access
Vol. 16 No. 2
Summer 2003

Executive Editor
Karen Green
kareng@ncsa.uiuc.edu

Managing Editor
J. William Bell
jbell@ncsa.uiuc.edu

Art Director
Carlton Bruett

Illustrators
Carlton Bruett
Ming-Yin Au-Yeung

Copyeditor
Trish Barker

On the Web
http://access.ncsa.uiuc.edu
Access is published on the Web every two weeks, covering the latest developments from the Alliance and NCSA.

National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign
152 Computing Applications Building
605 E. Springfield Ave.
Champaign, IL 61820-251
217-244-0072

ISSN 1064-9409
Copyright © 2003 Board of Trustees of the University of Illinois. All brand and product names are trademarks or registered trademarks of their respective holders.

Printed with soybean ink on recycled paper.

Who we are
The National Center for Supercomputing Applications (NCSA) opened its doors in January 1986. NCSA earned and maintains an international reputation in high-performance computing and networking and in developing innovative software applications—like NCSA Mosaic, the first readily-available graphical Web browser. Since 1997, NCSA has been the leading-edge site for the National Computational Science Alliance (Alliance), one of two partnerships of the NSF's Partnerships for Advanced Computational Infrastructure program. The Alliance is a partnership among about 50 academic, government, and industrial organizations from across the United States.

In August 2001, NCSA—as part of a four-institution team—was tapped by the National Science Foundation to build one of the world's first computational grids and put it to use. It will be the most comprehensive grid yet deployed for open scientific research, spanning the country and providing the backbone from which tomorrow's global grid can grow. With software developed to make it all work in concert, this TeraGrid will offer the fastest unclassified supercomputers as well as an unparalleled array of visualization tools, sensors and instruments, and mass storage devices. These resources will be linked via a network four times faster than today's fastest.

Major support for NCSA and the Alliance is provided by the National Science Foundation's Partnerships for Advanced Computational Infrastructure program. Additional funding for NCSA comes from the state of Illinois, the University of Illinois, industrial partners, and other federal agencies.

Cover
Model of a silicon nanocrystal's valence electron density. Physicists from Lawrence Livermore National Laboratory, North Carolina State University, and the University of Illinois at Urbana-Champaign have spent almost 300,000 hours of computing time on NCSA's Titan Linux cluster and SGI Origin2000 developing such models. The team is illuminating the causes of nanocrystals' amazing chemical, structural, and optical properties. Simulations for this visualization were conducted by North Carolina State University's Zachary Helms.
<table>
<thead>
<tr>
<th>Building nanocrystals</th>
<th>02</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using NCSA supercomputers, scientists calculate the electronic structures of molecules to predict how they will behave under a variety of conditions.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cruising with the top down</th>
<th>06</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of Illinois and NCSA researchers, using an uncommon mass spectrometer and the Alliance's Condor computing system, craft a new method of identifying proteins and characterizing changes in those proteins.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vortices unbound</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>By simulating the movements of microscopic bar magnets, a condensed matter physicist at the University of Kentucky sheds new light on the nature of a fundamental physical process.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The future of antibiotic ammunition</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Though they've been around for eons, antimicrobial peptides are among the &quot;new&quot; antibiotics being studied by a University of Minnesota research team.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A quieter blue yonder</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two University of Illinois experts in theoretical mechanics explore the origins of jet engine noise—without the screaming turbines.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Information into action</th>
<th>22</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allstate Insurance Company wins the NCSA Private Sector Program's 2003 Grand Challenge Award, honoring more than five years of data mining innovation.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A science of big numbers</th>
<th>26</th>
</tr>
</thead>
<tbody>
<tr>
<td>GADU, developed by members of the Alliance's expeditions, blasts through some of the acquisition, analysis, and storage challenges that surround bioinformatics.</td>
<td></td>
</tr>
</tbody>
</table>

| Parting shot | 29 |
Building nanocrystals

by Katherine A. Caponi

Using NCSA supercomputers, scientists calculate the electronic structures of molecules to predict how they will behave under a variety of conditions.

Model of the calculated electronic state of the excited electron of a silicon nanocrystal. Simulations by Zachary Helms, NCSU.
We normally think of building something as putting parts together to make a whole. Think of a carpenter building a piece of furniture, like a chair. She gathers her plans, tools, wood, and other hardware. Legs, seat, nails, and all, she uses those parts to make a whole, useable chair that she can then sit on. That's even how Webster's defines building—to make by putting together materials, parts, etc. By contrast, we typically think of taking something apart as demolition, destroying the purpose of the original object. But what if we could build something in such a way that we actually had to start with the "whole" and take parts away to get a useful object? Think of it like whittling or carving a delicate ice sculpture. That's what some scientists are doing with nanomaterials.

In a lab at the University of Illinois at Urbana-Champaign (UIUC), for example, scientists are building nanomaterials for future lasers, which will be one billionth of a meter long! To do this, the scientists first put a silicon wafer through a process called electrolysis reproduction. They gradually immerse the wafer in an etching bath of hydrofluoric acid and hydrogen peroxide. At the same time, the scientists apply an electrical current to the bath. Munir Nayfeh, the lead researcher and member of the Illinois experimental group, explains, "This process erodes the surface layer of the material, leaving behind a delicate network of weakly interconnected nanostructures." They then remove the wafer from the etching bath, immerse it in a liquid solvent bath, and subject it to ultrasound. The ultrasound bath causes the delicate nanostructures to crumble into particles. The particles, silicon nanocrystals, are easily sorted into different sized groups that will emit corresponding colors of light when hit with infrared or ultraviolet light.

Because the nanocrystals emit visible light and are made out of biologically benign silicon, they are useful for a variety of purposes such as sensors, semiconductor lasers, single electron transistors, and even for tagging cancer cells for study. In particular, they could provide a viable alternative to dye markers like the barium drinks used in upper gastrointestinal X-rays and some MRIs that have been used for decades. The nanocrystals are small enough to fit through the pores in a cell, and according to Nayfeh, they are also photostable—a quality that is quite valuable when researchers need to take repeated measurements under intense radiation.

But how do scientists find out which kinds of molecular structures have such special properties and how do they determine what uses can be made of them? The answer is in the electronic structure of molecules. If scientists can perform accurate calculations detailing molecules' electronic structures, they can predict how the molecules will behave under certain conditions. Lubos Mitas, assistant professor of physics at North Carolina State University and former NCSA research scientist; researchers from UIUC's physics department; and teams of scientists from other institutions have been working together to use supercomputers to calculate the molecular electronic structure of silicon nanocrystals and several other materials.

**Exciting Discoveries in Nanocrystals**

In the example of the tiny lasers, Nayfeh and his team are investigating the special properties of silicon nanocrystals. They emit light in the visible range from blue to red, they appear to be quite stable, and the electrolysis reproduction technology for homogenous crystals is inexpensive. "Moreover," Mitas states, "the observed nanoparticles tend to appear in 'magic' sizes, such as blue, yellow, orange, etc....This clearly points to the existence of 'sweet spots' in the cluster sizes that have enhanced structural and chemical stability."

