

# NewsBytes

## Automating Scientific Discovery

Robots already have a place in many labs, automating tedious tasks such as pipetting samples. But a new system designed at Aberystwyth University in the United Kingdom has taken laboratory automation a step further.

“The idea of using a robot is not news, but what’s different about ours is the robot was also involved in develop-

ing hypotheses and experiments on its own,” says **Ross King, PhD**, head of computational biology at Aberystwyth University’s computer science department. The work was published in the April 2009 issue of *Science*.

“The idea of using a robot is not news, but what’s different about ours is the robot was also involved in developing hypotheses and experiments on its own,” says **Ross King, PhD**, head of computational biology at Aberystwyth University’s computer science department. The work was published in the April 2009 issue of *Science*.

“orphan” enzymes in yeast. Armed with information from bioinformatics databases such as KEGG (the Kyoto Encyclopedia of Genes and Genomes), ADAM hypothesized, from sequence similarities, which genes could encode the enzymes.

ADAM owes its brainpower in part to databases of formalized knowledge. One component is a detailed model of yeast metabolism written in the logic

language Prolog; another is an ontology describing laboratory experiments, based on the open-source project EXPO. The robot also recorded its own experimental information as it worked. “One of the advantages of a robot scientist is that you get all that metadata for free,” says King. “We can under-

stand far more about the structure of the experiment than we would if only humans had been involved.” ADAM’s four computers directed the experiments, with robot arms moving yeast mutants from freezer to incubators to plate readers. Ultimately, it found 12 gene-enzyme pairings that the authors were able to confirm. In some cases, the link between gene and enzyme was found to be supported by

“The idea of using a robot is not news, but what’s different about ours is the robot was also involved in developing hypotheses and experiments on its own,” says **Ross King, PhD**, head of computational biology at Aberystwyth University’s computer science department. The work was published in the April 2009 issue of *Science*.

ing hypotheses and experiments on its own,” says **Ross King, PhD**, head of computational biology at Aberystwyth University’s computer science department. The work was published in the April 2009 issue of *Science*.

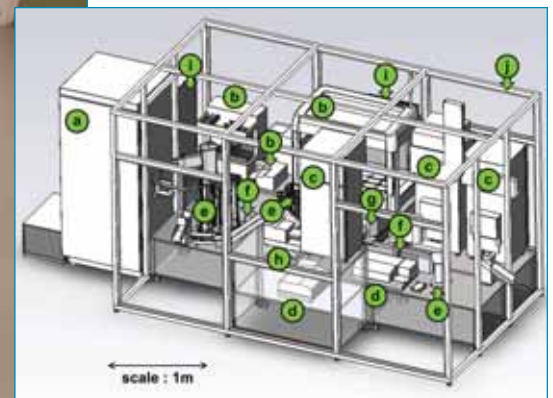
The robot, named ADAM, was programmed to find the genes that encode

language Prolog; another is an ontology describing laboratory experiments, based on the open-source project EXPO. The robot also recorded its own experimental information as it worked. “One of the advantages of a robot scientist is that you get all that metadata for free,” says King. “We can under-

stand far more about the structure of the experiment than we would if only humans had been involved.” ADAM’s four computers directed the experiments, with robot arms moving yeast mutants from freezer to incubators to plate readers. Ultimately, it found 12 gene-enzyme pairings that the authors were able to confirm. In some cases, the link between gene and enzyme was found to be supported by

literature even though it was missing from ADAM’s starting data. For others, the authors double-checked ADAM’s results by purifying and testing the protein themselves. The successful matches “are mostly to do with odd pieces of biochemistry that hadn’t been sorted out yet,” King says, which explains why the enzymes remained orphans for so long. Some were isozymes, with more than one gene encoding the same function, and others were promiscuous enzymes that catalyze more than one reaction.

King and his collaborators have started work on the next generation of robot scientists, beginning with a robot called EVE that will work to discover new drugs for tropical diseases.



