

IMG/M Spearheads Analysis of Human Microbial Communities

"We live in a microbial world, there are millions of organisms in one drop of water and even more in soil. Life on our planet cannot be sustained without the microbes, and the success of metagenomics will not only help us better understand human health, but may also help us address a variety of environmental challenges," says Nikos Kyrpides, Head of the Genome Biology Program at the Department of Energy's Joint Genome Institute (JGI) in Walnut Creek, Calif.

A metagenome is the collective genome, or DNA, of a microbial community. Initially released in 2006, the Integrated Microbial Genomics with Microbiome Samples (IMG/M) data management system has played a central role in helping scientists understand metagenomes in a variety of natural environments.

Now a new grant from the National Institute of Health (NIH) will expand the system's capabilities to include metagenomic data from humans, to give scientists valuable insights into how microbial communities affect human health.

"IMG/M was developed through a close collaboration of software engineers, computer scientists and biologists," says Victor Markowitz, head of the Berkeley Lab's Biological Data Management and Technology Center (BDMTC), and the technical lead for IMG/M. The system was developed in conjunction with JGI's Genome Biology Program.

Supporting the Human Microbiome Project

Within the body of a healthy adult, microbial cells are estimated to outnumber human cells by a factor of ten to one. These tiny organisms cover every surface and cavity of the human body, forming complex communities that help digest



Humans Hosting Metagenomes: The Human Microbiome Project will collect metagenome samples from individuals with a variety of health conditions. They will focus on microbial communities from these five areas of the human body.

food, break down toxins and fight off diseases.

"When the average person hears the word 'microbe,' they think of a disease or a disaster. However, the vast majority of microbes are our friends," says Kyrpides. "In fact entire microbial communities work in harmony with us to carry out essential functions, such as digestion in the human gut. When these communities are disturbed, people may get sick or catch infections. Microbes have won every major battle on our planet except that of the good impressions."

To understand how microbes affect human health and how they cause various diseases, researchers involved in the Human Microbiome Project will collect metagenome samples from different parts

continued on page 4

CRD Method for Understanding Nanostructures Is a Gordon Bell Finalist

The key to energy independence from petroleum, coal, and other fossil fuels could be tiny materials called nanostructures. At approximately 100,000 times finer than human hair, these structures may be microscopic individually, but in groups of thousands, they could revolutionize solar cell design by providing a cost-efficient resource for harvesting solar-energy.

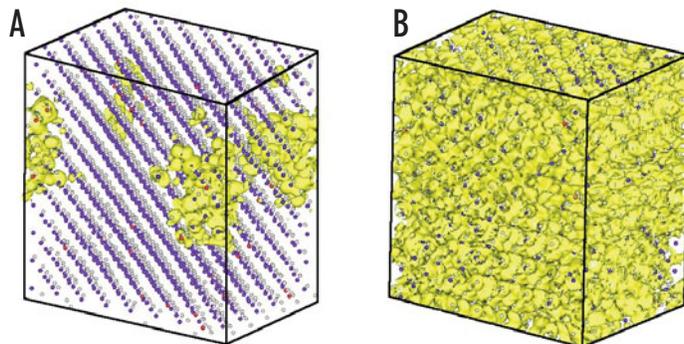
To theoretically understand and simulate the energy harnessing potential of nanostructures, a team of researchers in the Berkeley Lab's Computational Research Division (CRD) developed the Linear Scaling Three Dimensional Fragment (LS3DF) method. The computer algorithms in this method use a "divide-and-conquer" technique to efficiently gain insights into how nanostructures function in systems with 10,000 or more atoms.

"By incorporating the correct chemical formulas into efficient computer programs, scientists can learn a lot about the structures and properties of molecules and solids I like to think of computers as chemistry's 'third leg.' In most cases, computer simulations complement information obtained by chemical experiments, but in some cases it can predict unobserved phenomena," says Dr. Lin-Wang Wang, a CRD computational material scientist and leader of the LS3DF project.

The developers of LS3DF are finalists in the Association for Computing Machinery's (ACM) Gordon Bell Prize competition, which recognizes outstanding achievement in high-performance computing applications. The winners will be announced on November 20, 2008 at the SC08 Conference in Austin, Tex.

continued on page 2

Nanostructures *continued from page 1*



Images (A) and (B) show scientific results from the LS3DF method's test-run on Franklin, which included calculations for a 3,500-atom $\text{ZnTe}_{0.97}\text{Zn}_{0.03}$ alloy. The entire run took one hour on 17,000 processors. Isosurface plots (yellow) show the electron wavefunction squares for the bottom of the conduction band (A) and the top of the oxygen-induced band (B). The small gray dots are Zn atoms, the blue dots are Te atoms, and the red dots are oxygen atoms.

