Microsoft Research

BioHPC Helps Biologists Tap Latest Computational Tools

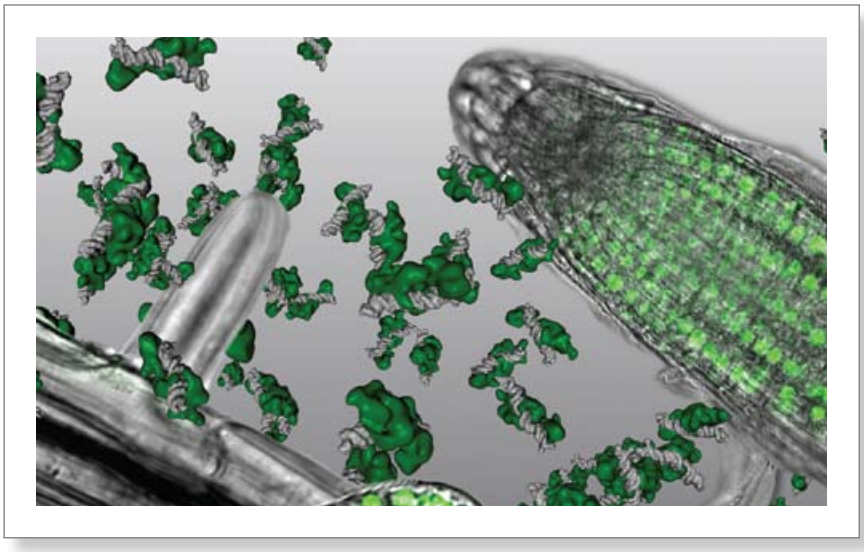
High-performance computing (HPC) is revolutionizing the life sciences by enabling biologists to compile and analyze data on a scale unimaginable a few decades ago. Many researchers, however, lack expertise in computational methods or access to the latest technology. Cornell University's Computational Biology Service Unit (CBSU) and Microsoft are collaborating on a project that is bringing the power of HPC to thousands of researchers worldwide.

Modern biologists are making huge strides toward solving many of life's great mysteries. From mapping the genomes of numerous species to predicting the structure of proteins, life scientists are gaining vast new understanding of how organisms function and evolve.

In large part, this era of discovery has been made possible by rapid advances in computer hardware and software. And perhaps nothing has played a bigger role than high-performance computing—the use of supercomputers and clusters of linked computers to tackle advanced computation problems.

But as the pace of discovery increases and new information multiplies at exponential rates, biologists face ever-greater challenges in managing and making sense of the data. For researchers to keep up, increased access to high-performance computing resources has become essential.

To help provide such access, the Computational Biology Service Unit at Cornell University in Ithaca, New York, is collaborating with Microsoft on a project—called BioHPC—that is giving thousands of researchers access to an array of powerful but easy-to-use bioinformatics tools.



A composite image showing micrographs of plant roots with an accumulation of FHY3 protein (green dots) in the cells' nuclei. The background of the image shows 3-D models of FHY3 protein (in green) bound to DNA (shown in gray). The image was developed as part of a groundbreaking study of how plant responses to light are regulated. Courtesy of Chris Pelkie, Daniel Ripoll and Rongcheng Lin.

Fast Facts

Project Principals:

Jaroslaw Pillardy, Director of the Computational Biology Service Unit (CBSU), Cornell University

Robert Bukowski, Senior Research Associate, CBSU, Cornell University

Web Site:

http://www.biohpc.org

Profile:

The Computational Biology Service Unit at Cornell University is collaborating with Microsoft on BioHPC, a project that is giving thousands of researchers easy access to high-performance computing resources. BioHPC supports nearly three dozen applications covering the major areas of computational biology, including population genetics, sequence analysis and data mining, and protein structure prediction.

Microsoft External Research

The Microsoft External Research Division within Microsoft Research partners with academia, government and industry to advance computer science, education and scientific research aimed at helping address some of the world's most urgent and significant social and technological challenges. Along with investing cash, software, hardware and research expertise to enable ground-breaking projects worldwide, Microsoft External Research is committed to providing the advanced technologies and services needed to support every stage of the research process. Efforts are focused in four research areas—including Health and Wellbeing, which explores technologies that advance healthcare and help people make better choices about their health.

