Université Paris Cité Master 2 Biologie-Informatique 2021-2022

Rapport de Stage M2

Annotation des protéines candidates à la régulation post-transcriptionelle des génomes des organismes photosynthétiques

Goulancourt Rebecca

Tutrice : Ingrid Lafontaine.

Site : IBPC - CNRS, 13 rue Pierre et Marie Curie, 75005 Paris

Table des matières

REMERCIEMENTS

Je remercie Angela Falciatore ainsi que toute l'équipe Photosynthèse de l'IBPC pour m'avoir accueillie dans leur laboratoire.

Je remercie particulièrement Ingrid Lafontaine et Céline Cattelin pour leur dévouement et leur accueil, ainsi que pour le suivi qu'elles m'ont apporté tout au long de mon stage.
Je remercie également Clotilde Garrido pour les conseils qu'elle a pu m'apporter.

I. Introduction & état de l'art

Au sein des eucaryotes le siège de la photosynthèse est le chloroplaste. Le chloroplaste, au même titre que la mitochondrie, est un organite de la cellule eucaryote. Ils proviennent de l'endosymbiose de bactéries ancestrales qui se sont maintenues dans leur cellule hôte au cours de l'évolution. Les cellules issues de l'endosymbiose primaire chloroplastique, les Archaeplastida, proviennent de l'internalisation (probablement par phagocytose) d'une cyanobactérie par une cellule ancestrale primaire hétérotrophe (figure 1B) [8]. La cyanobactérie se différencie ensuite en chloroplaste et la cellule acquiert une capacité photosynthétique, devenant ainsi autotrophe. Les autres eucaryotes photosynthétiques découlent d'une endosymbiose secondaire. C'est le cas des diatomées : des microalgues de diverses formes (figure 1A) dont la taille peut varier de 2 µm à 1mm et responsables de 20% de l'oxygène que nous respirons [36]. Elles font partie des *Stramenopila* (SAR) et résultent de l'internalisation d'une micro algue rouge (Rhodophyte, Archaeplastida) autotrophe possédant un chloroplaste par une autre cellule eucaryote hétérotrophe.

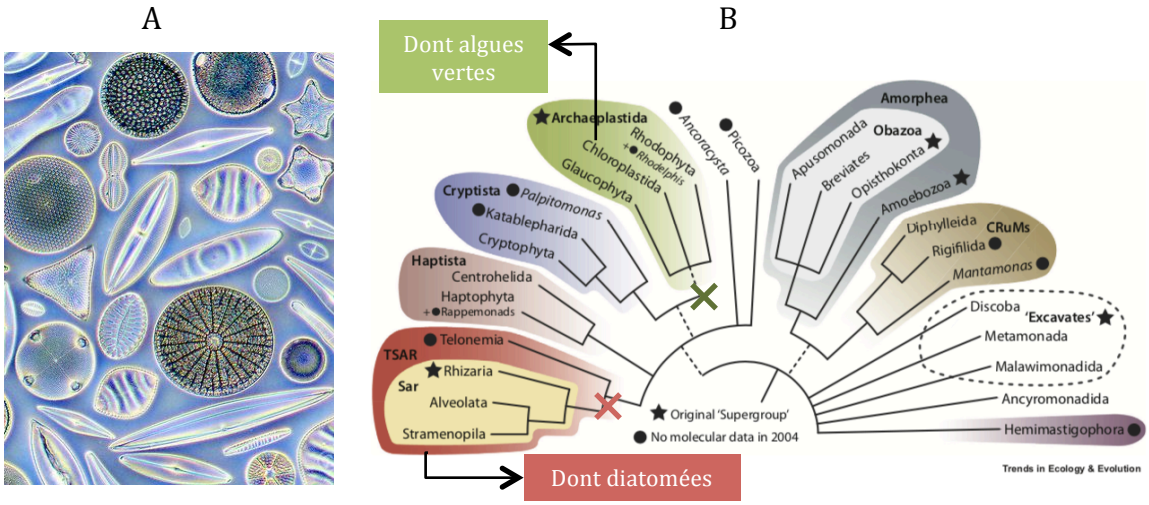


Figure 1 : (A) Diatomées observées en microscopie à contraste de phase. (B) Arbre phylogénétique des eukaryotes [8]. La croix verte indique la position de l'endosymbiose primaire avec une cyanobactérie. La croix rouge indique la position de l'événement d'endosymbiose primaire avec une rhodophyte dont découlent les diatomées.

Les événements d'endosymbiose primaires ont été accompagnés d'une perte d'ADN massive au sein du génome bactérien ancestral et de nombreux transferts vers le génome de l'hôte. Cependant une partie de ce génome d'origine bactérienne s'est maintenue et continue de s'exprimer. La régulation de l'expression du génome du chloroplaste et de la mitochondrie est

connue chez les Chloroplastida (plantes et algues vertes, faisant partie des Archaeplastida), issus d'une endosymbiose primaire mais beaucoup moins chez les autres eucaryotes photosynthétiques comme les diatomées. Ainsi chez les Chloroplastida la régulation de l'expression des génomes des organites s'effectue principalement au niveau post-transcriptionnel par des protéines codées dans le noyau, traduites dans le cytosol et adressées aux organites : chloroplaste et mitochondrie. La plupart de ces protéines possèdent une structure en α-solénoïde et interagissent de façon séquence-spécifique avec l'ARN messager (ARNm) du chloroplaste ou de la mitochondrie dans des processus de maturation, stabilisation, épissage, translation et dégradation génétique. Dans la suite de ce rapport nous désignerons par ROGEs, pour Regulator Organelle Gene Expression [42], ces protéines en α-solénoïdes adressées aux organites. Les ROGEs connues sont composées d'une succession de paires d'hélices α en antiparallèle. Ces paires d'hélices sont codées par des motifs répétés au sein de la séquence. Un motif de répétition constitue deux hélices antiparallèles. L'interaction des ROGEs avec l'ARN se fait au sein de la cavité concave formée par la structure en hélice α. Il semblerait que la méthode de reconnaissance de l'ARNm soit due à une spécificité des bases azotées et d'une reconnaissance dite « one repeat :one-nucleotide ». Cela signifie que la reconnaissance nucléotidique permet à un nucléotide de se fixer entre deux motifs d'hélices.

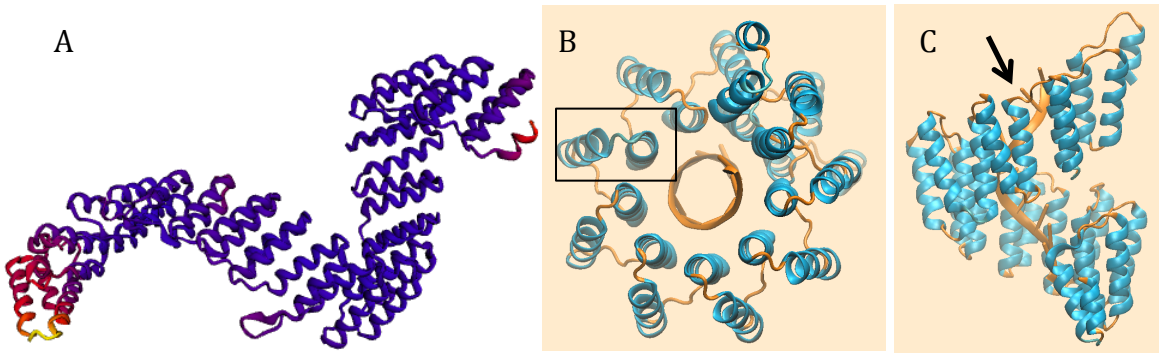


Figure 2 : (A) Protéine en α solénoïde prédite par alpha fold (code Uniprot : Q9LKV3, *Arabidopsis thaliana*). (B) Protéine en α-solénoïde résolue par diffraction de rayons-X. Elle est liée à un ARNm du côté 5', le carré noir montrant 1 motif de répétition d'hélice α en antiparallèle (code PDB : 5I9F). (C) Vue différente de la structure 5I9F. La flèche indique la position d'un nucléotides entre deux motifs : c'est ce que l'on appelle la reconnaissance par « one repeat : one nucleotide ». © Céline Cattelin, IBPC.

Parmi les protéines en α-solénoïdes impliquées dans la régulation de l'expression du génome des organites, on distingue trois types selon la nature des répétitions au sein de leur séquence [42] [48] :

- les TPR possèdent des répétitions de 34 acides aminés (tetracopeptide repeat), spécialisées dans les interactions protéine-protéine.
- les PPR possèdent des répétitions de 35 acides aminés (pentatricopeptide repeat), et les OPR des répétitions de 38 acides aminés (octatricopeptide repeat), qui interagissent avec ARN.

Ainsi seules les OPR et les PPR sont des ROGEs se liant à l'ARNm. Les PPR constituent de grandes familles multigéniques chez les plantes (environ 481 chez *Arabidopsis thaliana*) mais sont peu nombreuses chez les microalgues vertes (15 copies chez *Chlamydomonas reinhardtii*) [48]. Les PPR sont elles mêmes classifiées en trois sous-groupes selon leur composition en motifs : P-PPR, PPR-SMR et PLS-PPR.

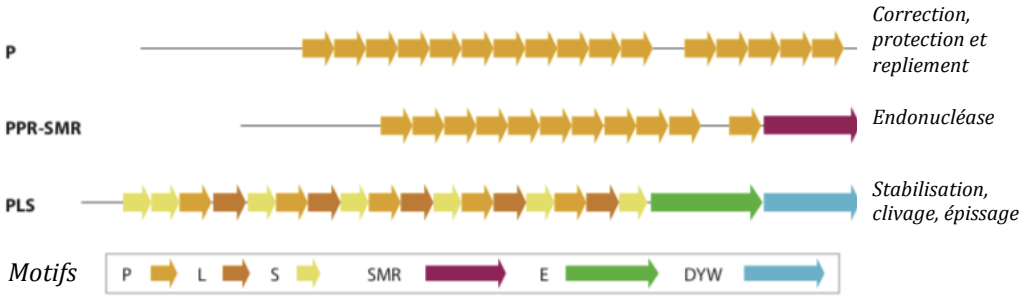


Figure 3 : Les différents motifs présents chez les PPR [48].

Ainsi les P-PPR ne sont constituées que de motifs P leur conférant des activités de correction, de protection face aux nucléases et de repliement (celui-ci jouant un rôle dans la traduction). Les PPR-SMR sont constitués de motifs P et d'un motif SMR. Ce motif SMR (MutS Related domain) se situant en fin de séquence (C-terminal) possède un rôle d'endonucléase. Pour ce qui est des PLS-PPR elles sont constituées de différents motifs : P, L, S ou encore E et DYW. Le motif DYW (Asp-Tyr-Trp) est aussi appelé le « cytidine deaminase-like domain » et permet ainsi la désamination des cytidines en uraciles. Toutefois les fonctions exactes des domaines E et DYW ne sont pas véritablement connus chez les PPR mais nous savons que les PLS-PPR possèdent un rôle de stabilisation, de clivage et d'épissage [48].

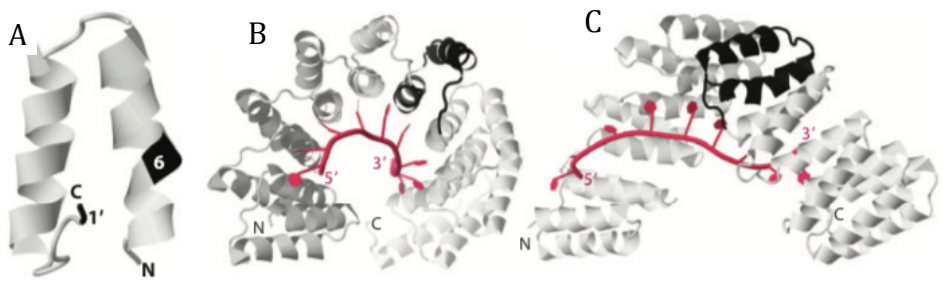


Figure 4 : Structure d'une PPR (PRORP1 RNaseP chez *A.thaliana*). (A) Un seul motif PPR de la protéine. Les positions 1' et 6' sont coloriées en noir et déterminent la spécificité de liaison des nucléotides aux motifs PPR. (B) & (C) L'ARNm se fixe au sein de la protéine et les nucléotides se lient entre chaque motif. Un motif est surligné en noir. Figure adaptée de l'article [48], Figure 1.

Les OPR restent peu étudiées et de ce fait moins connues que les PPR. Dans certaines OPR, il existe des motifs appelés RAP qui possèdent une activité endonucléolytique [42]. Les ROGEs étant peu conservées au niveau de leur séquence, il est difficile de les identifier par similarité de séquence dans l'ensemble des eucaryotes photosynthétiques en général, et chez les diatomées et les algues rouges en particulier.

