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*

I. Introduction & état de l’art 1

II. Matériel & méthodes 5

III. Résultats 9

III. 2) Résultats obtenus sur les protéomes de *Chlamydomonas reinhardtii* et d’*Arabidopsis thaliana* 16

III. 3) Résultats sur la diatomée *Phaeodactylum tricornutium* 18

IV. Conclusion & Discussion 20

*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.

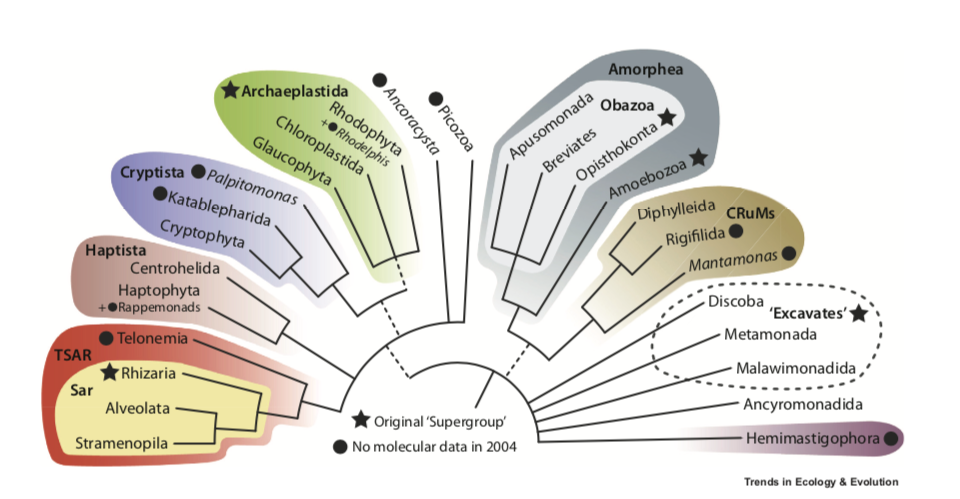
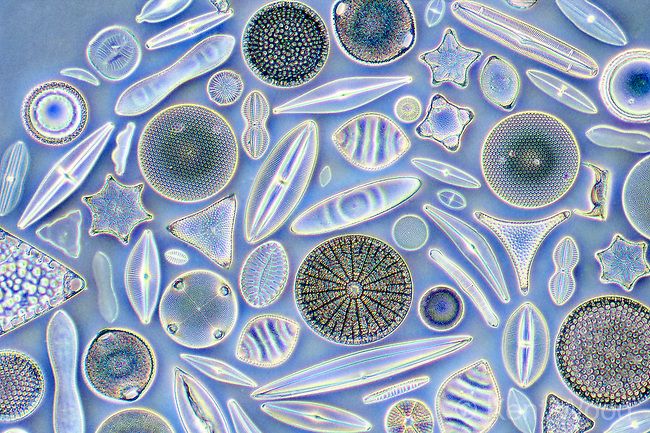
# I. Introduction & état de l’art

Au sein des eukaryotes 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 phagytose) d’une cyanobactérie par une cellule ancestrale primaire hétérotrophe (figure 1B). 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 [benoiston]. 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.

Dont algues vertes

A

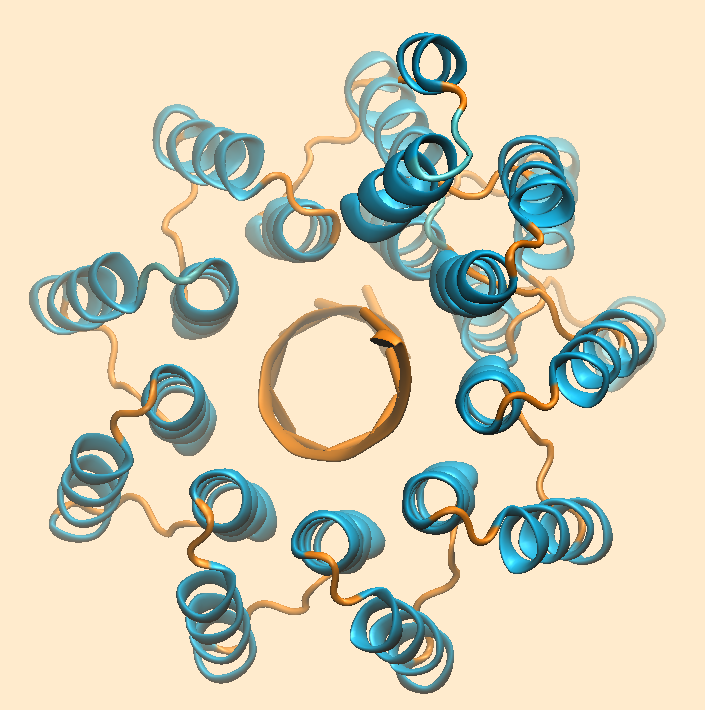
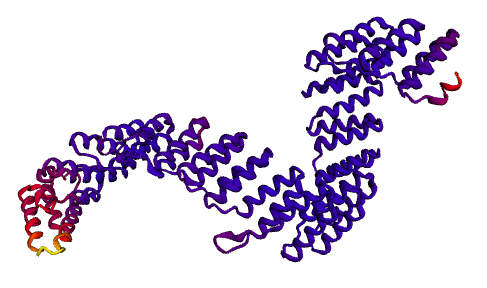
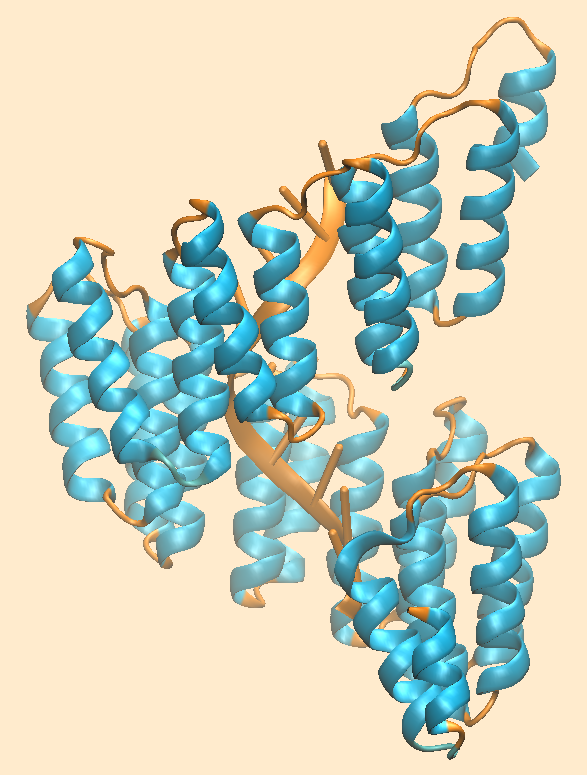
B



Dont diatomées

**Figure 1 :** (A) Diatomées observées en microscopie à contraste de phase. (B) Arbre phylogénétique des eukaryotes [Burki & al.]. Événement d’endosymbiose primaire avec une cyanobactérie. Événement d’endosymbiose primaire avec une rhodophyte.

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), issus d’une endosymbiose primaire mais beaucoup moins chez les autres eucaryotes photosynthétiques comme les diatomées. Ainsi chez les Archaeplastida 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) chloroplastique 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 [Boulouis & al.], 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.

A

C

B

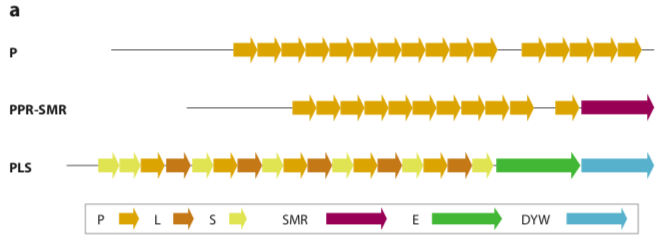
**Figure 2 :** (A) Protéine en α solénoïde prédite par alpha fold (code Uniprot : Q9LKV3, Arabidopsis thaliana). (B) Protéine en α solénoïde résolue par diffraction de rayons-X. Elle est liée à un ARNm du côté 5’, le carré noir montrant 1 motif de répétition d’hélice α en antiparallèle (code PDB : 5I9F). (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 alpha 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 [Boulouis & al., Barkan & Small] :

* 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, répétitions de 38 acides aminés (octatricopeptide repeat), qui interagissent avec ARN.

