Sujet de stage / projet – master bioinformatique

Encadrants : Fabien Sassolas (doctorant, Institut de Génomique Fonctionnelle de Lyon), Jean-Nicolas Volff (professeur, ENS de Lyon / IGFL), Benjamin Audit (directeur de recherche, laboratoire de physique de l'ENS de Lyon)

Contact : fabien.sassolas@ens-lyon.fr

Titre du projet : Création d'un pipeline d'alignement de génomes eucaryotes pour analyser l'évolution de séquences positionnant les nucléosomes

Description :

Les nucléosomes sont des complexes protéiques formés de 4 paires de protéines nommées histones sur lesquelles s'enroule la molécule d'ADN. Ils jouent un rôle essentiel dans la cellule en permettant la protection et le contrôle de l'expression de l'ADN.

Notre équipe a montré que la séquence d'ADN forme des maxima locaux d'énergie qui positionnent les nucléosomes en empêchant leur formation. La séquence au niveau de ces barrières (appelées NIEBs pour Nucleosome Inhibitory Energy Barrier) est trop coûteuse énergétiquement parlant pour être enroulée autour du nucléosome et ces derniers se positionnent donc contre ces barrières.

La caractérisation de ces régions a notamment impliqué l'étude de leur évolution, ce qui a été fait entre autres par Drillon et al. 2016. Cette étude a montré que les séquences autour des NIEBs évoluent de façon particulière, avec une baisse du taux de bases GC au niveau des NIEBs et à l'espace entre les deux nucléosomes, et une hausse du taux de GC à l'endroit où sont positionnés les nucléosomes.

Analyser l'évolution de séquences dans les génomes peut se faire de plusieurs façons. Une possibilité que nous avons exploité jusqu'à présent est de reconstruire le génome de l'ancêtre commun de plusieurs espèces en comparant leurs régions orthologues. Pour ce faire, il est nécessaire de disposer d'alignements deux à deux des génomes de ces espèces.

L'Université de Californie – Santa Cruz (UCSC) propose en téléchargement un certain nombre d'alignements de génome accessibles à cette adresse (« liftover files » car il s'agit de fichiers conçus originellement pour le logiciel du même nom).

Le pipeline en Perl servant à leur création est accessible sur Github, mais a été construit largement pour un usage interne à l'UCSC, impliquant notamment un logiciel de parallélisation développé spécialement pour leur cluster ainsi que nombre de dépendances obsolètes.

L'objectif de ce projet sera de reprendre le code de ce pipeline, de le comprendre, d'analyser les options techniques retenues et de développer un pipeline plus généraliste et dans un langage plus moderne, par exemple Snakemake, ce dernier étant dédié à la conception de pipeline et disposant d'options solides de parallélisation. Aucun niveau particulier en

Snakemake ou en Perl n'est nécessaire dans l'absolu, mais un intérêt important pour la programmation est attendu.

Drillon, Guénola, Benjamin Audit, Françoise Argoul, et Alain Arneodo. 2016. « Evidence of Selection for an Accessible Nucleosomal Array in Human ». *BMC Genomics* 17 (1): 1-20. https://doi.org/10.1186/s12864-016-2880-2.