Mitas and his team have joined Nayfeh to carry out an extensive study on the silicon nanocrystals, charged with the task of finding out what causes the enhanced chemical, structural, and optical properties. Because the traditional theories used to predict a material's properties based on electronic structure (which require solving highly complicated differential equations in many dimensions) tend to be less reliable when applied to new materials with unexpected physical and chemical properties, there is a clear opportunity for new theoretical alternatives.

One of the traditional methods, density functional theory (DFT), describes an interacting system of subatomic particles in terms of its density instead of its wave function and energy levels. The DFT method works well for determining geometries of molecules, but the Quantum Monte Carlo (QMC) method provides a significantly more accurate approach of calculating the energy differences, such as the optically active excitations.
Using QMC on NCSA supercomputers, Mitas and his team of researchers came to several interesting conclusions about the silicon nanocrystals. They identified the structures of those “magic” cluster sizes with strong photoluminescence in the blue, yellow, and orange ranges of the visible spectrum that were repeatedly obtained in experiments.

They also carried out calculations of the Stokes shift, or the difference between the energy absorbed by the molecule and the energy that the molecule emits afterward, for the excitation at the absorption edge of the molecule. Some of these calculations have been done in collaboration with a research group and supercomputers at Lawrence Livermore National Laboratory (LLNL) led by Giulia Galli. One of the most interesting questions was whether the excited state was localized within the molecule and how much the excitation could affect the atomic geometries. Because QMC was capable of describing the excitonic effects very well, the team found that the exited electrons were delocalized across the whole structure. They also found that the excitation causes very small changes to the structure of the molecule, and the Stokes shift calculation was in excellent agreement with the experimental data showing the actual difference in frequencies between absorbed and emitted energy.

Answering these two basic questions means that researchers now have a way to predict the color of the emitted light from silicon nanocrystals of varying sizes. It also tells them how any change of the electronic structure of the molecule would manifest itself in changes to the optical properties. In particular, they were able to predict changes in the color of the emitted light upon attaching additional molecules, such as methyl and ammonia groups, to nanocrystals. The results of their QMC simulations of chemical doping, which have been addressed in articles in the Physical Review Letters of the American Physical Society, were in excellent agreement with physical experiments.

One more fact that the researchers revealed is that hydrogen peroxide is absolutely crucial to the production of silicon nanocrystals in order for them to be stable. If hydrogen peroxide is not present in the etching bath during electrolysis, the molecules will not appear in the specific structural configurations necessary to make them a good alternative to the currently used dye markers.

The method to their magnets

The extensive calculations on silicon nanocrystals are just an example of the useful applications of calculating the electronic structures of molecules. Mitas and his team have logged close to 500,000 hours on the NCSA SGI Origin2000 and the NCSA Titan Linux cluster to help other researchers in a variety of projects.

For example, they applied QMC method to the properties of molecular magnets. Molecular magnets are of great importance to the microelectronic industry, opening a window of opportunity for prototyping increasingly smaller devices. Eventually the study of electronic structures of molecular magnets will provide the potential to tremendously increase storage capacity to just a few electrons/spins per bit of information. Additional projects involve applying QMC to ferroelectric and magnetic solid materials used in making transducers, capacitors, and sensors as well as biomolecules, which form complex systems that carry out vital tasks in our bodies.
These projects are diverse in nature yet they have something in common—the physics at work is incredibly complex. However, the development of new methods enables the scientists to reveal the electronic structure of materials on the quantum level and opens breathtaking opportunities to rapidly advance medical, microelectronic, and optical solutions in the era of supercomputing.

This research is supported by the Office of Naval Research, the University of Illinois at Urbana-Champaign, the Defense Advanced Research Projects Agency, the National Science Foundation, and NCSA.

**Access Online** [http://access.ncsa.uiuc.edu/CoverStories/electronicstructures/](http://access.ncsa.uiuc.edu/CoverStories/electronicstructures/)

**For further information:** [http://altair.physics.ncsu.edu/](http://altair.physics.ncsu.edu/)

**Team members**

**Lawrence Livermore National Laboratory**
- Giulia Galli
- Jeffrey Grossman
- Andrew Williamson

**North Carolina State University**
- Michael Bejdich
- Gabriel Droby
- Zachary Helms
- JiWoong Lee
- Lubos Mitas
- Prasenjit Sen
- Lucas Wagner

**University of Illinois at Urbana-Champaign**
- Munir Nayfeh
University of Illinois and NCSA researchers, using an uncommon mass spectrometer and the Alliance's Condor computing system, craft a new method of identifying proteins and characterizing changes in those proteins.

Cruising with the top down

by
J. William Bell
It should probably come as little surprise that James Watson—of Watson and Crick and their vision of DNA’s double-helical structure—understands proteomics’ power. At a biotechnology symposium in early 2003, Watson referred to DNA as the “script” and proteins as the “actors.”

Neil Kelleher, an assistant chemistry professor at the University of Illinois at Urbana-Champaign, couldn’t agree more.

“The realization of our society’s expectations for 21st century medicine depends on further insights into biology at the level of proteins,” he says, “not just DNA.” These insights will rely on “a new, dominant methodology for the collection and interpretation of proteomic data.”

Kelleher and his research team hope their “top-down” approach to identifying and characterizing proteins will become that new method. An advanced mass spectrometer, the Alliance’s Condor computing system at the University of Wisconsin, and NCSA’s expertise have all been integral to this nascent approach’s development.

Humans are too complex

Ian Brooks, an NCSA research programmer on the project who is also a biochemist, explains the importance of Watson’s distinction between actor and script: “For a simple organism—a bacteria, for example—you only need their genome [to address central questions about cell biology on a whole-system level]. For that organism, there is a one-to-one correspondence between a gene and the protein that it expresses.

“Humans are too complex,” he continues. “Their gene sequence doesn’t tell you what proteins are going to be made. Many proteins come from a single gene.” Though there’s much to be learned from the gene sequence (see “A science of big numbers” in this issue), it tells researchers far less when they’re studying mammals. In fact, the Kelleher group has found that about 10 percent of proteins are chemically different than the proteins expected when analyzing the gene sequence alone.

In complex critters like mammals, proteins might vary from those typically expressed due to variations in DNA transcription or RNA splicing during expression or due to chemical alterations to proteins’ amino acids following production. Thus, many different proteins can result from any single snippet of genetic code. Once expressed, and in some cases altered, the proteins go about their business—whether they’re enzymes that regulate biochemical activity, hormones such as insulin, or the structural basis for your body’s hair, muscles, and skin.

---

The Fourier-transform mass spectrometer used by the Kelleher team. This instrument includes a 9.4 tesla superconducting magnet and is one of only a half dozen such instruments in the world.

---

Schematic of the Kelleher team’s system of identifying and characterizing proteins.
The Kelleher team is out to characterize what are called post-translational modifications. Post-translational modifications change proteins' properties chemically by chopping up the proteins or "decorating" certain amino acids within the protein. Tack on a phosphate, for example, fundamentally changes the protein. It might make an active enzyme of the protein or flip a molecular logic switch from 0 to 1.

