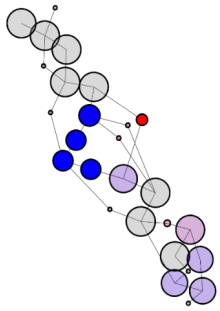
Algorithmes pour les études d'association à l'échelle du métagénome

Contexte

La métagénomique s'intéresse à l'ensemble des génomes retrouvés dans un même environnement naturel : un intestin, une terre agricole ou un échantillon d'eau. Les données sont obtenues en séquençant l'ensemble des molécules d'ADN présentes et mettent en évidence la diversité microbienne de chaque échantillon. Leur disponibilité croît rapidement, et **leur analyse a le potentiel d'améliorer nos connaissance sur des questions essentielles de santé, d'agriculture ou d'écologie**. Elle peut par exemple permettre d'identifier des espèces bactériennes dont l'absence dans le microbiote intestinal est cause d'intolérances alimentaire ou dont la présence dans le sol favorise le rendement agricole, ou encore de caractériser l'impact d'une pollution sur un écosystème.



Le graphe de de Bruijn construit sur la totalité des k-mers présents dans un ensemble de génomes bactériens permet d'identifier et de visualiser des variations génétiques associées à un phénotype tel que la résistance à un antibiotique (extrait de Jaillard et al. [2018]).

Projet

Les études d'association pangénomiques (GWAS) sont un outil essentiel pour identifier des variations génomiques associées à un trait phénotypique. Les approches standards, conçues pour l'étude de génomes uniques, recherchent des mutations le long du génome dont la présence est plus importante pour une valeur du phénotype que pour l'autre—par exemple, une mutation trouvée plus souvent dans le génome de sujets malades que dans celui de sujets sains. Cette approche n'est pas directement applicable aux métagénomes où chaque échantillon contient un ensemble inconnu d'espèces potentiellement différent. Un certain succès a été obtenu pour les GWAS sur une seule espèce bactérienne en s'appuyant sur le contenu en k-mer—tous les mots génomiques d'une longueur k prédéfinie, par exemple "TGATTAG" pour $k = 7$—des génomes plutôt que sur leurs mutations. Dans ce contexte, **le graphe de de Bruijn** , qui représente une succession de l'ensemble des k-mers de l'étude, s'est avéré être un outil essentiel pour leur manipulation—filtrage et compaction—ainsi que pour la visualisation des résultats du GWAS. Cependant, les méthodes existantes atteignent leurs limites sur les métagénomes, qui contiennent un nombre de k-mer bien plus important et pour lesquels les critères de filtrage employés sur une seule espèce sont inadaptés.

Il existe donc deux problématiques parallèles intéressantes pour améliorer les méthodes GWAS en se basant sur les k-mers, qui résument l'objectif du stage :

— traiter le problème de passage à l'échelle en nombre de k-mers
— passer des GWAS génomiques aux GWAS métagénomiques

Ce stage sera l'occasion d'acquérir une expertise sur des algorithmes fondamentaux pour l'analyse des données de séquençage, technologies incontournables en biologie moléculaire moderne. Il sera aussi l'occasion d'apporter une contribution pratique à de nombreux domaines d'application : thérapie microbienne, résistances aux antibiotiques, agronomie en lien avec les changements globaux... Notre méthode de GWAS bactériennes [Jaillard et al., 2018] a été utile à de nombreuses équipes de microbiologie, et plusieurs projets sont en attente d'une extension aux métagénomes. Les outils développés au cours du stage seront du reste utiles pour les GWAS bactériennes, et permettront une application à des génomes eucaryotes, pour lesquels une importante demande existe également.

La personne recrutée travaillera dans la continuité de progrès récents en algorithmique pour le fitrage des k-mers et la construction des graphes de de Bruijn. Il ou elle pourra s'appuyer sur les codes développés précédemment dans les équipes d'accueil, notamment sur le pipeline DBGWAS pour les GWAS bactériennes et les bibliothèques Bcool [Limasset et al., 2019] et Reindeer [Marchet et al., 2020] implémentant des avancées récentes en algorithmes pour la manipulation de k-mers. Elle bénéficiera également des conseils de chercheuses et de chercheurs spécalisé·e·s sur ces questions, et d'un réseau de collaboration déja établi pour les aspects applicatifs. Le stage pourra être localisé à Lille (UMR CRIStAL) ou à Lyon (UMR LBBE), à la préférence de la personne recrutée.

Contact

Camille Marchet (camille.marchet@univ-lille.fr)
Laurent Jacob (laurent.jacob@univ-lyon1.fr)

Références

Magali Jaillard, Leandro Lima, Maud Tournoud, Pierre Mahé, Alex van Belkum, Vincent Lacroix, and Laurent Jacob. A fast and agnostic method for bacterial genome-wide association studies : Bridging the gap between k-mers and genetic events. PLOS Genetics, 14(11) :1–28, 11 2018. doi : 10.1371/journal.pgen.1007758. URL https://doi.org/10.1371/journal.pgen.1007758.

Antoine Limasset, Jean-François Flot, and Pierre Peterlongo. Toward perfect reads : self-correction of short reads via mapping on de Bruijn graphs. Bioinformatics, 36(5) : 1374–1381, 02 2019. ISSN 1367-4803. doi : 10.1093/bioinformatics/btz102. URL https://doi.org/10.1093/bioinformatics/btz102.

Camille Marchet, Zamin Iqbal, Daniel Gautheret, Mikaël Salson, and Rayan Chikhi. REINDEER : efficient indexing of k-mer presence and abundance in sequencing datasets. Bioinformatics, 36 (Supplement_1) :i177–i185, 07 2020. ISSN 1367-4803. doi : 10.1093/bioinformatics/btaa487. URL https://doi.org/10.1093/bioinformatics/btaa487.