DIVERSE DISCIPLINES, ONE COMMUNITY

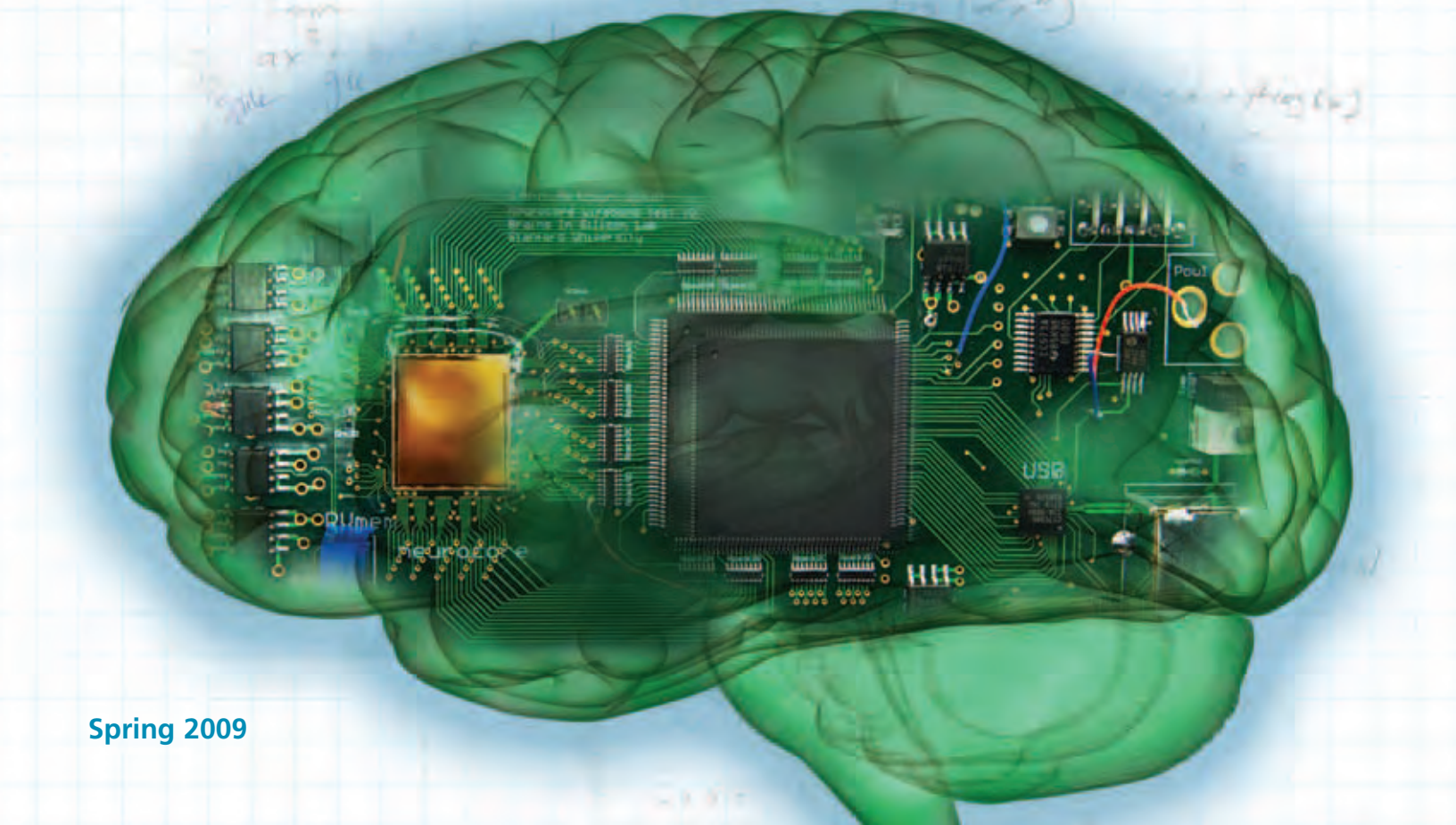
Biomedical Computation REVIEW

Published by Simbios, an NIH National Center for Biomedical Computing

Reverse Engineering the Brain

PLUS

Bringing the Fruits of Computation to Bear on Human Health: It's a Tough Job But the NIH Has To Do It

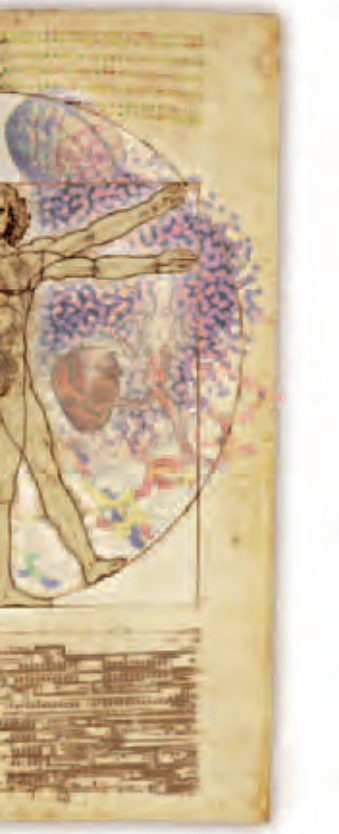


Spring 2009

contents

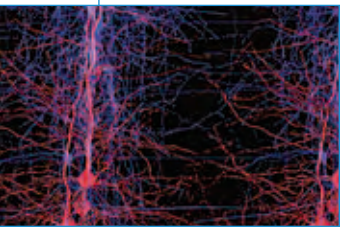
ContentsSpring 2009

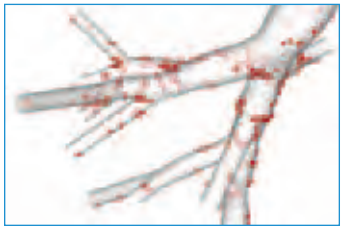
FEATURES



DEPARTMENTS

COVER ART: CREATED BY RACHEL JONES OF WINK DESIGN STUDIO USING A NEUROGRID CHIP, COURTESY OF RODRIGO ALVAREZ. BRAIN IS © YAKOBCHUK | DREAMSTIME.COM.
PAGE 16: ARTWORK CREATED BY RACHEL JONES OF WINK DESIGN STUDIO WITH IMAGES FROM THE NCBCs. VITRUVIAN MAN AND AGED PARCMENT IMAGES ARE © SQBACK | DREAMSTIME.COM. INTERNAL ORGANS BASED ON IMAGE © SEBASTIAN KAULITZKI | DREAMSTIME.COM.





Spring 2009
Volume 5, Issue 2
ISSN 1557-3192

Executive Editor David Paik, PhD

Managing Editor Katharine Miller

Associate Editor Joy Ku, PhD

Science Writers
Katharine Miller, Kristin Sainani, PhD, Chandra Shekhar, PhD, Roberta Friedman, PhD, Kaspar Mossman, PhD, Rachel Tompa, PhD

Community Contributors
Ron Kikinis, PhD, Tina Kapur, PhD, Bin Dong, Joy Ku, PhD, Russ B. Altman, MD, PhD, Isaac Kohane, MD, PhD

Layout and Design
Wink Design Studio

Printing
Advanced Printing

For general inquiries, subscriptions, or letters to the editor, visit our website at www.biomedicalcomputationreview.org

Office
Biomedical Computation Review
Stanford University
318 Campus Drive
Clark Center Room S231
Stanford, CA 94305-5444

Biomedical Computation Review is published quarterly by:

The NIH National Center for Physics-Based Simulation of Biological Structures

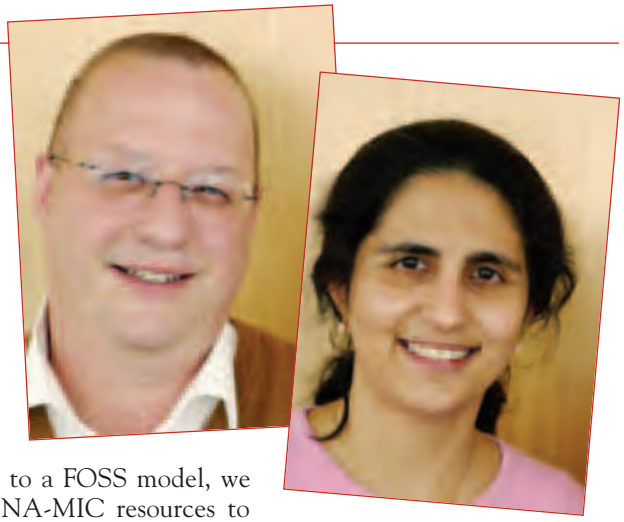
Publication is made possible through the NIH Roadmap for Medical Research Grant U54 GM072970. Information on the National Centers for Biomedical Computing can be obtained from http://nihroadmap.nih.gov/bioinformatics. The NIH program and science officers for Simbios are:

Peter Lyster, PhD (NIGMS)
Jennie Larkin, PhD (NHLBI)
Jennifer Couch, PhD (NCI)
Semahat Demir, PhD (NSF)
Peter Highnam, PhD (NCRR)
Jerry Li, MD, PhD (NIGMS)
Yuan Liu, PhD (NINDS)
Richard Morris, PhD (NIAID)
Joseph Pancrazio, PhD (NINDS)
Grace Peng, PhD (NIBIB)
Nancy Shinowara, PhD (NCMRR)
David Thomassen, PhD (DOE)
Ronald J. White, PhD (NASA/USRA)
Jane Ye, PhD (NLM)

GuestEditorial

BY RON KIKINIS, PhD AND TINA KAPUR, PhD

Open Solutions for Biomedical Research



What can open-source software do for biomedical research? Based on our experience at the National Alliance for Medical Image Computing (NA-MIC), we believe that open source software can be used very effectively to either directly solve biomedical research problems or improve the infrastructure that will enable others to solve it.

Research done nationwide by NA-MIC-associated researchers provides a case in point: Neuroscientists, clinicians, and biomedical researchers at Harvard Medical School, Johns Hopkins University, the University of North Carolina at Chapel Hill, and the University of New Mexico all use NA-MIC's open-source software tools routinely. They analyze anatomical brain connectivity abnormalities in patients with velo-cardio facial syndrome (VCFS); perform robot and image-guided biopsies for prostate cancer; monitor the progression of Lupus lesions in patients; and study cortical thickness as a predictor for autism. The needs of these specific biomedical problems drive the efforts of computational scientists and software engineers in the development of a Free and Open-Source Software (FOSS) toolkit, the NA-MIC Kit, to enable biomedical research.

At NA-MIC we've seen that biomedical research needs can more easily drive tool development when the tools are open source. That's because open source tools can be modified to suit researchers' needs. In accordance with our mandate as a National Center for Biomedical Computing, our mode of operation in NA-MIC has been to focus as a team on one set of biomedical problems at a time (we started with schizophrenia and expanded to include lupus, autism, prostate cancer, and VCFS) in order to identify gaps in the available image analysis algorithms and software tools, and then design solutions. To fill such gaps, we may create new computational technology, deploying them through the NA-MIC Kit; or when the gaps are better addressed by helping owners of existing computational solutions migrate to a FOSS model, we deploy NA-MIC resources to aid the migration. In either case, we consistently seek opportunities to enhance the NA-MIC platform by supporting "faster, better, cheaper" computation on increasingly complex medical images. This is possible because NA-MIC has adopted a BSD-style open source license for our infrastructure software, which allows research and commercial entities unrestricted access to build upon the NA-MIC software platform without any "give back" requirements.

And because our open source software is free, it also has the potential to impact a broad segment of the research community. To realize that potential, NA-MIC has applied significant resources to establish a dissemination program that has trained over 600 biomedical engineers and scientists from institutions around the country in the use of our tools. In addition to the detailed training materials that we provide through our Web site (http://www.namic.org), our hands-on training sessions are particularly effective in creating "super users" of our technology who in turn train groups in their home institutions and serve as NA-MIC ambassadors to the broader research community. We also have an active and open community process in place for those who choose to participate. This includes, wikis, teleconferences, mailing lists, and in-person events, all of which are publicly available through our Web site.

Biomedical research needs can more easily drive tool development when the tools are open source. That's because open source tools can be modified to suit researchers' needs.

We invite you to browse three specific end-to-end Open Science solutions that we have developed in NA-MIC recently for three separate biomedical research problems 1) Automatic Regional Cortical Thickness Assessment for Autism, 2) Segmentation for Non-human Primate Brains, and 3) Planning Therapy for Prostate Cancer at http://wiki.na-mic.org/Wiki/index.php/Events:TutorialContestJan2009. □

Clinical Data Repositories:

Less Than Meets the Eye

There has long been a belief that clinical data repositories are potential gold mines of untapped knowledge, and that with the appropriate electronic infrastructure, they would serve as a source of new information about the causes of disease, the identity of new biomarkers, and other unappreciated statistical correlations. Using the awesome power of modern data mining and machine learning methods, we would be able to troll years of clinical data and extract gold. I am afraid that this view is overly optimistic.

First, clinical data repositories are the historical record of physicians and other healthcare providers ordering tests, procedures, and documenting their inference in an extremely biased manner. The goal is not to objectively sample reality, but to build a story that convinces themselves and others that they have the right model of what is wrong with the patient, and that their actions are reasonable. This is not a bad thing—this is the exercise of the art of medicine (which is still very much an art, despite our attempts to codify and standardize)—but it is not a good basis upon which to build a discovery engine. Second, clinical records were not invented to support research and many of the elements required for good research are not present. All tests must be interpreted with knowledge of the prevalence as well as the sensitivity and specificity. Thus, taken out of context, the results of a test are very difficult to interpret with respect to their accuracy and information content. Physicians also sometimes order tests for medico-legal reasons, financial reasons, to document the course of a disease, because patients insist on the tests, and many other spurious reasons. These sources of bias, importantly, are typically not recorded anywhere in the medical record, and so the task of statistically untangling the data in order to generate a clean and believable dataset is incredibly difficult. Third, the practice of medicine and the use of medical tests, procedures and terminology are constantly evolving, and thus any attempt to combine data over a significant period of time (even as short as a few years) is likely to be confounded by changes in practice, the arrival of new therapeutic and diagnostic capabilities, and simple "medical fashion."

We should try our best to learn new things with the data, but our expectations should be low. More importantly, our skepticism about whatever is discovered should be very high.

Having expressed my pessimism, I believe we still should apply data mining algorithms to these data, and attempt to overcome these difficult challenges. In our initial investigations, we are likely to discover the obvious: typical patterns of test ordering associated with medical practice, and the trivial correlations between different data sources in the record. We should try our best to learn new things with the data, but our expectations should be low. More importantly, our skepticism about whatever is discovered should be very high. Data mining activities may suggest new hypotheses, and these can then be followed up with careful analysis of clean (ideally prospective) data. But to expect a goldmine of discoveries, at least today, is to underestimate the difficulty in preparing this data for serious use in discovery.

DETAILS

Russ B. Altman, MD, PhD, is principal investigator for Simbios, a National Center for Biomedical Computing, and professor of bioengineering, genetics and medicine at Stanford University.

SEND US IDEAS **Got your own opinions on this topic? Or have another topic you'd like to write about for these pages? Send us your thoughts on the Feedback page of our Web site: http://biomedicalcomputationreview.org/feedback.html.**

More Valuable Than You'd Expect

The recent federal stimulus package is injecting billions of dollars into electronic health record implementation. To determine whether such records will be useful for clinical research in the genomic era, three questions should be answered: 1) As compared to what? 2) To do what? 3) At what cost in time and treasure?

With regard to the first question, we know that even highly regarded large cohort studies such as those published regularly in the genome-wide association studies literature are highly prone to phenotypic misclassification. We also know that carefully selected populations exhibit different characteristics than the population of health-care recipients as a whole. For example, populations selected for a study may not manifest all the risks and interactions of exposure to a particular therapeutic drug. Moreover, when human beings extract medical records or ask questions, they inject variability and biases, evident in a review of any of the annotations in existing studies. In contrast, we can repeatedly run the entire corpus of suitably de-identified clinical records through different natural language processing methods and filters of varying stringency, and compare patient characteristics at one hospital to those at another. We can do so comprehensively, repeatedly, and have available large numbers of controls for confounding factors. Moreover, whereas classical recruited cohort studies usually face significant challenges in obtaining adequate representation of underrepresented minorities, comprehensive electronic medical records typically include more members of these groups.

With regard to the second question, there is no doubt one has to be sober and careful regarding what kinds of questions can be answered using data from electronic medical systems. Nevertheless, there are numerous "low hanging fruit" that enable electronic-health-record-based research projects to proceed productively. For example, any methodological and timely review of health-care system data could identify certain noteworthy high-magnitude epidemiological effects that currently go unrecognized in our health care system—often until it is too late. A prime case of this is the very large number of deaths attributable to Vioxx. Similarly important but rarer events such as the pancreatitis associated with exenatide can be sussed out even if the exact magnitude of the effect is in question. Such findings would then trigger further, perhaps better-controlled, studies. In addition, when a study would require hundreds of thousands of patients to measure a low-magnitude effect, electronic health records provide a unique resource, particularly when they are carefully mined using natural language processing techniques. Indeed much of the purely claims-based research is vulnerable to both the coarseness and reimbursement bias of the characterization provided by billing codes (e.g., a radiologist coding a "rule out rheumatoid arthritis" x-ray with the diagnosis of rheumatoid arthritis even if the patient does not have such disease).

