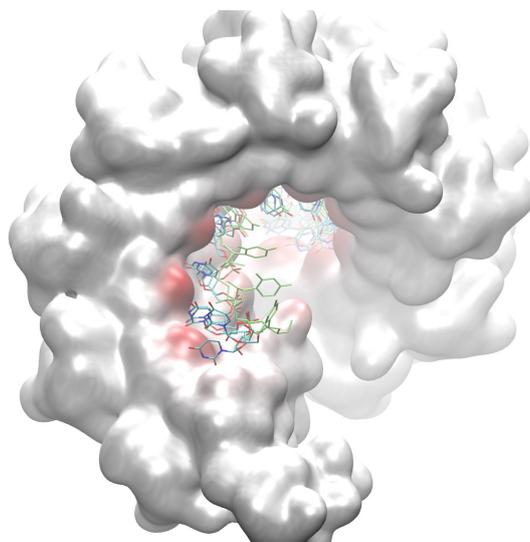


Figure 10: Surface representation of the PPR-protein complex colored white to red according to the average surface buried in protein-RNA contacts, and showing 3 different snapshots for the RNA in wireframe representation (colored by atom type) taken at successive moments in the MD simulation: RNA is fully bound, the first 3 Uracil bases have unbound, bases 2 and 3 have rebound to the binding site. [fig-struct-protein-RNA]

### Nucleotide base binding-site geometry

An important contributor to the RNA specificity in PPR systems is hydrogen-bonding, either direct or water-mediated, to the nucleotide base with amino acid residues at positions 5 and 35 of each motif. Indeed the amino-acids conserved in these two positions defines the current version of the PPR-RNA code Barkan *et al.* (2012), Coquille *et al.* (2014), Shen *et al.* (2016). How much of the binding site

geometry is preserved in the unbound PPR constructs?

Fig. 11 shows the chi1-chi2 distribution of the PPR code residue Asn5 for each motif in the apo PPR (left) and PPR-RNA complex (right) simulations. This Asn side chain in the bound protein remains in the rotamer observed in the crystal structure, which is consistent of course with the fact that this residue is engaged in a stable hydrogen bond with the uracil base. (An exception is the Asn at position 5 in motif 10 in the RNA-bound protein, which occupies a chi2 rotameric state adjacent to that seen in the crystal.)

On the other hand, the Asn side chain in the unbound protein is seen to explore many rotameric states. Yet in all motifs the dominant rotameric state is indeed that seen in the RNA-bound protein, suggesting that the nucleotide-base binding site, while not completely pre-structured as might be considered ideal for assuring an optimal free-energy of association, is at least partially structured, and has a significant probability of being in the appropriate orientation to form a hydrogen bond with an incoming uracil base.

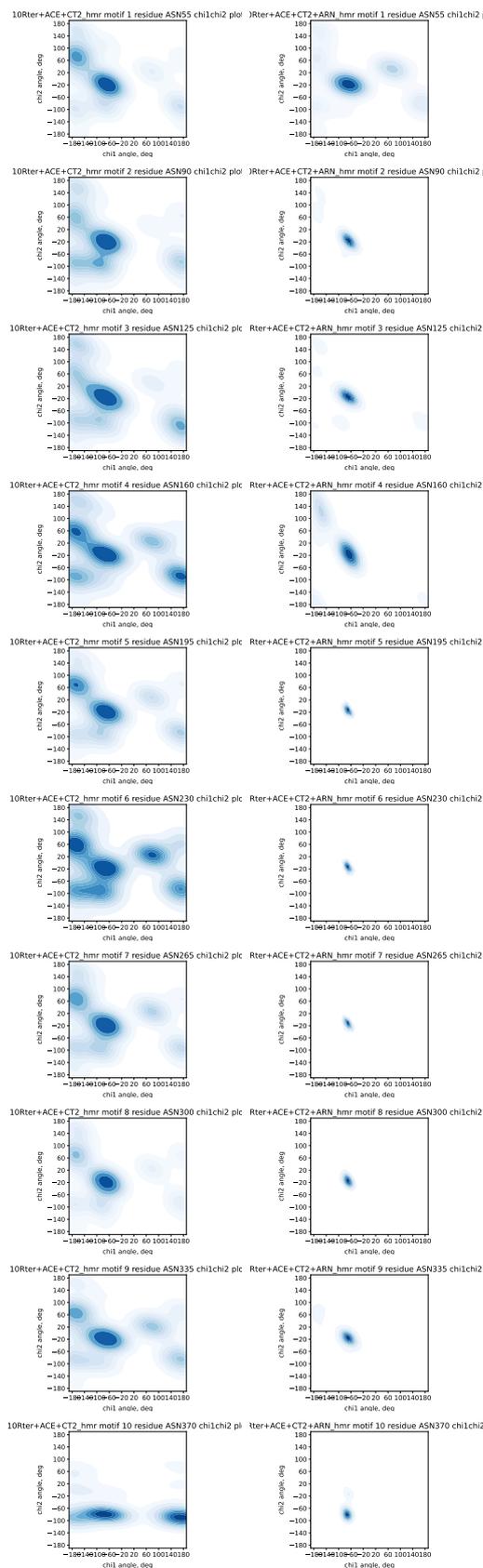


Figure 11:  $\chi_1$  -  $\chi_2$  dihedral angle distributions for the 10Rtern (apo) (left) and 10Rtern\_rna (holo) systems (right) [fig-chi1chi2-Asn]

### Val2 - Val2 distance

As revealed by the crystal structures of PPR-RNA complexes (Shen *et al.*, 2016, Yin *et al.*, 2013), the cognate RNA nucleotide base forms a hydrogen bond with the PPR code residue at position 5 deep in the binding site formed by neighboring motifs. In those structures, this hydrogen bond is sandwiched between two successive hydrophobic residues (Val, Phe) from the N-terminal end of helix  $\alpha$  in each motif. The distance between successive Val2 atoms CG2 in the RNA-bound PPR in the 5i9f structure is  $XX \pm YY$  Å. We note that this packing arrangement completely dehydrates the protein-RNA hydrogen bond, which would protect it from encroachment by solvent water molecules and raise the energy barrier to the charge separation needed to initiate breaking the Hbond.

The distance between atoms CG2 in Val2 of successive complete motifs in the unbound designed PPR crystal structure 4pjr of Coquille *et al.* (2014) is 9.77, 9.40, 9.65 Å.

## 5 Discussion

The simulations carried out in this work reveal the dynamics of a series of unbound PPR repeat-motif proteins and of a PPR-RNA complex modelled closely on the crystal structure of the 10-repeat consensus-design PPR protein and its cognate RNA  $U_{10}$  protein reported by Shen *et al.* (2016).

Although the state of the art of RNA and protein-RNA simulation is still undergoing significant evolution (*e.g.*, Krepl *et al.*, 2015), the PPR-RNA complex studied here proved to be remarkably stable in these MD simulations. Although the total surface buried in the protein-RNA complex was found to vary on the order of about 500 Å<sup>2</sup>, the value did not show a tendency to diminish on the timescale of the simulations (Fig. 9). We did however observe that the first three 5' RNA nucleotides, which we will call U1, U2, and U3 for simplicity, as well as U10, exhibited signs of comparatively less-tight binding, as they unbound and rebound to the protein during the simulation. The contributions of these bases to the overall affinity of the protein-RNA complex is not known and cannot be ascertained from these simulations. However, it is possible that the absence of the 5' phosphate on the first uracil in our simulations—and indeed the absence of any 5' nucleotides at all—may deprive the RNA of an additional stabilizing interaction, contributing to instability of the nucleotides at the 5' extremity of the RNA. Such local destabilization is unlikely to be encountered in the cell, as the PPR binding site in general is not located at an extremity of the mRNA target template.

The observed lability of the protein-uracil interaction at the 3' end of the targeted sequence cannot be explained in this way. However, we were unsurprised by the weaker interaction of the uracil in this position: the U10 base-binding site is only partially formed by the terminal motif R10 (which by definition has no matching consensus-sequence motif following it) and the hydrogen bond from the Asn5 code residue in motif 10 and the uracil base is thus not enclosed in the "Valine sandwich" described by Shen *et al.* (2016) present at the other base binding sites. Indeed, the two flanking Val residues would dehydrate and stabilize the hydrogen bond in motifs 1-9, similar to the hydrophobic dehydration stabilization of hydrogen bonds in the interior of globular proteins Fernández & Scheraga (2003).

The differences between the protein-RNA and the protein-only simulations are striking. In the 10-repeat protein-RNA complex, the protein remained close to the starting crystal structure (less than 4.5 Å IRMSD over the course of the simulation). On the other hand, the series of apo PPR constructs revealed in atomic detail how the non-globular, linear form of repeat-motif domains in PPR proteins introduce significant flexibility to the unbound protein as a whole.

The effect is particularly visible in the 10Rtern construct, consisting of the 10 repeat motifs (together with the capped N- and C-terminal regions). The protein attained IRMSD values of up to 20 Å  $C_{\alpha}$  lmsd with respect to the starting structure. Two factors in particular appear to contribute to these

large amplitude structural changes. First, the starting structure for the simulations is the holo molecule, consisting of protein and bound RNA, and the removal of the RNA results in a non-equilibrium state that relaxes to a more elongated shape of the PPR during the equilibration step of the simulations. The conformational changes we observed are entirely in line with experimental structures of unbound PPR molecules resolved by x-ray crystallography (e.g., Coquille *et al.*, 2014) and a model of the structure in the solution state (Gully *et al.*, 2015).

Second, our results showed that the flexibility of the extended array of repeat motifs confers marked variations in the overall conformation that are not present in globular proteins. Indeed, least-rmsd is a measure that is not well suited to such flexible systems, as it relies on superposition of structures that do not have a clear correspondence (e.g., Cazals & Tetley, 2019, Kufareva & Abagyan, 2012).

Still, the lrmsd permitted comparing the different constructs at the local level of motifs and motifpairs. This analysis showed that whereas the individual motifs varied remarkably little (0.7 Å on average) from the crystal structure in the MD simulations of both the apo PPR and the protein-RNA complex (Fig ??), the motif pairs showed much more variability Fig. 3. This suggests that the shape variability in the extended apo-PPR structures is the result of fluctuations in the motif-motif geometries in the thermal equilibrium simulated by the MD simulations.

Further, analysis of the motif-motif interaction geometries in terms of the equivalent helical twist (or screw) transformation, which can be used to describe any (quasi) rigid-body movement (Angelidis, 2004, Boyer *et al.*, 2015), show how such seemingly small changes in the motif-motif interfaces sampled in the MD simulations can substantially modify the shape of the linear assembly (Fig. ??).

Our results suggest that the shifts of neighboring motifs with respect to each other in the unbound PPR protein allow significant mobility to the side chain of the code residue Asn5 (motif numbering) compared to the RNA-bound state (see Fig. 11). In the crystal structure of the RNA-bound PPR protein, this residue is held in place by a hydrogen bond to Asp35, which allows it to form a second hydrogen bond to the uracil base (Coquille *et al.*, 2014, Shen *et al.*, 2016), and our MD simulations reflect this fact, indicating that this side-chain rotamer is very well-maintained throughout the PPR-RNA simulation.

On the other hand, in the unbound PPR protein, the side chain of this critical residue predominantly adopts the bound-state conformation while remaining free enough to visit alternative rotameric states as well. It is intriguing to hypothesize that a functional role for such fluctuations may be to establish a kinetic barrier to nucleotide base binding, sacrificing some affinity in order to allow higher specificity of interaction with the target mRNA.

## 5.1 Limitations of this work

The simulation timescales explored in this work appear to be largely sufficient to sample the principle mode of the distribution of motif-motif interface geometries. However, the sampling of the largest-amplitude interface fluctuations seen in our simulations are not yet converged, and appear to be on the  $\mu$  or tens of  $\mu$ s timescale, as seen for example in Figure ?? . Further investigations will be required to fully assess the relevance of these transitions.

Molecular dynamics simulations of protein-RNA complexes increasingly provide good correspondence to experimental measurements (e.g., Bochicchio *et al.*, 2018, ?). Still, the state of the art in protein-RNA simulations is in flux, and important contributions to the energetics of atomic interactions in the two types of macromolecules are not yet fully taken into account (e.g., Sponer *et al.*, 2018). For example, in simulations of a 9-motif PUF-RNA complex using protocols similar to those presented here (Krepl *et al.*, 2015), the Sponer group observed progressive unbinding of the RNA over a hundreds-of-ns time scale. In our simulations the PPR complex appears to be stable but not excessively so, as we

observed terminal nucleotide base unbinding and rebinding on the simulation timescale. However it is not possible to directly compare the PUF and PPR systems, which exploit different binding interactions.

## 6 Acknowledgements

We gratefully acknowledge the French CNRS 80|Prime (project DECRYPTOR) and Défi Infinity (project DecryProtARN) programs, as well as the "Initiative d'Excellence" program from the French State (Grant "DYNAMO", ANR-11-LABX-0011-01).