ADAM is a 5-meter-long robot whose equipment includes cameras, sensors, and computers in addition to (a) an automated -20°C freezer, (b) three liquid handlers, (c) three automated +30°C incubators, (d) two automated plate readers, (e) three robot arms, (f) two automated plate slides, (g) an

automated plate centrifuge, (h) an automated plate washer, (i) two air filters, and (j) a plastic enclosure. Diagram reprinted with permission from King, RD, et al., *The Automation of Science*, *Science*, 324:85 (2009). Photo: Courtesy of Aberystwyth University.

As King describes it, “ADAM and EVE are special purpose, but our goal for the future is to make more general purpose automation.”

“People ask if this is going to put scientists out of business, but the answer is no,” says **David Waltz, PhD**, director of the Center for Computational Learning Systems at Columbia University. Instead, he says, “this will make scientists more productive,” but they would also have to learn new skills. “Scientists would have to learn to be proficient in Artificial Intelligence and to create formal representations of knowledge.”

—By **Beth Skwarecki**

## The Function of DNA Form

According to a new computational analysis of DNA structure, variations in DNA shape—along the grooves of the double helix—may play an important role in defining how the genome works. The analysis revealed that six percent of the DNA ladder’s shape is conserved across a range of different mammals—even though the sequences that produce those conserved shapes could vary.

“We’ve found a new way that evolutionary selection is working in the human genome, beyond just preserving the strict sequence of nucleotides,” says **Tom Tullius, PhD**, chemistry professor at Boston University and one of the authors of the report, published April 17 in the journal *Science*. “I hope that this finding will open up some new ways of thinking about how the genome works. It’s more than just a collection of letters.”

A 2007 study by the ENCODE (Encyclopedia Of DNA Elements) research consortium hinted that something other than nucleotide sequence was at play in determining genome function. Looking at one percent of the human genome, the researchers found that only about half of the known functional regions (for example, sections of DNA where proteins bind) showed sequence conservation across a range of mammals (from mouse to human). “We

were struck by the fact that you may not be looking at the complete story if you only look at sequence conservation to define function,” Tullius says.

Tullius and his colleagues wondered if shape might be a factor. They had previously discovered, experimentally, that different DNA sequences can have similar structures. Using the reactive hydroxyl radical molecule, they had probed for subtle differences in DNA shape. Small variations in the radical’s accessibility to the DNA yield a detailed structural map. These variations are often in the DNA’s minor groove width, which can range from four angstroms at the narrowest to 11 at the widest, Tullius says. This finding led them to wonder whether sequences could diverge through evolution while form remained the same.

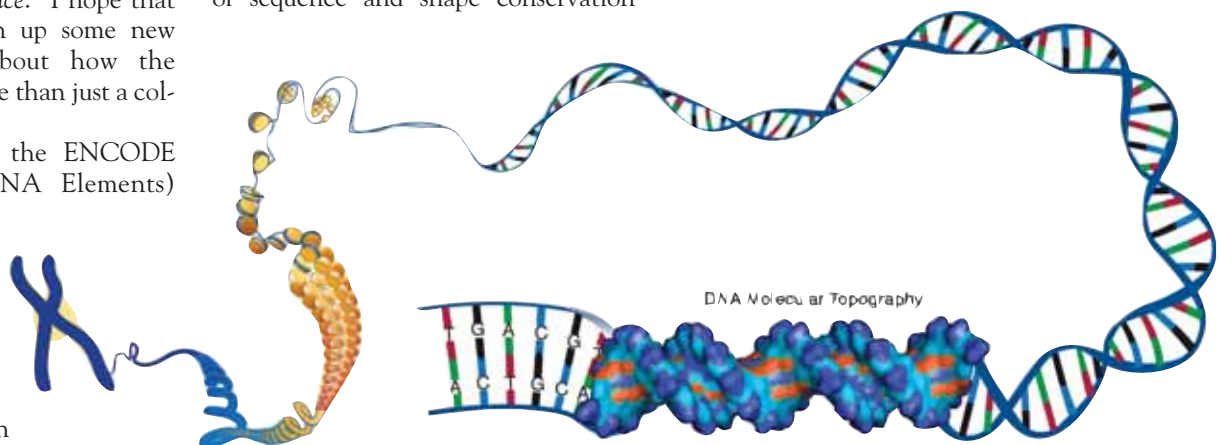
To answer that question, Tullius, **Elliott Margulies, PhD** of the National Institutes of Health, **Steve Parker** of Boston University and their colleagues created a computer program called Chai. The program compares computational predictions of DNA shapes from the same one percent of the human genome studied by ENCODE, and other mammalian genomes. They found that certain parts of the genome are conserved solely by structure, not sequence. Moreover, the combination of sequence and shape conservation

almost entirely covers the functional sites identified by the ENCODE study. Tullius and his colleagues also found that polymorphisms associated with disease are more likely to cause structural changes in DNA than neutral polymorphisms—meaning that these shape changes could be disrupting the binding of some essential protein.

