UMIACS
UNIVERSITY OF MARYLAND INSTITUTE FOR ADVANCED COMPUTER STUDIES

InterConnections

FALL/WINTER 2007

Contents

InterConnections is published by the University of Maryland Institute for Advanced Computer Studies (UMIACS).

Visit UMIACS on the web at: *www.umiacs.umd.edu*.

Director: V.S. Subrahmanian

Contact information
UMIACS
2119 A.V. Williams Bldg.
University of Maryland
College Park, MD 20742-3251
Phone: 301.405.6722

Lise Getoor: From Data to Information

Whether the problem is making sure that any given geographical location is represented once and only once in a database, figuring out who's who in a corporation based on e-mail correspondence, discovering the social structure among a group of dolphins, or mapping how genes regulate each other in human cells, Lise Getoor and her strategy of graph identification can help suggest a solution. Getoor, a member of UMIACS and an assistant professor of computer science, describes graph identification as a process that takes data, which is generally jumbled, noisy, and redundant, and yields information, which is tidy, logically organized, and well-annotated.

"Lise's very good at asking core questions and brainstorming," says Chris Diehl, a senior research scientist at the Johns Hopkins Applied Physics Laboratory who is collaborating with Getoor in an ongoing project in which they infer social relationships from on-line communications, particularly e-mail. "She has a unique way of breaking down problems and teasing out the essence of a question."

In Getoor's work, better understanding stems not so much from studying things and their attributes in isolation but from sorting out relationships. Whether the entities are people, events, research projects, or locations on a map, finding out how points connect to each other can yield a wealth of information about them.

Lise Getoor

"Traditionally, people have looked more at individuals instead of the links among them," says Getoor. She uses machine learning and probabilistic reasoning to clarify connections in data and their meaning. Getoor breaks down graph identification into four major parts: 1) link prediction, or finding connections among data points; 2) entity resolution or figuring out whether two separate data points in fact refer to the same underlying entity; 3) collective classification, or labeling points once data have been sorted and organized; and 4) group detection, or being able to group together points and links that share defined attributes. For her work on entity

continued on page 4...

Director's Message

UMIACS Director V.S. Subrahmanian

The preceding few months have been extraordinarily exciting for UMIACS. Our faculty has developed technology that not only has the potential to save millions of lives, but also has the potential to revolutionize how we will live our lives.

Let me first congratulate UMIACS Professor Rita Colwell on being named a recipient of the 2007 National Medal of Science by President Bush. Diarrheal diseases claim over a million lives each year – Rita's pioneering work on the study of diarrheal diseases has led to the development of interventions that can and have saved numerous valuable lives in countries such as Bangladesh.

UMIACS Professor Steven Salzberg and his colleagues in our Center for Bioinformatics and Computational Biology developed the first ever sequencing of the European strain of the avian flu H5N1 strain. Advances such as these are expected to play a major role in the development of vaccines against the flu and in possible treatments for possible flu pandemics. This issue also highlights the work of UMIACS and cell biology professor Najib El-Sayed whose work focuses on diseases such as sleeping sickness that kill over 100,000 people in Africa every year, but have not been studied with the intensity that they deserve. His attempts to sequence trypanasomes that cause this and various related diseases have the potential to save thousands of the poor and underprivileged.

One of the biggest scientific breakthroughs in computing in 2007 came from UMIACS Professor Uzi Vishkin. Professor Vishkin's research led to the development of a new desktop supercomputing paradigm that has the potential to revolutionize the computing market by speeding up desktops by a factor of up to 100. As computers are increasingly used for everything from monitoring climate change to tracking potential terrorists, such breakthroughs will play a critical role in both safeguarding US citizens, helping us understand climate change, and perhaps make possible a host of other applications that were impossible before.

These are just a few of several major advances made by UMIACS faculty during the past six months. This has not been an isolated phenomenon – since its very inception, UMIACS faculty have developed scientific breakthroughs that have helped shape a new generation of computing. However, most of these advances have not been made in isolation – rather, they are the product of a continuous collaboration with valuable partners from industry, academic, and government labs around the world. To them, I say "Thank You".

El-Sayed, continued from page 9

Metagenomics capitalizes on advances in nanotechnology and DNA sequencing technology to allow the analysis of small amounts of DNA with precision and at less expense than previous DNA sequencing techniques.

El-Sayed is interested in using metagenomics to compare the microbial communities inside the intestines of people with autism spectrum disorders versus people who are neurotypical. Genome sequencing would be a definitive way to address what some evidence suggests—that children with autism-spectrum disorders lack a proper balance in the flora of their gut. The same techniques could also be used to address whether microbial balances might be behind other conditions, such as inflammatory bowel syndrome and Crohn's disease, which can cause severe inflammation and bleeding of the intestines.

Of his new home in CBCB, El-Sayed says, "I'm surrounded by experts in genome assembly and gene finding." His work on protein interactions, analyzing protein functions, and graphically representing data also all depend on computer science, as does his lab's reliance on automation. "Computer science is an integral part of what we do," says El-Sayed.

News in Brief

Amy Karlson and **Ben Bederson** received one of three Brian Shackel Awards for "Outstanding Contribution with international impact in the field of HCI" at the INTERACT 2007 conference in Rio de Janeiro on September 12th.