According to Brooks, there are about 300 known post-translational modifications that can alter a protein and therefore influence its function or location in a cell. Formal, direct analyses have been "limited to date and totally unsystematic," says Kelleher.

A system such as the Kelleher team's would allow researchers to characterize all of the possible modifications that a specific protein might undergo. It would also allow researchers to catalog the various modification states so the next team through the breach wouldn't have to duplicate the effort.

**Separate, sort, and identify**

To begin the identification and characterization, cells are broken apart. Hundreds of incumbent proteins are separated by mass using gel electrophoresis, and a special soap allows further, improved separation using liquid chromatography. The proteins are sorted into groups of about five to 15 similarly sized particles. Proteins that are expressed from the same section of the genetic sequence but that differ due to post-translational modifications tend to fall in the same group.

Electrophoresis, which separates molecules using electrical current, and liquid chromatography, which separates molecules by passing them through sticky pores, are nothing special for chemists and biologists. Accusations of ordinariness, however, should probably end there.

Once the proteins are separated into like-sized sets, they are fed into a Fourier-transform mass spectrometer with a 9.4 tesla superconducting magnet. There are fewer than six research labs in the world with this type of instrument, according to Kelleher, and only the Kelleher team is pursuing the top-down approach to identifying and characterizing proteins.

The mass spectrometer ionizes the proteins and sets them spinning in the instrument's magnetic field. With a bit of computational analysis, the instrument's read on the spin can be converted into a mass measurement because the ionized protein's spin in the gas phase is a function of its mass. The faster the protein ion spins, the lower its mass. Masses, in turn, can be compared to a database and equated to a particular type of protein.

**Of traditional character**

Identification of proteins, therefore, is a straightforward enterprise that can be carried out using a more ho-hum mass spectrometer. Characterizing the post-translational modifications—tracing the steps between the protein initially expressed by the gene and the protein at work, or at times wreaking havoc, in your body—is a far more daunting prospect.

---

**Schematic highlighting the differences between the Kelleher team's top-down approach to protein characterization and the more traditional bottom-up approach. The top-down approach is shown in the lower portion of the diagram and usually cuts the protein ions into at most two pieces, making post-translational modifications (PTMs) easier to characterize.**

In a traditional characterization scheme, a protein is broken into many small pieces, typically fewer than 20 amino acids long. Various enzymes chew the protein into these pieces after separation but before the sample enters the mass spectrometer. Researchers predict the type and size of the resulting fragments based on the make-up of the protein and the types of enzymes used. They compare these predicted values to the values that result from the mass spectrometer. If the values match, that section of the protein is assumed to be unmodified. If they do not, it is assumed a modification has taken place.

This approach is powerful and time-tested. But since many fragments go missing, fingerprinting modifications is difficult. The smaller the piece, the less likely it is to be detected and properly identified. The method also tends to go south if numerous fragments have been modified because it becomes more difficult to determine which modified fragment is really the cousin of the predicted fragment.

**Screaming through samples, screaming for Condor**

With the new mass spectrometer online, the Kelleher team is eschewing tradition and using its top-down approach. Kelleher, then a PhD candidate, was part of the Cornell University team that first proposed the method in a 1999 *Journal of the American Chemical Society* article.
Inside the mass spectrometer, an infrared laser cuts each protein ion into two pieces (in almost every case). Predicted and derived masses are compared as they are in traditional methods. Because the fragments are much larger and, most importantly, because the fragments include one of the protein sequence's terminal ends, it is easier to determine what modifications have taken place where. By comparing data gathered from cutting different copies of the same protein in different places, the team can characterize multiple modifications more readily and do so with markedly better accuracy.

The Kelleher team first used the system to study the proteins of Methanococcus jannaschii, an autotrophic bacteria that lives at the bottom of the ocean, and Saccharomyces cerevisiae, also known as baker's yeast. Results were published in 2002 in the journals *Nature Biotechnology* and *Analytical Chemistry*, respectively.

Today, the targets are proteins from human cells. The team hopes to process some 100 million cells, identifying and characterizing the modifications of any protein that occurs more than 1,000 times in each cell and is less than 600 amino acids long. A Web portal, called ProSightPTM, serves as a clearinghouse for the data and provides tools for others doing protein analysis.

Once it's running at full capacity, the team's mass spectrometer will create about one gigabyte of data per day and will operate 24-7. With numbers like that, traditional methods of converting the newly produced data into masses and equating those masses to particular proteins and protein fragments are not an option.

The calculations that go into the analysis of an individual sample are not intense or time-consuming. "You can do it on a fast desktop in less than 10 minutes," says Brooks. "But, by the end of the year, they're going to be producing five to seven hundred datasets a day... At that rate, you're looking at one to four days of computing time if the calculations are run serially."

Rather than letting all that excess mass spectrometer capacity go to waste, Kelleher and company teamed up with NCSA to port the analysis software, called THRASH, to run on the Alliance's Condor system, which pools idle time on desktop systems to allow for high-throughput computing. One of Kelleher's graduate students, Jeff Johnson, completed part of the work—with Brooks and Peter Andrews of Eastern Illinois University—in a matter of days.

"It was a problem that just screamed out for Condor," says Brooks. "You don't need a huge amount of memory. The jobs have short run times and are easily crunched. What you want here is not one computer that can crunch one big problem, but a lot of computers that can crunch lots of little ones."

As a result of this natural fit, the portion of the analysis that converts the raw mass spectrometric data into database-ready queries is screaming along on Condor. With some 200 processors working in tandem, analyses that might have taken days to complete are now finished in 30 minutes. Other portions of the analysis process are likely to be moved to Condor in the future, according to Kelleher and Brooks, providing plenty of opportunities for interaction between NCSA and Kelleher's rapidly maturing group.

This research is supported by the Packard Foundation, the Burroughs-Wellcome Fund, the Chicago Community Trust's Searle Scholars Program, the Research Corporation's Cottrell Scholars Award, the University of Illinois at Urbana-Champaign, and the Sloan Foundation.

**For further information:**
- [http://kelleher.scs.uiuc.edu/](http://kelleher.scs.uiuc.edu/)
- [http://prosightptm.scs.uiuc.edu/](http://prosightptm.scs.uiuc.edu/)

**Team members**
- Andrew Birck
- Ian Brooks
- Yiyu Du
- Jonathan T. Ferguson
- Andy Forbes
- Nicole Friel
- Leslie M. Hicks
- Lihuai Jiang
- Yong-Bin Kim
- Neil Kelleher
- Ryan McCarthy
- Fanyu Meng
- Leah Miller
- Steven Patrie
- Dana Robinson
- Michael Roth
- Gregory K. Taylor
- Dan Wright
Vortices unbound

By simulating the movements of microscopic bar magnets, a condensed matter physicist at the University of Kentucky sheds new light on the nature of a fundamental physical process.

By Kathleen Ricker
The phrase "loss of order" brings to mind images of a momentous nature. Floods uproot enormous trees. Tornadoes lay waste to entire towns. Fires spectacularly blacken hundreds of acres of dry forest. Disorder goes hand in hand with catastrophic change.