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 450 chez *Arabidopsis thaliana*) mais sont peu nombreuses chez les microalgues vertes (1 seule copie chez *Chlamydomonas reinhardtii*) [Barkan & Small.]. Les PPR sont elles mêmes classifiées en trois sous-groupes selon leur composition en motifs : P-PPR, PPR-SMR et PLS-PPR.

*Correction, protection et repliement*



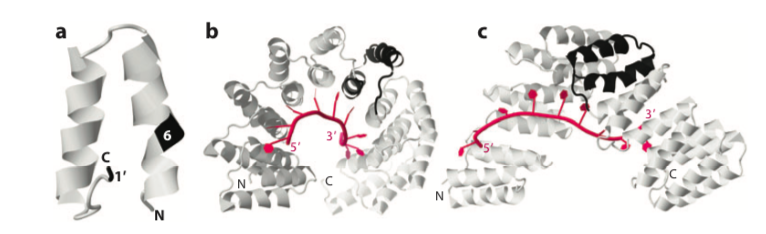
*Motifs*

*Stabilisation, clivage, épissage*

*Endonucléase*

**Figure 3 :** Les différents motifs présents chez les PPR [Barkan & Small.].

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 la fonction exacte 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 [Barkan & Small.].



B

C

A

**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 [Barkan & Small], Figure 1.

Les OPR restent peu étudiées et de ce fait moins connues que les PPR. Dans certaines OPR, il semblerait qu’il existe des motifs appelés RAP qui possèdent une activité endonucléolytique [Boulouis & al].

Les ROGEs étant peu conservées au niveau de leur séquence, il est difficile de les identifier sur la base de leur similarité 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-transcriptionnelle, 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 2). Il permet d’identifier des protéines candidates à l’aide d’un arbre décisionnel et d’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* etla plante terrestre *Arabidopsis thaliana*, puis nous l’appliquerons à la diatomée modèle *Phaeodactylum tricornutium* 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 1).

*Protéomes étudiés*

|  |  |  |  |
| --- | --- | --- | --- |
|  | Protéomes | | |
| *C. reinhardtii* | *A. thaliana* | *P. tricornutium* |
| Base de données | 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 7 logiciels :

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]](#footnote-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 adn 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 4 outils de prédiction de localisation intracellulaire :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Logiciel | **Targetp2** | **Deeploc** | **Wolfpsort** | **Localizer** |
| Méthode | Apprentissage profond (deep learning) | Réseaux neuronnaux récurrents | Machine learning | Machine learning |
| Version | 2.0 | 1.0 | 0.2 | 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*](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 [clotilde][S. Wold]. Pour chacun des 20 acides aminés on distingue trois valeurs de Z-scales (annexe 2.1). 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 [clotilde][S. Wold]. 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]](#footnote-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]](#footnote-3). Un graphe explicatif et les formules permettant de calculer les ACC sont donnés en annexe (annexe 2.2 et 2.3). 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 physique quant à la description d’un peptide.

*Localisation intracellulaire*

Targetp2 est entraîné et testé sur des modèles de séquences construits à partir de la matrice de substitution BLOSUM62. Il sert à détecter des séquences dites « signales » permettant l’adressage protéique dans différents compartiments cellulaires. Quant à Deeploc, Wolfpsort et Localizer ils permettent la prédiction d’adressage d’une protéine. En effet Deeploc distingue 10 localisations cellulaires différentes[[4]](#footnote-4) et sur la base de la séquence permet d’associer une probabilité d’adressage pour chaque localisation à une protéine. Wolpsort utilise la composition en acide aminé qu’il convertit en vecteurs pour être ensuite classifiés par méthode k-nearest neighbor[[5]](#footnote-5). L’algorithme permet donc de déterminer la localisation cellulaire d’une protéine par l’intermédiaire de sa proportion en acide aminé. Enfin, l’algorithme de Localizer a été entraîné pour prédire la localisation des protéines chez les plantes. L’adressage vers le chloroplaste ou la mitochondrie est prédite en déterminant la présence de peptide signal (ou peptide de transit) tandis que la localisation vers le noyau est prédite par un ensemble de signaux peptidiques nucléaires (NLSs). Il est donc déterminé par la présence d’un peptide de transit au niveau N-terminal de la protéine cible.

Tous ces logiciels ont été utilisés en stand alone.

*Comparaison de séquence et annotation*

Pour comparer les protéines nous utiliserons Blastp (https://www.ncbi.nlm.nih.gov/books/NBK279690/, ver sion 2.11.0), un outil de comparaison qui permet de déterminer un degré de similarité entre deux protéines. Il nous permet de récupérer une Evalue pour chaque comparaison protéine-protéine. Plus la 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[[6]](#footnote-6) 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 sur la base de séquences. Il reconnaît et utilise des séquences homologues de sa propre base de données *(*[*http://eggnog5.embl.de*](http://eggnog5.embl.de)*)* pour déterminer la fonction et les domaines des séquences d’entrées.

*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[[7]](#footnote-7). 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)[[8]](#footnote-8). 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 ces résultats ainsi que ceux des outils de Radar et Ard2 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 matrice de confusion. 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 |  |  |  |