Congrats to **Shuvra Bhattacharyya** on being a co-author of the paper "Low-overhead run-time scheduling for fine-grained acceleration of signal processing systems" co-authored with J. Boutellier and O. Silvén which won the best student paper award at the 2007 IEEE Workshop on Signal Processing Systems, Shanghai, China.

Rama Chellappa received the Oustanding Research Award from the A. James Clark School of Engineering.

Dr. Chellappa is also a recipient of the prestigious 2007 IBM faculty award.

UMIACS Professor **Rita Colwell** has been awarded the National Medal of Science by President Bush. The Medal of Science is the United States highest honor for science and is awarded to a small number of individuals whose work has had an unusually significant effect on the advancement of science. Dr. Colwell's pioneering research focuses on studying water borne pathogens and has led to demonstrable reductions in deaths due to diarrheal diseases such as cholera.

Dr. Colwell is also being recognized by NOAA as a Distinguished Scholar.

Bonnie Dorr and **Necip Fazil Ayan** won best paper at the North American ACL conference this year for their paper "Combining Out puts from Multiple Machine Translation Systems". The paper will appear in *Proceedings of the North American Chapter of the Association for Computational Linguistics.*

Graduate students **Vijay Gopalakrishnan** and **Ruggero Morselli** along with faculty members **Pete Keleher, Bobby Bhattacharjee** and **Aravind Srinivasan** received the best paper award at the 14th Annual IEEE International Conference on High Performance Computing this month for their paper on "Distributed Ranked Search".

David Jacobs was awarded a $50,000 advanced research grant from Honda for the study of enhanced facial recognition computer algorithms for humanoid robotics.

UMIACS researcher **Catherine Plaisant's** comments on visualization on Digg were reported by MIT Technology Review.

UMIACS researcher **Gang Qu's** work on saving power in hand-held devices such as iPhones has been featured in MIT Technology Review's August 30 issue.

UMIACS Professor **Steven Salzberg** has been named a Highly Cited Computer Science researcher in microbiology by ISIHighlyCited.com - a list of a little over 300 faculty members who have been most highly cited in their field.

An international team of researchers including **Steven Salzberg** reported the first ever sequencing of european g enomes of avian influenza virus, H5N1, in the May issue of Emerging Infectious Diseases. The sequences depict the lineages infecting wild and domestic birds in Europe and Africa and show relationships between all strains.

Ben Shneiderman's exhibit "Speculative Data and the Creative Imaginary" was featured on the WETA public TV web site

V.S Subrahamanian's report entitled, "Cultural Modeling in Real Time" appeared in Science Magazine.

An article in Science magazine on April 27 highlighted **V.S. Subrahmanian's** work on automated, real time methods to model the b ehaviors of foreign cultural groups and terror groups.

ECE and UMIACS Professor **Uzi Vishkin** introduces new "Desktop Supercomputing" prototype capable of computing speeds 100 times faster than current desktops, the technology is based on parallel processing on a single chip. The new desktop supercomputer was featured in Computerworld Magazine.

Dr. Vishkin was also named an "Innovator of the Year" by the Baltimore Daily Record.

Getoor, continued from front cover

resolution, Getoor received a Google Research Award in early 2007 and a Microsoft Gift in May 2007. In addition, she was an invited speaker at the National Conference on Artificial Intelligence held in July 2007 in Vancouver.

Getoor is interested in problems as mathematical challenges, but she also likes to address problems in practical terms. For example, she and her research group share an interest in applying algorithms in machine learning to analyzing social networks. Based on e-mail traffic, they worked to infer the organizational hierarchy in a corporation, testing to see whether they could identify managers and their subordinates based on the flow of e-mails. The work could have practical applications in helping design software to organize and prioritize e-mail, however, the study also served as a test case to ask what problems need to be resolved to yield a trustworthy solution.

Analyzing e-mail traffic flow was fruitful. The study showed that colleagues at the same level in an organization are more likely to e-mail each other than the boss who supervises them. "We may even be able to predict relationships when there is no observed communication between individuals," says Getoor. Even more informative have been subsequent analyses that also analyzed the text within e-mails. Promisingly, Getoor and her colleagues were able to rank how informative messages were to the question at hand. Once the researchers read messages tagged as important by the algorithm, they could see that the messages clearly brought up organizational relationships.

