desired phenotypes and then obtaining their anonymized leftover blood samples to test for genetic information.

"We showed that we can actually conduct full-blown association studies to find the right patients with the right phenotypes and connect them to the right samples," says **Isaac Kohane, MD, PhD**, professor at Harvard Medical School and director of i2b2 (Informatics for Integrating Biology and the Bedside), the National Center for Biomedical Computing that conducted the study published in the September 2009 issue of *Genome Research*. "It's soup to nuts work."

With the help of natural language processing (NLP), the i2b2 researchers set out to use a large, available, cheap data pool: the electronic medical record archives for 2.6 million patients at Partners Healthcare System in Massachusetts. Although doctor's notes are notoriously unstandardized, NLP tools can break them into their smallest components, analyzing parts of speech and how words are joined. The i2b2 team sought to identify pools of patients with rheumatoid arthritis, asthma, secondary illnesses and risk factors for asthma (for example, smoking history). Along the way, clinical experts gauged the accuracy of the process and helped refine search terms. "It takes three to four months of iteration with expert clinicians until we get it just right," Kohane says. In addition, the researchers developed a system to access anonymously saved leftover blood samples from the identified populations to use for future studies requiring genetic data.

And the NLP tools did a pretty good job: Of about 98,000 patients identified as having asthma, 82 percent of the time the experts reviewing the files concurred in that diagnosis; 90 percent of the patients identified with a history of smoking had such a history; and of the 4,618 NLP-identified rheumatoid arthritis sufferers, 92 percent had definite arthritis (according to expert review) while 98 percent probably did. By studying these electronic patients, the researchers successfully reproduced several results from past clinical research. And while the clinical studies had paid an average of $650 to characterize and obtain blood samples from each patient, i2b2 spent $20 to $100.

Medical Record Snippet	*Smoking History*
SOCIAL HISTORY: The patient is married with four grown daughters, uses tobacco, has wine with dinner.	Positive
SOCIAL HISTORY: The patient is a nonsmoker. No alcohol.	Negative
SOCIAL HISTORY: Negative for tobacco, alcohol, and IV drug abuse.	Negative
BRIEF RESUME OF HOSPITAL COURSE: 63 yo woman with COPD, 50 pack-yr tobacco (quit 3 wks ago), spinal stenosis, ...	Positive
SOCIAL HISTORY: The patient lives in rehab, married. Unclear smoking history from the admission note...	Insufficient data
HOSPITAL COURSE: ... It was recommended that she receive ...We also added Lactinax, oral form of Lactobacillus acidophilus to attempt a repopulation of her gut.	Insufficient data
SH: widow,lives alone,2 children,no tob/alcohol.	Insufficient data

After lengthy training, i2b2's natural language processing software scans clinical histories, tagging words and phrases that describe smoking history and making a diagnosis (right-hand column). With training, the NLP tools were able to equate "smoking history" with "smokes often," distinguishing both from "non-smoker." Clinical experts also reviewed random results and computer scientists refined the search terms to clarify ambiguities like "tob." Reprinted from S. Murphy, et. al, Instrumenting the health care enterprise for discovery research in the genomic era,* Genome Research *, 19(9): 1675–1681 (2009).

"This paper represents very encouraging results using free open-source software," says **Chunhua Weng, PhD**, assistant professor of biomedical informatics at Columbia University. She says the next step is to include information such as how long an individual smoked or when symptoms began in patient descriptions. Kohane agrees, noting that researchers are working to include time-varying data in i2b2's model.

—By Daniel Strain

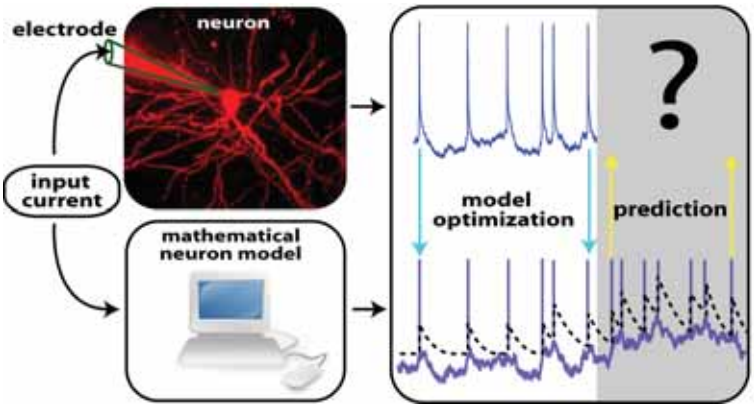
Neuron Models: Simpler Is Better

During the summer of 2009, the International Neuroinformatics Coordinating Facility in Stockholm dangled a nearly $10,000 cash prize in front of neuron modelers and challenged them to do better. And they did. The winners of the competition, which was described in the October 16, 2009 issue of *Science*, produced a neuron model that became more accurate as they stripped away pieces of a much more complex starting model.

"It was amazing for us physicists to see the description become simpler as we tried to make the performance better," says **Shigeru Shinomoto, PhD**, a physicist at Kyoto University in Japan who, along with two of his former students, snagged the grand prize.