Given the availability of clinical data obtained from our very expensive and intensive health care process, we must at least determine the extent to which electronic health record information can further science and improve diagnostic and treatment modalities.

And as for the cost in time and treasure, if we can conduct *in silico* observational studies in one hundredth of the time and for one tenth, hundredth or even thousandth of the cost of a conventional observational study, we should do so—with appropriate adjustments for bias, variation, and multiple hypothesis testing. We owe it to the public to at least explore what the results might be, especially if they identify dangerous drugs or promising new therapies. Given the availability of clinical data obtained from our very expensive and intensive health care process, we must at least determine the extent to which electronic health record information can further science and improve modes of diagnosis and treatment.

There is, of course, no doubt that many studies are answerable only through classically organized randomized controlled trials or tightly selected observational studies. However, the promise of electronic health records was never that they would be the only platform on which clinical research could be conducted in the future but "merely" an important component of the research agenda at the national and international level. □

DETAILS

Isaac Kohane, MD, PhD, is the principal investigator for Informatics for Integrating Biology and the Bedside (i2b2), as well as Lawrence J. Henderson Associate Professor of Pediatrics and Health Sciences and Technology at Harvard Medical School, and Chair of the Informatics Program at Children's Hospital, Boston.

SimbiosNews

BY JOY P. KU, PhD, DIRECTOR OF DISSEMINATION FOR SIMBIOS

Solving the 3-D RNA Structure Puzzle with NAST

For proteins, structure information leads to an understanding of function. The same turns out to be true for ribozymes, ribosomal RNAs, and some other recently discovered RNAs. But mapping out that three-dimensional (3-D) structure isn't always possible experimentally and structures that are obtained are often incomplete. The Nucleic Acid Simulation Tool (NAST) helps solve the 3-D puzzle, using what's known about a given RNA to generate a large number of plausible 3-D structures in a fully automatic way. In addition to advancing basic biology, such structural information could also potentially aid in the design of new RNAs—for example, as gene therapy tools.

"NAST is a powerful tool for exploring possible conformations of an RNA given a particular set of constraints," says **Alain Laederach, PhD**, a research scientist at the Wadsworth Center in New York and an assistant professor in the School of Biomedical Sciences at SUNY Albany. Laederach was one of the original developers of NAST.

Rather than model all the atoms in the RNA molecule, NAST uses what is called coarse-graining—it groups atoms together and represents them as a single particle. This means fewer computations are required so more results can be generated in a given time period.

But such a simplified system isn't accurately described by classical equations of physics. "We've lost a lot of information and can't use real physics because we no longer have real atoms," says **Magdalena Jonikas**, a graduate student in bioengineering at Stanford University and the lead developer of NAST. "So we built our own physics about how particles interact."

The NAST physics is based on the structural properties of coarse-grained representations of two ribosomal RNAs with known structures. That information was used to design an energy-based function that can produce realistic 3-D structures—helical parts of the molecule are identified and turned into helices, while non-helical portions are modeled after an average piece of ribosomal RNA.

Validation tests with NAST showed the average error for a prediction varied from 8-16 Angstroms, depending on the RNA and the available experimental data.

These results are encouraging, says **Nikolay Dokholyan, PhD**, an associate professor in biochemistry and biophysics at the University of North Carolina at Chapel Hill and the project lead for Discrete Molecular Dynamics (DMD), which has goals similar to NAST. "Predicting RNA structure is a very difficult task, perhaps more difficult than proteins."

NAST could also be useful for capturing snapshots of the RNA in motion, says Dokholyan. "RNA molecules can be quite dynamic and it would be important to have clusters of structures that show up during the lifetime of these RNA molecules. And in this case, NAST would do great."

"NAST is probably the best RNA structure prediction system based on molecular dynamics that's been published so far," says **Francois Major, PhD**, a principal investigator at the Institute for Research in Immunology and Cancer at the University of Montreal and project leader of MC-Fold and MC-Sym pipeline, an all-atoms approach for structure prediction. "If I had to use one such system today, I would use NAST." □

NAST generates a large number of 3-D structures that satisfy user-provided constraints. The image above shows the known crystal structure of a yeast tRNA and three subsets of five similar structures generated by NAST without reliance on crystal structure data. Reprinted from Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters, M. Jonikas et al., RNA, 2009, 15:189-199.

What You Need to Use NAST

NAST only needs information about the order of the A, U, G, C bases that make up RNA (the primary sequence) and the 2D map of the RNA, which shows how the bases pair up when the RNA folds back on itself (the secondary structure). Long-range tertiary interactions between bases can also be incorporated. NAST can be freely downloaded from http://simtk.org/home/nast.

Simbios (http://simbios.stanford.edu) is the National Center for Biomedical Computing located at Stanford University.

NewsBytes

Decoding Promotion

Despite their identical genomes, cells in the body develop distinct personalities—become neurons or liver cells, for instance—due to differences in gene expression. The mechanism that regulates this process has remained obscure, but a new study explains it using a simple thermodynamic model.

"Much of this phenomenon can be explained by a simple model of protein-protein and protein-DNA interactions," says principal author **Barak Cohen, PhD**, of the Washington University Medical School in St. Louis. "In our system there is no need to account for complicated chemical processes."

According to large-scale studies of eukaryotic genomes, gene expression is turned up and down when transcription factors interact with a zone of noncoding DNA located upstream from the gene—the gene's promoter. This interaction is complex, and can involve a variety of transcription factors operating in concert. Indeed, a typical promoter may include 20 or more sites that can each bind any one of about 250 known transcription factors. The number of possible promoters and their interactions is thus enormous, but data about their behavior is limited to a few thousand known promoters. "This makes it real hard to tease out the rules of gene regulation," says Cohen.

To make the problem tractable, Cohen and his collaborators built 2800 synthetic promoters each combining three to five transcription binding sites from about 20 known sites. Experiments on yeast cells showed that varying these mini-promoters for a gene yielded nearly three orders of magnitude variation in its expression. To analyze the promoter-expression relationship, the researchers invoked a thermodynamic model developed in earlier studies. In this model the interactions between proteins and their binding sites either help or hinder the recruitment of RNA polymerase—the molecule needed to build RNA from the DNA—to the promoters. The researchers "trained" the model using measured gene expression levels for a set of about 400 promoters, and tested it on an independent set of another 83 promoters.

Courtesy of Barak Cohen; graphic by Michela Hunt.

The trained thermodynamic model explained nearly 50 percent of the variation in gene expression for the training set, and about 44 percent of the variation for the independent set. In contrast, empirical models relying on genomic data explain less than 25 percent of the variation in gene expression, says Cohen. The system also showed how weak binding sites cooperate to regulate gene expression, an effect that prior models failed to address. When applied to actual yeast genome data, the system found that Mig1, a transcription factor associated with glucose metabolism, regulated several additional genes not previously known to be regulated by this protein. "This is remarkable because Mig1 is one of the most widely studied transcription factors," says Cohen.

In addition to shedding light on gene regulation, the findings could also facilitate *in silico* engineering of promoters with completely novel expression patterns, says Cohen. Such custom-designed promoters could be a boon for stem cell development, tissue engineering, regeneration, and similar areas. As a step towards this goal, the researchers plan to extend their work to mammalian cells, Cohen says.

"This paper is an important advance developing quantitative models for transcriptional regulation," says **Eran Segal, PhD**, of the Weizmann Institute of Science in Israel. "It shows on a large scale what has been demonstrated previously on smaller sets of genes in fly and bacteria." **Paturu Kondaiah, PhD**, of the Indian Institute of Science in Bangalore agrees with this assessment, but points out that transcription factors behave differently depending on their conformation, and can also recruit co-activators or co-repressors. "The next step is to take these effects into account," he says.

—By Chandra Shekhar, PhD

The Brain in Transition

Patients with schizophrenia and other psychotic disorders are known to have adverse brain changes, such as reduced volume—but it's unclear what comes first, the disease or the abnormality. Now, for the first time, researchers have shown that the brain is actually shrinking as psychosis unfolds. The results appear in the January 10 issue of *Schizophrenia Research*.

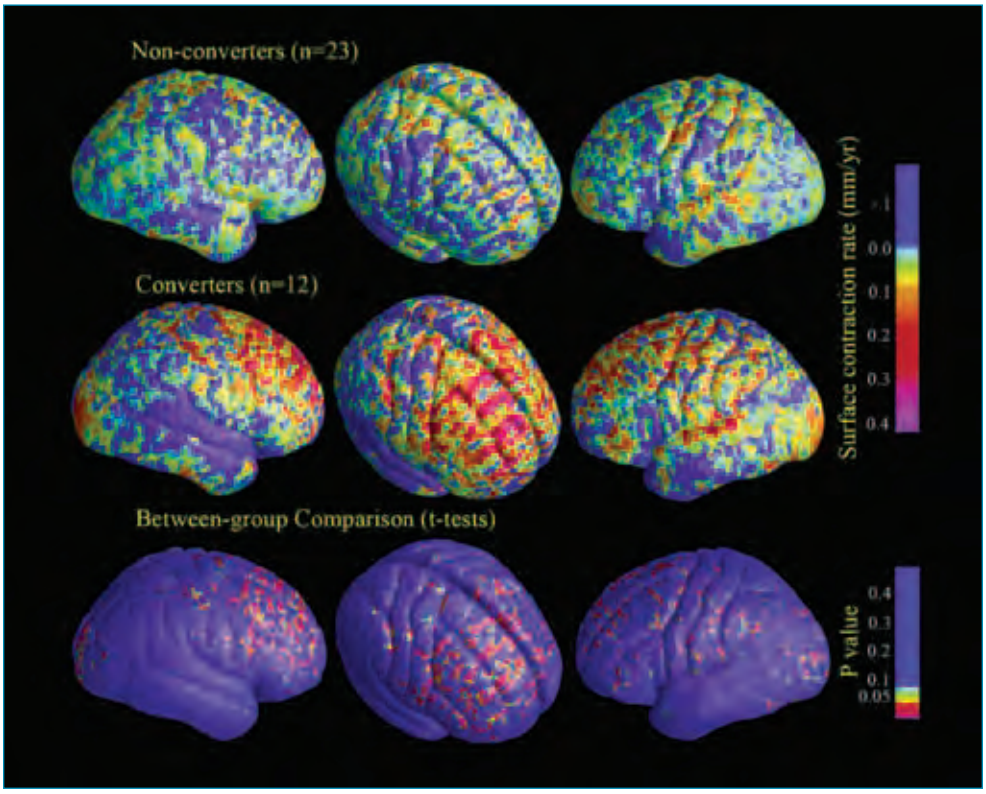
"We found that people who go on to develop psychosis have a different profile of neuroanatomical changes than those who do not," says **Tyrone D. Cannon, PhD**, professor of psychology, psychiatry, and biobehavioral sciences at the University of California, Los Angeles. The findings may have implications for predicting and preventing psychosis.

Cannon and his colleagues took pre-

"Our approach allowed us to detect more subtle anatomical changes in the brain, which is critical because we would not expect the changes associated with onset of psychosis to be so gross as to be detectable using standard voxel-based methods," Tyrone Cannon says.

morbid brain MRI images of 35 individuals who had never had a psychotic episode but were considered at "ultra-high" risk based on early symptoms or a strong family history. They re-scanned their brains after an average follow-up of 1.3 years—during which time 12 developed psychosis.

Previous studies had considered losses in brain tissue density, a voxel-level measure of brain volume (typical resolution on the order of 1 cubic millimeter). Cannon's team used a higher-resolution measure of volumetric brain change—the brain contraction rate. This parameter is calculated by transforming MRI scans into 65,000-point maps of the brain's surface and determining how fast the surface points are contracting between sequential scans. They found that the prefrontal lobes of subjects who progressed to psychosis were contracting significantly faster—by about 0.2 millimeters per year—than those of subjects who did not progress. "Our approach allowed us to detect more subtle anatomical changes in the brain, which is critical because we would not expect the changes associated with onset of psychosis to be so gross as to be detectable using standard voxel-based methods," Cannon says.



Shrinking Brains. *The top panels show the average rates of surface contraction in different regions of the brains of 12 high-risk subjects who went on to develop psychosis (converters) and 23 who did not (non-converters). Red and pink regions are contracting the fastest. The bottom panel shows regions where the converters' brains were contracting significantly faster than non-converters'. Yellow, red, and pink regions had the most statistically significant differences. Reprinted from Sun, D., et al., Progressive brain structural changes mapped as psychosis develops in 'at risk' individuals,* Schizophr. Res. *(2009) 108(1-3):85-92.*

Though the study is interesting, it is small and lacks a healthy control group—which makes it difficult to tell how much of the detected changes are due to random variation and normal aging versus disease, comments **R. Grant Steen, PhD**, associate professor of psychiatry at the University of North Carolina School of Medicine. Also, it took an average of eight months to re-scan subjects with disease after their initial psychotic episode, so the timing of the changes is not entirely clear and could be related to treatment, he says.

Medication is an unlikely explanation since its use was limited and unrelated to brain contraction rates, Cannon replies. Still, he agrees, "the full significance of the findings awaits confirmation in large, multisite, longitudinal imaging studies that are currently underway."

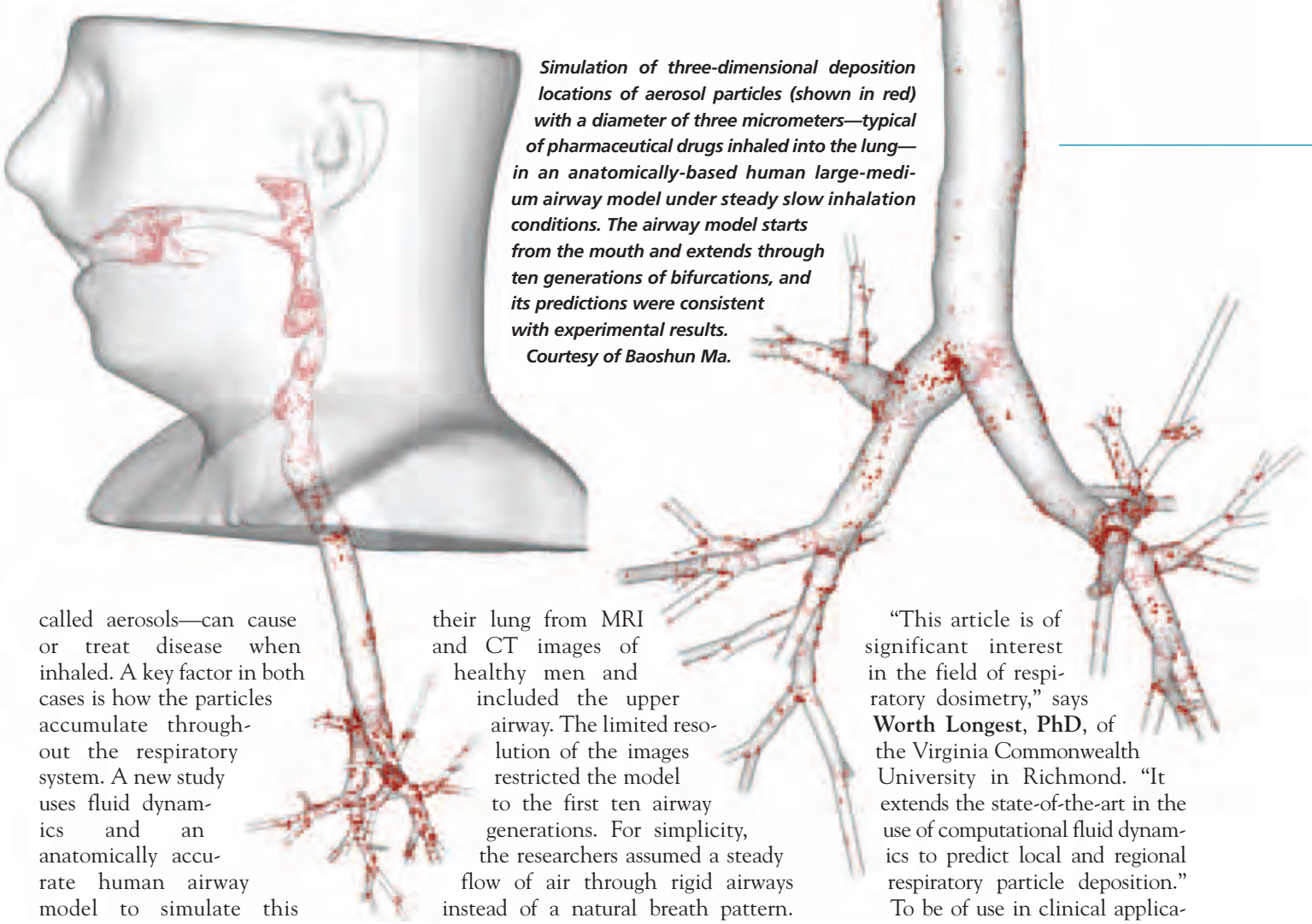
If the changes observed do turn out to be a cause of the onset of schizophrenia and associated disorders, "it may eventually be possible to provide treatment in high risk individuals—to delay or prevent the onset of psychosis," Cannon concludes.

