

D I V E R S E D I S C I P L I N E S , O N E C O M M U N I T Y

BiomedicalComputation

Published by Simbios, an NIH National Center for Biomedical Computing

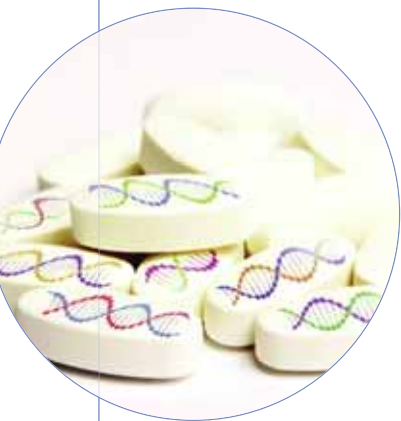
REVIEW

Evolution and HIV:

Using computational phylogenetics to close in on a killer

Summer 2009

PLUS:
From SNPs to
Prescriptions:
**Can Genes
Predict Drug
Response?**



FEATURES

10 From SNPs to Prescriptions: Can Genes Predict Drug Response?

BY CHANDRA SHEKHAR, PhD

20 Evolution and HIV: Using Computational Phylogenetics to Close In On a Killer

BY KRISTIN SAINANI, PhD

DEPARTMENTS

1 GUEST EDITORIAL | A VISION OF COMPUTATIONAL ANATOMY
BY ARTHUR W. TOGA, PhD

**2 POINT/COUNTERPOINT | SHOULD GRANT APPLICATIONS FOR THE
DEVELOPMENT AND MAINTENANCE OF SOFTWARE AND INFRASTRUCTURE:
COMPETE WITH BASIC RESEARCH APPLICATIONS OR
HAVE A SEPARATE FUNDING MECHANISM?**
SHANKAR SUBRAMANIAM, PhD, VS. JOEL R. STILES, MD, PhD

**4 NEWSBYTES | BY BETH SKWARECKI, RACHEL TOMPA, PhD,
KATHARINE MILLER, LOUISA DALTON, AND LIZ SAVAGE**

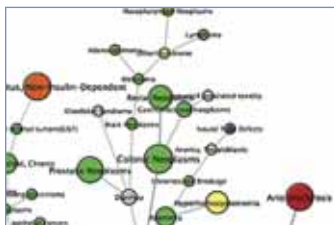
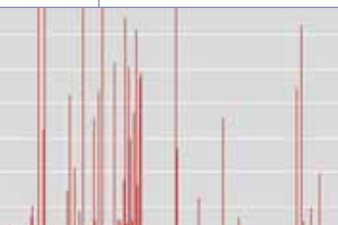
- Automating Scientific Discovery
- The Function of DNA Form
- Semantic Journals and Scientific Publishing
- Open Source Tools For Parsing Clinical Data
- Online Searches Warn of Flu Spikes
- Flowing Through the Interactome

**32 SIMBIOS NEWS | SIMPLIFYING THE SCIENCE AND
ART OF MOLECULAR DYNAMICS** BY JOY P. KU, PhD

33 UNDER THE HOOD | CANONICITY AND DISEASE ONTOLOGIES
BY ALBERT GOLDFAIN, PhD

34 SEEING SCIENCE | SWINE DYNAMICS BY KATHARINE MILLER

Cover Art: Created by Rachel Jones of Wink Design Studio using an evolutionary tree, courtesy of Andrew Rambaut. HIV virus is © Eraxion | Dreamstime.com. **Page 20:** Artwork created by Rachel Jones of Wink Design Studio. Pill bottle is © Webking | Dreamstime.com. **Page 31:** Artwork created by Rachel Jones of Wink Design Studio. Pills photo is © Creative_studios | Dreamstime.com, DNA graphics are © Kusuriuri | Dreamstime.com.



Summer 2009

Volume 5, Issue 3

ISSN 1557-3192

Executive Editor David Paik, PhD

Managing Editor Katharine Miller

Associate Editor Joy Ku, PhD

Science Writers

Katharine Miller, Chandra Shekhar, PhD,
Kristin Sainani, PhD, Rachel Tompa, PhD,
Beth Skwarecki, Liz Savage

Community Contributors

Arthur Toga, PhD, Joel Stiles, MD, PhD,
Shanka Subramaniam, PhD, Joy Ku, PhD,
Albert Goldfain, PhD

Layout and Design

Wink Design Studio

Printing

Advanced Printing

Editorial Advisory Board

Russ Altman, MD, PhD, Brian Athey, PhD,
Dr. Andrea Califano, Valerie Daggett, PhD,
Scott Delp, PhD, Eric Jakobsson, PhD,
Ron Kikinis, MD, Isaac Kohane, MD, PhD,
Mark Musen, MD, PhD, Tamar Schlick, PhD,
Jeanette Schmidt, PhD, Michael Sherman
Arthur Toga, PhD, Shoshana Wodak, PhD,
John C. Wooley, PhD

**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:



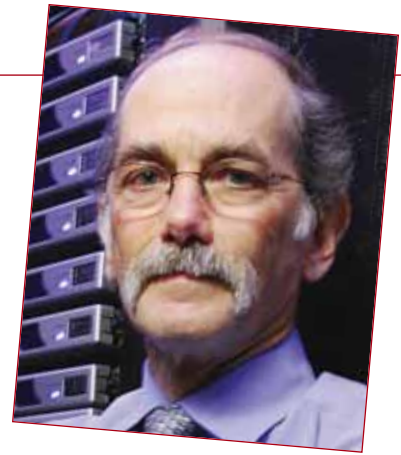
The NIH National
Center for Physics-
Based Simulation of
Biological Structures

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)

BY ARTHUR W. TOGA, PhD

A Vision of Computational Anatomy



Today, the knowledge, experience and memory of clinicians or scientists function as the exclusive resource for distinguishing normal from abnormal brain images; identifying signatures or biomarkers of disease in vast collections of images; and determining whether a particular surgical trajectory will help or damage a patient's brain. These experts decide, infer, interpret and estimate mostly qualitatively and often in reliance upon a personal knowledge base.

In my vision of the future, things will be quite different. We will use database information to reference and compare new or novel cases; to search for and compute biomarkers; and to determine the safest surgical course. How far are we from making this happen? Not as far as you'd think.

WHERE WE ARE NOW: HONING IN ON LOCATIONS AND INTEGRATING DATA

We are in the middle of one of the most significant scientific transformations in the study and treatment of the brain since the development of CT scans. Computational strategies that combine, compare, measure and visualize data-based images can provide enormous quantitative power to our understanding of brain structure and function in health and disease. Furthermore, we are now able to integrate disparate information from different modalities and at different scales, much as the brain itself apparently operates. And we can test relationships between data that comes from different cohorts, using different methods, to study different aspects of the brain, on different subjects, from different laboratories.

