Type d'offre : Offre en laboratoire

Date de publication : 21.03.25

CEA - LABGeM

PhD Offer - Language models at the scale of pangenome graphs for biological function prediction

Informations générales

Type de contrat : CDD

Durée du contrat : 3 ans

Contact :

Alexandra Calteau / **David Vallenet**

Date de prise de poste : mer 01/10/2025 - 12:00

Métier : PhD

Thématique : Autre

Précisez : Bioinformatics

CEA - LABGeM :

Le **LABGeM** est une équipe de bioinformatique de l'UMR 8030 Génomique Métabolisme qui fait partie de l'infrastructure France Genomique, situé à Evry. Les activités scientifiques du LABGeM sont centrées sur l'analyse bioinformatique : des (méta)génomes microbiens : dynamique et évolution des génomes bactériens, annotation fonctionnelle des (méta)génomes, attribution taxonomique des données métagénomiques ; du métabolisme bactérien & Biologie des systèmes : prédiction, curation et comparaison des réseaux métaboliques, recherche d'enzymes "orphelines", découverte de nouvelles activités enzymatiques. Ces activités de R&D participent à l'un des principaux axes de recherche de l'UMR "Génomique Métabolisme" : l'élucidation du métabolisme des procaryotes à travers la découverte de nouvelles réactions chimiques catalysées par le monde vivant.

Détail de l'offre (poste, mission, profil) :

Contexte de l'offre

Nous recherchons un(e) doctorant(e) enthousiaste avec un profil informatique/machine learning/statistic pour travailler sur le développement de nouvelles méthodes de pan-génomique comparative exploitant les modèles de langage. Cette thèse est financée par le CEA.

Les procaryotes (c'est-à-dire les bactéries et les archées) constituent un domaine fascinant d'organismes vivants, représentant une diversité et une ubiquité remarquables. Leur impact sur la biosphère est immense, influençant la santé humaine et animale, la biogéochimie des sols et des océans, et bien plus encore.

L'exploration à grande échelle des génomes microbiens a permis de découvrir les mécanismes moléculaires sous-jacents à leur diversité, et en particulier le rôle des éléments génétiques mobiles (EGM).

Ces dernières années, avec l'explosion des projets de séquençage, plusieurs approches bioinformatiques ont été développées basées sur le concept de pan-

génome, offrant des solutions pour gérer et exploiter efficacement de grandes quantités de données. La pan-génomique examine la variabilité génétique à travers tous les génomes disponibles d'un groupe donné, généralement une espèce, plutôt que de se baser sur un seul génome de référence ou de faire des comparaisons par paires. En termes de contenu génétique, on fait une distinction entre le génome de base, c'est-à-dire les gènes présents chez tous les individus, et les gènes accessoires (ou variables) qui sont plus ou moins conservés dans les génomes, et donc susceptibles d'expliquer les particularités phénotypiques. Le développement des méthodes pan-génomiques constitue ainsi une réponse au défi des données massives en biologie, contribuant à comprendre l'évolution des micro-organismes en lien avec des données épidémiologiques ou environnementales.

Depuis plusieurs années, le laboratoire LABGeM travaille sur un modèle pour représenter les données génomiques sous la forme d'un graphe de pan-génome au niveau des familles de gènes, permettant ainsi la compression des informations provenant de milliers de génomes tout en préservant l'organisation chromosomique des gènes. Les recherches ont abouti au développement des outils **PPanGGOLiN** et **PANORAMA.**

Les méthodes actuelles d'analyse des contextes génomiques ont démontré leur efficacité dans la prédiction des fonctions biologiques, mais souffrent de problèmes de mise à l'échelle pour exploiter pleinement la diversité des génomes disponibles dans les bases de données. PANORAMA offre l'une des premières perspectives d'analyse comparative du pan-génome des contextes génomiques dans des milliers de génomes, mais repose sur des règles algorithmiques prédéfinies pour identifier des systèmes biologiques similaires, ce qui limite sa capacité à découvrir des systèmes entièrement nouveaux. De nouvelles méthodes d'intelligence artificielle basées sur les transformateurs pour les modèles de langage ont montré leur efficacité dans la capture des relations sémantiques à grande échelle grâce à des mécanismes d'attention et commencent à être utilisées pour prédire et générer de nouveaux contextes génomiques.

Missions

Cette thèse propose d'exploiter les méthodes d'intelligence artificielle, en particulier les modèles de langage, appliquées aux graphes de pan-génome. En représentant leur contenu sous forme de séquences de phrases, où chaque mot correspond à une

unité fonctionnelle codée par une famille de gènes, cette approche ouvre de nouvelles perspectives pour révéler des motifs complexes grâce à l'apprentissage sur des ensembles de données à grande échelle. Cela permettra de prédire des annotations manquantes ou incertaines, offrant ainsi des informations sur la fonction des gènes et les processus biologiques non caractérisés. Les principaux objectifs de ce travail seront de :

- Construire un jeu de données de graphes de pan-génome annotés à différents niveaux fonctionnels, servant de base pour l'entraînement et la validation des modèles.

- Évaluer différentes méthodes d'apprentissage automatique, y compris les modèles de langage, afin d'identifier les approches les plus performantes.

- Appliquer la méthode développée à l'identification de nouveaux systèmes biologiques, tels que des voies métaboliques, des systèmes macromoléculaires ou de défense.

Candidatures

Pour postuler, envoyez un CV accompagné d'une lettre de motivation et de références avant le 27 avril 2025 à Alexandra CALTEAU (acalteau@genoscope.cns.fr) et David VALLENET (vallenet@genoscope.cns.fr).
Le poste sera situé au Genoscope à Évry.

En savoir plus :

Plus d'information

Date limite pour postuler : dim 27/04/2025 - 12:00

Lien vers l'offre sur le site dataia.eu :https://da-cor-dev.peppercube.org/node/1261