**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 68e 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.

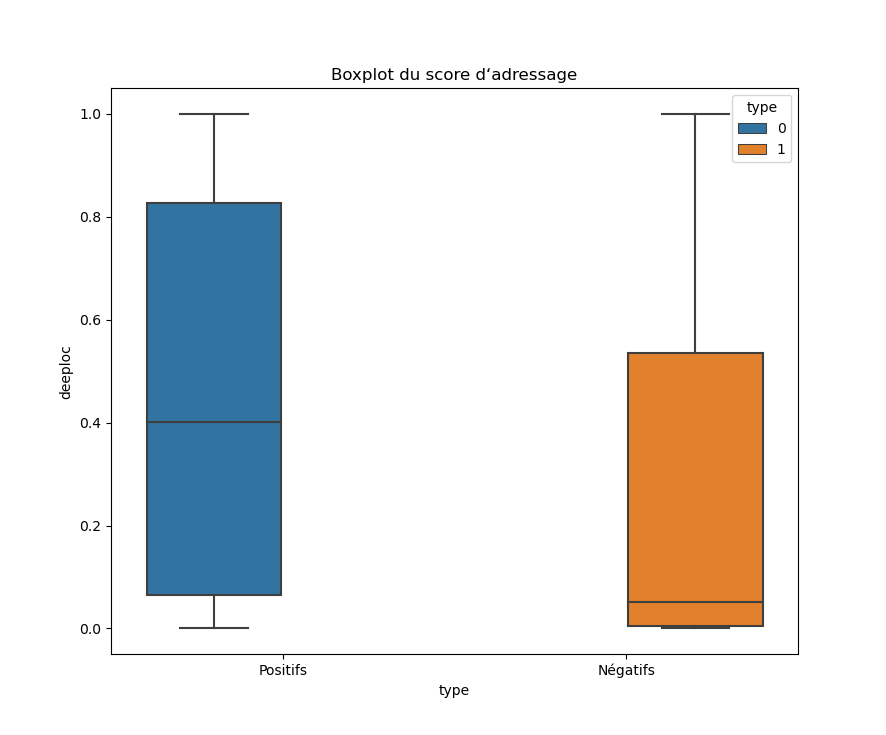
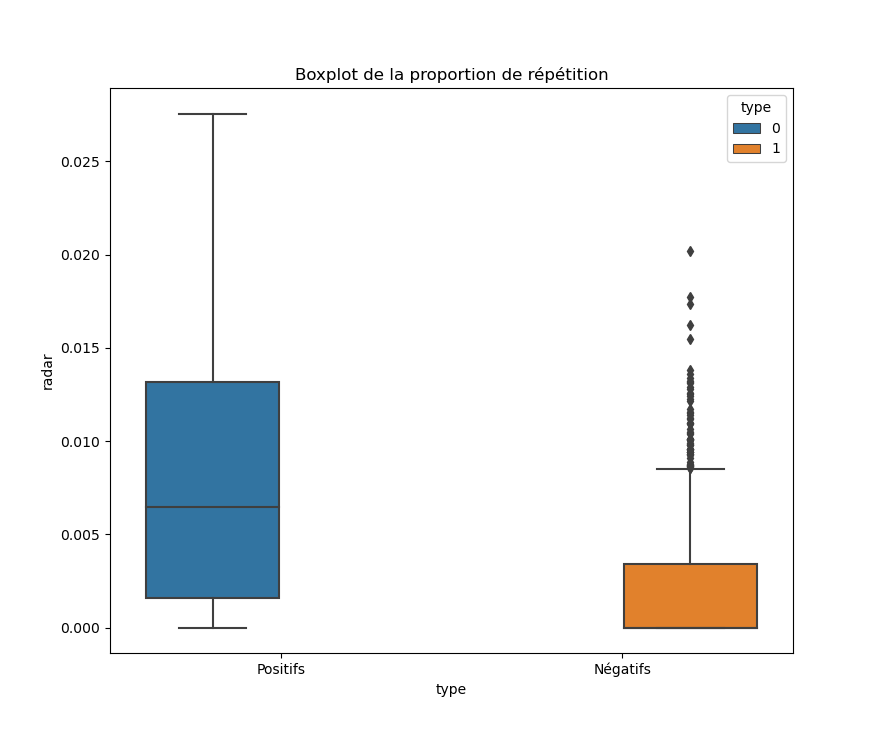
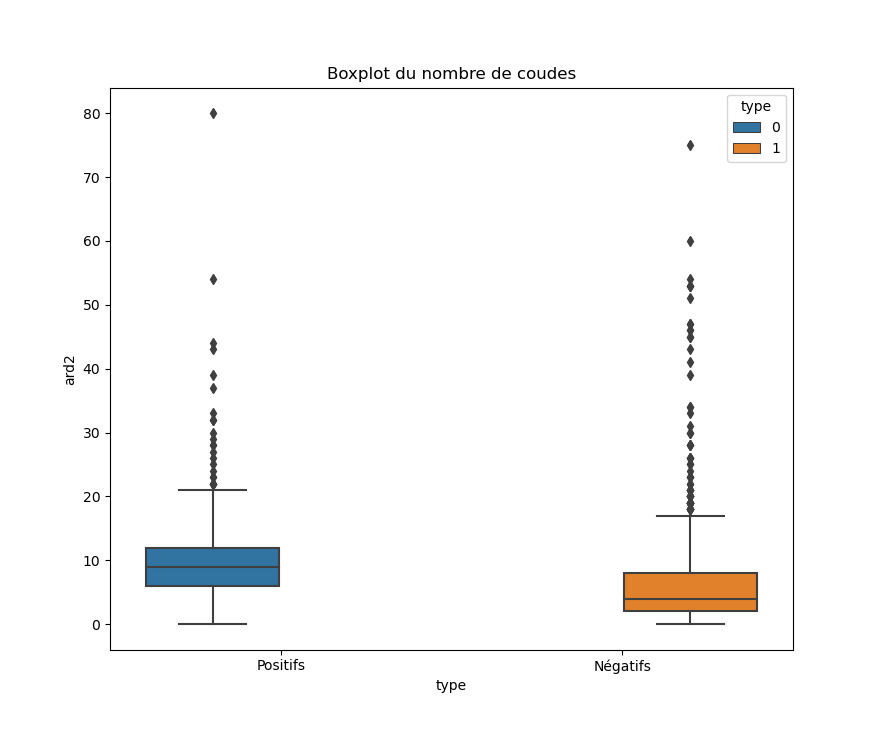
Le traitement des sorties des logiciels a été effectué comme suivant :

|  |  |
| --- | --- |
| **Outil** | **Propriété** |
| TMHMM | Identifiants des protéines sans domaine transmembranaire (après le 68e acide aminé) |
| Radar | Nombre de répétitions et proportion de la séquence qui possède des répétitions |
| Ard2 | Nombre de coudes, probabilités d’être coude |
| Targetp2 | Adressage[[9]](#footnote-9) |
| Deeploc | Probabilité d’adressage pour la mitochondrie et le chloroplaste |
| Wolfpsort | Score des protéines adressées à la mitochondrie ou au chloroplaste |
| Localizer | Résultat des protéines adressées au chloroplaste |

**Table 3 :** Tableau récapitulatif des outils utilisés et des informations recueillies des sorties de leur 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 68e 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 linker. 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 3). Pour Wolfpsort nous avons choisi de récupérer les scores associés aux adressages aux 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 normalisé 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 ces résultats que nous allons utiliser pour l’apprentissage et la prédiction de notre modèle. Chaque protéine sera ainsi décrite par X valeurs pour la construction du modèle RF.

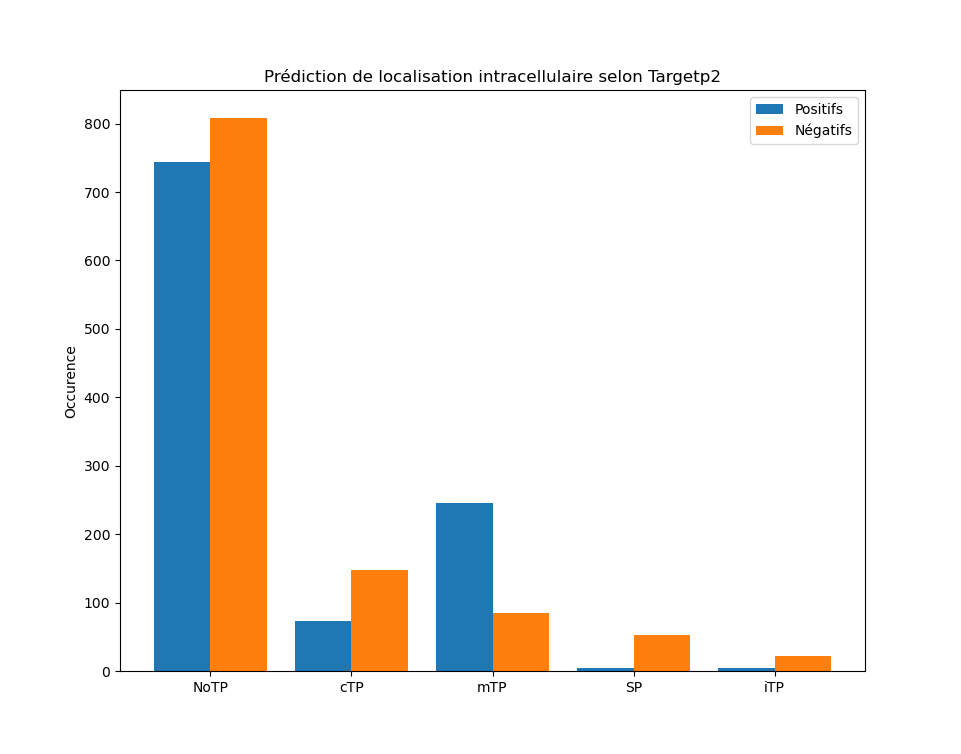
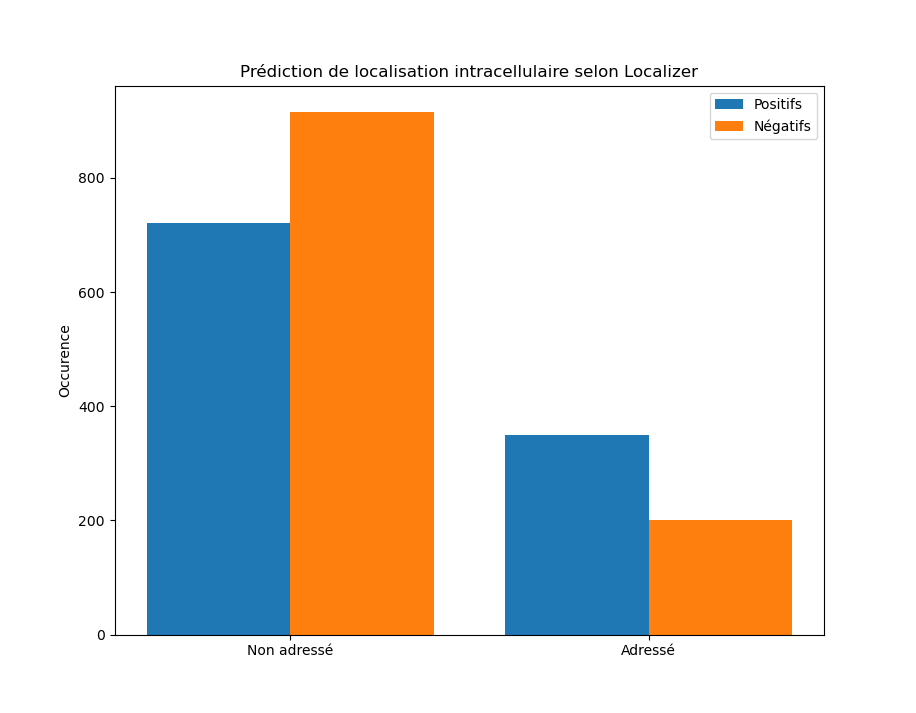
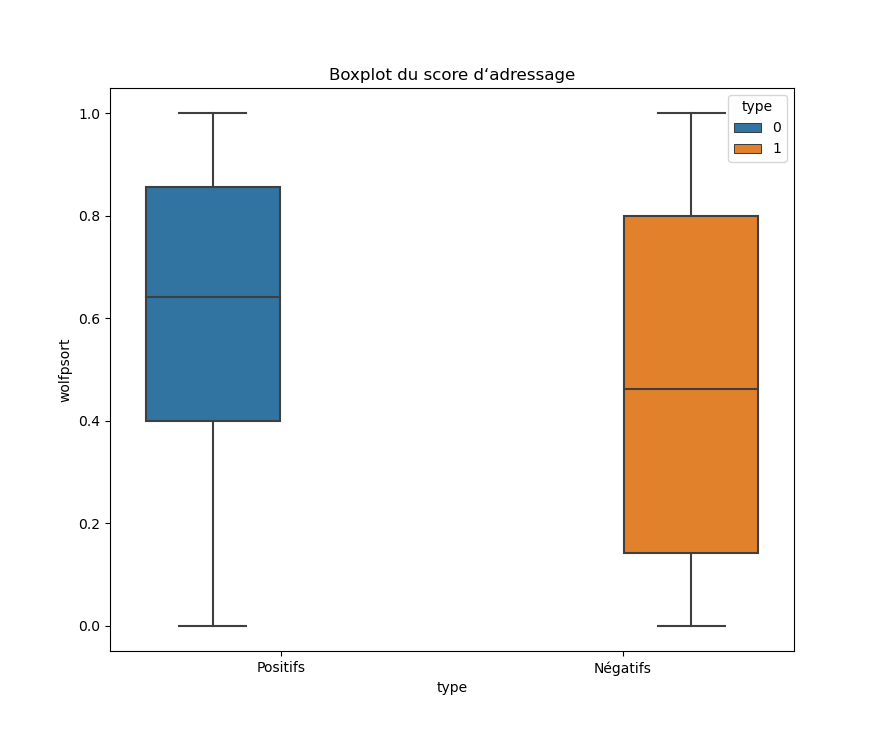
La distribution des propriétés ainsi définies est présentée figure 6. Les tests de comparaison de moyennes entre les deux jeux sont tous significatifs (p-value < 0.05) : les propriétés des protéines positives et négatives sont bien différentes. Ainsi les protéines du témoin positif comportent plus de 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.