Comparing and contrasting brain image data requires a complete description of 'where' things are happening in the brain. Sometimes the best approach is to point to things in different ways. Just as a house can be described by its address, GPS coordinates, neighbors, appearance, or proximity to an intersection, we can describe brain locations in a variety of ways. For example, we can locate an activation site by area (e.g., Brodmann area 46); by coordinate system (Talairach and Tournoux coordinate (x,y,z)); by brain region (e.g., the pars opercularis of the inferior frontal gyrus); or by proximity to a blood vessel or layer of the cortex. None of these individually is very precise but, collectively, these descriptions make it easier to identify and compare regions where an activation occurred. This is now being done by a number of computational labs, and refinements will one day make these approaches avail-

able for a broad range of investigative and clinical applications.

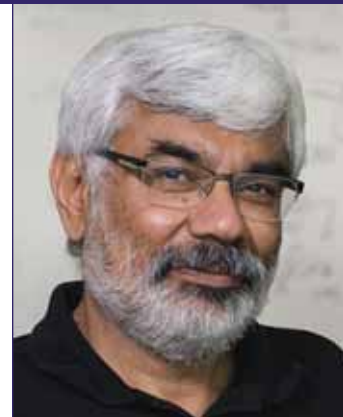
In addition to mapping locations in the brain, computational anatomists must deal with differences among individuals in a probabilistic way, retaining information regarding variability and other group statistics. Ongoing efforts are creating better canonical atlases to represent subpopulations—of, say, healthy 20-somethings or Alzheimer's patients—while retaining well-resolved anatomical features that are vital to assist automated algorithms for aligning data to them. These atlases can even have a time-varying component allowing subjects of different ages to be brought into the atlas using an age-appropriate transformation. Rather than average images together voxel-by-voxel to produce a blurred template, as was done in many "first generation" statistical atlases, many groups are developing practical methods to create well-resolved canonical atlas images that represent the statistical mean anatomy for a population, using deformation averaging and Lie group methods on deformation tensors or geodesics on groups of diffeomorphic flows. These approaches are complex, but are advantageous as they are close (in a strictly defined mathematical sense) to the brains being normalized to them and are likely to improve spatial accuracy and reduce sources of bias when comparing datasets in a canonical coordinate system.

WHERE WE NEED TO GO

We are now within striking distance of creating robust methods for atlasing the brain and for integrating data to build more diverse and specific atlases of subpopulations. But designing appropriate reference systems for brain data presents considerable ongoing challenges, since these systems must capture how brain structure and function vary in large populations, across age, gender, and different disease states as well as across imaging modalities, time and even across species. And to get to the point where computational anatomy has utility in both investigative and clinical scenarios, we need to put it all together in an intuitive way. Only then will it be adopted by the people who could take full advantage of it by letting it guide their judgments.

ACKNOWLEDGMENTS: This work was supported by the National Institutes of Health through the NIH Roadmap for Medical Research, grant U54 RR021813 (entitled Center for Computational Biology (CCB) (<http://loni.ucla.edu/CCB/>)). □

Grant applications for the development and maintenance of software and infrastructure should: Compete with basic research applications



Biomedicine has a strong tradition of being an experiment- and phenomenology-driven science. But over the course of the last century, the field has become both more quantitative and more interdisciplinary. And in the late part of the 20th century, it also became highly data-driven, owing in large part to genome sequencing and the emergence of high throughput and computer technologies that could measure components of living systems at large scale. Meanwhile, large advances in computers and computational methods inspired confidence in our ability to deal with biomedicine as a quantitative science. In other words, biomedicine began to be viewed as a constructionist science where data and knowledge could be integrated to reconstruct models of processes in living systems. This tight integration of biomedicine with computation and informatics is the fundamental reason grant applications for the development and maintenance of software and infrastructure should compete with basic research applications and be considered by study group panels that have sufficient biology expertise. I would also point out that:

1. Biomedical research is highly context-specific and in order to be considered mainstream, software and infrastructure proposals should be biology-rich, useful and have content as opposed to mere form. Competing with basic research applications will keep the applicants “honest” in learning the devil-in-the-details biology that serves as the underpinning for the software.

2. While computer professionalism and a high degree of digital sophistication is needed to generate “good-practices” software and tools, sophisticated software that does not embed biomedicine integrally is at best irrelevant. It is easy to see that one of the most-utilized software infrastructures, namely *GenBank*, did not begin with best-practices software engineering paradigms.

3. The price of software engineering is often cited as a reason for a separate funding mechanism since traditional biomedical research salaries are modest compared to those of computer professionals. Again, some of this is true, but most biomedical researchers appreciate aspects of this argument and in itself this does not warrant a

separate consideration for software engineering proposals.

4. The most important argument that has an intellectual basis is the fact that most biomedical researchers, while familiar with basic bioinformatics, are not computer science-savvy and hence would find it difficult to both assess and accept a higher proportion of software proposals. While acknowledging aspects of this argument, in the long run, having a “two-culture” paradigm would be detrimental for strong integration between computational sciences and biomedicine.

5. A significant argument that has been made repeatedly is that most basic research applications are “hypothesis-driven” while proposals that are computational and software in origin are “knowledge/discovery-driven.” It remains

“Competing with basic research applications will keep the applicants ‘honest’ in learning the devil-in-the-details biology that serves as the underpinning for the software,” says Subramaniam.

to be seen whether researchers will accept that analysis of data and incorporation of biological legacy knowledge to generate novel hypotheses is as good as, if not superior to, hypotheses that are stated based on experience and intuition. This question warrants significant discussion and in itself is not a reason for considering software and infrastructure proposals outside of basic research applications.

None of the arguments above are intended against the need for and importance of allocating significant resources for further developments in computational/software/informatics research. However, they are meant as a challenge to the computation, data and informatics-inclined researchers to embrace biology in a more intrusive manner and become polymaths who can participate in building next-generation biomedicine. □

DETAILS

Shankar Subramaniam,
PhD, professor of
bioengineering at
the University of
California, San Diego

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>.



Grant applications for the development and maintenance of software and infrastructure should: Have a separate funding mechanism

Although I agree with many of my esteemed colleague's points—especially the need for strong biological input into the review process—I disagree with his conclusions.

Enormous challenges and opportunities confront us. The growing overlap of structural biology, molecular biology, cell biology, genetics/genomics, synthetic biology, and systems biology will lead to a new era of quantitative physiological understanding, with breakthrough advances in preventive medicine, drug design, and medical interventions of all kinds. Computation is now inseparable from all of these fields, and new hardware developments (e.g., multicore and specialized processors) are increasing the complexity of computer engineering, adding even more complexity to software

“Shoe-horning development and maintenance of software and infrastructure into long-standing evaluation and funding mechanisms is not the answer,” says Stiles.

development. This, together with the scope of scientific problems, has changed the landscape of scientific computing dramatically. In general, the NIH and NSF, despite their roles as primary academic funding agencies, have yet to fully address the human and monetary expense of large-scale/high-performance software development, maintenance, and training, despite centers devoted to biomedical computation, supercomputing, and an often-stated desire for “hardened,” “integrated,” and “near-commercial quality” programming. Even the DOD, DOE, DARPA, and the national laboratories, with a longer-standing focus on large-scale computation (albeit often classified), are struggling to find effective mechanisms of support for increasingly complex software and computational infrastructure.

