NERSC
n e w s

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Office of
Science

U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

JANUARY 2007

Life Puzzle

PIECING TOGETHER MILLION PIECES OF GENETIC MATERIALS REQUIRES HEFTY COMPUTING POWER

The DOE Joint Genome Institute has received 1.5 million computing hours and 900,000 storage resource units at NERSC this year for processing and managing its fast-growing trove of genomic data emerging from the characterization of organisms involved in bioenergy, carbon cycling, and bioremediation.

The project marks a departure from how JGI has carried out its genome assembly and annotation work in the past. The institute has relied primarily on its own 400-processor computing cluster to assemble and annotate genomes after the genomes have been sequenced. But the on-site computing facility is running at capacity, while the amount of data generated have grown quickly in recent years thanks to new sequencing technologies, said Darren Platt, head of the Informatics Department at JGI.

"In the long run, we know we will continue to grow in CPU consumption. We

SPOTLIGHT

Two Researchers Join Analytics

Janet Jacobsen is returning to the Berkeley Lab after working at the Engineering Systems Research Center (ESRC) and the Institute for Quantitative Biomedical Research (QB3) at UC Berkeley, and before that in the Atmospheric

Janet Jacobsen

Next Generation Arrives

NERSC INSTALLS FIRST UNITS OF NEW CRAY XT4

The first installment of the Cray XT4 arrived at NERSC earlier this month, adding to the center's lineup of high-performance computers.

Although Cray delivered a small test system last fall, its shipment of 36 cabinets on January 16 marked the start of the installation of the full system, which is expected to go into production this summer.

Named "Franklin" after Benjamin Franklin, America's first scientist, the Cray XT4 will consist of more than 19,000 processor cores when fully installed. It will deliver sustained performance of at least 16 trillion calculations per second, with a theoretical peak speed of more than 100 teraflop/s. The new system was powered up on Jan. 17, 2007—the 217th anniversary of Franklin's birth.

"Franklin, the world's largest XT4 system, represents more than 10 times the sustained computing power of any other NERSC system," said Bill Kramer, NERSC's general manager. "We are extremely pleased to be able to make such a significant addition to the resources our computational scientists use."

Assembled and shipped from Chippewa Falls, Wisconsin, the cabinets feature a distinctive blue and gold color scheme reflecting their new home in Berkeley.

Learn more about Franklin here: http://www.lbl.gov/CS/Archive/news081006.html.

Start to finish: workers installed "Franklin" the new Cray XT4.

Packing DNA

RESEARCH PRODUCED IMAGES SHOWING HOW A VIRUS CONTROLS ITS GENETIC MATERIAL

After running calculations on NERSC resources for the past year, research by Luis R. Comolli and fellow scientists has yielded the three-dimensional architecture of the packaged dsDNA genome of bacteriophage Phi29. For the first time, the work rendered 3D images accounting for more than 95 percent of a bacterio-phage's genome.

The work combined cryo-EM and computer simulations while relying on independently obtained force spectroscopy measurements. Using Monte Carlo simulations, Comolli and fellow researchers tried to predict the conformation of

NERSC news

Puzzle *continued from page 1*

are interested in taking advantage of remote computing," Platt said. "The NERSC project is an ideal experiment for us to wean ourselves off doing everything locally."

Although the JGI researchers have used small allocations at other supercomputer centers before, the NERSC allocation is the largest to date and provides an opportunity for testing new ideas.

Genome assembly is akin to putting together a jigsaw puzzle. Because each sequencing process could produce 10,000 to 10 million pieces of information, identifying and putting them in order—and giving them descriptive titles after comparing them to other genome sequencing results—is a computationally intensive process.

Last year, the institute brought on line a second sequencing system, made by 454 Life Sciences. In early January, it added a third sequencing system by Solexa, which has the ability to produce one terabyte of raw data every three days.

JGI has been storing raw data at NERSC's High-Performance Storage System (HPSS) for years. The partnership formed for a similar reason: JGI's

JGI is home to over 100 DNA sequencing instruments. As the latest generation sequencers come on line, the amount of data generated daily will rise dramatically.

Photo credit: Joint Genome Institute

production capacity had produced more data than what its storage system could handle.

Staff from NERSC's Mass Storage Group and the JGI have since worked together to address two key issues facing the genome researchers. They turned HPSS into the archive for the JGI data, replacing the less-reliable local tape operation and freeing up disk space at the JGI for more immediate production needs. Click http://www.nersc.gov/news/nerscnews/NERSCNews_2005_10.pdf to learn about this collaboration.

"We have had a wonderful experience with NERSC, using its storage capability.

It's been a reliable and good partner," Platt said.

Now the institute needs additional computing power to process the raw data as well. The institute plans to complete 35 terabytes of assembly and annotation work this year, some of which will take place at NERSC.