Similarly, it has always been assumed that, on a microscopic level, dramatic change goes with the disordering of matter—what condensed matter physicists call a "phase transition." During a phase transition, matter, in changing from one state into another, loses its structure. Take, for example, the case of water: After it reaches its boiling point, it can only be converted to steam, regardless of how high the temperature is raised.

But when and where does this disordering start? Previously, it was believed that a phase transition was always marked by dramatic behavior, such as the transformation of water into steam. In the language of phase transitions, these are called "singularities," and the standard theory assumes that a singularity always marks a change in phase. But Herb Fertig, a condensed matter physicist at the University of Kentucky, argues that researchers may have been, in a sense, looking for a sign where there is not always one to be found.

Using the University of Kentucky's Hewlett-Packard Superdome cluster, an Alliance supercomputing resource, Fertig studies the behavior of vortices in thin film ferromagnets: tiny whirlpools within a layer of magnetic material so flat that it is nearly two-dimensional. These systems are analogous to many other systems—including superfluids, superconducting materials, and thin crystalline solids, all of which have measurable properties determined by the state of the vortices or similar objects. Thus, the study provides an interesting window into how many other systems work.

The vortices are minuscule. Contrary to the conventional wisdom of condensed matter physics, however, their phase transitions may be nearly imperceptible.

Magnets within magnets

The thinnest of magnets actually consist of trillions of atoms, which themselves behave like tiny bar magnets. For certain atoms—iron, nickel, and cobalt, for example—these bar magnets favor an ordered state at absolute zero, in which all the magnets are aligned. In two dimensions, when the temperature is above absolute zero but not too high, there is always some disordering, although not very strong. Condensed matter physicists say such states possess "quasi-long range order." The system only becomes truly disordered when vortices make their presence felt.

In most thin-film magnets, vortices come in two varieties. A vortex consists of a magnet rotating clockwise in a horizontal plane around a fixed core. By contrast, an antivortex consists of a magnet rotating counterclockwise. The two objects are, in a way, like bar magnets themselves, oriented perpendicular to the plane of the real magnet. When the north pole of such an imaginary magnet is above the plane, there is a vortex; when it is below the plane there is an antivortex.

Like simple magnets, vortex "magnets" that are all pointed in the same direction tend to repel one another, to move as far from one another as possible. Thus, a field full of them will take on a lattice pattern.

However, like two simple magnets positioned so that north and south poles are side by side, vortices that point in different directions will pair up, or bind. This is where things get really interesting, because the behavior of vortices is strikingly similar to that of elementary particles. When the temperature is high enough, the vortices are forced apart, or deconfined, despite their natural tendency to bind. It is only then that the thin-film magnet becomes truly disordered. This process was discovered in the early 1970s by a pair of researchers named Kosterlitz and Thouless, and it is accompanied by a weak but definite singularity—evidence of a phase transition.

Figure 1. Think of a magnet as a crystal of microscopically small bar magnets. The small magnets are the electrons in the atoms of the materials. Each bar magnet has a north pole (shown in red) and a south pole (blue). When these bar magnets align along a common direction—which happens for ferromagnetic materials when not too hot—the magnetization may be observed macroscopically. The tendency to align has to do with the way neighboring atoms interact and occurs only for certain materials, such as the iron atoms in a refrigerator magnet.
But what happens when a magnetic (or “symmetry-breaking”) field is turned on? “Its effect is to try to make the little bar magnets line up along a particular direction,” says Fertig. As the system attempts to force all the magnets to point in this direction, it conflicts with the natural tendency of the bar magnets in the film to rotate around the centers of individual vortices. “They have to rotate in a circle,” he explains, “but because of the extra magnetic field, the system doesn't want to do that, and the way it compromises is by confining the rotation to as small an area as possible.” The result is what Fertig calls a string—a line in the film across which the bar magnets rotate in spite of the force exerted by the magnetic field. The strings connect vortices to antivortices and enhance their binding. And because of this, researchers realized that the Kosterlitz-Thouless mechanism for vortex unbinding would not work in a symmetry-breaking field.

**Elemental changes**

If vortex pair unbinding in a magnetic field is a phase transition in the usual sense, it should be accompanied by a singularity—quantities with peaks or cusps when plotted as a function of temperature or magnetic field. “People were looking for those singularities,” says Fertig. But they were never found, and strong arguments were developed that said they could not be present. But if unbinding were always accompanied by singularities, this would mean vortices would always be paired. “No matter how hot you make the magnet, you can't make the vortices get to the completely disordered state. This seemed wrong to me.”

So Fertig created a physical model in which he could simulate the conditions that would lead to the phase transition, the state at which the vortices would unbind. “I found that this unbinding occurs, but that it has a very different character from [previous studies]—[it's] very subtle,” he explains. “The vortices do unbind, but you don't see singularities.”

Fertig realized that if this was true—if there really was a previously unknown unbound vortex state—it could have important implications not only for the study of magnetic vortices but for that of many other phenomena that undergo phase transitions. “It opens up the question of what is meant by a phase transition,” says Fertig, “if you don't have to go through a singularity.”

If no singularity occurs, the only way to understand what is happening is to look closely at the vortices themselves—a difficult task in the laboratory, where the particles are on the order of nanometers and can be seen only by scanning tunneling microscopes. So, to test his hypothesis, Fertig has created a simulation in which a lattice of bar magnets is subjected to a magnetic field. The simulation allows him to easily track the vortices. “As the simulation goes on, the configuration of the magnets is changing, and at any moment I can... identify where the vortices are.”

Fertig has developed a measure of how far apart the vortices are at any moment in the simulation. The idea is that because of the way the simulation is constructed, there's actually a maximum distance that the vortices can be apart. We run the simulation and we ask how many times during the course of the simulation we see vortices [separated by] the maximum possible distance.” Keeping track of the fraction of configurations with this maximum separation allows Fertig to identify when a phase transition might be taking place. If the fraction decreases to zero as the system size is increased, the vortices are in a bound state. However, if the fraction reaches a finite number, he can identify an unbound state.

In his simulation on the Superdome at the University of Kentucky, Fertig is able to show three phases. With a low-intensity magnetic field and a reasonably high temperature, the vortices can be unbound. He has also identified two phases in which a high-intensity magnetic field causes the vortices to bind together in two different ways.
favor alignment along the direction of a magnetic field, the region in which the bar magnets rotate around a vortex or antivortex configuration of microscopic bar magnets. Because bar magnets the enhanced binding would stop these pairs from separating, + indicates a vortex and * an antivortex.

As a result, Fertig is getting closer to identifying the boundaries between the three phases. "We have two 'knobs' that we can turn," he says, magnetic field and temperature. "If you make a graph of temperature on one axis and magnetic field on the other, there will be a line that separates the unbound vortex phase from the bound vortex phases, and we would like to know where that line is."

Fertig is currently conducting these same simulations on a much larger scale, trying to map out the phase boundary. To achieve this, and to confirm the correlation between the formation of strings and the unbinding of vortices, Fertig predicts that he will use 120,000 hours of computing time.