A

C

B



E

F

D

**Figure 6 :** Boxplots des résultats des logiciels selon les jeux positifs (bleu) et négatifs (orange). (A) ard2 (nombre de coude). (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 :

Jeu de données

10%

90%

20%

*Constituée des témoins positifs et négatifs*

Validation

Apprentissage

70%

Test

**Figure 5 :** 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 :

|  |  |  |
| --- | --- | --- |
|  | **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 paramètres autres paramètres non précisés sont laissés en défaut.

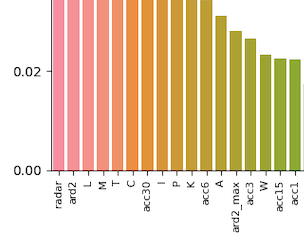
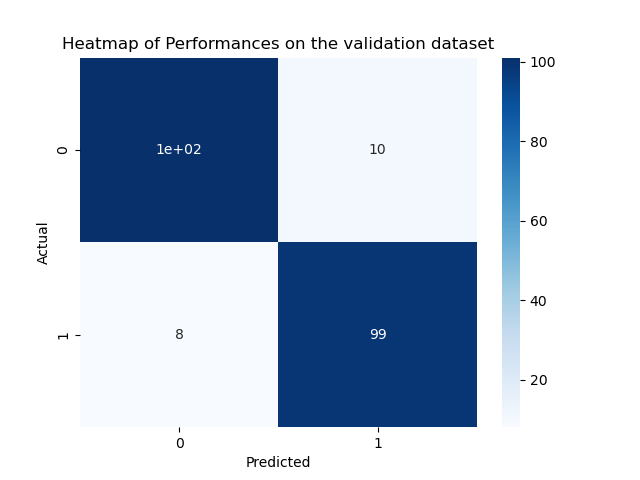
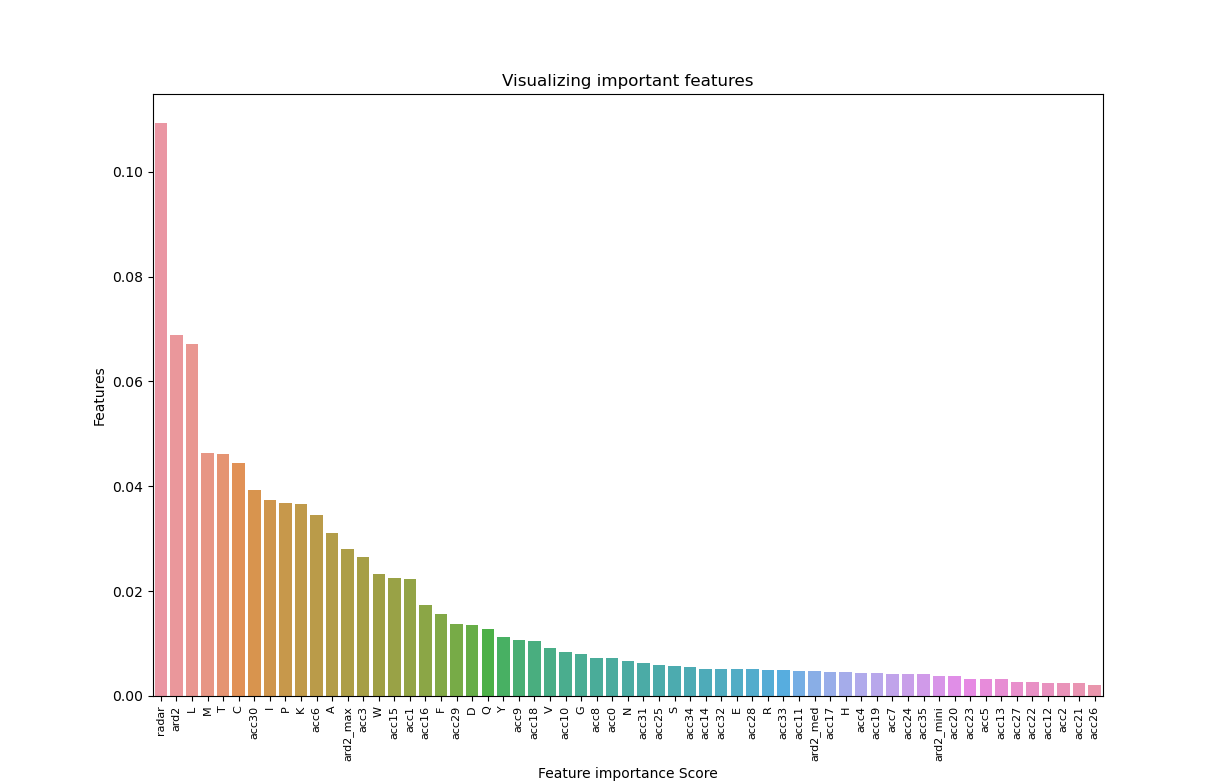
\*\* : Les paramètres utilisés sont ceux obtenus parmi une gamme paramétrique (de 100 à 800 pour n\_estimator et de 5 à 50 pour min\_sample\_split) permettant les meilleures performances possibles durant l’apprentissage du modèle.

Ensuite nous avons calculé les performances de notre modèle sur nos témoins, visible 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.

*Features importance score*

**Faux négatifs**

**Vrais positifs**

*Features*

B

A

**Vrais négatifs**

**Faux positifs**

**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 fiabilité de précision de 93% (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[[10]](#footnote-10), ce sont les résidus externes à ce cœur hydrophile qui sont reconnus par le modèle. Cela signifierait que ce sont les résidus qui aident au maintien de la structure en α-solénoïde (figure 8) qui permettent une bonne prédiction et donc que ces séquences sont conservées contrairement à celles du cœur de la protéine permettant la fixation à l’ARN.