Further compounding the biomedical problem, many experimental investigators still have no real frame of reference for software development and computing. Lab and home computers are seen largely as appliances, and software, regardless of origin, is expected to either run in a web browser applet or be downloadable, instantly installable, and intuitively usable.

Satisfying these expectations is difficult even for large commercial software and computing firms. For research projects, this is feasible only in a limited number of cases, and at the cutting edge can lead to counterproductive expectations from potential users. The future of biomedical computation spans enormous ranges of space and time, and lies at the interface of biology, engineering, chemistry, physics, mathematics, and computer science. Far-sighted multidisciplinary development is critical to success, and suitable mechanisms of evaluation and support must be developed and maintained in and of themselves.

Shoe-horning development and maintenance of software and infrastructure into long-standing evaluation and funding mechanisms is not the answer. Alternatives have been tried, including a standing special emphasis panel on software development and maintenance at the NIH. While this is a step in the right direction, it has not yet become a chartered study section primarily because the number of submitted proposals has remained too small. Why? Lack of interest or need? Hardly. Instead it reflects the still incomplete assimilation of leading-edge computing into mainstream biomedical research, and the poor fit of

the funding mechanism. All but the smallest projects exceed typical R01 scales, both in time and money. Even at the level of centers, a common point raised during program evaluations has been “underfunded.” This is not to say that bigger centers are the answer. No, the answer is probably expanded stable funding at the level of collaborating laboratories, with the primary emphasis on far-sighted multidisciplinary development and less on short-term hypotheses. Regardless, a sustained emphasis and support mechanism is essential. No commercial computing enterprise can survive without professional software and hardware engineers, and thus must provide corresponding salaries and stability. The same is true for the integrated multidisciplinary computational research needed for the future of biomedicine. □

DETAILS

Joel R. Stiles, MD, PhD, director of the National Resource for Biomedical Supercomputing at the Pittsburgh Supercomputing Center, and associate professor of biology at Carnegie Mellon University

NewsBytes

Automating Scientific Discovery

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

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

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

“The idea of using a robot is not news, but what’s different about ours is the robot was also involved in developing hypotheses and experiments on its own,” says **Ross King**.

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

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

language Prolog; another is an ontology describing laboratory experiments, based on the open-source project EXPO. The robot also recorded its own experimental information as it worked.

“One of the advantages of a robot scientist is that you get all that metadata for free,” says King. “We can understand far more about the structure of the experiment than we would if only humans had been involved.”

ADAM’s four computers directed the experiments, with robot arms moving yeast mutants from freezer to incubators to plate readers. Ultimately, it found 12 gene-enzyme pairings that the authors were able to confirm. In some cases, the link between gene and enzyme was found to be supported by

“The idea of using a robot is not news, but what’s different about ours is the robot was also involved in developing hypotheses and experiments on its own,” says **Ross King**.

literature even though it was missing from ADAM’s starting data. For others, the authors double-checked ADAM’s results by purifying and testing the protein themselves.

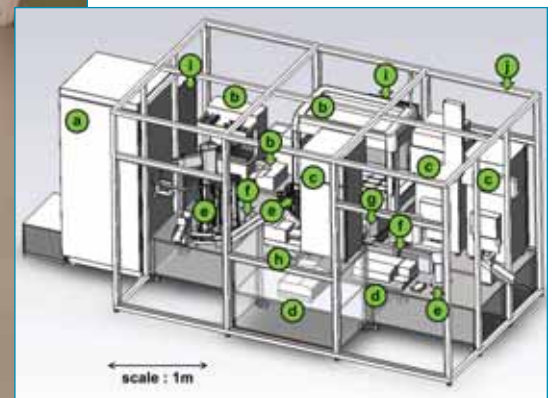
The successful matches “are mostly to do with odd pieces of biochemistry that hadn’t been sorted out yet,” King says, which explains why the enzymes remained orphans for so long. Some were isozymes, with more than one gene encoding the same function, and others were promiscuous enzymes that catalyze more than one reaction.

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

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

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

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



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

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

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

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

—By **Beth Skwarecki**

The Function of DNA Form

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

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

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

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

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

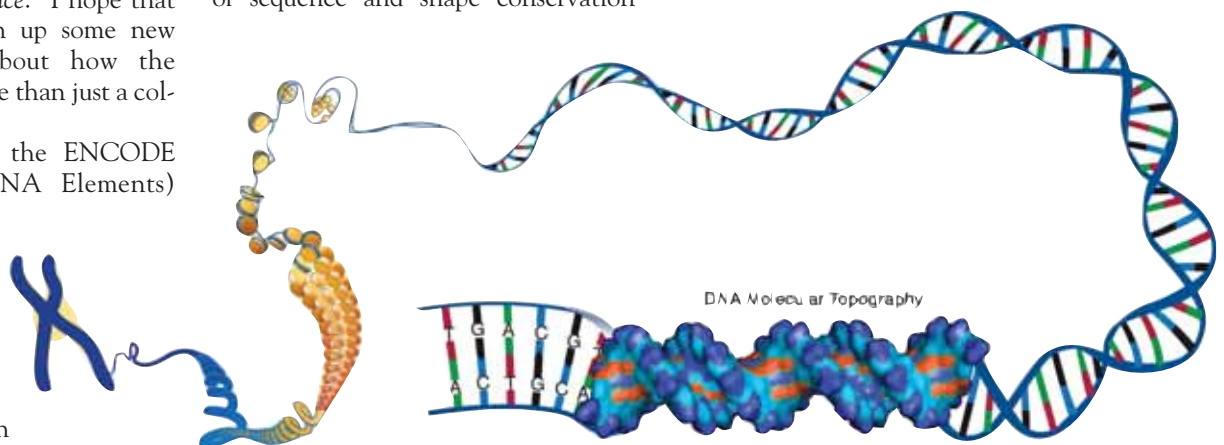
To answer that question, Tullius, **Elliott Margulies, PhD** of the National Institutes of Health, **Steve Parker** of Boston University and their colleagues created a computer program called Chai. The program compares computational predictions of DNA shapes from the same one percent of the human genome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

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

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

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

—By **Rachel Tompa, PhD**



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

Semantic Publishing and Scientific Journals

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

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

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

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

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

a helpful difference,” Shotton says.