“I hope that this finding will open up some new ways of thinking about how the genome works,” Tom Tullius says. “It’s more than just a collection of letters.”

In a 2007 report in the journal *Cell*, **Barry Honig, PhD** of Columbia University, had concluded that DNA shape influenced the binding of a homeodomain protein to developmental genes. “The combination of these two studies makes it clear that DNA shape is important in function,” Honig says. “This gives us a new avenue to study how DNA functions that we didn’t have before.”

—By **Rachel Tompa, PhD**



*An illustration of DNA organization from chromosome to double helix. Scientists have found subtle structural differences at the molecular level between different regions of DNA, often in the width of the helix’s minor groove. Surprisingly, different sequences can yield the same shapes in DNA. Tullius, Margulies and Parker found that these subtle shapes are conserved between humans and other mammals, meaning evolution is acting not only on our DNA sequence, but its form. Courtesy of Darryl Leja, NHGRI, NIH.*

## Semantic Publishing and Scientific Journals

Keeping up with the literature is a challenge for all scientists. But some researchers are making it easier by enhancing the usability and understanding of an article's contents in a variety of ways—an approach called “semantic publishing.” Recent efforts include a manual demonstration project published by the *Public Library of Science (PLoS)* as well as a number of automated tools being developed around the world. Combined, they provide an intriguing glimpse at scientific publishing's possible future.

“It's exciting to me that now there are the first stirrings of people who are doing this for real with semantic markup either manually or automatically,” says **David Shotton, PhD**, a reader in image bioinformatics at Oxford University and lead author of an April

2009 *PLoS Computational Biology* paper describing the demonstration project. “If researchers can find relevant papers faster and understand their import faster, that will assist their research.”

Shotton and his colleagues spent several weeks last year manually enhancing a paper (by Reis et al., 2008) published in *PLoS Neglected Tropical Diseases* (<http://dx.doi.org/10.1371/journal.pntd.0000228.x001>). Among other things, they added machine readable data (Excel spreadsheets rather than static images); provided ways to highlight various important terms in the paper; and added hyperlinks. In addition, scrolling over a text citation brings up a hover box showing the citation as well as relevant text from the original citation—so the reader can understand why it is cited without having to look it up.

“Many of the things we did are trivial, but cumulatively they make a difference. Perhaps a small difference, but

a helpful difference,” Shotton says.