Platt and his team also are interested in using NERSC resources to develop techniques that improve data access and handling for scientists, such as providing better correlations between the trace files and the sequence in which they appear.

The new techniques also will allow for reducing the number of files that researchers must retrieve in order to carry out close studies of a section of an assembled sequence.

In addition to Platt, other researchers in the project are Jonathan Carter, leader of the User Services Group at NERSC; Susan Lucas, head of JGI's Sequencing Department; and Dan Rokhsar, head of JGI's Computational Genomics Program.

Established in 1997, JGI brings together resources at the Berkeley Lawrence National Laboratory, Lawrence Livermore National Laboratory, Los Alamos National Laboratory, Pacific Northwest National Laboratory, Oak Ridge National Laboratory and the Stanford Human Genome Center. More information about JGI can be found at http://www.jgi.doe. gov/index.html.

SPOTLIGHT

continued from page 1

Sciences Division at Lawrence Livermore National Laboratory.

Previously at Berkeley Lab, Jacobsen was a staff scientist in the Earth Sciences Division, where she analyzed field data, developed finite element mesh generators for fracture data and numerical simulators for reactive chemical transport, and developed visualizations involving many types of field data. Jacobsen received both an M.A. in mathematics and a B.A. in applied mathematics and statistics from UC Berkeley.

Jacobsen's primary role on the NERSC Analytics Team will be to work with users to create data visualizations, but she also will be involved in helping users to simplify the process of getting their simulation output or experimental data into data visualiza-

tion software at NERSC and in working with other Analytics Team members to develop workflow and analysis tools.

Gunther Weber has recently joined NERSC's Analytics Team. Weber comes to the Lab from UC Davis, where he was an assistant project scientist at the Institute for Data Analysis and Visualization. At IDAV, Gunther's research focused around visualization and computer graphics. Weber

Gunther Weber

worked on visualization of three-dimensional gene expression data (with researchers in the Lab's Genomics and Life Sciences divisions), topology-based exploration of scalar data, and visualization of brain imaging data and experimental earthquake data.

Weber has worked at NERSC before, as a student employee and then a guest student assistant at NERSC between 2000 and 2003. He has also been a guest in the Lab's Life Sciences Division since 2003. He received his bachelor's, master's and doctorate degrees in computer science from the University of Kaiserslautern in Germany.

His research interests include scientific visualization, data analysis, bioinformatics, topological data analysis methods, hierarchical data representation methods and visualization of adaptive mesh refinement data. On the NERSC Analytics Team, Weber will help users with challenging visualization problems and help them on using data visualization software at NERSC for their simulation output and experimental data.

NERSC news

DNA *continued from page 1*

Bacillus subtilis' DNA in the interior of the viral capsid and the forces required to achieve this structure. Bacteriophage is a virus that attacks bacteria.

Understanding how bacteriophages can precisely control their DNAs can lead to many real-life applications, such as developing better medicines. Last August, the U.S. Food and Drug Administration approved the use of six bacteriophages in a mixture to kill Lysteria monocytogenes during meat preparations at processing plants. Lysteria monocytognes can cause meningitis, encephalitis and cervical infections.

Comolli has presented his work several times in the past year, including the Gordon Conference in Italy and the American Institute for Chemical Engineering conference last November in San Francisco. In March, he will present a poster at the 2007 Biophysical Society meeting in Baltimore.

"The state-of-the-art resources and friendly support provided by NERSC allowed us to make timely progress," Comolli said. He has received an 82,000-hour allocation at NERSC to further his work this year.

Bacteriophage mainly consists of a genome, a capsid—a protein outer shell covering a genetic material such as DNA—and the packaging machinery. In the process of replicating itself inside the host bacterium, an ATP-driven protein motor is put to work to fill each capsid with the DNA and must work against the resistance coming from packing a long, charged polymer chain into a small cavity. Bacteriophage Phi29 is roughly a cylinder ~54 nm long with a radius of ~22 nm, while the length of its DNA genome is approximately 6,500 nm.

While much is known about the genetic and physiological makeup of dsDNA bacteriophages, exactly how the genetic material is assembled and kept inside each capsid remains a great mystery. DNA packaging is a significant step in viral assembly and gene delivery, and determining the packaging structure of DNA in virus particles at different stages of the packaging process remains a formidable challenge.