Pour déterminer les possibles candidats à la régulation post-transcriptionelle, un protocole ne se basant pas sur la similarité de séquence a été développé avant mon arrivée au laboratoire par la doctorante Céline Cattelin (annexe 1). Il permet d'identifier des protéines candidates à l'aide d'un arbre décisionnel opérant sur une suite de prédictions de propriétés structurelles et physico-chimiques des protéines. Le but de ce projet est d'améliorer la détection de protéines candidates à la régulation post-transcriptionelle par méthode d'apprentissage de machine (machine learning) afin de pouvoir annoter l'ensemble des ROGEs chez les eucaryotes photosynthétiques.

Dans un premier temps nous développerons et utiliserons la méthode sur deux organismes modèles d'Archaeplastida : la microalgue verte *Chlamydomonas reinhardtii* et la plante terrestre *Arabidopsis thaliana*, puis nous l'appliquerons à la diatomée modèle *Phaeodactylum tricornutum* dont la régulation des génomes des organites est très peu étudiée.

II. Matériel & méthodes

Les codes permettant de répondre aux questions et problématiques ont été rédigés essentiellement en python (version 3.9), puis en R (annexe 2).

Protéomes étudiés

	Protéomes		
	C. reinhardtii	*A. thaliana*	*P. tricornutium*
Base de donnée	Phytozome	NCBI	NCBI
Identifiant	281	4	418
Version	5.6	assembly TAIR10.1	assembly ASM15095v2

Propriétés structurales

Nous utiliserons dans l'étude les résultats de prédictions de 8 logiciels, dont les paramètres ont été laissés par défaut :

Le logiciel TMHMM (version 2.0) permet de prédire la présence et la position de segments transmembranaires. Il fonctionne sur le modèle de chaînes de Markov[1]. Il permet ainsi de déterminer si une protéine est transmembranaire ou non.

Le logiciel Ard2 (Alpha-rod Reapeat Detector) permet l'identification de régions appelées coudes (ou « linker ») entre deux hélices α, permettant ainsi la détection de protéines possédant des structures en α-solénoïdes en utilisant un réseau neuronal.

Radar (Rapid Automatic Detection and Alignment of Repeats, version 1.3) permet la détection de motifs répétés au sein de séquences protéiques. Nous l'utilisons ici car les paires d'hélices α des ROGEs sont des structures aux motifs répétés. Ce logiciel fonctionne sur un algorithme capable d'identifier des répétitions de motifs de différents types.

Nous utilisons 5 outils de prédiction de localisation intracellulaire (table 1) :

[1] Chaque élément à i+1 ne dépend que de la valeur à la position i, les valeurs antérieurs ne permettent aucune ou qu'une très faible prédiction

Logiciel	**Targetp2**	**Deeploc**	**Wolfpsort**	**Localizer**	**Hectar**
Méthode	Apprentissage profond (deep learning)	Réseaux neuronaux récurrents	Machine learning	Machine learning	Support Vector Machines
Version	2.0	1.0	0.2	1.0	1.0

Table 1 : Logiciels de prédictions d'adressages intracellulaires, leurs méthodes et versions.

La prédiction de la structure tertiaire des protéines est réalisée par Alpha-fold *(https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb, version 1.4)*. C'est un logiciel de prédiction et de modélisation de structure protéique à partir de la séquence.

Composition et propriétés physico-chimiques

Pour notre modèle de prédiction nous utiliserons aussi des caractéristiques comme la fréquence en acide aminé, mais aussi les valeurs d'ACC (Auto Cross Correlation) sur les Z-scales de la protéine. Ces Z-scales résument les caractéristiques physico-chimiques des acides aminés [12][69]. Pour chacun des 20 acides aminés on distingue trois valeurs de Z-scales (annexe 3.1). Ces Z-scales permettent d'expliquer la variabilité de structure et d'activité des peptides. L'ACC correspond donc à la covariance entre Z-scales d'un ou plusieurs acides aminés pour un lag donné [12][69]. Nous prenons un lag (ou fenêtre) de 4 car cela est le plus conforme à un tour d'hélice de 3.6 acides aminés[2]. Comme un acide aminé se caractérise par 9 valeurs d'ACC, pour un lag de 4 un peptide sera caractérisé par 36 valeurs[3]. Un graphe explicatif et les formules permettant de calculer les ACC sont donnés en annexe (annexe 3.2 et 3.3). Ainsi les ACC combinent les informations de l'auto-cross variance entre mêmes facteurs à chaque position et la cross variance entre deux différents Z-scales à chaque positions. Les ACC constituent donc une approche physico-chimique quant à la description d'un peptide.

[2] Un lag de 1 représente donc la relation entre les résidus 1 et 2, un lag de 2 entre les résidus 1 et 3, ect. Pour un lag de 4 on considère les résidus 1 à 5. [Clotilde]

[3] Ainsi pour un même facteur ($z_1 - z_1$, $z_2 - z_2$ et $z_3 - z_3$) on obtient 3*r = 3*4 = 12 valeurs d'ACC. Tandis que pour des facteurs différents ($z_1 - z_2$, $z_1 - z_3$, $z_2 - z_3$, $z_3 - z_1$, …) on a 6*r = 6*4 = 24 ACC soit 12+24 = 36 termes d'ACC.

Localisation intracellulaire

L'adressage vers le chloroplaste ou la mitochondrie est prédite en déterminant la présence du peptide signal (ou peptide de transit) tandis que la localisation vers le noyau est prédite par un ensemble de signaux peptidiques nucléaires (NLSs). Nous utilisons cinq prédicteurs : Targetp2, Deeploc, Wolfpsort, Localizer et Hectar (table 1). Targetp2 est entraîné et testé sur des modèles de séquences créées à partir de la matrice de substitution BLOSUM62. Deeploc distingue 10 localisations cellulaires différentes[4] sur la base de la séquence. Wolpsort utilise la composition en acide aminé qu'il convertit en vecteurs pour être ensuite classifiés par méthode k-nearest neighbor[5]. L'algorithme permet donc de déterminer la localisation cellulaire d'une protéine par l'intermédiaire de sa proportion en acide aminé. L'algorithme de Localizer a été entraîné pour prédire la localisation des protéines chez les plantes. Enfin, Hectar utilise la méthode algorithmique « diviser pour mieux régner »[6] afin d'identifier la localisation intracellulaire des protéines chez les hétérokontes.
Tous ces logiciels ont été utilisés en stand alone.

Comparaison de séquence et annotation

Pour comparer les protéines nous utiliserons Blastp [68] un outil de comparaison qui permet de déterminer le degré de similarité entre deux protéines. Pour chaque comparaison protéine-protéine, une E-value est associée. Plus cette Evalue est faible plus la significativité statistique de la similarité entre les deux protéines comparée est élevée. Nous sélectionnons les similarités dont la E-value est inférieure[7] à 10^{-3}. Pour visualiser le réseau de similarité entre les protéines et donc leurs liens de parentés supposés, le logiciel Cytoscape (version 3.9.1) est utilisé. Pour annoter les protéines candidates nous utiliserons eggNOG. C'est un logiciel d'annotation fonctionnel basé sur la recherche de similarité de séquences. Il identifie les protéines d'entrée par similarité de séquence des protéines homologues de sa propre base de données *(http://eggnog5.embl.de)*, déjà annotée pour proposer une annotation de la séquence analysée.

---

[4] Noyau, Cytoplasme, extracellulaire, mitochondriale, membranaire, reticulum endoplasmique, chloroplaste, appareil de Golgi, Lysosome/vacuole et peroxysome.

[5] Méthode de classification par apprentissage supervisé.

[6] Méthode qui découpe un problème en sous problèmes pour le résoudre plus simplement.

[7] Le seuil dépend de la taille du protéome. Pour un protéome de taille de l'ordre de grandeur de N on prend un seuil de Evalue de 10^{-N}. Pour un protéome de 100 protéines on a donc une Evalue de 10^{-3}.

La classification Random Forest

Nous utilisons un modèle de type « Random Forest » (RF) pour classifier nos protéines. C'est un algorithme utilisé en machine learning se basant sur le principe d'arbres décisionnels. Le résultat du modèle nous amènerait à deux décisions possibles : protéine ROGEs ou non. La RF est en réalité un ensemble de plusieurs arbres décisionnels pour combiner les résultats afin qu'ils soient plus précis. Chaque arbre décisionnel est entraîné sur un sous-ensemble de la dataset (échantillon) et donnera son propre résultat. Ensuite dans le cas d'une classification toutes les décisions sont combinées et le résultat final sera celui qui aura eu la plus grande chance d'être piochée aléatoirement parmi toutes les décisions de chacun des arbres[8]. Le RF possède divers avantages comme une robustesse au sur-apprentissage grâce à sa « forêt » d'arbres décisionnels, mais surtout à l'issue de l'apprentissage il nous sera possible de déterminer le poids de chacun des descripteurs peptidiques.

Jeux de données témoins

En premier lieu deux protéomes témoins ont été fournis au format fasta. Ces protéomes contiennent les séquences de protéines que l'on sait être en α-solénoïde (1081 protéines pour le protéome de témoins positifs) et de protéines qui ne le sont pas (1196 protéines pour le protéome de témoins négatifs)[9]. Nous ferons tourner les logiciels cités plus haut sur ces deux jeux de données protéomiques. Nous calculerons aussi sur ces séquences les valeurs d'ACC et de fréquences d'acides aminés. Ce sont les résultats des outils de Radar et Ard2, ainsi que les ACC et la fréquence en acide aminé qui serviront à faire apprendre notre modèle. Les étapes d'apprentissage et de test du modèle nous permettrons de réguler au mieux les paramètres du modèle RF tandis que l'échantillon de validation nous permettra de vérifier la qualité de notre modèle. Enfin nous pourrons déterminer l'efficacité du modèle par sa précision (taux de bonnes prédictions), sensibilité (probabilité d'avoir un vrai positif) et spécificité (probabilité d'avoir un vrai négatif). Ces informations seront résumées dans ce que l'on appelle une

8 Tandis que pour une approche de régression le résultat constituera la moyenne des prédictions à travers tous les arbres de l'ensemble de la RF.

9 Ces protéomes sont constitués de protéines issues d'espèces de Chloroplastida suivantes : *Chlamydomonas reinhardtii, Arabidopsis thaliana, Chlorella variabilis, Volvox carteri f. nagariensis, Tetradesmus obliquus, Chlamydomonas eustigma, Gonium pectorale, Monoraphidium neglectum, Tetrabaena socialis, Raphidocelis subcapitata, Dunaliella salina*.

matrice de confusion (table 2). Le but est donc de maximiser ces performances afin d'avoir le meilleur modèle prédictif possible.

	Précision	**Sensibilité**	**Spécificité**
Définition	Taux de bonnes prédictions	Probabilité d'avoir un vrai positif	Probabilité d'avoir un vrai négatif
Formule	$\frac{\text{Vrais positifs} + \text{Vrais négatifs}}{\text{Total}}$	$\frac{\text{Vrais négatifs}}{\text{Vrais négatifs} + \text{Faux positifs}}$	$\frac{\text{Vrais positifs}}{\text{Vrais positifs} + \text{Faux négatifs}}$

Table 2 : Formules du calcul des performances du modèle. Plus la précision est élevée plus le modèle est capable de prédire avec exactitude la nature de la protéine (ROGE ou non).

III. Résultats

Nous cherchons à pouvoir identifier des protéines en α-solénoïdes impliquées dans la régulation post-transcriptionelle du génome des organites sur la base de leurs propriétés structurales et physico-chimiques. Nous savons que ces protéines ont des caractéristiques communes :

- Elles sont adressées aux organites : chloroplaste et/ou mitochondrie.
- Elles ne possèdent pas de domaines transmembranaires. Si un tel domaine est détecté avant le 68^{e} acide aminé il pourrait en fait correspondre au peptide d'adressage qui permet aux protéines codées dans le noyau et traduites dans le cytoplasme d'être importées aux organites.
- De par leur structure en α-solénoïdes elles possèdent des paires d'hélices α.

Pour commencer, pour être sûr de la pertinence des logiciels choisis, nous avons défini ces propriétés pour les séquences des protéomes témoins.
Le traitement des sorties des logiciels a été effectué comme indiqué dans la table 3:

Outil	**Propriété**
TMHMM	Identifiants des protéines sans domaine transmembranaire (après le 68^{e} acide aminé)
Radar	Nombre de répétitions et proportion de la séquence en répétition

Ard2	Nombre de coudes, probabilités d'être coude
Targetp2	Adressage[10]
Deeploc	Probabilité d'adressage pour la mitochondrie et le chloroplaste
Wolfpsort	Scores des protéines adressées à la mitochondrie et chloroplaste
Localizer	Résultat des prédictions d'adressage des protéines

Table 3 : Tableau récapitulatif des outils utilisés et des informations recueillies des sorties.

