Integrative High Performance BigData mining : Application to metagenomics and metabolomics

Mots clés :
- Directeur de thèse : Jean-Daniel ZUCKER
- Co-encadrant(s) :
- Unité de recherche : Unité de modélisation mathématique et informatique de systèmes complexes
- Ecole doctorale : École Doctorale Informatique, Télécommunications, Électronique de Paris
- Domaine scientifique principal: Divers

Résumé du projet de recherche (Langue 1)

The main goal of this PhD is to modify/improve algorithms that facilitate the scale-up the integration and the mining of Big OmicsData. A concrete application is to support the extensive use of large catalogs available in metagenomics and metabolomics to efficiently perform annotation, clustering and prediction of NGS data from the Metacardis project. The subject of this thesis is technically contextualized in three main areas regarding NGS Data: i) DBMS (dataset optimization and indexing), ii) Data Transformation (NGS pipeline), iii) High Performance Computing. 1 - Dataset optimization and indexing The PhD candidate should explore and propose new methods based on indexation, map-reduced or other solutions to make possible the exploration of any big data matrix, regardless of its size. R as an important and increasingly used statistical platform is not adapted for big data mining due to size and computational limitations, but offers nevertheless very powerful analytical capacities. The metagenomic and metabolomic data formatted as very large matrixes are thus unadapted for R analysis. Such approaches will open the way for parallelized computing. 2 - Bioinformatics pipeline improvement The bioinformatic processing from reads to counts and from counts to frequencies is still under research in the metagenomics field. Even though much progress has been made, issues such as filtering, normalization and dimensional reduction are still to be improved. Other issues includes identification of metagenomic ecosystems based on domain specific network metrics, Functional annotations of species and ecosystems, de novo metagenome assembly based on de Bruijn graph with high requirement of memory. 2 - High performance scientific computing Big data mining comes with the need for high performance computing. The PhD candidate will also explore different computation solutions adapted for the different bioinformatics pipelines. Local solutions such as HPC, GPU or grid computing but also distant solutions such as cloud computing will analyzed. The exploration of such solutions has not been done before and is of major interest in the field.