Simulations created by Comolli and his team showed a good qualitative agreement with cryo-EM results during the
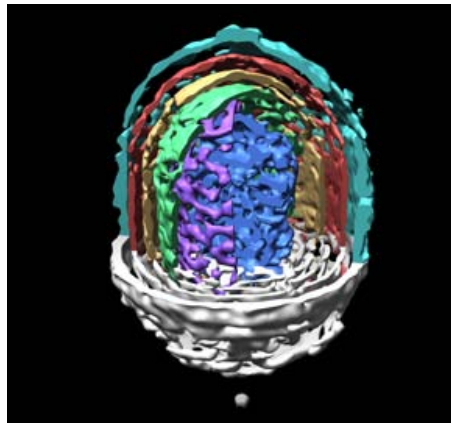
Three-dimensional reconstruction of bacteriophage Phi29 fully packaged DNA genome. We can observe, for the first time, a layered architecture that accounts for more than 95 percent of the total genome.

packaging of the first 80 percent of the total genome. The predicted forces are consistent with previous single-molecule measurements of the packaging forces.

The team of researchers includes Andrew Spakowitz, an assistant professor of chemical engineering at Stanford University and Cristina Siegerist, a researcher on NERSC's Analytics Team.

"The computational resources provided by NERSC greatly facilitate our collaborative research," Spakowitz said. "One great benefit of having a major computational resource like NERSC is its adaptability to a wide range of applications, which greatly aids our interdisciplinary research project."

The project's next phase will involve modeling the compaction step, which occurs during the last step of the packaging process, between the packaging of 80 percent and 100 percent of the total genome.

Using NERSC's Jacquard supercomputer, Comolli's team was able to render powerful images that illustrate details of the genetic material inside a capsid, as shown in figure 1. Cryo-EM imaging provided direct measurements of physical parameters, such as inter-strand distances, local curvatures and local- and long-range degree of order. Optical tweezers measured the pressures gener-
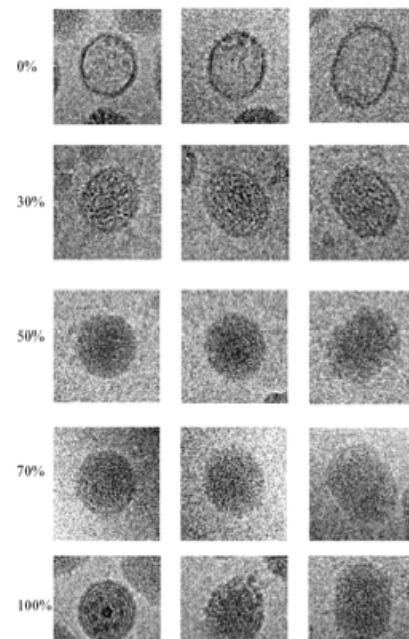
Cryo-EM images of bacteriophage particles with precise amounts of packaged DNA. The completely empty capsids shown in row zero percent are allowed to package only ~30 percent, ~50 percent or ~70 percent of the total genome. The DNA fully occupies the whole volume of the capsids, adopting a layered structure only towards the end of the packaging process. Compare particles with ~70 percent and particles with 100 percent of the genome packaged.
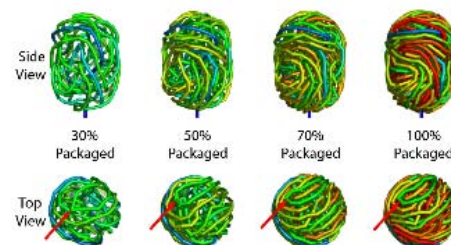
Snapshots of a Monte Carlo simulation of DNA packaging in _29. Comolli's theoretical model consists of a semiflexible polyelectrolyte within a confinement (cylindrical with hemispherical caps). The conformation fluctuations of the chain reflect the inherent Brownian noise, capturing the equilibrium behavior (as predicted from statistical mechanics). These images show typical snapshots throughout a packaging simulation. The chain is colored blue at the end entering the confinement and red at the free end within the capsid at full packaging.

3

NERSC news

INCITE Grants Spur Innovations
NATIONAL INITIATIVE ALLOCATES NEARLY 9 MILLION HOURS AT NERSC

The U.S. Department of Energy announced early this month that it has allocated a large amount of supercomputing resources at NERSC as part of an initiative to accelerate scientific research and promote innovations in public institutions and private industry.

Supported by DOE's Office of Science, seven research projects will receive nearly nine million processor hours at NERSC in 2007. The projects range from studying the behavior of a supernova to designing more energy-efficient cars.

The allocations are part of a fast-growing program called Innovative and Novel Computational Impact on Theory and Experiment (INCITE), which launched in 2003. INCITE selects projects that not only require large-scale and intensive use of computers but also promise to deliver a significant advance in science and engineering. For 2007, the program awarded 95 million processor hours for 45 projects overall, a five-fold increase in computing time awarded from 2006.

"We believe there is a great opportunity for significant design analysis productivity improvements through the use of HPC. Due to the extremely large capital cost required, this has been traditionally difficult to justify, and thus never attempted," said Paul Bemis from Fluent Inc., an engineering software firm based in Lebanon, New Hampshire and an INCITE award recipient working with General Motors. "The INCITE award provides the opportunity to realize the potential productivity improvements and allows a factual and more quantitative analysis of the HPC benefits."