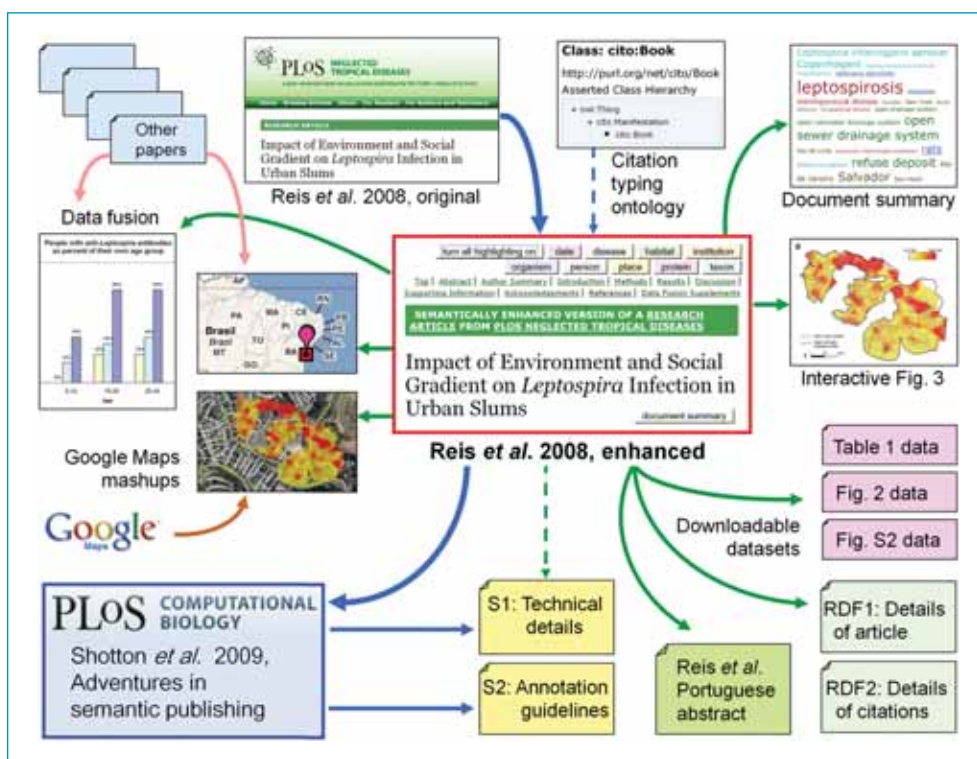
Shotton and his colleagues manually curated the paper—a slow process that could be improved via automa-

“Static PDFs are antithetical to the spirit of the web,” David Shotton says.

tion. Automation of some of Shotton's manual tasks has already occurred through the Elsevier Grand Challenge (where Shotton served as a judge)—a contest created to improve the way scientific information is communicated and used. One of the runners-up this year—a team from Australia—built a tool that automatically creates the kind of citation hover boxes that Shotton's group built by hand. It uses very standard reliable text mining algorithms to extract words from the citing reference, looks at the cited reference for similar conjunctions of words, and pulls back the most relevant sentences. “And it works,” Shotton says.

This year's Challenge's winners (announced in April) developed a browser plug-in called Reflect (freely downloadable at <http://reflect.ws>). Clicking on the REFLECT button in any Web browser automatically marks up an online document to show instances of protein, gene and chemical names—in just seconds. Next, a click on the highlighted term brings up a box with all sorts of information about that gene/protein or chemical. Soon, the group hopes to add other categories, such as diseases and cell types.

The journal *Nature* is starting to implement some semantic publishing approaches, says **Timo Hannay, PhD**, the publishing director at Nature.com. Still, he says, there remains the question of which enhancements to implement first, given the state of technology; and how to get authors to buy in, especially if they will have to do extra work. “We're just at the beginning, but I'd like to see as much of our information as pos-



*As a demonstration of what's possible, Shotton and his colleagues manually enhanced a paper by Reis, et al. (2008) in PLoS Neglected Tropical Diseases. As shown here, the enhancements included (among other things) mash-ups of maps with data from several papers; a citation ontology; a document summary in word cloud format; and conversion of PDFs into downloadable datasets (“Static PDFs are antithetical to the spirit of the web,” Shotton says). Reprinted from Shotton, D, et al., Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article. PLoS Computational Biology 5(4): e1000361. doi:10.1371/journal.pcbi.1000361, (2009).*

sible provided in structured, standard, machine-readable form,” Hannay says.

—By *Katharine Miller*

## Open Source Tools for Parsing Clinical Records

Researchers at the Mayo Clinic and IBM have each built computer pipelines for extracting useful information from unstructured notes in patient charts, such as physician’s notes and pathology reports. And they’ve now partnered to make these best-of-breed natural language annotators freely available through the Open Health Natural Language Processing (OHNLP) Consortium (<http://ohnlp.org>).

“While each of us [IBM and Mayo] contributed a whole pipeline, the more important contribution was that we were starting to feed the shelves with annotator widgets that other people could take and assemble in different and interesting ways,” says **Christopher Chute, MD, DrPh**, Mayo Clinic bioinformatics expert and senior consultant on the OHNLP Consortium project. If someone wants to create a complex natural language processing (NLP) pipeline to address a particular research question, “maybe they write the tough little piece that goes in the middle, but 90 percent of the work is already written,” he says.