According to Wang, traditional methods for calculating the energy potential of nanostructure systems containing 10,000 or more atoms can be very time consuming and resource intensive. Because these techniques calculate the entire structure as a whole system, the compute time, disk space and memory required to determine the energy potential of these structures grows to the third power of the system's size. That means calculating a 1000-atom system will be a thousand times more expensive than calculating a 100-atom system.

He notes that LS3DF offers a more efficient way for calculating energy potential because it is based on the observation that the total energy of a large nanostructure system can be broken down into small pieces, and each piece can be calculated separately. Wang refers to this technique as "divide-and-conquer."

The total energy of the large system has two components: electrostatic energy and quantum mechanical energy. To determine the structure's total quantum mechanical energy, the LS3DF method breaks the entire structure into small fragments, applies its algorithm to each individual piece, and then combines the results of the pieces to get a total for the whole system. Scientists say that under the traditional density functional theory methods, the quantum mechanical energy calculation typically requires the most compute time and resources. By breaking up the big problem into small pieces, LS3DF can solve it a lot more quickly and efficiently, making the computational cost proportional to the total number of the atoms in the system.

Meanwhile, the electrostatic energy of large-scale nanostructure systems is not

as resource intensive to solve. Scientists calculate this classical energy by looking at the whole system, which may contain tens of thousands of atoms. This problem is solved separately from the quantum mechanical energy. In the end, both energy results are combined to get the structure's total energy potential.

When team members tested the LS3DF method on supercomputers at the Department of Energy's (DOE) National Energy Research Scientific Computing Center (NERSC) in Oakland, Calif, National Center for Computational Sciences (NCCS) at Oak Ridge National Laboratory in Oak Ridge, Tenn., and Argonne Leadership Computing Facility in Argonne, Ill., they found that the LS3DF method can work hundreds to thousands of times faster than traditional density functional theory calculations for systems with tens of thousands of atoms, and yielded essentially the same results.

"The core of LS3DF is a novel patching scheme that cancels out the artificial boundary effects caused by dividing the system into smaller fragments," says Wang. "This cancellation is what gets us the same results as the traditional method."

Because LS3DF scales almost perfectly with the number of compute cores, it is the first electronic structure code that runs efficiently on computer systems with tens to hundreds of thousands of cores. On 17,280 cores of the dual-core Cray XT4 (Franklin) at NERSC, LS3DF achieved 32 Tflop/s or 32% of the peak floating-point performance of the machine. On 30,720 cores of the quad-core Cray XT4 (Jaguar) at NCCS, LS3DF reached 60 Tflop/s or 23% of the theoretical peak. In a later run on the IBM BlueGene/P system

(Intrepid) at Argonne, the code achieved 107.5 Tflop/s on 131,072 cores, or 24.2% of peak.

Energy Independence from Fossil Fuels

Scientists agree that a fundamental understanding of nanostructure behaviors and properties could provide a solution for curbing our dependence on petroleum, coal, and other fossil fuels.

According to Wang, nanostructure systems are cheaper to produce than the crystal thin films used in current solar cell designs, and offer the same material purity. In addition, nanostructures are extremely versatile. They can act as electrodes to carry electric currents, or active materials that absorb sunlight and convert it to electricity.

One type of nanostructure, called quantum dots, actually changes color with size. Scientists say this color, or band gap, affects the type of light that the structure absorbs, which will be very useful for designing solar cells.

"We still don't quite understand how the electron moves around in a nanostructure, and how such properties depend on the size, geometry, composition, and surface passivations... Understanding such dependence will allow us to design nanostructures for desired applications, and LS3DF can help us to understand and predict these properties with computers," says Wang.

Other authors on the Gordon Bell paper include the Berkeley Lab's David H. Bailey, Zhao, Byoungchak, Zhengji Lee, Juan Meza, Hongzhang Shan, and Erich Strohmaier.

Meet the Fellows

Kamesh Madduri, 2008 Alvarez Fellow

Kamesh Madduri's interest in computers ignited at age 10, when the first wave of PCs arrived in India and gave him a new medium to play games on. Eventually this fascination grew to include programming, computer science, and finally supercomputers.

