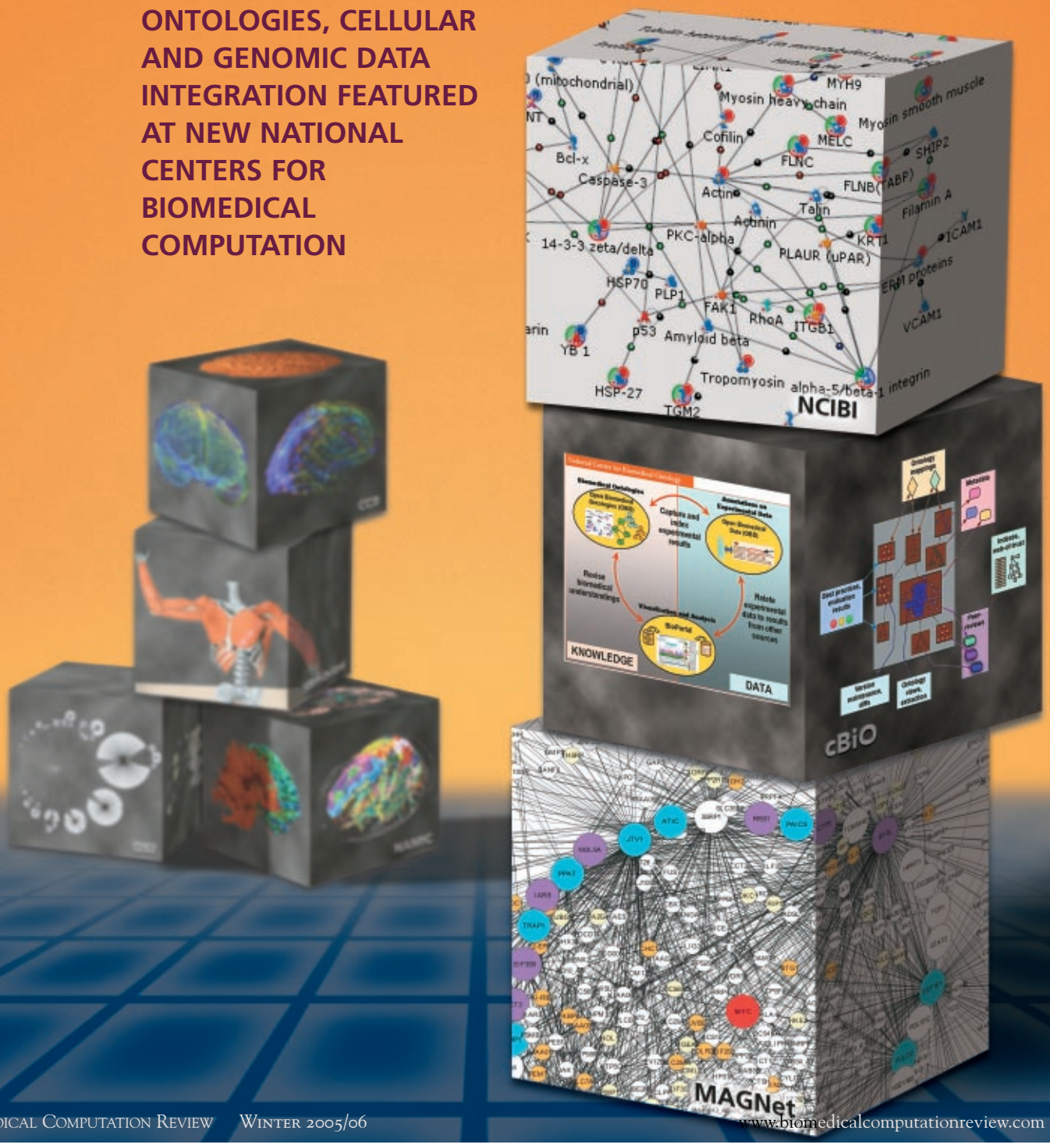
Three New CENTERS

BY KATHARINE MILLER WITH AN INTRODUCTION BY JOHN WHITMARSH, PhD

ONTOLOGIES, CELLULAR AND GENOMIC DATA INTEGRATION FEATURED AT NEW NATIONAL CENTERS FOR BIOMEDICAL COMPUTATION



The NIH Roadmap for Medical Research Increases Support for Biomedical Computing

JOHN WHITMARSH, PhD
ACTING DIRECTOR,
CENTER FOR BIOINFORMATICS AND COMPUTATIONAL BIOLOGY,
NATIONAL INSTITUTE OF GENERAL MEDICAL SCIENCES, NATIONAL INSTITUTES OF HEALTH