Les ROGEs ne sont pas transmembranaires, ainsi nous avons effectué les prédictions TMHMM sur les deux protéomes et conservé uniquement les protéines ne possédant pas de domaines transmembranaires après le 68^{e} acide aminé (si un domaine transmembranaire est prédit avant le 68ème acide aminé, cela peut correspondre à un peptide d'adressage qui sera clivé et perdu à l'entrée de l'organite). Pour Targetp2, Deeploc et Localizer nous avons récupéré les prédictions et probabilités d'adressage. Ard2 donne pour chaque séquence la probabilité P de chaque acide aminé d'être un coude. Ainsi nous sélectionnons les acides aminés avec P > 0.10 qui est la valeur par défaut. De plus, si l'on dispose d'une « région de coudes » nous considérons qu'il faut sélectionner l'acide aminé avec la plus grande probabilité de coude. Pour cela nous avons utilisé une fenêtre glissante. Cette fenêtre parcourt chaque acide aminé et son résultat par pas de 6 acides aminés. Au sein de cette fenêtre nous prenons l'acide aminé qui possède la plus forte probabilité lorsqu'elle est supérieure à 0.10, ainsi que sa position au sein de la séquence. Pour le logiciel Radar il a fallu regarder si les répétitions se chevauchent, auquel cas c'est l'union des deux répétitions chevauchant qui est considérée comme une répétition unique (annexe 4). Pour Wolfpsort nous avons choisi de récupérer les scores associés aux adressages des organites lorsqu'il y avait une prédiction de ce type, la valeur 0 est prise dans le cas contraire. Ces scores ont ensuite été additionnés puis normalisés sur chaque séquence en le divisant par le score total, différent pour chaque protéine. Finalement, pour regrouper tous nos résultats nous avons fabriqué une matrice contenant par lignes les identifiants protéiques et pour chaque colonne les résultats des descripteurs (parsing des logiciels Radar et Ard2, ACC et fréquences en acide aminé). Ce sont

[10] Sorties possibles : SP = signal peptide, mTP = mitochondrial transit peptide, cTP = chloroplast transit peptide, iTP = thylakoid transit peptide, NoTP = aucun signal

ces résultats que nous allons utiliser pour l'apprentissage et la prédiction de notre modèle. Chaque protéine sera ainsi décrite par 61 valeurs pour la construction du modèle RF. Ensuite nous utiliserons les résultats de prédiction pour filtrer les candidats prédits par le modèle.

La distribution des propriétés ainsi définies est présentée figure 5. Les tests de comparaison de moyennes entre les deux jeux sont tous significatifs (p-value < 0.05) : les propriétés des protéines positives et négatives sont bien différentes. Ainsi les protéines du témoin positif comportent plus de coudes entre paires d'hélices (figure 5A), plus de régions répétées (figure 5B) et sont globalement plus prédites comme adressées au chloroplaste et/ou mitochondrie que les protéines du témoin négatif (figure C, D, E et F). On notera des efficacités variables pour la prédiction d'adressage.

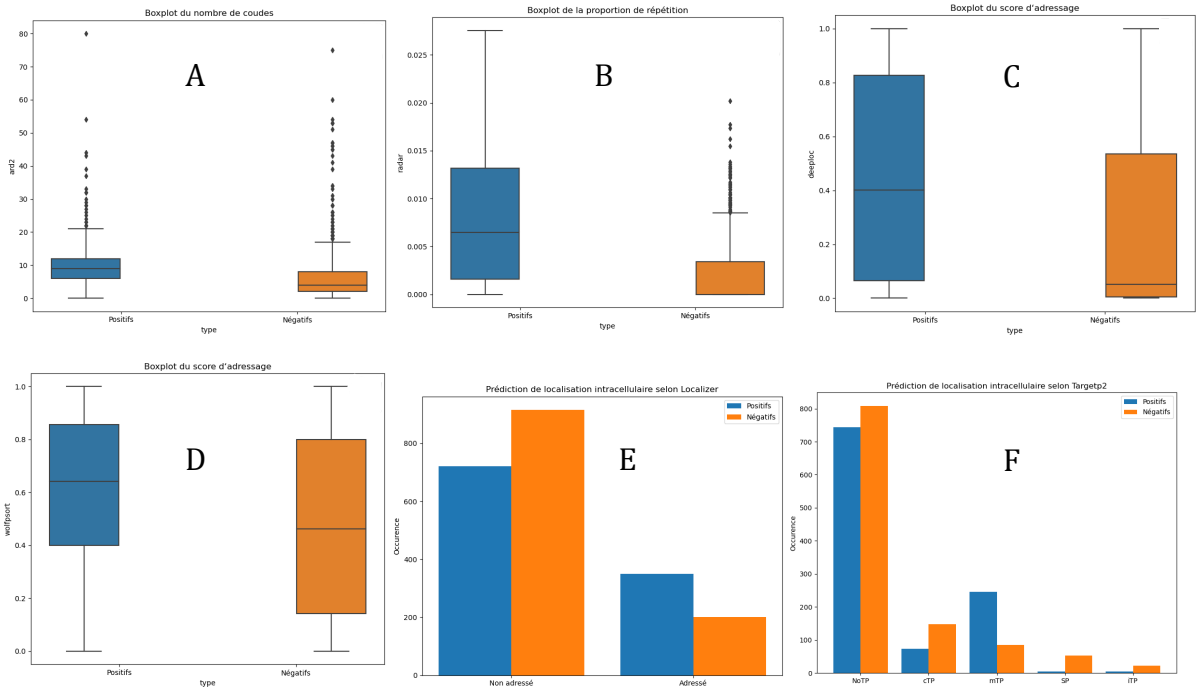


Figure 5 : Boxplots des résultats des logiciels selon les jeux positifs (bleu) et négatifs (orange). (A) Ard2 (nombre de coudes). (B) Radar (proportion de séquence en répétition). (C) Deeploc (probabilité d'adressage aux organites). (D) Wolfpsort (score d'aressage). (E) Localizer (occurrence des protéines adressées ou non aux organites). (F) Targetp2 (occurrence des protéines adressées aux différents compartiments).

III. 1) Mise en œuvre du modèle

Nous utiliserons les données comme présentées ci-dessous :

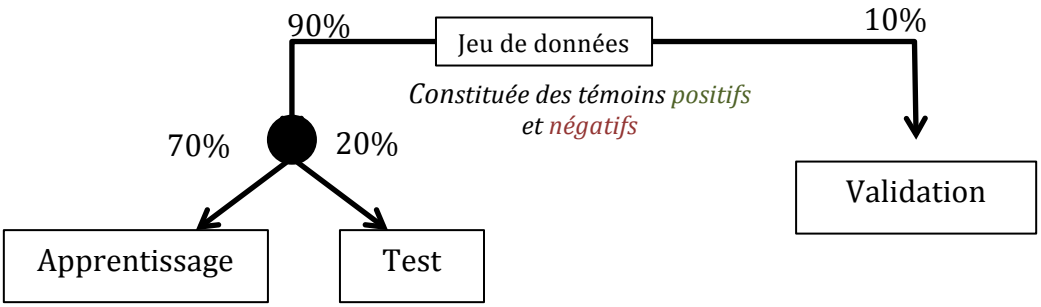


Figure 6 : Méthode d'échantillonnage des données positives et négatives. Les 10% de l'échantillon de validation permettront le calcul des performances du modèle.

Pour ce qui est du modèle en soit nous avons décidé d'utiliser les paramètres suivants qui permettaient d'avoir les meilleures performances possibles (table 4) :

	Définition	**Valeur (ou attribut)**
criterion	Fonction de mesure la qualité d'un échantillon	« gini » (défault)
n_estimator	Nombre d'arbres de la forêt	500**
min_samples_split	Nombre minimum de données dans l'échantillon provenant de la dataset	30**
min_samples_leaf	Nombre minimum d'échantillons à un nœud	1 (défault)
max_features	Nombre de descripteurs pris en compte par échantillons	« auto » (défault)

Table 4 : Paramètres utilisés pour la RF. Les autres paramètres non précisés sont laissés en défaut.
** : Les paramètres utilisés sont ceux obtenus parmi une gamme paramétrique (de 100 à 800 pour n_estimator et de 5 à 50 pour min_sample_split) permettant les meilleures performances possibles durant l'apprentissage du modèle.

Ensuite nous avons calculé les performances de notre modèle sur nos témoins, visibles ci-dessous :

	Précision	**Sensibilité**	**Spécificité**
Moyenne du modèle	0.94	0.94	0.92

Table 5 : Performances du modèle RF. Moyenne des performances calculées sur le jeu de validation obtenues sur 1000 essais du modèle.

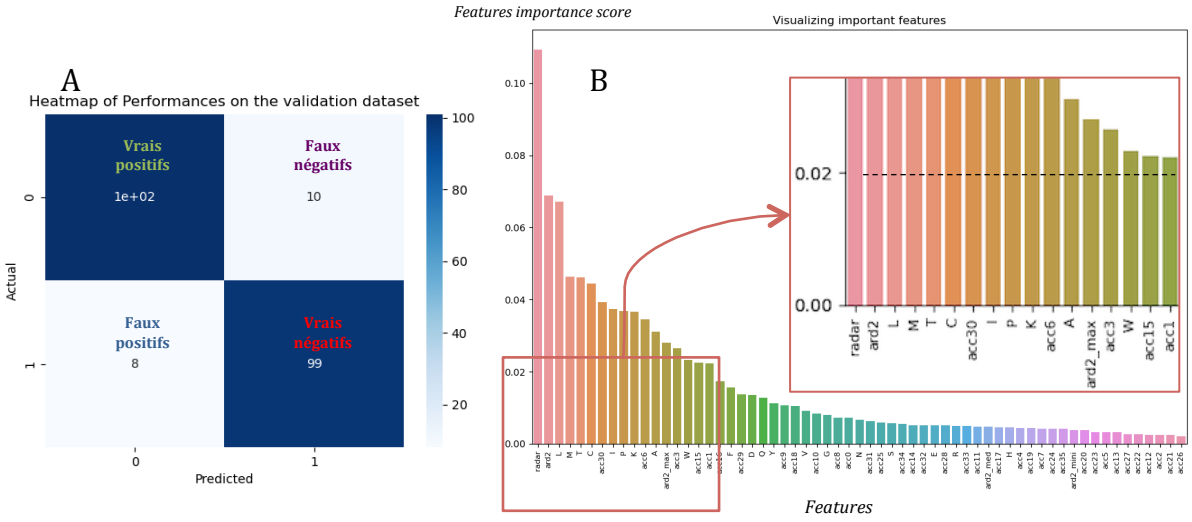


Figure 7 : (A) Heatmap des performances sur le jeu de validation. Actual = vraie nature de la protéine. Predicted = nature prédite de la protéine par le modèle RF. 0 = positif et 1 = négatif. (B) Barplot de l'importance des descripteurs. Les descripteurs ayant un score élevé contribuent le plus aux bonnes prédictions.

Ainsi notre modèle a de bonnes performances avec un faible taux de mauvaises prédictions et une précision de 94% (figure 7A et table 5). De plus à l'aide de l'attribut python « .feature_importances_() » nous pouvons déterminer les descripteurs les plus importants pour le modèle, donc ceux qui sont les plus décisifs pour classifier nos protéines en α-solénoïdes (figure 7B). Nous remarquons alors que les propriétés les plus importantes pour le modèle (score > 0.02) sont les répétitions de séquences (Radar) et les coudes d'hélice α (Ard2) ce qui est cohérent puisque les répétitions au sein des ROGEs forment des paires d'hélice α. Il est aussi intéressant de noter que la proportion de certains acides aminés au sein des séquences est importante pour les prédictions : L (Leucine), C (Cystéine), T (Thréonine), M (Méthionine), K (Lysine), I (Isoleucine), P (Proline), A (Adénine), W (Tryptophane). Ces acides aminés sont pour la majorité apolaires (L, M, I, P, A, W). Une hypothèse est que les séquences permettant la fixation à l'ARN étant composée d'acides aminés polaires[11], ce sont les résidus externes à ce cœur hydrophile qui sont reconnus par le modèle.

[11] L'ARN étant chargé négativement l'interaction avec la protéine reste logique.

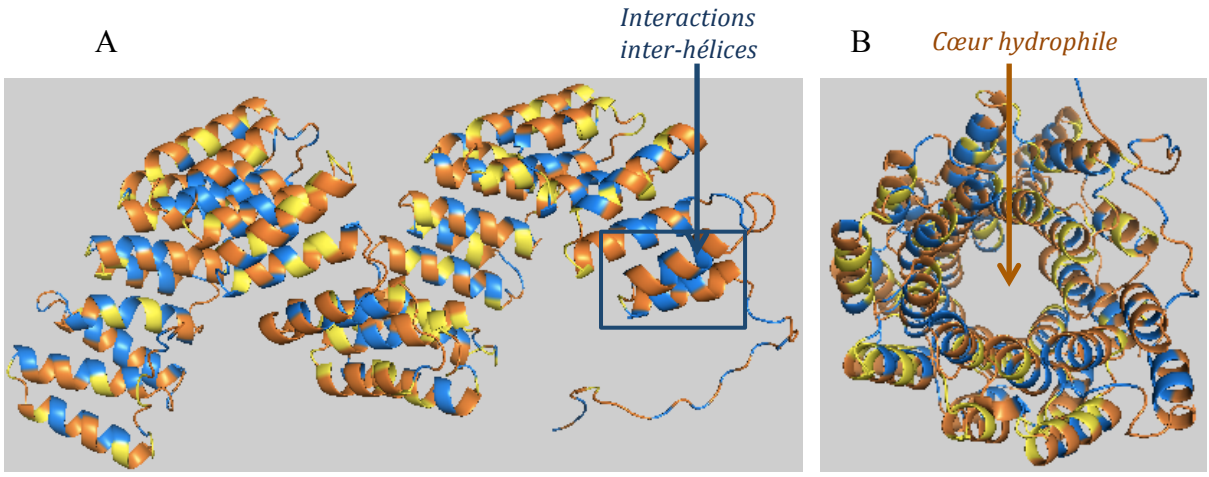


Figure 8 : Protéine ROGEs en α-solénoïde PPR de *A. thaliana* (NP_1718531) et résolue par Alpha fold. (A) Vue de côté. (B) Vue du dessus. Bleu : acides aminés apolaires (score > 0.02). Orange : acides aminés polaires (score > 0.02). Jaune : autres acides aminés. Les résidus apolaires sont situés à l'intérieur des hélices d'une même paire et permettent ainsi leur interaction et le maintient de la structure de la protéine. Les résidus polaires sont situés au niveau de la cavité de la protéine et permettent la liaison à l'ARN. Les résidus apolaires aideraient donc à la formation et au maintient d'un motif d'hélice α.