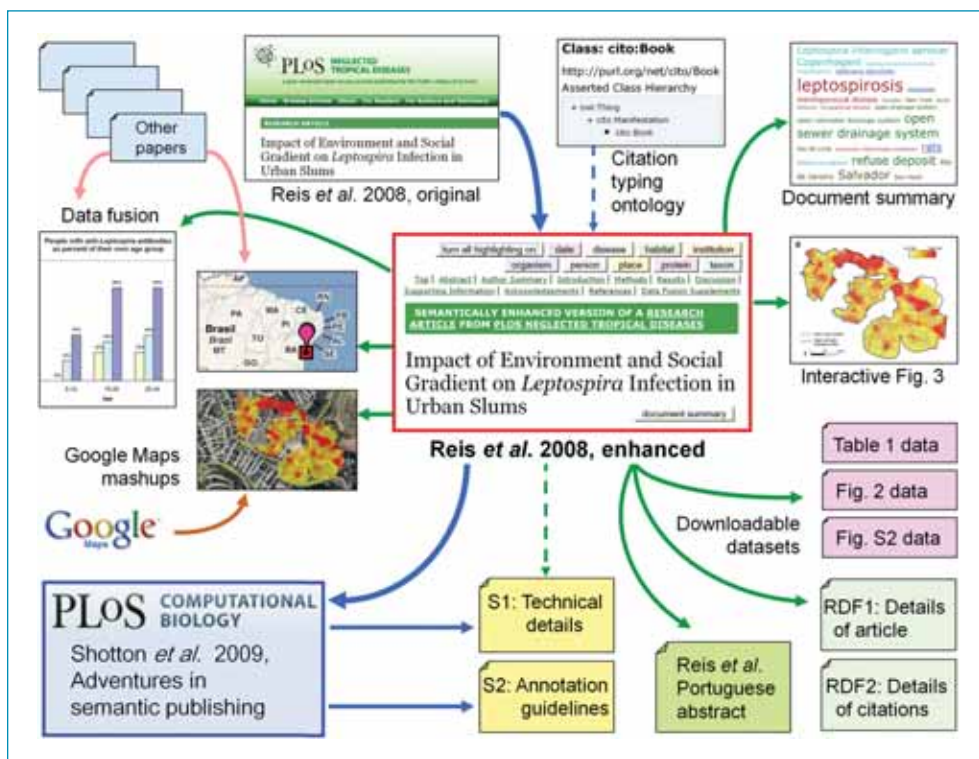
Shotton and his colleagues manually curated the paper—a slow process that could be improved via automa-

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

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

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

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



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

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

—By *Katharine Miller*

Open Source Tools for Parsing Clinical Records

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

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

Until recently, researchers could only

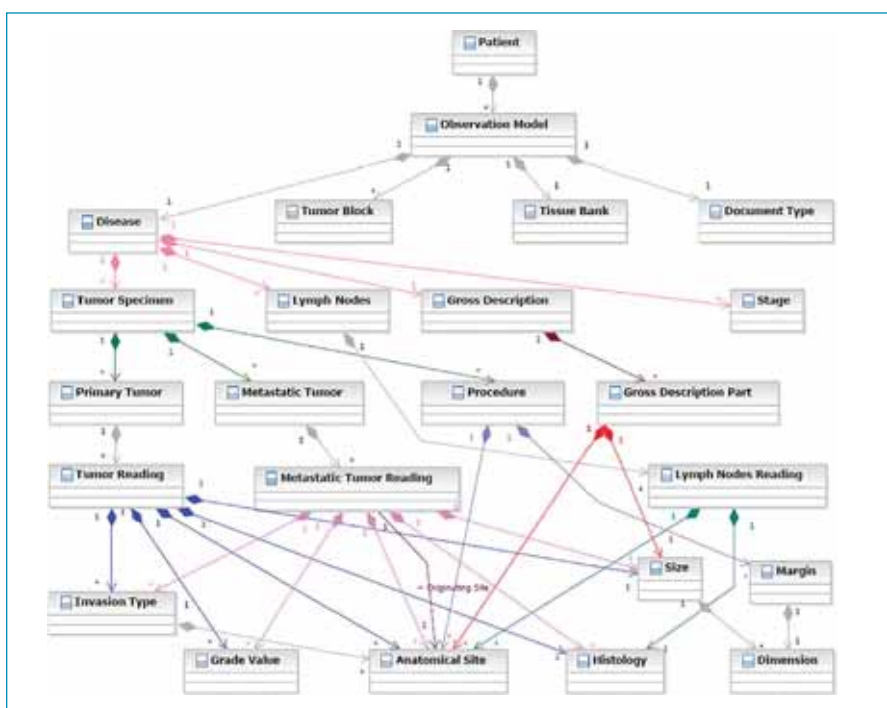
access valuable data within medical records by hiring medical professionals to read charts and abstract the information out to a case report form. And while general architectures for building NLP pipelines and automating this process of extracting information from medical records aren’t new, and some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

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

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

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

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



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

ple, the size of the primary tumor.

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

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

—By Katharine Miller

Online Searches Warn of Flu Spikes

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

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

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

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

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

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

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

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

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

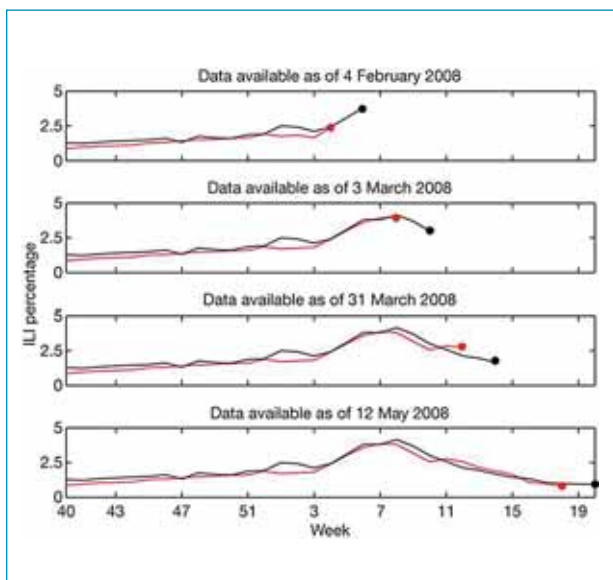
the up and down of seasonal flu spikes over five years. Those that correlated best (the top 45) were all flu-related.

