2018-00291 - PhD offer: Search engine for genomic sequencing data

Contract type : Public service fixed-term contract  
Level of qualifications required : Graduate degree or equivalent  
Fonction : PhD Position

About the research centre or Inria department

Inria, the French National Institute for computer science and applied mathematics, promotes “scientific excellence for technology transfer and society”. Graduates from the world's top universities, Inria's 2,700 employees rise to the challenges of digital sciences. With its open, agile model, Inria is able to explore original approaches with its partners in industry and academia and provide an efficient response to the multidisciplinary and application challenges of the digital transformation. Inria is the source of many innovations that add value and create jobs.

Context

The main objective is to produce a model and a prototype dedicated to allowing users to directly query large unassembled raw sequencing data on the fly in order to tap into the largest underexploited resource in life sciences.

Assignment

For a better knowledge of the proposed research subject:

We are currently witnessing a deep knowledge revolution due to the availability of exponentially expanding sequence databases made possible by the continuously accelerating throughput of sequencing techniques. Sequencing data is accumulating faster than Moore's Law, bringing fundamental new insights, conjecture, and understanding, with impacts in medicine, agronomy and ecology. Today, the INSDC SRA raw data archive stores more than 10^{10} (10 000 PB) nucleotides, in the form of short sequences (<1000 PB) which represent fragments from generally unknown genomic location (the "reads"). However, the overwhelming majority of those sequences have only been analyzed within the context of single projects addressing each a small fraction of the total resource. It is therefore of primary importance to maintain this trace of diversity for future studies and to develop technologies to interrogate these data. Moreover, providing fast access to the sum of all data would open the doors to novel discoveries that a single or a limited number of read sets do not have the power to address.

Assignments:

The recruited person will be taken to design and propose new indexing scheme, scaling up very large DNA collection (assembled or not), and offering a way to query in real time input sequences of interest. There exist methods such as Sequence Bloom Tree and as Bloom Filter Trie, that index and compress (lossless or not) such banks. In this project, we will explore the novel idea of representing the bank in a global incremental compressed index using a graph representation of all corrected reads from the whole bank read sets.

Main activities

Main activities

- Datastructure model design (bloom filters, minimal perfect hash functions, ...)
- Prototype developments (C/C++)
- Tests on simulated and real data
- Interface with biologist users
- Diffusion (publications, talks)

Skills

Languages: French - English
Other valued appreciated: Pedagogy, writing skills.

Benefits package
- Subsidised catering service
- Partially-reimbursed public transport
- Social security
- Paid leave
- Sports facilities

Remuneration
Fix term contract
Duration: 36 months

Gross Salary: 1 982€/month (before taxes) during the first 2 years, 2 085€/month (before taxes) during the third year