Until recently, researchers could only

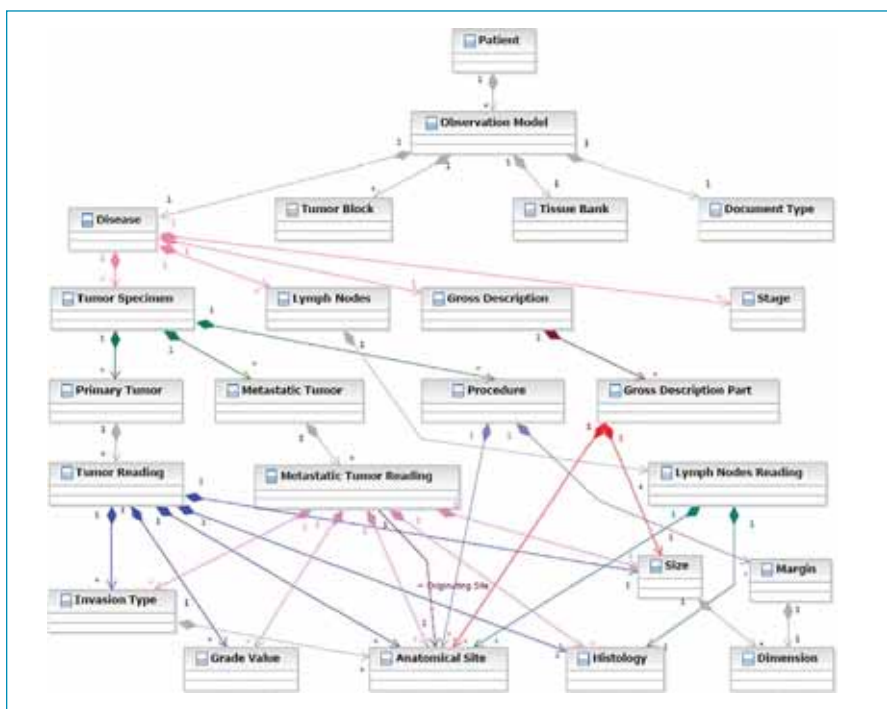
access valuable data within medical records by hiring medical professionals to read charts and abstract the information out to a case report form. And while general architectures for building NLP pipelines and automating this process of extracting information from medical records aren’t new, and some of the pieces of the OHNLP pipelines, such as standardized vocabularies, have existed for some time, the Mayo and IBM pipelines weave many pieces together to accomplish real world tasks. “And that’s what we so desperately need,” says **Rebecca Crowley, MD**, associate professor of biomedical informatics at the University of Pittsburgh, a researcher who has developed a separate open source pipeline (caTIES) for extracting information from pathology reports.

In an NLP pipeline, unstructured text goes through a series of annotators that work step-by-step toward identifying meaningful entities or phrases and the relationships between them. For instance, the first annotator distinguishes letters from punctuation and other marks, the next identifies

“While each of us [IBM and Mayo] contributed a whole pipeline, the more important contribution was that we were starting to feed the shelves with annotator widgets that other people could take and assemble in different and interesting ways,” says Christopher Chute.

words, and the next that identifies parts of speech. Ultimately this might lead to an annotator that could assign meaning to phrases or entities.

For the most part, Chute says, Mayo’s pipeline (cTAKES) stops at the stage of entity recognition—identifying specific symptoms, diseases, and drugs. Once you have the entities or phrases, he says, “then you can start doing all kinds of fun things either with subsequent annotators or as a post-NLP process.” IBM’s medKAT pipeline also includes annotators that identify relations between named entities. For example, a pathology record might mention multiple sites and sizes of tumors, but medKAT identifies relationships among those pieces of information in order to identify, for exam-



*The OHNLP resource (<http://ohnlp.org>) includes IBM’s NLP pipeline (medKAT), which can automatically extract cancer disease characteristics from pathology reports in order to populate a cancer disease knowledge base with the structure shown here. Reprinted from Coden A, et al., Automatically extracting cancer disease characteristics from pathology reports, *Journal of Biomedical Informatics* (2009) doi:10.1016/j.jbi.2008.12.005.*

ple, the size of the primary tumor.

