Université de Rennes

Direction de la Recherche et de l'Innovation

Appel à projets - Année 2024

Contrats doctoraux «politique scientifique de l'établissement»

ECOLE DOCTORALE : Matisse (& EGAAL)

POLE DE RECHERCHE : Mathématique-Numérique & Environnement

Unité de recherche d'accueil

Unité de Recherche : IRISA/centre Inria de l'université de Rennes

Equipe : Genscale

Encadrement

Directeur de thèse : Claire Lemaitre **Unité de recherche** : UMR 6074 IRISA/centre Inria de l'université de Rennes

Employeur/Établissement d'origine : Inria

HDR : x oui ☐ non **Nb de thèses en cours** : 3 (250%)

Mail : claire.lemaitre@inria.fr **Tél**. : 0299847116

Co-directeur de thèse : Claire Mérot **Unité de recherche** : UMR 6553 ECOBIO

Employeur/Établissement d'origine : CNRS

HDR : ☐ oui x non (inscription en cours pour passage en 2024) **Nb de thèses en cours** : 1 (50%)

Mail : claire.merot@univ-rennes.fr **Tél.** :

Sujet

Titre : Développement de méthodes bioinformatiques pour la caractérisation des variations structurales dans les génomes avec des données linked-read.

Mots clés : informatique – biologie – bioinformatique – algorithmique des séquences – comparaison de séquences – indexation de séquences – graphes de pangenome – génomique – variants de structure – populations – évolution - conservation

Descriptif (15-30 lignes) :
Contexte socioéconomique et scientifique ; Les hypothèses et questions posées ; Les grandes étapes de la thèse et démarche ; Approches méthodologiques et techniques envisagées

Une question fondamentale en biologie est d'interpréter les variations entre les génomes aussi bien pour comprendre l'évolution des espèces, que pour conserver la biodiversité et optimiser l'utilisation des ressources naturelles. Les progrès du séquençage ont mis en valeur la prévalence des variations de structure du génome (segments d'ADN dupliqués, supprimés, inversés ou déplacés) qui couvrent 5 à 10 fois plus de bases dans le génome que les mutations ponctuelles analysées couramment. C'est l'essor du séquençage de 3ème génération (lectures longues) qui a permis ces 5 dernières années d'enfin caractériser et cataloguer toute la gamme de SVs dans de nombreux génomes. Mais cette technologie reste chère et cette caractérisation est limitée souvent à un faible nombre d'individus ou à des espèces modèles, telles que l'homme. Une nouvelle technologie de séquençage permet de baisser les coûts en séquençant simultanément des centaines d'individus avec une approche « linked-reads » (lectures courtes associées à des barcodes donnant une information longue distance). Si cette technologie est extrêmement prometteuse pour les études à l'échelle des populations, l'approche par barcodes rend les méthodes existantes pour les longues lectures inadaptées et nécessite le développement de nouvelles méthodes informatiques pour détecter et analyser les SVs.

L'objectif de ce doctorat sera de développer des méthodes de détection et génotypage des SVs combinant l'information de barcodes et de séquence des linked-reads, puis de tester les outils développés sur des données réelles issues des recherches empiriques en biologie afin de répondre aux questions suivantes. Quelle est la précision et la puissance de détection par données linked-reads pour différentes gammes et types de variations génomiques structurales ?

Comment optimiser ces outils pour l'application à l'échelle de grands jeux de données (espèces, populations) d'une part et à des variants complexes d'autre part ?

Principales étapes de la thèse:
(i) analyse des signaux de barcodes générés par les différents types de SV dans les données disponibles de mouches du varech et développement d'algorithmes pour les détecter.
(ii) développement d'un logiciel automatisant la détection des SVs avec des données linked-reads et application à grande échelle sur différentes espèces ainsi que différents types et complexités de SVs.
(iii) améliorations méthodologiques pour le passage à l'échelle sur un grand nombre d'individus et pour la détection de SVs plus complexes (caractérisation fine des faux négatifs et faux positifs de la méthode)

Approches méthodologiques et techniques:
Les méthodes développées seront basées sur l'algorithmique du texte et des graphes (alignement de séquences, parcours de graphes de séquences) ainsi que sur les structures d'indexation de séquences. Les logiciels seront implémentés en Python, C++ ou Rust. Le développement méthodologique sera guidé et évalué par les applications et analyses sur données réelles.

S'agit-il d'un projet de thèse en cotutelle internationale (oui/non) : non

Si oui, préciser l'établissement et le pays de rattachement :

Critères« politique doctorale générale»

☐ **International**
☐ **Attractivité**
x Interdisciplinarité *(validation par les unités et ED concernées)*

Sollicitation d'un cofinancement EDUC (oui/non) : non

Argumentaire et adéquation du sujet avec les critères de l'appel à projet :
(Pour le critère international : préciser les engagements de réciprocité attendus du partenaire international)

Le sujet de thèse se rattache à deux pôles : Mathématique-Numérique et Environnement. Il relève d'une coopération entre deux unités, IRISA et ECOBIO.

Pour étudier les variations génomiques en médecine, écologie ou biologie, il faut être capable de les détecter avec puissance et précision. C'est là que l'informatique et la bioinformatique apportent aujourd'hui les innovations nécessaires à une meilleure compréhension des génomes. Le doctorat proposé s'inscrit à cette interface entre informatique et biologie. Il vise à développer des outils informatiques applicable aux analyses biologiques. L'échange entre développements bioinformatiques et applications biologiques tout au long du doctorat permettra de développer un corpus de recherche interdisciplinaire intégrant aussi bien des questions d'algorithmique des séquences que d'évolution de la variation biologique.

Le ou la doctorant·e bénéficiera de l'expertise complémentaire et des ressources des deux équipes d'accueil.

- L'équipe Genscale à l'IRISA (rattachement principal), avec Claire Lemaitre, possède une forte expertise sur les méthodes de détection et d'analyse des SVs et a déjà initié des développements sur ce type de données linked-reads avec la librairie C++ LRez permettant notamment l'indexation par barcode de ces données. Ces données étant très volumineuses, cette structure de données dédiée est un atout pour le développement de méthodes d'analyses efficaces en temps et en mémoire.
- Au sein de l'équipe Evo-Adapt à ECOBIO (rattachement secondaire), Claire Mérot, étudie l'impact des SVs sur l'adaptation au changement climatique des insectes avec des approches expérimentales et génomiques. Elle a généré des données de re-séquençage par linked-reads et par lectures longues pour plusieurs populations de

mouches du varech. Ces données sont inédites et idéales pour une évaluation fine et réaliste des performances des méthodes développées en comparaison avec la stratégie plus coûteuse par lectures longues. Les données disponibles (et en cours d'acquisition grâce à un financement ERC-Stg) garantissent une mise en œuvre immédiate et efficace du projet bioinformatique.

Par ailleurs, C. Lemaitre et C. Mérot font partie d'un groupe de chercheurs français et européens qui s'intéressent à l'utilisation des données de type linked-reads pour des questions de conservation (démographie récente), d'amélioration des pangénomes (phasing des haplotypes), et d'association génétique-trait, en incluant des séquences chez des poissons, plantes, coquillages et reptiles. Ces interactions compléteront le champ d'application des travaux de thèse et offriront un cadre interdisciplinaire élargi bénéfique au/à la doctorant.e.