NEWS RELEASE
UC San Diego
University Archives. RSS 6020

SDSC, Calit2 Awarded $1.4 Million NSF Grant for New Bioinformatics Tools

Project Aimed at Harnessing Next-Generation DNA Sequencing and Analysis

October 18, 2011

Jan Zverina

Researchers at the San Diego Supercomputer Center (SDSC) and the California Institute for Telecommunications and Information Technology (Calit2) at the University of California, San Diego, have been awarded a three-year, $1.4 million grant from the National Science Foundation (NSF) to create a Kepler Scientific Workflow System module. Researchers will develop new tools to help manage ever-growing data sets used in next-generation DNA sequencing.

"Next-generation DNA sequencing is now creating such a large amount of sequence data that it is overwhelming current computational tools and resources," said Ilkay Altintas, director of the Scientific Workflow Automation Technologies (SWAT) Lab within SDSC's Cyberinfrastructure Research, Education And Development (CI-RED) group, and Principal Investigator for the project. "New computational techniques and efficient implementation mechanisms for this data-intensive workload are needed to enable rapid analysis of these next-generation sequence data."

The project receiving the NSF award is called *Advances in Biological Informatics Development: bioKepler: A Comprehensive Bioinformatics Scientific Workflow Module for Distributed Analysis of Large-Scale Biological Data*. Bioinformatics refers to a field of science that combines biology, information technology, computers and statistical techniques to create research-driven solutions such as customized medications and treatments to help prevent disease, three-dimensional models of genomes and proteins, and advanced agricultural technologies.

"The enormous growth in data-intensive research means that as these data sets get larger, moving data over the network becomes more complicated, error-prone and costly to maintain," said Altintas, who also serves as SDSC's deputy coordinator for research.

The bioKepler project is motivated by the following three challenges that remain unsolved:

How can large-scale sequencing data be analyzed systematically in a way that incorporates and enables reuse of best practices by the scientific community?
How can such analysis be easily configured or programmed by end-users with various skill levels to formulate actual bioinformatics workflows?
How can such workflows be executed in computing resources available to scientists in an efficient and intuitive manner?

To create such an environment, the bioKepler project will create scientific workflow components to develop an array of bioinformatics tools using distributed execution techniques. Once customized, these components will be used on multiple distributed platforms, including various cloud and grid computing platforms. The tools will be selected to meet the diverse needs of researchers, and organized into eight groups covering most aspects of bioinformatics applications: sequence database searches; mapping; sequence assembly; gene prediction; clustering; multiple sequence alignment, phylogeny and taxonomy; protein annotation; and other miscellaneous utilities such as data format transformation and parsing.

Training Next-Generation Scientists "These tools will be applicable to a wide range of bioinformatics and computational biology problems," said Altintas, noting that "a key part of this project will also focus on education and outreach efforts, underscoring the importance of training next-generation scientists, as well as the need to narrow the gap between bioinformatics and technology."

All the resources, materials, and open-source software products produced by the bioKepler project will be integrated with Calit2's Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA), a data repository and a bioinformatics resource for metagenomic analysis.

"The Kepler workflow system has already been used comprehensively in the CAMERA project," said project co-investigator Weizhong Li, a research scientist at Calit2 and the Center for Research in Biological Systems (CRBS), and Bioinformatics group leader for CAMERA. "With the proposed developments in bioKepler, the CAMERA project and its large user communities will benefit from a larger set of next generation sequence analysis tools with much better scalability and flexibility. Other projects that heavily rely on next-generation sequencing, such as various microbiome projects, can also take advantage of the bioKepler software."

Moreover, bioKepler will be packaged to be installed on diverse, distributed execution environments (e.g., as a Web service and as virtual machines tuned for various Grid and Cloud systems), which in turn will enable deployment of bioKepler on public and private clusters and clouds.

In addition to Altintas and Li, the bioKepler research team includes Eric E. Allen, assistant professor of marine biology at the Scripps Oceanography Institute (SIO); Jianwu Wang, project scientist with SWAT; Daniel Crawl, workflow specialist with SWAT; and Shulei Sun and Sitao Wu, bioinformaticians at CRBS.

The bioKepler project is funded by NSF DBI-1062565 under the CI Reuse and Advances in Bioinformatics programs.

Media Contacts: Jan Zverina, SDSC Communications, 858 534-5111 or jzverina@sdsc.edu Warren R. Froelich, SDSC Communications, 858 822-3622 or froelich@sdsc.edu Doug Ramsey, Calit2 Communications, (858) 822-5825 or dramsey@ucsd.edu