Modeling the electrical behavior of individual neurons is crucial to understanding how thought and other cognitive functions arise in complex neuronal networks. Current neuron models can predict some neuron behavior, but with limited accuracy and at high computational cost.

The international competition has grown from eight entrants in 2007 to 33 this year and included teams around the world. "We had different people from different backgrounds using methods we would never have thought of," says **Wulfram Gerstner, PhD**, a computational neuroscientist at the Ecole Polytechnique Federale in Lausanne, Switzerland who co-authored the *Science* paper.



To set up one of the challenges for the neuron modeling competition, an artificial current was injected into a live neuron (upper left) and the resulting electrical activity was recorded for 60 seconds (blue trace, top right). Competitors used data from the first 38 seconds of the recording to fine-tune the parameters of a mathematical neuron model receiving an identical current injection (purple trace, lower right). Model performance was measured by the percentage of spikes correctly predicted in the final 22 seconds of the recording. Graphic courtesy of Richard Naud.

Contestants had to predict the precise timing of electrical spikes in individual neurons from different parts of the brain. Since different neurons can respond differently to the same signal, competitors used the first 38 seconds of data from a neuron to adjust their model parameters to better fit that neuron. They used the freshly tuned model to predict spikes in the subsequent 22 seconds of data. Shinomoto's winning model predicted 59.6 percent and 81.6 percent, respectively, of the spikes from two different neurons.

Electrical activity in a real neuron spikes when its membrane potential passes a set threshold value. Shinomoto's model neuron has an adapting threshold that increases immediately after a spike and decays exponentially to its initial value. The decay is modulated by two time constants of 10 ms and 200 ms, chosen to reflect the timing of ion currents in the neuron membrane.

The competition will evolve with the field, Gerstner says. Computational neuroscientists will soon draw on an emerging body of molecular knowledge to improve their models, says **Erik De Schutter, PhD**, a professor of computational neuroscience at the Okinawa Institute of Science and Technology. Advanced molecular techniques should reveal the physical structures and electrical properties of neurons in much greater detail than is currently known. These data may help modelers account for the effects of variations in temperature and chemical conditions and in the physical structures of the neurons.

"Neuron modeling is still a work in progress," De Schutter says. "It's much more difficult than we thought."

—By Sandra M. Chung

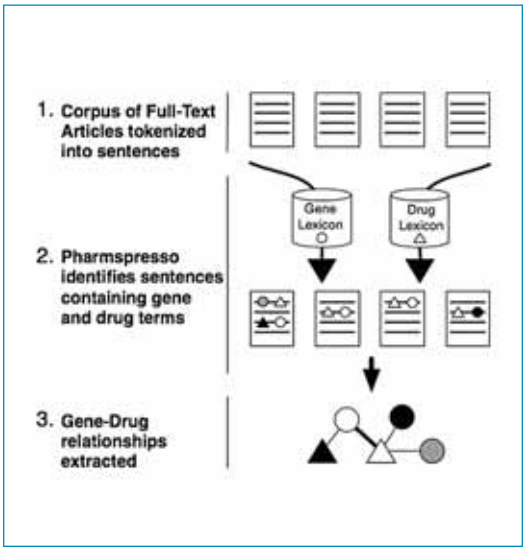
Trawling for Drug-Gene Relationships

When a drug saves one person but makes another ill, a bitter lesson in genetic differences often follows. With many such lessons already under our collective belts, researchers are using existing knowledge to predict additional drug-gene relationships as a way to forestall future calamities. A new software program can trawl published papers for gene-drug relationships, plug those relationships into known genetic networks, and predict which genes are likely to affect a patient's response to a drug.

"Our contribution is using text mining and taking decades of research and folding that in to inform the prediction," says **Yael Garten**, biomedical informatics PhD candidate in the lab of **Russ B. Altman, MD, PhD**, at Stanford University and a lead author of the work. "We showed that this is as good as and sometimes even better than manual curation," in which scientists painstakingly enter published drug-gene interactions into a database. Garten will present the team's research January 2010 at the Pacific Symposium on Biocomputing in Hawaii.

The previous version of the algorithm, designed by Altman and others, relied more heavily on manual labor. Called PGxPipeline, it employed a database of gene-drug relationships manually compiled from scientific articles by a team of scientists at Stanford Medical School. PGxPipeline wove these relationships into an orderly web, along with a database of gene-gene interactions and other data, to predict how strongly each of 12,460 genes affects response to a specific drug.

The team has now cut PGxPipeline loose from the manually created drug-gene database, automatically mining the information from published papers. This faster, cheaper method will inform the drug-gene rankings with constant updates from new literature. The manual-curation- and text-mining-based



The text-mining-based version of PGxPipeline automatically dissects journal articles into component sentences and marks where a drug or a gene is mentioned. Reading the sentence syntax and vocabulary, it tracks the interactions between drugs and genes. A network/web of interactions is established (bottom), in which the thickness of each edge corresponds to the number of articles that support the interaction. The web of relationships is later enhanced using a database of gene-gene interactions and other information. Image reprinted from Garten, Y., Tatonetti, N., & Altman, R., Improving the prediction of pharmacogenes using text-derived drug-gene relationships, Pacific Symposium on Biocomputing, Hawaii, January 2010.