Compression d'informations de qualité dans les données de séquençage (NGS)

Encadrants:
Raluca Uricaru (contact)
Mail: raluca.uricaru@labri.fr

Guillaume Rizk
Mail: guillaume.rizk@inria.fr

Mots-clés: compression, NGS

Description:
Les séquenceurs nouvelles génération (NGS) permettent de lire l'ADN des cellules d'organismes vivants. Ces dernières années, le tarif du séquençage couplé avec l'augmentation vertigineuse de la capacité de production de ces machines ont profondément modifié le paysage de la biologie cellulaire.

Un problème majeur est maintenant l'énorme masse de données généréé par ces machines, qui produisent en sortie des fichiers en format FASTQ, contenant des centaines de millions de petites séquences nucléiques (appelées reads), suivies par des scores de qualité associés à chacune des bases de ces séquences. Les problèmes de coût de stockage et de transfert de ces données est un réel handicap pour l'utilisation des données NGS. Il y a donc un fort besoin de mettre en place des outils de compression adaptés.

Les algorithmes de compression génériques fonctionnent mais ne donnent qu'un taux de compression faible. Une méthode de compression exploitant les spécificités des données NGS pourrait compresser davantage ces fichiers de type FASTQ. Pour cela, il est nécessaire de traiter indépendamment la partie contenant les séquences nucléiques et celle avec les scores de qualité.

Dans le cadre de ce projet, on s'intéresse au deuxième volet de ce problème: la compression de données de qualité dans les fichiers FASTQ. Le travail comporte une première partie de comprehension et d'analyse des techniques existantes, en se basant, entre autres, sur la publication suivante:

Transformations for the compression of FASTQ quality scores of next-generation sequencing data
Raymon Wan, VO Ngoc Anh and Kiyoshi Asai. Bioinformatics (2012)

La deuxième partie du travail porte sur l'évaluation des méthodes existantes sur des vraies données. Un travail de recherche sur l'amélioration de ces techniques est également envisageable. Dans ce contexte, le sujet pourra être étendu et proposé comme sujet de stage de recherche pour le deuxième semestre.

Bibliographie supplémentaire:
Compression of next-generation sequencing reads aided by highly efficient de novo assembly
Daniel C. Jones, Walter L. Ruzzo Xinxia Peng and Michael G. Katze, Nucl. Acids Res. (2012)

NGC: lossless and lossy compression of aligned high-throughput sequencing data
Niko Popitsch* and Arndt von Haeseler, Nucleic Acids Research (2013)
http://nar.oxfordjournals.org/content/41/1/e27.full