The National Institutes of Health Roadmap for Medical Research has recently completed the first stage of an ambitious program to expand the computational infrastructure and software tools needed to advance biomedical, behavioral and clinical research. At the core of this effort are seven National Centers for Biomedical Computing. The centers, each funded at nearly $20 million over five years, are part of a coordinated effort to build the computational framework and resources that researchers need to gather and analyze the massive amounts of biomedical data currently being generated by their labs and clinics. This infrastructure will help the research community translate their data into knowledge that ultimately improves human health.

Dynamic partnerships among many different types of researchers form the core of the effort. Those involved include computer scientists, who invent and develop efficient and powerful languages, data structures, software architectures, hardware and algorithms; computational biologists, mathematicians and physical scientists, who adapt and deploy computational resources and mathematical formalisms; and experimental, clinical and behavioral researchers, who work on problems that can be transformed by computational biology.

To ensure that the products and infrastructure created by the centers truly serve the needs of the biomedical community, each center has identified biological projects to drive the computational efforts. This approach puts leading bench scientists and clinical researchers side-by-side with computational and quantitative scientists to develop easy-to-use software programs that address research needs across disciplines.

In a major effort to further expand the breadth of these multidisciplinary centers, NIH recently announced a Roadmap-related program for Collaboration with National Centers for Biomedical Computing (PAR-05-063). The announcement invites applications from individuals or groups of investigators to work with one or more of the centers on projects that broaden the national centers' biological and computational expertise. The overarching goal of these collaborations is to provide biomedical, behavioral, and clinical researchers access to cutting edge expertise and technology in computer science, as well as offer an opportunity for individuals and teams with computational expertise to help build a vital biomedical computing environment.

But there's a major challenge: How do we get the new computational tools and resources into the hands of researchers who will use and test them? The National Centers for Biomedical Computing have developed plans for sharing and distributing software, resources, and data. For example, the centers will ensure the interoperability of software and data that's accessible to the entire biomedical community. In addition to disseminating computational tools and resources, the centers actively engage in educating and training researchers. These programs support NIH efforts to develop a new generation of multidisciplinary biomedical computing scientists.

The seven National Centers for Biomedical Computing are key players in the integrated Roadmap vision to deepen our understanding of biology, stimulate interdisciplinary research, and accelerate medical discovery to improve people's health. >

"The seven National Centers for Biomedical Computing are key players in the integrated Roadmap vision to deepen our understanding of biology, stimulate interdisciplinary research, and accelerate medical discovery to improve people's health."

—John Whitmarsh, NIH

Interviews with the principal investigators of

The Three New National Centers for Biomedical Computing

The National Center for Biomedical Ontology (cBiO) at Stanford University

According to researchers at the National Center for Biomedical Ontology, ontologies are a very basic piece of the biomedical computing infrastructure because they promote meaningful use and re-use of the biomedical data that researchers are generating in ever-greater quantities. Ontologies convey the biomedical meaning of experiments in a computer-accessible format, and they permit integrating data and knowledge from many sources. Until now, the biomedical community has produced ontologies at the grassroots level, with little coordination or peer review to ensure consistency and quality. The Center will address that problem by creating three open-access resources: an online library of open-content ontologies and terminologies called Open Biomedical Ontologies (OBO); a database resource called Open Biomedical Data (OBD) where scientists can create appropriate data annotations for experimental data using OBO ontologies and terminologies; and Web-based access to these resources via a system called BioPortal.