## References

- Angelidis, A. (2004). Hexanions: 6D space for twists. Technical report Oxford University Computing Services.
- Barkan, A., Rojas, M., Fujii, S., Yap, A., Chong, Y. S., Bond, C. S., & Small, I. (2012). A combinatorial amino acid code for RNA recognition by pentatricopeptide repeat proteins. *PLoS Genet* **8**(8), e1002910.
- Barkan, A. & Small, I. (2014). Pentatricopeptide repeat proteins in plants. *Annu Rev Plant Biol* **65**, 415–42.
- Bochicchio, A., Krepl, M., Yang, F., Varani, G., Sponer, J., & Carloni, P. (2018). Molecular basis for the increased affinity of an rna recognition motif with re-engineered specificity: A molecular dynamics and enhanced sampling simulations study. *PLoS Comput Biol* **14**(12), e1006642.
- Boyer, B., Ezelin, J., Poulain, P., Saladin, A., Zacharias, M., Robert, C. H., & Prévost, C. (2015). An integrative approach to the study of filamentous oligomeric assemblies, with application to reca.. *PLoS One* **10**(3), e0116414.
- Cazals, F. & Tetley, R. (2019). Characterizing molecular flexibility by combining least root mean square deviation measures. *Proteins: Structure, Function, and Bioinformatics* **87**(5), 380–389.
- Coquille, S., Filipovska, A., Chia, T., Rajappa, L., Lingford, J. P., Razif, M. F. M., Thore, S., & Rackham, O. (2014). An artificial ppr scaffold for programmable rna recognition. *Nat Commun* **5**, 5729.
- Eberhard, S., Loiselay, C., Drapier, D., Bujaldon, S., Girard-Bascou, J., Kuras, R., Choquet, Y., & Wollman, F.-A. (2011). Dual functions of the nucleus-encoded factor tda1 in trapping and translation activation of atpa transcripts in chlamydomonas reinhardtii chloroplasts. *Plant J* **67**(6), 1055–66.
- Fernández, A. & Scheraga, H. A. (2003). Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. *Proc Natl Acad Sci U S A* **100**(1), 113–8.
- Filipovska, A. & Rackham, O. (2012). Modular recognition of nucleic acids by puf, tale and ppr proteins. *Mol Biosyst* **8**(3), 699–708.
- Gully, B. S., Cowieson, N., Stanley, W. A., Shearston, K., Small, I. D., Barkan, A., & Bond, C. S. (2015). The solution structure of the pentatricopeptide repeat protein PPR10 upon binding atpH RNA. *Nucleic Acids Res* **43**(3), 1918–26.
- Hall, T. M. T. (2016). De-coding and re-coding RNA recognition by PUF and PPR repeat proteins. *Curr Opin Struct Biol* **36**, 116–121.

- Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD - Visual Molecular Dynamics.. *J. Molec. Graphics* **14**, 33–38.
- Hénin, J., Lopes, L. J. S., & Fiorin, G. (2022). Human learning for molecular simulations: The collective variables dashboard in vmd. *J Chem Theory Comput* **18**(3), 1945–1956.
- Jo, S., Kim, T., Iyer, V. G., & Im, W. (2008). Charmm-gui: a web-based graphical user interface for charmm. *J Comput Chem* **29**(11), 1859–65.
- Krepl, M., Havrila, M., Stadlbauer, P., Banas, P., Otyepka, M., Pasulka, J., Stefl, R., & Spöner, J. (2015). Can we execute stable microsecond-scale atomistic simulations of protein-rna complexes?. *J Chem Theory Comput* **11**(3), 1220–43.
- Kufareva, I. & Abagyan, R. (2012). Methods of protein structure comparison. *Methods Mol Biol* **857**, 231–57.
- Lee, J., Cheng, X., Swails, J. M., Yeom, M. S., Eastman, P. K., Lemkul, J. A., Wei, S., Buckner, J., Jeong, J. C., Qi, Y., Jo, S., Pande, V. S., Case, D. A., Brooks, 3rd, C. L., MacKerell, Jr, A. D., Klauda, J. B., & Im, W. (2016). Charmm-gui input generator for namd, gromacs, amber, openmm, and charmm/openmm simulations using the charmm36 additive force field. *J Chem Theory Comput* **12**(1), 405–13.
- Marx, C., Wünsch, C., & Kück, U. (2015). The octatricopeptide repeat protein raa8 is required for chloroplast trans splicing. *Eukaryot Cell* **14**(10), 998–1005.
- Miranda, R. G., Rojas, M., Montgomery, M. P., Gribbin, K. P., & Barkan, A. (2017). Rna-binding specificity landscape of the pentatricopeptide repeat protein ppr10. *RNA* **23**(4), 586–599.
- Nakamura, T., Yagi, Y., & Kobayashi, K. (2012). Mechanistic insight into pentatricopeptide repeat proteins as sequence-specific rna-binding proteins for organellar rnas in plants. *Plant Cell Physiol* **53**(7), 1171–9.
- Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M., & Jensen, J. H. (2011). Propka3: Consistent treatment of internal and surface residues in empirical pka predictions. *J Chem Theory Comput* **7**(2), 525–37.
- Paladin, L., Bevilacqua, M., Errigo, S., Piovesan, D., Mičetić, I., Necci, M., Monzon, A. M., Fabre, M. L., Lopez, J. L., Nilsson, J. F., Rios, J., Menna, P. L., Cabrera, M., Buitron, M. G., Kulik, M. G., Fernandez-Alberti, S., Fornasari, M. S., Parisi, G., Lagares, A., Hirsh, L., Andrade-Navarro, M. A., Kajava, A. V., & Tosatto, S. C. E. (2021). Repeatsdb in 2021: improved data and extended classification for protein tandem repeat structures. *Nucleic Acids Res* **49**(D1), D452–D457.
- Phillips, J. C., Hardy, D. J., Maia, J. D. C., Stone, J. E., Ribeiro, J. V., Bernardi, R. C., Buch, R., Fiorin, G., Hénin, J., Jiang, W., McGreevy, R., Melo, M. C. R., Radak, B. K., Skeel, R. D., Singharoy, A., Wang, Y., Roux, B., Aksimentiev, A., Luthey-Schulten, Z., Kalé, L. V., Schulten, K., Chipot, C., & Tajkhorshid, E. (2020). Scalable molecular dynamics on cpu and gpu architectures with namd. *J Chem Phys* **153**(4), 044130.
- Shen, C., Zhang, D., Guan, Z., Liu, Y., Yang, Z., Yang, Y., Wang, X., Wang, Q., Zhang, Q., Fan, S., Zou, T., & Yin, P. (2016). Structural basis for specific single-stranded RNA recognition by designer pentatricopeptide repeat proteins. *Nat Commun* **7**, 11285.

- Sponer, J., Bussi, G., Krepl, M., Banáš, P., Bottaro, S., Cunha, R. A., Gil-Ley, A., Pinamonti, G., Poblete, S., JureÅka, P., Walter, N. G., & Otyepka, M. (2018). RNA structural dynamics as captured by molecular simulations: A comprehensive overview. *Chemical Reviews* **118**(8), 4177–4338.
- Wang, M., Ogé, L., Perez-Garcia, M.-D., Hamama, L., & Sakr, S. (2018). The puf protein family: Overview on puf rna targets, biological functions, and post transcriptional regulation. *Int J Mol Sci* **19**(2).
- Yin, P., Li, Q., Yan, C., Liu, Y., Liu, J., Yu, F., Wang, Z., Long, J., He, J., Wang, H.-W., Wang, J., Zhu, J.-K., Shi, Y., & Yan, N. (2013). Structural basis for the modular recognition of single-stranded RNA by PPR proteins. *Nature* **504**(7478), 168–71.
- Zhao, Y.-Y., Mao, M.-W., Zhang, W.-J., Wang, J., Li, H.-T., Yang, Y., Wang, Z., & Wu, J.-W. (2018). Expanding rna binding specificity and affinity of engineered puf domains. *Nucleic Acids Res* **46**(9), 4771–4782.



## 2.3 Discussion de l'article 2

### 2.3.1 Interaction ARN-PPR

Grâce à l'analyse de nos simulations, nous avons montré que le complexe protéine-ARNsb est très stable au cours de la simulation. En effet, l'ARN reste en complexe avec la protéine tout le long de la simulation. Ce résultat ne pouvait pas forcément être anticipé car la simulation des complexes protéine-ARN n'est pas encore aussi répandue que les simulations des protéines seules ou dans la membrane. Par exemple, une précédente étude d'une protéine PUF liée à un ARNsb avait montré que l'ARN se détachait progressivement le long d'une simulation plus courte que celles que nous avons effectuées ici. Dans nos simulations, seules les bases des nucléotides U1, U2, U3 et U10 ont tendance à être moins bien fixées à la protéine : on observe ainsi que la fixation de ces bases nucléotidiques a un caractère plus labile que les autres.

Du côté 5' de l'ARNsb, un phosphate est manquant car il a été résolu de façon incomplète dans la structure cristalline (il est présent dans la nature car le site de fixation de la protéine sur l'ARNm n'est pas au bout de ce dernier). Ceci cause possiblement la déstabilisation de la liaison du nucléotide U1 à son site de fixation. Cette déstabilisation peut ensuite se propager aux nucléotides U2 et U3 et causer leur instabilité.

À l'extrémité 3' de l'ARNsb, la fixation plus faible de U10 par rapport aux autres nucléotides n'est pas vraiment surprenante. En effet, bien que le code PPR soit complet (le motif 10 est entier) et les acides aminés 5 et 35 disponibles, il manque une partie du site de liaison puisque le motif 10 est le dernier. Ainsi, le faisceau d'hélices supposé accueillir la base est hétérogène car la première hélice du motif suivant n'existe pas dans notre structure. L'interaction avec les résidus provenant du motif  $i+1$  dans les autres motifs ne peut donc pas se faire ici et le nucléotide est ainsi déstabilisé.

Les simulations par dynamique moléculaire permettent également l'analyse détaillée de la conformation rotamérique de la chaîne latérale du résidu Asn5 du code PPR (Asn5 dans la numérotation relative à chaque motif) et de ses changements le long des simulations. Dans la structure cristalline, ce résidu forme une liaison hydrogène avec la base uracile du nucléotide SHEN et al. 2016 et fait également partie d'une seconde liaison avec l'Asp35, l'autre résidu du code. Les simulations de dynamique moléculaire du complexe PPR-ARNsb indiquent un état rotamérique de la chaîne latérale de Asn5 de chaque motif très majoritaire, et ce même état rotamérique domine aussi dans la protéine apo. Néanmoins dans

la protéine apo, l'état rotamérique de la chaîne latérale de ce résidu se révèle beaucoup plus variable, ce qui suggère l'existence d'un "prix à payer" en terme d'énergie libre lors de la fixation de la base.

### 2.3.2 Comparaison avec la protéine apo

Par ailleurs, on observe de fortes différences dans les comportements de la protéine seule et de la protéine lié à l'ARNsb. En effet, dans la simulation de la protéine liée, on remarque que celle-ci se comporte d'avantage comme une protéine globulaire, c'est-à-dire changeant peu de conformation. On observe l'inverse dans la simulation de la protéine non liée à l'ARNsb : un fort mouvement global. La IRMSD (pour least RMSD ou RMSD minimum) globale de chaque protéine permet de constater la différence de comportement des protéines et les forts mouvements de la protéine non liée (les constructions 10R sont bien représentatives de ce constat, certaines simulations atteignant les 20Å de IRMSD par rapport à la structure initiale). En analysant la IRMSD des motifs composant la protéine, on remarque alors que ce ne sont pas les mouvements des motifs eux-mêmes qui en sont responsables car ils s'avèrent très stables.

On peut expliquer ce changement de conformation de la façon suivante : la structure de départ de toutes nos simulations est une structure liée à l'ARNsb. Lorsqu'on simule la protéine sans cet ARN, il est normal de s'attendre à observer un changement de conformation. De plus, il existe plusieurs structures de protéines PPR sans ARNsb lié et ces structures sont plus ouvertes que la structure initiale dont nous sommes partis (COQUILLE et al. 2014; GULLY et al. 2015). C'est d'ailleurs pour cette raison que nous avons imposé un chauffage très doux à la protéine avant les étapes de simulation à part entière.

Ces structures cristallines d'une forme plus ouverte ne permettent cependant pas de juger de la variabilité de la structure dans le temps. Nos simulations suggèrent que cette variabilité est très conséquente. Il faut d'ailleurs noter que la IRMSD de la protéine entière est peu adaptée et informative dans cette situation car la protéine s'éloigne rapidement de la structure initiale. Les motifs composant la protéine bougent peu (les valeurs de IRMSD des motifs dans les simulations avec et sans ARN sont très similaires : environ 0,7Å) et c'est assez remarquable quand on sait que la IRMSD de la protéine entière varie jusqu'à 20Å.

Pour comprendre le changement conformationnel de la protéine non liée, on peut se tourner vers les paires de motifs. Ceux-ci montrent des variations conséquentes dans leurs mouvements (en moyenne 1,5Å et jusqu'à 3Å) alors que

dans la simulation où la protéine est liée à l'ARNsb, la IRMSD est en moyenne aux alentours de 1,0Å.

