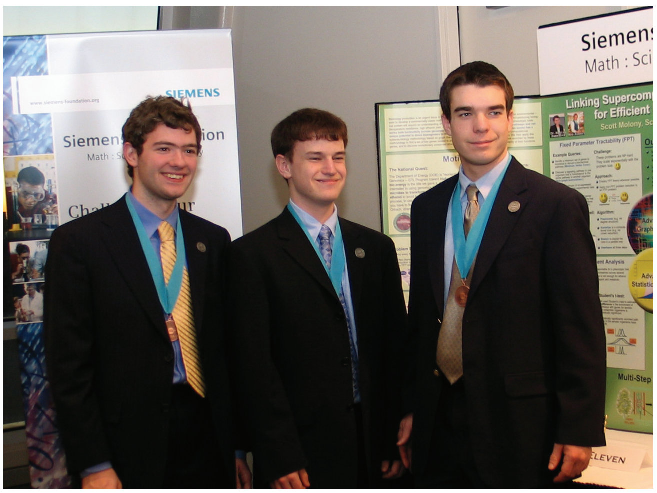
High-School Researchers at ORNL Land National Honor

Three Oak Ridge High School seniors recently used National Center for Computational Sciences (NCCS) supercomputers to improve the process for producing biofuel. As a consequence, they have also won a national math and science competition and pulled in a $100,000 scholarship.



Oak Ridge High School students Scott Horton (left), Steven Arcangeli, and Scott Molony won the Grand Prize Scholarship at the 2006–2007 Siemens Competition in Math, Science and Technology.

The students—Steven Arcangeli, Scott Molony, and Scott Horton—took the Grand Prize Scholarship at the 2006–2007 Siemens Competition in Math, Science and Technology for their work with ORNL researchers in the project "Data-Intensive Computing for Complex Biological Systems (BioPilot)."

The award grows out of a collaboration between the high school and Oak Ridge National Laboratory (ORNL). The students enrolled in a two-semester course entitled "Math, Science, Computer Science Thesis" that teamed them up with ORNL researchers.

The three worked intensively in the BioPilot program with senior ORNL researchers Tatiana Karpinets, Hoony Park, Chris Symons, and Nagiza Samatova. The program is funded by the Office of Advanced Scientific Computing Research and conducted through ORNL's Computer Science and Mathematics Division.

The students worked full time at the lab through the summer of 2006 and about 1 day a week since mid-August. Along the way, their ORNL partners instructed them in the fundamentals of graph theory, statistical theory, systems biology, bioinformatics, artificial intelligence, and programming in C and C++. The students also presented research papers related to the project to members of their team.

The three were naturals as individuals and as a team.

"They are really bright," said Samatova, who served as their mentor. "What really mattered, though, is that they really worked very nicely together as a team. They are all talented, but in a complementary way."

Their research focused on 28 microbes to elucidate how microbes can turn grasses and other biological materials into fuel, specifically ethanol. The students were interested in making a connection between desirable traits (e.g., resistance to high temperature, the ability to convert more than one form of sugar, and high ethanol yield) and specific genes and biological pathways in the microbes. To find these genes, they classified the organisms according to the presence or absence of a trait; the responsible genes would presumably be found in the group of microbes that had the trait but not in the other group. The students formulated a graph-theory problem to find the cliques of genes (pair-wise connected) that are conserved through evolution among the organisms that possess a beneficial trait.

As they looked into the utility of this method, the students ran into a challenge faced by their science brethren around the world: The problem they wanted to solve threatened to overwhelm the resources they had for solving it. Even for 28 microbes classified as aerobes (growing with oxygen) or anaerobes

OAK RIDGE
National Laboratory

ORNL 2006-G01791/aas_Jan.

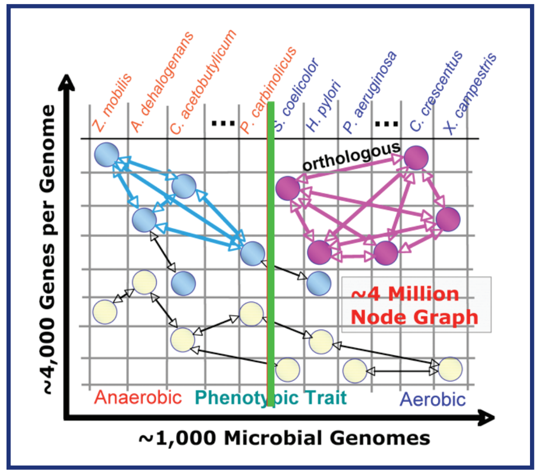
(growing without oxygen), the existing codes took too much time running on ORNL's SGI Ram supercomputer. The size of the graph was only 66,000 vertices (genes), but the problem required scaling the codes to potentially millions of genes (thousands of microbes with about 4,000 genes each).

The students took a creative approach to the challenge, which hinged on the data-intensive nature of a problem requiring access to terabytes of memory. They extended the existing codes through prototyping techniques that would ease hierarchical memory management, processor group utilization to hide memory latency, and data compression to reduce the size of the data stored in memory at any particular moment.

Their results are preliminary but promising. By looking at a reasonably well-studied trait (aerobic versus anaerobic growth) and using only a small number of microbes, the students identified a target list of important genes. Many of these were known, thus confirming the validity of their idea. Others provide fodder for the formulation of future hypotheses and their experimental validation

Samatova expects to see the three back at ORNL next summer. Their future work will focus on a database larger than the original 28 microbes and a richer set of traits directly related to ethanol production and will take them a step closer to engineering biofuels.

For NCCS information contact:
Dr. Nagiza Samatova
Oak Ridge National Laboratory
Computing & Computational Sciences Directorate
(865) 241-4351
samatovan@ornl.gov



A schematic of the method that seeks to identify genes important to a trait of interest. Aerobic versus anaerobic growth is given as an example.