In the long run, Chute says, the OHNLP will be most valuable for building a community of people who use shared tools. IBM's manager of medical text and image analysis, **Anni Coden, PhD**, who leads work on the IBM pipeline (medKAT), agrees. "We [IBM and Mayo] decided to put this out there in open source because it takes a whole community to make progress in this field," says Coden. "If we put our efforts together we may be able to solve it."

Crowley says the long-term value of NLP pipelines is clear: "So much of the data we want to work with is available only in text," she says. "Data mining, identifying new hypotheses, translational research and clinical trials can all benefit greatly from being able to access data in text."

—By Katharine Miller

## Online Searches Warn of Flu Spikes

Current methods of tracking the flu all come with a bit of a time lag—which is unfortunate when trying to monitor for potential pandemics like today's swine flu crisis. There is a faster way: According to a February 2009

report in *Nature*, Google researchers can track flu incidence in real time by monitoring online search queries. The Google model catches a flu outbreak one to two weeks earlier than the Center for Disease Control's current reporting methods.

"Having that one to two week advantage of knowing that something may be developing can have a signifi-

influenza spike, and offered the data a week or so faster than traditional methods. His study was reported in the *American Medical Informatics Association Annual Symposium Proceedings*.

Google built on the work of Eysenbach and others. The Google researchers started with 50 million of the most common searches. They compared the weekly frequency of each with

"Having that one to two week advantage of knowing that something may be developing can have a significant impact on the public health outcome," says Kumanan Wilson.

cant impact on the public health outcome," says **Kumanan Wilson, MD**, an investigator of public health policy at the University of Toronto.

Public health officials in the United States and Canada now depend on sentinel doctor's offices to regularly report the number of people who walk through the door with "influenza-like illness" (ILI) symptoms. But this approach is slow, prone to human error, and relatively costly, says **Gunther Eysenbach, MD, MPH**, a senior scientist at the Centre for Global eHealth Innovation and professor at the University of Toronto in Canada.

In 2006, Eysenbach, who first had the idea of using Internet search queries to track the flu, performed a pilot study showing that he could largely eliminate the reporting lag with an automated system. Eysenbach's strategy tracked how often folks searched for "flu" or "flu symptoms" online, and then noted how many users subsequently clicked on an informational ad about seasonal influenza. The number of users who clicked on the ad closely traced Canada's seasonal

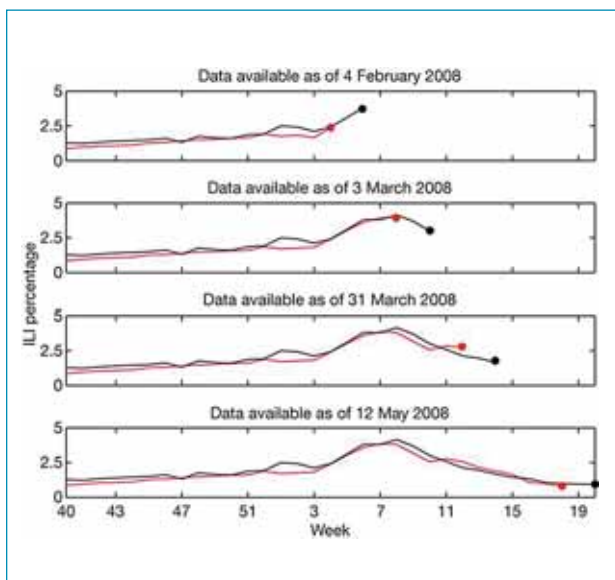
the up and down of seasonal flu spikes over five years. Those that correlated best (the top 45) were all flu-related.

With those top 45 search queries, Google created a linear model for tracking the flu in real time. Current data can be found at Google Flu Trends (<http://www.google.org/flutrends/>), which, since April, is also tracking flu trends in Mexico.

Internet queries can pick up a flu spike quickly because they immediately register any increased interest in the flu. That is both a strength and a weakness of what Eysenbach calls "infodemiology." The downside, he says, is that in a pandemic situation, you may be monitoring more panic than actual flu cases. "Our current swine flu data demonstrate that it can be difficult to separate the signal from the noise," he says.