Il est difficile d'apprécier l'ampleur des mouvements globaux de la protéine à motifs répétés qui sont impliqués par ces mouvements d'interface motif-motif. Partant de ce constat, nous avons mis en place des analyses utilisant les propriétés hélicoïdales qui permettent de décrire un mouvement en (quasi) corps-rigide entre deux motifs successifs. Nous analysons les interfaces entre les motifs grâce à des analyses plus globales comme le calcul du pitch et de l'angle de rotation d'un motif au suivant. Le pitch correspond à la longueur nécessaire à la protéine pour faire un tour complet de l'axe hélicoïdal si toutes les interfaces entre les motifs étaient identiques à celle analysée (cf. figure 45). Ces analyses montrent que la protéine avec ARN semble avoir deux modes principaux pour le pitch (environ 40Å et 60Å). Sans ARN, il semble aussi y avoir deux modes mais les valeurs de pitch sont plus grandes (60Å et 130Å). Le changement de conformation de la protéine non liée peut donc être expliqué par les mouvements importants des interfaces entre les paires de motifs.

## 2.4 Conclusion et perspectives de l'article 2

La simulation de complexe protéine-ARN est actuellement en expansion mais il s'agit d'un champs récent de la recherche et de la simulation et les champs de force peuvent donc être encore imparfaits. Les résultats obtenus pour la construction PPR-ARN ne doivent alors pas être considérés comme définitifs. Nos études sur les constructions de PPR dans leur forme apo, elles, suggèrent fortement que dans le changement d'état conformationnel de la protéine il n'y a pas une unique forme apo de ces protéines mais une multitude. En effet, la forme de la molécule change constamment en solution. Pourtant, les motifs répétés eux-mêmes changent très peu par rapport à la structure initiale, les valeurs de IRMSD restent très proches des valeurs obtenues dans les simulations du complexe protéine-ARNsb. Ce sont les interfaces entre les paires de motifs qui varient.

Nous avons caractérisé ces mouvements grâce à des paramètres hélicoïdaux qui permettent d'observer leurs effets et leurs importance. Ces études et analyses de la protéine simulée suggèrent que les mouvements des interfaces entre les motifs peuvent jouer un rôle dans la cinétique de la fixation de la protéine à l'ARNsb ainsi que leur rôle potentiel dans la spécificité de la reconnaissance. La continuité de ce travail portera d'ailleurs principalement sur ce sujet.



## Quatrième partie

# Conclusion et perspectives

"Celui qui excelle à résoudre les difficultés les résout avant qu'elles ne surgissent."

*L'art de la guerre*, Sun Tzu (durant la période des Printemps et Automnes de l'histoire de Chine (771-481/453 av. J.C.), traduit du chinois)

"Celui qui n'a pas d'objectifs ne risque pas de les atteindre."

*L'art de la guerre*, Sun Tzu (durant la période des Printemps et Automnes de l'histoire de Chine (771-481/453 av. J.C.), traduit du chinois)



## Conclusion et perspectives

Tout au long de ce manuscrit, nous nous sommes plongés dans un univers microscopique où des familles de protéines adoptant une structure en solénoïde alpha sont porteuses de rôles essentiels au bon fonctionnement des machineries énergétiques cellulaires au sein des mitochondries et des chloroplastes : des rôles dans la régulation de l'expression des gènes des organites. Les objectifs principaux de ce travail de thèse étaient de construire une approche permettant d'identifier de nouvelles protéines à solénoïde alpha candidates dans une optique d'élaboration d'un catalogue des protéines susceptibles de jouer un rôle de régulation de l'expression des gènes des organites en se liant aux ARNm en étant issus ; mais aussi de produire des données de simulation d'un système proche des protéines à solénoïde alpha naturelles afin de tenter de mieux appréhender les mécanismes de la liaison spécifique de la protéine à l'ARNm cible.

Les méthodes présentées dans le premier article ont permis d'identifier des protéines à solénoïde alpha au sein de 48 espèces d'Archaeplastida. Ces méthodes sont efficaces et nous ont permis de constater de grandes variations dans la distribution des différentes familles de protéines à solénoïde alpha connues (comme les protéines OPR et PPR) dans ces espèces. Ceci nous amène donc à penser que ces deux familles de protéines ont subi des expansions conséquentes dans certains groupes d'espèces (les protéines OPR chez les Chlorophyceae ou chez la Streptophyte *Mesostigma viride* par exemple). Cette constatation pourrait refléter une adaptation à un environnement ou à un nouveau mode de vie et il sera nécessaire de comparer les données du catalogue des protéines à solénoïde alpha candidates à la régulation des organites que nous réalisons à des données portant sur les milieux de vie et les caractères physiologiques des espèces que nous étudions. Ce travail est actuellement en cours et se poursuivra également après la thèse.

Par ailleurs, il reste possible d'améliorer encore les approches construites. En effet, nous envisageons d'implémenter l'utilisation d'autres algorithmes d'apprentissage machine (comme SVM ou encore la régression linéaire) mais nous prévoyons aussi de tester l'assouplissement des filtres de la méthode DT pour identifier des protéines candidates plus lointaines ou d'utiliser d'autres propriétés des protéines pour ajouter des filtres et ainsi mieux identifier spécifiquement les protéines à solénoïde alpha. Il s'agira d'un travail qui suivra également la thèse pour éventuellement ajouter certaines nouvelles fonctionnalités à nos méthodes prochainement.

Parallèlement, il existe d'autres lignées d'espèces eucaryotes possédant éga-

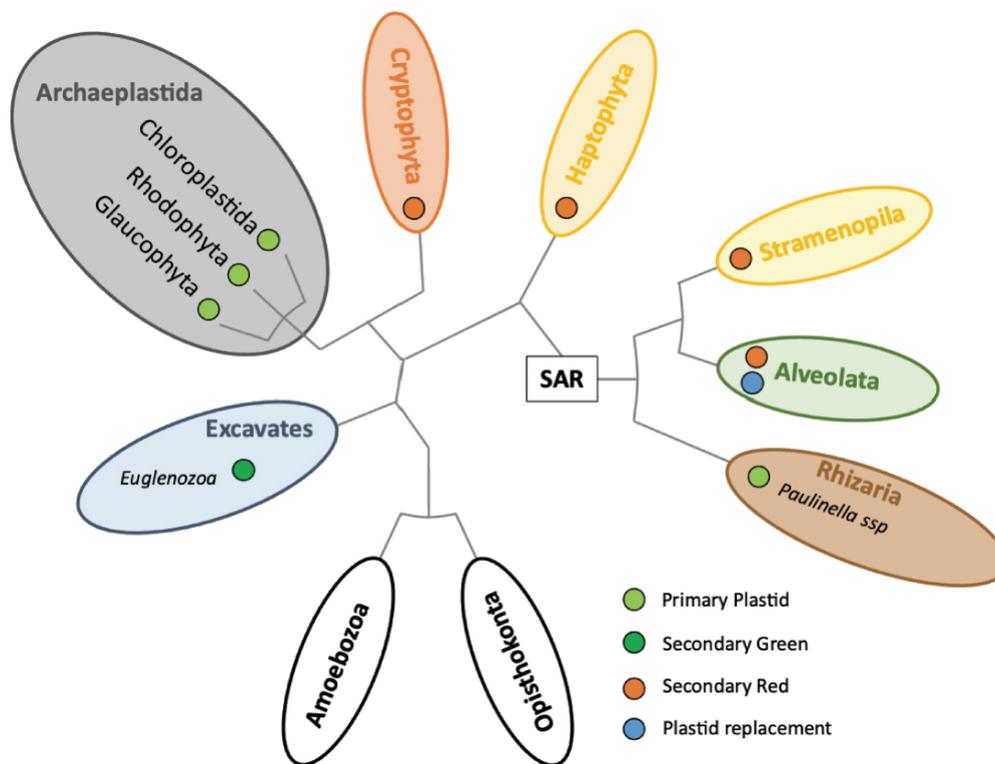


FIGURE 49 – Arbre regroupant différentes lignées eucaryotes photosynthétiques ayant subi une ou plusieurs endosymbioses (cf. FALCIATORE et al. 2022).

lement des plastes et provenant de diverses endosymbioses comme présenté en introduction (cf. figure 49). Il est possible d'appliquer notre approche à ces espèces pour estimer la distribution de nos protéines d'intérêt et proposer un catalogue complet des protéines susceptibles de réguler l'expression des génomes des organites au sein du monde photosynthétique eucaryote. On peut notamment citer l'endosymbiose à l'origine de la lignée des diatomées (Straménopiles) (pour lesquelles il est possible de tester *in vivo* au laboratoire les protéines à solénoïde alpha candidates qui seraient nouvellement identifiées) mais aussi l'endosymbiose spécifique au genre *Paulinella* au sein des Rhizaria car il s'agit d'une endosymbiose primaire qui se rapproche de l'endosymbiose primaire à l'origine du groupe des Archaeplastida.

En outre, explorer la diversité en protéines à solénoïde alpha au sein d'autres espèces permettra de mieux cerner les événements de contraction et d'expansion des différentes familles et ainsi de commencer le travail de reconstruction de l'histoire évolutive de chacune.

Grâce à nos simulations de dynamique moléculaire, nous avons constaté et caractérisé des changements de conformation de la protéine PPR synthétique non liée à l'ARNsb dus aux mouvements des interfaces au sein de chaque paire de motifs. Nous avons émis l'hypothèse qu'il s'agit possiblement d'un phénomène jouant un rôle dans la fixation de l'ARNsb à la protéine et de nouvelles études à ce propos seront mises en place pour mettre à l'épreuve cette hypothèse comme la simulation d'une protéine sur laquelle moins de la moitié des nucléotides sont fixés initialement. En comparaison, les simulations de la protéine liée à l'ARNsb ont permis de montrer la stabilité de la liaison des deux partenaires et la capacité d'une base à se défixer puis se fixer à nouveau.

Trois autres structures cristallines de protéines PPR synthétiques liant un brin d'ARNsb sont disponibles sur la base de données PDB. Celles-ci ressemblent fortement à la structure que nous avons utilisée mais elles diffèrent de deux motifs PPR et deux nucléotides de l'ARNsb au milieu de leurs séquences (cf. figure 50). Ces structures proviennent de la même étude sur le principe du code PPR (SHEN et al. 2016) que celle qui a été utilisée durant ce travail. Les différences entre les quatre structures peuvent donc permettre d'étudier le code PPR en comparant les simulations entre elles, notamment du point de vue de la stabilité de la liaison de l'ARNsb à la protéine et des liaisons impliquées dans la fixation des différents nucléotides.

On note finalement que le code PPR mérite encore d'être perfectionné. En effet, seuls les résidus 5 et 35 ont pour l'instant été démontrés essentiels à la reconnaissance spécifique de l'ARNsb par la protéine. Il ne semble cependant pas déraisonnable de proposer qu'au moins une partie des autres résidus du motif PPR soit aussi impliquée et importante dans la fixation du nucléotide à la protéine. On peut par exemple citer la paire de valines du motif répété de la structure PPR synthétique que nous avons utilisée.

Les deux aspects de ce travail peuvent sembler éloignés l'un de l'autre mais ils se complètent et surtout se rejoignent en un même objectif : décrypter les règles de la régulation de l'expression des génomes au sein des organites. Les connaissances actuelles sur les familles de protéines à solénoïde alpha ont pu être améliorées grâce à ce travail, qui ouvre en plus de nombreuses nouvelles voies de recherche.