Ainsi la détection de la structure en hélice α (dont les séquences sont plus conservées que le cœur de la protéine) par l'intermédiaire des acides aminés apolaires permet de mieux reconnaître nos protéines cibles. Il y a donc une meilleure conservation des propriétés des résidus d'interaction hélice-hélice d'un même motif que ceux de la liaison avec l'ARN. Le modèle se base préférentiellement sur la face apolaire des hélices et démontre le caractère amphipatique des hélices de chacun des motifs des protéines en α-solénoïdes avec une face hydrophile (permettant la liaison à l'ARN) et une face hydrophobe (maintenant sa structure caractéristique), figure 8. [12] Une hypothèse concernant l'importance de la Lysine (K, unique polaire chargée positivement) pour le modèle est qu'elle aurait ainsi un rôle important quant à l'interaction avec l'ARN qui est chargé négativement. Une image de la position de la Lysine au sein de la PPR de la figure 8 est donnée en annexe 5. On remarque que la lysine est présente non seulement dans la « région cœur » de la protéine mais aussi au sein des hélices externes à ce cœur, principalement aux extrémités des chaînes. Nous n'avons pas d'hypothèse quant à ce constat.

Les ACC dont l'importance est supérieure à 0.02 (figure 7B) sont présentés dans la table 6 (la correspondance entre numéros d'ACC, lag et Z-scales est donnée en annexe 6).

	Acc30	Acc15	Acc6	Acc3	Acc1
Facteurs	Z_1-Z_2	Z_2-Z_1	Z_1-Z_1	Z_1-Z_1	Z_2-Z_2
Lag	4	1	3	2	1
Face	A	B	A	B	B

Table 6 : ACC déterminants pour la prédiction du modèle. En vert : Z-scale le plus récurrent. En rouge : lag le plus récurrent. Fond bleu : mêmes facteurs de Z-scales. Les faces A et B représentent les deux faces d'une même hélice α.

Nous remarquons que la covariance au lag 1 est plus présente que celle des autres lags. De même pour la covariance entre même facteurs de Z-scales. De plus dans quasiment tous les cas de figures le premier Z-scales (Z_1) est impliqué dans ces ACC déterminants. Ces observations signifieraient que ces ACC permettent de différencier plus facilement les protéines en α-solénoïdes des autres. Le Z_1 caractérisant l'hydrophobicité (annexe 3.1), cela est cohérent avec les hypothèses faites plus haut à savoir que le caractère hydrophobe a un fort pouvoir discriminant.

Il semble que la récurrence de facteurs de Z-scales et lag précis reflète le caractère amphiphile des hélices d'une protéine [12]. Il existe ainsi des combinaisons de Z-scales dans le calcul d'ACC qui mettent en évidence des patterns et propriétés des résidus. Étant donné qu'un tour d'hélice est constitué de 3.6 acides aminés les résidus aux lags 1 et 2 sont positionnés sur la même face d'une hélice α. De la même manière les résidus aux lags 3 et 4 sont positionnés sur les faces opposées [12]. Dans l'article [12, supplementary data] il a été montré que les propriétés hydrophobes et stériques correspondant aux ACC : Z_1-Z_1, Z_2-Z_2, Z_1-Z_2 et Z_2-Z_1 permettent de distinguer les propriétés amphipatiques d'une hélice α. Nous retrouvons les mêmes résultats dans la table 6 (face B). Ces termes d'ACC reflètent alors l'hydrophobicité et l'encombrement stérique sur une même face d'hélice. Cela nous permet de valider notre constat : les acides aminés apolaires d'une même face permettent le maintien de la structure de la protéine. De plus ils ont un fort taux de conservation au sein des séquences et sont ainsi d'une grande utilité pour la prédiction des ROGEs.

III. 2) Résultats obtenus sur les protéomes de *Chlamydomonas reinhardtii* et d'*Arabidopsis thaliana*

Ensuite nous avons utilisé notre modèle sur *Chlamydomonas reinhardtii* et *Arabidopsis thaliana*, à partir d'une matrice de 51 755 lignes et 61 colonnes contenant les propriétés de chacune des protéines de ces deux organismes dont un exemple de résultat est visible en annexe 7. Nous avons alors pu prédire au sein des deux protéomes les protéines α-solénoïdes. Ensuite pour ne garder que les protéines ROGEs (α-solénoïdes adressées aux organites) nous n'avons conservé que celles prédites comme adressées au moins par deux logiciels de prédiction d'adressage intracellulaire. Les résultats sont résumés dans la table ci-dessous (table 7). Enfin pour avoir une meilleure compréhension de nos résultats nous les avons comparés avec les jeux de donnés positifs et négatifs ainsi que ceux obtenus avec l'arbre de décision développé précédemment, ainsi qu'avec une approche basée sur la recherche de similarité des motifs répétés (annexe 8.1 et 8.2).

	A. thaliana	***C. reinhardtii***	**Total**
Taille du protéome	36699	15081	51780
Nombre d'α-solénoïdes prédites (RF)	1301	1896	3197
Dont prédites aux organites	752	1015	1767
Dont annotées	736	453	1189
Contenant les mots-clés (parmi les annotées)	578	146	724
Dont non annotées	16	562	578

Table 7 : Résultats du modèle de prédiction sur les organismes modèles *A. thaliana* et *C. reinhardtii*. En mettant en commun les résultats des différentes approches on obtient in fine 1448 nouvelles candidates de protéines ROGEs sur 51780 (soit ~1.4% du génome d'*A. thaliana* et ~6.25% du génome de *C. reinhardtii*). En vert : protéines prédites par le modèle. En bleu : protéines candidates ROGEs finales (après filtrage). En rouge : résultats des annotations par eggNOG.

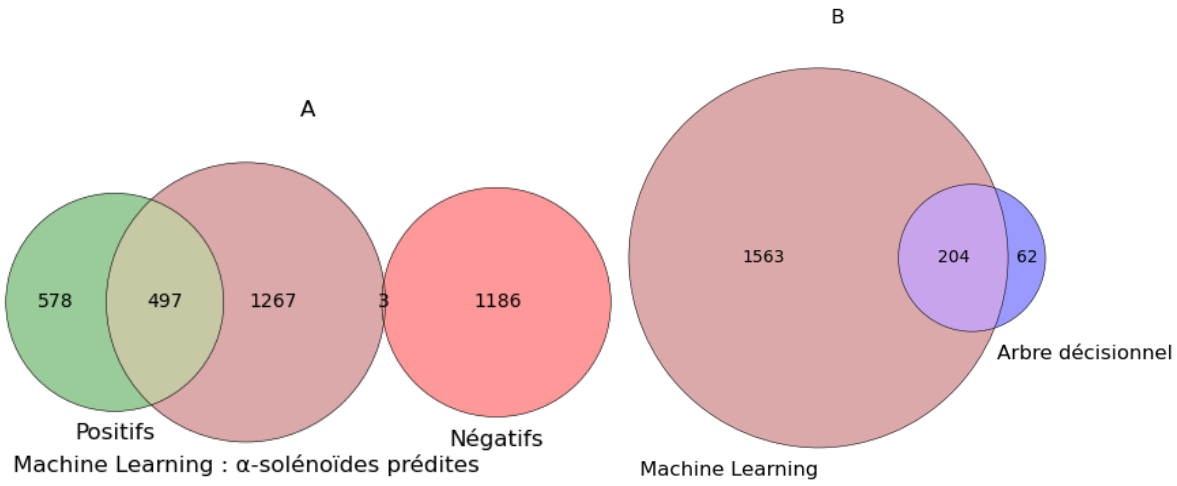


Figure 9 : Comparaison des résultats (A) avec les témoins positifs et négatifs. Seulement les organismes de *C. reinhardtii* et *A. thaliana* ont été pris en compte dans le jeu positif de ce diagramme. (B) avec les résultats obtenus avec l'arbre décisionnel. En rose : résultats obtenus par machine learning. En vert : témoin positif. En rouge témoins négatif. En bleu : résultats obtenus par arbre décisionnel.

Ces résultats ont montré que parmi les 1767 protéines candidates ROGEs déterminées par RF et prédites comme adressées aux organites, il y avait déjà 497 protéines connues en α-solénoïde et 3 que l'on sait ne pas être en α-solénoïde (figure 9A). Cela correspond aux performances de notre modèle (94% de précision) donc nous pouvons penser que les résultats sont cohérents. Il s'agit donc d'investiguer sur les 1267 protéines restantes. Notre modèle trouve beaucoup d'OPR et de PPR connues chez les deux organismes (annexe 9) et la majorité des protéines candidates ROGEs identifiées par l'arbre décisionnel (figure 9B).

Afin de déterminer s'il existait des familles de paralogues dans chacune des deux espèces parmi les protéines prédites nous avons regroupé les protéines sur la base de leur similarité de séquence. Nous avons lancé un blast sur les protéomes contre eux même. Chacune des protéines sera comparée avec toutes les protéines du protéome. Ensuite nous remplaçons la valeur de la Evalue par le $-\log_{10}$(Evalue) pour visualiser le réseau de similarité avec Cytoscape[12]. Les images des clusters sont disponibles en annexe (annexe 10). La figure montre qu'il y a des clusters comprenant à la fois des OPR (ou des PPR) et des protéines nouvellement identifiées par notre approche, ce qui suggère que ces protéines sont aussi des OPR (ou PPR) non annotées précédemment. Surtout, il est très intéressant de voir qu'il existe aussi des clusters constitués uniquement de protéines nouvellement identifiées par notre

[12] Cela permet d'avoir des poids entre protéines pour permettre à Cytoscape de faire son clustering.

approche, suggérant qu'il existe d'autres familles de ROGEs, en plus des OPR (ou PPR) présentes chez *C. reinhardtii* (ou *A. thaliana*).

Enfin pour avoir une annotation de nos protéines candidates ROGEs nouvellement identifiés (en rouge table 7) nous utilisons eggNOG-mapper. Nous récupérons donc pour chaque candidate ROGE, une protéine orthologue[13], la Evalue entre la protéine étudiée et l'orthologue trouvée, la description de cette protéine ainsi que sa composition en domaines protéiques conservés dans PFAM (base de données de domaines protéiques). Nous avons également recherché une liste de mots clés en lien avec la fixation à l'ARN[14] parmi ces annotations. Les résultats sont visibles en rouge table 7.
Enfin sur les protéines non annotées nous effectué une prédiction de la structure 3D avec Alpha-fold dont un exemple de résultat est visible en annexe 11.

III. 3) Résultats sur la diatomée *Phaeodactylum tricornutum*

Nous avons ensuite appliqué le modèle pour explorer le protéome de la diatomée *Phaeodactylum tricornutum*. Avant tout nous avons voulu vérifier qu'il n'y avait pas de différences fondamentales entre les diatomées, *C.reinhardtii* et *A. thaliana* en terme d'ACC et de fréquences d'acides aminés, sans quoi notre modèle ne serait pas adapté. D'après la figure 10 ci-dessous on remarque qu'il n'y a pas de cluster spécifique des espèces sur la base des ACC et des fréquences en acides aminés pour chacun des protéomes. Nous pouvons donc appliquer notre modèle à *P. tricornutum*. Les logiciels d'adressage utilisés plus haut ne sont pas adaptés aux diatomées pour la localisation au chloroplaste car les séquences d'adressage au chloroplaste des diatomées sont différentes de celles des Archaeplastida. Ceci s'explique sur le fait qu'ils ne sont pas issus du même évènement d'endosymbiose. Nous utiliserons donc Hectar (outil de prédiction d'adressage au chloroplaste adapté diatomées) et Deeploc pour la prédiction à la mitochondrie comme filtre final sur les protéines déterminées comme ROGEs par le modèle.

[13] Deux protéines sont orthologues si elles sont issues d'un ancêtre commun. Cela signifie qu'elles sont similaires.

