Projet de stage d’été en bio-informatique

Version -22 octobre 2021

DETECTION *in silico* DE PROTEINES du type PPR

**Bien-fondé**

Les protéines PentatricoPeptide Repeat (PPR) sont caractérisées par de longues répétitions en tandem d’un motif de 35 acides aminés, appelés PPR. Elles jouent un rôle important dans la stabilité, la régulation, et la maturation de l'ARN chez de nombreux eucaryotes. Chez l’humain, la mutation dans un membre de la famille des protéines PPR est responsable pour le syndrome de Leigh type Canadien français. Les protéines PPRs se lient à des séquences spécifiques dans l'ARN, et ceci via les interactions de deux acides aminés critiques dans chaque motif avec un nucléotide particulier dans l'ARN.

Le groupe d’eucaryotes avec la plus grande famille PPR sont apparemment les plantes chez lesquelles ces protéines ont été étudiées d’une façon intégrale. Basé sur les motifs de séquences protéiques trouvés chez les plantes, on a développé des profils probabilistes (profils modèle Markov cachés ou Hidden Markov Model (HMM)) pour trouver ces protéines chez d’autres organismes. Or, les profils HMM sont biaisés : il ne reconnaisse qu’une fraction de PPR chez l’homme ou la levure, par exemple.

Notre but est de développer une procédure qui adapte le profil HMM de façon itérative pour découvrir tous les membres de la famille PPR dans n’importe quel organisme --hors des plantes--.

Le **système de choix** pour développer l’approche est le groupe d’eucaryotes unicellulaires nommé ‘diplonemids’ car leur génome encode un nombre de protéines PPR exceptionnel, notamment ~ 70. Les protéines PPR des diplonemides diffèrent considérablement en séquence de leurs homologues végétaux.

**Objectifs**

Le projet vise à développer une méthodologie qui sert à identifier et caractériser, par des moyens bio-informatiques, les protéines PPR de n’importe quel eucaryote incluant l’humain. La première étape consiste en la classification des *motifs* PPR pour ensuite regrouper les protéines selon la composition et l’arrangement des motifs qu’elles comprennent. Les tâches à compléter lors du stage sont :

1. Le développement d’un script qui automatise les itérations servant à la construction de profils HMM;

2. Le développement d’une procédure qui surveille la performance des profils tout au long des itérations et guide leur optimisation ;

Le but à long-terme est de prédire la séquence d'ARN à laquelle les protéines PPRs peuvent se lier. Cette information est encapsulée dans l’ensemble des motifs de PPR répétés en tandem d’une protéine PPR donnée.

**Méthodologie**

On utilisera le progiciel HMMER afin de développer des profils HMM pour différentes sous-classes du motif PPR. Des logiciels tels que JPred4 ou Phyre seront utilisés pour valider au niveau de la structure secondaire et tertiaire les motifs PPR identifiés.

**Impact prévu**

Le travail du stage contribuera à l’identification fiable et la caractérisation de protéines PPR et à la découverte des gènes sous le contrôle de ces protéines. Ceci promet d’ouvrir une nouvelle perspective sur les circuits complexes de la régulation des gènes en santé et maladie.