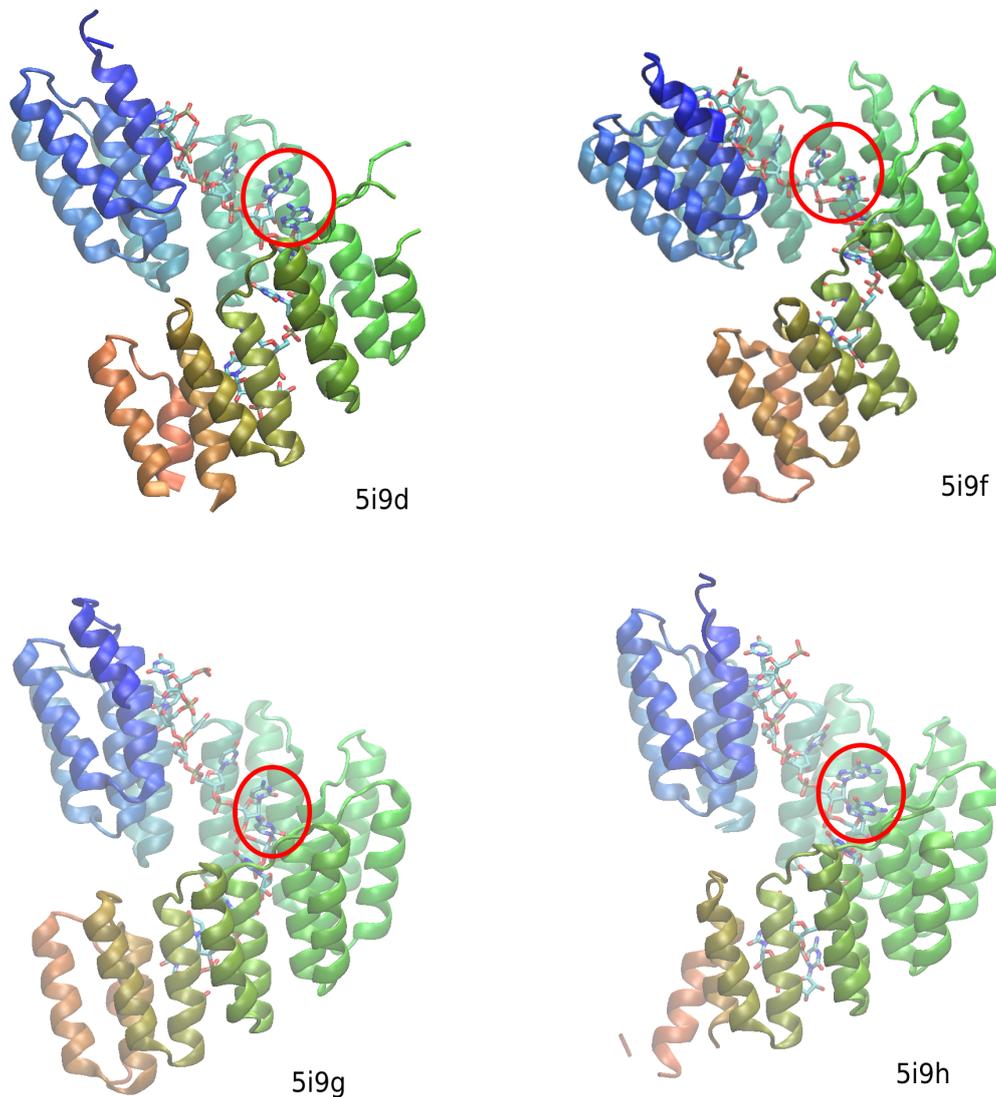


FIGURE 50 – Les quatre structures des protéines synthétiques se lient à des ARNs différents (cf. SHEN et al. 2016). La structure 5i9F correspond à celle utilisée dans ce travail mais il est possible d'utiliser la dynamique moléculaire sur les trois autres. En rouge sont entourés les acides nucléiques qui varient d'une structure à l'autre : 5i9d contient un ARN composé de 4 uraciles, 2 adénines et 4 uraciles ; 5i9f contient un ARN composé de 10 uraciles ; 5i9g contient un ARN composé de 4 uraciles, 2 cytosines et 4 uraciles ; 5i9h contient un ARN composé de 4 uraciles, 2 guanines et 4 uraciles.





## Références

- ALBAREDE, Francis (2009). “Volatile accretion history of the terrestrial planets and dynamic implications”. In : *Nature* 461, p. 1227-1233. DOI : [10.1038/nature08477](https://doi.org/10.1038/nature08477).
- ALBERTS, Bruce et al. (2015). *The RNA World and the Origins of Life*. T. 4th edition. New York : Garland Science. DOI : Available from : <https://www.ncbi.nlm.nih.gov/books/NBK26876/>.
- ALLEN, John F. (2015). “Why chloroplasts and mitochondria retain their own genomes and genetic systems : Colocation for redox regulation of gene expression”. In : *Proceedings of the National Academy of Sciences* 112.33, p. 10231-10238. DOI : [10.1073/pnas.1500012112](https://doi.org/10.1073/pnas.1500012112).
- ALLWOOD, Abigail C. et al. (2006). “Stromatolite Reef from the Early Archaean Era of Australia”. In : *Nature* 441.7094, p. 714-718. URL : <https://doi.org/10.1038/nature04764>.
- ALTSCHUL, Stephen F. et al. (1990). “Basic local alignment search tool”. In : *Journal of Molecular Biology* 215.3, p. 403-410. ISSN : 0022-2836. DOI : [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).
- ANDRADE, Miguel A et al. (2000). “Homology-based method for identification of protein repeats using statistical significance estimates” Edited by J. Thornton”. In : *Journal of Molecular Biology* 298.3, p. 521-537. DOI : <https://doi.org/10.1006/jmbi.2000.3684>.
- ANET, Frank AL. (2004). “The Place of Metabolism in the Origin of Life”. In : *Current Opinion in Chemical Biology* 8.6, p. 654-659. URL : <https://doi.org/10.1016/j.cbpa.2004.10.005>.
- ANGELIDIS, A. (2004). Rapp. tech.
- ANTHONISEN, Inger Lill, Maria L. SALVADOR et Uwe KLEIN (2001). “Specific sequence elements in the 5’ untranslated regions of rbcL and atpB gene mRNAs stabilize transcripts in the chloroplast of *Chlamydomonas reinhardtii*.” In : *RNA*, p. 1024-1033. DOI : [10.1017/s1355838201001479](https://doi.org/10.1017/s1355838201001479).
- ARCHER, E. Kathleen et Kenneth KEEGSTRA (1990). “Current views on chloroplast protein import and hypotheses on the origin of the transport mechanism”. In : *Journal of Bioenergetics and Biomembranes*. URL : <https://doi.org/10.1007/BF00786931>.
- ARCHIBALD, John M. (2015). “Endosymbiosis and Eukaryotic Cell Evolution”. In : *Current Biology* 25.19. URL : <https://doi.org/10.1016/j.cub.2015.07.055>.
- ASTATOURIAN, Alexis (2021). “Détection des protéines alpha-solénoides impliquées dans la régulation du génome des organites : amélioration de la détection et application aux protéomes de diatomées”. In : *Mémoire de master*.

- AUBOURG, Sébastien et al. (2000). “In *Arabidopsis thaliana*, 1% of the genome codes for a novel protein family unique to plants.” In : *Plant molecular biology* 42.4, p. 603-613. DOI : 10.1023/a:1006352315928..
- BAHR, Anne et al. (2001). “BALiBASE (Benchmark Alignment dataBASE) : enhancements for repeats, transmembrane sequences and circular permutations”. In : *Nucleic Acids Research* 29.1, p. 323-326. DOI : 10.1093/nar/29.1.323.
- BAI, Yun et al. (2007). “Crystal Structure of Murine CstF-77 : Dimeric Association and Implications for Polyadenylation of mRNA Precursors”. In : *Molecular Cell* 25.6, p. 863-875. DOI : <https://doi.org/10.1016/j.molcel.2007.01.034>.
- BALCH, William E. et al. (1977). “An ancient divergence among the bacteria”. In : *Journal of Molecular Evolution* 9. DOI : <https://doi.org/10.1007/BF01796092>.
- BARKAN, Alice, Margarita ROJAS et al. (2012). “A Combinatorial Amino Acid Code for RNA Recognition by Pentatricopeptide Repeat Proteins”. In : *PLOS Genetics* 8.8, p. 1-8. DOI : 10.1371/journal.pgen.1002910. URL : <https://doi.org/10.1371/journal.pgen.1002910>.
- BARKAN, Alice et SMALL (2014). “Pentatricopeptide Repeat Proteins in Plants”. In : *Annual Review of Plant Biology* 65.1, p. 415-442. DOI : 10.1146/annurev-arplant-050213-040159.
- BECKER, Sidney et al. (2019). “Unified Prebiotically Plausible Synthesis of Pyrimidine and Purine RNA Ribonucleotides”. In : *Science* 366.6461, p. 76-82. DOI : <https://doi.org/10.1126/science.aax2747>.
- BELL, Elizabeth A. et al. (2015). “Potentially biogenic carbon preserved in a 4.1 billion-year-old zircon”. In : *Proceedings of the National Academy of Sciences* 112.47, p. 14518-14521. DOI : 10.1073/pnas.1517557112. eprint : <https://www.pnas.org/doi/pdf/10.1073/pnas.1517557112>. URL : <https://www.pnas.org/doi/abs/10.1073/pnas.1517557112>.
- BERCEL, T.L. et S.A. KRANZ (2019). “Insights into carbon acquisition and photosynthesis in *Karenia brevis* under a range of CO<sub>2</sub> concentrations”. In : *Progress in Oceanography* 172, p. 65-76. DOI : <https://doi.org/10.1016/j.pocean.2019.01.011>.
- BHATTACHARYA, Debashish, Hwan Su YOON et Jeremiah D. HACKETT (2004). “Photosynthetic eukaryotes unite : endosymbiosis connects the dots”. In : *BioEssays* 26.1, p. 50-60. DOI : <https://doi.org/10.1002/bies.10376>.
- BIEGERT, Andreas et Johannes SODING (2008). “De novo identification of highly diverged protein repeats by probabilistic consistency”. In : *Bioinformatics* 24.6, p. 807-814. DOI : 10.1093/bioinformatics/btn039.
- BJORKHOLM, Patrik et al. (2015). “Mitochondrial genomes are retained by selective constraints on protein targeting”. In : *Proceedings of the National Academy of Sciences* 112.33, p. 10154-10161. DOI : 10.1073/pnas.1421372112.

- BOHN, Jennifer A et al. (nov. 2017). “Identification of diverse target RNAs that are functionally regulated by human Pumilio proteins”. In : *Nucleic Acids Research* 46.1, p. 362-386. DOI : 10.1093/nar/gkx1120.
- BOTTKE, William F. et Marc D. NORMAN (2017). “The Late Heavy Bombardment”. In : *Annual Review of Earth and Planetary Sciences* 45.1, p. 619-647. URL : <https://doi.org/10.1146/annurev-earth-063016-020131>.
- BOULOUIS, Alix, Dominique DRAPIER et al. (2015). “Spontaneous Dominant Mutations in Chlamydomonas Highlight Ongoing Evolution by Gene Diversification”. In : *The Plant Cell* 27.4, p. 984-1001. DOI : 10.1105/tpc.15.00010.
- BOULOUIS, Alix, Cécile RAYNAUD et al. (2011). “The Nucleus-Encoded trans-Acting Factor MCA1 Plays a Critical Role in the Regulation of Cytochrome f Synthesis in Chlamydomonas Chloroplasts”. In : *The Plant Cell* 23.1, p. 333-349. DOI : 10.1105/tpc.110.078170.
- BOYER, Benjamin et al. (2015). “An Integrative Approach to the Study of Filamentous Oligomeric Assemblies, with Application to RecA”. In : *PLOS ONE* 10, p. 1-25. DOI : 10.1371/journal.pone.0116414.
- BURKI, Fabien et al. (2020). “The New Tree of Eukaryotes”. In : *Trends in Ecology Evolution* 35.1, p. 43-55. DOI : <https://doi.org/10.1016/j.tree.2019.08.008>.
- BYRNES, James et al. (2016). “Base Flipping by MTERF1 Can Accommodate Multiple Conformations and Occurs in a Stepwise Fashion”. In : *Journal of Molecular Biology* 428.12, p. 2542-2556. DOI : <https://doi.org/10.1016/j.jmb.2015.10.021>.
- CALDER, K. M. et J. E. MCEWEN (1991). “Deletion of the COX7 gene in *Saccharomyces cerevisiae* reveals a role for cytochrome c oxidase subunit VII in assembly of remaining subunits”. In : *Molecular Microbiology* 5.7, p. 1769-1777. DOI : <https://doi.org/10.1111/j.1365-2958.1991.tb01926.x>.
- CÁMARA, Yolanda et al. (2011). “MTERF4 Regulates Translation by Targeting the Methyltransferase NSUN4 to the Mammalian Mitochondrial Ribosome”. In : *Cell Metabolism* 13.5, p. 527-539. DOI : <https://doi.org/10.1016/j.cmet.2011.04.002>.
- CANSIZOGLU, Ahmet E. et Yuh Min CHOOK (2007). “Conformational Heterogeneity of Karyopherin2 Is Segmental”. In : *Structure* 15.11, p. 1431-1441. DOI : <https://doi.org/10.1016/j.str.2007.09.009>.
- CAVAIUOLO, Marina et al. (2017). “Small RNA profiling in Chlamydomonas : insights into chloroplast RNA metabolism”. In : *Nucleic Acids Research* 45.18, p. 10783-10799. DOI : 10.1093/nar/gkx668.
- CEA (2012). “Le Soleil, Notre Etoile”. In : URL : <https://www.cea.fr/comprendre/Pages/matiere-univers/soleil.aspx>.