NERSC, where Bemis will carry out his research, is one of the four supercomputer centers providing resources for these INCITE projects. As the flagship facility for the Office of Science, NERSC provided the only computing resources available during the first two years of the program. Here are short descriptions of the seven projects awarded computing time at NERSC:

- The project by Fluent Inc., in partnership with General Motors, will use its computational fluid dynamics software to perform intensive calculations for designing different parts of a car.

- Gilbert Compo from the University of Colorado, along with researchers from the National Oceanic and Atmospheric Administration, will use a new technique to create a more representative dataset for validating certain climate models. The Intergovernmental Panel on Climate Change is using the climate models to make 21st-century projections in its fourth assessment report, due out later this year.

- Hong Im of the University of Michigan will lead the work on developing three-dimensional simulations of turbulent nonpremixed flames in the presence of a mean flow strain and fine water droplets.

- Warren Mori from the University of California at Los Angeles will lead research on using computer simulations to answer questions about plasma-based particle accelerators that currently cannot be answered through experiments. The project will contribute to the development of better acceleration methods.

Plugged In

Andrew Karsner, DOE's Assistant Secretary for Energy Efficiency and Renewable Energy, toured the NERSC facility during his trip to the Bay Area last month to learn about how DOE and the tech industry can work together to develop energy-efficient computers and data centers. NERSC's Associate General Manager Howard Walter (center, with a hard hat in hand), described the center's supercomputing resources to Karsner while Horst Simon (far left), NERSC Director and Associate Lab Director of Computing Sciences, looked on.

NERSC news

NERSC users published more than 1,400 research papers

Researchers reported publishing 1,434 refereed papers based on their NERSC computing work in the year leading up to their ERCAP allocation requests for 2007.

The publications, which consist of journal articles and conference papers, reflect a wide range of projects that rely on NERSC resources. The subjects include supernovae, dark energy, global warming, particle accelerator designs, turbulent combustion and nano dots.

"The number and quality of publications demonstrate once again that NERSC is a premier facility for computational research," said NERSC Division Director Horst Simon. "NERSC consistently provides strong system and software support and user services to its users."

Check out the complete list of publications here: http://www.nersc.gov/news/reports/ERCAPpubs06.php.

Innovations *continued from page 4*

- Chuang Ren at the University of Rochester, along with Mori from UCLA, will create large-scale particle-in-cell simulations to test fast ignition, one of the most promising new schemes to improve the viability of inertial confinement fusion as a practical energy source.

- Lawrence Pratt at Fisk University plans to unravel the mysteries of several lithium compounds that are among the best reagents for forming carbon-carbon bonds in organic synthesis, which can lead to the development of powerful medicines. The project will look at the structures and reactions of these compounds.

- Don Lamb, a researcher at the University of Chicago's ASC/Alliance Flash Center, will create three-dimensional simulations of a stellar explosion known as a Type Ia supernova. Understanding the mechanism of such explosion has been done in two-dimensional studies in the past.

"The INCITE award will make it possible for us to do huge computer simulations, whose goal is to pin down the explosion mechanism for Type Ia supernovae," said Lamb. "A better understanding of these supernovae will enable astronomers to learn more about the properties of dark energy."

Aside from NERSC, supercomputer centers where other INCITE projects will be carried out are located at Oak Ridge National Laboratory in Tennessee, Argonne National Laboratory in Illinois and Pacific Northwest National Laboratory in Washington. Industrial firms receiving the 2007 INCITE awards also include Corning Inc., General Atomics, DreamWorks, Pratt and Whitney, the Boeing Co. and Procter and Gamble. More information about INCITE can be found at http://www.sc.doe.gov/ascr/INCITE/index.html.

DNA *continued from page 3*

ated inside the capsid by the DNA. These experimentally derived values can be used to evaluate predictions and models based on the current theory.

The scientists also used Monte Carlo simulations to predict the structure and behavior of the DNA during the initial stages of the packaging process, shown in figures 2 and 3. The result: a good agreement between their simulations and the cryo-EM experimental data, and the force measurements obtained with optical tweezers. However, further progress is necessary before a complete understanding of the last step of DNA compaction is reached.

So far, the research has led Comolli to determine that "there is no unique, deterministic DNA packaging pathway," he said. "Rather, the fully packaged DNA fits into an organized architecture that is filled by different DNA trajectories in different particles, quite possibly with a considerable number of defects."

WHAT IS NERSC NEWS?

NERSC News publishes every other month and highlights the cutting-edge research performed using the National Energy Research Scientific Computing Center, the flagship supercomputer facility for DOE's Office of Science. Find previous NERSC News articles at http://www.nersc.gov/news/nerscnews.