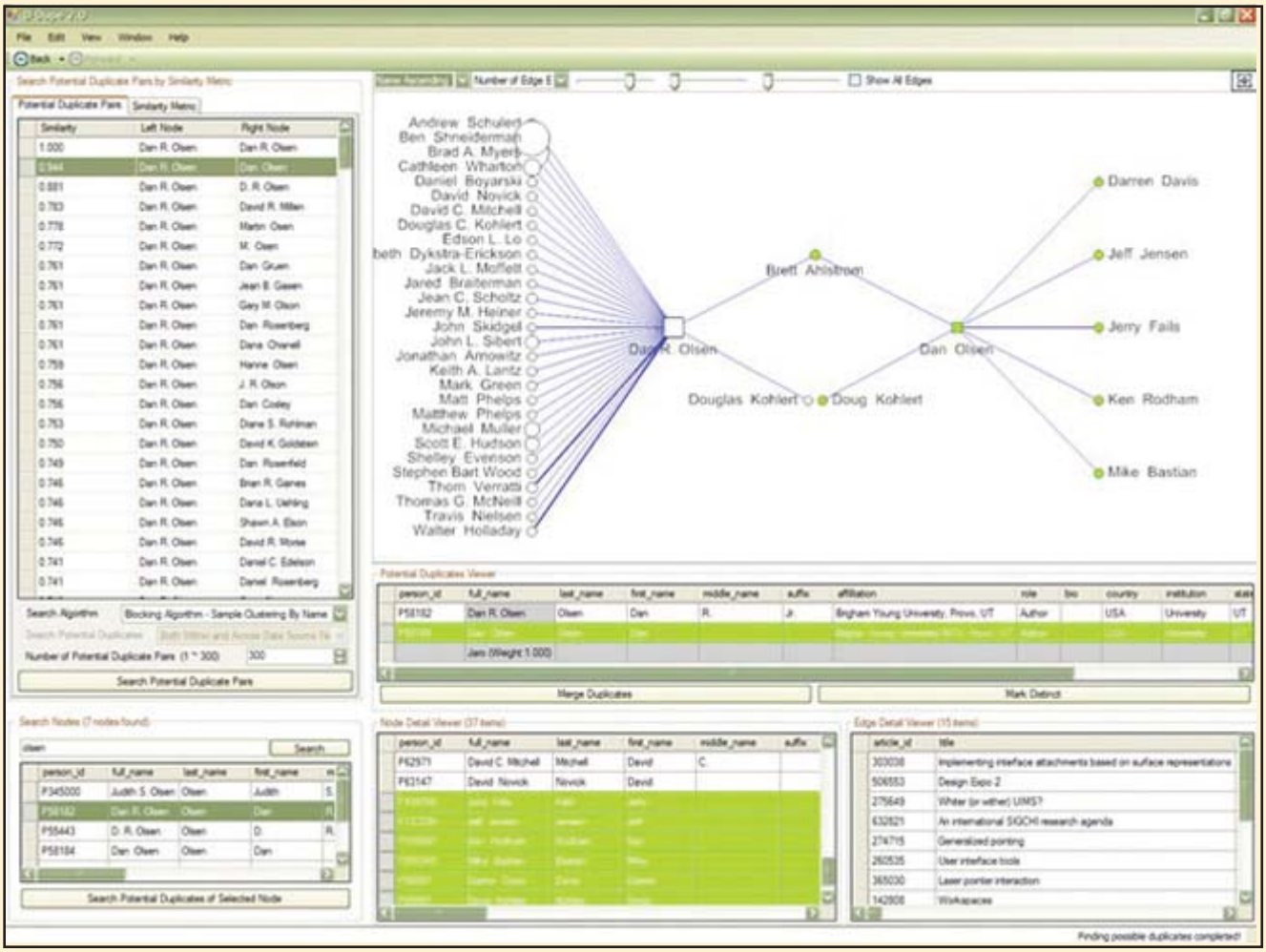
In another example of studying social relationships, Getoor has analyzed how research publications are connected to each other based on their citations and coauthors. Such bibliographic data can identify collaborative groups, a type of social group. Going a step further, Getoor has looked at how social relationships among academics, deduced from evidence such as how professors are grouped in conference organizing committees, affect individuals' social capital and, in turn, publication rates. The same kind of analysis can be used to explore relationships among lawmakers and lobbyists, she notes.

In addition to link mining, one of Getoor's main interests is entity resolution, or developing tools to deduce whether two separate mentions of a person, place, object, or event refer in fact to the same thing. For example, are references to a John Smith and another to a Jonathan Smith talking about the same person? Getoor and her Ph.D. student Indrajit Bhattacharya, who has since graduated and works at IBM Research Delhi, developed new algorithms for entity resolution in social networks. One of their papers won the best paper award at the 2006 Society for Industrial and Applied Mathematics, or SIAM, data mining conference.

Two of Getoor's other graduate students, Mustafa Bilgic and Louis Licamele, together with Hyunmo Kang and Ben Shneiderman of the University of Maryland's Human-Computer Interactions laboratory have developed D-Dupe, a tool for interactive entity resolution. D-Dupe helps users cope with large databases that may potentially contain redundant information by making common relationships easy to identify visually. More information, including a video demonstration and downloadable software, is available at http://www.cs.umd.edu/linqs/ddupe. The demonstration at this web site shows how a user can sort through a bibliographic database to sort among the authors of academic research papers, helping the user eliminate repetitive information and chart relationships among authors. The same kind of analysis can be applied to a wide range of data.

For example, entity resolution is an important problem in geospatial data analysis and one that Getoor is addressing in a collaboration with the U.S. government's National Geospatial-Intelligence Agency, or NGA. Getoor is helping develop ways to map references to the same locations in NGA databases. The agency receives updates from local governments and would benefit from a more efficient way of distinguishing whether new information should be added to an existing entry in a database or merits a new entry. The problem is challenging because it's common for people to use names or coordinates that do not overlap exactly, notes Getoor. Any case of entity resolution requires algorithms that assess the commonalities and distinctions between two records. Clearly, if the attributes and relationships of two entities do not overlap extensively, they are more likely to be distinct units and not duplicates.