- CEBALLOS, Gerardo, Paul R. EHRLICH, Anthony D. BARNOSKY et al. (2015). “Accelerated modern human-induced species losses : Entering the sixth mass extinction”. In : *Science Advances* 1.5. DOI : [10.1126/sciadv.1400253](https://doi.org/10.1126/sciadv.1400253).
- CEBALLOS, Gerardo, Paul R. EHRLICH et Rodolfo DIRZO (2017). “Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines”. In : *Proceedings of the National Academy of Sciences* 114.30, E6089-E6096. DOI : [10.1073/pnas.1704949114](https://doi.org/10.1073/pnas.1704949114).
- CHADEE, Avesh et al. (2021). “The Complementary Roles of Chloroplast Cyclic Electron Transport and Mitochondrial Alternative Oxidase to Ensure Photosynthetic Performance”. In : *Frontiers in Plant Science* 12. DOI : [10.3389/fpls.2021.748204](https://doi.org/10.3389/fpls.2021.748204).
- CHAKRAVARTY, Devlina et al. (2013). “Reassessing buried surface areas in protein–protein complexes”. In : *Protein Science* 22.10, p. 1453-1457. DOI : <https://doi.org/10.1002/pro.2330>.
- CHATTON, Edouard (1925). *Pansporella perplexa : amoebien à spores protégées parasite des daphnies : réflexions sur la biologie et la phylogénie des protozoaires*. Masson.
- CHEN, Guanglong et al. (2018). “Genome-wide analysis of the rice PPR gene family and their expression profiles under different stress treatments”. In : *BMC Genomics*. DOI : [10.1186/s12864-018-5088-9](https://doi.org/10.1186/s12864-018-5088-9).
- CHENG, Shifeng et al. (2016). “Redefining the structural motifs that determine RNA binding and RNA editing by pentatricopeptide repeat proteins in land plants”. In : *The Plant Journal* 85.4, p. 532-547. DOI : <https://doi.org/10.1111/tpj.13121>.
- CHOQUET, Yves, David B. STERN et al. (1998). “Translation of cytochrome f is autoregulated through the 5' untranslated region of petA mRNA in Chlamydomonas chloroplasts.” In : *Proceedings of the National Academy of Sciences of the United States of America*, p. 4380-4385. DOI : [10.1073/pnas.95.8.4380](https://doi.org/10.1073/pnas.95.8.4380).
- CHOQUET, Yves et Francis-André WOLLMAN (2002). “Translational regulations as specific traits of chloroplast gene expression”. In : *FEBS Letters* 529.1, p. 39-42. DOI : [https://doi.org/10.1016/S0014-5793\(02\)03260-X](https://doi.org/10.1016/S0014-5793(02)03260-X).
- (2009). “Chapter 29 - The CES Process”. In : *The Chlamydomonas Sourcebook (Second Edition)*. Sous la dir. d'Elizabeth H. HARRIS, David B. STERN et George B. WITMAN. Second Edition. London : Academic Press, p. 1027-1063. DOI : <https://doi.org/10.1016/B978-0-12-370873-1.00037-X>.
- CHOQUET et al. (1988). “Mutant phenotypes support a trans-splicing mechanism for the expression of the tripartite psaA gene in the C. reinhardtii chloroplast”. In : *Cell* 52.6, p. 903-913. DOI : [https://doi.org/10.1016/0092-8674\(88\)90432-1](https://doi.org/10.1016/0092-8674(88)90432-1).

- CHOTEWUTMONTRI, P., K. HOLBROOK et B.D. BRUCE (2017). “Chapter Six - Plastid Protein Targeting : Preprotein Recognition and Translocation”. In : sous la dir. de Lorenzo GALLUZZI. T. 330. *International Review of Cell and Molecular Biology*. Academic Press, p. 227-294. DOI : <https://doi.org/10.1016/bs.ircmb.2016.09.006>.
- CLINE, Sara G., Isaac A. LAUGHBAUM et Patrice P. HAMEL (2017). “CCS2, an Octatricopeptide-Repeat Protein, Is Required for Plastid Cytochrome c Assembly in the Green Alga *Chlamydomonas reinhardtii*”. In : *Frontiers in Plant Science* 8. DOI : [10.3389/fpls.2017.01306](https://doi.org/10.3389/fpls.2017.01306).
- COFFIN, John W. et al. (1997). “The *Neurospora crassa* cya-5 nuclear gene encodes a protein with a region of homology to the *Saccharomyces cerevisiae* PET309 protein and is required in a post-transcriptional step for the expression of the mitochondrially encoded COXI protein”. In : *Current Genetics* 32.4, p. 273-280. DOI : <https://doi.org/10.1007/s002940050277>.
- COMTE, Denis et al. (2023). “Glycine Peptide Chain Formation in the Gas Phase via Unimolecular Reactions”. In : *The Journal of Physical Chemistry A* 127.3, p. 775-780. URL : <https://doi.org/10.1021/acs.jpca.2c08248>.
- COQUILLE, Sandrine et al. (2014). “An artificial PPR scaffold for programmable RNA recognition”. In : *Nature Communications* 5.1. DOI : <https://doi.org/10.1038/ncomms6729>.
- DABERDAKU, Sebastian (2018). “Identification of protein pockets and cavities by Euclidean Distance Transform”. In.
- DACKS, Joel B. et al. (2016). “The changing view of eukaryogenesis – fossils, cells, lineages and how they all come together”. In : *Journal of Cell Science* 129.20, p. 3695-3703. URL : <https://doi.org/10.1242/jcs.178566>.
- DANIEL, Vincent (2003). “Le rayonnement thermique et la loi du Corps Noir”. In : *ENS de Lyon*. URL : <https://planet-terre.ens-lyon.fr/ressource/bilan-radiatif-terre1.xml>.
- DELANNOY, E. et al. (2007). “Pentatricopeptide repeat (PPR) proteins as sequence-specificity factors in post-transcriptional processes in organelles”. In : *Biochemical Society Transactions* 35.6, p. 1643-1647. DOI : [10.1042/BST0351643](https://doi.org/10.1042/BST0351643).
- DELAYE, Luis, Cecilio VALADEZ-CANO et Bernardo PÉREZ-ZAMORANO (2016). “How Really Ancient Is *Paulinella Chromatophora*?” In : *PLoS currents* 129.20, p. 3695-3703. URL : <https://doi.org/10.1371/currents.tol.e68a099364bb1a1e129a17b4e06b0c6b>.
- DEMIRDJIAN, Hagop (2007). In : URL : <https://culturesciences.chimie.ens.fr/thematiques/chimie-physique/photochimie/chimie-atmospherique-1-absorption-des-uv-par-l-ozone>.

- DRAPIER, Dominique, Jacqueline GIRARD-BASCOU et Francis-André WOLLMAN (1992). “Evidence for Nuclear Control of the Expression of the atpA and atpB Chloroplast Genes in Chlamydomonas.” In : *The Plant Cell* 4.3, p. 283-295. DOI : [10.1105/tpc.4.3.283](https://doi.org/10.1105/tpc.4.3.283).
- DRAPIER, Dominique, Blandine RIMBAULT et al. (2007). “Intertwined translational regulations set uneven stoichiometry of chloroplast ATP synthase subunits”. In : *The EMBO Journal* 26.15, p. 3581-3591. DOI : <https://doi.org/10.1038/sj.emboj.7601802>.
- EBERHARD, Stephan et al. (2011). “Dual functions of the nucleus-encoded factor TDA1 in trapping and translation activation of atpA transcripts in Chlamydomonas reinhardtii chloroplasts”. In : *The Plant Journal* 67.6, p. 1055-1066. DOI : <https://doi.org/10.1111/j.1365-313X.2011.04657.x>.
- FALCIATORE, Angela et al. (2022). “Light-driven processes : key players of the functional biodiversity in microalgae”. In : *Comptes Rendus. Biologies* 345.2, p. 15-38. DOI : [10.5802/crbio1.80](https://doi.org/10.5802/crbio1.80).
- FAUCHÈRE, J. et Vladimir PLISKA (1983). “Hydrophobic parameters II of amino acid side-chains from the partitioning of N-acetyl-amino acid amides”. In : *Eur. J. Med. Chem.* 18, p. 369-375.
- FENG, Changli, Quan ZOU et WANG (2021). “Using a low correlation high orthogonality feature set and machine learning methods to identify plant pentatricopeptide repeat coding gene/protein”. In : *Neurocomputing* 424, p. 246-254. ISSN : 0925-2312. DOI : <https://doi.org/10.1016/j.neucom.2020.02.079>.
- FIGUEROA, Rosa Isabel et al. (2009). “The Life History and Cell Cycle of Kryptoperidinium foliaceum, A Dinoflagellate with Two Eukaryotic Nuclei”. In : *Protist* 160.2, p. 285-300. DOI : <https://doi.org/10.1016/j.protis.2008.12.003>.
- FISK, Dianna G., Macie B. WALKER et Alice BARKAN (1999). “Molecular cloning of the maize gene crp1 reveals similarity between regulators of mitochondrial and chloroplast gene expression”. In : *The EMBO Journal* 18.9, p. 2621-2630. DOI : <https://doi.org/10.1093/emboj/18.9.2621>.
- FOURNIER, David et al. (2013). “Functional and Genomic Analyses of Alpha-Solenoid Proteins”. In : *PLOS ONE* 8.11, p. 1-13. DOI : <https://doi.org/10.1371/journal.pone.0079894>.
- GÁNTI, Tibor (2003). *The principles of life*. Oxford University Press.
- GAO, Haishan et al. (2012). “Crystal structure of a TALE protein reveals an extended N-terminal DNA binding region”. In : *Cell Research* 22.12, p. 1716-1720. DOI : <https://doi.org/10.1038/cr.2012.156>.
- GARGAUD, Muriel et al. (2009). *Le Soleil, la terre... la vie. La quête des origines*.
- GARRIDO, Clotilde (déc. 2021). “De l’origine des peptides d’adressage aux organites (mitochondries et chloroplastes)”. Theses. Sorbonne Université. URL : <https://theses.hal.science/tel-03905223>.

- GARRIDO, Clotilde et al. (2020). “Evidence Supporting an Antimicrobial Origin of Targeting Peptides to Endosymbiotic Organelles”. In : *Cells* 9.8. URL : <https://www.mdpi.com/2073-4409/9/8/1795>.
- GEORGE, Richard A. et Jaap HERINGA (2000). “The REPRO server : finding protein internal sequence repeats through the Web”. In : *Trends in Biochemical Sciences* 25.10, p. 515-517. DOI : [https://doi.org/10.1016/S0968-0004\(00\)01643-1](https://doi.org/10.1016/S0968-0004(00)01643-1).
- GOLDSCHMIDT-CLERMONT et al. (1990). “Trans-splicing mutants of *Chlamydomonas reinhardtii*”. In : *Molecular and General Genetics MGG* 223, p. 417-425. DOI : <https://doi.org/10.1007/BF00264448>.
- GRANT, Barth et Iva GREENWALD (1996). “The *Caenorhabditis elegans* sel-1 Gene, a Negative Regulator of lin-12 and glp-1, Encodes a Predicted Extracellular Protein”. In : *Genetics* 143.1, p. 237-247. DOI : [10.1093/genetics/143.1.237](https://doi.org/10.1093/genetics/143.1.237).
- GREENWOOD, Richard C. et al. (2018). “Oxygen isotopic evidence for accretion of Earth’s water before a high-energy Moon-forming giant impact”. In : *Science Advances* 4.3, eaao5928. DOI : [10.1126/sciadv.aao5928](https://doi.org/10.1126/sciadv.aao5928).
- GREY, Aubrey D.N.J. de (2005). “Forces maintaining organellar genomes : is any as strong as genetic code disparity or hydrophobicity?” In : *BioEssays* 27.4, p. 436-446. URL : <https://doi.org/10.1002/bies.20209>.
- GRUBER, Markus, Johannes SÖDING et Andrei N. LUPAS (2005). “REPPER—repeats and their periodicities in fibrous proteins”. In : *Nucleic Acids Research* 33. DOI : [10.1093/nar/gki405](https://doi.org/10.1093/nar/gki405).
- GUHARROY, Mainak, Joël JANIN et Charles H. ROBERT (2010). “Side-chain rotamer transitions at protein-protein interfaces”. In : *PROTEINS : Structure, Function, and Bioinformatics* 78, p. 3219-3225. DOI : [10.1002/prot.22821](https://doi.org/10.1002/prot.22821).
- GULLY, Benjamin S. et al. (2015). “The solution structure of the pentatricopeptide repeat protein PPR10 upon binding atpH RNA”. In : *Nucleic Acids Research*. DOI : [10.1093/nar/gkv027](https://doi.org/10.1093/nar/gkv027).
- GUTMANN, Bernard et al. (2020). “The Expansion and Diversification of Pentatricopeptide Repeat RNA-Editing Factors in Plants”. In : *Molecular Plant* 13.2, p. 215-230. DOI : <https://doi.org/10.1016/j.molp.2019.11.002>.
- HAMMANI, Kamel, Géraldine BONNARD et al. (2014). “Helical repeats modular proteins are major players for organelle gene expression”. In : *Biochimie* 100, p. 141-150. DOI : <https://doi.org/10.1016/j.biochi.2013.08.031>.
- HAMMANI, Kamel, William B. COOK et Alice BARKAN (2012). “RNA binding and RNA remodeling activities of the half-a-tetratricopeptide (HAT) protein HCF107 underlie its effects on gene expression”. In : *Proceedings of the National Academy of Sciences* 109.15, p. 5651-5656. DOI : [10.1073/pnas.1200318109](https://doi.org/10.1073/pnas.1200318109).