We spoke with the center's principal investigator, **Mark Musen, MD, PhD**, who is a professor of medicine at Stanford University School of Medicine.

Q: What are ontologies and why do they matter to biomedicine?

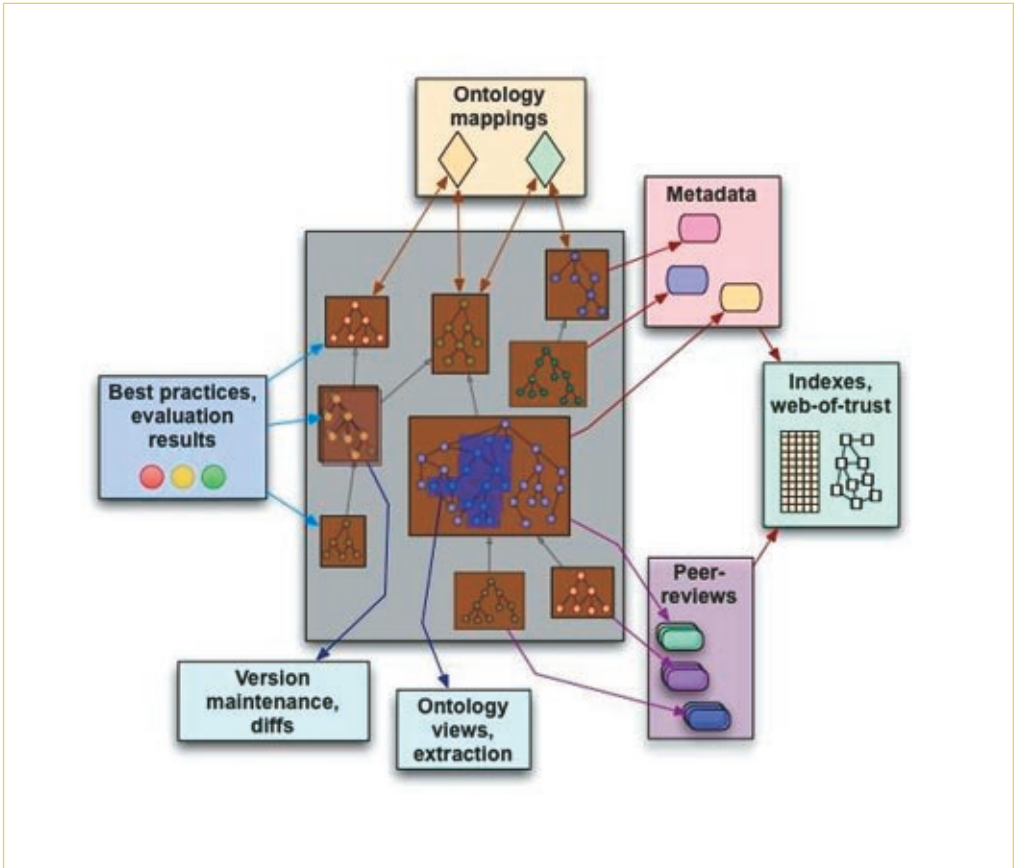
Musen: Ontologies provide a way to capture the structure of knowledge. Aristotle was probably the first philosopher to start thinking critically about the categories of things that exist in the world. And although ontology has traditionally been the province of people who study metaphysics, it has turned out that creating a standard way to refer to something, and being able to enumerate the entities that matter in a particular endeavor are extremely valuable in biomedicine as well as in a number of other fields.

The kinds of applications that tell Amazon what books to recommend to you would not be possible without ontologies. The whole notion of electronic commerce depends on ontologies that define services on the Internet so that a user can find an appropriate match. Ontologies have become big business and are very important outside

of biomedicine. And in biomedicine, they're incredibly important, particularly as biologists have generated so much data and are now going back and trying to figure out what to do with it all.

Q: How do ontologies help us grapple with the vast quantities of biomedical data now being created?