Before the search query approach can be adopted as an early warning signal on the national or international level, its effectiveness needs to be better proven, says Wilson. But he likes the idea of a freely available, Internet-based system that would likely encourage more transparent reporting by governments and health officials.

Eysenbach is now investigating many other ways of using the Internet to observe and influence people's health. He wants to interact with those



Google's model (black) uses Internet search queries about the flu to estimate current flu levels a week or two faster than the CDC (red). Reprinted by permission from MacMillan Publishers, LTD, Jeremy Ginsberg, et al., *Detecting influenza epidemics using search engine query data*, *Nature* 457:1012-1014, copyright 2009.

who search online through questionnaires, and he is seeing what he can gather from microblogs such as Twitter.

—By *Louisa Dalton*

## Flowing through the Interactome

High-throughput experimental methods are widely used today to identify genes and proteins involved in a particular process, but not all molecules in a pathway can be identified in this manner. To fill the gaps, a new computer program called ResponseNet follows the path of least resistance—like water flowing from sources to sinks in a terrain—to find the most efficient path through the maze of interacting molecules in a cell (the “interactome”). The work was published in the March 2009 issue of *Nature Genetics*.

ResponseNet “is a step toward a much more realistic and mechanistic view of what’s going on in cells that could ultimately do much better in terms of predicting what’s important in diseases,” says co-author **Ernest Fraenkel, PhD**, a biological engineer at the Massachusetts Institute of Technology (MIT). Indeed, Fraenkel and his colleagues have already produced the first cellular map of the proteins and genes that respond to alpha-synuclein—a key protein linked to Parkinson’s disease.

Two important types of high-throughput experiments are commonly used to identify genes and proteins that are important in a particular condition or disease: mRNA profiling, which measures changes in gene expression under various conditions; and genetic screening, which finds genes that, when deleted or altered, change how cells respond to stimuli. But some components of signaling pathways don’t show up in these experiments. In addition, there’s surprisingly little overlap between the genes

identified via these two techniques: Genes found by genetic screening tend to be involved in regulating other genes while genes found by mRNA profiling are often part of metabolic processes. The team hypothesized that the two might be connected; that is, the genes found in genetic screens might be controlling those found by mRNA profiling.

To test their idea, the team turned to the yeast interactome, a massive and complex network of all known yeast protein-gene and protein-protein interactions. “The data are very noisy and incomplete, which means that everything can be connected to everything,” says team member **Esti Yeager-Lotem, PhD**, an MIT postdoc. Using a flow algorithm—an approach commonly used in the telecommunications industry—they sought the most efficient path from the regulators (genetic screen results) to the differentially expressed genes (mRNA profiling results). “By doing that, ResponseNet identifies intermediary proteins that are predicted to be part of response pathways but are not found by high-throughput methods,” says **Laura Riva, PhD**, also a postdoc at MIT.

The researchers tested their approach in cells that overexpress alpha-synuclein, a protein that is associated with Parkinson’s disease. “ResponseNet was able to provide the first cellular map of the proteins and genes responding to alpha-synuclein expression,” Riva says.

“Their solution is

*By flowing through the interactome from genes identified in genetic screening experiments (orange diamonds—usually regulators) to proteins identified in mRNA profiling (green squares—usually regulatees involved in metabolism), ResponseNet identifies what other components (gray circles) might be involved in the pathway and evaluates their likely importance within the pathway (heaviness of the arrows). Image reprinted by permission from MacMillan Publishers LTD, Esti Yeager-Lotem et al., Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity, Supplementary Notes, Nature Genetics 41:316-323 (2009).*

“I think this is a step toward a much more realistic and mechanistic view of what’s going on in cells that could ultimately do much better in terms of predicting what’s important in diseases,” says co-author Ernest Fraenkel.

novel and makes an important step,” comments **Aviv Regev, PhD**, a computational and systems biologist at MIT and the Broad Institute who was not involved in the work. While the research team hopes to apply this technique to mammalian cells, “the key challenge in applying it to higher organisms is the lack of interaction data to the same scale and coverage as in yeast,” Regev says. For now, though, ResponseNet will make the yeast model a more powerful tool for studying neurodegenerative and other diseases.

—By *Liz Savage* □