It requires a lot of computing power to study something so infinitesimal that it went unnoticed for 25 years.

This research is supported by the National Science Foundation.

Access Online http://access.ncsa.uiuc.edu/CoverStories/vortices/

For further information:
http://www.pa.uky.edu/bios.Fertig.html

Team members
Herb Fertig
Joe Straley

Figure 5 and Figure 6. Snapshot of a simulation for a clean system. + indicates a vortex and * an antivortex. Most of these are bound tightly into pairs. Occasionally, however, one can see a relatively separated pair (Inset, Fig. 6).

The bar magnets are shown as lines with a dot indicating the south pole of the magnet. The magnets are color coded to indicate how much they deviate from the direction favored by the applied magnetic field. Red magnets point nearly antiparallel to the direction favored by the applied field. The region of red connecting the vortex-antivortex pair is the string of overturned spins tethering them together.

The simulation demonstrates that when the system is heated up, the fluctuations of the string—its entropy—can become great enough to effectively allow pairs of vortices to unbind. The nature of this phase transition is unique in that it does not lead to singularities of the type seen in other phase transitions.
The future of antibiotic ammunition

by

Jennifer Allerson

Though they've been around for eons, antimicrobial peptides are among the "new" antibiotics being studied by a University of Minnesota research team.
Antibiotics are indispensable to the modern health care system, assisting and complementing our immune systems. Over the past 10 years, the rapid emergence of bacteria strains that are resistant to multiple drugs has heightened the need to develop new classes of antibiotics. A particularly promising class of these new antibiotics is called antimicrobial peptides (AMPs).

Hundreds of AMPs of vertebrate and invertebrate origin have been discovered in the past decade, but they have existed since prehistoric times. Gene-encoded antimicrobial peptides are now well known to be a pervasive component of the immune defense system throughout the animal kingdom. They work by attacking the bacterial cell membrane. This process is called cell lysis. Formation of pores in the protective layer causes cell death by allowing the flow of ions and molecules into and out of the cell non-selectively. Most other antibiotics attack bacteria by attacking specific molecules that are part of the bacterial cellular machinery, against which bacteria can develop resistance. Yet bacteria have not been able to develop a resistance to AMPs that have existed for millennia. In order to develop resistance to AMPs, the bacterium must reengineer its outer membrane, a very difficult and complex proposition.

Unfortunately, most naturally occurring AMPs are toxic. They cause hemolysis, or the premature breakdown of blood cells, and are thus inappropriate for therapeutic purposes. The exact mechanism behind AMPs bringing about cell lysis, whether in blood cells or targeted microbial cells, is not yet understood. Hence, further efforts are needed to understand and engineer AMPs that are less toxic and have improved therapeutic qualities.

Yiannis Kaznessis, who leads one such research project at the University of Minnesota, is focusing on a class of AMPs called cathelicidins. He and his graduate student, Himanshu Khandelia, are trying to understand how these antimicrobial peptides work at the molecular level. They are using computational modeling on NCSA’s SGI Origin2000 and new IBM p690 supercomputing system known as Copper to quantify interactions with lipid molecules, which are an integral part of membranes in both pathogens and friendly cells. How exactly AMPs disrupt the membrane structure remains unclear. This lack of a molecular-level picture hampers the engineering of peptide antibiotics. However, the Kaznessis team’s observations will help clarify this picture and enable the design of a novel class of antibiotics.

Small, fast, and lethal

Kaznessis has chosen cathelicidins as the focus of his research for many reasons, but perhaps most important is their pervasiveness. They are a major antimicrobial peptide family found in many mammals, including pigs, guinea pigs, mice, goats, cattle, sheep, rabbits, and monkeys. They are also one of the two major classes of AMPs produced in the human body. Cathelicidins typically are found in white blood cells produced in myeloid organs, but they also can be found in the spleen, stomach, airways, and intestines.

Cathelicidins are also characterized by their ease of synthesis, the speed at which they work, their small size, and their remarkable structural diversity. They are capable of destroying pathogens in minutes whereas current antibiotics may need days or even weeks to overcome an infection. Speed is particularly important for fighting viruses. Many viral infections mutate very rapidly. That’s why a new strain of influenza emerges every year.
"AMPs will be very useful against viral infections once we come up with a good design of a peptide that can attack the outer viral envelope," Kaznessis adds. It is not as easy for viruses to mutate to stop an AMP attack on their lipid bilayers. Peptides have alpha-helical, extended helical, or beta-sheet conformations. This structural diversity, coupled with their small size, make them a threat to a wide range of microbes. For example, bacteria, viruses, and fungi are often attacked with multiple peptides of different structural types. Therefore, the microbe cannot rest after surviving the attack of one or two peptides. This wide number of AMP sequences and structures makes difficult for pathogens to develop resistance.

Promiseing research

Promising data about the activities and specifics of many of the cathelicidins are already available. Current experiments have quantified and analyzed the interaction of several AMPs with lipids. These experiments have shown that cathelicidins in particular are very active against a host of pathogenic organisms. They have also shown that mutations of at least one cathelicidin, known as ovispirin-1, can reduce harmful effects while preserving the peptide’s antimicrobial effects.

As a result, the Kaznessis team is working to understand how this peptide, among others, causes remarkable alterations in lipid structure. They have already implicated specific electrostatic interactions between specific amino acids and lipid head groups. "We will help design better and more rational experiments from our knowledge and observations of molecular-scale phenomena," Kaznessis says.

The research team’s initial results are also valuable because they corroborate experimental data. "We have seen two peptides disrupt model lipid monolayers and bilayers in our simulations," Kaznessis says. Simulations with other peptides in the next two years will lead to a more complete, global understanding of the nature of peptide-lipid interactions.

In order to create simulations of empirical research, the Kaznessis team is standardizing a procedure to carry out dynamic molecular simulations of peptide-membrane systems that will run the CHARMM (Chemistry at HARvard Molecular Mechanics) software package with the PARAM28b parameter set. This CHARMM package has allowed the team to model the lipid molecule system, water molecules, and peptides using an empirical potential field. This field is used to calculate forces on individual particles in the system. Newton’s equations of motion are then solved for all atoms of the system to obtain their time-dependent trajectory in phase space. Thus, they are able to see how these molecules behave and interact in the presence of each other. Ultimately, these trajectories can be used to calculate several important properties like surface tension, lipid order parameters, and peptide dihedral angles, which can be correlated to experimentally observed quantities.

Kaznessis has simulated more than 30 systems and five different peptides on the SGI Origin2000 at NCSA. Because NCSA is phasing out this system, his team is moving to the new IBM p690, transitioning its version of CHARMM to the new architecture. Though they use the machines to run processor- and memory-intensive molecular dynamics simulations of large lipid-peptide systems, they also plan to simulate metabolic and gene networks, as well as protein folding.
Kaznessis' team has just begun to better understand how cathelicidins work. Ultimately, they hope to demonstrate systematically that "fine-tuning," or manipulating peptide sequences to achieve various results such as reduced toxicity, is possible. Their research will also enhance the understanding of how other AMPs work. A global understanding of the interaction of peptides with model lipid systems will lead to the design of some of the most effective and long-lasting antibiotic medicines ever known.