Musen: Ontologies are a way of representing knowledge so that it is accessible to machines for processing. We sometimes talk about "lite" ontologies that are really nothing more than taxonomies—classifications without much more information. What becomes interesting is that when you start adding definitions in computer-understandable form and start defining relationships among concepts, you've also added an element of richness that allows a computer to reason about the concepts and to understand relationships that would otherwise not be intuitive. And with the quantities of biomedical data now being generated, this is precisely what we need computers to do—to be able to evaluate complex relationships between, for example, phenotypes of mutant model organisms and human homologs that are associated with diseases.



A graphical display showing some of the cBiO activities that will become part of the web-based BioPortal. In the center is a repository of the various ontologies contributed to the Center, some of them interlinked and related to one another. Outside the ontology collection are the value-added software components and metadata that will be created in support of the center's goals: a repository of mappings between ontologies; collections of best practices; evaluation metadata that describe the conformance of individual ontologies with the best practices; a version-maintenance mechanism; a repository of views extracted from large ontologies; peer reviews of ontologies and metadata with provenance, usage information, and quality metrics that contributed to various indexes and to the web-of-trust.

Q: Why is there a need for a national center for biomedical ontologies?

Musen: When you look at how ontologies are being created and deployed currently, what you see is a tremendous cottage industry. You see groups of professionals in biology and medicine working in relative isolation, sometimes without knowledge of standards for knowledge representation or standards for best practices for representing concepts, and without the tools that make it possible for large communities to work together on building and evaluating ontologies. Right now, no one really knows what ontologies are out there, which ones are good, and whether problems exist with some of the most commonly used ontologies.

What our center hopes to do more than anything else is to move ontologies from this cottage industry to the industrial age, both by making them accessible and by facilitating the development, evaluation, indexing and peer review of ontologies. These are the kinds of things that we associate with the publication and distribution of knowledge when it's in textual form, but such practices haven't been carried over to the publication and distribution of knowledge that is stored electronically in the form of ontologies.

Q: How do you convince people who are creating ontologies that they should make use of the resources you will provide?

Musen: We have to offer something that they can't achieve any other way. We are all happy to search Web

Mark Musen, MD, PhD, principal investigator of cBiO and professor of medicine at Stanford University School of Medicine.

pages using the Yahoo ontology because it allows us to do things we couldn't do in a blind fashion. And we're all happy to use the ontologies that categorize things in the Yellow Pages because it's a resource that's there and it's useful. So our center has to produce tools that are helpful to people, and then they'll use them. In addition, an important part of all the NCBCs is outreach. We'll have workshops that we'll host in association with scientists and biomedical investigators to help them with the ontologies they are building and to teach best practices. At the same time, we will learn what their needs are. We construe the process as a cycle, with outreach being part of that cycle because it brings in new ontologies and ontology modifications that will be needed to drive the next set of experiments.

Q: Does the ontology cottage industry interfere with the prospect of comparing data from different data sets?

Musen: Yes it does. To be able to relate one gene in one database with the same gene in another database and to feel confident that they are the same, scientists must annotate databases using the same terms. You need ontologies to be sure that you're comparing apples to apples. Right now, there's no requirement to use ontologies when annotating data. Using an ontology forces consistency, which allows you then to do queries and to feel confident that you're getting back everything that you should.

There is also a lot of work to be done to make sure that ontologies make the kinds of distinctions that are most relevant and important to investigators. The numbers of distinctions you can make in the world is infinite. A national center can be helpful in reaching consensus about which distinctions are the right ones to make in order for ontologies to be most useful.

"You need ontologies to be sure that you're comparing apples to apples."—**Mark Musen, cBiO**

The National Center for Mulitscale Analysis of Genomic and Cellular Networks (MAGNet), at Columbia University

MAGNet's goal is to understand how all the genes and proteins inside cells interact to implement specific biological processes over a wide range of scales. This challenge will be tackled using a combination of structural and systems biology approaches.