—By Kristin Sainani, PhD

The Fate of Inhaled Particles

New computational model simulates how particles in the air get deposited in the lungs during breathing

Depending on their nature, microscopic particles suspended in air—



Simulation of three-dimensional deposition locations of aerosol particles (shown in red) with a diameter of three micrometers—typical of pharmaceutical drugs inhaled into the lung—in an anatomically-based human large-medium airway model under steady slow inhalation conditions. The airway model starts from the mouth and extends through ten generations of bifurcations, and its predictions were consistent with experimental results. Courtesy of Baoshun Ma.

called aerosols—can cause or treat disease when inhaled. A key factor in both cases is how the particles accumulate throughout the respiratory system. A new study uses fluid dynamics and an anatomically accurate human airway model to simulate this process, potentially paving the way for improved disease understanding and patient-specific drug delivery.

"It is one of the first computational studies that uses anatomically correct models to predict aerosol deposition," says principal author **Kenneth Lutchen, PhD**, of Boston University, principal author of the study published in the February 2009 issue of *Annals of Biomedical Engineering*.

Starting with the windpipe, airways in the human respiratory system branch out, producing about 23 levels of branching or "generations." The resulting structure includes nearly 10 million microscopic airways, making it hard to study aerosol deposition. According to Lutchen, experimental methods relying on *in vivo* rat studies or lung-shaped casts have yielded useful, but preliminary, data. Prior computational studies have dealt with more complex respiratory structures, but typically used idealized lung models instead of the actual anatomy. Further, many of them ignore the upper airways where most of the deposition occurs, Lutchen says.

In contrast, Lutchen and his collaborator **Baoshun Ma, PhD**, modeled their lung from MRI and CT images of healthy men and included the upper airway. The limited resolution of the images restricted the model to the first ten airway generations. For simplicity, the researchers assumed a steady flow of air through rigid airways instead of a natural breath pattern. They then used a computational fluid dynamics framework with a standard turbulence model to simulate aerosol deposition for different particle sizes and airflow rates. Results indicated that large particles (with a diameter of 30 micrometers—about the width of a human hair) end up mostly in the mouth and upper throat, whereas small (1 micrometer) particles typical of pharmaceutical drugs inhaled into the lung spread out more evenly. Typically, the left lung absorbed more particles—as much as 5 times more for some parameter settings—compared to the right lung. "These predictions are consistent with experimental data," says Lutchen.

Inhaled aerosols have emerged as an important method for delivering drugs for lung-related conditions ranging from asthma to cystic fibrosis. However, proper dosing requires accurate, patient-specific prediction of aerosol deposition patterns under a variety of conditions. Lutchen hopes that the new approach will eventually facilitate this task. "This model will tell you what particle sizes and inhaled volumes you need to get the desired dose for a specific patient," he says.

"This article is of significant interest in the field of respiratory dosimetry," says **Worth Longest, PhD**, of the Virginia Commonwealth University in Richmond. "It extends the state-of-the-art in the use of computational fluid dynamics to predict local and regional respiratory particle deposition." To be of use in clinical applications, however, the system should be extended to include transient effects over a breathing cycle, effects of airway wall motion, and a more robust turbulence model, he adds.

—By Chandra Shekhar, PhD

RNA Families Set Up House in Wikipedia

For scientists submitting to the journal *RNA Biology*, the publishing guidelines now include a new task: Submit a Wikipedia entry. In collaboration with the RNA database Rfam, the journal recently launched a new section, RNA Families, that requires a corresponding peer-reviewed Wikipedia article along with each article published in the section.

"It is so globally important to have knowledge accessible to everybody," says **Renée Schroeder, PhD**, editor-in-chief of the journal.

The new section, dedicated to descriptions of non-coding RNA families, debuted in the January/February/March issue of the journal, with one article and its corresponding Wikipedia entry. The entries are not meant to exactly mirror the scientific literature, Schroeder says. "In a research article

you have the way you acquired the knowledge, and in Wikipedia you have the results," she says.

This is the first instance of such a link between Wikipedia and a scientific journal, says **Alex Bateman, PhD**, co-director of Rfam, an open-access database of non-coding RNA families coordinated by the Wellcome Trust Sanger Institute in Cambridge, UK. In 2007, Bateman and his colleagues linked the database to Wikipedia. Editing of the Wikipedia articles automatically updates the database. In fall 2008, he brought his idea for a new publishing paradigm to *RNA Biology*.

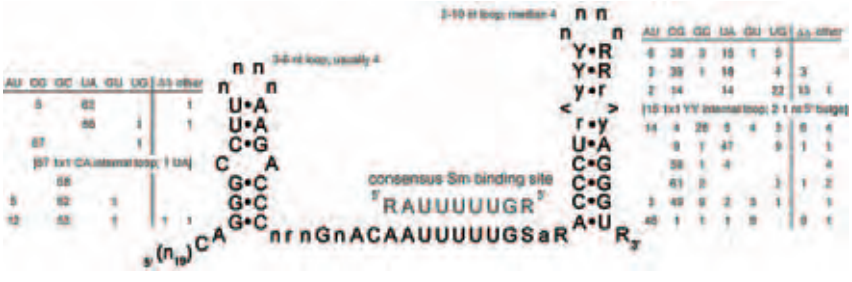
Bateman thinks this is an exciting step for a scientific journal. "It wouldn't be reasonable to claim that these articles were going to change the world," he says. "But the important thing is that the model is really interesting. Hopefully this can be an experiment that other journals can follow in other areas of science."

Schroeder points out that this model won't work for every scientific journal. The article subjects must easily translate to the encyclopedic format, and the site is meant to be a source of accurate information, so "there shouldn't be too much that hasn't been tested and retested," she says.

The Rfam and *RNA Biology* entries fall under the rubric of the Molecular and Cellular Biology wikiproject, which is working to improve all molecular biology, biochemistry and cellular biology entries. **Tim Vickers, PhD**, director of the wikiproject and postdoctoral fellow at Washington University, thinks the decision by *RNA Biology* is a step in the right direction toward getting more scientists involved with updating and maintaining Wikipedia. "Obviously we scientists all like to publish papers, but if you just do that and don't reach out and tell people why your work is important, that's a big chunk missing," he says. "Editing Wikipedia and giving the general public a good summary of the science in your field, that's almost as important as publishing scientific papers."

—By Rachel Tompa, PhD

"Obviously we scientists all like to publish papers, but if you just do that and don't reach out and tell people why your work is important, that's a big chunk missing," Tom Vickers says.



A figure from the first Wikipedia entry tied to an RNA Biology article, entered into Wikipedia in November 2008. The entry and article, published in the January/February/March 2009 issue of* RNA Biology *describe the SmY family of non-coding RNA molecules found in some nematode species.

Predicting Vaccine Efficacy

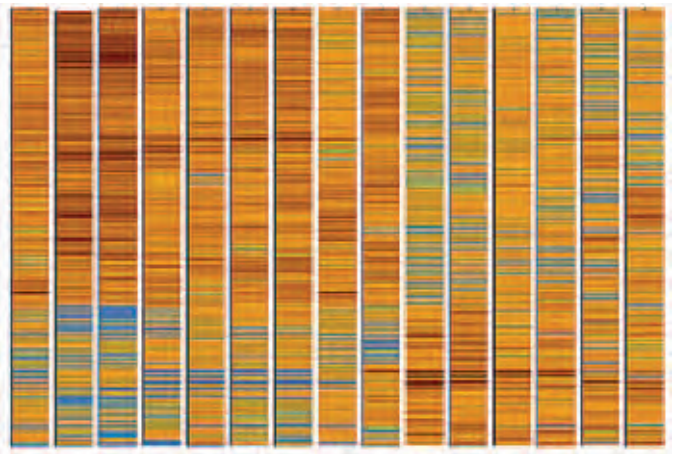
Researchers developing a new vaccine currently have no direct way of predicting its efficacy short of exposing patients to the disease. A new study that combines gene expression data with advanced computational analysis provides the first evidence that the vaccine-induced immune response can be predicted.

"We develop vaccines but can never say how effective they will be," says **Bali Pulendran, PhD**, a researcher at the Emory Vaccine Center in Atlanta who led the study, published in *Nature Immunology* in November 2008. "Only after exposure do we really know."

To gauge a vaccine's effectiveness, scientists evaluate indicators of the so-called "adaptive" immune response, which develops over time. The titer—a measurement of the concentration—of long-term antibodies in the blood is one indicator. The number of killer T cells is another. But a more complete profile of the early or "innate" immune system reaction could help researchers screen vaccine candidates or help identify individuals whose adaptive immune systems don't respond.

To develop such a profile of the immune response, Pulendran and his colleagues monitored patients given the yellow fever vaccine—a vaccine that has been given to more than 600 million people and is considered one of the most powerful ever developed, proving effective 80 to 90 percent of the time. In two sets of volunteers (15 in the first group and 10 in the second) Pulendran's group sought to correlate patients' innate (early) immune response to the vaccine with the later T cell response. Several cytokines and 65 genes responded to the vaccine in significant ways, but there was no apparent link between this innate signature and the subsequent T cell reaction.

To zero in on what was evidently a subtle connection, the researchers looked more broadly at the gene expression signatures for the first set of patients. They found 839 genes whose expression correlated with the T cell response. Using these data and a supervised learning algorithm developed by **Eva Lee, PhD**, at the



The researchers used correlation cluster analysis of expressed genes to confirm that subjects could be sorted clearly into two categories: "high" or "low" responders to the vaccine, based on the strength of T cell response. Courtesy of Bali Pulendran. Reprinted from Querec TD et al., Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans, **Nature Immunology** *(2009) 10(1):116-125 with permission from Macmillan, publishers.*

Georgia Institute of Technology, they pulled out eight different genetic signatures from data from the first group that strongly predicted the T cell response in the second group of volunteers. The researchers also used that algorithm to generate signatures predicting the titer of long-term antibodies.

In the case of both T cells and antibodies, the researchers were particularly interested in a small number of genes that featured in all the predictive signatures. These genes form a core set that doctors could potentially monitor to predict how effective a vaccine will be in a patient. Pulendran also hopes that by working to replicate the innate reaction to yellow fever, scientists may be able to make potent vaccines against other pathogens.

"If the approach could be extended to development of vaccines against different sorts of pathogens, it would be a real advance," says **Larry Stern, PhD**, an immunologist at the University of Massachusetts. "The key here is whether the same signature would be induced by other pathogens," he says, noting that even if the method works only for related pathogens, such as dengue fever and West Nile virus, that would still be a very valuable contribution.

—By Kaspar Mossman, PhD

A Model Neuron

For patients suffering from nerve damage, neural regeneration is a faint hope. It rarely happens naturally, and attempts to coax new growth often fail. Researchers are trying to develop scaffolds to guide regenerating neurons in the body. But the best way to guide neural growth on these substrates remains unknown. So *in vitro* studies of neuronal behavior on these templates are a key first step. But such studies largely rely on trial and error rather than engineering principles.

Now, scientists have developed a computational model to predict the first stage of neural development, neuron polarization. Their model, published in the February issue of *Annals of Biomedical Engineering*, could yield powerful predictions for better scaffold design in neural tissue engineering.

"Our work is unique as it is the first effort of its kind to quantitatively model the interactions of the neuron with the substrate," says **Muhammad Zaman, PhD**, assistant professor of biomedical engineering at the University of Texas at Austin.

Directing neuron growth on an artificial substrate is no easy feat. To lead to nerve regeneration, the neurons must polarize in the same direction, but immature cells send out multiple tendrils in all directions initially. The projection that grows the longest eventually becomes the axon, the path for sending out electrical signals; the others become dendrites, the stimulus receptors for the neuron. These projections' fates can be influenced by various external cues, both chemical and physical.

For unknown reasons, physical factors such as ridges dominate over chemical cues *in vitro*. That is, if an immature neuron is faced with chemical cues on one side and ridges on the other, it will tend to polarize toward the ridged side, extending its axon along one of the grooves.

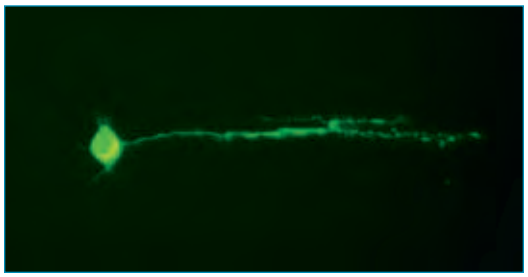
To model the cell's reaction to its surroundings, Zaman and his colleagues broke neuron polarization into several small steps, using probabilities at each step to predict the cell's next choice in projection growth. They introduced parameters based on known factors, such as the physics of the internal forces acting on the projections, how projections behave on different substrates and how they react to different chemical cues.

Their model accurately reproduced known results, and also revealed that ridge size is important to a neuron. If the ridges are too small or too wide, the neural projections view them as a continuous surface, and chemical cues will win out. For the kinds of cells in Zaman's experiments, the best ridges were between two and 10 microns wide.

"The cells seem to like persistence," Zaman says. Once a projection starts down a ridge, it is like a car on a one-way road. With only one direction to travel, growth is much faster. But if the ridge is too wide or too narrow, the cell no longer sees the road.

"There is a lack of engineering rigor in the whole area of tissue and regenerative engineering," says **Gabriel Silva, PhD**, assistant professor in bioengineering at the University of California, San Diego. "I think the approach that these authors have taken is exactly what's needed, which is a systematic, quantitative, rigorous engineering-type model that can guide the design of experiments and materials."

—By Rachel Tompa, PhD □

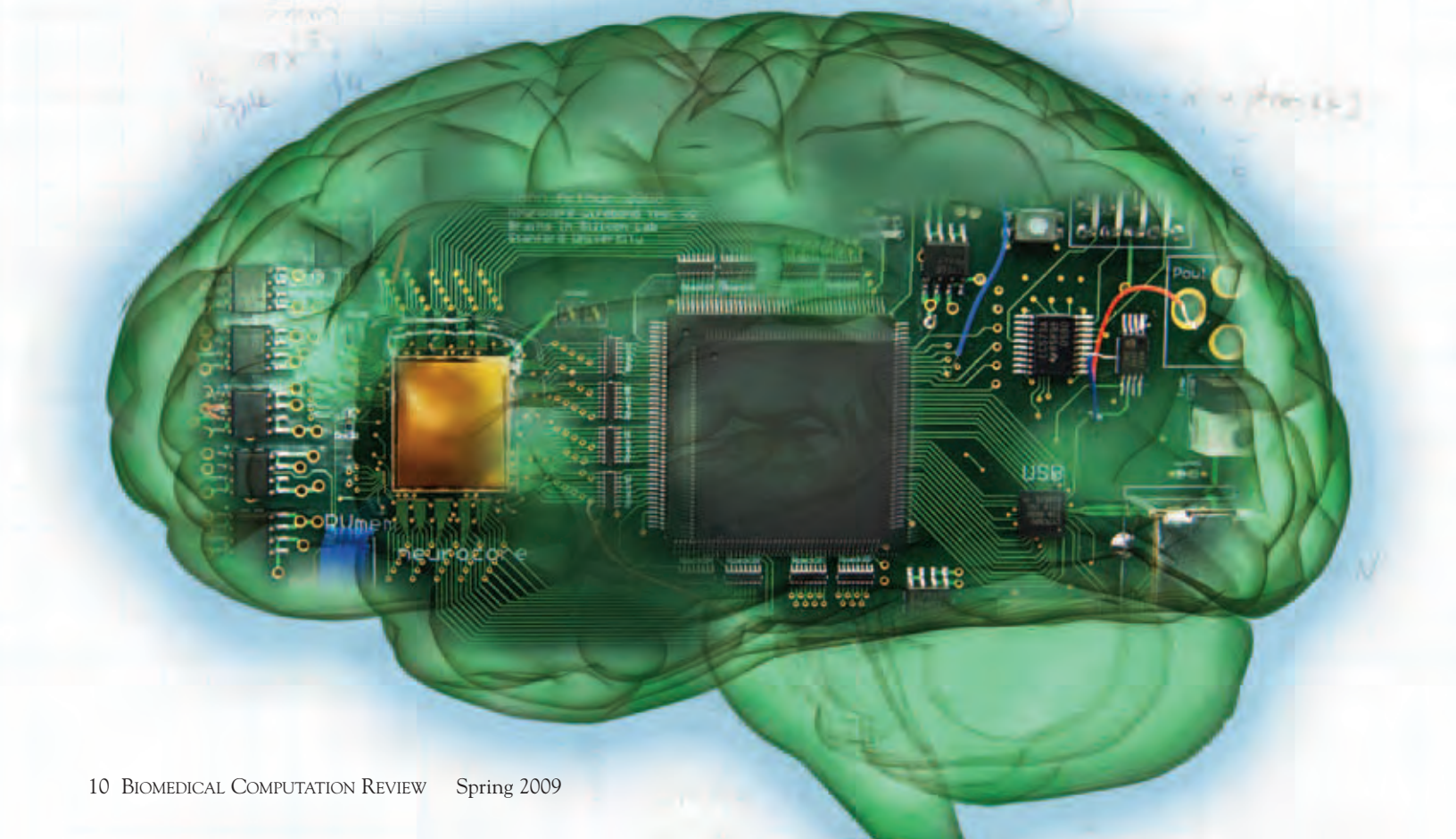


A single rat neuron has a decision to make. When this cell was immature, it was placed on an artificial substrate in between immobilized nerve growth factor (on the left) and a surface of two micron-wide ridges (on the right). The projection that eventually turned into an axon grew along the ridged side. Zaman's computational model of neuron growth reproduced this outcome and also identified axons' preferred ridge size. Photo courtesy of* Natalia Gomez, PhD, *formerly at the University of Texas at Austin.