"I'm also interested in how all this relates to privacy and what you can say about the privacy you can guarantee," says Getoor. Technology that can help distinguish and identify individuals can provide practical benefits, such as thwarting identity theft, but it can also be used to ensure that identifying particular individuals or relationships remains impossible. The process of

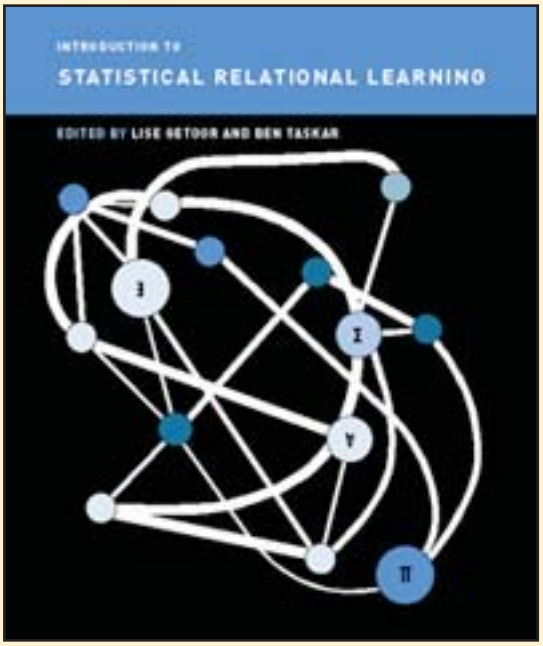


A screen shot from D-Dupe, Getoor's tool for interactive entity resolution. The program helps users sort through databases that may contain redundant information by making common relationships easy to spot graphically. For more information, see http://www.cs.umd.edu/linqs/ddupe.

graph identification could help guarantee the privacy of medical records and other sensitive information, says Getoor.

"She's very creative and is very good at defining problems and coming up with rigorous solutions for them. That's what she has done in a number of cases, and the results speak for themselves in the impact she's making," says Getoor's colleague Chris Diehl.

For her research, Getoor receives funding from the National Science Foundation, the National Geospatial-Intelligence Agency and other government funding sources.



Lise Getoor is currently editing a book on statistical relational learning.

Jonathan Katz: MURI Award Funds Study of Security in Mobile, Ad-Hoc Networks

Mobile ad-hoc networks, or MANETs, link wireless routers that are free to move around in unpredictable ways. MANETs could link everything from laptops in a coffee shop to soldiers in a battlefield or emergency workers in a disaster zone. Mobile networks could help track the movements of animals in a herd or reveal when traffic is congested on roadways. In order to become common, mobile networks will first need to become reliable and secure.

UMIACS member Jonathan Katz, an assistant professor of computer science and of electrical and computer engineering, is on a team that has just received a five-year $6.25 million Multidisciplinary University Research Initiative, or MURI, award from the Department of Defense to study how to build robust and secure MANETs. Katz, an expert in cryptography, will help to construct secure protocols for mobile networks. Virgil Gligor, a University of Maryland professor of electrical and computer engineering, is the principal investigator on the MURI award. The award-winning team unites Maryland researchers with scientists from three other institutions—Carnegie Mellon University, University of Illinois Urbana-Champaign, and University of Washington, Seattle. The project gets underway in August with a kick-off meeting at the University of Maryland.

"For us, this was a great honor," says Gligor. The proposal was selected as the winner from among 56

Jonathan Katz

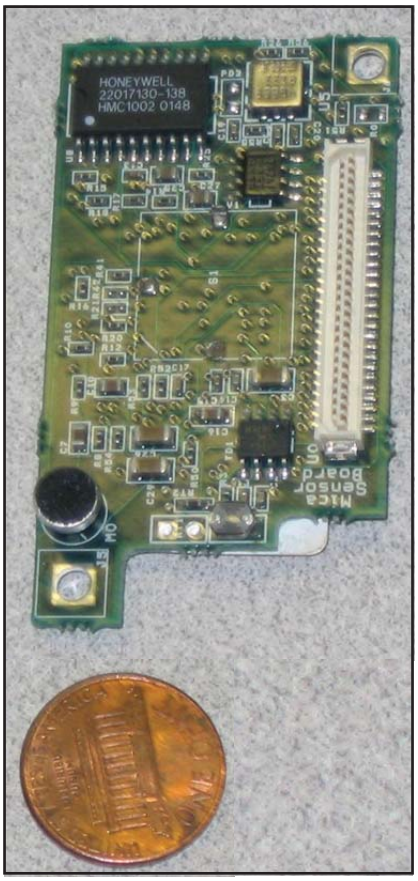
that were submitted. Katz provides important expertise in cryptographic protocols for the team. "He is well-known in cryptography. He is young and very energetic. Obviously he was an asset to the team" in preparing a winning proposal, says Gligor.

Simply creating a viable network from randomly-scattered nodes is a challenge. For example, sensors may be scattered from a plane onto a battlefield and then have to organize themselves efficiently into a network. The individual nodes are often very small and have limited power. Any given node may be destroyed at any time. The nodes must organize a network among themselves and must be able to reorganize continually to deal with failed nodes, Katz explains.

Reliability is a particularly tricky issue given the harsh environments for which sensors are generally envisioned. Beyond battlefields and natural disaster zones, sensors would be exposed to the elements when they're used to track animals or when they are scattered on bridges to monitor for possible stress fractures. Inside buildings, sensors will have to be durable to fulfill expectations that they will be used report fires or chemical hazards.