The center will use a variety of algorithms and databases to anchor molecular interaction clues within an integrated genomics framework. Clues will be principally drawn from four sources: sequence and structure analysis; novel reverse engineering algorithms; literature datamining; and experimental results stored in a variety of existing databases. Clues about individual protein-protein or protein-DNA interactions will be further associated with a specific organism, tissue, cell differentiation stage, and cellular phenotype (i.e., normal vs. disease). Gathering and integrating all these clues will

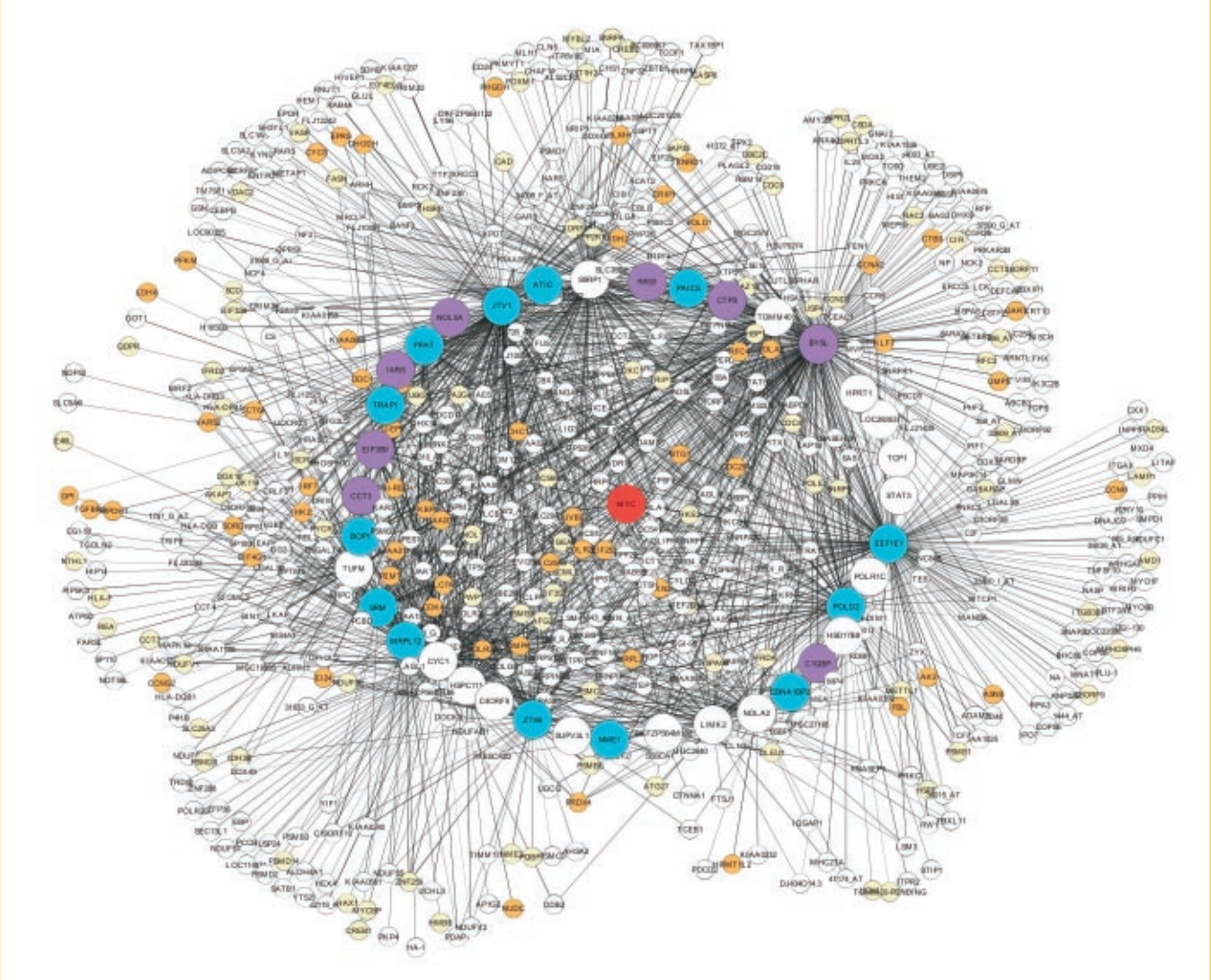
require the development, integration and use of many computational methods and frameworks, including natural language processing, machine learning, information theory, and ontology-based representation models. The center's tools will be made available to the biomedical research community as components of a software platform called the Genomic Workbench (geWorkbench).