This research is supported by an Amundson Fellowship from the University of Minnesota's chemical engineering and materials science department, the Digital Technology Center at the University of Minnesota, and 3M.

Access Online  http://access.ncsa.uiuc.edu/CoverStories/cathelicidins/

For further information:  
http://www.cems.umn.edu/research/kaznessis

Team members
Yiannis Kaznessis
Himanshu Khandelia
Benjamin Anderson
Two University of Illinois experts in theoretical mechanics explore the origins of jet engine noise—without the screaming turbines.

A quieter blue yonder

by

Kathleen Ricker
Anyone who has lived under the flight pattern of a major metropolitan or regional airport can attest that constant exposure to jet engine noise is a serious problem. Numerous studies have shown that the relentless roar from a busy runway poses all sorts of threats to the health and quality of life of its nearby neighbors, from sleep disruption to low birth weight. The reverberations from jet engine noise can become so intense that they have even been known to cause metal fatigue in aircraft components.

Aircraft manufacturers have made tremendous strides in making jet engines quieter in the past several years by focusing on the shape of the nozzle, the outlet at the back of the jet engine where compressed air mixed with jet fuel is released to propel the airplane forward. However, these improvements are achieved through a time-consuming, expensive trial-and-error process that involves comparing the way different kinds of nozzles alter the flow of air through the engine, resulting in different noise levels.

Moreover, while the experimental apparatus may achieve the desired results, it can't explain them. At NASA Glenn Research Center in Cleveland, Ohio, “researchers conducting the mechanical experiments have a table covered with different attachments to put on nozzles,” explains Jonathan Freund, assistant professor in theoretical and applied mechanics at the University of Illinois. “They try one after another, and if one works, they’re happy. They don’t know why; they don’t know if it’s the best they can do.”

However, what currently can't be easily determined directly from mechanical trial-and-error experiments very well could be simulated, and that's precisely what Freund and his graduate student, Mingjun Wei, are doing. Using a code that Wei has written and that forms the basis of his dissertation, Freund and Wei are using NCSA's SGI Origin2000 and Platinum Linux cluster to clarify the nature of the mechanism of aerodynamic noise by working backwards. Work is under way to run the problem on the Titan cluster and the new IBM p690 system called Copper.

First principles

Sound is a traveling pressure disturbance in a fluid, which can be envisioned as an invisible ripple in a pond. Closer to the center of the disturbance, the waves are shorter and their vibrations are more powerful. Farther from the disturbance, the waves grow longer and the vibrations fainter. However, noise is a specific kind of sound, produced by friction between layers of air that results in irregular vibrations that are extremely irritating to listeners.

The air leaving a jet engine is highly unstable. Its unsteadiness creates the irregular vibrations that we recognize as noise, just as the unsteady motions of a loud speaker make noise, only a good deal louder. The unsteady turbulent flow exiting a jet engine is very complex, so the underlying mechanisms of jet noise are unclear. A simplifying principle has not been identified. “There are some models that explain the phenomenon,” Freund says, “but not to the fidelity that’s needed to actually do something with them. So we’re circumventing this lack of clear understanding with the method we’ve developed.”

Thus, the noise produced by a jet engine can be described by the same set of equations that describe how the aircraft itself can fly. Beginning at the nozzle, where the noise is emitted with the jet exhaust, Freund and Wei compute the sound intensity at a distance from first principles—the most fundamental equations for a given phenomenon. In this case, first principles mean the Navier-Stokes equations, which describe the motion of a compressible fluid, in this case air. The Navier-Stokes equations relate the rates of change of fluid density, momentum, and energy to pressure differences and viscous forces in the fluid (the last of which are created by the extreme friction generated when fluid layers are perturbed).
Putting it in reverse

Freund and Wei start with a simulation of a two-dimensional mixing layer, a thin turbulent region where two streams of fluid mix and create a layer of instability. This is phenomenologically similar to flow in the region just downstream of a nozzle. They numerically solve the Navier-Stokes equations for this near-nozzle region without modeling a physical approximation of the field. This approach is called direct numerical simulation (DNS). They then solve the adjoint of the Navier-Stokes equations in reverse. The adjoint is an equation that superimposes the perturbations caused by changes in pressure on their numerical model of the near-nozzle region. Because Freund and Wei know both the initial conditions (the airflow past the nozzle) and the end result (sound intensity), they are able to identify the remaining piece of the puzzle: how sensitive the noise is to even the slightest change in nozzle conditions.

This last piece of information is especially crucial. “After solving the equations in reverse,” says Wei, “we know how to change the control [conditions] at the nozzle to reduce the noise further.” Simulating these processes numerically allows Freund and Wei to work with enormous numbers of individual data points. “Because each control point in space and time is optimized individually, we are actually managing three million space/time control parameters.” The runtime for a single simulation on Platinum is equally staggering—22,000 CPU hours. However, Wei has put a great deal of effort into parallelizing his code and estimates that on Copper, it will run five times as fast.

“We have something unique right now,” says Freund. “We have simultaneously a flow that makes a lot of noise, and a flow that’s been perturbed slightly and is a lot quieter, and so we can go on from there and try to figure out what’s changed. Hopefully we will be able to generalize that to more practical problems and learn from it.”
Freund and Wei emphasize that the simulations they're performing are intended to complement aerodynamical experiments going on at Glenn, Boeing, and other federal and commercial research facilities. While these simulations can't singlehandedly produce a quieter engine design, they can go a long ways toward helping aerospace engineers who are directly involved in mechanical experiments understand the mechanism of aerodynamic noise caused by turbulence. "We're hoping to guide the experimental process eventually through a better understanding of how noise is generated," says Freund.

This research is supported by the U.S. Air Force Office of Scientific Research.

**For further information:**
http://www.tam.uiuc.edu/directory/faculty/freund.html

**Team members**
Jonathan Freund
Mingjun Wei
Information into action

Allstate Insurance Company wins the NCSA Private Sector Program's 2003 Grand Challenge Award, honoring more than five years of data-mining innovation.
There’s value to be found in any pile of data. When you’re facing a molehill, it’s easier to dig up. Patterns, if they exist, can be found with little effort. Predictions can be made based on those patterns. And any deviation from the norm catches your eye. But for Allstate Insurance Company, molehills gave way to mountains not long after the company paid its first claim on a broken car door handle in 1931. Today, with more than 30 million policies in force, finding patterns and deviations in the vast customer, geographic, economic, and other data requires more than just a keen actuary and a slide rule.

To glean more insight from the data that they collect and have access to, Allstate has worked with members of the NCSA’s Automated Learning Group for more than five years in a field known as data mining. Data mining combines statistics, machine learning, pattern recognition, database management, and high-performance computing to produce automated methods for analyzing data sets and discovering new knowledge. It’s a burgeoning and challenging field that no company can afford to ignore in today’s rapidly evolving business environment.