A B

*Interactions inter-hélices*

*Cœur hydrophile*

**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. Gris : acides aminés apolaires (score > 0.02). Rouge : acides aminés polaires (score > 0.02). Orange : 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. Une hypothèse probable est qu’il n’y a qu’une seule face des hélices qui permet la prédiction du modèle : c’est la face apolaire.

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 une seule face des hélices : celle qui est apolaire. Cela démontre, aussi à l’aide de la figure 8, le caractère amphipatique des hélices de chacun des motifs des protéines en α-solénoïdes qui sont donc caractérisées par une face hydrophile (permettant la liaison à l’ARN) et d’une face hydrophobe (maintenant sa structure caractéristique) [Clothilde]. 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 protéine-ARN. 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.

Les ACC dont l’importance est supérieure à 0.2 (figure 7B) sont présentés dans le tableau 7 (la correspondance entre numéros d’ACC, lag et Z-scales est donnée en annexe 6).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Acc30** | **Acc15** | **Acc6** | **Acc3** | **Acc1** |
| **Facteurs** | Z1-Z2 | Z2-Z1 | Z1-Z1 | Z1-Z1 | Z2-Z2 | |
| **Lag** | 4 | 1 | 3 | 2 | 1 |
| **Face** | A | B | A | B | B |

**Table 7 :** 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 (Z1) est impliqué dans l’importance de l’ACC comme descripteur. Ces observations apparaissent au moins un cas sur deux chacune et signifieraient qu’elles permettent de différencier plus facilement les protéines en α-solénoïdes des autres. Le Z1 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.

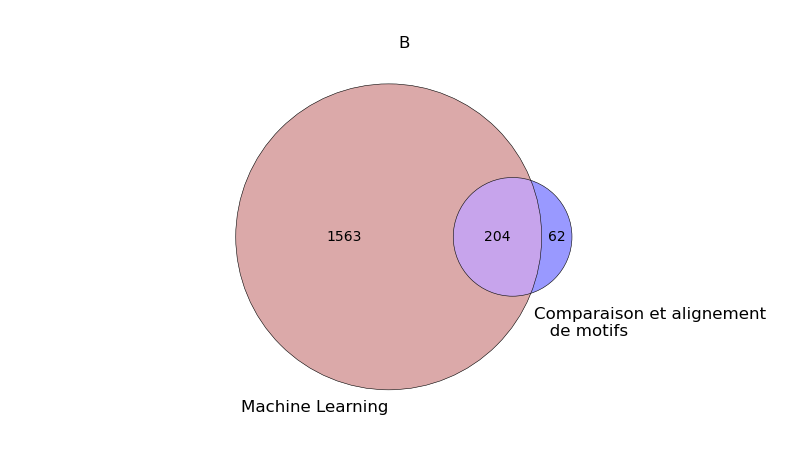
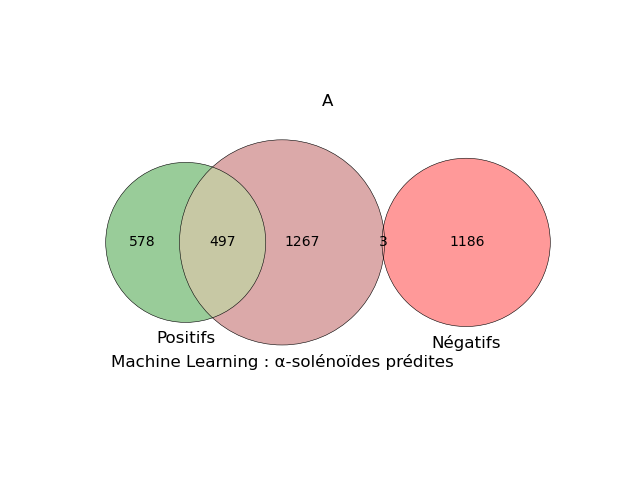
Il semblerait que la récurrence de certains facteurs de Z-scales et de lag précis reflèteraient le caractère amphiphile des hélices d’une protéine [Clotilde]. Il existe des combinaisons d’ACC qui mettent en évidence des patterns et propriétés des résidus. Étant donné qu’un tour d’hélice est constitué 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 [Clotilde]. En poursuivant notre hypothèse que le modèle reconnaît principalement la face apolaire d’une l’hélice, cela pourrait correspondre par exemple à la face B (table 7). Dans l’article [Clothide, supplementary data] il a été montré que les propriétés hydrophobes et stériques correspondant aux ACC : Z1-Z1, Z2-Z2, Z1-Z2 et Z2-Z1 permettent de distinguer les propriétés amphipatiques d’une hélice α. Nous retrouvons les mêmes résultats dans la table 7. Comme dit plus haut, ces termes d’ACC reflètent l’hydrophobicité et l’encombrement stérique sur une même face d’hélice. Cela nous permet de valider notre hypothèse que les acides aminés apolaires d’une même face, permettant le maintient de la structure de la protéine et ayant un fort taux de conservation au sein des séquences sont d’une grande utilité quant à la prédiction des ROGEs. Aussi, les résultats sur l’importance des ACC reflète l’encombrement stérique, la polarité des résidus et le caractère amphiphile des hélices chez les protéines ROGEs.

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

Ensuite nous avons utilisé notre modèle sur *Chalmydomonas 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 deux fois par les logiciels de prédiction d’adressage intracellulaire. Les résultats sont résumés dans la table ci-dessous (table 9). Enfin pour avoir une meilleure compréhension de nos résultats nous les avons comparé 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 |
| Après filtrage sur les résultats RF | 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 9 :** Résultats du modèle de prédiction sur les organismes modèles *A. thaliana* et *C. reinhardtii*. En mettant en commun les résultats 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 8 :** Comparaison des résultats avec (A) 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) les résultats obtenus avec l’arbre décisionnel. En rose : résultats par machine learning. En vert : témoin positif. En rouge témoins négatif. En bleu : résultats par comparaison et alignements de motifs.

Ces résultats ont montré que parmi les 1767 protéines candidates ROGEs déterminées par RF il y avait déjà 497 protéines connues en α-solénoïde et 3 dont on sait qu’elles ne le sont pas (figure 8A). Cela correspondrait aux performances de notre modèle (93% 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 la majorité des protéines ROGEs identifiés par l’arbre décisionnel (figure 8B) ainsi que beaucoup de OPR et PPR connues chez les deux organismes (annexe 9).

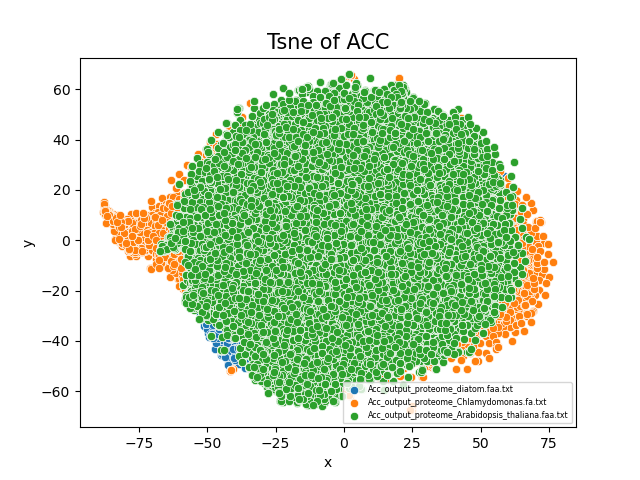
Afin de déterminer s’il existait des familles de paralogues dans chacune des deux espèces parmi les protéines prédites nous avons regroupé les protéines sur la base de leur similarité de séquence. Nous avons lancé un blast sur les protéomes contre eux même. Chacune des protéines sera comparée avec toutes les protéines du protéome. Ensuite nous remplaçons la valeur de la Evalue par le –log10(Evalue) pour visualiser le réseau de similarité avec Cytoscape[[11]](#footnote-11). Les images des clusters sont disponibles en annexe (annexe 10). La figure montre que qu’il y a des clusters d’OPR et/ou de PPR parmi les familles protéiques et donc que les protéines faisant partie de ce cluster (non OPR ou PPR) sont de potentielles ROGEs en OPR ou PPR.