Microsoft External Research
http://research.microsoft.com/en-us/collaboration/

In 2006, Cornell was chosen as one of 10 universities worldwide to establish a Microsoft Institute for High Performance Computing, with the goal of providing bioinformatics and research support for the university community. More recently, Microsoft External Research funded a two-year CBSU project to expand BioHPC's capabilities and make it available to more researchers.

CBSU Director Jaroslaw Pillardy says access to high-performance computing can be prohibitively expensive. He also argues that biologists shouldn't have to be experts in computer science or know the latest programming techniques. The goal of BioHPC is to help remove these barriers, Pillardy says.

"Most biologists are focused on the experimental aspects of their research and are not familiar with HPC," Pillardy says. "They often know the algorithms and programs required for analysis of their data, but they have no expertise in how to use them efficiently in HPC environments."

Researchers affiliated with Cornell have broad access to BioHPC as registered users. BioHPC is also available, at no cost, as downloadable open-source software to researchers around the world. Four universities are currently working with CBSU to set up local BioHPC installations. In addition, researchers anywhere can sign up as guest users and, as resources permit, submit data-processing jobs directly to the BioHPC installation at Cornell.

"Most biologists ... often know the algorithms and programs required for analysis of their data, but they have no expertise in how to use them efficiently in HPC environments."

Jaroslaw Pillardy, director, Computational Biology Service Unit, Cornell University

BioHPC's popularity is growing rapidly. Since it was deployed in 2003, BioHPC has processed about 80,000 computationally intensive data-processing jobs submitted by more than 7,200 researchers from 80 countries.

Two major plant genome projects are using BioHPC for data analysis. One is the SOL Genomics Network (SGN), an international database containing genomic information for the families *Solanaceae* (which includes tomato, potato, eggplant and pepper) and *Rubiaceae* (coffee). The other is the Arabidopsis Information Resource (TAIR) project, which maintains a database of genetic and molecular biology data for *Arabidopsis thaliana*, a small flowering plant used widely as a model organism in botany.

Meanwhile, a research team at Cornell's Boyce Thompson Institute has been using BioHPC for a groundbreaking study of how plant responses to light are regulated. The research, funded by the National Science Foundation, recently revealed that plants—while still in the dark as seeds and roots—produce proteins that prepare the plant to respond to light.

The BioHPC system architecture consists of an easy-to-use Web-based interface written in ASP.NET, Microsoft® SQL Server® 2005 and compute clusters running Microsoft Windows Server® 2003 and Windows Compute Cluster Server 2003. Under the new project, Pillardy's team plans to fully integrate BioHPC with Microsoft HPC Server 2008, the successor software to Compute Cluster Server 2003.

BioHPC supports nearly three dozen applications covering all major areas of computational biology, including population genetics, sequence analysis and data mining, and protein structure prediction. Most of BioHPC's current software was installed at the request of users. But as experimental techniques and research focuses change, computational biology applications evolve. And, Pillardy notes, there are many potentially useful software tools that are unknown to most biology researchers.

The new support from Microsoft will enable CBSU to continually upgrade BioHPC with new bioinformatics applications, adapt other applications to the BioHPC platform and continue supporting local installations of the program. Pillardy estimates that CBSU will add at least five new applications and oversee about three new local installations annually.

Another aspect of the collaboration with Microsoft will be to integrate various Microsoft Research computational biology tools into BioHPC. Microsoft Research collaborates with academic communities to create open tools and services based on Microsoft platforms and productivity software.

Through CBSU's affiliation with the Microsoft Institute for High Performance Computing program, BioHPC users have free access to a 256-processor cluster at Microsoft's headquarters in Redmond, Washington. CBSU plans to seek additional support from the National Science Foundation and/or the National Institutes of Health to expand BioHPC's free services to the scientific community.

Pillardy says researchers without high-performance computing resources or expertise face significant obstacles. He describes a typical scenario in which a researcher finds the application he needs, spends a long time learning how to use it and then has to wait weeks while running a single job. "When you go to the lab, you see a desktop PC off in the corner with a sign taped on the screen that says, 'Please do not touch this computer for the next month,'" Pillardy says.

Pillardy, who was initially trained in quantum chemistry before switching to protein structure prediction, also says he has too often seen researchers and biology students wasting valuable time and energy learning computer science. "They should be focusing on their research, and with BioHPC they are able to do that," Pillardy says.

Part No. 098-111134