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

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

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

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



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

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

—By *Louisa Dalton*

Flowing through the Interactome

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

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

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

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

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

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

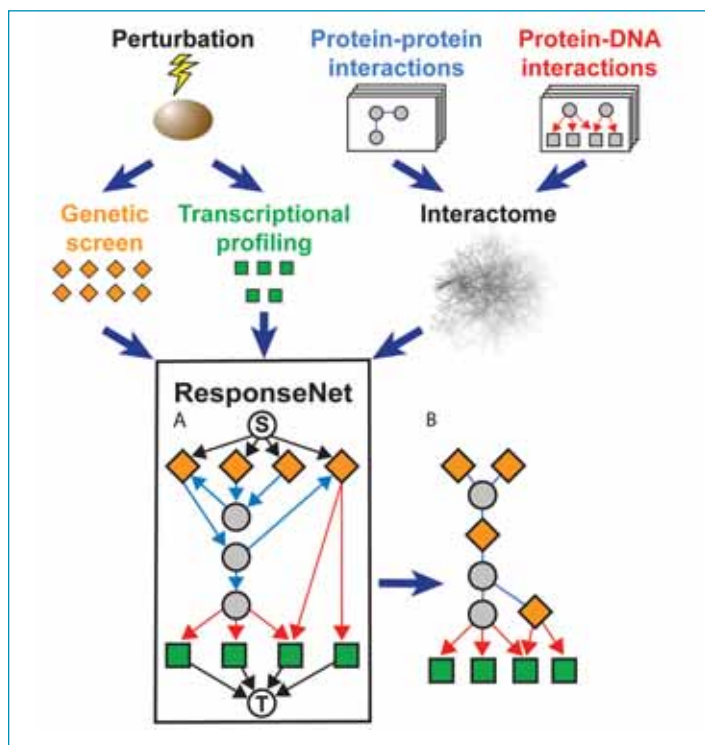
“Their solution is

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

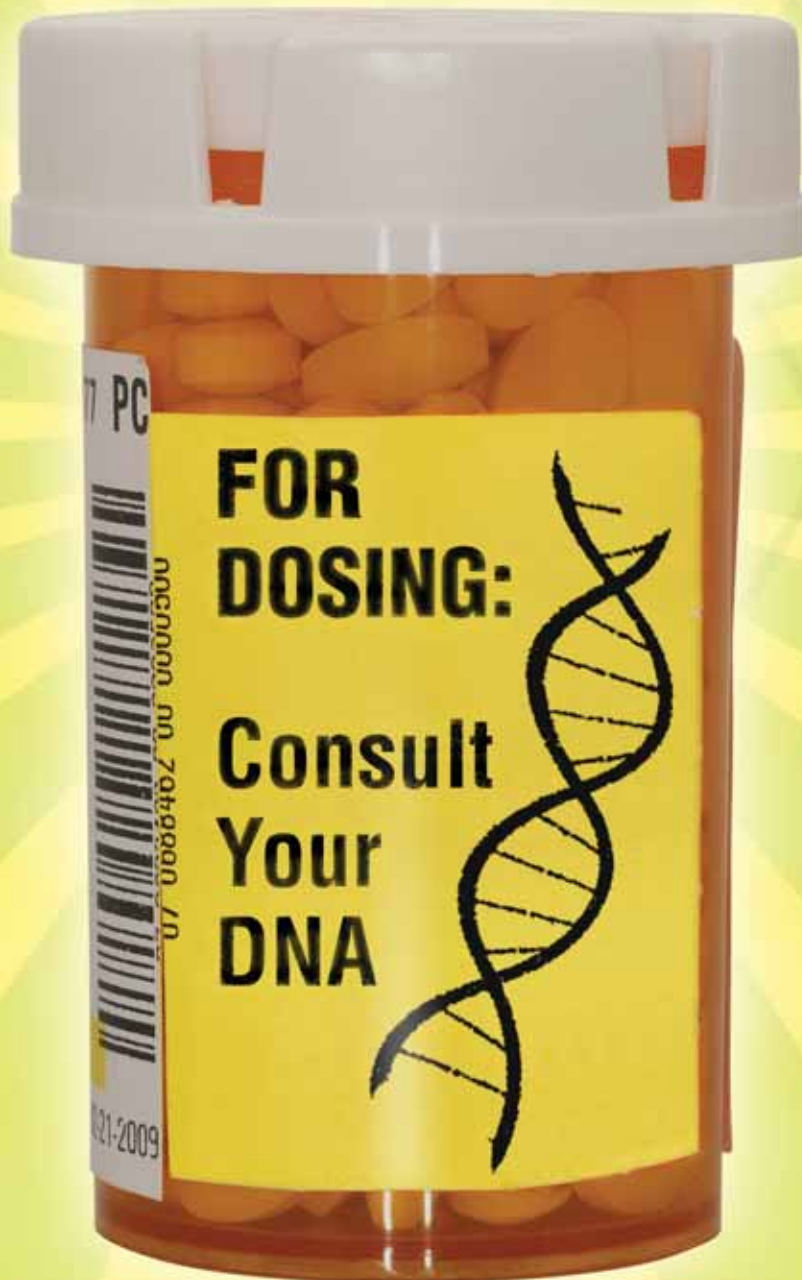
“I think this is a step toward a much more realistic and mechanistic view of what’s going on in cells that could ultimately do much better in terms of predicting what’s important in diseases,” says co-author Ernest Fraenkel.

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

—By *Liz Savage* □



From SNPs to Can genes



Prescriptions: predict drug response?

Decades of steady progress in pharmacogenetics have unearthed hundreds of associations between genes and drug response. But the field has to solve some theoretical and practical issues before it can deliver on the promise of personalized drug therapy.

As algorithms go, it's deceptively simple. Just add together eight weighted pieces of patient information—age, height, weight, race, data about two genes, and a pair of clinical parameters. Yet this straightforward linear equation could mark a watershed moment in medical science.

The algorithm in question helps physicians prescribe a safe and effective dose of the blood thinner warfarin. Currently, because warfarin's optimal dose varies tenfold among patients, physicians prescribe an intermediate dose and make adjustments over the course of several weeks to achieve the desired effect. But this approach carries huge risks: Too high a dose could trigger fatal bleeding while an insufficient dose might allow dangerous blood clots to form. And it's hard to tell how a patient will react. "If you gave warfarin to a huge football player and a tiny grandma, the football player could bleed uncontrollably at a dose much smaller than what you give grandma," says **Balaji Srinivasan, PhD**, a


Stanford University statistician who worked with the International Warfarin Pharmacogenetics Consortium—an unprecedented collaboration of 21 research groups that jointly developed the new dosing algorithm.

is nearly 50 percent better at identifying patients who need low doses and more than 3 times better at identifying those who need high doses. Given warfarin's wide use—up to two million new patients take it each year—the con-

"If you gave warfarin to a huge football player and a tiny grandma, the football player could bleed uncontrollably at a dose much smaller than what you give grandma," says Balaji Srinivasan.

According to the Consortium's study of about 5,000 patients from a variety of ethnic backgrounds, the new algorithm can significantly lower the risk of under- or over-dosing compared to using clinical information alone. It

crete benefits are obvious. But the greater significance is symbolic: the genetic dosing algorithm for warfarin could be a major milestone in the evolution of drug prescription from a trial-and-error strategy to an exact science.



“It’s an amazing story at many levels,” says Stanford University computational biologist **Russ Altman, MD, PhD**, one of the organizers of the Consortium and a senior author of its warfarin study. “You had 21 research groups in nine countries who pooled all their data together to come up with this algorithm. And we clearly show that genotype-based dosing can be a vast improvement over the guessing game physicians have to play now.”

So is personalized drug therapy—

“You had 21 research groups in nine countries who pooled all their data together to come up with this algorithm. And we clearly show that genotype-based dosing can be a vast improvement over the guessing game physicians have to play now,” Russ Altman says.

prescribing the right drug at the right dose for an individual patient—about to become a reality?

