HIGHLIGHT LECTURE

Data-Driven genomic computing: Making sense of the signals from the genome

Next Generation Sequencing (NGS) allows the production of the entire human genome sequence at a cost of about 1000 US $; many algorithms exist for the extraction of genome features, or "signals", including peaks (enriched regions), mutations, or gene expression (intensity of transcription activity). The missing gap is a system supporting data integration and exploration, giving a "biological meaning" to all the available information. The GeCo Project (Data-Driven Genomic Computing, ERC Advanced Grant currently undergoing the contract preparation) has the objective of revisiting genomic computing through the lens of basic data management. Starting from an abstract data model, we already developed a system that can be used to query processed ENCODE, TCGA, and Roadmap Epigenomics data; the system employs internally the Spark, Flink, and SciDB data engines, and prototypes can already be accessed from CINECA servers or be downloaded from PoliMi servers.

During the five-years of the ERC project, the system will be enriched with data analysis tools and environments and will be made increasingly efficient. Among the objectives of the project, the creation of an "open source" system available to biological and clinical research; while the GeCo project will provide public services which only use public data (anonymized and made available for secondary use, i.e., knowledge discovery), the use of the GeCo system within protected clinical contexts will enable personalized medicine, i.e. the adaptation of therapies to specific genetic features of patients. The most ambitious objective is the development, during the 5-year ERC project, of an "Internet for Genomics", i.e. a protocol for collecting data from Consortia and individual researchers, and a "Google for Genomics", supporting indexing and search over huge collections of genomic datasets.