Together, the team has taken advantage of data mining to improve customer service, experiment with new pricing systems, and develop better ways for the company’s analysts to keep up with changes in the competitive environment. These efforts allow the company to work more efficiently and better target its products to its customers—whether those customers are buying a new car, insuring their first home, or recovering from a fire.

In April NCSA’s Private Sector Program presented its 2003 Grand Challenge Award to Allstate, honoring this expansive data-mining relationship.

Wide range of implications

NCSA’s Private Sector Program gives partners access to all of NCSA’s leading-edge technology and knowledge. It creates an environment in which companies are free to experiment with their most daring ideas. The annual Grand Challenge Award honors breakthrough research completed by private sector partners while working with NCSA. These breakthroughs ensure America’s leadership in global business.

“Our relationship with NCSA is...judged on one critical dimension and that is our ability to execute our objectives in the marketplace better than anyone else. [We’re going to] be able to take reams and reams of customer data, and reams and reams of customer transactions, and turn those into actions that the customers value,” says Ronald McNeil, Allstate’s senior vice president of product operations. “The partnership between NCSA and Allstate has allowed us to turn information into action quicker than our competitors.”

NCSA Director Dan Reed says that turning information into action has wide-ranging implications. “Many of the problems in data management and data mining cut across a broad range of business environments and business sectors. The common theme is business intelligence. And the way you glean that intelligence is by taking advantage of the raw data that your business produces in everyday practice. Data mining...gives you the power to be earlier to market, to shape products that more correctly and more accurately match the expectations of your customer base. It allows you to tailor products, to respond more nimbly to the business environment.”

More than 100 data-mining modules

Allstate’s work with NCSA relies heavily on the Automated Learning Group’s premiere software, called D2K. Short for Data to Knowledge, D2K integrates more than 100 modules representing both common and unique data-mining algorithms. These algorithms can, among other things, clean up data sets and prepare them for computations, search for patterns, make predictions, identify unusual features, and visualize the data for further analysis. Within a visual programming environment, users can connect these modules and create powerful problem-solving systems.

From the data, they can forge knowledge. And from that knowledge companies like Allstate can make better decisions—about who and what to insure, how much to charge, and how to deliver what their customers may want and need.

Roger Einbecker, Darryl Osman, and Michael Sullivan, all of Allstate, receive the 2003 Grand Challenge Award from NCSA’s Dan Reed.
"NCSA is one of the only places in the world that can offer a computational resource—the TeraGrid—a data-mining framework—D2K—and access to application expertise through the Automated Learning Group to address these large-scale data-mining applications," says Michael Welge, director of the Automated Learning Group.

**Tactical Competitive Intelligence Network**

One of Allstate's greatest data-mining successes has come with its Tactical Competitive Intelligence Network, or TCIN. Relying on a subset of D2K modules that represent sophisticated text-analysis algorithms, TCIN is a boon for the company's analysts. Analysts search the Internet and other information sources for data on things like competitors' pricing structures or the safety ratings of a particular make of car, then they refine the cost of insurance policies offered by the company.

The TCIN system relies on another piece of NCSA software, called VIAS, to retrieve data from the Web, newsgroups, and public mailing lists to analyze competitive information. Algorithms selected from the suite of D2K tools and easily combined using D2K's visual programming environment then filter the significant portions of the data retrieved, group similar and related pieces of information, and classify those groupings. The system even visualizes the results.

And all of this is done automatically.

According to Jeff Deigl, Allstate's assistant vice president of product operations, research, and development, analysts use "three quarters of the time spent on a project getting the data and getting it prepared." Only that last 25 percent is spent understanding what the data are telling them. "By shrinking the time that the analysts need to spend on data prep, it allows them to get to more projects and to be more effective in the analysis end product. Which is where, really, the key decisions are made."

**Territorial Rate Making**

Allstate's Territorial Rate Making project gives the company another opportunity to refine the methods used to set the price of insurance policies in a given region. Traditionally, companies have simply looked at the historical data for a given territory. With the help of NCSA and D2K, however, Allstate is taking a more complicated view.
"In setting our territorial rates, we've relied only on our internal loss data to estimate what we should charge in different geographical locations...NCSA and their tools have allowed us to supplement that information with external data," says Deigl. "For example, in assessing the fire exposure we face, understanding the forest cover and the amount of precipitation in an area is very useful and helps us better estimate the losses we'll incur."

By combining all of this data—and more importantly, by drawing new knowledge from this data—Allstate is doing what it's always done best: delivering competitively priced products that help meet the needs of its customers. It's a simple goal. And, through Allstate's partnership with NCSA's Private Sector Program, it's a goal that's better realized every day.

"Our relationship with NCSA has grown over the years, and I would consider it now to be an extraordinary partnership. NCSA provides us a research environment that includes leading-edge hardware, software, and analytic methods that, when coupled with the research staff, allows us to address Allstate business problems in ways that we can't in our own computing environment," says Roger Einbecker, Allstate's assistant vice president of information technology.

"NCSA helps us to better understand the knowledge of our business that's locked in our data."

Access online: http://access.ncsa.uiuc.edu/CoverStories/allstate03/

For further information:
http://www.ncsa.uiuc.edu/About/PSF/
http://www.ncsa.uiuc.edu/Divisions/DMV/ALG/
A science of big numbers

by J. William Bell

GADU, developed by members of the Alliance’s expeditions, blasts through some of the acquisition, analysis, and storage challenges that surround bioinformatics.
Natalia Maltsev calls bioinformatics a "science of big numbers." Instead of focusing on a single cell, protein, or organism in the lab or using computer simulation, bioinformatics looks at numerous organisms computationally. It is the search for similarities and differences in hundreds of thousands of genome sequences, protein structures, and other features of biological systems and particles. When properly understood, these variations can show researchers a given piece of the system's function.

Labs around the globe constantly churn out new sequences, running the gamut from the most simple virus to hugely complex organisms like ourselves. According to GOLD, an online guide to published genomic data, some 140 species' genomes have been completed and nearly 600 other organisms are currently being sequenced. In August 2002, GenBank, which is one of the main sequence databases, contained maps representing 22 billion nucleotide bases, the individual building blocks that make up a gene sequence.

These data are a boon for bioinformatics experts like Maltsev; they're the big numbers that make a science of big numbers possible. But working with them requires tedious search sessions and cumbersome analysis procedures.

A new "analysis pipeline" that relies on grid computing promises to automate the genome-analysis process and make it much easier. The Genome Analysis and Database Update system, or GADU, is being developed by Argonne National Laboratory's computational biology group. The team includes Maltsev, Dinanath Sulakhe, and Alex Rodriguez, a PhD student in the University of Illinois at Chicago's bioengineering department, and is part of the Alliance's data quest expedition. The data quest expedition builds tools for data-intensive applications, like those in bioinformatics, that run on Alliance and TeraGrid resources.

Further help, support, and guidance come from throughout the Alliance by way of the scientific workspaces for the future expedition and the scientific portals expedition.

"The amount of data is increasing exponentially," says Maltsev. "It dictates a need to really be able to scale up the analysis capabilities...The grid and distributed computing provide an ideal match for the type of problems that bioinformatics is facing."

