Proposition de sujet de stage

Développement d'une stratégie pour l'alignement de lectures courtes à partir de données de sRNAseq

Nom de l'entreprise : Institut National de la Recherche Agronomique
Adresse : INRA – Unité Mathématiques et Informatique Appliquées-Toulouse (MIA-T). Chemin de Borde Rouge. BP 52627 - 31326 Castanet-Tolosan cedex
Personnes à contacter : Christine Gaspin ou Matthias Zytnicki
Téléphone : 05 61 28 52 82 / 05 61 28 50 71

Contact email : christine.gaspin@toulouse.inra.fr, matthias.zytnicki@toulouse.inra.fr

Description du sujet

Les petits ARN non codants, sont des acteurs majeurs de la régulation de l'expression des gènes. Ces ARNnc sont notamment impliqués dans la transmission de signaux environnementaux, la maturation de la cellule ou encore certaines pathologies.

Le séquençage haut-débit, et plus particulièrement le sRNA-Seq, est une technique aujourd'hui largement utilisée pour trouver ces petits ARNs. Ce séquençage peut produire des centaines de millions de séquences à un prix abordable.

La première étape d'analyse de ces séquences est l'alignement sur un génome de référence (lorsqu'il est connu), qui prédit les loci qui peuvent avoir produit les séquences. Alors qu'il existe beaucoup d'algorithmes spécifiques pour les séquences ADN ou les longs ARN, aucun n'a été conçu pour traiter spécifiquement les petits ARN.

Pourtant, le sRNA-Seq présente beaucoup de particularités, notamment des séquences de petites tailles avec un très fort taux de répétition dans les séquences obtenues (qui rendent compte de l'expression des gènes associés).

Dans le cadre de ce stage de M2R, nous concevrons un outil simple permettant d'aligner rapidement des lectures de sRNA-Seq sur un génome. Tout d'abord, le stagiaire concevra un premier module permettant de réduire la complexité des données en faisant disparaître les séquences dupliquées avant de les proposer à un outil d'alignement déjà existant. Un second module se chargera de dupliquer les loci trouvés de manière à conserver l'information d'expression.

Dans un second temps, le stagiaire implémentera un outil simple d'alignement sans erreur en utilisant les filtres de Bloom en utilisant des bibliothèques déjà existantes.

Ce sujet pourra donner lieu à une poursuite en thèse dans le cadre d'un financement demandé pour les années 2015-2018 (non acquis).