We spoke with **Andrea Califano, PhD**, principal investigator for MAGNet and professor of biomedical informatics at Columbia University.

Q: You use the terms "evidence" and "clues" a lot, as though you're trying to solve a big mystery, which I suppose you are. What do you mean by these terms in the biomedical context?

Califano: There is nothing more mysterious and intriguing than the processes that make a cell work. And, unfortunately, these are so complex and daunting that they cannot be tackled using a single type of data or methodology.

The idea of integrating several distinct clues together comes from the field of visual recognition. When we interpret a visual scene, we really process and integrate many different kinds of visual clues at once, such as texture, edge, color, shape, and motion. Each could be interpreted independently, but that's not how



A reverse-engineered sub-network of co-regulated genes derived from gene expression profiles in human B-cells using the ARACNE algorithm. The proto-oncogene MYC appears at the center of the hub. MAGNet will integrate such reverse engineering tools as part of its Genomic Workbench (geWorkbench).

Andrea Califano, PhD, principal investigator for MAGNet and professor of biomedical informatics at Columbia University.

we do it because individual clues are obviously less informative than the whole. Thus, our brain has developed strategies to process information from multiple clues at once.

In biology, it's important to integrate information in a similar way. As an example, consider the number and variety of clues that can help determine where on the DNA molecule a particular transcription factor will bind to activate or repress the expression of a gene. One clue might be the short, specific DNA sequence (4-6 base pairs) to which a particular transcription factor binds. But such sequences recur in the genome with great frequency. So, based on this one clue, you'll get a long list of potential binding sites. To shorten that list, you need to exploit additional clues, such as the fact that similar regulatory sequences in different organisms may have the same protein binding properties, or that genes with similar regulatory regions will be likely co-expressed in the cell, or that the binding sites tend to self-organize into functional modules. By integrating these low-dimensional biological clues, you are truly combining different bits of information to solve your biological mystery, and the statistical significance of your prediction can be significantly increased.

Q: Your Center will be working on three biological problems in which the computational approaches are interwoven with experimental approaches. Is that part of your thinking about how biomedical computation will become most useful in biomedical research—as an integral part of experimental biology?

Califano: The interplay between computation and experimentation is exactly where our center wants to be. Our goal is to predict key interactions in the cell using structure prediction and systems biology methods and then to validate them using biochemical approaches. My personal opinion is that, in the future, the only successful way to investigate biological processes will be by combining computational and experimental methods—either by forming tight collaborations among dry and wet labs or by doing both in a single lab. These days, biologists increasingly seek a truly engaging and equal-footed interaction with computational scientists, and an increasing number of labs are closely coupling experimental and theoretical work. For instance, I have a great deal of respect for people such as George Church at Harvard, Jim Collins at Boston University, or Stan Leibler at Rockefeller, just to mention a few, who have managed to successfully tackle both aspects of this research in their own labs.

"The interplay between computation and experimentation is exactly where our center wants to be."

—Andrea Califano, MAGNet

Q: What is your model for developing MAGNet's software platform, geWorkbench (Genomic Workbench) so that it can be used by bench scientists as well as computational biologists?

Califano: The foundation of our NCBC software platform is caWorkbench (Cancer Workbench), which is truly a product of the Cancer Biomedical Informatics Grid (caBIG) initiative funded by the NIH/NCI. In promoting caWorkbench, caBIG has really strived to create a model such that all cancer researchers can exploit each others' software and data within a research grid that links all the cancer centers.