Perhaps not right away. Pharmacogenetics, the study of genetic factors that influence drug response—and its younger sibling pharmacogenomics, which adopts large-scale genome-wide methods—are indeed hot research areas. (A PubMed search with the two terms brings up nearly 9,000 entries, most from this decade.) As with research in general, however, some studies in pharmacogenetics have turned out to be poorly designed. Others are well designed but don’t give a biologically significant result. And still others make clear-cut biological predictions, but with limited clinical value. Finally, even interventions of proven worth are struggling to reach the clinic. So it may be a while before insurers stop using the dreaded “experimental” adjective when referring to these techniques.

Despite these challenges, the flood of pharmacogenetics results pouring in gives hope that at least a few will

make it into everyday use. The need is obvious: nearly 90 percent of drugs don’t work for half the people; worse, adverse reactions to drugs send millions of patients each year to the hospital and cause more than 100,000 deaths. In most cases, genetic factors seem to play a role. An editorial that accompanied the warfarin study in the February *New England Journal of Medicine* says, “A better understanding of individual differences in the response, either positive or negative,

to medicines should be an overarching goal for pharmacotherapy over the next decade.”

BEANS TO GENES

To see where pharmacogenetics is headed, it is instructive to take a step back and see how it emerged. Although the field has gained much of its prominence this century, it has a long and eventful history. Some researchers credit Pythagoras, in the 6th century BC, with making the first contribution to it when he noted that eating fava beans made some people sick. (Two-and-half millenia later, scientists would discover the cause: a variant in a red blood cell enzyme that also causes abnormal responses to anti-malarial drugs.) One facetious researcher gives the credit to Karl Marx—“didn’t he say to each according to his need?” A more serious claimant for the honor is chemist **Arthur Fox**. While working at the DuPont laboratories in 1931, Fox accidentally released a cloud of a chemical that he was working on. He felt noth-

ing, but a colleague complained of a bitter taste sensation. Intrigued, Fox investigated this further and discovered that the ability to taste the compound is an inherited trait, later shown to be due to variants in a bitter taste receptor gene. In the 1950s, University of Toronto medical scientist **Werner Kalow, MD**, found that people who suffocated to death after getting certain muscle relaxants had inherited a variant of the gene for pseudocholinesterase, an enzyme involved in nerve function. Kalow would go on to pen a monograph in 1962 on pharmacogenetics, defining the term as the study of heredity and the response to drugs.

The work of Fox and Kalow set the template for pharmacogenetics that lasted until the mid-90s: identify a peculiarity in drug response and look for inherited variations in a relevant gene or enzyme. During the next few decades, researchers used this approach to explain atypical responses to the tuberculosis drug isoniazid, the malaria drug primaquine, and the heart arrhythmia drug sparteine. In 1964, even good old alcohol got a response enzyme; people with a variant of aldehyde dehydrogenase can get violently ill after even a tiny sip from the flask, an effect the alcohol aversion drug Antabuse achieves in other people by blocking the enzyme.

While these pioneering studies showed that drug response could be a hereditary trait, they dealt with relatively simple problems. “Classically, in pharmacogenetics the focus was on cases where we thought there’s a single gene and single mutation that was going to explain drug response,” explains **Marylyn Ritchie, PhD**, a geneticist at Vanderbilt University.

In most cases however, drug response is influenced by many genes that each have only a modest impact. These “pharmacogenes” come in two flavors: At the “front end” stand the so-called pharmacokinetic genes that determine how quickly the body breaks the drug down and eliminates it; at the “back end” lurk the so-called pharmacodynamic genes whose function the drug targets. Patients with a sluggish front end but touchy targets could have an excessive, even toxic, response—such as warfarin-induced bleeding—while those with an unusually brisk front end and apathetic tar-

gets could barely respond.

Another drawback of many early efforts was that they relied on family-based data to locate genetic markers that were inherited along with the trait being studied. But such studies require huge sample sizes to produce valid results—especially when the effect of the gene variant may be small, as is the case for many genetic drug responses. In a 1996 paper in *Science*, Neil Risch, PhD, and Kathleen Merikangas, PhD, showed one would need to look at 70,000 families in order to pinpoint a gene variant that occurs 10 percent of the time and increases the relative risk of a certain trait by 50 percent—which would be high by drug response standards. To detect genes of such modest influence, the authors recommend a different approach—the association study—“even if one needs to test every gene in the genome.” This type of study compares the gene variants of patients who respond normally to a drug (controls) with those of patients who have an adverse reaction (cases). (If the drug response can be quantified, as with warfarin, the study may look at patients with a range of responses instead.)

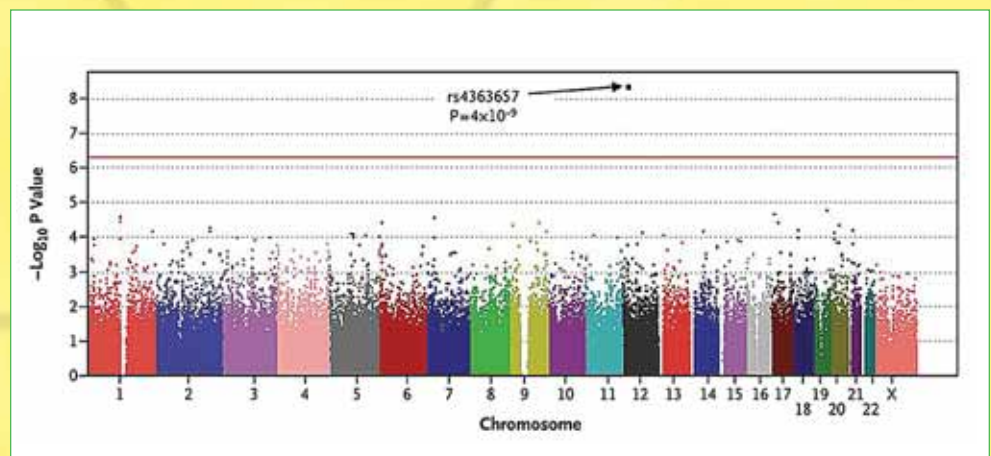
Early pharmacogenetic association studies didn't have the resources to look at the entire genome; instead, they focused on “candidate” genes—those suspected of modulating response to the drug. A 2002 effort comparing 18 cases and 167 controls taking the HIV drug abacavir identified a variant of an immune system-related gene (*HLA-B*) as a possible culprit in causing a toxic skin reaction. In another study, the cancer drug irinotecan was found to be toxic in patients with a variant in the promoter of a gene (*UGT1A1*) that helps the body break down and eliminate certain foreign compounds. Later studies have strongly validated both findings, and genetic testing for sensitivity to abacavir and irinotecan is now widely available.

Although the candidate gene study is a useful tool, it has an important limitation: “You make *a priori* assumptions about which genes may be important,” says Eileen Dolan, PhD, a cancer pharmacologist at the University of Chicago interested in finding the genetic basis of toxicity to cancer drugs. “In the process, you risk missing some important ones.”