[14] Liste de mots clés (avec et sans majuscule) : chlo (pour chloroplaste), mito (pour mitochondrie), RNA, binding, GTP, ATP, Synthase, helicase, transferase, protease, maturation, exonuc, endonuc, ribo, armadillo, tetra, penta, octa, transcript, traduction, histone, rubis (pour rubisco qui participe à la photosynthèse), repeat, ribo (ribosomal), mTERF (facteur 1 de la transcription mitochondriale), PPDK (protéine du chloroplaste), AMPK (enzyme du métabolisme énergétique), GRAS (facteur de transcription régulant le développement des plantes).

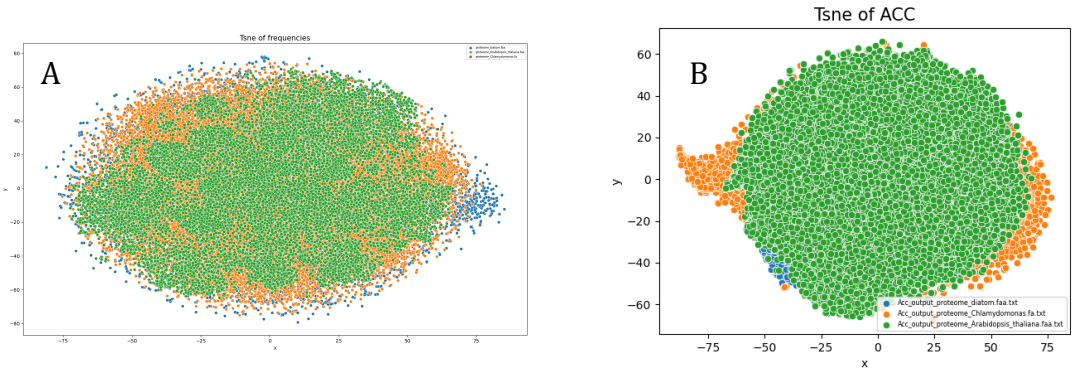


Figure 10 : (A) t-sne sur les fréquences. Vert : *Chlamydomonas reinhardtii*, orange : *Aradopsis thaliana*, bleu : protéome de différentes diatomées. (B) Tsne sur les ACC. Vert : *Arabidopsis thaliana*, orange : *Chlamydomonas reinhardtii*, bleu : protéome de plusieurs diatomées. Chaque couleur correspond à un protéome et chaque point à une protéine de ce protéome.

À la suite de cela nous avons pu appliquer notre modèle sur la matrice pour *P. tricornutum* (8396 lignes et 61 colonnes) dont les résultats sont disponibles ci-dessous (table 8).

	P. tricornutum
Taille du protéome	10681
α-solénoïdes totales	107
Dont prédites aux organites	26
Dont annotées	13
Dont non annotées	13

Table 8 : Résultats obtenus du modèle sur *P. tricornutum*. On trouve 26 protéines ROGEs, soit ~0.24% du protéome.

Ces résultats sont en accord avec les résultats préliminaires obtenus par Alexis Astatourian, un ancien stagiaire de M2 qui avait utilisé la méthode basée sur la similarité de séquences pour annoter les OPR et les PPR chez les diatomées. Il trouvait 3 OPR et 55 PPR dans le protéome de *P. tricornutum*, mais n'avait pas effectué la prédiction d'adressage aux organites. Notre méthode est donc plus stringente mais plus précise.

Les autres protéomes de diatomées pourront ainsi être explorés avec le même protocole afin d'y déterminer la présence de ROGEs. Le modèle pourra aussi être testé sur différents organismes. Le but étant d'annoter ainsi l'ensemble des protéomes disponibles des eucaryotes photosynthétiques.

IV. Conclusion & Discussion

Nous avons donc pu développer un bon modèle de prédiction des protéines ROGEs, d'après les résultats obtenus pour deux organismes modèles d'Archaeplastida : *A. thaliana* et *C. reinhardtii*. Toutefois il se peut qu'il fasse encore des erreurs ou ne sélectionne pas les protéines voulues. Par exemple, il existe la PPR7 de *C. reinhardtii* [45] qui n'est pas reconnue par le modèle. Il existe aussi une protéine intéressante : la TCA1 (Translation of Cytochrome b(6)f complex pet1 mRNA) [46], qui ne fait pas partie des OPR ni des PPR, mais qui est également impliquée dans la régulation de l'expression du génome du chloroplaste de *Chlamydomonas reinhardtii*. Celle-ci est reconnue par notre modèle et est bien prédite comme adressée aux organites. Ainsi nos résultats montrent que TCA1 se replie en α-solénoïde et est donc une ROGE. Aussi les méthodes de prédiction de localisation subcellulaire ne sont pas toujours fiables, et peuvent donc nous amener à éliminer des protéines de notre jeu de données qui sont pourtant des ROGEs. En parallèle la comparaison de nos résultats avec ceux obtenus par l'arbre de décision et la recherche par similarité de séquence montrent que l'on peut récupérer des protéines ROGEs différentes. Cela nous amène à penser que nos méthodes se complètent et permettent d'identifier une large gamme de protéines ROGEs candidates.

Nous avons fait tourner ce modèle 1000 fois sur *C. reinhardtii* et *A. thaliana* et regardé la moyenne des performances sur le jeu de validation. Nous avons aussi déterminé la fréquence de prédictions ROGEs pour chaque protéine, afin d'éviter les biais de prédiction du modèle. Parmi les protéines prédites majoritairement nous retrouvons bien TCA1. Par contre, TCA1 ne faisait partie d'aucun cluster de protéines. L'étude des clusters obtenus a montré une similarité entre les OPR d'une part et les PPR d'autre part des différents organismes, et a permis d'identifier des familles de protéines non connues comme ROGEs. Elles seraient de nouvelles candidates ROGEs probables.

Parmi les protéines candidates dont nous avons prédit la structure avec Alpha fold, quelques unes ne ressemblaient pas à des α-solénoïdes. Cela signifierait qu'il existe des protéines ayant des mêmes caractéristiques structurelles que les protéines recherchées mais qui n'en sont pas. Les prédictions d'Alpha fold confirment, comme attendu, que les motifs répétés ne coïncident pas forcément avec les paires d'hélices α et soulignent la pertinence de combiner la détection

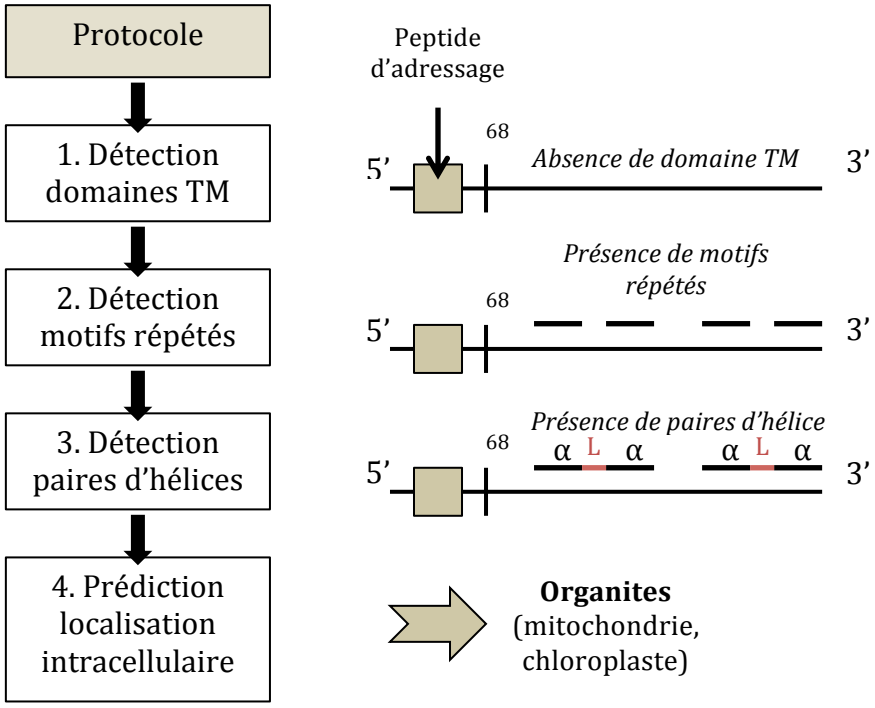
des motifs répétés dans la séquence (ici Radar) et des répétitions d'hélices α (ici Ard2) pour la construction du modèle.

Enfin, notre modèle peut être utilisé sur tous types d'organismes, avec un filtrage final selon la localisation intracellulaire en dernier lieu avec un logiciel adapté[15].
Le fait que le modèle retrouve encore moins de ROGE dans *Phaeodactylum tricornutum* que l'approche par similarité de séquence laisse supposer que très peu de ROGEs adressées aux organites sont présentes dans les diatomées. Il a été observé que seulement ~0.24% du protéome de *Phaeodactylum tricornutum* contient des ROGEs, contre ~1.4% chez *Arabidopsis thaliana* et ~6.25% chez *Chlamydomonas reinhardtii*. Cela peut être dû au fait que *A. thaliana* et *C. reinhardtii* font partie des Archaeplatida et que leur chloroplaste leur provient d'une endosymbiose primaire. Tandis que *P. tricornutum* a acquis son chloroplaste d'une endosymbiose secondaire. Il est donc probable qu'il existe d'autres protéines que celles que nous connaissons impliquées dans la régulation post-transcriptionelle du chloroplaste chez les diatomées.

La prédiction systématique de structure 3D des ROGEs candidats est en cours, et permettra de valider ces prédictions. Les meilleurs candidats adressés au chloroplaste chez *C. reinhardtii* et *P. tricornutum* pourront être caractérisés fonctionnellement au laboratoire, afin de confirmer leur adressage, et de préciser leur rôle au sein du chloroplaste.

[15] Pour *P. tricornutum* nous avons utilisé Hectar.

ANNEXES



Annexe 1 : Protocole mis en place pour identifier les protéines en α-solénoïdes. Si une protéine passe tous ces filtres alors elle est candidate pour faire partie des protéines ROGEs. La présence d'hélice α est détectée par l'intermédiaire d'acides aminés « linker » (L) ou coudes qui relient deux hélices entre elles.

Python	**R**
Panda, Numpy, os, glob, seaborn, basename, matplotlib, statistics, glob, operator, itemgetter, sklearn, TSNE scipy, dist, shapiro, f, RandomForestClassifier, GridSearchCV, confusion_matrix	*Protr, Biostrings, optparse, stringr, M3C, Rtsne, RcolorBrewer, seqinr*

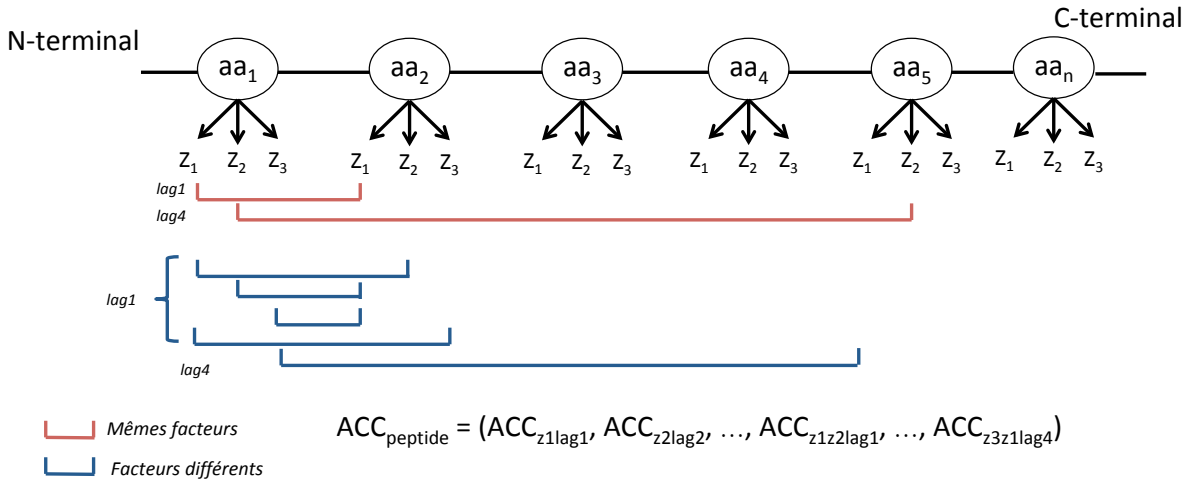
Annexe 2.1 : Modules utilisés pour les analyses bio-informatiques.

Lien Github : https://github.com/Rebbekkah/Stage_M2.git

Annexe 2.2 : Lien vers le Github du stage contenant les scripts.

	Z_1	Z_2	Z_3
Caractéristique	Hydrophobicité (caractère apolaire)	Encombrement stérique	Propriétés électroniques

Annexe 3.1 : Z-scales et les propriétés qu'elles caractérisent.



Annexe 3.2 : Représentation du calcul des ACC pour un peptide donné au lag 4. Un peptide est caractérisé par r*9 = 4*9 = 36 valeurs d'ACC.