After a series of test runs on the Alliance's Condor system at the University of Wisconsin and the Chiba City cluster at Argonne, GADU was fully put through its paces in April 2003. The application analyzed 59 microbial genomes in about a day. This process required that more than 10,000 jobs be submitted and represented a five-fold improvement in turnaround time, according to Maltsev. The runs were completed on the Department of Energy's Science Grid using NCSA network bandwidth and storage space. Another run in June further solidified the system's value. The team compared 1.8 million protein sequences to one another using 200 hundred processors on a cluster at Argonne. What would have taken about seven years to complete on a single desktop system was finished in about three days.

"This is a great success for the field of bioinformatics—one of the first examples of the discipline taking full advantage of a grid-based system," says Dan Reed, director of NCSA and the Alliance. "One of the things we have learned over the life of the Alliance is that we're at our best when we form multidisciplinary teams and give those teams a clear mission as they focus on the deployment of technology. So this is also a great example of what the Alliance expeditions teams can do and how those teams can make contributions to a project from end to end."

**Tedious and difficult becomes automatic and reliable**

Before analysis can start, GADU has to cull the sequence data from specialized databases such as GenBank. Via a Web-based interface, GADU users select the databases that they are interested in, the bioinformatics tools they want to employ, and the frequency with which they want their data analyzed. On that schedule and with that list of goals, GADU compares the content of the selected databases to the data that are already stored on the user's local system or a designated public server. Any new content is downloaded, old genome records are updated, and files for any new genomes are added to the user's library. Alternatively, the user may choose to be notified by email of new content and manually select data to be taken from the databases.

Using Argonne's Globus toolkit, the sequence data are formatted so they can be easily passed to grid-based machines for analysis. Annotation data—notes on what other scientists have learned about the sequence—are parsed and stored separately.

"What was tedious and difficult becomes automatic and reliable," says Maltsev.
A change in the sociology of science

A prototype of a GADU-based genome analysis server, including a public portal that will provide access to genome data that have already been analyzed, will be released by early fall. But the tool’s implications are already fully realized in the minds of its creators. Most obviously, it makes the systems that bioinformaticists need for their massive calculations easy to use.

The system’s public portal, which is being developed by the Alliance’s scientific portals expedition and scientific workplaces for the future expedition, expands GADU’s benefits.

Bioinformatics has historically been something of a cottage industry. Research groups build their own in-house computers to crunch their data. This situation leaves those at smaller universities and groups with less funding to struggle. With GADU’s portal, however, “anyone can use precomputed results [that are in GADU’s database] or they can use the secure facilities of the public server to process their data. It will eliminate huge amounts of redundant work that people are doing,” according to Maltsev. “I personally believe [this sort of thinking] will completely change the sociology of science.

“It will provide access for a lot of scientists who couldn’t even dream about this to the capabilities of large computations. For the sciences that rely on the computation of huge amounts of data or complicated simulations or huge models, it will provide the framework. People will be free to quit thinking about the framework and implement the scientific part.”

Ian Foster—associate director of Argonne’s mathematics and computer science division, co-leader of the data quest expedition, and a member of the Alliance’s Executive Committee—concurs, pointing out that a framework like GADU has the ability to impact other disciplines. In a recent issue of BioInform, a bioinformatics newsletter published by GenomeWeb, he said, “Everyone thinks they’re special [when they’re moving their applications to grid-based systems]... In some sense they’re not because the basic technology requirements are the same.”

This research is supported by the National Science Foundation, the National Institutes of Health, the National Computational Science Alliance, and the University of Chicago.

Access Online http://access.ncsa.uiuc.edu/CoverStories/GADU/

For further information:
http://www.ncsa.uiuc.edu/Expeditions/DataQuest/
http://www-fp.mcs.anl.gov/pdq/
http://www-unix.mcs.anl.gov/~maltsev/

Team members
Susan Coghlan
Terry Disz
Natalia Maltsev
Zach Miller
Michael Milligan
Nika Nefedova
Alex Rodriguez
Dinanath Sulakhe
Jens Voeckler
Gregor von Laszewski
Von Welch
Mike Wilde

WIT3, which allows researchers to model the metabolism of the organism genomic sequence data. WIT3 is developed by Argonne’s computational biology group and will benefit from GADU. By automating genome analysis, GADU will allow for the timely incorporation of data into clearinghouses and for the further interpretation of that data using systems like WIT3.

Once the new sequence data are found and taken from the public databases, GADU submits them to the grid’s computational resources for analysis. GADU currently supports three of the most common bioinformatics tools that reside on machines around the world: Blast, Blocks, and Pfam. These tools, which are variations on a theme, do the painstaking work of assessing the sequences and finding variations. Biological sequences for which the functions are unknown are compared to those with known functions using computationally intensive algorithms. The analysis is automatically checked at various points in the process. If any of these checks fails, the analysis of that segment is aborted and restarted.

“We’re talking about a humongous number of comparisons,” says Maltsev. “An average bacteria genome has 4,000 genes that encode proteins.” Compounding that, “six or seven genomes are acquired every month, and the pace is ever increasing.”

Storage represents GADU’s third function. The system places the annotations that accompany the public data in permanent storage, holds the files that are to be analyzed, and formats the results so they can be searched easily. An archive is also kept as GADU’s acquisition module collects updated versions of the same gene sequence over time. Finally, the system needs temporary storage where intermediate versions of the data are kept during analysis. This space was provided by NCSA during the Argonne team’s April run. Chimera, a data-management tool developed as part of NSF’s Grid Physics Network (GriPhyN), controls the flow of these data and properly labels and catalogs them for future search.
On April 29, NCSA celebrated the 10th anniversary of the center’s release of Mosaic, the world’s first widely used graphical Web browser. The highlight of the event was a panel discussion called “The Future Frontier: Computing on NCSA Mosaic’s 10th Anniversary,” which included Dan Reed, NCSA’s director; Groove Networks Founder, Chairman, and CEO Ray Ozzie; Vinton Cerf, co-designer of the TCP/IP protocols that are integral to the Internet’s architecture and a senior vice president at MCI; Rick Rashid, the senior vice president of Microsoft Research; and David Kuch, director of Intel’s KAI Software Lab.

Back in 1993, people saw Mosaic as an exciting new tool, but no one could have predicted that its wide adoption would lead to e-commerce, online classrooms, downloadable music and films, and new worldwide communities of people with shared interests. Likewise, no one today can say for sure where the next generation of computing and networking technologies—grids connecting people, businesses, powerful computers, tools, applications, and data—will lead us. Telemedicine could save lives. Worldwide cyberclassrooms might bring cutting-edge technologies to the most remote parts of the planet. Real-time, online weather forecasting could help people avoid tornadoes and other deadly storms before they strike.

The wide-ranging discussion at “The Future Frontier” explored these possibilities and the technologies and ideas that will make them a reality while reflecting on the impact that Mosaic and the popularization of the Internet have had on science and society.

To view the archived Webcast of the event, visit: http://www.ncsa.uiuc.edu/Conferences/MosaicEvent.

http://www.ncsa.uiuc.edu/Conferences/MosaicEvent