For a century, neuroscientists have dissected, traced, eavesdropped on, and are now compiling a seemingly endless cast of players in the nervous system. As we keep gathering more and more molecular details, how do we know when we know enough?

Reverse Engineering the Brain

By Roberta Friedman, PhD



Some have decided it's time to just go ahead and create a brain *in silico*. And to a surprising extent, they've done it: Labs around the world are populated with autonomously functioning brains based on what we know so far. These simulations match what happens at the cellular level in the brain when the nerve cells, or neurons, that make up the brain pump ions and produce electrochemical activity that propagates across the synapse from one neuron to another. Robots or avatars activated by these engineered brains are directing movement, perceiving visual objects, and even responding to rewards—exhibiting behaviors associated with our "thinking" brains.

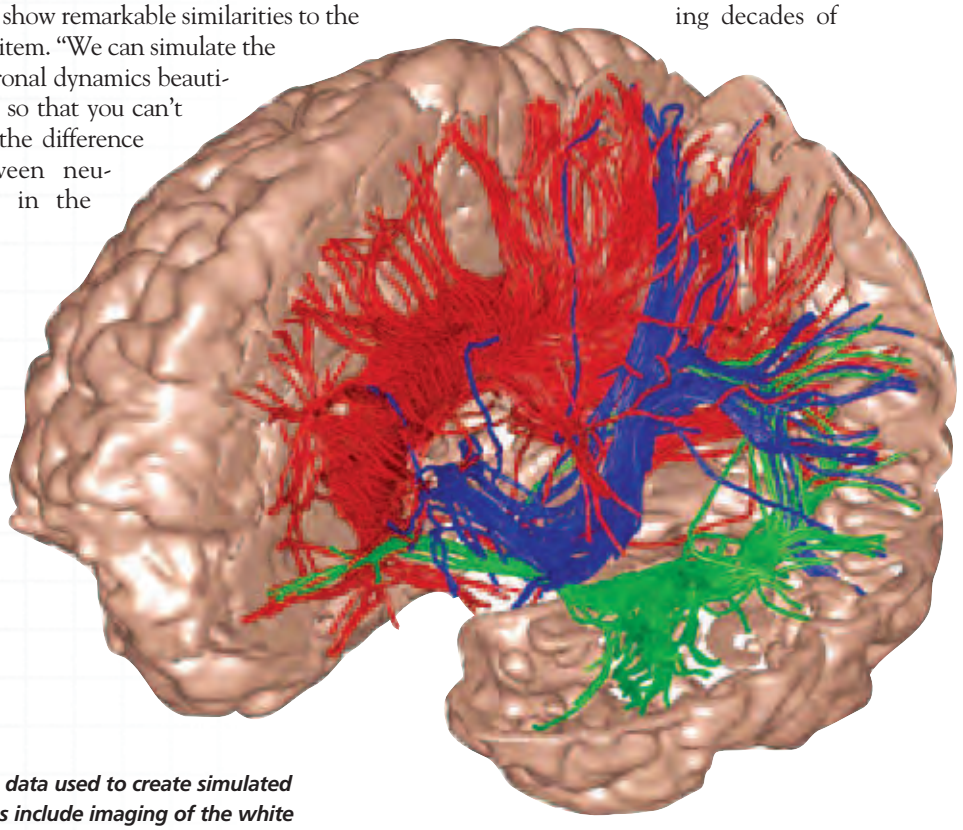
Eerily, the most recent simulations show the same oscillating rhythms seen when physicians record human brain waves using an electroencephalogram (commonly known as an EEG).

Computer simulations of the brain already allow experiments impossible to carry out with animals. "As good as modern neuroscience is—and it has been brilliant over the last two decades—we can't really sample every neuron and every synapse as they are performing a behavior," notes consciousness researcher **Gerald Edelman, MD, PhD**, director of the Neurosciences Institute and chair of neurobiology at the Scripps Research Institute in San Diego, California.

Researchers are looking to develop even more efficient simulated brains to help produce computers that can think while at the same time accelerating neuroscience. Ultimately brain simulations promise the ability to study the effect of drugs and disease and aid in the design of new therapeutic strategies.

HOW TO BUILD A BRAIN 101

To build a simulated brain requires a vast amount of detailed information about this complex organ, starting from its basic unit (the neuron) and building up to the complex network of connections between them that produces perception and cognition. None of this information is available from any single species. Much of the data on how individual neurons behave comes from rat studies. Observations of primates have provided data about how neurons are wired together across brain regions. And cat and human research led to an understanding of the finer, local circuitry in specific areas of the brain. Nevertheless, the basics of the nervous system are similar enough across mammals that Edelman and others have cobbled together chimeric, rudimentary brain simulations that show remarkable similarities to the real item. "We can simulate the neuronal dynamics beautifully so that you can't tell the difference between neurons in the model and real neurons," Edelman says.



Brain data used to create simulated brains include imaging of the white matter fibers in the brain using a technique called diffusion tensor magnetic resonance imaging (DTMRI). Reprinted from Izhikevich et al., Large-scale model of mammalian thalamocortical systems,* Proceedings of the National Academy of Sciences *(2008) 105:3593-3598.

To build a simulated brain, Edelman and others start with what's known about the neuron, a cell that actively maintains a separation of charged ions across its membrane. Specific channels in the membrane allow certain ions in, and these are quickly pumped back out, or sequestered internally. But when a certain threshold of charge is reached the neuron fires a spike of current toward an adjacent neuron.

Here, at the synapse—a microscopic gap between each nerve cell—current becomes chemistry (and here is where drugs alter that chemistry). A spike wave arriving at the synapse triggers the release of neurotransmitters—to activate the next cell—provided enough inputs arrive in a very short time. Sufficient impulses strengthen the synapse. Neglected, the synaptic strength weakens and the particular connection is diminished. In the developing brain, synapses are ruthlessly pruned. This is what neuroscientists have uncovered during decades of listening in with electrodes a hundred times finer than a human hair. And this is the basic information that Edelman and others use to construct their simulated neurons.

To determine how these neurons are connected, simulators turn to microscopists and their latest technologies. Techniques from immunology have brought incredible resolution on the molecular level: cells containing particular molecules can be tagged by dye-bearing antibodies so that researchers can distinguish them from from their

"We can simulate the neuronal dynamics beautifully so that you can't tell the difference between the model and real neurons," says Gerald Edelman.

The simulation of major brain centers and their microcircuits is able to generate its own inherent activity—similar to what is seen in real brains.

fellows and follow their links to one another. Scanning electron microscopy has been able to home in on the fine molecular scale at the synapse.

Knowing how individual neurons function and how they're connected will not make a brain work. Simulators need to know the bigger picture of brain area networks. To understand the function of brain regions, neuroscientists initially used data from scalp EEG and depth electrodes placed within the brains of living patients and animals, as well as observational reports such as from accidents that selectively damaged specific brain areas. These days computer-analyzed imaging can reveal additional details of the normal brain. Simulators employ all of these lines of evidence, and still seek more.

But none of this data could produce an engineered brain without huge advances in computer simulation. Alan Turing's idea for a calculating machine at the end of World War II laid the groundwork. Warren McCulloch and Walter Pitts set forth the initial properties of an electronic replica of a neuron in 1943. In the mid 50s, IBM researchers ran a simulation of 512 neurons.

These are the lines of investigation picked up by Edelman who entered the field of reverse brain engineering after receiving a Nobel Prize in Physiology or Medicine (for immunology research) in 1972.

ACHIEVING AUTONOMY: EDELMAN'S SIMULATED BRAIN RHYTHMS

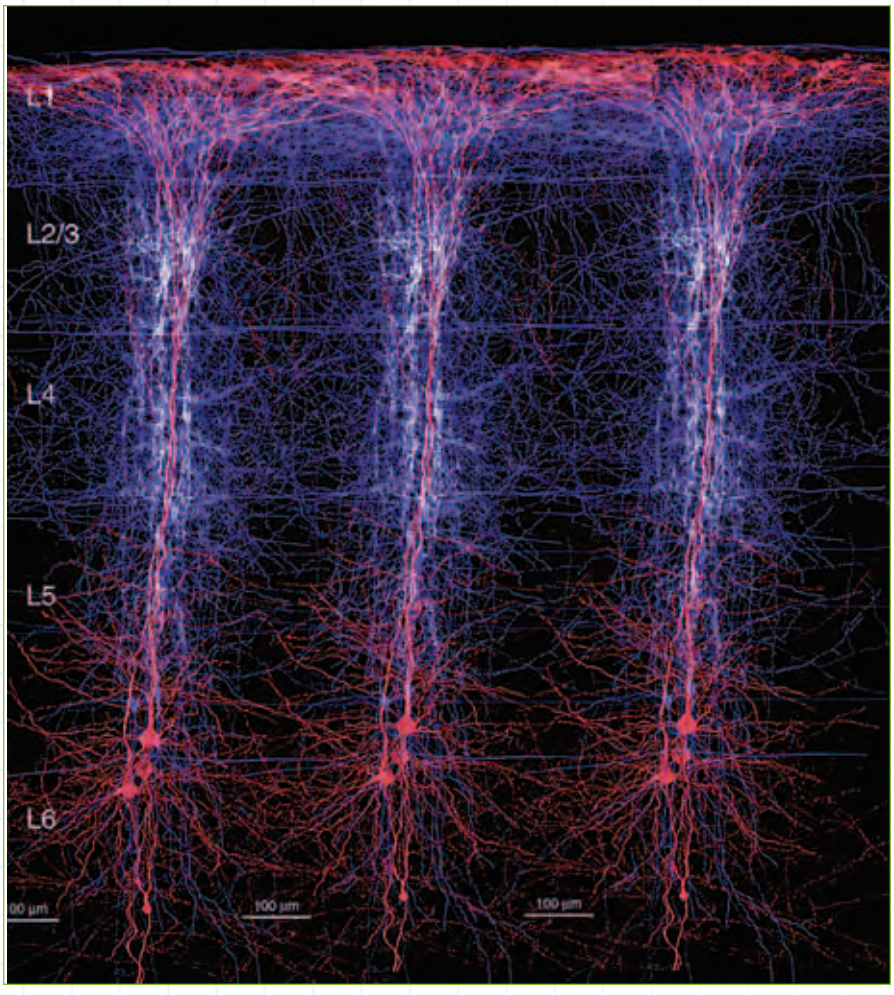
The latest of Edelman's simulations incorporates the known circuitry from the thalamus, a central command post in the core of the brain. The thalamus drives the cortex (the brain's covering layers—also modeled) into and out of sleep and through various levels of alertness. When the thalamus no longer talks to the cortex, vegetative states result. The model also includes circuitry of the hippocampus, a seahorse shaped curl of brain tissue beneath the temples, which is crucial for long-term memory, a region attacked in Alzheimer's disease.

Once enough of the brain's macro and microcircuitry is simulated, the *in silico* model is able to generate its own inherent activity—similar to what is seen in real brains. "When you stimulate the neural model, it takes off on its own and is constantly active," Edelman says. "We've never succeeded in doing this before." Moreover, oscillating waves of synchronous neural firing not explicitly built-in emerged spontaneously, the researchers reported in the March 4, 2008, *Proceedings of the National Academy of Sciences*. The researchers also were able to induce and reproduce spontaneous, low-level activity at the synapses—called miniature postsynaptic potentials or minis. The results suggest that, as a real brain develops in a fetus, minis like these might prime neurons for action.

EAVESDROPPING ON SIMULATED NEURONS: THE BLUE BRAIN PROJECT

Edelman is not alone in simulating the brain. **Henry Markram, PhD**, co-Director of École Polytechnique Fédérale de Lausanne (EPFL), in Lausanne, Switzerland, directs the data-intensive Blue Brain project.

Edelman's group relied on a top-down approach based on global network properties of the brain and mathematical formulas to reproduce known types of neuron behavior. In a complementary approach, Blue Brain focuses



Cells in the cortex form columns. In this image the red neurons, called pyramidal cells, are revealed to be entwined by blue fibers from other, inhhibitory neurons that slow their firing. The layers of the column are indicated by the numerals to the left: L1, at the surface of the brain, through L6, the deepest cortical layer. Pyramidal cells, which receive messages along their extensively branched fibers, and send long fibers out to other brain areas or down to the spinal cord, are crucial in movement control and in cognition. They have their cell bodies in layer 5 of the cortex, and the main receiving fiber, the apical dendrite, rises up to the surface, layer 1. ©BBP/EPFL

on exact structural and molecular details to model a particular piece of the brain, building up from exact details of individual neurons, Markram says. "We are constrained by biology. There are so many theoretical ways to do it you would be lost forever. Biology has chosen a certain way and when you choose that, it becomes easier, not more difficult."

"We are constrained by biology," says Henry Markram. "There are so many theoretical ways to do it you would be lost forever. Biology has chosen a certain way and when you choose that, it becomes easier, not more difficult."

Data for the Blue Brain project was gathered using a key innovation: the ability to record ion signals from many neurons at once using what's called a multiple unit patch clamp technique. By eavesdropping on the interactions among neurons, researchers learned what synaptic currents were being generated and where.

In addition, they gathered data on gene activity within neurons—as an indicator of which discrete ion channels are present. In most neurons, a dozen or more types of these pores regulate ion flow. The Blue Brain simulation specifies which ones are present in each neuron. They also captured the precise connecting points of each neuron, by injecting dye once they were done recording the electrical activity. "The details are accurate, down to the micron," for each contact point of each nerve fiber, adds **Phil Goodman, MD**, professor of Internal Medicine and Biomedical Engineering at the University of Nevada, Reno, who collaborated on Blue Brain.

"It is a simulation, in time and space, of cells in real life."

So far, the project has reproduced the architecture and electrical properties of a single cortical column of a two-week-old rat. The living cerebral cortex is comprised of millions of such columns, with each column consisting of a vertical stack of six layers of over 400 types of neurons. The cortex column is has a blueprint which is quite similar from mouse to man and across brain regions with only subtle variations.

The Blue Brain researchers can probe the simulated cortical column with simulated electrodes. As in Edelman's lab, once a few stimulations are fed in, the simulation keeps going with its own intrinsic activity. For example, if thalamic fibers arriving at a deep layer of the cortical column are stimulated, the activity spreads, and finally the most superficial layer lights up. Markram notes that laboratory experiments failed to make this observation because they failed to listen in at the superficial layer. Thus, the simulation has already generated observations that could easily be missed in the lab, suggesting how simulations can guide brain research.

In the next six months, the Blue Brain project plans to publish "key insights never seen before in the neocortical column," Markram says. "By the end of the year we will publish the entire circuit with the blueprint. It's like the genome map—it's a comprehensive description of the neocortical column."

"It took 15 years to get the data for this small piece of brain," Markram says. "Every week the model becomes more biological," he adds. "It's very much like a real little bit of tissue." And now that they've built one cortical column, building another is a simple task. "We can (now) push a button and build an unlimited amount of neurons automatically." The goal is to build up the brain from this discrete piece.

There's still a need for more data about brain anatomy, Markram says. Some neuroanatomists are working toward a map to locate every single neuron in the human brain. This so-called "connectome," says Markram, will undoubtedly help the next generation of brain simulation. **Javier DeFelipe, PhD**, from the Cajal Institute in Madrid has joined the project to provide Blue Brain with data at the electron microscopic level. "Blue Brain is hungry for data," Markram says.