Mêmes facteurs

$$ACC_{Zj,lagr} = \sum_{i}^{N-r} \frac{z_{j,i} * z_{j,i} + r}{N - r}$$

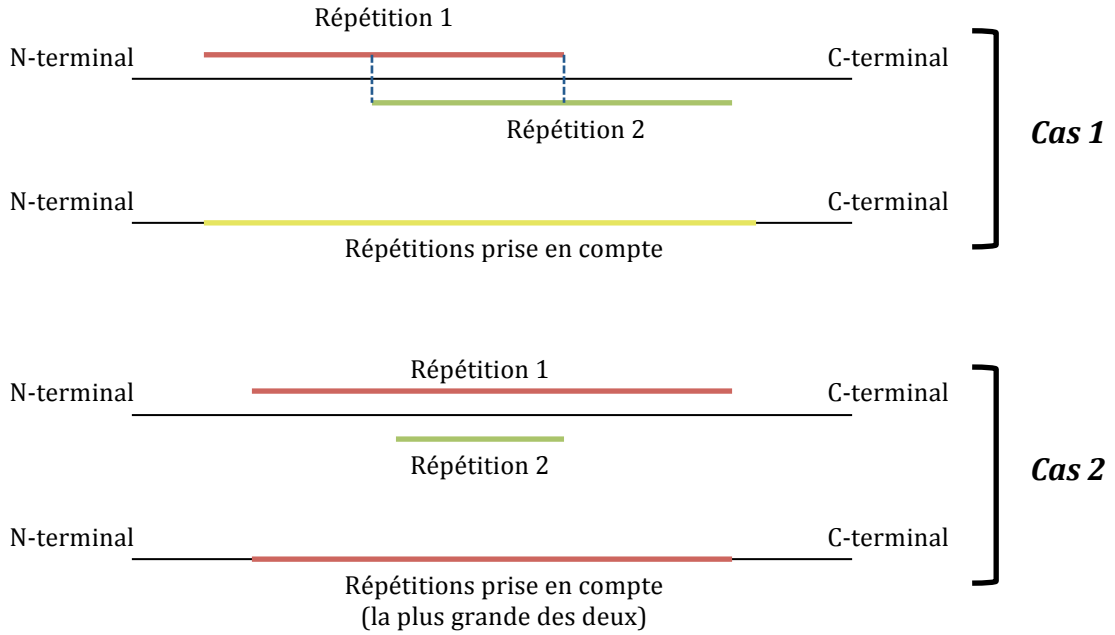
Même facteur j, position i

Facteurs différents

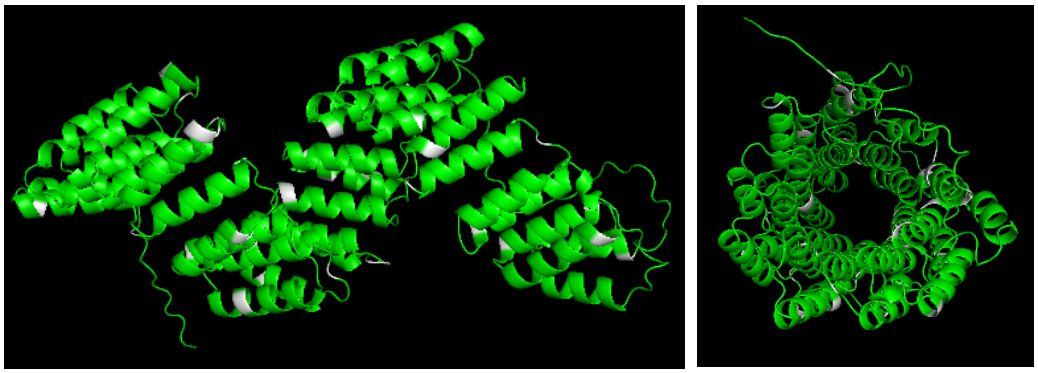
$$ACC_{ZjZk,lagr} = \sum_{i}^{N-r} \frac{z_{j,i} * z_{k,i} + r}{N - r}$$

Facteurs j et k, position i

Annexe 3.3 : Formule du calcul des ACC.



Annexe 4 : Exemple de répétitions chevauchantes trouvées par Radar. Les répétitions 1 et 2 sont deux types de répétitions différentes détectées.



Annexe 5 : Position de la Lysine (K) en gris au sein de la PPR NP_1718531 chez *A. thaliana*. Il y a des résidus Lysine dans le cœur hydrophile de la protéine, et sont situées principalement en bout de chaînes.

Acc0	Z1-Z1 lag1	**Acc18**	Z1-Z2 lag2
Acc1	Z2-Z2 lag1	**Acc19**	Z1-Z3 lag2
Acc2	Z3-Z3 lag1	**Acc20**	Z2-Z3 lag2
Acc3	Z1-Z1 lag2	**Acc21**	Z2-Z1 lag2
Acc4	Z2-Z2 lag2	**Acc22**	Z3-Z1 lag2

Acc5	Z3-Z3 lag2	**Acc23**	Z3-Z2 lag2
Acc6	Z1-Z1 lag3	**Acc24**	Z1-Z2 lag3
Acc7	Z2-Z2 lag3	**Acc25**	Z1-Z3 lag3
Acc8	Z3-Z3 lag3	**Acc26**	Z2-Z3 lag3
Acc9	Z1-Z1 lag4	**Acc27**	Z2-Z1 lag3
Acc10	Z2-Z2 lag4	**Acc28**	Z3-Z1 lag3
Acc11	Z3-Z3 lag4	**Acc29**	Z3-Z2 lag3
Acc12	Z1-Z2 lag1	**Acc30**	Z1-Z2 lag4
Acc13	Z1-Z3 lag1	**Acc31**	Z1-Z3 lag4
Acc14	Z2-Z3 lag1	**Acc32**	Z2-Z3 lag4
Acc15	Z2-Z1 lag1	**Acc33**	Z2-Z1 lag4
Acc16	Z3-Z1 lag1	**Acc34**	Z3-Z1 lag4
Acc17	Z3-Z2 lag1	**Acc35**	Z3-Z2 lag4

Annexe 6 : Table de correspondance des ACC.

Index	**Radar**	**Ard2_nombre_linker**	**Ard2_proba_min**	**Ard2_proba_med**	**Ard2_proba_max**
Identifiant Protéique	Proportion de la séquence en répétition[16]	Nombre de linker	Probabilité minimale d'un acide aminé d'être linker	Médiane des probabilités pour chaque acide aminé d'être linker	Probabilité maximale d'un acide aminé d'être linker
>Cre06.g251750.t1.2 (C. reinhardtii)	0.0	0.0	0.0	0.0	0.0
>Cre15.g640101.t1.1 (C. reinhardtii)	0.0012406947890818	24.0	0.11	0.175	0.88

(Suite de la dataframe)

Acc_0	**Acc_1**	**…**	**Acc_{35}**	**M**	**Q**
Première valeur d'ACC sur Z-scales	Seconde valeur d'ACC sur Z-scales	…	36ième valeur d'ACC sur Z-scales	Fréquence d'apparition au sein de la séquence	Fréquence d'apparition au sein de la séquence
1.25371729323308	0.557835338345865	…	ACC_{35}	0.887958461538462	F_Q
0.571304930095659	1.8240744665195	…	ACC_{35}	0.0811762799263352	0.179464198895028

(Suite de la dataframe)

A	**…**	**Acide aminé X**
Fréquence	Fréquence	Fréquence d'apparition

[16] Pour des répétitions de longueur comprises entre 29 et 46. En effet ce sont des types de répétitions caractéristiques des OPR, TPR et PPR.

d'apparition au sein de la séquence	d'apparition au sein de la séquence	au sein de la séquence
F_A	…	F_x
F_A	…	F_x

Annexe 7 : Exemple pour deux protéines de la matrice finale obtenue. >Cre06.g251750.t1.2 est une protéine non prédite en α solénoïde par le modèle et >Cre15.g640101.t1.1en est une. Il existe 4 descripteurs liés à Ard2 qui caractérisent le nombre de coudes (« linker »), la probabilité minimale d'un acide aminé de la séquence d'être linker, la médiane des probabilités et la probabilité maximale trouvée pour un acide aminé d'être un coude.

	A. thaliana	**C. reinhardtii**	**Total**
Taille du protéome	36699	15081	51780
Nombre d'α-solénoïdes prédites (RF)	1301	1896	3197
Après filtrage sur les résultats RF	752	1015	1767
Nombre d'α-solénoïdes prédites (Céline Cattelin)	192	74	266
Nouvelles protéines déterminées (RF)[17]	107	244	351
Nouvelles protéines déterminées (RF + Céline Cattelin)	505	943	1448

Annexe 8.1 : Tableau complet des comparaisons de méthodes.

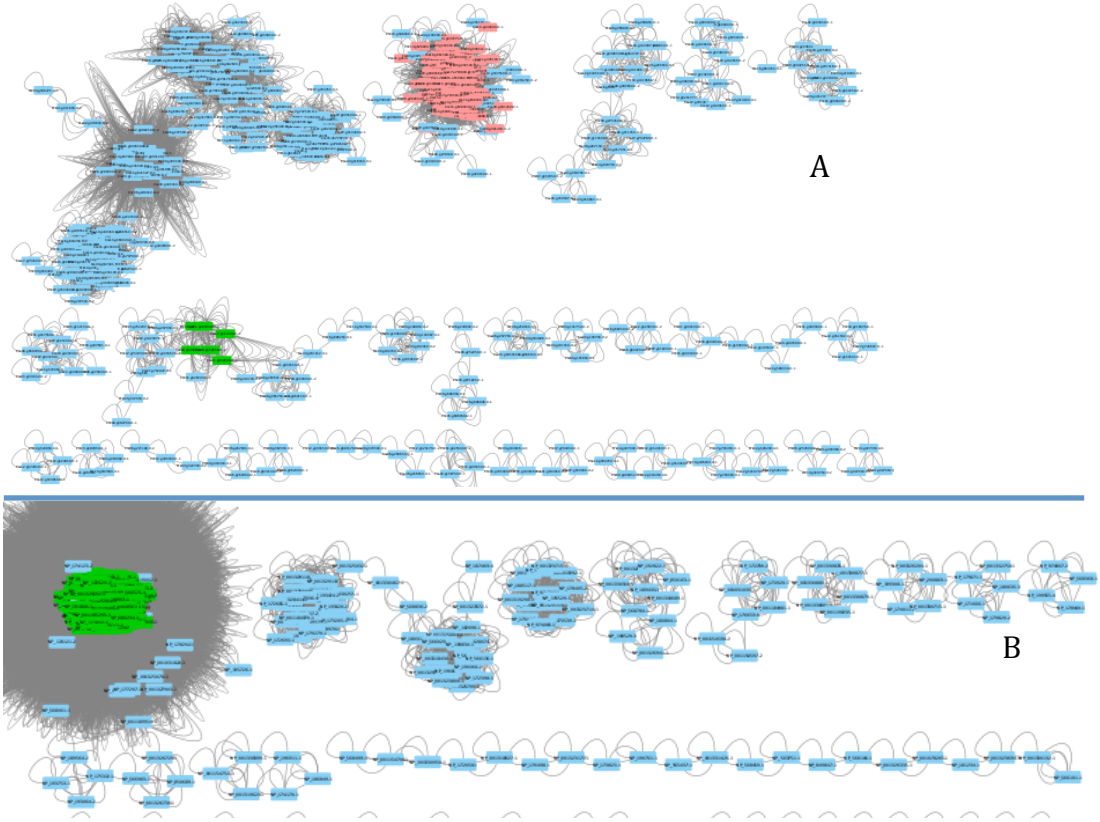
1. Obtention de jeux de motifs OPR & PPR.
2. Comparaison par Blast. On récupère les protéines qui ont une Evalue inférieure au seuil.
3. Clustering des motifs.
4. Alignement des motifs au sein de chaque cluster.
5. Construction de profils HMM grâce aux alignements puis recherche de motifs similaires dans les protéomes.
6. Filtre sur l'adressage aux organites.
7. Est-ce une protéine inconnue ? Si oui on l'ajoute au jeu de données de base utilisé sinon le programme se termine. On obtient au final un nouveau jeu de protéines candidates.

Annexe 8.2 : Protocole de Céline Cattelin pour la détection d'α-solénoïdes ROGE sur la base de leurs motifs.

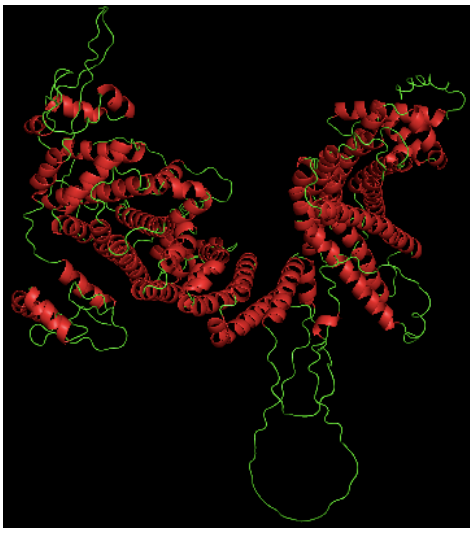
[17] Protéines qui n'étaient pas connues comme ROGEs jusqu'alors (non comprises dans le jeu de données positif ni dans les résultats de Céline Cattelin). Ce sont les réelles nouvelles protéines candidates.