- HARRISON, Thomas et al. (2016). “aPPRove : An HMM-Based Method for Accurate Prediction of RNA-Pentatricopeptide Repeat Protein Binding Events”. In : *PLOS ONE* 11. DOI : [10.1371/journal.pone.0160645](https://doi.org/10.1371/journal.pone.0160645).
- HASHIMOTO, Mihoko et al. (2003). “A nucleus-encoded factor, CRR2, is essential for the expression of chloroplast *ndhB* in Arabidopsis”. In : *The Plant Journal* 36.4, p. 541-549.
- HAWKING, Stephen (2018). *Brèves réponses aux grandes questions*. Éditions Odile Jacob.
- HEGER, Andreas et Liisa HOLM (2000). “Rapid automatic detection and alignment of repeats in protein sequences”. In : *Proteins : Structure, Function, and Bioinformatics* 41.2, p. 224-237. DOI : [https://doi.org/10.1002/1097-0134\(20001101\)41:2\\$<224::AID-PROT70>3.0.CO;2-Z\\$](https://doi.org/10.1002/1097-0134(20001101)41:2$<224::AID-PROT70>3.0.CO;2-Z$).
- HILLEBRAND, Arne et al. (2018). “Identification of clustered organellar short (cos) RNAs and of a conserved family of organellar RNA-binding proteins, the heptatricopeptide repeat proteins, in the malaria parasite”. In : *Nucleic Acids Research* 46.19, p. 10417-10431. DOI : [10.1093/nar/gky710](https://doi.org/10.1093/nar/gky710).
- HIRAKAWA, Yoshihisa et al. (2011). “Morphological diversity between culture strains of a chlorarachniophyte, *Lotharella globosa*.” In : *PloS one*. DOI : [10.1371/journal.pone.0023193](https://doi.org/10.1371/journal.pone.0023193).
- HIRSH, Layla et al. (2018). “RepeatsDB-lite : a web server for unit annotation of tandem repeat proteins”. In : *Nucleic Acids Research* 46.W1, W402-W407. DOI : [10.1093/nar/gky360](https://doi.org/10.1093/nar/gky360).
- HUD, Nicholas V. et David M. FIALHO (2019). “RNA Nucleosides Built in One Prebiotic Pot”. In : *Science* 366.6461, p. 32-33. DOI : <https://doi.org/10.1126/science.aaz1130>.
- HUG, Laura A. et al. (2016). “A new view of the tree of life”. In : *Nature Microbiology* 1. DOI : <https://doi.org/10.1038/nmicrobiol.2016.48>.
- IPGP (2018). “Le champ magnétique de la Terre”. In : URL : <https://www.ipgp.fr/fr/obsmag/champ-magnetique-de-terre>.
- ISHIKAWA, Masakazu, Hiroshi SHIMIZU et al. (2016). “Two-step evolution of endosymbiosis between hydra and algae”. In : *Molecular Phylogenetics and Evolution* 103, p. 19-25. ISSN : 1055-7903. DOI : <https://doi.org/10.1016/j.ympev.2016.07.010>.
- ISHIKAWA, Masakazu, Ikuko YUYAMA et al. (2016). “Different Endosymbiotic Interactions in Two Hydra Species Reflect the Evolutionary History of Endosymbiosis”. In : *Genome Biology and Evolution* 8.7, p. 2155-2163. DOI : [10.1093/gbe/evw142](https://doi.org/10.1093/gbe/evw142).
- JANIN, Joël et al. (1978). “Conformation of amino acid side-chains in proteins”. In : *Journal of Molecular Biology* 125.3, p. 357-386. DOI : [https://doi.org/10.1016/0022-2836\(78\)90408-4](https://doi.org/10.1016/0022-2836(78)90408-4).

- JARRIGE, Domitille (déc. 2019). “Deciphering the ”OPR code” to further assess the physiological role of OPR proteins”. Theses. Sorbonne Université. URL : <https://theses.hal.science/tel-03349187>.
- JENKINS, Huw T., Rosanna BAKER-WILDING et Thomas A. EDWARDS (2009). “Structure and RNA binding of the mouse Pumilio-2 Puf domain”. In : *Journal of Structural Biology* 167.3, p. 271-276. DOI : <https://doi.org/10.1016/j.jsb.2009.06.007>.
- JENSEN et al. (1986). “Biogenesis of photosystem II complexes : transcriptional, translational, and posttranslational regulation.” In : *The Journal of cell biology*, p. 1315-1325. DOI : 10.1083/jcb.103.4.1315..
- JOHNSON, Lisa K., Harriet ALEXANDER et C. Titus BROWN (2018). “Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes”. In : *GigaScience* 8.4. DOI : <https://doi.org/10.1093/gigascience/giy158>.
- KEELING, Patrick J. (2004). “Diversity and evolutionary history of plastids and their hosts”. In : *American Journal of Botany* 91.10, p. 1481-1493. DOI : <https://doi.org/10.3732/ajb.91.10.1481>.
- KIVELSON, Margaret G., Krishan K. KHURANA et Martin VOLWERK (2002). “The Permanent and Inductive Magnetic Moments of Ganymede”. In : *Icarus* 157.2, p. 507-522. DOI : <https://doi.org/10.1006/icar.2002.6834>.
- KLEINKNECHT, Laura et al. (2014). “RAP, the Sole Octotricopeptide Repeat Protein in Arabidopsis, Is Required for Chloroplast 16S rRNA Maturation”. In : *The Plant Cell* 26.2, p. 777-787. DOI : 10.1105/tpc.114.122853.
- KLOECKENER-GRUISSEM, Barbara, J. E. MCEWEN et Robert O. POYTON (1987). “Nuclear functions required for cytochrome c oxidase biogenesis in *Saccharomyces cerevisiae* : multiple trans-acting nuclear genes exert specific effects on expression of each of the cytochrome c oxidase subunits encoded on mitochondrial DNA”. In : *Current Genetics* 12, p. 311-322. DOI : <https://doi.org/10.1007/BF00405753>.
- KOTERA, Emi, Masao TASAKA et Toshiharu SHIKANAI (2005). “A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts”. In : *Nature*, p. 326-330. DOI : <https://doi.org/10.1038/nature03229>.
- KUCHKA, GOLDSCHMIDT-CLERMONT et al. (1989). “Mutation at the *Chlamydomonas* nuclear NAC2 locus specifically affects stability of the chloroplast psbD transcript encoding polypeptide D2 of PS II”. In : *Cell* 58.5, p. 869-876. DOI : [https://doi.org/10.1016/0092-8674\(89\)90939-2](https://doi.org/10.1016/0092-8674(89)90939-2).
- KUCHKA, MAYFIELD et ROCHAIX (1988). “Nuclear mutations specifically affect the synthesis and/or degradation of the chloroplast-encoded D2 polypeptide of photosystem II in *Chlamydomonas reinhardtii*”. In : *The EMBO Journal* 7.2, p. 319-324. DOI : <https://doi.org/10.1002/j.1460-2075.1988.tb02815.x>.

- KURTZ, S et C SCHLEIERMACHER (1999). “REPuter : fast computation of maximal repeats in complete genomes.” In : *Bioinformatics* 15.5, p. 426-427. DOI : 10.1093/bioinformatics/15.5.426.
- LAURO, Sebastian E. et al. (2022). “Using MARSIS signal attenuation to assess the presence of South Polar Layered Deposit subglacial brines”. In : *Nature Communications* 13, p. 2041-1723. DOI : <https://doi.org/10.1038/s41467-022-33389-4>.
- LEACH, Andrew (2001). *Molecular modeling : Principles and application*.
- LIGHTOWLERS, Robert N. et Zofia M. A. CHRZANOWSKA-LIGHTOWLERS (2013). “Human pentatricopeptide proteins”. In : *RNA Biology* 10.9, p. 1433-1438. DOI : 10.4161/rna.24770.
- LIPINSKI, Kamil A. et al. (2011). “Revisiting the Yeast PPR Proteins—Application of an Iterative Hidden Markov Model Algorithm Reveals New Members of the Rapidly Evolving Family”. In : *Molecular Biology and Evolution* 28.10, p. 2935-2948. DOI : 10.1093/molbev/msr120.
- LOISELAY, Christelle et al. (2008). “Molecular Identification and Function of *cis*- and *trans*-Acting Determinants for *petA* Transcript Stability in *Chlamydomonas reinhardtii* Chloroplasts”. In : *Molecular and Cellular Biology* 28.17, p. 5529-5542. DOI : 10.1128/MCB.02056-07.
- LOWN, F. J., A. T. WATSON et S. PURTON (2001). “Chlamydomonas nuclear mutants that fail to assemble respiratory or photosynthetic electron transfer complexes”. In : *Biochemical Society Transactions* 29.4, p. 452-455. DOI : 10.1042/bst0290452.
- LURIN, Claire et al. (2004). “Genome-Wide Analysis of Arabidopsis Pentatricopeptide Repeat Proteins Reveals Their Essential Role in Organelle Biogenesis[W]”. In : *The Plant Cell* 16.8, p. 2089-2103. DOI : 10.1105/tpc.104.022236.
- MACEDO-OSORIO, Karla S., Agustino MARTÍNEZ-ANTONIO et Jesús A. BADILLO-CORONA (2021). “Pas de Trois : An Overview of Penta-, Tetra-, and Octo-Tricopeptide Repeat Proteins From *Chlamydomonas reinhardtii* and Their Role in Chloroplast Gene Expression”. In : *Frontiers in Plant Science* 12. DOI : 10.3389/fpls.2021.775366.
- MAIRE, Justin, Linda L. BLACKALL et Madeleine J. H. van OPPEN (2021). “Intracellular Bacterial Symbionts in Corals : Challenges and Future Directions”. In : *Microorganisms* 9.11. DOI : 10.3390/microorganisms9112209.
- MALDE, Ketil et Tomasz FURMANEK (2013). “Increasing Sequence Search Sensitivity with Transitive Alignments”. In : *PLOS ONE* 8.2, p. 1-7. DOI : 10.1371/journal.pone.0054422.

- MANNA, Sam (2015). “An overview of pentatricopeptide repeat proteins and their applications”. In : *Biochimie* 113, p. 93-99. DOI : <https://doi.org/10.1016/j.biochi.2015.04.004>.
- MANTHEY, G. M. et J. E. MCEWEN (1995). “The product of the nuclear gene PET309 is required for translation of mature mRNA and stability or production of intron-containing RNAs derived from the mitochondrial COX1 locus of *Saccharomyces cerevisiae*.” In : *The EMBO Journal* 14.16, p. 4031-4043. DOI : <https://doi.org/10.1002/j.1460-2075.1995.tb00074.x>.
- MARIN, Birger, Eva C. M. NOWACK et Michael MELKONIAN (2005). “A Plastid in the Making : Evidence for a Second Primary Endosymbiosis”. In : *Protist* 156.4, p. 425-432. DOI : <https://doi.org/10.1016/j.protis.2005.09.001>.
- MARSELLA, Luca et al. (2009). “REPETITA : detection and discrimination of the periodicity of protein solenoid repeats by discrete Fourier transform”. In : *Bioinformatics* 25.12. DOI : [10.1093/bioinformatics/btp232](https://doi.org/10.1093/bioinformatics/btp232).
- MCCUTCHEON, John P., Bret M. BOYD et Colin DALE (2019). “The Life of an Insect Endosymbiont from the Cradle to the Grave”. In : *Current Biology* 29.11. DOI : <https://doi.org/10.1016/j.cub.2019.03.032>.
- MCDOWELL, Rose, SMALL et Charles S. BOND (2022). “Synthetic PPR proteins as tools for sequence-specific targeting of RNA”. In : *Methods* 208, p. 19-26. DOI : <https://doi.org/10.1016/j.ymeth.2022.10.003>.
- MCEWEN, J. E. et al. (1986). “Nuclear functions required for cytochrome c oxidase biogenesis in *Saccharomyces cerevisiae*. Characterization of mutants in 34 complementation groups.” In : *Journal of Biological Chemistry* 261.25, p. 11872-11879. DOI : [https://doi.org/10.1016/S0021-9258\(18\)67323-5](https://doi.org/10.1016/S0021-9258(18)67323-5).
- MEIERHOFF, Karin et al. (2003). “HCF152, an Arabidopsis RNA Binding Pentatricopeptide Repeat Protein Involved in the Processing of Chloroplast psbB-psbT-psbH-petB-petD RNAs ”. In : *The Plant Cell* 15.6, p. 1480-1495. DOI : [10.1105/tpc.010397](https://doi.org/10.1105/tpc.010397). URL : <https://doi.org/10.1105/tpc.010397>.
- MEYER, Eli et Virginia M. WEIS (2012). “Study of Cnidarian-Algal Symbiosis in the “Omics” Age”. In : *The Biological Bulletin* 223.1, p. 44-65. DOI : [10.1086/BBLv223n1p44](https://doi.org/10.1086/BBLv223n1p44).
- MICHAELY, Peter et al. (2002). “Crystal structure of a 12 ANK repeat stack from human ankyrinR”. In : *The EMBO Journal* 21.23, p. 6387-6396. DOI : <https://doi.org/10.1093/emboj/cdf651>.
- MILLER, Stanley L. (1953). “A production of amino acids under possible primitive earth conditions”. In : *Science* 117.3046, p. 528-529.
- MOOTHA, Vamsi K. et al. (2003). “Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics”. In : *Proceedings of the National Academy of Sciences* 100.2, p. 605-610. DOI : [10.1073/pnas.242716699](https://doi.org/10.1073/pnas.242716699).

