

**Rapport de Stage M2**

*Goulancourt Rebecca*

*M2 Biologie-Informatique*

*Université de Paris*

*2021-2022*

*Github : https://github.com/Rebbekkah/Stage\_M2.git*

*Tutrice : Ingrid Lafontaine, IBPC.*

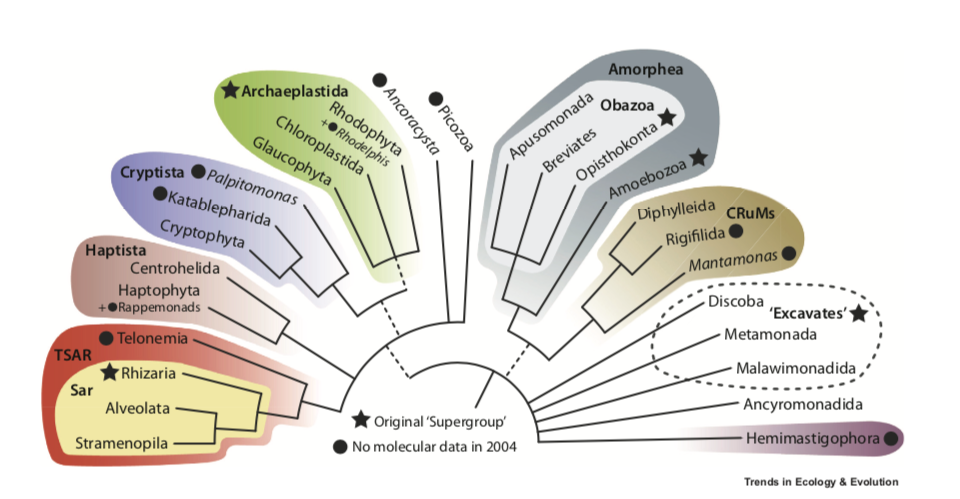
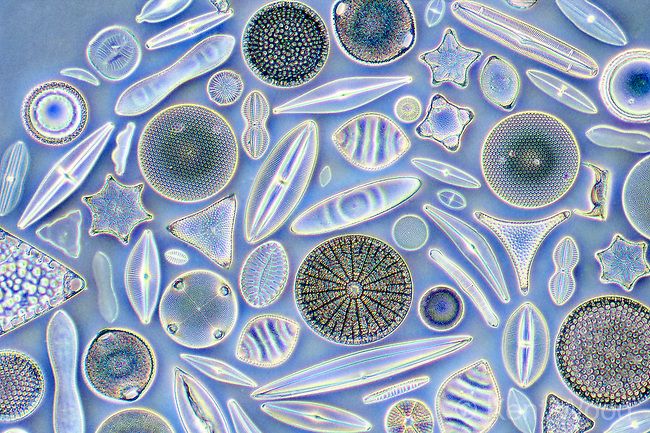
*Mots clés : Chloroplaste, protéine, adressage, régulation post-transcriptionelle, alpha-solénoïde, prédiction, modèle de prédiction « Random Forest ».*

*SOMMAIRE*

1. Introduction

Les diatomées sont des organismes marins, microalgues eucaryotes photosynthétiques de diverses formes c’est-à-dire possédant différents axes de symétrie. Leur taille peut varier de 2 µm à 1mm et en tant qu’espèce photosynthétique ils sont dotés de mitochondrie et surtout de chloroplastes. Ce chloroplaste leur vient d’une endosymbiose secondaire. L’endosymbiose secondaire est caractérisée par l’internalisation d’une cellule issue d’une endosymbiose primaire. Les cellules issues de l’endosymbiose primaire, comme les algues vertes, proviennent de la phagytose d’une cyanobactérie par une cellule ancestrale primaire hétérotrophe. La cyanobactérie se différencie ensuite en chloroplaste et la cellule acquiert une capacité photosynthétique, devenant ainsi autotrophe. Cependant les diatomées, faisant partis de la famille des *Stramenopila* (algues rouges) découlent d’une internalisation d’une cellule autotrophe possédant un chloroplaste par une autre cellule hétérotrophe, obtenant ainsi des caractéristiques photosynthétiques. +chlo et mito = info génétique

Algues vertes



Endosymbiose 1aire

Endosymbiose 2aire

Algues rouges

**Figure x :** k,eklvkve

La régulation génomique du chloroplaste est connue chez les algues vertes, mais beaucoup moins chez les algues rouges dont font partis les diatomées. Chez les algues vertes ce type de mécanisme s’effectue par des protéines possédant une structure particulière appelée alpha-solénoïde. Elles sont aussi connues pour interagir particulièrement avec l’ARN messager chloroplastique dans des processus de maturation, stabilisation, épissage, translation et dégradation génétique.

Alpha sol 🡪 adressées au chloroplaste et mitochondries

1. Matériel et méthodes

Les codes permettant de répondre aux questions et problématiques ont été rédigés essentiellement en python, puis en R.