	Modèle		Connues	
	C. reinhardtii	*A. thaliana*	*C. reinhardtii*	*A. thaliana*
OPR	90	0	106	1
PPR	6	476	8	686
Total	96	476	114	687

Annexe 9 : OPR et PPR déterminées par le modèle. Chez *C. reinhardtii* le modèle RF détecte ~85% des OPR et ~75% des PPR connues. Chez *A. thaliana* il détecte 0% des OPR et ~69% des PPR connues.



Annexe 10 : Résultats obtenus avec la recherche de cluster OPR et PPR chez (A) *C. Reinhardtii* et (B) *A. thaliana*. En rouge : les protéines OPR. En vert : les protéines PPR.



Annexe 11 : Prédiction par Alpha-fold de Cre17.g704400.t1.1 (*C. reinhardtii*).

RÉFÉRENCES

[1] Guo, Manyun; Ma, Yucheng; Liu, Wanyuan; Yuan, Zuyi (2022) "A computational method for predicting nucleocapsid protein in retroviruses", Scientific Reports, Numéro 1, Volume 12, 10.1038/s41598-021-03182-2.

[2] Sandaruwan, Pahalage Dhanushka; Wannige, Champi Thusangi (2021) "An improved deep learning model for hierarchical classification of protein families", PLOS ONE, Numéro 10, Volume 16, 10.1371/journal.pone.0258625.

[3] Zhao, Bi; Katuwawala, Akila; Oldfield, Christopher J; Dunker, A Keith; Faraggi, Eshel; Gsponer, Jörg; Kloczkowski, Andrzej; Malhis, Nawar; Mirdita, Milot; Obradovic, Zoran; Söding, Johannes; Steinegger, Martin; Zhou, Yaoqi; Kurgan, Lukasz (2021) "DescribePROT: database of amino acid-level protein structure and function predictions", Nucleic Acids Research, Numéro D1, Volume 49, 10.1093/nar/gkaa931.

[4] Takenaka, Mizuki; Takenaka, Sachi; Barthel, Tatjana; Frink, Brody; Haag, Sascha; Verbitskiy, Daniil; Oldenkott, Bastian; Schallenberg-Rüdinger, Mareike; Feiler, Christian G.; Weiss, Manfred S.; Palm, Gottfried J.; Weber, Gert (2021) "DYW domain structures imply an unusual regulation principle in plant organellar RNA editing catalysis", Nature Catalysis, Numéro 6, Volume 4, 10.1038/s41929-021-00633-x.

[5] Macedo-Osorio, Karla S.; Martínez-Antonio, Agustino; Badillo-Corona, Jesús A. (2021) "Pas de Trois: An Overview of Penta-, Tetra-, and Octo-Tricopeptide Repeat Proteins From Chlamydomonas reinhardtii and Their Role in Chloroplast Gene Expression", Frontiers in Plant Science, Volume 12, 10.3389/fpls.2021.775366.

[6] Hollin, Thomas; Jaroszewski, Lukasz; Stajich, Jason E.; Godzik, Adam; Le Roch, Karine G. (2021) "Identification and phylogenetic analysis of RNA binding domain abundant in apicomplexans or RAP proteins", Microbial Genomics, Numéro 3, Volume 7, 10.1099/mgen.0.000541.

[7] Gutmann, Bernard; Royan, Santana; Schallenberg-Rüdinger, Mareike; Lenz, Henning; Castleden, Ian R.; McDowell, Rose; Vacher, Michael A.; Tonti-Filippini, Julian; Bond, Charles S.; Knoop, Volker; Small, Ian D. (2020) "The Expansion and Diversification of Pentatricopeptide Repeat RNA-Editing Factors in Plants", Molecular Plant, Numéro 2, Volume 13, 10.1016/j.molp.2019.11.002.

[8] Burki, Fabien; Roger, Andrew J.; Brown, Matthew W.; Simpson, Alastair G.B. (2020) "The New Tree of Eukaryotes", Trends in Ecology & Evolution, Numéro 1, Volume 35, 10.1016/j.tree.2019.08.008.

[9] Falciatore, Angela; Jaubert, Marianne; Bouly, Jean-Pierre; Bailleul, Benjamin; Mock, Thomas (2020) "Diatom Molecular Research Comes of Age: Model Species for Studying Phytoplankton Biology and Diversity", The Plant Cell, Numéro 3, Volume 32, 10.1105/tpc.19.00158.

[10] Falciatore, Angela; Jaubert, Marianne; Bouly, Jean-Pierre; Bailleul, Benjamin; Mock, Thomas (2020) "Diatom Molecular Research Comes of Age: Model Species for Studying Phytoplankton Biology and Diversity", The Plant Cell, Numéro 3, Volume 32, 10.1105/tpc.19.00158.

[11] DEI, University of Padua, viale Gradenigo 6, Padua, Italy; Nanni, L.; Brahnam, S.; Information Technology and Cybersecurity, Missouri State University, 901 S. National, Springfield, MO 65804, USA (2020) "Set of Approaches Based on Position Specific Scoring Matrix and Amino Acid Sequence for Primary Category Enzyme Classification", Journal of Artificial Intelligence and Systems, Numéro 1, Volume 2, 10.33969/AIS.2020.21004.

[12] Garrido, Clotilde; Caspari, Oliver D.; Choquet, Yves; Wollman, Francis-André; Lafontaine, Ingrid (2020) "Evidence Supporting an Antimicrobial Origin of Targeting Peptides to Endosymbiotic Organelles", Cells, Numéro 8, Volume 9, 10.3390/cells9081795.

[13] Chatzimparmpas, Angelos; Martins, Rafael M.; Kerren, Andreas (2020) "t-viSNE: Interactive Assessment and Interpretation of t-SNE Projections", IEEE Transactions on Visualization and Computer Graphics, Numéro 8, Volume 26, 10.1109/TVCG.2020.2986996.

[14] McInnes, Leland; Healy, John; Melville, James (2020) "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction", arXiv:1802.03426 [cs, stat] .

[15] Graham, Jill B.; Canniff, Nathan P.; Hebert, Daniel N. (2019) "TPR-containing proteins control protein organization and homeostasis for the endoplasmic reticulum", Critical Reviews in Biochemistry and Molecular Biology, Numéro 2, Volume 54, 10.1080/10409238.2019.1590305.

[16] Ponce‐Toledo, Rafael I.; López‐García, Purificación; Moreira, David (2019) "Horizontal and endosymbiotic gene transfer in early plastid evolution", New Phytologist, Numéro 2, Volume 224, 10.1111/nph.15965.

[17] Ponce‐Toledo, Rafael I.; López‐García, Purificación; Moreira, David (2019) "Horizontal and endosymbiotic gene transfer in early plastid evolution", New Phytologist, Numéro 2, Volume 224, 10.1111/nph.15965.

[18] Lv, Zhibin; Jin, Shunshan; Ding, Hui; Zou, Quan (2019) "A Random Forest Sub-Golgi Protein Classifier Optimized via Dipeptide and Amino Acid Composition Features", Frontiers in Bioengineering and Biotechnology, Volume 7, 10.3389/fbioe.2019.00215.

[19] Hakala, Kai; Kaewphan, Suwisa; Bjorne, Jari; Mehryary, Farrokh; Tolvanen, Martti; Salakoski, Tapio; Ginter, Filip (2019) "Neural network and random forest models in protein function prediction".

[20] Luttrell, Joseph; Liu, Tong; Zhang, Chaoyang; Wang, Zheng (2019) "Predicting protein residue-residue contacts using random forests and deep networks", BMC Bioinformatics, Numéro S2, Volume 20, 10.1186/s12859-019-2627-6.

[21] Luttrell, Joseph; Liu, Tong; Zhang, Chaoyang; Wang, Zheng (2019) "Predicting protein residue-residue contacts using random forests and deep networks", BMC Bioinformatics, Numéro S2, Volume 20, 10.1186/s12859-019-2627-6.

[22] Hakala, Kai; Kaewphan, Suwisa; Bjorne, Jari; Mehryary, Farrokh; Tolvanen, Martti; Salakoski, Tapio; Ginter, Filip (2019) "Neural network and random forest models in protein function prediction".

[23] Almagro Armenteros, Jose Juan; Salvatore, Marco; Emanuelsson, Olof; Winther, Ole; von Heijne, Gunnar; Elofsson, Arne; Nielsen, Henrik (2019) "Detecting sequence signals in targeting peptides using deep learning", Life Science Alliance, Numéro 5, Volume 2, 10.26508/lsa.201900429.

[24] Rovira, Aleix Gorchs; Smith, Alison G. (2019) "PPR proteins – orchestrators of organelle RNA metabolism", Physiologia Plantarum, Numéro 1, Volume 166, 10.1111/ppl.12950.

[25] Dong, Shanshan; Zhao, Chaoxian; Zhang, Shouzhou; Wu, Hong; Mu, Weixue; Wei, Tong; Li, Na; Wan, Tao; Liu, Huan; Cui, Jie; Zhu, Ruiliang; Goffinet, Bernard; Liu, Yang (2019) "The Amount of RNA Editing Sites in Liverwort Organellar Genes Is Correlated with GC Content and Nuclear PPR Protein Diversity", Genome Biology and Evolution, Numéro 11, Volume 11, 10.1093/gbe/evz232.

[26] Salomé, Patrice A.; Merchant, Sabeeha S. (2019) "A Series of Fortunate Events: Introducing Chlamydomonas as a Reference Organism", The Plant Cell, Numéro 8, Volume 31, 10.1105/tpc.18.00952.

[27] Hillebrand, Arne; Matz, Joachim M; Almendinger, Martin; Müller, Katja; Matuschewski, Kai; Schmitz-Linneweber, Christian (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", Nucleic Acids Research, 10.1093/nar/gky710.

[28] Seo, Seokjun; Oh, Minsik; Park, Youngjune; Kim, Sun (2018) "DeepFam: deep learning based alignment-free method for protein family modeling and prediction", Bioinformatics, Numéro 13, Volume 34, 10.1093/bioinformatics/bty275.

[29] Kathuria, Charu; Mehrotra, Deepti; Misra, Navnit Kumar (2018) "Predicting the protein structure using random forest approach", Procedia Computer Science, Volume 132, 10.1016/j.procs.2018.05.134.

[30] Taherzadeh, Ghazaleh; Zhou, Yaoqi; Liew, Alan Wee-Chung; Yang, Yuedong (2018) "Structure-based prediction of protein– peptide binding regions using Random Forest", Bioinformatics, Numéro 3, Volume 34, 10.1093/bioinformatics/btx614.

[31] Schietgat, Leander; Vens, Celine; Cerri, Ricardo; Fischer, Carlos N.; Costa, Eduardo; Ramon, Jan; Carareto, Claudia M. A.; Blockeel, Hendrik (2018) "A machine learning based framework to identify and classify long terminal repeat retrotransposons", PLOS Computational Biology, Numéro 4, Volume 14, 10.1371/journal.pcbi.1006097.

[32] Zieliński, Bartosz; Plichta, Anna; Misztal, Krzysztof; Spurek, Przemysław; Brzychczy-Włoch, Monika; Ochońska, Dorota (2017) "Deep learning approach to bacterial colony classification", PLOS ONE, Numéro 9, Volume 12, 10.1371/journal.pone.0184554.

[33] Sarica, Alessia; Cerasa, Antonio; Quattrone, Aldo (2017) "Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review", Frontiers in Aging Neuroscience, Volume 9, 10.3389/fnagi.2017.00329.

[34] Sperschneider, Jana; Catanzariti, Ann-Maree; DeBoer, Kathleen; Petre, Benjamin; Gardiner, Donald M.; Singh, Karam B.; Dodds, Peter N.; Taylor, Jennifer M. (2017) "LOCALIZER: subcellular localization prediction of both plant and effector proteins in the plant cell", Scientific Reports, Numéro 1, Volume 7, 10.1038/srep44598.

[35] Sperschneider, Jana; Catanzariti, Ann-Maree; DeBoer, Kathleen; Petre, Benjamin; Gardiner, Donald M.; Singh, Karam B.; Dodds, Peter N.; Taylor, Jennifer M. (2017) "LOCALIZER: subcellular localization prediction of both plant and effector proteins in the plant cell", Scientific Reports, Numéro 1, Volume 7, 10.1038/srep44598.

[36] Benoiston, Anne-Sophie; Ibarbalz, Federico M.; Bittner, Lucie; Guidi, Lionel; Jahn, Oliver; Dutkiewicz, Stephanie; Bowler, Chris (2017) "The evolution of diatoms and their biogeochemical functions", Philosophical Transactions of the Royal Society B: Biological Sciences, Numéro 1728, Volume 372, 10.1098/rstb.2016.0397.