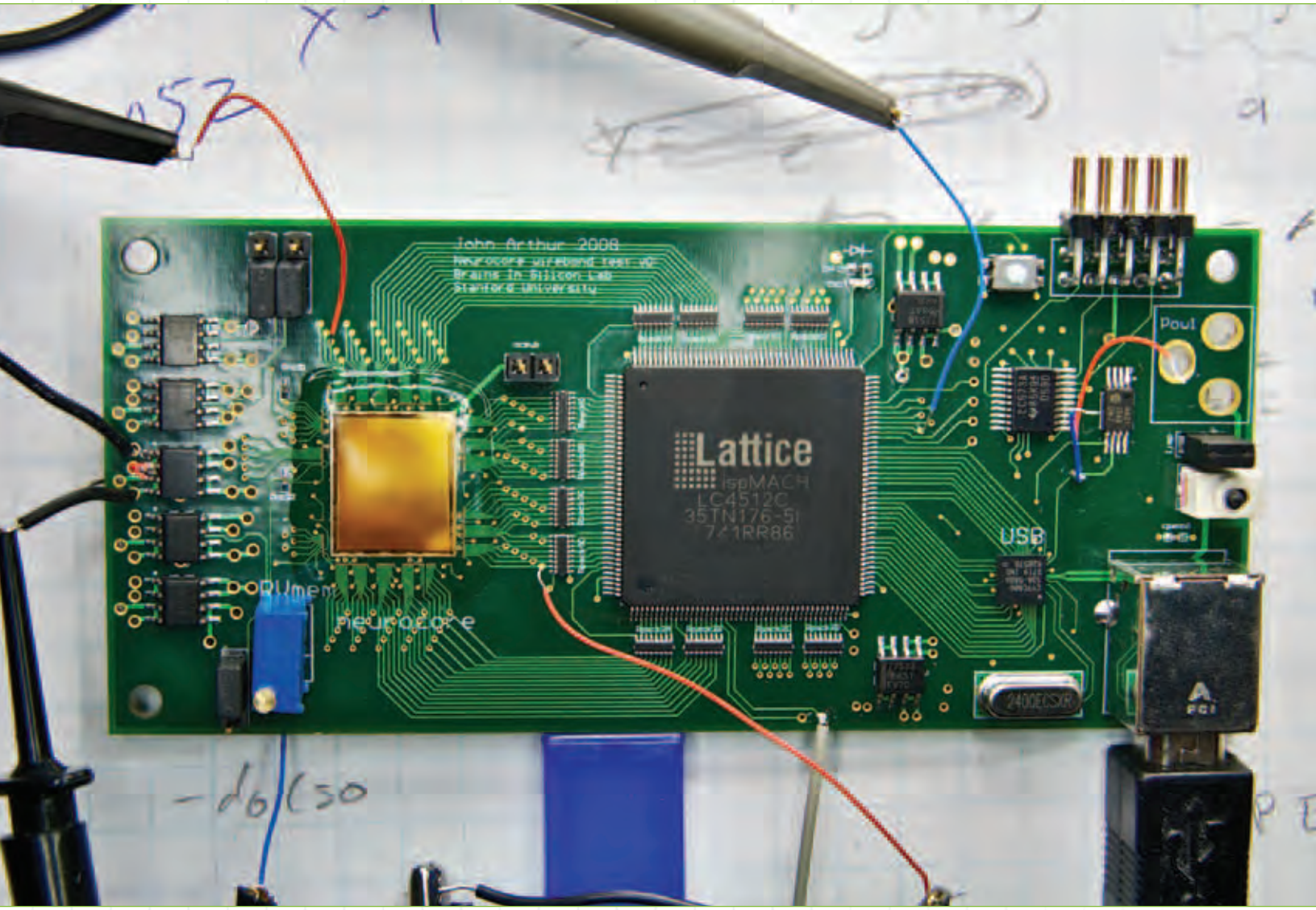
"We can (now) push a button and build an unlimited amount of neurons automatically," Markram says.

A POCKET-SIZED SIMULATED BRAIN: NEUROGRID CHIPS

To simulate the human brain, to really know how we think, is not a research problem many can take on. Electricity alone for a supercomputer to simulate a million neurons eats through $200,000 a month, restricting brain simulations to the very few able to get that kind of funding. "This is something we want to change," says **Kwabena Boahen, PhD**, associate professor of bioengineering and the principal investigator for Brains in Silicon at Stanford University. "If we can create a tool to allow a lot of people to play at this scale, as a community, we will progress faster."

To that end, Boahen and members of his lab have developed the Neurogrid chip with funding from the NIH Director's Pioneer Award Program. No bigger than a fingernail, 16 of these chips will be assembled in an iPod-sized device that can do what a supercomputer does—simulate a million neurons—at only $40,000. The Neurogrid chips have been received from the silicon foundry and should allow the group to emulate a million neurons in the cortex in real time at a thousandth of the cost of supercomputing. "Everybody can play now," says Boahen. "Not just IBM."

The Neurogrid chip works the same way the brain does, Boahen says. Its circuitry is analog because that is the way neurons compute: They sum their inputs continuously, not discretely. It is only past a certain threshold that the process



A Neurogrid chip (Neurocore) mounted on a test printed circuit board. Each Neurocore has 65,536 programmable neurons in 162 mm² of silicon. Sixteen Neurocores connected together will form the first hardware system with over one million model neurons operating in real-time, while consuming less than 10 Watts and taking up less space than a paperback book. Courtesy of Rodrigo Alvarez and Kwabena Boahen, Brains in Silicon, Stanford University.

becomes digital, generating a spike of electrical activity—all or nothing (the spike is like a one; lack of spike, a zero).

"Instead of using transistors as switches, I can build a capacitor and sum currents and get the same voltage on the capacitor that a neuron makes," Boahen says. With one transistor and a capacitor, he says, you can solve a differential equation that would take a thousand transistors in the traditional arrangement in a computer.

Dharmendra S. Modha, PhD, manager of cognitive computing at IBM's Almaden Research Center in San Jose, and a collaborator of Boahen's, says "Neurogrid is a genuine technical breakthrough. It has the potential to transform computational neuroscience."

Modha cites the mouse cortex model that his team has created as a prime example. Their simulation shows the oscillations present in living brains just as Edelman's do and runs "in near real time" on a 4096 processor BlueGene/L supercomputer with a terabyte of memory. Modha explains that even so, that was still seven to ten times slower than the action in the rodent brains.

Obviously, the requirements for brain simulation outstrip the available hardware unless alternatives such as Neurogrid or others are achieved. "The

"If we can create a tool to allow a lot of people to play at this scale, as a community, we will progress faster," says Kwabena Boahen, who, with colleagues, has developed the Neorogrid chip.

brain that Mother Nature has created is enormously complex," Modha says. "Any attempt to emulate it is always a radical simplification."

Edelman, too, is looking for ways to simulate the brain using less computing power. He can currently simulate 10 million neurons and half a billion synapses. But the human cortex has at least 3,000 times that many neurons and almost a million times more connections. He says his group has designed and built their own completely new computer architecture in order to be able to add regional microcircuit details into their generic cortex simulated so far. Their simulations to-date have used a Beowulf cluster of 64 interactive processors. "Beowulf is seven feet high and 250 to 300 pounds," Edelman says. The new system—which hasn't yet been described in a published paper—"is about 10 inches by three inches and weighs a few pounds. It can be stuffed inside a brain-based device and is more powerful."

Markram is also starting to feel constrained. "Our BlueGene supercomputer is only just enough to launch this project. It is enough to simulate about 50,000 fully complex neurons close to real-time. Much more power will be needed to go beyond this."

SIMULATED BRAINS IN THE REAL WORLD: THROW IT A BONE

Simulated brains on computers may be interesting research, but like real brains, they are best understood by how they respond to the real world. To test simulated brains in real world settings, some researchers, such as Edelman, use robot-like devices; others use computer avatars; and still others, with a focus on computer vision, struggle to achieve object recognition.

Edelman emphasizes that real world interactions have shaped brain evolution. He has formulated a theory he calls neural Darwinism focused on the role of reward as a driver of brain evolution. "The brain is embodied, and the body and brain are embedded in the real world environment," Edelman says. "And that environment, enormously rich, provides the reward that drives real brains to make choices."

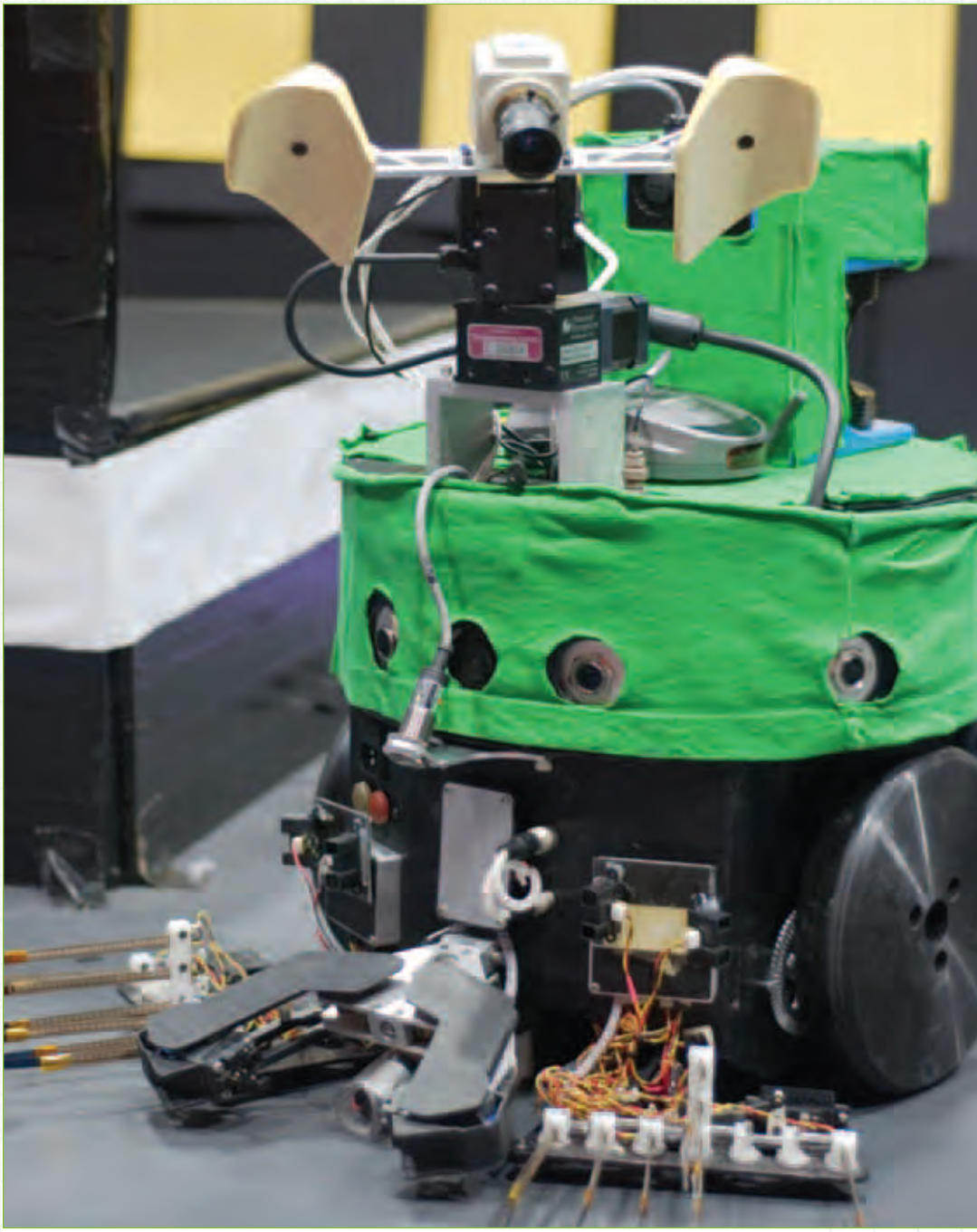
Edelman has tested this theory using "embodied" devices run by a brain-based network. These brain-based devices, called "the Darwin series," are fitted with cameras and microphones that serve as their eyes and ears, and they can sense conductance ("taste") between their grippers.

Darwin VII ran on a brain simulation that included elements of the reward circuitry of the mammalian brain. The knee-high device started out randomly picking up little blocks placed in its roaming zone. One set had stripes, the others, spots. One pattern had high conductance, the other, low conductance. High conductance was arbitrarily rewarded, strengthening the appropriate connections in the brain-based network. This eventually led the Darwin to pick up only this one type of block and ignore the other.

As in the brain, strengthening and weakening of synapses determines if neurons next in line will fire, Edelman says. "Just like synapses act in real brains, the next one won't necessarily fire unless enough stimulation occurs."

Experience changed the synaptic strength. In other words, the Darwin learns.

This reinforced behavior is exactly equivalent to how mammals learn to choose what to eat based on taste.



As Darwin XI learned to navigate mazes, its hippocampus exhibited responses similar to those seen in rats engaged in the same task. Courtesy of Jason Fleischer/The Neurosciences Institute.

After all, taste is a random function of the chemicals in food detected by the olfactory system. The Darwin's sensing of the conductance was equivalent to the mammal's ability to taste food.

"The world is not a coded piece of tape. It can't be explicitly contained in an algorithm," Edelman says. A brain-based device, with a value system, learns by making mistakes. "Hook that to a Turing machine and what you will get is not artificial intelligence, but an entirely new machine," he says—for example, an aerial drone that could decide on its own about threats.

Though one might think the Darwin device hovers on the brink of consciousness, a lot still separates these simulations from actual brain processes. Phil Goodman emphasizes the role of intention and emotion in mammalian brain action. He embodies simulated brain circuitry through projection of a virtual device, an avatar, similar to what video gamers are used to seeing and controlling.

One of his avatars is a dog with pre-programmed behaviors: It starts out lying down, gives a threatening bark while sitting up, or engages in panting and tail-wagging while standing. Sensors allow the simulated brain that is steering the avatar to see and hear. So much of communication of emotions is subliminal that Goodman says, "if (an avatar) is to be social, it needs to interact with our own bodies." So his model incorporates aspects of the emotional processing regions of the brain, the so called limbic system.

A supercomputer runs programs that process sensory input, producing probabilities of neuronal firing, which in turn trigger behavior. A stranger's posture and actions elicit the appropriate reaction of the projected canine. Upright posture with a raised arm will trigger the threat behavior. The happy reception is elicited by crouching with soothing words—and petting on a touch pad.

FEEDING THE WORLD INTO THE BRAIN AND BACK AGAIN

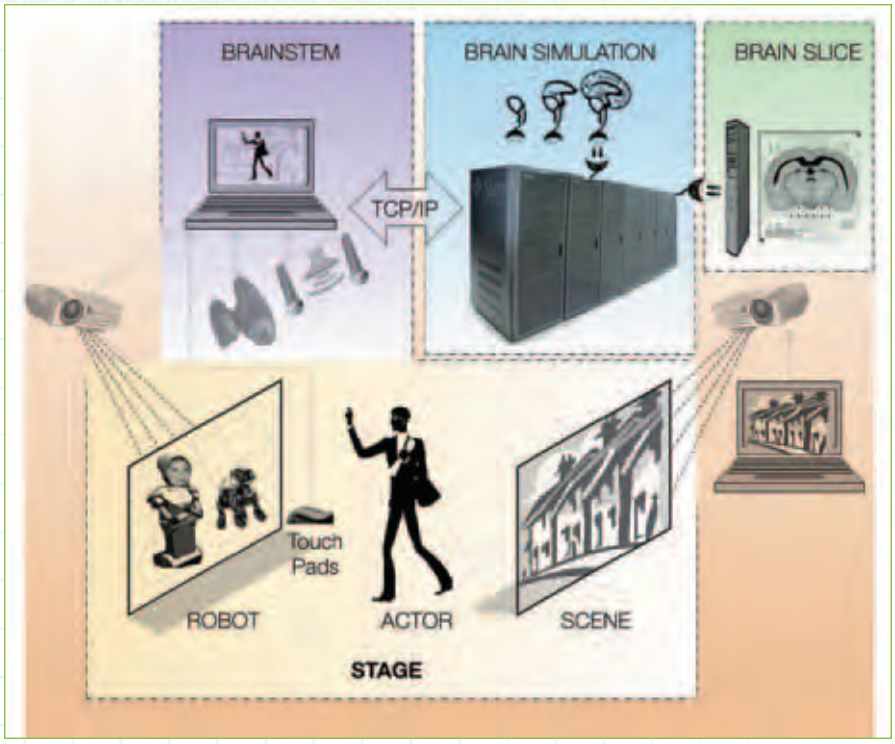
The brain remodels itself in response to perceptions through its sense organs. Thus, simulators need to tackle these brain accessories as well in order to recreate cognition.

Object recognition is vital for a virtual or a material brain-based device such as the Darwin series or Goodman's avatars. Yet it has been one of the most challenging tasks for artificial intelligence. Goodman uses fairly primitive visual processing in his model, but **Thomas Serre, PhD**, a postdoc working with **Tomaso Poggio, PhD**, at the Massachusetts Institute of Technology, has recreated in a machine the ability to perceive objects when flashed at the threshold of human visual perception. Remarkably, the simulation performs as well as people (as described in a News Byte in the Summer 2007 issue of *Biomedical Computation Review* http://biomedicalcomputationreview.org/3/3/4.pdf).