As Columbia University's representative on caBIG, I was involved in caWorkbench, starting in 2003. The primary goal was to create a software platform where bioinformatics tools could inter-operate freely and be assembled into complex workflows. Initially it was all about microarray expression profile data. Today, caWorkbench works with a variety of data from different modalities—eg., gene expression, sequence, SNPs, pathways, etc.—allowing the creation of complex data analysis pipelines without any need to write code every time, thanks to a common interoperability model called BISON (Biomedical Informatics Structured Ontology). The caWorkbench was also designed so that our biomedical colleagues can build pipelines and workflows themselves, without a computational biologist's intervention. MAGNet will extend the cancer workbench model (caWorkbench) into a genomic workbench (geWorkbench) to allow the same type of approach outside the cancer area.

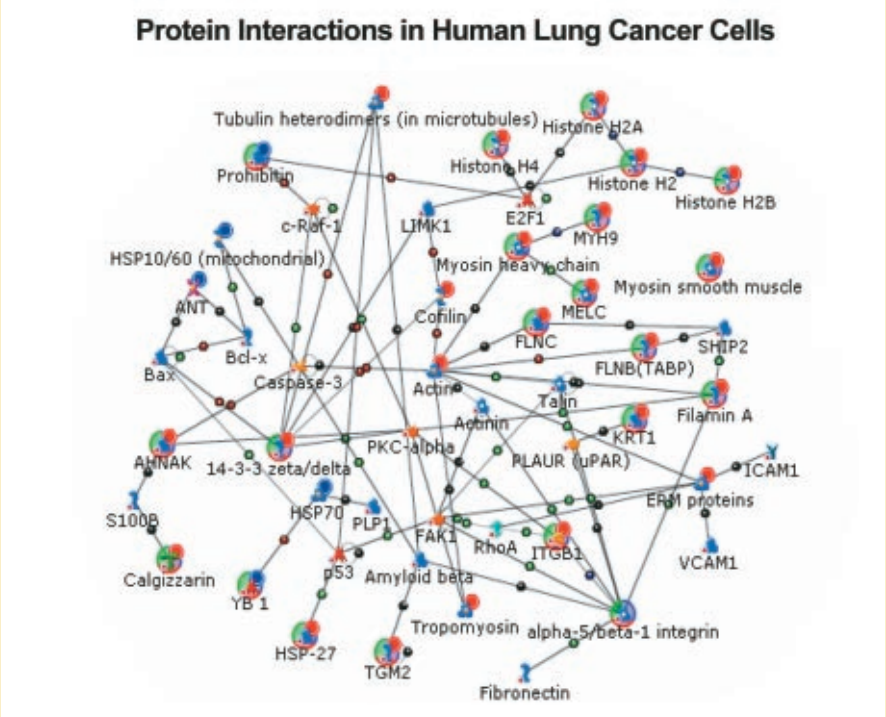
Q: What is MAGNet's biggest challenge going forward?

Califano: Our biggest challenge is the ability to make the biomedical research community aware of the vast array of resources that we hope to create and to make these accessible to people with a strong biomedical training but relatively limited computational expertise. We plan to spend a very significant amount of effort and resources to reach out to the community to make sure that what we build is useful to other scientists in the lab.

The National Center for Integrative Biomedical Informatics (NCIBI), at the University of Michigan

The goal of NCIBI is to facilitate the efficient scientific exploration of complex disease processes on a much larger scale than is currently feasible. The center will deeply integrate data from multiple sources including emerging experimental data, international genomic databases, models, and the published literature. The user will then be presented with a single unified information stream that can be subjected to queries. Such capabilities will allow researchers to rapidly and reliably proceed through what is ordinarily a lengthy and laborious process of developing a testable model: from exploratory data analysis through model formulation, evaluation, revision and refinement. A flexible query interface will allow researchers to select the data and literature resources they're interested in and to iteratively construct, refine or alter complex queries of those resources as well as repeat the same analyses in many different contexts.

We spoke with the NCIBI's principal investigator, **Brian D. Athey, PhD**, who is also an associate professor of biomedical informatics at the University of Michigan and director of the Michigan Center for Biological Information.



Interactions between differentially expressed proteins in human lung cancer. Courtesy of V.G. Keshamouni.

Q: NCIBI deals with what you call deep data integration. What is that and why is it valuable?