[37] Turesson, Hjalmar K.; Ribeiro, Sidarta; Pereira, Danillo R.; Papa, João P.; de Albuquerque, Victor Hugo C. (2016) "Machine Learning Algorithms for Automatic Classification of Marmoset Vocalizations", PLOS ONE, Numéro 9, Volume 11, 10.1371/journal.pone.0163041.

[38] Fučíková, Karolina; Lewis, Paul O.; Lewis, Louise A. (2016) "Chloroplast phylogenomic data from the green algal order Sphaeropleales (Chlorophyceae, Chlorophyta) reveal

complex patterns of sequence evolution", Molecular Phylogenetics and Evolution, Volume 98, 10.1016/j.ympev.2016.01.022.

[39] Wu, Yunzhe; Xun, Qingqing; Guo, Yi; Zhang, Jinghua; Cheng, Kaili; Shi, Tao; He, Kai; Hou, Suiwen; Gou, Xiaoping; Li, Jia (2016) "Genome-Wide Expression Pattern Analyses of the Arabidopsis Leucine-Rich Repeat Receptor-Like Kinases", Molecular Plant, Numéro 2, Volume 9, 10.1016/j.molp.2015.12.011.

[40] Cheng, Shifeng; Gutmann, Bernard; Zhong, Xiao; Ye, Yongtao; Fisher, Mark F.; Bai, Fengqi; Castleden, Ian; Song, Yue; Song, Bo; Huang, Jiaying; Liu, Xin; Xu, Xun; Lim, Boon L.; Bond, Charles S.; Yiu, Siu‐Ming; Small, Ian (2016) "Redefining the structural motifs that determine editing by pentatricopeptide repeat proteins in land plants".

[41] Provart, Nicholas J.; Alonso, Jose; Assmann, Sarah M.; Bergmann, Dominique; Brady, Siobhan M.; Brkljacic, Jelena; Browse, John; Chapple, Clint; Colot, Vincent; Cutler, Sean; Dangl, Jeff; Ehrhardt, David; Friesner, Joanna D.; Frommer, Wolf B.; Grotewold, Erich; Meyerowitz, Elliot; Nemhauser, Jennifer; Nordborg, Magnus; Pikaard, Craig; Shanklin, John; Somerville, Chris; Stitt, Mark; Torii, Keiko U.; Waese, Jamie; Wagner, Doris; McCourt, Peter (2016) "50 years of Arabidopsis research: highlights and future directions", New Phytologist, Numéro 3, Volume 209, 10.1111/nph.13687.

[42] Boulouis, Alix; Drapier, Dominique; Razafimanantsoa, Hélène; Wostrikoff, Katia; Tourasse, Nicolas J.; Pascal, Kevin; Girard-Bascou, Jacqueline; Vallon, Olivier; Wollman, Francis-André; Choquet, Yves (2015) "Spontaneous Dominant Mutations in Chlamydomonas Highlight Ongoing Evolution by Gene Diversification", The Plant Cell, Numéro 4, Volume 27, 10.1105/tpc.15.00010.

[43] Marx, Christina; Wünsch, Christiane; Kück, Ulrich (2015) "The Octatricopeptide Repeat Protein Raa8 Is Required for Chloroplast Splicing", Eukaryotic Cell, Numéro 10, Volume 14, 10.1128/EC.00096-15.

[44] Libbrecht, Maxwell W.; Noble, William Stafford (2015) "Machine learning applications in genetics and genomics", Nature Reviews Genetics, Numéro 6, Volume 16, 10.1038/nrg3920.

[45] Jalal, Abdullah; Schwarz, Christian; Schmitz-Linneweber, Christian; Vallon, Olivier; Nickelsen, Jörg; Bohne, Alexandra-Viola (2015) "A Small Multifunctional Pentatricopeptide Repeat Protein in the Chloroplast of Chlamydomonas reinhardtii", Molecular Plant, Numéro 3, Volume 8, 10.1016/j.molp.2014.11.019.

[46] Wang, Fei; Johnson, Xenie; Cavaiuolo, Marina; Bohne, Alexandra-Viola; Nickelsen, Joerg; Vallon, Olivier (2015) "Two Chlamydomonas OPR proteins stabilize chloroplast mRNAs encoding small subunits of photosystem II and cytochrome b6f", The Plant Journal, Numéro 5, Volume 82, 10.1111/tpj.12858.

[47] Gruber, Ansgar; Rocap, Gabrielle; Kroth, Peter G.; Armbrust, E. Virginia; Mock, Thomas (2015) "Plastid proteome prediction for diatoms and other algae with secondary plastids of the red lineage", The Plant Journal, Numéro 3, Volume 81, 10.1111/tpj.12734.

[48] Barkan, Alice; Small, Ian (2014) "Pentatricopeptide Repeat Proteins in Plants", Annual Review of Plant Biology, Numéro 1, Volume 65, 10.1146/annurev-arplant-050213-040159.

[49] Jahandideh, Samad; Jaroszewski, Lukasz; Godzik, Adam (2014) "Improving the chances of successful protein structure determination with a random forest classifier", Acta Crystallographica Section D Biological Crystallography, Numéro 3, Volume 70, 10.1107/S1399004713032070.

[50] Mohan, Abhilash; Divya Rao, M.; Sunderrajan, Shruthi; Pennathur, Gautam (2014) "Automatic classification of protein structures using physicochemical parameters", Interdisciplinary Sciences: Computational Life Sciences, Numéro 3, Volume 6, 10.1007/s12539-013-0199-0.

[51] Zhao, Nan; Han, Jing Ginger; Shyu, Chi-Ren; Korkin, Dmitry (2014) "Determining Effects of Non-synonymous SNPs on Protein-Protein Interactions using Supervised and Semi-supervised Learning", PLoS Computational Biology, Numéro 5, Volume 10, 10.1371/journal.pcbi.1003592.

[52] Ge, Changrong; Spånning, Erika; Glaser, Elzbieta; Wieslander, Åke (2014) "Import Determinants of Organelle-Specific and Dual Targeting Peptides of Mitochondria and Chloroplasts in Arabidopsis thaliana", Molecular Plant, Numéro 1, Volume 7, 10.1093/mp/sst148.

[53] Fournier, David; Palidwor, Gareth A.; Shcherbinin, Sergey; Szengel, Angelika; Schaefer, Martin H.; Perez-Iratxeta, Carol; Andrade-Navarro, Miguel A. (2013) "Functional and Genomic Analyses of Alpha-Solenoid Proteins", PLoS ONE, Numéro 11, Volume 8, 10.1371/journal.pone.0079894.

[54] Tourasse, Nicolas J; Choquet, Yves; Vallon, Olivier (2013) "PPR proteins of green algae", RNA Biology, Numéro 9, Volume 10, 10.4161/rna.26127.

[55] He, Ronghai; Ma, Haile; Zhao, Weirui; Qu, Wenjuan; Zhao, Jiewen; Luo, Lin; Zhu, Wenxue (2012) "Modeling the QSAR of ACE-Inhibitory Peptides with ANN and Its Applied Illustration", International Journal of Peptides, Volume 2012, 10.1155/2012/620609.

[56] Leliaert, Frederik; Smith, David R.; Moreau, Hervé; Herron, Matthew D.; Verbruggen, Heroen; Delwiche, Charles F.; De Clerck, Olivier (2012) "Phylogeny and Molecular Evolution of the Green Algae", Critical Reviews in Plant Sciences, Numéro 1, Volume 31, 10.1080/07352689.2011.615705.

[57] Fukui, Kenji; Kuramitsu, Seiki (2011) "Structure and Function of the Small MutS-Related Domain", Molecular Biology International, Volume 2011, 10.4061/2011/691735.

[58] Lin, Hao; Chen, Wei (2011) "Prediction of thermophilic proteins using feature selection technique", Journal of Microbiological Methods, Numéro 1, Volume 84, 10.1016/j.mimet.2010.10.013.

[59] Fukui, Kenji; Kuramitsu, Seiki (2011) "Structure and Function of the Small MutS-Related Domain", Molecular Biology International, Volume 2011, 10.4061/2011/691735.

[60] Fujii, Sota; Small, Ian (2011) "The evolution of RNA editing and pentatricopeptide repeat genes", New Phytologist, Numéro 1, Volume 191, 10.1111/j.1469-8137.2011.03746.x.

[61] Madera, M.; Calmus, R.; Thiltgen, G.; Karplus, K.; Gough, J. (2010) "Improving protein secondary structure prediction using a simple k-mer model", Bioinformatics, Numéro 5, Volume 26, 10.1093/bioinformatics/btq020.

[62] Jain, Pooja; Hirst, Jonathan D (2010) "Automatic structure classification of small proteins using random forest", BMC Bioinformatics, Numéro 1, Volume 11, 10.1186/1471-2105-11-364.

[63] Kappel, Christian; Zachariae, Ulrich; Dölker, Nicole; Grubmüller, Helmut (2010) "An Unusual Hydrophobic Core Confers Extreme Flexibility to HEAT Repeat Proteins", Biophysical Journal, Numéro 5, Volume 99, 10.1016/j.bpj.2010.06.032.

[64] Gschloessl, Bernhard; Guermeur, Yann; Cock, J Mark (2008) "HECTAR: A method to predict subcellular targeting in heterokonts", BMC Bioinformatics, Numéro 1, Volume 9, 10.1186/1471-2105-9-393.

[65] Horton, P.; Park, K.-J.; Obayashi, T.; Fujita, N.; Harada, H.; Adams-Collier, C.J.; Nakai, K. (2007) "WoLF PSORT: protein localization predictor", Nucleic Acids Research, Volume 35, 10.1093/nar/gkm259.

[66] Krogh, Anders; Larsson, Björn; von Heijne, Gunnar; Sonnhammer, Erik L.L (2001) "Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen", Journal of Molecular Biology, Numéro 3, Volume 305, 10.1006/jmbi.2000.4315.

[67] Heger, Andreas; Holm, Liisa (2000) "Rapid automatic detection and alignment of repeats in protein sequences", Proteins: Structure, Function, and Genetics, Numéro 2, Volume 41, 10.1002/1097-0134 (20001101) 41:2<224::AID-PROT70>3.0.CO;2-Z.

[68] Altschul, S. (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Research, Numéro 17, Volume 25, 10.1093/nar/25.17.3389.

[69] Wold, S.; Jonsson, J.; Sjörström, M.; Sandberg, M.; Rännar, S. (1993) "DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures", Analytica Chimica Acta, Numéro 2, Volume 277, 10.1016/0003-2670 (93) 80437-P.

[70] Hellberg, Sven; Sjoestroem, Michael; Skagerberg, Bert; Wold, Svante (1987) "Peptide quantitative structure-activity relationships, a multivariate approach", Journal of Medicinal Chemistry, Numéro 7, Volume 30, 10.1021/jm00390a003.

RESUMÉ

Un modèle de prédiction protéique par machine learning a été conçu pour annoter, sans recourir à la recherche de similarité de séquence, des protéines en α-solénoïdes adressées au chloroplaste et à la mitochondrie chez les organismes eucaryotes photosynthétiques ; candidates à la régulation post-transcriptionelle du génome des organites. Le modèle possède de bonnes performances sur un jeu de ces protéines déjà connues chez les organismes modèles *Arabidopsis thaliana* et *Chlamydomonas reinhardtii*. L'analyse du modèle a montré que les propriétés relatives au caractère amphipathique des hélices, permettant la formation d'une cavité hydrophile au sein de la structure en α-solénoïde sont déterminantes pour la classification. Nous avons démontré la validité du modèle chez *A. thaliana* et *C. reinhardtii*, qui retrouve bien les α-solénoïdes impliquées dans la régulation des organites, et nous avons également identifié de nouvelles candidates dont la fonction reste encore inconnue à ce jour. Le faible nombre de candidates identifiées dans la diatomée *Phaeodactylum tricornutum* suggère que la régulation du génome des organites est variable entre les différents groupes d'eucaryotes photosynthétiques. Tous les nouveaux candidats identifiés chez les microalgues *C. reinhardtii* et *P. tricornutum* pourront être caractérisés fonctionnellement au laboratoire.

RESUME

A machine learning prediction model was designed to annotate, without the use of sequence similarity search, α-solenoid proteins addressed to chloroplasts and mitochondria in photosynthetic eukaryotes; candidates for transcriptional regulation of the organelle genome. The model performs well on a control dataset including such proteins already known in the model organisms, *Arabidopsis thaliana* and *Chlamydomonas reinhardtii*. Analysis of the model showed that properties related to the amphipathic nature of the helices, allowing the formation of a hydrophilic cavity within the α-solenoid structure are crucial for classification. We have demonstrated the validity of the model in *A. thaliana* and *C. reinhardtii* that retrieves known α-solenoids involved in organelle regulation, and we have also identified new candidates whose function remains unknown to date. The low number of candidates identified in the diatom *Phaeodactylum tricornutum* suggests that organelle genome regulation is variable among different groups of photosynthetic eukaryotes. All identified new candidates in the microalgae *C. reinhardtii* and *P. tricornutum* will further be functionally characterized in the laboratory.