Sensors by definition have to be able to sense something about the environment around them and must be able to communicate that information. They also serve as routing nodes to help other nodes communicate. A single node on its own may not have enough power to communicate directly with a central reporting point, but collectively they can transmit messages.

"Because you're not dealing with a fixed network but with paths being formed on the fly, it introduces challenges," Katz says. One challenge he deals with is how to communicate securely. Security strategies developed for desktop computers are far too inefficient to be applied to MANETs. The small, mobile sensors have no external sources of power, and each unit must be relatively inexpensive. With the devices' limited capabilities, protocols for them must be extremely efficient. The battery power that runs each sensor node



Small, battery-powered sensors such as this one can be used to send information about their surroundings via mobile, ad-hoc networks. Photo courtesy of Ajay Gupta, Western Michigan University.

must be conserved. "Many cryptographic algorithms require a lot of power, which will drain the battery of these devices rather quickly," Katz says. A second security challenge, obviously in a battlefield but also in many other applications, is that the sensors are distributed in locations where they are unprotected, to say the least. Someone could potentially re-engineer or take over a system and send corrupted information, Katz notes.

Efficient Cryptographic Protocols

In a MANET organized around laptops in, say, a coffee shop, the problem is establishing a way of communicating securely even when parties lack prior knowledge of each other. With MANETs linking small sensors, cryptographic algorithms also have to be energy efficient. To create ways for sensor nodes to prove they are legitimate rather than imposters, Katz is working to develop authentication protocols that are "so simple that even a human could run them," Katz says.

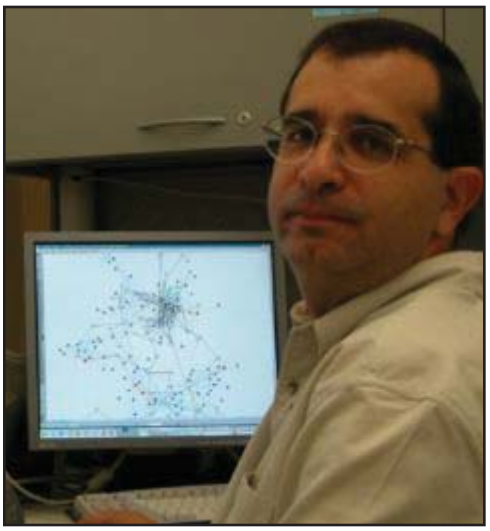
Katz has extensive experience in designing and analyzing tools for encryption and in developing ways to securely transmit information over insecure networks. One of his innovations has been to develop encryption keys that can evolve over time. He and his coworkers were the first to design a system where even if an outsider decodes an encryption key, the intruder will not also be able to decode past keys from that information. Katz also developed the first provably secure system for encryption based on individuals' identities alone, a system that could work well in large, open institutions such as universities.

"In general, I'm interested in things you can prove something about," says Katz. As such, Katz tends to take a formal, mathematical approach to security problems. For Katz, designing secure protocols goes hand in hand with proving they're secure.

With MANETs, ultimately, the goal is to transmit information from distributed sensors to a base station. Two possible barriers could impede the transmission of information: the loss of nodes could lead to lost data, or an adversary could interfere with data transmission, possibly even using kidnapped nodes to send erroneous information. Secure routing focuses on transmitting data despite occasional equipment failure among nodes. Secure data transmission makes sure trustworthy data reaches its destination unchanged.

The MURI proposal Katz and his colleagues wrote suggests designing a system with a core set of nodes that can be trusted. Katz will work to come up with how to define and achieve such a core. He will also come up with definitions for modeling threats to MANETs. Along with everyone else collaborating in the project, he will also be involved in experiments implementing the ideas the group develops.

Najib El-Sayed: Genome Studies of Neglected Diseases



Najib El-Sayed

Trypanosomes are single-celled parasites transmitted by insects that infect plants and vertebrates, including humans. In humans, trypanosomes cause three distinct diseases, mostly in warmer parts of the world. Najib El-Sayed, who joined UMIACS's Center for Bioinformatics and Computational Biology (CBCB) last November, is one of the foremost experts of trypanosome biology. Notably, he led a number of studies published in the journal Science in July 2005 reporting the genome sequences for the major disease-causing trypanosomes.

"At a number of levels, Najib has made a significant contribution to the science of these important pathogens," says Michael Gottlieb, formerly the chief of parasitology and international programs at the National Institute of Allergy and Infectious Diseases, which continues to fund El-Sayed's research. The sequences and their analysis "have been a tremendous achievement and resource for the community," Gottlieb adds. "This project required a fair amount of coordination. Najib offered not only scientific expertise in sequencing and bioinformatics and a strong background in trypanosome biology. He also has a real talent for diplomacy and coordination."

El-Sayed, who has a Ph.D. in molecular parasitology, started studying trypanosomes 22 years ago. He joined CBCB and the University of Maryland department of cell biology and molecular genetics after eight years at The Institute for Genomic Research in Rockville, Md. Trypanosomes have been the subject of research interest for a long time, partly because of unusual aspects of their molecular biology. "Many dogmas have been broken by these parasites," says El-Sayed. "They have interesting biology as well as human impact."

The human impact is undeniably great. Different species of trypanosomes cause African sleeping sickness in sub-Saharan Africa, Chagas disease in Central and South America, and leishmaniasis in tropical and subtropical regions around the world. Collectively, the diseases infect an estimated 100 million people each year and cause many levels of hardship. The diseases kill more than 100,000 people annually and also lead to a great deal of disability and disfigurement. Because African sleeping sickness infects livestock as well as people, the trypanosomes also inflict economic damage.

