Type d'offre : Laboratory offer

Post date : 21.03.25

CEA - LABGeM

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

Informations générales

Contract type : Fixed-term contract

Contract length : 3 years

Contact :

Alexandra Calteau / **David Vallenet**

Starting date : Wed 01/10/2025 - 12:00

Trade : PhD

Topic : Autre

Précisez : Bioinformatics

CEA - LABGeM :

The LABGeM is a bioinformatics team of the UMR 8030 Genomics Metabolics, the basic research structure of Genoscope (the national sequencing center), now part of the France Génomique infrastructure. Scientific activities of the LABGeM are centered on the bioinformatics analysis of Microbial (meta)genomes: dynamics and evolution of bacterial genomes, functional annotation of (meta)genomes, taxonomic assignation of metagenomic data Bacterial metabolism & System Biology: prediction, curation and comparison of metabolic networks, investigation for 'orphan' enzymes, discovery of new enzymatic activities. These R&D activities participate to one of the main research topic of the UMR "Genomics Metabolics": the elucidation of the metabolism of prokaryotes through the discovery of new chemical reactions catalyzed by the living world.

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

General Context

We are looking for an enthusiastic Ph.D student to work on the development of new methods for comparative pangenomics exploiting language models. This thesis is funded by the CEA.
Prokaryotes (i.e. bacteria and archaea) constitute a fascinating field of living organisms, representing remarkable diversity and ubiquity. Their impact on the biosphere is immense, influencing human and animal health, soil and ocean biogeochemistry, and much more. Large-scale exploration of microbial genomes has helped uncover the molecular mechanisms underlying their diversity, and particularly the role of Mobile Genetic Elements (MGE). In recent years, with the explosion of sequencing projects, several bioinformatics approaches have been developed based on the pangenome concept, offering solutions for efficiently managing and exploiting large quantities of data. Pangenomics examines genetic

variability across all available genomes of a given group, usually a species, rather than relying on a single reference genome or making pairwise comparisons. In terms of gene content, a distinction is made between the core genome, i.e. the genes present in all individuals, and the accessory (or variable) genes that are more or less conserved in the genomes, and therefore likely to explain phenotypic particularities. The development of pangenomic methods is thus a response to the challenge of massive data in biology, helping to understand the evolution of microorganisms in relation to epidemiological or environmental data.

For several years now, the LABGeM laboratory has been working on a model to represent genomic data as a pangenome graph at the gene family level, enabling the compression of information from thousands of genomes while preserving the chromosomal organization of genes. The research works have resulted in developping tools such as **PPanGGOLiN** and **PANORAMA.**

Current methods for analyzing genomic contexts have shown their effectiveness in predicting biological functions, but suffer from problems of scaling up to fully exploit the diversity of genomes available in databases. PANORAMA offers one of the first perspectives in comparative pangenomic analysis of genomic contexts in thousands of genomes, but relies on predefined algorithmic rules to identify similar biological systems, which limits its ability to discover completely new ones. New Transformer-based artificial intelligence methods for language models have shown their effectiveness in capturing large-scale semantic relationships through attention mechanisms and are beginning to be used to predict and generate new genomic contexts.

Missions

This thesis proposes to exploit artificial intelligence methods, in particular language models, applied to pangenome graphs. By representing their contents as sequences of sentences, where each word corresponds to a functional unit encoded by a gene family, this approach opens up new prospects for revealing complex patterns through learning on large-scale datasets. This will make it possible to predict missing or uncertain annotations, offering insights into gene function and uncharacterized biological processes. The main objectives of this work will be to :

- build a dataset of annotated pangenome graphs at different functional levels, serving as a basis for model training and validation,

- evaluate different machine learning methods, including language models, in order to identify the best performing approaches,

- apply the developed method to the identification of new biological systems, suchas metabolic pathways, macromolecular or defense systems.

Application

To apply, send a CV with a covering letter and references before 27 April 2025 to Alexandra CALTEAU (acalteau@genoscope.cns.fr) and David VALLENET (vallenet@genoscope.cns.fr).
The position will be located at the Genoscope in Evry.

Learn more :

More information

Closing date for submitting applications : Sun 27/04/2025 - 12:00

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