Enfin pour avoir une annotation de nos protéines candidates ROGEs (en bleu table 9) nous utilisons eggNOG-mapper. Nous récupérons donc pour chaque candidate ROGE, une protéine orthologue[[12]](#footnote-12), 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 « recherchés » en lien avec la fixation à l’ARN[[13]](#footnote-13). Les résultats sont visibles en rouge table 9.

**Enfin sur les protéines non annotées nous avons décidé de lancer un Alpha-fold dont un exemple de résultat est visible en annexe 11.**

## III. 3) Résultats sur la diatomée *Phaeodactylum tricornutium*

**Nous avons ensuite appliqué le modèle pour explorer le protéome de la diatomée *Phaeodactylum tricornutium*. D’après la figure 9 ci-dessous on remarque qu’il n’y a pas de clustering entre les différents organismes sur la base des ACC et des fréquences en acides aminés pour chacun des protéomes. Il n’y a donc pas de différences fondamentales entre les diatomées, *C.reinhardtii* et *A.thaliana* en terme d’ACC et de fréquences d’acides aminés. Notre modèle est donc adaptable à *P. tricornutium*. Les logiciels d’adressage utilisés plus haut ne sont pas adaptés aux diatomées, à l’exception de Deeploc. En effet les séquences d’adressage au chloroplaste des diatomées sont différentes de celles des Archaeplastida car ils ne sont pas originaires du même évènement d’endosymbiose. Des logiciels de prédictions d’adressages adaptés à cet organisme seront utilisés en tant que filtre final pour ne garder que les protéines adressées aux chloroplastes des diatomées. Nous utiliserons donc Hectar (outil de prédiction de localisation pour diatomées) et Deeploc comme filtre final sur les protéines déterminées comme ROGEs par le modèle.**

**

B

A

**Figure 9 :** (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 *Phaeodactylum tricornutium* (8396 lignes et 61 colonnes) dont les résultats sont disponibles ci-dessous. On trouve alors 26 protéines ROGEs sur 10681 chez *Phaeodactylum tricornutium.***

|  |  |
| --- | --- |
|  | **Prédictions** |
| Taille du protéome | 10681 |
| α-solénoïdes totales | 107 |
| Après filtrage (protéines finales) | 26 |
| Dont annotées | 13 |
| Dont non annotées | 13 |

**Table 11 :** Résultats obtenus du modèle sur *Phaeodactylum tricornutium*. On trouve 26 protéines ROGEs, soit ~0.24% du protéome.

Ces résultats sont en accord avec ceux 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 (annexe 8.2). Il trouvait 3 OPR et 55 PPR dans le protéome de *P. tricornutium*, 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 suivre 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 [jaled] qui n’est pas reconnue par le modèle. Il existe aussi une protéine intéressante, la « TCA1 », qui est connue pour ne pas être une α-solénoïde mais qui possède la même structure et les mêmes rôles au sein des cellules chloroplastiques chez *Chlamydomonas reinhardtii* que les ROGEs. Celle-ci est reconnue par notre modèle et retenue après filtrage. Aussi les méthodes de filtrage comme Deeploc ne sont pas toujours fiables, et de la même manière peut se tromper et nous amène 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 ainsi de récupérer par une méthode ou l’autre les protéines ROGEs éliminées par un des modèles ou au cours des filtres.

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[[14]](#footnote-14). Nous avons fait tourner ce modèle un certain nombre de fois sur *C. reinhardtii* et *A. thaliana* et regardé la moyenne des performances sur le jeu de validation mais nous avons aussi vérifié quelles étaient les protéines qui étaient souvent prédites comme ROGEs par les différents lancements du modèle. Nous avons ainsi attribué à chacune des protéines un score selon le nombre de fois où elles étaient prédites comme ROGEs afin d’éviter les biais de prédiction du modèle. Parmi ces protéines nous avons noté une certaine redondance de TCA1. Quant aux clustering de cytoscape, TCA1 ne faisait parti d’aucun cluster. L’étude des clusters obtenus a montré une similarité entre les OPR et PPR des différents organismes ainsi que d’autres protéines non connues comme ROGEs. Elles seraient de nouvelles candidates ROGEs probables.

De plus nous avons pu conclure avec les prédictions d’Alpha fold que nous avons fait que radar ne distingue pas les répétitions d’helices α et de brins. Ainsi il sera toujours nécessaire de le combiner avec ard2 pour être sûre de la présence de répétitions d’hélices au sein de la protéine. Il est donc dans notre intérêt de garder la combinaison de ces deux logiciels dans notre modèle, sans parler de l’importance que ceux-ci occupent dans les prédictions. Toutefois nous avons remarqué avec les prédictions d’Alpha fold que des protéines finales qui ont été prédites comme ROGEs et passées le filtre d’adressage 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..

Le fait que le modèle retrouve encore moins de ROGE dans *Phaeodactylum tricornutium* 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 tricornutium* 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 partis des Archaeplatida et que leur chloroplaste leur provient d’une endosymbiose primaire. Tandis que *P. tricornutium* a acquis son chloroplaste d’une endosymbiose secondaire. Évolutivement parlant il est donc probable qu’il y ai déjà d’autres protéines que celles que nous recherchons impliquées dans la régulations post-transcriptionnelle du chloroplaste chez les diatomées.

# *ANNEXES*

|  |  |
| --- | --- |
| **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 1.1 :** Modules utilisés pour les analyses bio-informatiques.

*Lien Github :* [*https://github.com/Rebbekkah/Stage\_M2.git*](https://github.com/Rebbekkah/Stage_M2.git)

**Annexe 1.2 :** Lien vers le Github du stage contenant les scripts.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Z1** | **Z2** | **Z3** |
| Caractéristique | Hydrophobicité  (caractère apolaire) | Encombrement stérique | Propriétés électroniques |

**Annexe 2.1 :** Z-scores et les propriétés qu’elles caractérisent.

**Annexe 2.2 :** 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.



*Facteurs j et k, position i*

*Même facteur j, position i*

**Annexe 2.3 :** Formule du calcul des ACC.

Peptide d’adressage

68

3’

*Absence de domaine TM*

5’

1. Détection domaines TM

Protocole

*Présence de motifs répétés*

3’

5’

68

2. Détection motifs répétés

α

α

α

α

3’

5’

*Présence de paires d’hélice*

L

L

68

4. Prédiction localisation intracellulaire

3. Détection paires d’hélices

**Organites** (mitochondrie, chloroplaste)

**Annexe 4 :** Protocole mis en place pour identifier les protéines en α-solénoïdes. Si une protéine passe tout ces filtre 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.

Répétition 1

C-terminal

N-terminal

***Cas 1***

Répétition 2

N-terminal

C-terminal

Répétitions prise en compte

Répétition 1

C-terminal

N-terminal

Répétition 2

***Cas 2***

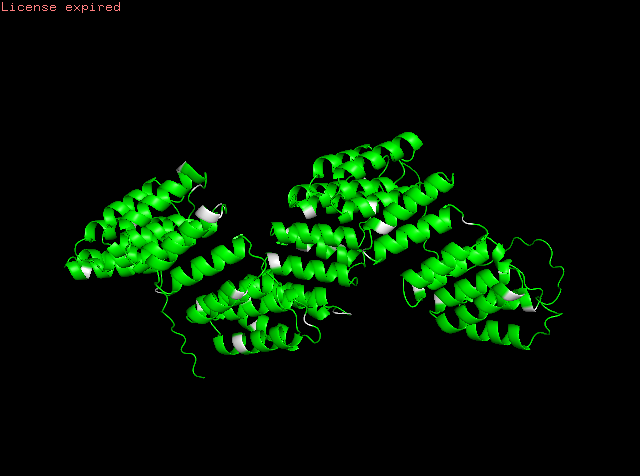
C-terminal

N-terminal

Répétitions prise en compte

(la plus grande des deux)

**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, situés 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[[15]](#footnote-15) | 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)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Acc0** | **Acc1** | **…** | **Acc35** | **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 | … | ACC35 | 0.887958461538462 | FQ |
| 0.571304930095659 | 1.8240744665195 | … | ACC35 | 0.0811762799263352 | 0.179464198895028 |

*(Suite de la dataframe)*

|  |  |  |
| --- | --- | --- |
| **A** | **…** | **Acide aminé X** |
| Fréquence d’apparition au sein de la séquence | Fréquence d’apparition au sein de la séquence | Fréquence d’apparition au sein de la séquence |
| FA | … | Fx |
| FA | … | Fx |