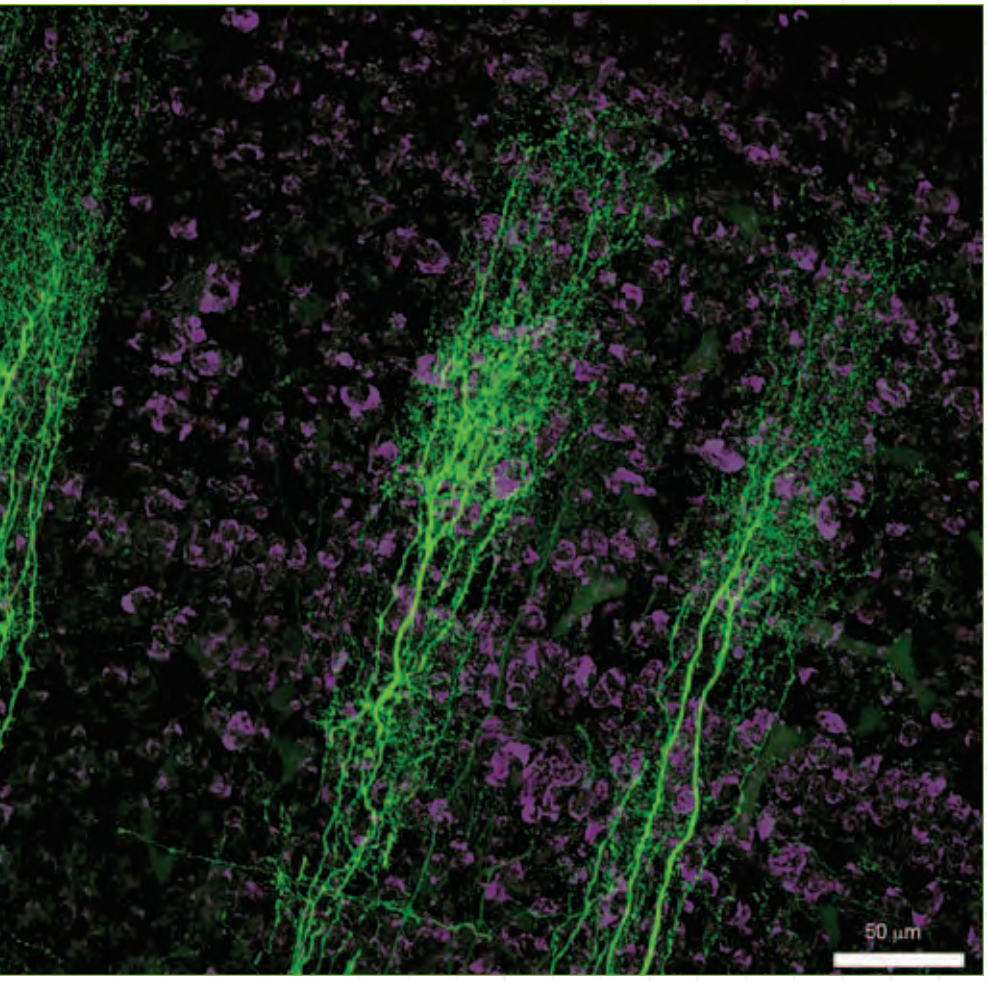
Serre's experiment was limited, however, to the brain's response to an image flashed for less than 150 milliseconds. Thus, it provides just a skeleton of a complete theory of vision, Serre says. He's now working on what happens beyond the first 150 milliseconds of visual processing—"when you move your eyes and shift attention."

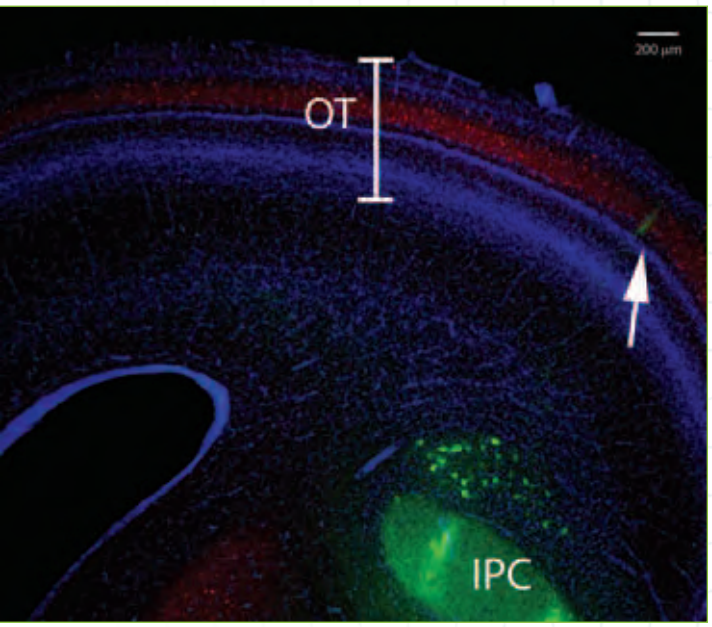
The visual system involves a complex of more than 30 brain areas propagating signals from the retina through the visual cortex to the region of motor cortex that controls how the person (or the simulator) responds. Living brain also contains back projections, echoing all the way back to the primary visual area that receives the initial signals

"The brain is embodied, and the body and brain are embedded in the real world environment," Edelman says. "And that environment, enormously rich, provides the reward that drives real brains to make choices."



Schematic of a virtual neurorobotic system. By creating an avatar of a robot, Goodman and his colleagues avoid the complex engineering of the physical robot. The virtual robot can still respond to environmental stimuli provided through a mouse pad, microphone and camera. Reprinted from Goodman et al., Virtual Neurorobotics (VNR) to Accelerate Development of Plausible Neuromorphic Brain Architectures,* Frontiers in Neurorobotics, *(2008) 2:123:128.





The optic tectum (OT) is a brain structure important for gaze in birds and still present in mammalian brains. Boahen's collaboration is building a model of the OT on a silicon microchip while parallel efforts by Eric Knudsen's team attempt to uncover its biological properties in living brain slices. Below left, we see a cross-section of the bird brain through both the OT and an area that connects with it, the isthmic nucleus (ipc, stained green). The arrow points to a green line marking the location of the close-up image shown at left. The close-up shows nerve fibers in the OT (stained purple) with cell fibers (axons) in green arriving from the ipc. Images courtesy of Dr. Alex Goddard, postdoctoral researcher in the Knudsen laboratory.

from the retina. Vision researchers suspect these back connections may be the way that the visual system can pick out a target object from complex scenes. "By adding back projections to the model, and allowing one shift in attention, to one part of the image, we are (now) able to mimic the next level of performance of a human observer when the image is left just 30 ms longer on the screen, just enough for people to shift their attention once," Serre says.

Boahen at Stanford heads a team working on recreating the basics of different parts of the perceiving brain. Much of the circuitry they plan to model will include back projections. Boahen agrees that feedback likely mediates attention, as competing firing is suppressed. As with other brain simulations, his also shows synchrony, the living rhythms of the brain, including gamma waves with attention.

To find out what the oscillations mean for visual attention, team member **Sridharan Devarajan** and Stanford neurobiologist **Eric Knudsen, PhD**, are working to understand the wiring in a barn owl's tectum, the brain area that controls gaze. Other collaborators are working with Boahen to simulate the wiring of the frontal eye fields in monkeys, an area that allows primates to gauge where attention is needed. This brain area evolved in the social setting of primate life, allowing monkeys to suppress a direct gaze at a superior monkey while still attending to what needs to be watched—covertly. These brain regions feed forward as well as back to higher and more basic levels of visual processing in the brain. Thus, simulations of this area will help researchers to understand the role of feedback circuits in perception.

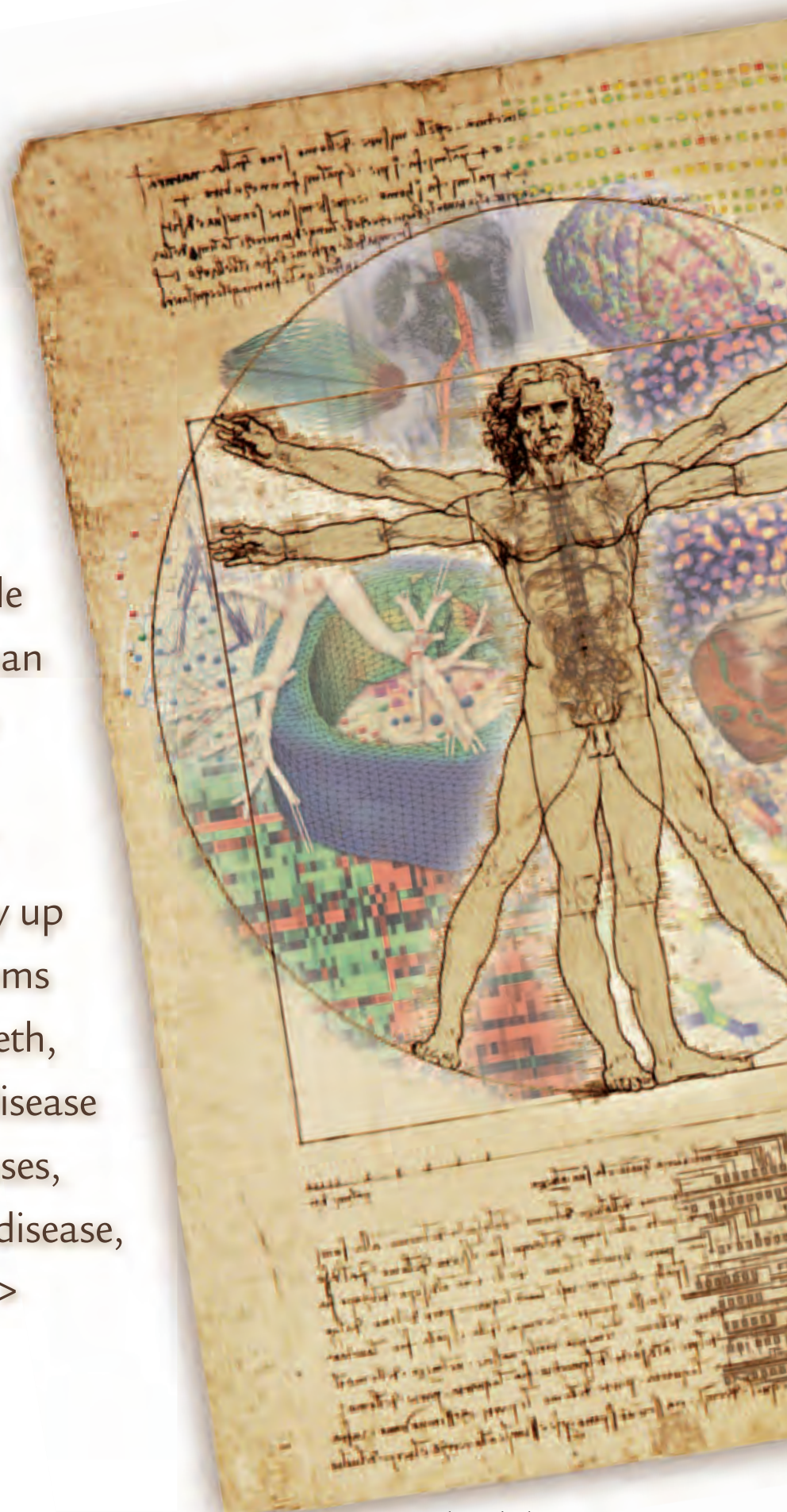
COMPUTER CONSCIOUSNESS

Where are the eavesdroppers and engineers going with all this? Better business machines may be IBM's mantra. Modha's favorite saying is that the mind arises from the wetware of the brain. "The quickest and cheapest way to engineer mind-like intelligence into machines is to reverse engineer the structure, function, and dynamics of the brain" with its low power consumption and compact size, Modha says. "This is our quest."

Some may be scared by this quest. Others eagerly await the emergence of machine intelligence. Eventually, brain scientists hope to simulate the effect of strokes, tumors, or neurological disorders such as Alzheimer's or Parkinson's disease to understand how they derail brain dynamics.

Edelman states frankly his intention: to craft a conscious artifact. "Philosophers have owned the field of consciousness research from time immemorial. What could be more romantic, remarkable or valuable," Edelman says, "than to take on their quest? Right now, you might say, I am going for broke." □

The National Institutes of Health are on a mission: To understand and tackle the problems of human health. To make that daunting problem approachable, 15 of the 20 institutes divvy up human health problems by body part (eye, teeth, heart/lung, etc.) or disease type (infectious diseases, cancer, neurological disease, mental health, etc.). >



BRINGING THE *Fruits of Computation* TO BEAR ON Human Health: *It's a Tough Job But the* NIH *Has to Do It*

By Katharine Miller

These so-called categorical institutes, driven as they are by a desire to understand their chunk of the health puzzle, invest in computational biology research almost incidentally. "Many institute people might say: We want to fund good science and if it happens to require computation then we'll fund computation," says **Karin Remington, PhD**, director of the Center for Bioinformatics and Computational Biology within the National Institute for General Medical Sciences (NIGMS).

But because computation provides tools that can be useful in many categories of biology and medicine, a large portion of the computational research portfolio (how much is actually unclear) is funded by the non-categorical institutes and centers. These include the National Library of Medicine (NLM), the National Human Genome Research Institute (NHGRI), NIGMS, and the National Institute of Biomedical Imaging and Bioengineering (NIBIB), and the National Center for Research Resources (NCRR).

How can the NIH ensure that its investment in computational resources across all the institutes—categorical and non-categorical—really serves the NIH mission? "We have to do a better job of connecting the R01 investigator—the bread and butter investigator of the NIH in general—with these computational approaches," says **Dan Gallahan, PhD**, deputy director for the Division of Cancer Biology and a program officer at the National Cancer Institute (NCI).

Here, we've interviewed a small group of NIH staff people who are immersed in computational research: They're all part of the computational choir, if you will. Their perspective provides, we hope, an interesting peek into the way the NIH as an institution thinks about how to bring the fruits of computation home to the categorical sciences.

It will be a tough job, they say, requiring that computational biologists reach across institute boundaries as well as discipline boundaries. For its part, the NIH must facilitate such inter-disciplinary cooperation and find better ways to coordinate computational research across its institutes. It must avoid duplication of effort without stifling innovation, and lead the development of common approaches, including common vocabularies and common data repositories. And, say some, there should be a far greater investment in computational resources to deal with the flood of high-throughput data.

In the end, when these challenges are met, it will have been worth the effort, says Gallahan. "If you can give researchers a bioinformatics tool that will allow them to replace an animal model or allow them to assay something in multi-dimensional research rather then on just one parameter, then that's going to help everybody."

"We have to do a better job of connecting the R01 investigator—the bread and butter investigator of the NIH in general—with these computational approaches," says Dan Gallahan.

BRIDGE THE CULTURE GAP

Right now, computational and biomedical research travel largely on uncoordinated parallel tracks. On the one hand, many biomedical scientists don't understand the ways that computation could potentially help their research, so they don't know what to ask of computational scientists. On the other hand, computational scientists aren't exactly sitting around waiting for the biomedical researchers to brainstorm good questions. They have their own research aims.

"People might make an algorithmic advance that will eventually have some impact in biomedical research but it's not a coordinated effort," Remington says. The two fields speak different languages, "so it's really tough to translate state of the art developments in computer science and math into things that will be useful in biomedical research."

So the question for the NIH is how to leap over this sociological barrier. Of course there are a few people who do both computation and biomedical research. The National Centers for Biomedical Computing (NCBCs) are rich with people who do both, Remington says. "Immersed in NCBC-land, you get a different perspective. But NCBC-land is a very biased and blessed community. It's a good model to

follow, but it's not the way that most of our community works."

Ideally, says Remington, the NCBCs could serve as a prototype for the kind of environment that the NIH wants to build, where people are talking together in a common language. "What we want to do is make it more of the standard operating procedure, that the experimentalists will be able to communicate what they need to the math and computer science people and really forge relationships and communication structures to advance things in a more coordinated way," Remington says. "We could really accelerate progress in our basic biology research efforts if we could drive the computer algorithm development to fit the needs of these science areas more directly."

"What we want to do is make [the NCBCs] more of the standard operating procedure, that the experimentalists will be able to communicate what they need to the math and computer science people and really forge relationships and communication structures to advance things in a more coordinated way," Karin Remington says.

Another way to bridge the culture gap is to train the next generation of scientists in multiple disciplines. Right now, it's hard to find people with the mathematical skills necessary to support NLM's computational projects, says **Michael Ackerman, PhD**, assistant director for High Performance Computing and Communications at the NLM. "Mathematicians end up in the area of biocomputation sideways—for all the right reasons; but I'm not sure if we have a program that sponsors biomathematics," he says. The NIH could initiate training programs in cross-disciplines like biomathematics to increase the pool of people who can cut across the divide, proposes Ackerman.

Cross-training is not just about exposing mathematicians and computer scientists to biology. It also goes the other way—clinicians and biologists need to understand and be comfortable with the technical side. Toward that end, the NIH has established a program that exposes medical residents and clinical physicians to biomedical engineering research for a year or two. "It keeps it real, to have clinicians interacting closely with the biomedical imaging and bioengineering research groups," says **Zohara Cohen, PhD**, a program director with NIBIB. While not exclusively computational, Cohen points out, "It involves computation in that a lot of our grantees are doing computational work."