Now, Madduri's lifelong hobby has brought him to the Berkeley Lab in California, as a prestigious Luis W. Alvarez Fellow in Computing Sciences. He will spend the next few years working with Arie Shoshani, John Wu, and the Scientific Data Management group, developing algorithms that will allow researchers to efficiently sift through large datasets for the bits of information that are relevant to their work.

"Great technological advances allow scientists to do a lot more, which means that the amount of scientific data that needs to be managed is constantly growing," says Madduri. "It is really exciting to be working on large-scale applications and solving cutting-edge problems that affect the larger science community."

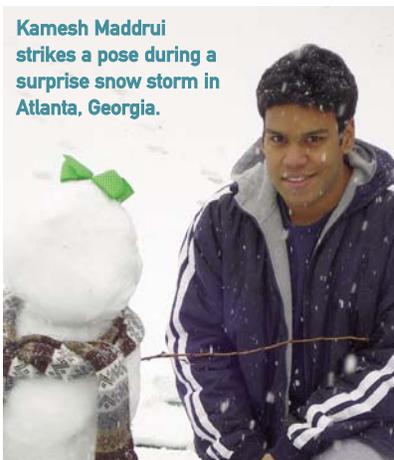
Madduri's work will allow researchers to quickly search for relevant information when data is presented in a graph-like structure. Similar to the way social networks represent "friends," this graph-like perspective shows how individual elements are connected in complex ways.

"On social networking sites like Facebook, you are connected to all sorts of people in different ways — you are connected to 'friend B' because you were former schoolmates,

connected to 'friend C' through mutual friends, and met 'friend D' at a nightclub. A graph data structure allows you to see all of those different and complex connections," says Madduri. "If scientists are able to study or analyze their data in a similar fashion, with a systems perspective, they may find unexpected connections, and that is very interesting."

This past August, Madduri received a Ph.D. specializing in computational science and engineering from the Georgia

Kamesh Madduri strikes a pose during a surprise snow storm in Atlanta, Georgia.



Institute of Technology in Atlanta, Ga. Prior to that, he completed a bachelor's degree in electrical engineering at the Indian Institute of Technology Madras, in Chennai, India.

He currently lives in south Berkeley. Outside of work, Madduri enjoys reading, sipping coffee in cafes around downtown Berkeley, and exploring San Francisco.

The Alvarez fellowship, named for Dr. Luis W. Alvarez, the Nobel Laureate and physicist who worked at Berkeley Lab, was established to encourage the development and application of tools to advance scientific research.

Maciej Haranczyk, 2008 Seaborg Fellow

Maciej Haranczyk's interest in computational chemistry has taken him around the world, across Europe and all over the Western United States. Now, this interest and a new idea have brought him to the Lawrence Berkeley National Laboratory as a distinguished 2008 Seaborg Fellow. He is one of only three Lab-wide fellows for 2008.

Upon appointment, Seaborg Fellows are given time to review

will bring virtual combinational chemistry into design of molecules and materials. He got the idea while characterizing isomers of small fragments of DNA for his PhD thesis project.

"At first I thought that the number of interesting structures would not exceed 20 or 30, but after few months of work the number of structures I was looking at reached about 300. I realized that reaching final results is prohibited not by the time required to run calculations within computer, but by the

human time required to generate structures, submit calculations and analyze the results," says Haranczyk.

"That was the moment I became involved in development of automated tools that allow quantum chemical characterization of combinatorial generated libraries of molecules. Later on, I realized that these could have very



Maciej Haranczyk catches some rest and relaxation in Carmel, California, before starting his fellowship in Berkeley.

the Laboratory's research program and can participate in any aspect of it for three years. Haranczyk opted to join the Scientific Computing Group.

"I chose to be affiliated in this group because it provides unique research opportunity, in terms of both scientific expertise of the group members and access to one of the most powerful computing facilities at NERSC," says Haranczyk. "I am interested in tools and approaches that combine quantum chemistry, cheminformatics, combinatorial chemistry and computer science."

As part of his fellowship, Haranczyk hopes to spend the next three years developing algorithms and software that

broad application. For example, I have recently used a variation of these tools to characterize families of organic pollutants, demonstrating usefulness of my tools in environmental science."

