2020-03187 - PhD Position F/M Dynamic pan-genome graphs

**Contexte et atouts du poste**

**Alpaca project presentation**

Funded by the European Commission through the Horizon 2020 Marie Sklodowska-Curie ITN Programme, the ALPACA network offers a high level fellowship for joint research on new data structures, algorithms and statistical / machine learning approaches to store, arrange, process and analyze millions of individual genomes and genetic profiles. The most talented and motivated students will be selected for advanced multidisciplinary research training, preferably starting September / October 2021.

**Genomes are strings over the letters A,C,G,T, which represent nucleotides, the building blocks of DNA. In view of genome sequence data emerging from ever more and technologically rapidly advancing genome sequencing devices, amounting to exabytes in the meantime, the driving question is:**

How can we arrange and analyze these data masses in a computationally / mathematically / statistically appropriate way such that we can redeem the biomedical promises of these data masses, with respect to understanding cancer rare genetic diseases, and the development, the virulence and resistance patterns of pathogens?

The individual variation that affects evolutionarily related, large collections of genomes (termed pan-genomes) follows patterns that the laws of genetics systematically imply. This explains why graph-based data structures, which focus on highlighting the individual variation, while summarizing redundancies in a compact way, have clear benefits over the na{"i}ve idea to store genomes as strings.

However, although having proven to hold great promises, research about graph-based data structures that enable to capture (exabytes of) genetic data can be considered to still be in its infancy. The goal of the research conducted in this project is to make substantial progress with respect to the design and development of such graph-based data structures.

**ABOUT THE PROJECT**

The move from sequence- to graph-based pan-genome data structures is unavoidable when seeking to exploit the wealth of genetic data, instead of having devices massively congested. Putting the paradigm shift (from sequences to graphs) in effect requires new ways of thinking about genomes, as well as computer programs and mathematical models that reflect this.

However, developing, maintaining and computationally / statistically exploiting graph-based pan-genomes requires skills that common-day education does not yet provide. The goal of this project is to train a new "class" of researchers / operators / administrators who are able to deal with the (exabyte-scale) masses of genome data in terms of the progressive, graph-based approaches the research of this project deals with.

ESR's (early-stage researchers = PhD students) will carry out corresponding research at Inria Rennes. To acquire further practical skills, the ESR will also spend time at partner institutions, among which leading industrial players, in the frame of month-long secondments.

For more information on the available position and more detailed project description, please visit https://alpaca-itn.eu/.

**Ph.D. specific presentation**

We propose to explore distinct approaches when creating or adding information to a pan-genome graph. The simplest approach is to map new sequences, indicating newly discovered variants and annotating existing ones. However, when the graph is getting too complex and/or too big, we may have interest to split it into two (or more) sub-graphs. The objective of this ESR will be to determine the best strategy to adopt depending on data size and complexity, from high-quality trustable sequences (perfectly assembled genomes) to lower quality sequences (badly assembled data) or even unassembled sequences.

**Context**

The individual variation that affects evolutionarily related, large collections of genomes (termed pan-genomes) follows patterns that the laws of genetics systematically imply. This explains why...
graph-based data structures, which focus on highlighting the individual variation, while summarizing redundancies in a compact way, have clear benefits over the naive idea to store genomes as strings. However, although having proven to hold great promises, research about graph-based data structures that enable to capture the diversity of genetic data can be considered to still be in its infancy. The goal of the research conducted in this project is to make substantial progress with respect to the design and development of such graph-based data structures.

In the global context of the Alpaca project (https://alpaca-itn.eu/), the specific aim of the Ph.D. is to propose and develop strategies related to the dynamic of pan-genome graphs. Current state of the art approaches [1,2] propose solutions based on sequence mapping on genome graphs. Dynamic of pan-genome graphs is an open problem, as mentioned in [3] “Deciding which variation should be added to a graph is nontrivial”, and existing methods make use of heuristics while adding new sequences or variants.

Mission confiée

Assignments:

The recruited Ph.D. will explore new ways for incrementing pan-genome graph(s) with novel sequences or pre-determined variants, while maintaining graph features (annotations, indexation). A special care will also be provided regarding the graph splitting strategies: when the graph is getting too complex and/or too big, we may have interest to split it into two (or more) sub-graphs. Sub-graphs can also be seen as hierarchical, storing distinct level of informations (variants, strains, species, metadata, ...)

The work will include pure algorithmic tasks (indexation, data representation, mapping) and application to real biological data-sets in collaboration with the CEA Genoscope and Deinove company, who will co-supervise the work.

Collaboration:

The recruited Ph.D. student will work in the GenScale team at Inria Rennes, France. She/he will work in close collaboration with the Alpaca partners, in particular with CEA Genoscope France, Deinove France, and Bielefeld University Germany. Several stays in those remote places are planned. Camille Marchet (CNRS Lille) will be the main supervisor.

Principales activités

All activities needed during a Ph.D.

Compétences

To comply with the funding rules of the Horizon 2020 Marie Sklodowska-Curie programme:

- You qualify as an Early Stage Researcher, meaning that - on the starting date of your employment with the host institute – you are in the first four years of your research career and have not (yet) been awarded a doctoral degree.
- You have not resided and/or have had your main activity (study, work, etc.) in France for more than 12 months during the 3 years prior to the starting date of your employment with the respective host institute.

Candidates must have strong interest and expertise in algorithmics, data structures and languages such as C/C++/Rust.

Knowledge or at least a deep motivation in genomics and biology is a prerequisite.

Candidate should be proficient in English language (academic level).

Avantages

- Subsidized meals
- Partial reimbursement of public transport costs

Rémunération

Monthly gross salary amounting to 1982 euros for the first and second years and 2085 euros for the third year.