There's also another bridge that needs to be crossed, says **Jennie Larkin, PhD**, a program officer with the National Heart, Lung and Blood Institute (NHLBI): the bridge to the open source movement. If tools are freely available, researchers funded by categorical institutes will be more likely to make use of them. Thus, says Larkin, "the NIH needs better clarity about how to support people who are trying to develop and support freely available open-access computing tools and resources." Since this sort of open-source

Karin Remington would love to ask biomedical types: "If computer cycles and algorithms weren't an issue and you could do anything you wanted to do, what would you want to know?"

model is accepted in other areas of computer science and self-sustaining (albeit through different mechanisms), Larkin says it's time for the NIH to think through ways to achieve the same.

LEARN FROM SUCCESSES

To date, successful efforts to bring computation to the categorical sciences have flowed mostly from the many large centers that have received and continue to receive NIH support, say the NIH staffers interviewed.

"These Center programs are a good approach," Gallahan says. "I would extend that to make it more diverse and have more of them."

Michael Marron, PhD, director of the Division of Biomedical Technology at the NCRR, agrees: There's a need for much greater investment in infrastructure and enabling technologies, and large centers are a great way to achieve that.

Marron points to the Biomedical Technology Research Resources (BTRRs) as a long-standing example. For fifty years, the NCRR has invested in these centers, a subset of which is devoted exclusively to Informatics Resources. The "Resources," as they're commonly dubbed, develop and disseminate a wide range of software for the biomedical community, including—to mention just a few—software for molecular dynamics (such as the widely-used AMBER and CHARMM), visualization (such as the popular VMD), and genetic epidemiology (SAGE). And new centers are still being created, including one established at UCSD in 2008 to develop computational ways to analyze mass spectrometry data.

> "These Center programs are probably a good approach," Gallahan says. "I would extend that to make it more diverse and have more of them."

Significantly, Marron says, the BTRRs are evaluated and renewed based on whether their products are disseminated and adopted for biomedical research. "The important thing is to get the science done," Marron says. Many of the Resources offer help-desks as well as software training—not only for students, but for senior biomedical researchers as well. And their principal investigators attend a broad range of conferences—including categorical science conferences—to spread the word about their tools.

When the NIH Roadmap came out, many at NCRR saw that as an affirmation of what they'd been doing for a long time. "The National Centers for Biomedical Computing (NCBCs) are very similar to the [BTRRs] that we've funded for years and that's of course by design—because it's a model that works," Marron says. "The NCBC grants complement the BTRRs. And together, they represent the most coordinated NIH activity for support of computation in medicine."

The NCBCs promote collaborations between computational researchers and biologists by focusing the computational research around specific driving biological problems (DBPs). The NCBC program is a national network that's envisioned as having "both hubs and spokes," says Cohen. People from outside the NCBC communities can draw on the resources created at these hub centers, expanding the community.

To further that aim, the NIH established a program to fund collaborating R01s—individual investigator-driven research projects that would collaborate with the NCBCs to develop tools with a specific biomedical focus. As a result, the NCBCs have succeeded in connecting up with a variety of collaborating R01s, many of which are funded by categorical institutes. "This is a great model for the future," Cohen says.

At the NCI, a similar program—the Collaborative Research in Integrative Cancer Biology and the Tumor Microenvironments Program—was modeled on the NCBC collaborative R01s. It's just starting now, and mandates that individual computational biology investigators within the Integrative Cancer Biology Program and the Tumor Microenvironment Network program collaborate with people who are not in those groups—ensuring that the program expands to new people and communities. "That again is an example of an active program trying to bring the rest of the community in," Gallahan says. "That's exactly what it's designed to do."

The NIH is also learning to structure program announcements to bring together computational and biomedical researchers. For example, a program for Collaborative Research in Computational Neuroscience (CRCNS, for which the National Science Foundation serves as lead, with NIH participation) mandates participation by key person-

NIH Challenge Grants: Stimulating Biomedical Computation

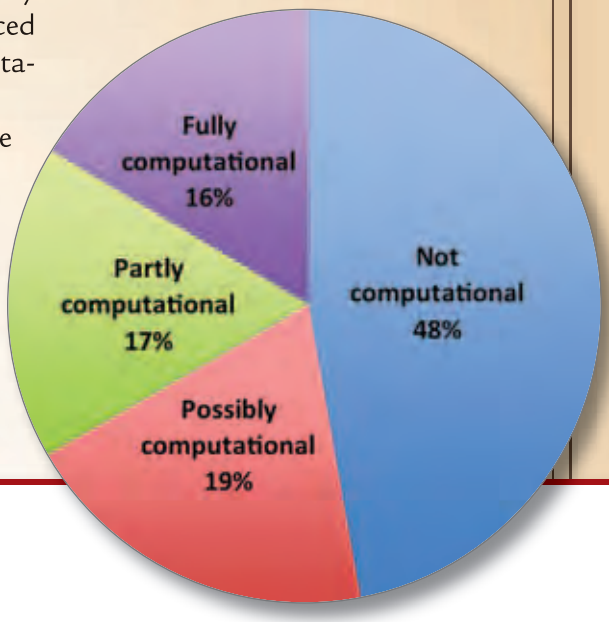
The federal government's recently passed stimulus package—the American Recovery and Reinvestment Act (ARRA)—provides $10.4 billion to the National Institutes of Health, all dollars that must be spent in 2009-2010. The goal: to stimulate the U.S. economy through support of scientific research. And, if recently announced Challenge Grants are any indication, ARRA will also stimulate computational research in biomedicine.

On March 5, the NIH announced that at least $200 million of the ARRA funds will go to a new program called the NIH Challenge Grants in Health and Science Research. According to the announcement, the idea is to give a two-year "jumpstart" to specific scientific and health research challenges in biomedical and behavioral research.

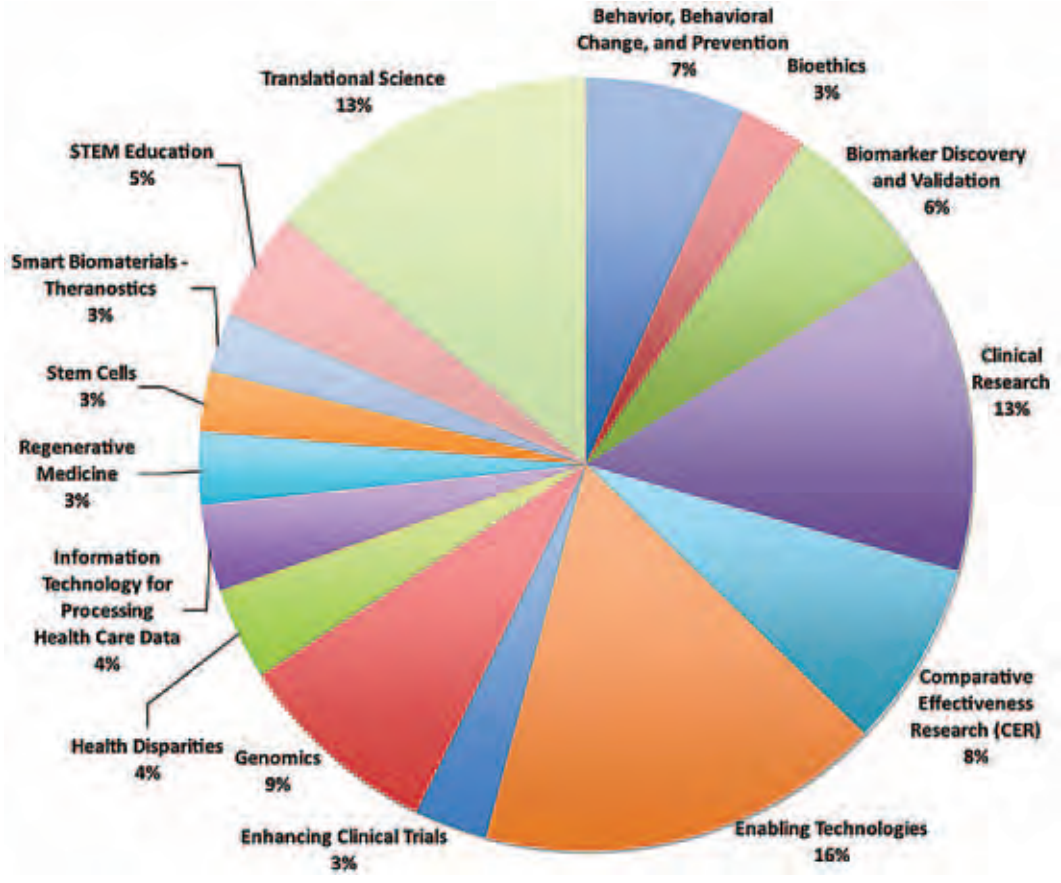
The announcement identified 15 "challenge areas" that encompassed 878 "challenge topics," 207 of which are deemed "high priority." One of the 15 broad challenge areas is entirely computational. Called "Informational Technology for Processing Health Care Data," it covers four percent of the topics. But many of the other topics—under other broad challenge areas—are highly computational as well, as shown by the charts on this page. "It's a step in the right direction," says NCRR's Michael Marron.

In the two charts shown here, **Biomedical Computation Review (BCR)** ***conducted our own review of the Challenge Grant RFA. The chart at right shows all 878 topics categorized by broad challenge area. The chart above shows the 207 "high priority" topics categorized by the extent to which they are fully computational, partially computational (i.e., the solution will involve both computational and non-computational approaches), or potentially computational (i.e., the topic addresses a problem that could be addressed by computational approaches if someone with that expertise applies). Courtesy of*** ***David Paik, PhD,*** ***assistant professor of Radiology at Stanford University and executive editor for*** **BCR.**

Computation and the NIH Challenge Grant High Priority Topics



NIH Challenge Grant Topics by Broad Challenge Area



nel from both computation and neuroscience. Likewise the Physiome program announcement required leadership representation from both the modeling and biomedicine communities. The Bioengineering Research Partnership (BRP) program establishes interdisciplinary partnerships between people from both biomedicine and engineering, with the aim that they create a deliverable for the biomedical research community within a ten-year time frame. And each of the 34 Clinical Translational Science Awards, which are geared toward remaking the clinical research enterprise, includes a bioinformatics focus.

The Biomedical Information Resource Network (BIRN) and the Cancer Biology Grid (caBIG) serve up a different model of connecting computation and biomedicine—by providing biologists and physicians with platforms that allow them to share data and tools.

The BIRN involved categorical sciences from the get-go, says **Michael Huerta**, **PhD**, director of the National Database for Autism Research and associate director at the National Institute of Mental Health (NIMH). NCRR launched the BIRN in 2001 to develop a national infrastructure for biomedical research using neuroscience as a test-bed. When NCRR was just starting to put BIRN together, Huerta says, the institute's leadership engaged people in the categorical institutes to find out what the biomedical research community needed. "BIRN has grown but always with the categorical institutes kept posted, invited to meetings, invited to the review of grant applications and so forth," Huerta says. "That proactive effort has transformed BIRN from a good idea to an infrastructure that is increasingly important to neuroscience."

BIRN confederates data and tools so that users can access them from across the network regardless of where they are stored or housed. "And it does so in an invisible way," Huerta says. "You don't necessarily know where you're getting things from." Nowadays, with the BIRN platform reaching production mode, NCRR is very interested in expanding BIRN to other domains, Huerta adds. "I'm sure they'd be delighted if folks doing diabetes or heart/lung research would start to increasingly use BIRN."

The NCI's caBIG platform is similar, with a set of standardized rules and a common vocabulary for applications, tools, and data shared through its infrastructure. Both BIRN and caBIG were launched within specific research communities (neuroscience and cancer, respectively) but have potentially broader applicability—and may eventually link up to one another. "As time has gone on, these two platforms are getting closer to each other," Huerta says, "so that before too long, I think in the next generation, you'll be able to work across them."

INCREASE FUNDING FOR COMPUTATION

Despite all the great work being done by the various centers, Marron says, there simply isn't enough money devoted to this area. "We could easily fund two to three times what we do and still not be exhausting high quality areas," he says.

The NIH, Marron says, is really behind in spending on computation and informatics compared to the National Science Foundation, the Department of Energy, and many pharmaceutical companies. According to Marron's best guess, less than two to four percent of the NIH budget goes to computation and bioinformatics grants. "It's peanuts," he says. He thinks the investment should be closer to 25 percent. "There are many challenging computational areas where we could see rapid advances if we could capture the computational tools to do it."

Specifically, Marron favors much more funding for the development of enabling activities like software and infrastructure. He'd like to see more efforts like BIRN, that create processes, protocols, sharing agreements and middleware, "all the stuff that makes formation of a virtual organization barrier-free."

And the NCBCs, he says, are a good start. "We would like to fund more and I think there should be more."

In addition, Marron believes the NIH should invest more to ensure the effective use of massive amounts of data, Marron says. "The vast amount of data collected today will never be viewed by humans, so you have to have tools to do this," he says. And that raises huge questions of quality control, "How do you know what you're even looking at?" he asks. "You need to have that built into your software tools so it's not just garbage-in/garbage-out.

Marron would like to see the NIH invest in new ways to build databases; discover data; visualize data; and analyze data. "We haven't begun to firm up tools for that. We're almost still at the level of spreadsheets." Just finding usable data is nearly impossible. "We should be supporting the development of machine-readable registries of data, so your machine can find it," he says. And then there's the problem of combining data of various sorts (gene expression, proteomics and tissue mapping, for example) to come up with meaningful analyses. "It's clear that getting a handle on the etiology of disease is a multi-dimensional problem," he says.

And more computer scientists need to be engaged with the NIH, Marron says. "If we support 100 to 200 computer scientist awards at the NIH, I'd be surprised. We should be supporting thousands, from natural language, to database and networking experts," he says, all with a devotion to improving biomedical research, of course. Plus, there's a need for computation

associated with new experimental tools, such as analyzing fluid dynamics for advanced ultracentrifugation techniques; or applying radar-imaging techniques to high frequency MRI.

Marron also points to a lack of computer expertise in the biomedical research community at large compared to, say, the physics or astronomy communities. Clinicians and clinical institutions are particularly skittish about things like choosing a computer system or accidentally leaking out private data, which can cause them a huge amount of grief, he says. "All of this is more of an argument for why one needs to invest in Centers like the NCBCs—to provide centers of expertise so that everyone doesn't need to develop them on their own."

AVOID DUPLICATION WHILE LETTING A THOUSAND FLOWERS BLOOM

In general, the NIH avoids funding redundant research, say the interviewed program officers. Even for computation, which is widely dispersed, the grant review process weeds out proposals that duplicate existing grants. And communication and coordination efforts ensure that where funded research has enough in common, NIH program officers will bring researchers together to make alliances or to work together.

On the other hand, it can be hard to avoid duplication for computational pieces that are not the main focus of a grant, says Remington. "I hesitate to even think how much money NIH invests in software engineers on R01 grants to reproduce the same sorts of basic database dissemination Web site tools over and over again when really we could have one central repository and do that sort of thing easily as a service for the research community." These infrastructure problems may eventually be a thing of the past as more people migrate to common platforms like BIRN or caBIG, Huerta says. But other kinds of duplicated effort remain that are tougher to tackle, such as redundancy in algorithms.

A few years ago, when the NLM asked people to rewrite their algorithms in a standard format to be archived and maintained by NLM, they got a surprising number of different algorithms that did the same thing, Ackerman says. Redundancy arises because people think they can do better than what already exists. "So you're stuck with the redundancy to find out whether it can be done better," he says. And it's nearly impossible to discover why a person chooses one algorithm over another. "Is it better for one type of data than another? That's a nut we've never cracked."

The same problem exists at the NCI, Gallahan says. For example, a number of groups have independently developed microarray analysis programs. "The NCI Center for Bioinformatics and its director, Ken Buetow, really have had to come to grips with what to do with all of these programs that we're supporting," he says. It's a daunting task partly because scientists by their nature want to explore things in their own ways and are wedded to their application and their own research area. "So that can spawn redundancy with an 'I can do that better' sort of attitude," he says. "You have a lot of people pursuing different avenues, all with the best intentions."

Larkin discovered the same phenomenon when she asked her NHLBI systems biology grantees whether they'd be interested in some way to facilitate sharing of code, software or models. Could they leapfrog off others' work in order to go farther and faster? The response was mixed. In addition to intellectual property concerns, the researchers had another problem: Sometimes solutions are over-specified so that in fact it's meaningless to share code. "Just because it works perfectly for one researcher doesn't mean it will be helpful to anyone else," she says. "There also may be more than one

"We were surprised at the number of different algorithms that do the same thing," Ackerman says. "And then the question, which we've never gotten an answer to—why did you you choose Max's algorithm rather than Joe's algorithm? Is it better for one type of data than another? That's a nut we've never cracked."

good solution to a problem. Sometimes it's a real loss to reinvent the wheel because it's a wasted effort. But other times you end up with a better wheel—a radial instead of a Conestoga wagon wheel, for example."

One of the best ways to reduce duplication, Larkin says, is for the NIH to develop efficient ways to share software and other computational tools—which remain hard to find even when they are posted to the web. The NIH already supports a variety of such efforts.