El-Sayed's research group sequenced half of the genome of the African trypanosome, Trypanosoma brucei, with the other half sequenced by collaborators in the Sanger Center in the United Kingdom. El-Sayed's group also sequenced most of the genome for T. cruzi, which causes Chagas disease, and his team led the comparative analysis of the three major trypanosomes' genomes.

Toward Rational Drug Design

Besides helping answer some long-standing biological puzzles, the comparative genome analysis has been an important step toward developing medicines for these neglected diseases. "If we see shared pathways that differ from those in humans, we can start thinking about the rational design of drugs that target all three diseases," El-Sayed explains. Finding a drug that could treat all three trypanosome diseases would make them a more attractive target for pharmaceutical companies. Indeed, El-Sayed and his coworkers found that one-third of the genes shared among all three species are also unique to these parasites.

By comparing the genome maps for all three organisms, El-Sayed found that they not only share a large number of genes, they often share genes in the same order on their chromosomes. The three trypanosomes are estimated to have diverged in evolution about 200 million years ago, but there are far more rearrangements in the chromosomes of vertebrate species that diverged at a similar time.

Among the unusual biological features of trypanosomes, the species that causes African sleeping sickness has long been known to have a dense surface coat consisting of molecules that turn over regularly. The preprogrammed variation of the surface helps the parasites evade the immune systems of infected animals. "Now that we've sequenced the genome, we know that 10 to 15 percent of genome resources are devoted to antigenic variation," says El-Sayed. Depending on the strain, 1000 to 1500 genes code for variant coat proteins, with only one coat protein being expressed at any given time. Since sequencing the genome of the agent causing Chagas disease, El-Sayed and his team have

also discovered what looks to be a family of surface proteins in that trypanosome.

Because of the great variation of trypanosomes' surface coats, at least in T. brucei, the organisms have not been a good target for vaccines. One promising alternative target, El-Sayed says, is the process for making the proteins that anchor the cells' surface proteins. The enzymes that make these anchors are unique to the parasites. "If we could interfere with these, we could find an Achilles heel for these organisms," El-Sayed says.

Trypanosomes have also interested biologists because they have served as a good model for studying RNA editing. Like humans and unlike yeast, the single-celled eukaryote most commonly studied in biology labs, trypanosomes edit their RNA after it is copied from DNA. Researchers can engineer gene expression in trypanosomes, so the organisms have served as a good model for studying RNA editing machinery.

Biologists have also been intrigued by trypanosomes' unusual ways of expressing their genes. Instead of transcribing one gene at a time onto RNA, trypanosomes can copy over 200 genes onto a single strand of RNA. After sequencing the organisms' genomes, El-Sayed and his collaborators could see that the genes transcribed together do not appear to be related by function. El-Sayed is working to find the DNA regions where RNA transcription begins. He is also trying to figure out how trypanosomes regulate the expression of individual genes. The African trypanosome genome sequence revealed that genes encoding the variant surface proteins exist in large clusters on chromosomes but need to be copied and moved to the ends of chromosomes in order to be expressed, raising questions about how exactly all this happens. "This is a typical story of a genome project where you start with a hundred questions and end with six hundred," says El-Sayed.

Having the genome sequences in hand is also helping El-Sayed and others probe how trypanosomes differ. The single-celled creatures look alike and appear to share many characteristics at the level of genes and proteins yet cause distinct diseases, infecting different cells, attacking different organs and causing different symptoms.Parasitologists are working to dissect how the genes among the organisms differ as well as how they are similar. About two-thirds of the genes in each organism have unknown functions, with their sequences not matching those of any previously studied genes. To try to figure out the roles of the proteins these genes encode, El-Sayed and his research group are conducting large-scale studies to find interactions among the proteins. They are starting by looking for interactions among the proteins within each trypanosome. To systematically study the thousands of genes with unknown functions, El-Sayed has set up a robotic system in his lab to help isolate and examine the genes and the proteins they encode. Eventually, El-Sayed says, he wants to study interactions among proteins of trypanosomes and humans to shed light on how parasites infect host cells.

Trypanosoma brucei, which causes African sleeping sickness, is shown in a thin film of blood. Courtesy of WHO/TDR/Stammers.

El-Sayed is also studying the class of surface proteins he discovered in the trypanosome that causes Chagas disease. Even if the surface proteins cycle like in the African trypanosome, "anything on the surface is a possible target for intervention," El-Sayed says. He and his team are seeking to answer a number of questions about these previously unknown proteins, among them: Do all parasites express the same surface proteins when they're in humans? Do different surface proteins direct the trypanosomes to infect different cell types? Do the surface proteins vary over the course of the trypanosome's life cycle?