Nous cherchons à pouvoir identifier des protéines en alpha solénoïdes sur la base de leurs propriétés structurales et physico-chimiques. Nous savons que ces protéines ont des caractéristiques communes :

* Elles sont adressées au choloroplaste ou à la mitochondrie.
* Elles ne possèdent pas de domaines transmembranaires après le 68e acides aminé (s’il y en a avant cela correspondrait à un peptide d’adressage, ce qui correspond à nos attentes).
* De part leur structure en hélice α elles possèdent des acident aminés dit « linker » situés entre deux hélices et faisant le lien; mais aussi des répétitions de séquences caractéristiques de ces hélices.

Pour déterminer les possibles candidats à la régulation post-transcriptomique nous disposons d’un protocole au préalablement conçu qui permet d’identifier nos protéines à l’aide de différentes étapes et d’outils de prédictions utilisés en stand alone. Ces outils et leurs prédictions sont visibles dans le tableau ci-dessous.[[1]](#footnote-1)

À la suite de cela nous pouvons identifier des protéines en alpha solénoïdes afin de créer un jeu d’apprentissage de témoins positifs (alpha solénoïde) et négatifs (non alpha solénoïde). Il faut noter que l’on utiliser plusieurs outils de prédiction d’adressage car il peut y avoir des biais de résultats selon les outils. Nous utilisons donc plusieurs logiciels pour avoir des résultats plus fiables en combinant les résultats. Chaque logiciel prends en entrée un protéome, et a un output formaté qu’il sera ensuite possible de parser pour récupérer les informations qui nous intéressent.

|  |  |  |
| --- | --- | --- |
| **Outil** | **Prédiction** | **Information** |
| TMHMM | Structure transmembranaire | Identifiants des protéines sans domaine transmembranaire (après le 68e acide aminé) |
| Targetp2 | Présence de pré-séquence en Nter (peptide d’adressage) | Adressage1 |
| Ard2 | Acide aminés linker d’hélice α | Nombre de linker, probabilités d’être linker |
| Radar | Répétitions de séquences | Nombre de répétitions, longueur et fréquence d’acide aminé pour chaque répétition, proportion de la séquence qui possède des répétitions |
| Deeploc | Adressage | Probabilité d’adressage pour la mitochondrie et le chloroplaste |
| Wolfpsort | Adressage | Identifiants des protéines adressées à la mitochondrie ou au chloroplaste |
| Localizer | Adressage | Identifiants des protéines adressées au chloroplaste |

Table x : nfiezongeizgneizneiz

Expliquer les logiciels et output

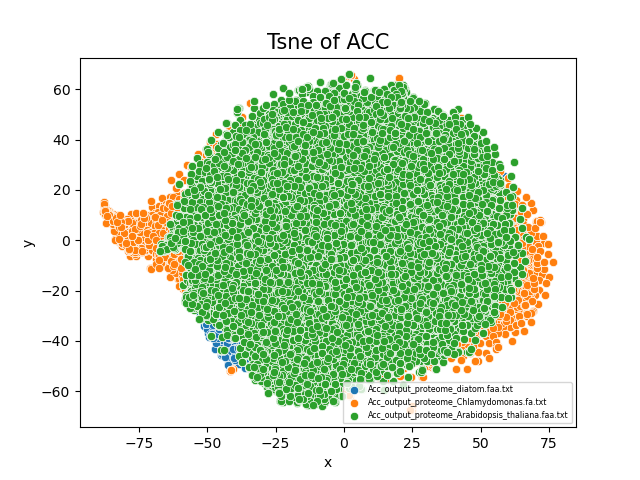
En premier lieu deux protéomes ont été fournis. Ces protéomes contiennent les séquences au format fasta de protéines que l’on sait être en alpha solénoïde (pour le protéome de témoin positif) et de protéines qui ne le sont pas (pour le protéome de témoin négatif). Nous avons donc fait tourner en premier lieu TMHMM sur ces protéomes et récupéré en effectuant un parsing des output les identifiants des protéines ne possédant pas de domaines transmembranaires après le 68e acide aminé. Ensuite nous avons reconstitué de nouveaux protéomes positifs et négatifs avec les protéines sélectionnées à l’issu du filtrage. Ces nouveaux protéomes ont ensuite été donnés comme input aux autres logiciels. Chacun des autres logiciels a donné des outputs que l’on a filtré pour ne garder que l’essentiel, visible ci-dessus dans le tableau récapitulatif.

Rappeler but de l’étude

* les outils de prédictions v
* protéomes à disposition v
* témoins positifs/ négatifs 🡪 ce que c’est, d’où ca vient, ce que j’ai fait avec, le parsing ect
* la même chose avec les protéomes de diatomées que céline m’a fournit
* Acc et freq

1. Résultats



**

🡪 on veut discriminer les protéomes 🡪 est-ce qu’ils y a des différences significatives ? 🡪 si non on peut continuer l’analyse dessus

ici l’acc semble ne pas être différent sur chacun des protéomes

***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*.

[] 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.*

1. 1. Sorties possibles : SP = signal peptide, mTP = mitochondrial transit peptide, cTP = chloroplast transit peptide, iTP = thylakoid transit peptide, NoTP = aucun signal [↑](#footnote-ref-1)