"Sometimes it's a real loss to reinvent the wheel because it's a wasted effort," Larkin says, "But other times you end up with a better wheel—a radial instead of a Conestoga wagon wheel, for example."

The Biositemaps project launched by the NCBCs has been discussed in this magazine before (see Winter 2008/09 Issue of *BCR*). "It might be a real lightweight, decentralized solution," Larkin says. Many of the categorical institutes also have their own solutions. caBIG connects resources for NCI researchers. In the neuroscience arena there's the Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC), designed to facilitate the dissemination and adoption of neuroimaging informatics tools and resources; and the Neuroinformatics Framework (NIF) provides a concept-based query language for locating all types of neuroscience resources—including computational ones. And heart researchers can turn to the Cardiovascular Research Grid, while PhysioNet houses analysis tools for looking at medical time series data. There are also repositories for various categories of tools—for example, Simtk.org for physics-based models and simulations; or the ITK/VTK repository for visualization tools.

"So there are many sorts of solutions due to different cultures in different areas," Larkin says. And like other duplicative efforts, we don't yet know which ones are the best ones. "I'd be loath to restrict the solution set now," Larkin says, "because we might guess wrong."

STRATEGIZE TRANS-NIH

Bioinformatics and biocomputation cut across all of the institutes' research programs. This makes coordination among the institutes somewhat challenging. Over the years, a variety of informal coordinating groups have managed different trans-NIH programs involving computation. The Biomedical Information Science and Technology Initiative Consortium (BISTI) is perhaps the best-known and longest-lived. Launched in 2000, it brought together program officers from across all the institutes. And it developed the NCBC program.

"The NCBCs are our best example of how to do things together, but it's a teeny-tiny example," Remington says. "It needs to be taken up a notch…to achieve synergy in an area like computation that cuts across so many fields."

What's missing, as Remington sees it, is a data-driven, comprehensive strategy for coordinating computation across the NIH. To Remington, the problem has two components: The lack of information about the trans-NIH investment in computation; and the lack of a coordinating group vested with power to act strategically.

"We have precious little understanding of what our real investment is across the institutes," Remington says. In February 2009, the NIH launched a database (called Research Condition and Disease Categorization) that, for the first time, allows trans-NIH portfolio analysis. This will help the NIH deliver mandatory reports to Congress about its investment in specific disease areas. "It stands to be a really big improvement in our relationship with the public," Remington comments. But it's not likely to help identify the trans-NIH investment in computation because, Remington says, it's built largely on a biomedical vocabulary. "So to do an analysis on computational networks, the word networks will come into play but it won't turn up as computer networks. It will give networks of genes or hospital networks," she says. "The system is ill-tuned towards anything related to computation."

Another portfolio-analysis tool called the Electronic Scientific Portfolio Assistant, or eSPA, which has been evolving within the National Institute of Allergy and Infectious Diseases, allows users to dynamically probe and poke things to see how money is being spent. In May of 2008, it was opened up to a pilot program involving 17 Institutes and Centers. According to Huerta, "NIH really is investing in giving us the informatics capabilities that we need to know what's happening." And though he hasn't tried eSPA yet, Huerta has been told it's quite powerful. "In the future, this will empower program officers to know what's going on beyond their own portfolio and beyond what they happen to hear about."

Huerta and Remington both hope eSPA will prove useful for a trans-NIH computational portfolio analysis. "That's the piece that has been missing from this BISTI consortium," Remington says. "As functional as it has been over the years, it has really been unable to look across institutes in a real data-driven way, to analyze across the NIH where our investments are going."

NIH also lacks a coordination effort vested with actual authority, says Remington. "BISTI is really more ad hoc than I think is called for given the need." BISTI relies on voluntary participation by program officers at multiple institutes, and some institutes participate more than others. "It doesn't have the same sort of strategic-planning capability as would be best-suited, I think, for moving us forward in this area," Remington says.

> "The NCBCs are our best example of how to do things together, but it's a teeny-tiny example," Remington says. "It needs to be taken up a notch… to achieve synergy in an area like computation that cuts across so many fields."

Gallahan agrees. While trans-NIH programs like BISTI help communicate what's going on among the various institutes, "they don't have the same sort of gravitas of resources and public awareness as things that come from the Office of the Director, like the Roadmap, or even some specific programs at the NCI," he says. Admittedly, he says, NCI is less dependent on BISTI, partly because its internal resources and overall scope allow it to frequently act independently. "There might be some benefit to more of a top-down review of computation with some power behind it. But where do you define the point of asking? Sometimes it's at the Health and Human Services Department level, the NIH level, or we might think it's at the NCI level."

To Remington, the next logical progression from BISTI is to have a consortium, perhaps based in the Office of the Director, that strategizes carefully about where NIH investments are going and tries to leverage things that are clearly trans-NIH. "A group that leverages no-brainers for us to do together instead of funding over and over again the same thing, institute by institute," she says. Of course each institute will have its own strategic plans and its own things they need to do. "But in a cross-cutting area like informatics and computation," she says, "we could really leverage that effort better if we came together to develop a strategic plan that's coordinated, that's not institute by institute."

The Roadmap was eye-opening for many at the NIH, Remington says. "The Institutes start-

> "As functional as [BISTI] has been over the years, it has really been unable to look across institutes in a real data-driven way, to analyze across NIH where our investments are going," says Remington.

ed to realize how much potential savings there could be in sharing intellectual capital and resources…. And bioinformatics and biocomputation are really a sweet spot of that potential."

DEVELOP COMMON APPROACHES

To Huerta, one of the things that's hindering the success of biocomputation is the lack of common approaches—common data formats, common vocabularies, common ontologies and common long-term data repositories. The NIH needs to take the lead on this because individual communities won't do it on their own, Huerta says. "They're interested in the research. They're interested in what genes are involved in autism or what peptides are involved in myocardial infarction. They are not driven by 'what should we call the peptide,' or 'what data format should we use?'"

"The Institutes started to realize how much potential savings there could be in sharing intellectual capital and resources…. And bioinformatics and biocomputation are really a sweet spot of that potential," Remington says.

It's particularly problematic for communities that are organized around a particular data type that might cross institute boundaries—for example, signaling data, which is relevant to NIMH, NHLBI, NCI and others. "How does NIH encourage the development and use of common approaches by such research communities?" he wonders. "They are not going to organize around these things, and there really isn't a way to do this right now. So I see this as a major need that NIH has. I call it community-based solutions for community-wide needs because the solutions are going to come from the community to serve the needs of the community."

Getting communities to rally around common approaches requires a different mechanism than a typical research grant. "Really what folks need is organizational and operational support. We could serve as a way to organize around these issues where they wouldn't be self-organized," Huerta says. "That's kind of on the horizon of what NIH needs to start paying attention to. And in fact we're doing some of that. We haven't gotten there yet."

The NIH also has to address the fact that common approaches are dynamic and require ongoing support to be updated, Huerta says. Cohen agrees, emphasizing the particular need for long-term data management. "NIH has to answer some questions about how to support large datasets and make them available after a grant ends," Cohen says. Perhaps with the development of common approaches, this will become an easier problem to handle.

Taking it to the next level, says Gallahan, common approaches in the computational area will also help researchers explore commonality among diseases, which will in turn help guide ways to interfere with disease. Gallahan points to a paper published last year by **Albert-Laszlo Barabasi**, [covered in the Fall 2007 News Bytes section of this magazine] that was able to find this sort of interconnection among diseases. Thus, Gallahan says, "Modeling might be able to do scientifically what we're unable to do administratively."

COMPUTATION IS THE FUTURE

Ask Gallahan why computation matters to the NCI, and he'll tell you that it's the future. "Much as molecular biology opened the world at that scale to manipulation, I think computational biology is going to bridge many of the challenges we have in dealing with biological complexity."

The effort to cure cancer is particularly on point. Over the last 15 years, Gallahan says, they haven't seen as many advances as the institute would like. "And I think that's partly because it is such a complex disease," he says. The greatest advances have tended to be very targeted therapies that affect a limited (albeit important) population. And after treatment with these therapies, sometimes the tumors reappear, having gained resistance to the drug. The lesson: The problem of cancer requires a better understanding of the disease's complexity. "And in order to understand and integrate that, we're going to need these computational approaches." □

Under TheHood

BY BIN DONG

The Implicit Representation of Biological Shapes and Forms

Imaging, geometric modeling, representation and computing of shapes and forms are important components of modern computational biology. These processes apply across wide spectra of scales, genotypes and phenotypes, from microarray imaging for genomics, to neuroimaging of human brains. One of the most fundamental image processing problems is the representation of shapes and forms. There are two popular ways of representing biological shapes: parameterization-based (explicit) representation and implicit representation. Parameterization-based representation codes important shape information into geometric variables (such as the vertices of a triangle in a triangular mesh, and how the vertices relate to one another—i.e., whether they are connected by an edge) and topological variables (such as whether there are holes in the shape or not, e.g. the difference between a ball and a donut). In contrast, implicit representations are frequently described via level set functions and their siblings. A level set function is usually defined to take negative values inside the shape and positive values outside, and hence its zero level set (i.e., the set of points on which the function takes zero values) describes the shape. Both types of representations have their own advantages and disadvantages.

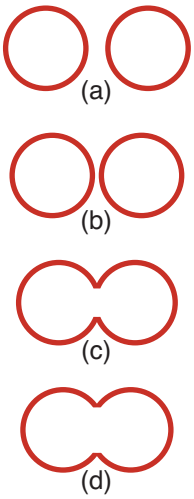


Figure 1. Topological changes induced by merging two bubbles.

The major advantage of using triangular meshes to represent biological shapes is the efficiency of data storage and algorithmic development. We can represent and process a high-resolution high-accuracy shape without using excessive physical memory. However, one drawback of using a triangulated surface is its inflexibility in terms of topological changes (e.g., merging of two bubbles). This is rather critical for some cases. Topological changes affect many shape-processing procedures, e.g., shape restoration and segmentation. Whenever topological alterations occur, the original triangular mesh becomes degenerate and demands retriangulation or surface correction. Take shape restoration, for example: Topological changes may happen so often that it demands constant shape retriangulation, which makes processing algorithms computationally inefficient.

As for implicit representations of shapes, taking level set functions as an example, the major advantage is their flexibility in terms of topological changes. Whenever shape-processing introduces topological changes, implicit representations are more flexible and convenient than parameterization-based representations. Let us look at the simple example in Fig. 1, where we are animating the merging of two bubbles in 2-D. From (a) to (d) in Fig. 1, the two bubbles are growing at a constant speed with their centers fixed. Topological change happens at (c), where the two bubbles touch and then merge into one bubble. It is very easy to implement this motion when the bubbles are represented by a level set function. All we need to do is solve a certain differential equation. If the bubbles are parameterized, however, we would need to constantly check if merging is about to happen, and when it does, reparameterize the shape. This makes the computational implementation rather complicated.

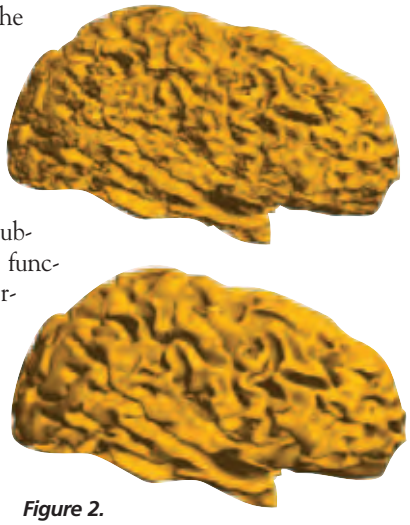


Figure 2. top: original, noisy cortex data; below: NLM processed cortex.

In addition to topological flexibility, implicit representations are more natural in representing biological shapes for practical purposes, because the shapes extracted from modern imaging data, e.g., MRI, CT and ultrasound images, are intrinsically implicitly represented in the first place. Also, since level set functions are usually defined on standard Euclidean grids, most level-set based algorithms are very easy to implement. However, in contrast to the parameterization-based representations, implicit representations are usually not very efficient in storing the data. Indeed, surface data, which is essentially 2-D, is implicitly saved as a 3-D function. The problem becomes more severe when we are dealing with high-resolution shapes. In general, efficient storage and manipulation of high-resolution implicit shapes is challenging when we need to keep all the existing features of the representation.

One example where the level set representation may be more appropriate is cortical surface restoration, where topological changes are unavoidable. In Fig. 2, we show how nonlocal means (NLM) [1], where the cortex is represented by a level set function, can automatically remove the many small isolated bubbles that arise from segmentation errors.

REFERENCES:

[1]. Bin Dong, Jian Ye, Stanley Osher, Ivo Dinov. Level-set-based nonlocal surface restoration. In *Multiscale Modeling and Simulation*, 7(2), 589-598, 2008. □

DETAILS

Bin Dong is a graduate student in the department of Mathematics at the University of California, Los Angeles (UCLA). He is an investigator in the Center for Computational Biology and the Laboratory of Neuro Imaging at UCLA. His work focuses on the application of variational methods and partial differential equations in medical image processing.

DIVERSE DISCIPLINES, ONE COMMUNITY
BiomedicalComputation REVIEW
Published by Simbios, an NIH National Center for Biomedical Computing

Nonprofit Org.
U.S. Postage Paid
Permit No. 28
Palo Alto, CA

Biomedical Computation Review
Simbios An NIH National Center for Biomedical Computing
Stanford University
318 Campus Drive
Clark Center Room S231
Stanford, CA 94305-5444

seeing science

SeeingScience

BY KATHARINE MILLER

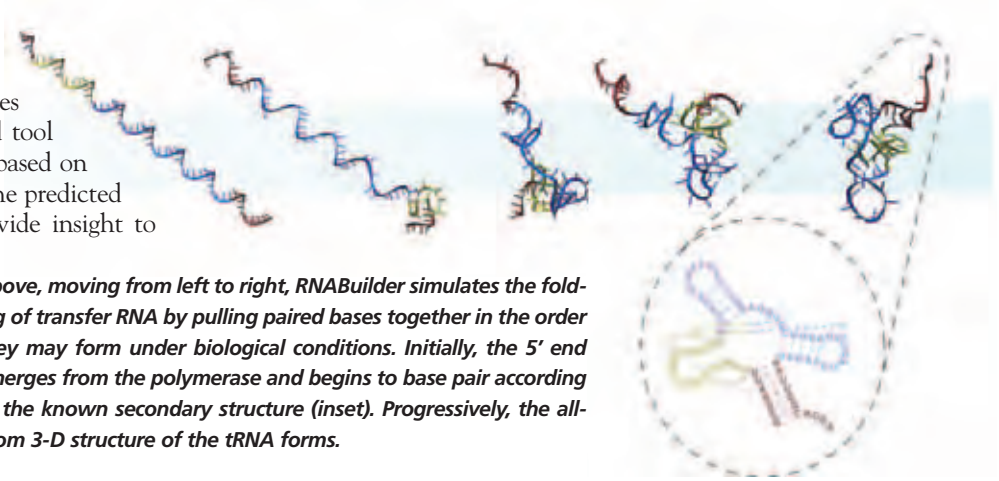
Building RNA 3-D Structure

The structure of RNA is an important key to its function—including its role in disease. However, the structure of most RNAs is unknown because their extreme flexibility and high charge make them difficult to crystallize. In addition, prediction of RNA structure based only on its nucleotide sequence remains elusive for all but the smallest molecules.

Aiming to bridge the gap between successful sequence-based structure prediction codes that predict the structure of small RNAs and unsuccessful attempts to predict larger ones, **Samuel Flores, PhD**, a postdoc in bioengineering at Stanford University, and his colleagues have developed a rigid body dynamics software program called RNABuilder. Because the software relies only on readily available information such as base-pairing contacts (which are often known even when the full 3-D structure of the molecule is unknown), it provides experimentalists with a long-awaited tool to quickly model possible structures based on limited experimental information. The predicted structures and folding pathways provide insight to guide further experiments.

RNABuilder mimics how RNA is made in nature—enforcing base-pairing starting at the 5' end and finishing with the 3' end. It uses Simbody, SimTK's Multibody Dynamics code library. Simbody's Contact subsystem is used to economically account for steric and Coulomb repulsion. Selected bonds are rigidified to reduce the number of bodies for greater economy. For tRNA and the P4/P6 domain of the *Tetrahymena* ribozyme, the program has been shown to recover the correct topology, base-pairing contacts, and overall structure using only the base-pairing information that was available before the three-dimensional structure was known. □

DETAILS: RNABuilder is an RNA modeling program based on Simbios' Simbody code for multi-body mechanics, which is freely available as part of the SimTK toolkit (http://simtk.org/home/simtkcore). A workshop on using RNABuilder and NAST (see Simbios News column in this issue) will be held at Stanford University on June 19, 2009. For more information, contact Blanca Pineda, bpineda@stanford.edu.



Above, moving from left to right, RNABuilder simulates the folding of transfer RNA by pulling paired bases together in the order they may form under biological conditions. Initially, the 5' end emerges from the polymerase and begins to base pair according to the known secondary structure (inset). Progressively, the all-atom 3-D structure of the tRNA forms.