Athey: Deep data integration allows researchers to make complex queries of multiple sources and types of data. In a normal data integration situation, you see success linking information together, but it doesn't scale out because it doesn't have full text literature search capability as well as modeling and simulation capability, which is what we are proposing to add. So, for example, a normal data integration exercise would utilize existing databases to ask about expression mRNAs in prostate cancer, and link these elements to other data from the literature. That's an example of what can be done already, and is currently leading to novel results. With deep data integration you might be able to ask a more complex question such as "Do androgen-responsive promoter elements of TMPRSS2 and ETS mediate the overexpression of ETS transcription factor genes in prostate cancer?" This could lead to an answer like "Yes, it is observed that there is a recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer."

It's an answer that's not conceivable under current approaches. Deep data integration would let researchers do something much more structured and knowledge-based than is currently available.

Brian Athey, PhD, principal investigator for NCIBI and associate professor of biomedical informatics at the University of Michigan and director of the Michigan Center for Biological Information.

Q: Ultimately, will researchers out in the world have access to the NCIBI's deep data integration tools or are you really designing tools for people who are collaborating with you?

Athey: We absolutely have to make our tools available to researchers in the real world of NIH-funded researchers. If we can't get to that level, we're going to fail. We're creating what we call a "problem posing architecture" at the front end so that users can have the resources of our center accessible. So we're user- and problem-centric. This can't be all hand-curated or specialized. It must be done by computer. Otherwise we wouldn't be able to scale it up for general use. In the first three years, we'll be testing our tools on bipolar disorder, prostate cancer, and type-1 and type-2 diabetes. Scalability will mean bringing in other researchers first in those diseases and then later in other complex diseases that could benefit from our modeling, data integration and literature search capabilities.

Scalability is, in fact, one of the biggest challenges faced by all of the national centers. When we do computational biology and informatics research, every problem and researcher has a different focus and intention. The challenge is to have our capability usable by people at all levels—from graduate students to experts.

Q: Do you view the NCIBI, to some extent, as a sociological experiment in interdisciplinary problem-solving?

Athey: It's something we've been working on here for some time—to acknowledge the sociological aspects of working together in large interdisciplinary teams. The technical problems are often only a fraction of the solution. The rest is collaborative. We're acknowledging this by studying these human factors to improve communication and improve outcomes.

The School of Information at the University of Michigan has developed an interdisciplinary approach to studying how collaboration occurs. We've learned from that.

I've been involved in a lot of interdisciplinary projects, and there is always a great deal of hope and promise. But the scale of these large projects, the different disciplines and environments involved, and the terms and conditions of sponsorship, all lead to special challenges. If you understand the challenges you can address them. We're trying to boil down some of the lessons learned, so that we can interact with researchers in ways that will seem natural to them and will encourage them to be creative and work collaboratively with us and others on these new capabilities.

One of the things we've learned here is that it's very difficult to get researchers to try new things. You might think you could give researchers free software or Web-based tools and they'd try them, but researchers are often so busy and committed to what they're doing that they won't experiment with a new tool. So you have to present the complex capabilities of the center in such a way that they're intuitive to use. That doesn't come for free. You need to study and try prototypes, repeatedly.

"You have to present the complex capabilities of the center in such a way that they're intuitive to use. That doesn't come for free. You need to study that."

—Brian Athey, NCIBI

Q: In five years, how will you know if NCIBI has succeeded in its mission?

Athey: The entire NCBC program is intended to enhance NIH research and further discoveries in certain driving biological problems. At the end of the day, it will be the researchers using the NCIBI infrastructure who will help us determine whether we've provided a capability that exceeds the scope of their laboratories and capabilities.

The success of NCIBI will be in our ability to translate researchers' questions into higher-level, more sophisticated queries of the data, literature and sets of models. In our case, the value-added is bringing a set of capabilities—natural language processing of full text scientific literature; modeling normal and diseased states; integrating high-throughput or other databases focused on a specific disease—that will allow researchers to ask and answer questions now beyond their ability. Measuring our success may or may not be difficult. Ideally, researchers will say they've made a class of discoveries they wouldn't have thought of if NCIBI didn't exist. □