**Annexe 5 :** 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)[[16]](#footnote-16) | 107 | 244 | 351 |
| Nouvelles protéines déterminées (RF + Céline Cattelin) | 505 | 943 | 1448 |

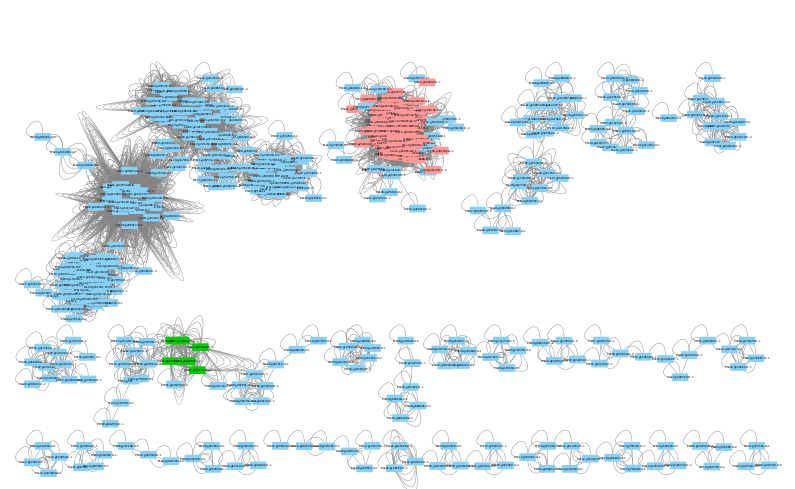
**Annexe 8.1 :** Tableau complet des comparaisons de méthodes.

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.

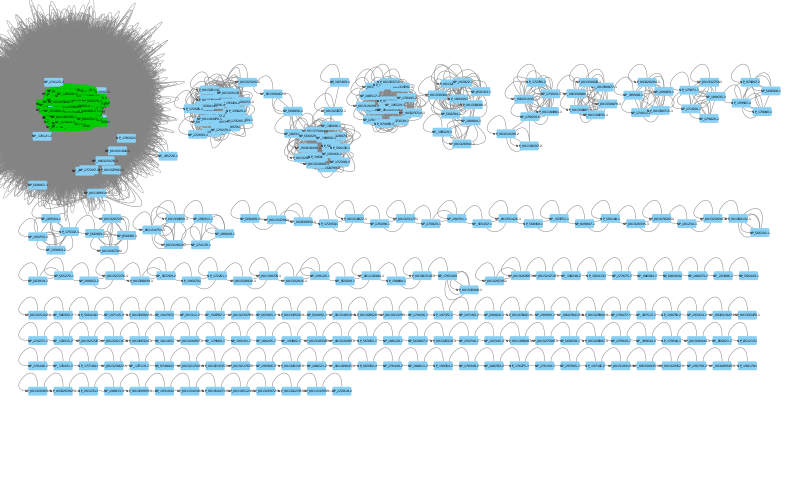
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Modèle** | | **Connues** | |
| *C. reinhardtii* | *A. thaliana* | *C. reinhardtii* | *A. thaliana* |
| **OPR** | 90 | 0 | 106 | 1 |
| **PPR** | 6 | 476 | 8 | 686 |
| **Total** | 96 | 476 | 114 | 687 |

**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.



B

A



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

# *RÉFÉRENCES*

[] Guo & al. (2022) « A computational method for predictiong nucleocapsid protein in retroviruses », *Scientific Reports, Volume 12, page 524, 10.1038/s41598-021-03182-2.*

[] Sandaruwan & al. (2021) « An improved deep learning model for hierarchical classification of protein families », *PloS ONE, Volume 16, Numéro 10, 10.1371/journal.pone.0258625*.

[] Takenaka & al. (2021) « DYW domain structures imply an unusual regulation principle in plant organellar RNA editing catalysis », *Nature Catalysis, Volume 4, Numéro 6, pages 510-522, 10.1038/s41929-021-00633-x.*

[] Zhao & al. (2021) « DescribePROT : database of amino acid-level protein structure and function predictions », *Nucleic Ancids Reasearch, Volume 49, Numéro D1, pages D298-D308, 10.1093/nar/gkaa931.*

[] Chatzimparm & al. (2020) « t-viSNE : Interactive Assessment and Interpretation of t-SNE Projections », *IEE Trans. Visual. Comput. Graphics, Volume 26, Numéro 8, pages 2696-2714, 10.1109/TVCG.2020.2986996.*

[] Garrido & al. (2020) « Evidence Supporting an Animicrobial Origin of Targeting Peptides to Endosymbiotic Organelles », *Cells, Volume 9, Numéro 8, pages 1795, 10.3390/cells9081795.*

[] Falciatore & al. (2020) « Diatom Molecular Research Comes of Age : Model Species for Studying Phytoplaktion Biology and Diversity », *The Plant Cell, Volume 32, Numéro 3, pages 547-572, 10.1105/tpc.19.00158.*

[] Gutmann & al. (2020) « The Expansion and Diversification of Pentratricopeptide Repeat RNA-Editing Factors in Plants », *Molecular Plant, Volume 13, Numéro 2, pages 215-230, 10.1016/j.molp.2019.11.002.*

[] Burki & al. (2020) «  The new Tree of Eukaryotes », *Trends in Ecology and Evolution, Volume 35, Numéro 1, pages 43-55, 10.1016/j.tree.2019.08.008.*

[] Nanni & Brahman (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, Volume 2, Numéro 1, pages 38-52, 10.33969/AIS.2020.21004.*

[] Ponce-Toledo, Lopez-Garcia & Moreira (2019) « Horizontal and endosymbiotic gene transfer in early plastid evolution », *New Phytologist, Volume 224, Numéro 2, pages 618-624, 10.1111/nph.15965.*

[] Lv & al. (2019) «  A random Forest Sub-Golgi Protein Classifier Optimized via Dipeptide and Amino Acid Composition Features », *Frontiers in Bioengineering and Biotechnology, Volume 7, pages 215, 10.3389/fbioe.2019.00215*.

[] Graham & al (2019) « A Random Forest Sub-Golgi Protein Classifier Optimized via Dipeptide and Amino Acid Composition Features », *Critical Reviews in Biochemistry and Molecular Biology, Volume 54, Numéro 2, pages 103-118, 10.1080/10409238.2019.1590305.*

[] Luttrel & al. (2019) « Predicting protein residue-residue contacts using random forests and deep networks », *BMC Bioinformatics, Volume 20, Numéro S2, pages 100, 10.1186/s12859-019-2627-6.*

[] Hakala & al. (2019) « Neural network and random forest models in protein function prediction », *bioRxiv, pages 18.*

[] Hillebrand & 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 », *Nucleic Acids Research, 10.1093/nar/gky710.*

[] Seo & al. (2018) « DeepFam: deep learning based alignment-free method for protein family modeling and prediction », *Bioinformatics , Volume 34, Numéro 13, pages i254-i262, 10.1093/bioinformatics/bty275.*

[] Taherzadeh & al. (2018) « Structure-based prediction of protein– peptide binding regions using Random Forest », *Bioinformatics, Volume 34, Numéro 3, pages 477-484, 10.1093/bioinformatics/btx614.*

[] Kathuria & al. (2018) « Predicting the protein structure using random forest approach », *Procedia Computer Science, Volume 132, pages 1654-1662, 10.1016/j.procs.2018.05.134.*

[] Schietgat & al. (2018) « A machine learning based framework to identify and classify long terminal repeat retrotransposons », *PLOS Computational Biology, Volume 14, Numéro 4, pages e1006097, 10.1371/journal.pcbi.1006097.*

[] Sarica & al. (2017) « Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review », *Frontiers in Aging Neuroscience, Volume 9, pages 329, 10.3389/fnagi.2017.00329.*

[] Zielinski & al. (2017) « Deep learning approach to bacterial colony classification », *PLOS ONE, Volume 12, Numéro 9, 10.1371/journal.pone.0184554.*

[] Fučíková & al. (2016) « Chloroplast phylogenomic data from the green algal order Sphaeropleales (Chlorophyceae, Chlorophyta) reveal complex patterns of sequence evolution », *Molecular Phylogenetics and Evolution, Volume 98, pages 176-183, 10.1016/j.ympev.2016.01.022.*

