DIPLÔME NATIONAL DE DOCTORAT

(Arrêté du 25 mai 2016)

Date de la soutenance : **23 avril 2019**

Nom de famille et prénom de l'auteur : **ISHI SOARES DE LIMA Leandro**

Titre de la thèse : « *Algorithmes de novo pour l'identification de motifs associés à des événements biologiques dans les graphes de De Bruijn construits à partir de données NGS* »

Résumé

L'objectif principal de cette thèse est le développement, l'amélioration et l'évaluation de méthodes de traitement de données massives de séquençage, principalement des lectures de séquençage d'ARN courtes et longues, pour éventuellement aider la communauté à répondre à certaines questions biologiques, en particulier dans les contextes de transcriptomique et d'épissage alternatif.

Notre objectif initial était de développer des méthodes pour traiter les données d'ARN-seq de deuxième génération à l'aide de graphes de De Bruijn afin de contribuer à la littérature sur l'épissage alternatif, qui a été exploré dans les trois premiers travaux. Le premier article (Chapitre 3, article [76]) a exploré le problème que les répétitions apportent aux assembleurs de transcriptome si elles ne sont pas correctement traitées. Nous avons montré que la sensibilité et la précision de notre assembleur local d'épissage alternatif augmentaient considérablement lorsque les répétitions étaient formellement modélisées. Le second (Chapitre 4, article [11]) montre que l'annotation d'événements d'épissage alternatifs avec une seule approche conduit à rater un grand nombre de candidats, dont beaucoup sont importants. Ainsi, afin d'explorer de manière exhaustive les événements d'épissage alternatifs dans un échantillon, nous préconisons l'utilisation combinée des approches mapping-first et assembly-first. Étant donné que nous avons une énorme quantité de bulles dans les graphes de De Bruijn construits à partir de données réelles d'ARN-seq, qui est impossible à analyser dans la pratique, dans le troisième travail (Chapitre 5, articles [1, 2]), nous avons exploré théoriquement la manière de représenter efficacement et de manière compacte l'espace des bulles via un générateur des bulles. L'exploration et l'analyse des bulles dans le générateur sont réalisables dans la pratique et peuvent être complémentaires aux algorithmes de l'état de l'art qui analysent un sous-ensemble de l'espace des bulles.

Les collaborations et les avancées sur la technologie de séquençage nous ont incités à travailler dans d'autres sous-domaines de la bioinformatique, tels que: études d'association à l'échelle des génomes, correction d'erreur et assemblage hybride. Notre quatrième travail (Chapitre 6, article [48]) décrit une méthode efficace pour trouver et interpréter des unitigs fortement associées à un phénotype, en particulier la résistance aux antibiotiques, ce qui rend les études d'association à l'échelle des génomes plus accessibles aux panels bactériens, surtout ceux qui contiennent des bactéries plastiques. Dans notre cinquième travail (Chapitre 7, article [75]), nous évaluons dans quelle mesure les méthodes existantes de correction d'erreur ADN à lecture longue sont capables de corriger les lectures longues d'ARN-seq à taux d'erreur élevé. Nous concluons qu'aucun outil ne surpasse tous les autres pour tous les indicateurs et est le mieux adapté à toutes les situations, et que le choix devrait être guidé par l'analyse en aval.

Les lectures longues d'ARN-seq fournissent une nouvelle perspective sur la manière d'analyser les données transcriptomiques, puisqu'elles sont capables de décrire les séquences complètes des ARN messagers, ce qui n'était pas possible avec des lectures courtes dans plusieurs cas,

même en utilisant des assembleurs de transcriptome de l'état de l'art. En tant que tel, dans notre dernier travail (Chapitre 8, article [74]), nous explorons une méthode hybride d'assemblage d'épissages alternatifs qui utilise des lectures à la fois courtes et longues afin de répertorier les événements d'épissage alternatifs de manière complète, grâce aux lectures courtes, guidé par le contexte intégral fourni par les lectures longues.

MOTS-CLEFS en français
ARN-seq, Lectures courtes, Lectures longues, Épissage alternatif, Graphes de De Bruijn, Bulles, Études d'association à l'échelle des génomes, Correction d'erreurs, Assemblage hybride.