Born and raised in Gdansk, Poland, Haranczyk's interest in chemistry ignited as a child watching fireworks. With time he came to realize that it is much more fun to create things with chemistry, rather than light them on fire. He also had an interest in computers, and when he realized that it was possible to perform chemical experiments with computers, his career as a computational chemist was sealed.

continued on page 5

IMG/M *continued from page 1*

of the human body, on individuals with a variety of health conditions. They will then use IMG/M to analyze the metagenome datasets generated from these samples.

According to Kyrpides, the field of metagenomics is relatively new. Until a few years ago scientists studied individual microbes by growing them in laboratories, extracting their DNA, and then examining the sequence of their genes in order to understand the organism's genetic makeup. While this approach was somewhat successful, he notes that it had substantial limitations because most microbes cannot be grown in laboratories. When scientists extract DNA from an entire microbiome sample containing potentially hundreds of different microbial species, they don't know which individual organism the genes come from or the function

these genes carry out in the context of the community.

"The IMG/M system is an invaluable tool in the quest of finding how communities function," says Kyrpides. "The system allows us to analyze metagenomic datasets in the rich context of all available individual microbial genomes, and provides scientists with tools to compare and identify the functional capabilities of microbial communities."

This past year, researchers used IMG/M to learn how microbes in Seattle's Lake Washington enable the oxidation of methane, methanol and methylated amines, which are compounds contributing to the greenhouse effect and the global carbon cycle.

The system's track record in analyzing metagenomes from these types of natural

environments inspired scientists working on the Human Microbiome Project to include IMG/M in their NIH proposal to create a Data Analysis Coordination Center (DACC). This center will act as a central repository for all the human metagenome data collected by the project.

The principal investigator on the NIH grant is Dr. Owen White of the Institute for Genome Sciences at the University of Maryland's School of Medicine, Baltimore, Md. In addition to Kyrpides and Markowitz, other investigators include Dr. Gary Andersen of Berkeley Lab's Earth Sciences Division and Robin Knight of the Department of Chemistry and Biochemistry at the University of Colorado, Boulder, Colo.

For more information on the DACC, please visit: <http://nihroadmap.nih.gov/hmp>.

LBNL Researchers Contribute Expertise in All Aspects of SC08 Conference



Researchers from Lawrence Berkeley National Laboratory are making significant contributions to the SC08 Conference Technical Program, contributing four technical papers and one research poster, organizing two workshops, participating in two panel discussions and hosting or co-hosting six birds-of-a-feather sessions (BoFs). The four technical papers are an indication of the strength and quality of the LBNL program in computational science, since this conference usually is very competitive and has an acceptance ratio of only one in six to seven papers. SC08, the international conference on high performance computing, networking, storage and analysis, will be held Nov. 15-21 in Austin, Texas.

Additionally, UC Berkeley Prof. David Patterson, who has a joint appointment in LBNL's Future Technologies Group, is one of four invited speakers, and Dale Sartor of the Environmental Energy Technologies Division will give a Masterworks presentation on energy-efficient computing.

One of the thrusts of SC08 is Energy and Computing. This thrust was co-organized by Steve Hammond from the National Renewable Energy Laboratory and Horst Simon, LBNL's Associate Laboratory Director for Computing Sciences. As leaders in energy-efficient computing, LBNL staff are making several contributions to the thrust area. John Shalf co-organized the Nov. 16 workshop, "Power Efficiency and the Path to Exascale Computing," and will participate in the panel discussion "Will Electric Utilities Give Away Supercomputers with the Purchase of a Power Contract?" Sartor's presentation will tell how to "Save Energy Now in Computer Centers"; and Sartor and Bill Tschudi, along with Pacific Northwest National Laboratory's Moe Khaleel, are co-hosting a BoF looking at "High Energy Performance for High Performance Computing."

LBNL staff will also be on hand in Booth 540 to demonstrate and discuss technologies behind "Green Flash," a new concept for energy-efficient high-performance scientific computer systems. The joint effort with Tensilica is focused on novel processor and systems architectures using large numbers of small processor cores, connected together with optimized links, and tuned to the requirements of highly parallel applica-

tions such as climate modeling.

Here is a list of contributions to SC08 by LBNL staff.

Technical Papers

"Stencil Computation Optimization and Autotuning on State-of-the-Art Multicore Architectures," Kaushik Datta, et al, 11-11:30 a.m. Tuesday, Nov. 18.