- MORTAZA, Shogofa (2018). “Evolution des gènes OPR chez les microalgues”. In : *Mémoire de master*.
- MOYEN, Jean-François et Pierre THOMAS (2007). “Les stromatolithes”. In : *ENS de Lyon*. URL : <https://planet-terre.ens-lyon.fr/ressource/stromatolithes.xml>.
- MURPHY, Bonnie J. et al. (2019). “Rotary substates of mitochondrial ATP synthase reveal the basis of flexible F<sub>1</sub>-F<sub>o</sub> coupling”. In : *Science* 364.6446, eaaw9128. DOI : [10.1126/science.aaw9128](https://doi.org/10.1126/science.aaw9128).
- MURRAY, Kevin B., Denise GORSE et Janet M. THORNTON (2002). “Wavelet transforms for the characterization and detection of repeating motifs”. Edited by G. von Heijne”. In : *Journal of Molecular Biology* 316.2, p. 341-363. DOI : <https://doi.org/10.1006/jmbi.2001.5332>.
- MURRAY, Kevin B., William R. TAYLOR et Janet M. THORNTON (2004). “Toward the detection and validation of repeats in protein structure”. In : *Proteins : Structure, Function, and Bioinformatics* 57.2, p. 365-380. DOI : <https://doi.org/10.1002/prot.20202>.
- NASA (2022). “Sun”. In : URL : <https://solarsystem.nasa.gov/solar-system/sun/overview/>.
- NIMMO, Francis et Robert T. PAPPALARDO (2016). “Ocean worlds in the outer solar system”. In : *Journal of Geophysical Research : Planets* 121.8, p. 1378-1399. DOI : <https://doi.org/10.1002/2016JE005081>. URL : <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016JE005081>.
- ODENWALD, Sten (s. d.). “How long does it take light to get out from the inside of the Sun?” In : (). URL : <http://sten.astronomycafe.net/faqs/>.
- OIKONOMOU, Catherine M., Yi-Wei CHANG et JENSEN (2016). “A new view into prokaryotic cell biology from electron cryotomography”. In : *Nature Reviews Microbiology* 14, p. 205-220. DOI : <https://doi.org/10.1038/nrmicro.2016.7>.
- PARKS, Donovan H. et al. (2018). “A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life”. In : *Nature Biotechnology*. DOI : <https://doi.org/10.1038/nbt.4229>.
- PERALTA, Susana, WANG et Carlos T. MORAES (2012). “Mitochondrial transcription : Lessons from mouse models”. In : *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1819.9, p. 961-969. DOI : <https://doi.org/10.1016/j.bbagrm.2011.11.001>.
- PETERSEN, Jörn et al. (2014). “Chromera velia, Endosymbioses and the Rhodoplex Hypothesis—Plastid Evolution in Cryptophytes, Alveolates, Stramenopiles, and Haptophytes (CASH Lineages)”. In : *Genome Biology and Evolution* 6.3, p. 666-684. DOI : [10.1093/gbe/evu043](https://doi.org/10.1093/gbe/evu043).

- PHILLIPS, James C. et al. (2020). “Scalable molecular dynamics on CPU and GPU architectures with NAMD”. In : *The Journal of Chemical Physics* 153.4, p. 044130. DOI : [10.1063/5.0014475](https://doi.org/10.1063/5.0014475).
- PONCE-TOLEDO, Rafael I. et al. (2017). “An Early-Branching Freshwater Cyanobacterium at the Origin of Plastids”. In : *Current Biology* 27.3, p. 386-391. DOI : <https://doi.org/10.1016/j.cub.2016.11.056>.
- PORTER, J.R. (1976). “Antony van Leeuwenhoek : tercentenary of his discovery of bacteria.” In : *Bacteriological reviews*, p. 260-269. DOI : [10.1128/br.40.2.260-269.1976](https://doi.org/10.1128/br.40.2.260-269.1976).
- PREKER, Pascal J. et Walter KELLER (1998). “The HAT helix, a repetitive motif implicated in RNA processing”. In : *Trends in Biochemical Sciences* 23.1, p. 15-16. DOI : [https://doi.org/10.1016/S0968-0004\(97\)01156-0](https://doi.org/10.1016/S0968-0004(97)01156-0).
- QU, Kaiyang et al. (2019). “Identifying Plant Pentatricopeptide Repeat Coding Gene/Protein Using Mixed Feature Extraction Methods”. In : *Frontiers in Plant Science* 9. DOI : [10.3389/fpls.2018.01961](https://doi.org/10.3389/fpls.2018.01961).
- RAHIRE, Michèle et al. (2012). “Identification of an OPR protein involved in the translation initiation of the PsaB subunit of photosystem I”. In : *The Plant Journal* 72.4, p. 652-661.
- REUELL, Peter (2019). “Spreading seeds of life”. In : *The Harvard Gazette*. URL : <https://news.harvard.edu/gazette/story/2019/07/harvard-study-suggests-asteroids-might-play-key-role-in-spreading-life/>.
- RIS, Hans et Walter PLAUT (1962). “Ultrastructure of DNA-containing areas in the chloroplast of Chlamydomonas”. In : *Journal of Cell Biology* 13.3, p. 383-391. DOI : [10.1083/jcb.13.3.383](https://doi.org/10.1083/jcb.13.3.383).
- ROBERTSON, Hugh D., Sidney ALTMAN et John D. SMITH (1972). “Purification and properties of a specific Escherichia coli ribonuclease which cleaves a tyrosine transfer ribonucleic acid precursor”. In : *Journal of Biological Chemistry* 247.16, p. 5243-5251.
- ROCHAIX (1996). “Post-transcriptional regulation of chloroplast gene expression in Chlamydomonas reinhardtii.” In : *Plant molecular biology* 42.4, p. 327-341. DOI : [10.1007/BF00039389](https://doi.org/10.1007/BF00039389).
- ROCHAIX et al. (1989). “Nuclear and chloroplast mutations affect the synthesis or stability of the chloroplast psbC gene product in Chlamydomonas reinhardtii.” In : *The EMBO Journal* 8.4, p. 1013-1021. DOI : <https://doi.org/10.1002/j.1460-2075.1989.tb03468.x>.
- RODRIGUES-OLIVEIRA, Thiago et al. (2023). “Actin cytoskeleton and complex cell architecture in an Asgard archaeon”. In : *Nature* 613. DOI : <https://doi.org/10.1038/s41586-022-05550-y>.

- ROGER, Andrew J., Sergio A. MUÑOZ-GÓMEZ et Ryoma KAMIKAWA (2017). “The Origin and Diversification of Mitochondria”. In : *Current Biology* 27.21, R1177-R1192. DOI : <https://doi.org/10.1016/j.cub.2017.09.015>.
- ROVIRA, Aleix Gorchs et SMITH (2019). “PPR proteins – orchestrators of organelle RNA metabolism”. In : *Physiologia Plantarum* 166.1, p. 451-459. DOI : <https://doi.org/10.1111/ppl.12950>.
- RÜDINGER, Mareike et al. (2011). “Plant-type mitochondrial RNA editing in the protist *Naegleria gruberi*.” In : *RNA*, p. 2058-2062. DOI : [10.1261/rna.02962911](https://doi.org/10.1261/rna.02962911).
- SAGAN, Lynn (1967). “On the origin of mitosing cells”. In : *Journal of Theoretical Biology* 14.3. DOI : [https://doi.org/10.1016/0022-5193\(67\)90079-3](https://doi.org/10.1016/0022-5193(67)90079-3).
- SALADINO, Raffaele et al. (2012). “Genetics First or Metabolism First? The Formamide Clue.” In : *Chemical Society Reviews* 41.16. URL : <https://doi.org/10.1039/c2cs35066a>.
- SCHALLENBERG-RÜDINGER, Mareike et al. (2013). “A survey of PPR proteins identifies DYW domains like those of land plant RNA editing factors in diverse eukaryotes”. In : *RNA Biology*, p. 1549-1556. DOI : [10.4161/rna.25755](https://doi.org/10.4161/rna.25755).
- SCHMITZ-LINNEWEBER, Christian et SMALL (2008). “Pentatricopeptide repeat proteins : a socket set for organelle gene expression”. In : *Trends in Plant Science* 13.12, p. 663-670. DOI : <https://doi.org/10.1016/j.tplants.2008.10.001>.
- SHEN, Cuicui et al. (2016). “Structural basis for specific single-stranded RNA recognition by designer pentatricopeptide repeat proteins”. In : *Nature Communications* 7.1. DOI : <https://doi.org/10.1038/ncomms11285>.
- SHIMIZU, Hikaru et al. (2017). “Structure-based analysis of the guanine nucleotide exchange factor SmgGDS reveals armadillo-repeat motifs and key regions for activity and GTPase binding”. In : *Journal of Biological Chemistry* 292.32, p. 13441-13448. DOI : <https://doi.org/10.1074/jbc.M117.792556>.
- SIBBALD, Shannon J. et John M. ARCHIBALD (2020). “Genomic Insights into Plastid Evolution”. In : *Genome Biology and Evolution* 12.7, p. 978-990. URL : <https://doi.org/10.1093/gbe/evaa096>.
- SIEBURTH, L.E. et al. (1991). “Chloroplast RNA Stability in *Chlamydomonas* : Rapid Degradation of psbB and psbC Transcripts in Two Nuclear Mutants.” In : *The Plant cell*, p. 175-189. DOI : [10.1105/tpc.3.2.175](https://doi.org/10.1105/tpc.3.2.175).
- SIEVERS, Dane et Günter VON KIEDROWSKI (1994). “Self-replication of complementary nucleotide-based oligomers”. In : *Nature* 369.6477, p. 221-224.
- SIKORSKI, Robert S. et al. (1990). “A repeating amino acid motif in CDC23 defines a family of proteins and a new relationship among genes required for mitosis and RNA synthesis”. In : *Cell* 60.2, p. 307-317. DOI : [https://doi.org/10.1016/0092-8674\(90\)90745-Z](https://doi.org/10.1016/0092-8674(90)90745-Z).

- SISQUELLAS, Marion (2018). “Analyses des interactions spécifiques entre des protéines alpha-solénoides et l’ARN messager”. In : *Mémoire de master*.
- SMALL et Nemo PEETERS (2000). “The PPR motif – a TPR-related motif prevalent in plant organellar proteins”. In : *Trends in Biochemical Sciences* 25.2, p. 45-47. ISSN : 0968-0004. DOI : [https://doi.org/10.1016/S0968-0004\(99\)01520-0](https://doi.org/10.1016/S0968-0004(99)01520-0).
- SODING, Johannes, Michael REMMERT et Andreas BIEGERT (2006). “HHrep : de novo protein repeat detection and the origin of TIM barrels”. In : *Nucleic Acids Research* 34.suppl<sub>2</sub>, W137-W142. DOI : [10.1093/nar/gkl1130](https://doi.org/10.1093/nar/gkl1130).
- SPENCER, John R. et Francis NIMMO (2013). “Enceladus : An Active Ice World in the Saturn System”. In : *Annual Review of Earth and Planetary Sciences* 41.1, p. 693-717. DOI : [10.1146/annurev-earth-050212-124025](https://doi.org/10.1146/annurev-earth-050212-124025). URL : <https://doi.org/10.1146/annurev-earth-050212-124025>.
- STANLEY, George D. et Peter K. SWART (1995). “Evolution of the coral-zooxanthellae symbiosis during the Triassic : a geochemical approach”. In : *Paleobiology* 21.2, p. 179-199. DOI : [10.1017/S0094837300013191](https://doi.org/10.1017/S0094837300013191).
- STILLMAN, David E. et al. (2022). “Partially-Saturated Brines Within Basal Ice or Sediments Can Explain the Bright Basal Reflections in the South Polar Layered Deposits”. In : *Journal of Geophysical Research : Planets* 127.10. DOI : <https://doi.org/10.1029/2022JE007398>. URL : <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022JE007398>.
- SUGITA, Mamoru et al. (2013). “Architecture of the PPR gene family in the moss *Physcomitrella patens*”. In : *RNA Biology* 10.9, p. 1439-1445. DOI : [10.4161/rna.24772](https://doi.org/10.4161/rna.24772).
- SZKLARCZYK, Radek et Jaap HERINGA (2004). “Tracking repeats using significance and transitivity”. In : *Bioinformatics* 20.suppl<sub>1</sub>, p. i311-i317. DOI : [10.1093/bioinformatics/bth911](https://doi.org/10.1093/bioinformatics/bth911).
- TAIB, Najwa et al. (2020). “Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm transition”. In : *Nature Ecology Evolution* 4, p. 1661-1672. DOI : <https://doi.org/10.1038/s41559-020-01299-7>.
- TOMECKOVA, Lucia et al. (2020). “The Lipid Composition of *Euglena gracilis* Middle Plastid Membrane Resembles That of Primary Plastid Envelopes”. In : *Plant Physiology* 184.4, p. 2052-2063. DOI : [10.1104/pp.20.00505](https://doi.org/10.1104/pp.20.00505).
- VERLET, Loup (1967). “Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules”. In : *Phys. Rev.* 159 (1), p. 98-103. DOI : [10.1103/PhysRev.159.98](https://doi.org/10.1103/PhysRev.159.98).
- VIOLA, Stefania et al. (2019). “MDA1, a nucleus-encoded factor involved in the stabilization and processing of the atpA transcript in the chloroplast of *Chlamydomonas*”. In : *The Plant Journal* 98.6, p. 1033-1047. DOI : <https://doi.org/10.1111/tpj.14300>.

