Genomic Big Data Management, Modeling And Computing

Masseroli M[1,§], Pinoli P[1], Canakoglu A[1], Bernasconi A[1], Gulino A[1], Nanni L[1], Orlova O[1], Pallotta S[1], Ceri S[1]

[1] *Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy*

[§] Email:marco.masseroli@polimi.it

Motivation

Modern genomics promises to answer fundamental questions for biological and clinical research, e.g., how cancer develops, how driving mutations occur. Unprecedented efforts in genomics are made possible by Next Generation Sequencing (NGS), a family of technologies that is progressively reducing the cost and time of reading the DNA. Huge amounts of sequence data are continuously collected by research laboratories, often organized through world-wide consortia (such as ENCODE, TCGA, the 1000 Genomes Project, and Epigenomics Roadmap); precision medicine, based on genomic information, is becoming a reality. So far, the bioinformatics research community has been mostly challenged by primary analysis (production of sequences in the form of short DNA segments, or "reads") and secondary analysis (alignment of reads to a reference genome and search for specific genomic features on the reads, such as variants/mutations and peaks of expression). The most important emerging problem is the so-called tertiary analysis, concerned with sense making, e.g., discovering how heterogeneous regions interact with each other, by integrating heterogeneous DNA features, such as variants or mutations in a DNA position, or signals and peaks of expression, or structural properties of the DNA, e.g., break points (where the DNA is damaged) or junctions (where DNA creates loops). According to many biologists, answers to crucial genomic questions are hidden within genomic data already available in public repositories, but suitable tools for processing them are lacking.

Methods

We previously proposed a paradigm shift in genomic data management, based on the Genomic Data Model (GDM), which mediates existing data formats, and the GenoMetric Query Language (GMQL), a high-level, declarative query language required by tertiary data analysis. Its enhancement and implementation using a cloud-based technologies provided a GMQL-based system with enhanced accessibility, portability, scalability and performance.

Results

Our new GMQL system is publicly available at http://www.bioinformatics.deib.polimi.it/GMQLsystem/); it has a modular architecture including an intermediate representation supporting operations over genomic regions and metadata that are executed by a Spark engine, a data framework on the cloud that proved to be extremely efficient in supporting massive genomic queries, a high-level technology-independent repository abstraction supporting different repository types (e.g., local file system, Hadoop File System, or others), several system interfaces, including an intuitive Web-based interface and a Web Service interface publicly available at http://www.gmql.ue/, as well as two programmatic interfaces: a

pyGMQL library (https://pygmql.readthedocs.io/en/latest/) for Python language and a RGMQL package (https://bioconductor.org/packages/release/bioc/html/RGMQL.html) for R/Bioconductor environment. The GMQL system is associated with an integrated repository of heterogeneous datasets from multiple public sources. It includes datasets form ENCODE, Roadmap Epigenomics and TCGA, whose metadata have been homogenized through a rule-based framework that we specifically developed for automatize the maintenance of the repository and its extension with the inclusion of new data sources. Data and metadata integrated in the repository can be comprehensively queried through GMQL scripts, as demonstrated by several biological use case examples, and extracted data can then be directly processed and analyzed using analytical Python libraries or R/Bioconductor packages. Further extensions of our GMQL system will support data sharing and federated processing of data located in distinct GMQL instances running on different systems/clouds, by automatically splitting and transferring the processing where the data to be evaluated are located. This will avoid as much as possible expensive downloads and transfers of big data quantities between repositories, also allowing taking advantage of processing resources available in different organizations. Furthermore, and more important, it will ease the integrated processing of valuable data made available by different sources, for their comprehensive evaluation toward biomedical knowledge discoveries.