While El-Sayed and his research group continue intensive studies on trypanosomes, he is also seeking to expand the labs' direction into a field called metagenomics. In the past, genome studies have focused on a single species at a time, but metagenomics seeks to study the DNA sequences of whole communities of microorganisms at once. For example, other researchers have applied such studies to microbial communities in hydrothermal vents in the ocean. "The potential of the method is vast," says El-Sayed. "We will be able to address questions that we wouldn't otherwise be able to address." The technique can be applied to complex, mixed populations and organisms that researchers don't know how to raise in the laboratory.

continued on page 2

Varshney, continued from back cover

Amitabh Varshney

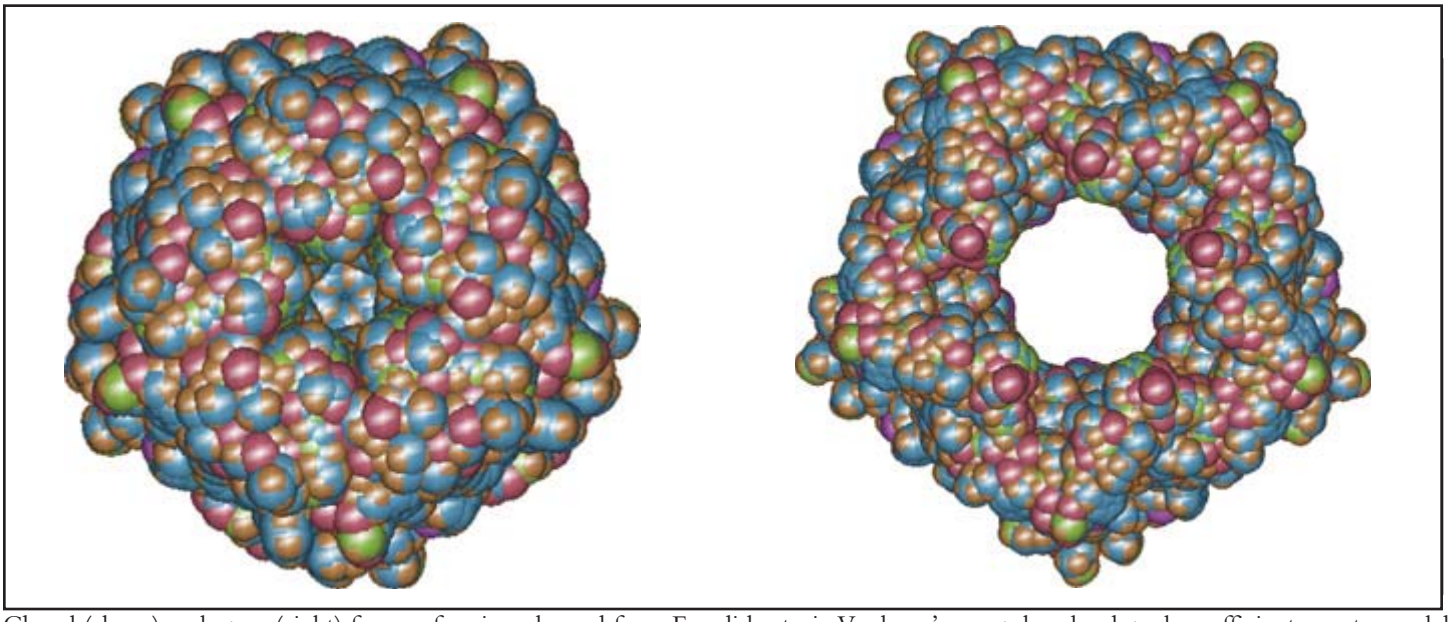
Graphics processing units, or GPUs, are highly parallel processors developed for video games.

"Three years ago, we went to NSF and said we want to explore applications for these cheap GPUs," says Amitabh Varshney, a professor of computer science and member of UMIACS. Since then, with funding from the National Science Foundation and the University of Maryland, Varshney has been exploring applications for GPUs and the cases where they can be more useful than conventional CPUs, or central processing units. "In general, CPUs are better. We are asking what applications would GPUs be better for," says Varshney. GPUs are especially good at performing the same operation repeatedly on large streams of data. GPUs appear particularly well-suited, for example, to applications in audio or video surveillance and computational biology. Two graduate students taking a class with Varshney—Michael Schatz and Cole Trapnell—showed that GPUs can be 10 times faster than CPUs in mapping genes in large stretches of DNA sequence data.