- VIRCHOW, Rudolf Ludwig Karl (1859). *Die Cellularpathologie in ihrer Begründung auf physiologische und pathologische Gewebelehre*. A. Hirshwald.
- VO, An, Nha NGUYEN et Heng HUANG (2010). “Solenoid and non-solenoid protein recognition using stationary wavelet packet transform”. In : *Bioinformatics* 26.18. DOI : [10.1093/bioinformatics/btq371](https://doi.org/10.1093/bioinformatics/btq371).
- WÄCHTERSCHÄUSER, Günter (1992). “Groundworks for an evolutionary biochemistry : The iron-sulphur world”. In : *Progress in Biophysics and Molecular Biology* 58.2, p. 85-201. DOI : [https://doi.org/10.1016/0079-6107\(92\)90022-X](https://doi.org/10.1016/0079-6107(92)90022-X).
- WEINER, Alan M. (1993). “mRNA splicing and autocatalytic introns : distant cousins or the products of chemical determinism?” In : *Cell* 72.2, p. 161-164.
- WETHERBEE, Richard et al. (2019). “The golden paradox – a new heterokont lineage with chloroplasts surrounded by two membranes”. In : *Journal of Phycology* 55.2, p. 257-278. DOI : <https://doi.org/10.1111/jpy.12822>.
- WIEDEMANN, Nils et Nikolaus PFANNER (2017). “Mitochondrial Machineries for Protein Import and Assembly”. In : *Annual Review of Biochemistry* 86.1, p. 685-714. DOI : [10.1146/annurev-biochem-060815-014352](https://doi.org/10.1146/annurev-biochem-060815-014352).
- WOLLMAN, Francis-André (2016). “An antimicrobial origin of transit peptides accounts for early endosymbiotic events”. In : *Traffic* 17.12, p. 1322-1328. DOI : <https://doi.org/10.1111/tra.12446>.
- WOLLMAN, Francis-André, Limor MINAI et Rachel NECHUSHTAI (1999). “The biogenesis and assembly of photosynthetic proteins in thylakoid membranes<sup>1</sup>This article is dedicated to the memory of our colleague and friend, Alma Gal. Her presence played a major role in our decision to prepare this review article.<sup>1</sup>”. In : *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 1411.1, p. 21-85. DOI : [https://doi.org/10.1016/S0005-2728\(99\)00043-2](https://doi.org/10.1016/S0005-2728(99)00043-2).
- WOODSON, Jesse D. et Joanne CHORY (2008). “Coordination of gene expression between organellar and nuclear genomes”. In : *Nature Reviews Genetics* 9.5, p. 383-395. DOI : <https://doi.org/10.1038/nrg2348>.
- WOOLFSON, Michael (fév. 2000). “The origin and evolution of the solar system”. In : *Astronomy & Geophysics* 41.1, p. 1.12-1.19. ISSN : 1366-8781. DOI : [10.1046/j.1468-4004.2000.00012.x](https://doi.org/10.1046/j.1468-4004.2000.00012.x). URL : <https://doi.org/10.1046/j.1468-4004.2000.00012.x>.
- WOSTRIKOFF, Katia et al. (2004). “Biogenesis of PSI involves a cascade of translational autoregulation in the chloroplast of *Chlamydomonas*”. In : *The EMBO Journal* 23.13, p. 2696-2705. DOI : <https://doi.org/10.1038/sj.emboj.7600266>.
- YAMAZAKI, Hiroyuki, Masao TASAKA et Toshiharu SHIKANAI (2004). “PPR motifs of the nucleus-encoded factor, PGR3, function in the selective and distinct steps of chloroplast gene expression in *Arabidopsis*”. In : *The Plant Journal*

- 38.1, p. 152-163. DOI : <https://doi.org/10.1111/j.1365-313X.2004.02035.x>.
- YIN, Ping et al. (2013). “Structural basis for the modular recognition of single-stranded RNA by PPR proteins”. In : *Nature* 504.7478, p. 168-171. DOI : <https://doi.org/10.1038/nature12651>.
- YOON, Hwan Su, Jeremiah D. HACKETT, Claudia CINIGLIA et al. (2004). “A Molecular Timeline for the Origin of Photosynthetic Eukaryotes”. In : *Molecular Biology and Evolution* 21.5, p. 809-818. DOI : 10.1093/molbev/msh075.
- YOON, Hwan Su, Jeremiah D. HACKETT, Frances M. VAN DOLAH et al. (2005). “Tertiary Endosymbiosis Driven Genome Evolution in Dinoflagellate Algae”. In : *Molecular Biology and Evolution* 22.5, p. 1299-1308. DOI : 10.1093/molbev/msi118.
- YUZAWA, Satoru et al. (2011). “Structural basis for interaction between the conserved cell polarity proteins Inscuteable and Leu-Gly-Asn repeat-enriched protein (LGN)”. In : *Proceedings of the National Academy of Sciences* 108.48, p. 19210-19215. DOI : 10.1073/pnas.1110951108.
- ZAMBRANO, Andrea et al. (2007). “Aberrant translation of cytochrome c oxidase subunit 1 mRNA species in the absence of Mss51p in the yeast *Saccharomyces cerevisiae*.” In : *Molecular biology of the cell* 23.13, p. 523-535. DOI : 10.1091/mbc.e06-09-0803..
- ZERGES et ROCHAIX (1994). “The 5' leader of a chloroplast mRNA mediates the translational requirements for two nucleus-encoded functions in *Chlamydomonas reinhardtii*”. In : *Molecular and Cellular Biology* 14.8, p. 5268-5277. DOI : 10.1128/mcb.14.8.5268-5277.1994.
- ZHAO, Xudong, Qing JIAO et al. (2020). “ECFS-DEA : an ensemble classifier-based feature selection for differential expression analysis on expression profiles”. In : *BMC Bioinformatics*. DOI : <https://doi.org/10.1186/s12859-020-3388-y>.
- ZHAO, Xudong, WANG et al. (2021). “Identifying Plant Pentatricopeptide Repeat Proteins Using a Variable Selection Method”. In : *Frontiers in Plant Science* 12. DOI : 10.3389/fpls.2021.506681.
- ZIMMER, Christophe, Krishan K. KHURANA et Margaret G. KIVELSON (2000). “Subsurface Oceans on Europa and Callisto : Constraints from Galileo Magnetometer Observations”. In : *Icarus* 147.2, p. 329-347. DOI : <https://doi.org/10.1006/icar.2000.6456>.
- ZIMMERMANN, Lukas et al. (2018). “A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core”. In : *Journal of Molecular Biology* 430.15, p. 2237-2243. DOI : <https://doi.org/10.1016/j.jmb.2017.12.007>.



"- Je connais une planète où il y a un Monsieur cramoisi. Il n'a jamais respiré une fleur. Il n'a jamais regardé une étoile. Il n'a jamais aimé personne. Il n'a jamais rien fait d'autre que des additions. Et toute la journée il répète comme toi : "Je suis un homme sérieux ! Je suis un homme sérieux !" et ça le fait gonfler d'orgueil. Mais ce n'est pas un homme, c'est un champignon !

- Un quoi ?

- Un champignon !

Le petit prince était maintenant tout pâle de colère.

- Il y a des millions d'années que les fleurs fabriquent des épines. Il y a des millions d'années que les moutons mangent quand même les fleurs. Et ce n'est pas sérieux de chercher à comprendre pourquoi elles se donnent tant de mal pour se fabriquer des épines qui ne servent jamais à rien ? Ce n'est pas important la guerre des moutons et des fleurs ? Ce n'est pas plus sérieux et plus important que les additions d'un gros Monsieur rouge ? Et si je connais, moi, une fleur unique au monde, qui n'existe nulle part, sauf dans ma planète, et qu'un petit mouton peut anéantir d'un seul coup, comme ça, un matin, sans se rendre compte de ce qu'il fait, ce n'est pas important ça !

Il rougit, puis reprit :

- Si quelqu'un aime une fleur qui n'existe qu'à un exemplaire dans les millions et les millions d'étoiles, ça suffit pour qu'il soit heureux quand il les regarde. Il se dit : "Ma fleur est là quelque part..." Mais si le mouton mange la fleur, c'est pour lui comme si, brusquement, toutes les étoiles s'éteignaient ! Et ce n'est pas important ça !"

*Le Petit Prince*, Antoine de Saint-Exupéry (1943)

# Résumé

Au sein des Archaeplastida (eucaryotes photosynthétiques ayant acquis un chloroplaste suite à une endosymbiose avec une cyanobactérie ancestrale) les génomes chloroplastiques et mitochondriaux des algues vertes et des plantes terrestres sont régulés de manière post-transcriptionnelle, principalement par des protéines à solénoïde alpha codées dans le noyau. Ces facteurs nucléaires sont composés de motifs répétés dégénérés (protéines PPR et OPR, respectivement pentatricopeptide repeat et octatricopeptide repeats) interagissant de façon spécifique avec une partie de la séquence de leur ARN cible et forment de grandes familles de paralogues. Les protéines PPR sont très abondantes chez les plantes terrestres tandis que les OPR le sont chez les algues vertes. Ces expansions différentielles, en parallèle de l'évolution du métabolisme des ARN dans les organites pourraient refléter des adaptations génétiques préservant la phototrophie dans diverses conditions et niches écologiques. Chez les autres Archaeplastida (algues rouges et Glaucophytes) et chez les eucaryotes issus d'une endosymbiose avec une microalgue ancestrale comme les Diatomées, la régulation des génomes des organites reste peu explorée.

Un premier objectif de ma thèse a été de décrire la diversité et la dynamique évolutive des protéines à solénoïde alpha connues ou candidates pour la régulation de l'expression du génome des organites et ce, dans l'ensemble des eucaryotes photosynthétiques. Pour les identifier, j'ai développé une approche combinant détection d'homologie lointaine de séquence et classification indépendante de la similarité entre séquences. J'ai validé cette approche en retrouvant et complétant les familles OPR et PPR connues chez les espèces modèles *Chlamydomonas reinhardtii* et *Arabidopsis thaliana*. J'ai montré que les expansions d'OPR étaient restreintes au sein des Chlorophytes et qu'en dehors des algues vertes et des plantes terrestres, les protéines à PPR et à OPR étaient peu nombreuses, suggérant que d'autres acteurs de la régulation de l'expression des génomes des organites restent à découvrir. J'ai également identifié plusieurs dizaines d'autres familles de protéines à solénoïde alpha adressées aux organites dans tous les protéomes étudiés, certaines aux fonctions encore inconnues et dont la caractérisation expérimentale dans des organismes modèles serait pertinente.

Dans un second temps, j'ai utilisé des approches de dynamique moléculaire pour mieux comprendre l'affinité et la spécificité des liaisons entre les PPR et leurs ARN cibles. J'ai notamment étudié la dynamique des motifs répétés et la géométrie des sites de liaison des nucléotides en fonction de leur position dans la séquence des motifs PPR, y compris les effets du nombre de répétitions et de la présence ou non des domaines N- et C-terminaux, en plus de l'évolution de la conformation globale de la protéine. Nos résultats suggèrent le rôle de la flexibilité des protéines PPR, tant au niveau de la protéine que du motif dans la liaison à sa cible ARN et sa pertinence pour l'affinité et la spécificité de la reconnaissance des nucléotides.

# Abstract

In Archaeplastida (photosynthetic eukaryotes that acquired a chloroplast following endosymbiosis with an ancestral cyanobacterium) the chloroplast and mitochondrial genomes of green algae and land plants are regulated post-transcriptionally, mainly by alpha-solenoid proteins encoded in the nucleus. These nuclear factors are composed of degenerate repeat motifs (PPR and OPR proteins, respectively pentatricopeptide repeat and octatricopeptide repeats) that interact specifically with part of their target RNA sequence and form large families of paralogs. PPR proteins are very abundant in terrestrial plants while OPRs are abundant in green algae. These differential expansions, in parallel with the evolution of RNA metabolism in organelles, may reflect genetic adaptations that preserve phototrophy under different conditions and ecological niches. In other Archaeplastids (red algae and Glaucophytes) and in eukaryotes that originate from endosymbiosis with an ancestral microalga such as the Diatoms, the regulation of organelle genomes remains poorly explored.

A first objective of my thesis was to describe the diversity and evolutionary dynamics of known or candidate alpha-solenoid proteins for the regulation of organelle genome expression in all photosynthetic eukaryotes. To identify them, I developed an approach that combines distant sequence homology detection and sequence similarity independent classification. I validated this approach by finding and completing the known OPR and PPR families in the model species *Chlamydomonas reinhardtii* and *Arabidopsis thaliana*. I showed that OPR expansions were restricted within Chlorophytes and that outside of green algae and land plants, PPR and OPR proteins were few in number, suggesting that other players in the regulation of organelle genome expression remain to be discovered. I also identified several dozen other families of organelle-addressed alpha-solenoid proteins in all the proteomes studied, some of which have as yet unknown functions and whose experimental characterisation in model organisms would be relevant.

In a second step, I used molecular dynamics approaches to better understand the affinity and specificity of binding between PPRs and their target RNAs. In particular, I studied the dynamics of the repeat motifs and the geometry of the nucleotide binding sites as a function of their position in the PPR motif sequence, including the effects of the number of repeats and the presence or absence of N- and C-terminal domains, in addition to the evolution of the overall conformation of the protein. Our results suggest the role of PPR protein flexibility, both at the protein and motif level, in binding to its RNA target and its relevance to the affinity and specificity of nucleotide recognition.