"Accelerating Configuration Interaction Calculation for Nuclear Structure," Philip Sternberg et al., 2-2:30 p.m. Tuesday, Nov. 18.

"Characterizing and Predicting the I/O Performance of HPC Applications using a Parameterized Synthetic Benchmark Hongzhang Shan et al., 11-11:30 a.m. Wednesday, Nov. 19.

"High Performance Multivariate Visual Data Exploration for Extremely Large Data," Oliver Rübél et al., 2:30-3 p.m. Thursday, Nov. 20.

ACM Gordon Bell Prize Finalist
 "Linear Scaling Divide-and-Conquer Electronic Structure Calculations for Thousand Atom Nanostructures," Lin-Wang Wang et al., 11:30-noon Thursday, Nov. 20.

continued on page 5

Fellows *continued from page 3*

He moved out of Gdansk after the third year of college upon receiving a summer research fellowship at Utrecht University in Holland. Since then, he has received prestigious research fellowships

at Pacific Northwest National Laboratory, University of Southern California, and University of Sheffield. Haranczyk recently received a doctoral degree in chemistry from University of Gdansk.

On his spare time, Haranczyk enjoys flying kites, skiing, hiking, and anything that has to do with marine biology. "Luckily the Bay Area provides great opportunities to do all of these," says Haranczyk.

SC08 Conference *continued from page 4*

Workshops

"Power Efficiency and the Path to Exascale Computing," co-organized by John Shalf, 8:30 a.m.–5 p.m. Sunday, Nov. 17.

"The Fourth International Workshop on High Performance Computing for Nano-Science and Technology" (HPCNano08), co-organized by Andrew Canning and Lin-Wang Wang, 8:30 a.m.–5 p.m. Friday, Nov. 21.

Invited Speakers/Masterworks

"Parallel Computing Landscape: A View from Berkeley," invited talk by David Patterson, University of California Berkeley and LBNL, 9:15–10 a.m. Wednesday, Nov. 19.

"Save Energy Now in Computer Centers," Masterworks presentation by Dale Sartor, Berkeley Lab Environmental Energy Technologies Division, 3:30–4:15 p.m. Thursday, Nov. 20.

Panels

"Applications for Heterogeneous, Massively Parallel Systems: Can Developing Applications for Massively Parallel Systems with Heterogeneous Processors Be Made Easy(er)?" including David Patterson and John Shalf, 3:30–5 p.m. Tuesday, Nov. 18.

"Will Electric Utilities Give Away Supercomputers with the Purchase of a Power Contract?" including Bill Tschudi, 10:30 a.m.–noon, Wednesday, Nov. 19.

"Exa and Yotta Scale Data: Are We Ready?" moderated by Bill Kramer, 10:30 a.m.–noon Friday, Nov. 21.

Research Poster

"When Workflow Management Systems and Logging Systems Meet: Analyzing Large-Scale Execution Traces," Dan Gunter et al., poster reception 5:15–7 p.m. Tuesday, Nov. 18.

Birds-of-a-Feather Sessions

(BoFs)

"High Energy Performance for High Performance Computing," led by Bill Tschudi, 12:15–1:15 p.m. Tuesday, Nov. 18.

"Network Measurement," led by Jon Dugan, 12:15–1:15 p.m. Tuesday, Nov. 18.

"SP-XXL: Large IBM HPC Systems," co-led by David Paul, 12:15–1:15 p.m. Tuesday, Nov. 18.

"Tools for High Productivity Supercomputing," co-led by David Skinner, 12:15–1:15 p.m. Tuesday, Nov. 18.

"TOP500 Supercomputers," led by Erich Strohmaier, 5:50–7 p.m. Tuesday, Nov. 18.

"The Challenges, Risks and Successes of Integrating Petascale Systems into Science Environments," led by Bill Kramer, 12:15–1:15 p.m. Thursday, Nov. 20.

About CRD Report

CRD Report, which publishes every other month, highlights the cutting-edge research conducted by staff scientists in areas including turbulent combustion, nanomaterials, climate change, distributed computing, high-speed networks, astrophysics, biological data management and visualization. CRD Report Editor Linda Vu can be reached at 510 495-2402 or LVu@lbl.gov. Find previous CRD Report articles at <http://crd.lbl.gov/html/news/CRDreport.html>.

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California. Ernest Orlando Lawrence Berkeley National Laboratory is an equal opportunity employer.