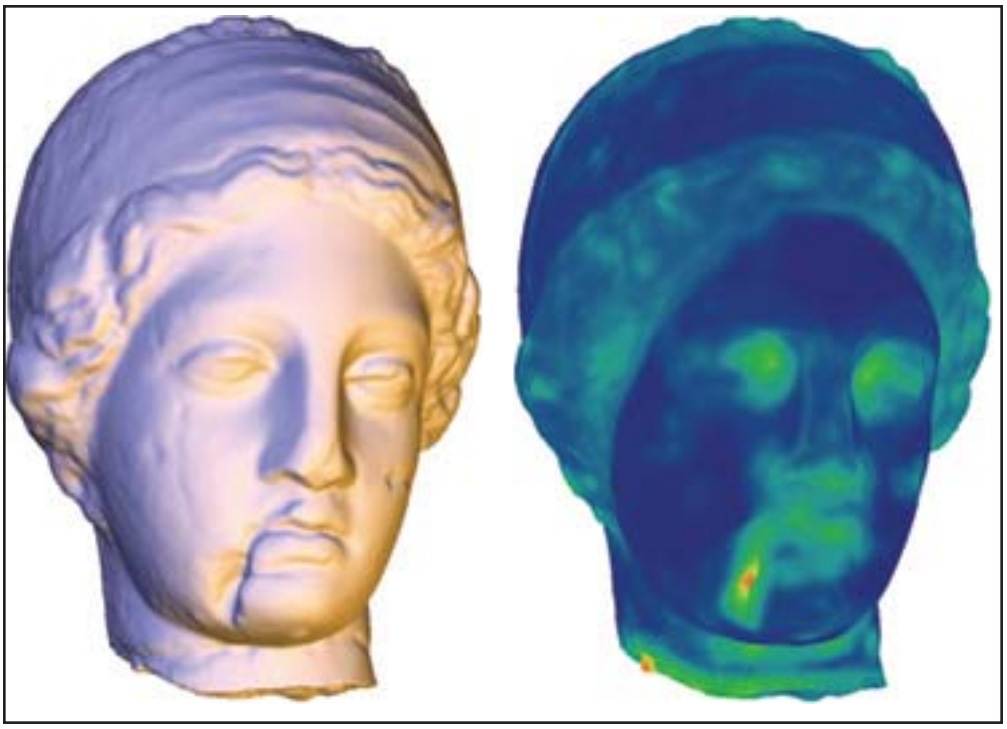
CPUs are indispensable for solving problems without well-defined structures, but GPUs work better for problems that involve applying the same or similar computational codes repeatedly on large streams of data, Varshney says. One example of the latter is modeling the mixing of two gases over time. "Most problems have a regular and structured component and an irregular and unstructured component," notes Varshney. CPUs can be used on irregular and unstructured aspects of problems, and GPUs on regular and structured aspects. CPUs and GPUs can be used in combination, explains Varshney.

One of many problems Varshney and his team have worked on is presenting a way to visualize the mixing of gases. His group collaborated with researchers at the Lawrence Livermore National Laboratory, who studied the physics of gas mixing, and with Joseph JaJa, a member of UMIACS who also works on scientific visualization. Varshney's group helped process the large amounts of data involved efficiently and worked to display the information in a vivid and informative way.

Videos, such as those recorded by surveillance cameras, are another source of streaming data. In collaboration with UMIACS researchers Rama Chellappa and Larry Davis of the Computer Vision Lab, Varshney has been working to analyze videos by dividing them into meaningful segments, which GPUs can



Closed (above) and open (right) forms of an ion channel from E. coli bacteria. Varshney's group has developed an efficient way to model the behavior of the more than 10,000 atoms in the channel protein.



Varshney and collaborators have shown that the human eye is drawn to areas of high curvature in an object. For the sculpture shown at left, the graphic at right shows where people are less (blue) and more (green, yellow, red) likely to look.

then process efficiently.

One of Varshney's interests is in mathematically representing complicated volumes. Equations that represent irregular three-dimensional shapes can be very long and complex, but a formulation known as implicit representation can describe shapes in a relatively compact way. Succinctly representing the large data sets that define shapes could be invaluable in creating a wide range of applications that require real-time responsiveness. One such example is image-guided surgery. Taking data from imaging techniques such as ultrasound or magnetic resonance imaging, implicit representation could help process information quickly enough and in enough detail to create interactive applications for, for example, needle-guided biopsy surgery. In addition to applications in medicine, implicit representation is already used in design and entertainment.

Existing implicit representations focus on surfaces, but Varshney is crafting equations that represent volumes, so for example in a medical application, a user would see not just the surface of a body but be able to explore within it. For another example, volume representations of atoms would show not just electron shells but electron densities. By representing density, one can better represent internal structures.

"Dr. Varshney is extending implicit shape representation beyond surface representations and trying to capture the notion of implicit solids. This is a bold step. It should be especially useful for representations in medicine and biology," comments Terry Yoo, a computer scientist at the National Library of Medicine.

A large aspect of visualizing complex data involves being able to display the information. To create a very high resolution display, Varshney has tiled 25 LCD displays in his laboratory, creating a composite display with 50 million pixels. On this merged display, he can show, for example, the approximately 10,000 atoms of an ion channel protein from the bacterium E. coli and how each atom might move when the channel opens and closes. Varshney is collaborating on this project with Sergei Sukharev, an associate professor of biology at the university. "You can see small-scale and large-scale motion at once," Varshney says. "We're trying to allow real-time exploration, so scientists can examine a lot of what-if scenarios. Interactivity changes how scientists approach problems."

In collaboration with UMIACS members David Jacobs and Francois Guimbretiere, Varshney is also studying how people look at things. Tracking eyes, the researchers have shown that people tend to look most at areas where curvature changes abruptly in a picture. This sort of research fits into Varshney's overall goal of better modeling data for human use. "If you can model how people look at things, you can represent things better for comprehension," he concludes.

Amitabh Varshney:
Fast, Responsive Models for Large Amounts of Data

Tiled LCD displays help Varshney and his students look at images in very high resolution. Here, graduate students examine a simulation of how two gases mix.

It's a perennial problem: how to represent vast amounts of data in ways that help users understand them. To help users interact with data, models should be responsive to user input, but responding quickly can be a challenge when models represent large amounts of information. Fortunately, a large market for video games has helped fund technology for speedy information processing.

continued on page 10

UNIVERSITY OF MARYLAND
Institute for Advanced Computer Studies
2119 A.V. Williams Bldg.
University of Maryland
College Park, MD 20742-3311

Nonprofit Org.
U.S. Postage
PAID
ICM