LOOKING FAR AND WIDE

To cast the net wider, many present-day association studies look at the entire genome, or a large fraction of it, using new high throughput platforms such as the “SNP-chip” that can simultaneously probe millions of single point mutations in the DNA (also called single nucleotide polymorphisms, or SNPs). This method has already paid off handsomely in disease studies—discovering a strong suspected link between age-related macular degeneration and a specific variant of a gene involved in inflammation. In pharmacogenetics too, useful results have started trickling in. In April last year a study of warfarin response that examined about half a million SNPs in 181 patients confirmed findings from earlier, smaller studies that the two genes included in the Consortium's dosing algorithm (*VKORC1* and *CYP2C9*) are indeed the ones with the greatest impact. Another warfarin study published the same month hunted among about 1,200 markers within 170 pharmacogenes and caught a third gene of interest (*CYP4F2*). Knowing the patient's version of this gene can improve dosing accuracy by an extra five percent (about one milligram per day), the authors report. A study last August used a genome-wide scan of 175 subjects to find that a variant of a membrane transporter gene (*SLCO1B1*) is associated with toxic reactions to the

“In a genome-wide study you come with an open mind,” says Eileen Dolan. “When you study that way, you often come up with genes that you didn't even conceptualize could be important to the disease or drug response.”



Statin Response Genes. Some people who take statins to reduce cholesterol levels end up with a new problem: a type of muscular damage called myopathy. According to a 2008 genome-wide association study of 85 cases and 90 controls, this reaction is strongly associated with a variant of an anion transporter gene (*SLCO1B1*) on chromosome 12. In this map of 300,000 SNPs, the horizontal axis shows the genomic positions of SNPs grouped by chromosome, and the vertical axis shows the probability of error (p-value) for each SNP-response association. A SNP within *SLCO1B1* (the dot above the horizontal line across the chart) had the strongest score of 4×10^{-9} . An individual with two copies of this variant has a 17-fold higher risk of statin-induced myopathy. Reprinted with permission from Massachusetts Medical Society: *New England Journal of Medicine* 359:789-799 (2008).

“It is hard to find large sample sizes of patients who are receiving pretty uniform treatment and for whom we have adequate follow-up,” says Mary Relling. “That’s really the rate-limiting step.”

cholesterol-lowering drug simvastatin. “In a genome-wide study you come with an open mind,” says Dolan. “When you study that way, you often come up with genes that you didn’t even conceptualize could be important to the disease or drug response.”

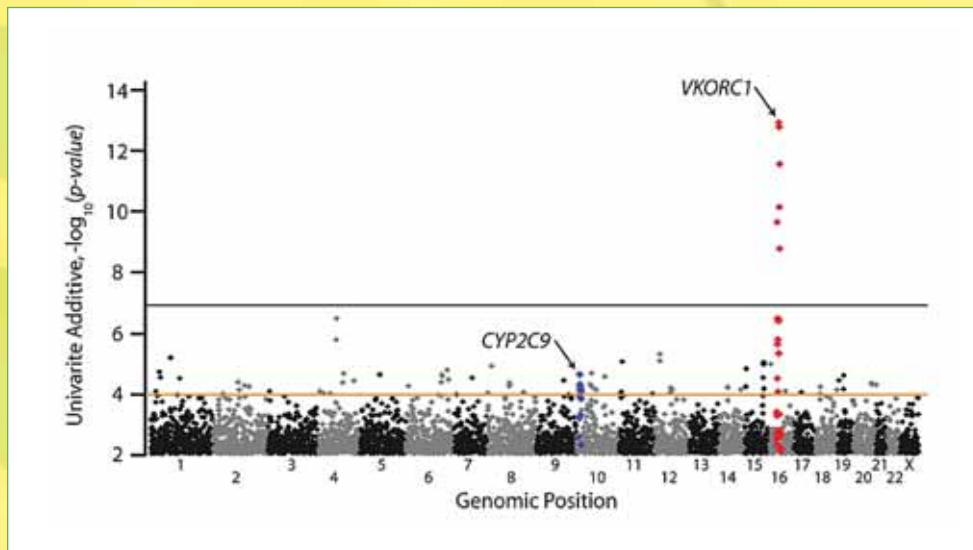
To be able to discriminate reliably between true and false associations, a study that looks at a large number of genetic markers needs a very large number of subjects. This is hard enough to achieve in a disease study—the biggest ones to date boast of a few thousand subjects at most. For drug response studies, a few hundred would be a luxury; follow-up studies to replicate results may have even more difficulty finding subjects. “Not only do you need people with the same disease, you need people treated the same way with the same drug,” says Ritchie. And unlike a disease study, Ritchie points out, “all your controls should have the disease as well—they are actually cases, too.” As a result, most genome-wide studies tend to be under-powered from a statistical perspective. “It is hard to find large sample sizes of patients who are receiving pretty uniform treatment and for whom we have adequate follow-up,” says Mary Relling, PharmD,

who uses pharmacogenetics to improve drug therapy for children with leukemia at St. Jude’s Children’s Research Hospital in Memphis. “That’s really the rate-limiting step.”

One way to gain statistical power is to choose samples carefully. For instance, a study to predict the degree of response to a drug could be better off with 300 low responders and 300 high responders rather than 200 each of low, medium, and high responders. This trick, however, doesn’t apply to studies of drug toxicity, where a person has either a normal or an adverse reaction. And for rare, potentially serious adverse responses to drugs, finding enough samples gets even harder. Examples include Stevens-Johnson syndrome, a serious skin rash; rhabdomyolysis, a muscle-destroying condition; QT prolongation, an abnormal heart condition; and drug-induced liver injury. These reactions may strike as few as one patient among several tens of thousands. “Because of the low incidence of such events, it is almost impossible to study them within a single academic institution or at individual pharma companies,” says **Andrea Califano, PhD**, a professor of biomedical informatics at Columbia University.

Califano is the head of the data analysis and coordination center of the International Serious Adverse Event Consortium, a pharmaceutical company and Wellcome Trust funded effort to identify the genetic determinants of rare adverse drug reactions. For their Serious Skin Rash study in 2007, the Consortium set up about 20 centers in the United States, the United Kingdom and Canada to enroll study subjects. “Even this massive effort yielded only 71 cases and 135 matched controls,” says Califano.

Finding enough samples for a genome-wide study that may examine up to a million SNPs is one challenge; making sense of the huge amount of data generated is another. “It is the typical sorting the wheat from the chaff problem,” says **Howard McLeod, PharmD**, of the University of North Carolina at Chapel Hill. “Where there’s a whole lot of chaff, there’s got to be some wheat in it somewhere.” From a statistical per-



Warfarin Response Genes. This figure from the genome-wide study by Cooper and his colleagues shows the strength of association of about 500,000 genetic markers with warfarin response. The horizontal axis shows the genomic positions of SNPs grouped by chromosomes. SNPs lying within 500 kilobases of the VKORC1 and CYP2C9 genes are shown in red and blue, respectively. The vertical axis shows the probability of error (p-value) of each SNP-response association. Overall, after accounting for all related SNPs and data from all replication experiments, VKORC1 scored a stunning 4.7×10^{-24} , CYP2C9 a more modest, but still convincing 6.2×10^{-12} . This research was originally published in *Blood*. Cooper GM, et al., “A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose,” *Blood* 112: 1022-1027 (2008). © American Society of Hematology.