[] Turesson & al. (2016) « Machine Learning Algorithms for Automatic Classification of Marmoset Vocalizations », *PLOS ONE, Volume 11, Numéro 9, 10.1371/journal.pone.0163041.*

[] Marx, Wunsch & Kück (2015) « The Octatricopeptide Repeat Protein Raa8 Is Required for Chloroplast *trans* Splicing », *Eukaryotic Cell, Volume 14, Numéro 10, pages 998-1005, 10.1128/EC.00096-15.*

[] Libbrecht & Noble (2015) « Machine learning applications in genetics and genomics », *Nature Reviews Genetics, Volume 16, Numéro 6, pages 321-332, 10.1038/nrg3920.*

[] Boulouis & al. (2015) « Spontaneous Dominant Mutations in Chlamydomonas Highlight Ongoing Evolution by Gene Diversification », *The Plant Cell, Volume 27, Numéro 4, pages 984-1001, 10.1105/tpc.15.00010.*

[] Mohan & al. (2014) « Automatic classification of protein structures using physicochemical parameters », *Interdisciplinary Sciences: Computational Life Sciences, Volume 6, Numro 3, pages 176-186, 10.1007/s12539-013-0199-0.*

[] Barkan & Small (2014) « Pentatricopeptide Repeat Proteins in Plants », *Annual Review of Plant Biology, Volume 65, Numéro 1, pages 415-442, 10.1146/annurev-arplant-050213-040159.*

[] Ge & al. (2014) « Import Determinants of Organelle-Specific and Dual Targeting Peptides of Mitochondria and Chloroplasts in Arabidopsis thaliana », *Molecular Plant, Volume 7, Numéro 1, 10.1093/mp/sst148.*

[] Zhao & al. (2014) « Determining Effects of Non-synonymous SNPs on Protein-Protein Interactions using Supervised and Semi-supervised Learning », *PLoS Computational Biology, Volume 10, Numéro 5, pages e1003592, 10.1371/journal.pcbi.1003592.*

[] Jahandideh & al. (2014) « Improving the chances of successful protein structure determination with a random forest classifier », *Acta Crystallographica Section D Biological Crystallography, Volume 70, Numéro 3, pages 627-635, 10.1107/S1399004713032070.*

[] Fournier & al. (2013) « Functional and Genomic Analyses of Alpha-Solenoid Proteins », *PLOS ONE, Volume 8, Numéro 11, 10.1371/journal.pone.0079894.*

[] Tourasse & al. (2013) «  PPR proteins of green algae », *RNA Biology, Volume 10, Numéro 9, pages 1526-1542, 10.4161/rna.26127.*

[] He & al. (2012) « Modeling the QSAR of ACE-Inhibitory Peptides with ANN and Its Applied Illustration », *International Journal of Peptides, Volume 2012, pages 1-9, 10.1155/2012/620609.*

[] Leliaert & al. (2012) « Phylogeny and Molecular Evolution of the Green Algae », *Critical Reviews in Plant Sciences, Volume 31, Numéro 1, pages 1-46, 10.1080/07352689.2011.615705.*

[] Fukui & Kuramitsu (2011) « Structure and Function of the Small MutS-Related Domain », *Molecular Biology International, Volume 2011, pages 1-9, 10.4061/2011/691735.*

[] Lin & Chen (2011) « Prediction of thermophilic proteins using feature selection technique », *Journal of Microbiological Methods, Volume 84, Numéro 1, pages 67-70, 10.1016/j.mimet.2010.10.013.*

[] Jain & Hirst (2010) « Automatic structure classification of small proteins using random forest », *BMC Bioinformatics, Volume 11, Numéro 1, pages 364, 10.1186/1471-2105-11-364.*

[] Madera & al. (2010) « Improving protein secondary structure prediction using a simple k-mer model », *Bioinformatics, Volume 26, Numéro 5, pages 596-602, 10.1093/bioinformatics/btq020.*

[] Wold & al. (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, Volume 277, Numéro 2, pages 239-253, 10.1016/0003-2670(93)80437-P.*

[] Hellberg & al (1987) « Peptide quantitative structure-activity relationships, a multivariate approach », *Journal of Medicinal Chemistry, Volume 30, Numéro 7, 10.1021/jm00390a003.*

*RESUMÉ*

Un modèle de prédiction protéique par machine learning a été conçu pour explorer, déterminer et annoter des protéines en α-solénoïdes adressées aux organites chez les eucaryotes photosynthétiques.  Ces protéines impliquées dans la régulation post-transcriptionelle du chloroplaste ont été déterminées sur la base d’un arbre décisionnel par l’intermédiaire des caractéristiques structurelles des protéines. Le modèle possède de bonnes performances et les résultats ont montré des conservations de séquences permettant le maintien de la structure tridimensionnelle de ces protéines ainsi que des schémas récurrents quand à leurs propriétés physico-chimiques. Celui-ci a pu retrouver des protéines connues comme α-solénoïdes mais aussi de nouvelles protéines encore inconnues à ce jour chez les organismes étudiés. Le modèle peut être adapté et applicable à tout type d’organisme possédant les protéines étudiées. Le principe d’arbres décisionnels permet une large gamme de choix de descripteurs ainsi qu’une robustesse au sur-apprentissage.

*RESUME*

A predictive model for proteins has been developed by machine learning in order to explore, determine and annotate α-solenoid proteins that are addressed to cells organelles within the photosynthetic eukaryotes. These proteins are involved in the post-transcriptional regulation of the chloroplast and have been determined with the approach of decisional trees through their structural characteristics. The model has good performances and the results have shown that there exists sequences in these α-solenoid that are conserved and serves to the tri-dimensional structure maintenance. Also, it appeared that there are typical recurrent schemes about their physic-chemical properties. It was able to find well-known α-solenoid proteins but also undiscovered ones within the studied organisms. The principle of decisional trees allows a wide spectrum of choices for the descriptors and a robustness concerning the over-fitting.

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 [↑](#footnote-ref-1)
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] [↑](#footnote-ref-2)
3. Ainsi pour un même facteur (z1 – z1, z2 – z2 et z3 – z3) on obtient 3\*r = 3\*4 = 12 valeurs d’ACC. Tandis que pour des facteurs différents (z1 – z2, z1 – z3, z2 – z3, z3 – z1, …) on a 6\*r = 6\*4 = 24 ACC soit 12+24 = 36 termes d’ACC. [↑](#footnote-ref-3)
4. Noyau, Cytoplasme, extracellulaire, mitochondriale, membranaire, reticulum endoplasmique, chloroplaste, appareil de Golgi, Lysosome/vacuole et peroxysome. [↑](#footnote-ref-4)
5. Méthode de classification par apprentissage supervisé. [↑](#footnote-ref-5)
6. Le seuil dépend de la taille du protéome. Pour un protéome de taille N on prends un seuil de Evalue de 10-N. [↑](#footnote-ref-6)
7. 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. [↑](#footnote-ref-7)
8. Ces protéomes sont constitués de protéines issues d’espèces de Chloroplastida suivantes : *Chlamydomonas reinhardtii, Arabidopsis thaliana, Chlorella variabilis, Volvox carteri f. nagarientis, Tetradesmus obliquus, Chlamydomonas eustigma, Gonium pectorale, Monaraphidium neglectum, Tetrabaenca socialis, Rophidocelis subcapitata, Dunalella salina*. [↑](#footnote-ref-8)
9. Sorties possibles : SP = signal peptide, mTP = mitochondrial transit peptide, cTP = chloroplast transit peptide, iTP = thylakoid transit peptide, NoTP = aucun signal [↑](#footnote-ref-9)
10. L’ARN étant chargé négativement l’interaction avec la protéine reste logique. [↑](#footnote-ref-10)
11. Cela permet d’avoir des poids entre protéine pour permettre à Cytoscape de faire son clustering. [↑](#footnote-ref-11)
12. Deux protéines sont orthologues si elles sont issues d’un ancêtre commun. Cela signifie qu’elles sont similaires. [↑](#footnote-ref-12)
13. 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). [↑](#footnote-ref-13)
14. Pour *P. tricornutium* nous avons utilisé Hectar. [↑](#footnote-ref-14)
15. Pour des répétitions de longueur comprises entre 29 et 46. En effet ce sont des types de répétitions caractéristiques des OPR, TPR et PPR. [↑](#footnote-ref-15)
16. 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. [↑](#footnote-ref-16)