European GeCo Project Aims to Replace Dominant Genomic Computing Methods

Aug 29, 2016 | [Justin Petrone](#)

Premium

NEW YORK (GenomeWeb) – The European Research Council has awarded researchers at Politecnico di Milano €2.5 million ($2.8 million) to carry out a new project that could upend standard genomic computing approaches.

Called "Data-Driven Genomic Computing," or GeCo, the effort will commence Sept. 1 and run through Aug. 2021. Investigators hope to develop tools that could enable integrated access to large repositories of sequence data and the building of an "Internet of genomic computing services" that provides Google-like processing.

"We wanted to reach a level of abstraction, a modeling language that would be more powerful than conventional approaches, and at the same time would be able to work over thousands or even millions of samples," lead investigator Stefano Ceri told GenomeWeb. "We are now proposing this modeling language as a new approach to genomics and we have five years to work with it," he said.

Ceri is a computer engineer and professor of database management at Polytechnic University of Milan, and is one of the inventors of WebML, a modeling language for designing web applications. He is also the founder of university spinoff WebRatio, which commercializes tools based on WebML.

Between 2013 and 2016, Ceri also headed GenData 2020, a consortium of Italian universities formed to build abstractions, models, and protocols for supporting a network of genomic data hosted on genome servers operated by the world's largest biological laboratories.

Ceri's experience with GenData 2020 and other projects has served as a catalyst for the new GeCo project which aims, according to the grant's abstract, to "rethink genomic computing through the lens of basic data management." According to the abstract, while next-generation sequencing technologies have led to the creation of large repositories of well-curated data, genomic computing "has not comparatively evolved."

"Bioinformatics has been driven by specific needs and distracted from a foundational approach; hundreds of methods solve individual problems, but miss the broad perspective," according to the abstract.

"Currently people use programs and scripts," Ceri said. "They use a system such as R and Python and so on, and essentially do a lot of low-level coding, which is difficult to manage and difficult to write," he said.

In response, Ceri aims to design a new genomic computing model based on the principle that just as data should express high-level properties of DNA regions and samples, high-level data management languages should express biological questions with simple, powerful, orthogonal abstractions, he said.

"Although this idea is very simple, putting it in action is far from trivial, as it requires a radical change of the dominant approach," said Ceri. He argued, though, that he aims to build a "progressive revolution of genomic computing" by integrating access to large repositories of sequence data and building a network of searchable genomic computing services.

His team aims to accomplish this with the design of a new query language with orthogonal, domain-specific abstractions for genomics. As laid out in the grant's abstract, the new query processing will trace its metadata analysis and computation steps, enabling descriptive statistics and high-level data analysis.

The developers will also achieve computational efficiency by using parallel computing clusters and public clouds, according to the abstract, and the resulting technology will be applicable to individual and federated repositories, allowing investigators from scientific consortia to query curated data via user-friendly search features.

The team's "most far-fetching vision," though, is the "Internet of Genomes," a protocol for collecting data from consortia and individual researchers, as well as a "Google for Genomics," supporting indexing and search over vast repositories of genomic datasets.

"It is much more powerful," Ceri said of the concepts behind GeCo. "Before, you had to essentially take your favorite programming language and write a lot of code, and now it can be done at a very high level."

Ceri intends to make the language developed within GeCo available as open source, enabling other developers to hone and refine it. He ruled out the idea of parlaying GeCo into commercial software.

"I made a company myself 16 years ago and I know what it means to build a company, but I think open source is a key for success," said Ceri of the decision to go open source.

He said that while much of the language has already been crafted, it will take about two years before an open-source system could take off. This adoption would then support other, "more complicated, visionary ideas," Ceri said, such as the Internet of Genomes.

Over the next five years, the envisioned system will also be enriched with data analysis tools and environments, and will be made increasingly efficient, Ceri noted. In addition to making the system openly available to biological and clinical researchers, using public data, Ceri said that the use of the GeCo system in "protected clinical contexts" could also enable personalized medicine, such as the adaptation of therapies to specific genetic features of patients.

Ceri himself is interested in using the GeCo system for personal research. "I am attracted by the possibility of doing research in biology," he said. "We are already attacking several problems in terms of how to evaluate situations in cells, such as comparing normal and tumor gene expression in relationship to the tridimensional structure of the genome," he said.

Filed Under   **Informatics**   **Research Funding**   **research grant**   **ERC**   **bioinformatics**

genomic datasets   **NGS**

✉   **Get Weekly Informatics Updates**

Related Articles

Dec 17, 2015

Transplant Genomics Grants Four Institutes Early Access to Kidney Transplant Test

Sep 10, 2015

U of Michigan Unveils $100M Data Science Investment; Includes Support for Biomedical, Precision Med Efforts

Dec 15, 2015

NIH Awards Iowa Researchers $3M for Whole-Genome Sequencing Study of Language Impairment

Nov 16, 2015

Correction: NIH Earmarks $48.2M for Progenitor Cell Translational Research Consortium

Dec 07, 2015

FDA Grants 510(k) Clearance, CLIA Waiver to Cepheid Flu/RSV Test

Aug 09, 2016

Current Methods for Counting HIV Proviruses May Be Inaccurate

We recommend

U Penn Cancer Center Initiative Builds Bridge Between Clinical and Genomic Data
GenomeWeb, 2002

BioDiscovery Turns to the Cloud to Support Collaborative Copy Number Variation Database
GenomeWeb, 2010

UK Firm Repositive Secures Funds to Develop First Commercial Software Solution
GenomeWeb, 2015

Community-Bred GO Terms Aimed to Help Scientists Manage Microbe-Host Interactions
GenomeWeb, 2009

Challenge of Developing IBMs Dr. Watson Not Technical, as Much as Cultural, Researchers Say
GenomeWeb, 2011

Stanford-Led Team Awarded $6M to Help Northrop Grumman Expand Immunology Repository for NIAID
GenomeWeb, 2013

Powered by