Offre n°2024-07673

Doctorant F/H Limitation de la taille des structures de données indexant les séquences génomiques.

Type de contrat : CDD

Niveau de diplôme exigé : Bac + 5 ou équivalent

Fonction : Doctorant

Niveau d'expérience souhaité : Jeune diplômé

A propos du centre ou de la direction fonctionnelle

Le centre Inria de l'Université de Rennes est l'un des neuf centres d'Inria et compte plus d'une trentaine d'équipes de recherche. Le centre Inria est un acteur majeur et reconnu dans le domaine des sciences numériques. Il est au cœur d'un riche écosystème de R&D et d'innovation : PME fortement innovantes, grands groupes industriels, pôles de compétitivité, acteurs de la recherche et de l'enseignement supérieur, laboratoires d'excellence, institut de recherche technologique.

Contexte et atouts du poste

Cette thèse s'inscrit dans le PEPR "AgroEcologie Numérique". Elle sera encadrée par l'équipe GenScale. Le travail se fera de concert avec les membres du PEPR, en particulier du "flagship AgroDiv".

Pour faire face aux contraintes du changement climatique, tout en répondant aux objectifs de l'agroécologie, ce groupe, principalement composé de biologistes et de bio-analyses, a pour ambition de caractériser efficacement la diversité génétique inexploitée, stockée et disponible dans les collections. Il s'agit de 20476 espèces animales (lapin, abeille, truite, poulet, porc, chèvre, mouton, bovins...) et de 7466 espèces végétales (blé, maïs, tournesol melon, choux, navet, abricotier, pois, fèverole, luzerne, tomate aubergines, pommier, cerisier, pêcher, vigne...) majeures de l'Agriculture Française.

Ce poste s'inscrit dans l'un des axes de recherche de ce groupe, consistant à développer des moteurs de recherche conviviaux pour filtrer rapidement et efficacement les données des collections et des essais sur le terrain afin d'évaluer « fonctionnellement » les accessions ou les populations d'intérêt.

Mission confiée

Objectifs de la thèse et méthodes :

Les moteurs de recherches actuels permettant de faire des requêtes sur des données génomiques sont principalement basés sur la notion de k-mers (mots de taille k). Il est nécessaire d'indexer les k-mers de tous les jeux de données que l'on souhaite pouvoir requêter.

Les meilleurs index actuels [1] nécessitent environ 10 à 15% de la taille des données brutes. Il est nécessaire de réduire leur taille afin de pouvoir indexer des données à l'échelle généralisée, atteignant plusieurs pétaoctets.

Les directions de recherches seront alors pleinement consacrées à cette réduction.

- possibilité d'organiser les données pour mieux compresser les indexes [2]
- possibilité de ne pas indexer tous les k-mers, au prix de résultats déteriorés [3]
- possiblité de proposer de nouvelles structures de données, avec une compression intrinsèque des données de jeux de séquences similaires
- etc...

Objectifs
La doctorante ou le doctorant aura pour mission d'explorer de nouvelles approches permettant d'améliorer les résultats existants en terme d'indexation de données génomiques. Il pourra s'agir d'une ou plusieurs contributions majeures parmi les points suivants :
1/ améliorer le passage à l'échelle : indexer plus de jeux de données, ou des jeux de données de plus en plus complexes en terme de diversité;
2/ développer de nouvelles approches pour associer des métadonnées aux kmers (abondance, séquence et position dont ils sont issus, annotations connues, ...);
3/ limiter l'impact environnemental de la construction, du stockage et de la requête des indexes

proposés.
4/ interagir avec les membres du PEPR AgroDiv en particulier et bien sur la communauté des utilisateurs en général.

[1] Lemane, Téo, et al. "Indexing and real-time user-friendly queries in terabyte-sized complex genomic datasets with kmindex and ORA." *Nature Computational Science* 4.2 (2024): 104-109.

[2] Břinda, Karel, et al. "Efficient and Robust Search of Microbial Genomes via Phylogenetic Compression." bioRxiv (2023).

[3] Darvish, Mitra, et al. "Needle: a fast and space-efficient prefilter for estimating the quantification of very large collections of expression experiments." Bioinformatics 38.17 (2022): 4100-4108.

Principales activités

Principales activés :

- Etudes de l'état de l'art
- Recherche et développement algorithmique
- Tests et validations
- Implémentation
- Rédactions

Activités complémentaires :

- Vie de l'équipe
- Veille technologique
- Lien avec les utilisateurs

Compétences

Compétences techniques et niveau requis :

- expérience significative en programmation (si possible en C++ ou rust)
- expérience et gout pour l'algorithmique et les structures de données
- connaissances en développement de logiciel
- présentations, redaction et lecture en anglais

Avantages

- Restauration subventionnée
- Transports publics remboursés partiellement
- Possibilité de télétravail à hauteur de 90 jours annuels
- Prise en charge partielle du coût de la mutuelle

Informations générales

- **Thème/Domaine :** Algorithmique, calcul formel et cryptologie
  Calcul Scientifique (BAP E)
- **Ville :** Rennes
- **Centre Inria :** Centre Inria de l'Université de Rennes
- **Date de prise de fonction souhaitée :** 2024-10-01
- **Durée de contrat :** 3 ans
- **Date limite pour postuler :** 2024-06-30

Contacts

- **Équipe Inria :** GENSCALE
- **Directeur de thèse :**
  Peterlongo Pierre / pierre.peterlongo@inria.fr

A propos d'Inria

Inria est l'institut national de recherche dédié aux sciences et technologies du numérique. Il emploie 2600 personnes. Ses 215 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3900 scientifiques pour relever les défis du numérique, souvent à l'interface d'autres disciplines. L'institut fait appel à de nombreux talents dans plus d'une quarantaine de métiers différents. 900 personnels d'appui à la recherche et à l'innovation contribuent à faire émerger et grandir des projets scientifiques ou entrepreneuriaux qui impactent le monde. Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 200 start-up. L'institut s'efforce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l'économie.

L'essentiel pour réussir

Master d'informatique avec motivation pour les questions biologiques ou master de bioinformatique

Attention: Les candidatures doivent être déposées en ligne sur le site Inria. Le traitement des candidatures adressées par d'autres canaux n'est pas garanti.

Consignes pour postuler

Merci de déposer en ligne CV, lettre de motivation et éventuelles recommandations

Sécurité défense :
Ce poste est susceptible d'être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L'autorisation d'accès à une zone est délivrée par le chef d'établissement, après avis ministériel favorable, tel que défini dans l'arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l'annulation du recrutement.

Politique de recrutement :
Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.