For GWAS studies, says Howard McLeod, “The statistical tools out there are pretty crude. They’re geared towards preventing false positives, whereas initially what we need is some method that enriches true positives, or helps minimize false negatives.”

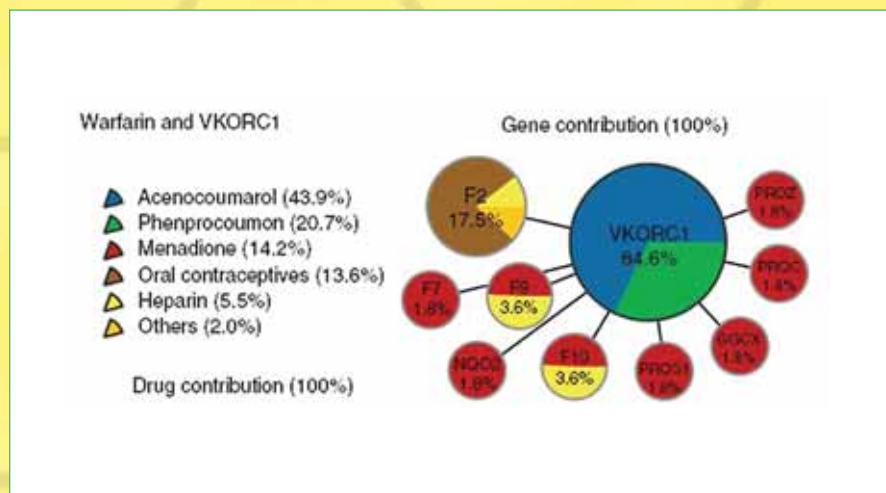
spective, this is a classic multiple-comparison problem with a high risk of false positives—chaff can masquerade as wheat. Several high-profile failures attest to this risk, such as a 2005 study of Parkinson’s disease that later studies could not replicate. “The problem with whole genome studies is dealing with extremely wide data matrices,” says Srinivasan. “Unless the signal is so ridiculously strong that it jumps out at you, there are strong theoretical reasons to think that such datasets will be highly under-determined.”

Studies often overcompensate by being excessively cautious. Consider one that compares n genetic markers between cases and controls. To keep the overall probability of making a wrong association below p , the study would typically try to keep the risk of getting a false positive from a single comparison below p/n . To ensure that a study with 100,000 SNPs has an error rate of 0.01, the error rate for individual SNP comparisons has to be 1.0×10^{-7} —each test would need to be 99.99999 percent reliable. Known as the Bonferroni adjustment, this method assumes the worst-case scenario of the tests being independent. In reality, however, SNPs are usually inherited in bunches and so the


tests are not independent. The Bonferroni adjustment ends up being too harsh on the type of moderate-effect association that typifies most gene-drug ties. Alternative scoring schemes exist, but none is adept at the delicate balancing act of finding true gene-drug associations while avoiding false matches. “The statistical tools out there are pretty crude,” says McLeod. “They’re geared towards preventing false positives, whereas initially what we need is some method that enriches true positives, or helps minimize false negatives.”

GETTING IT RIGHT: ADDING KNOWN BIOLOGICAL DATA

One way to enrich true positives is to bring in prior biological information, says Altman. As he points out, a vast amount of data about biological pathways and mechanisms of drug action already exists. To use this information, Altman makes some common-sense assumptions about the interactions of genes and drugs. For instance, genes whose proteins interact with each other are more likely to interact with the same small molecule drugs. At the same time, drugs that have a similar chemical structure, or drugs that are



Biological Priors. For a given query drug and indication, Altman and his colleagues use pre-existing biological knowledge in the form of gene-drug, drug-target, and gene-gene interactions to rank genes in the order of pharmacogenetic relevance. This preliminary ranking can then help a genome-wide study focus on important candidates and avoid making false associations. For the blood-thinning drug warfarin, pre-existing knowledge from other vitamin K agonists (acenocoumarol, phenprocoumon), a vitamin K2 precursor (menadione), other blood thinners (heparin), and important genes in the anticoagulation pathway (F2, F9, etc.) leads to a high ranking for VKORC1. The figure shows how different drugs as well as different genes contribute to this ranking. Oral contraceptives may seem out of place, but Altman explains: “Oral contraceptives can cause clotting...and have distant structural similarity to warfarin, and that’s what gets them on the list.” Reprinted by permission from MacMillan publishers: *Clinical Pharmacology & Therapeutics*, Hansen, NT, et al., *Generating Genome-Scale Candidate Gene Lists for Pharmacogenomics*, (2009).



chemically dissimilar but treat the same disease, are likely to interact with the same genes. For a given query drug and indication, Altman uses these principles to rank-order all the 12,000 or so genes whose interactions are reasonably well understood. For a blood pressure drug such as nadolol, for instance, his method would prioritize genes that interact directly or indirectly with other beta blockers as well as with other blood pressure medications such as ACE inhibitors or calcium channel blockers. Altman then combines this result with data from a genome-wide study. Doing so helps to move the analysis from the purely statistical realm to one that takes biological mechanisms into account, he says. “Now you are beginning to tell a story other than ‘they are correlated.’”

Altman and his team get their prior biological information from three public databases—PharmGKB, DrugBank, and InWeb—for gene-drug, drug-target, and gene-gene interactions, respectively. (It turns out that each gene interacts with about 25 other genes on average while a typical pharmacogene interacts with about 3 drugs.) Based on the prior biological information alone, the now-famous warfarin gene duo (*VKORC1* and *CYP2C9*) ranked 11 and 16, respec-

“You should get your prior, put it in an envelope, and get it time-stamped by the post office,” says Russ Altman. “Otherwise you are going to have a rough idea of what genes are involved, and that is going to poison your prior.”

tively. Based purely on raw experimental data from a genome-wide association study, the genes would rank 25 and 34, says Altman. Combining the prior scores with the experimental data boosted the rankings to 1 and 2. (The genome-wide study too got these final rankings, but using other, less general criteria.)

Altman and his colleagues get similarly encouraging results for some other common drugs using the same approach. “It works amazingly well,” says Altman. The key, he says, is to avoid infusing any bias into the analysis of the association data. “You should get your prior, put it in an envelope, and get it time-stamped by the post office,” he says. “Otherwise you are going to have a rough idea of what genes are involved, and that is going to poison your prior.”

GETTING IT RIGHT: ADDRESSING COMPLEXITY

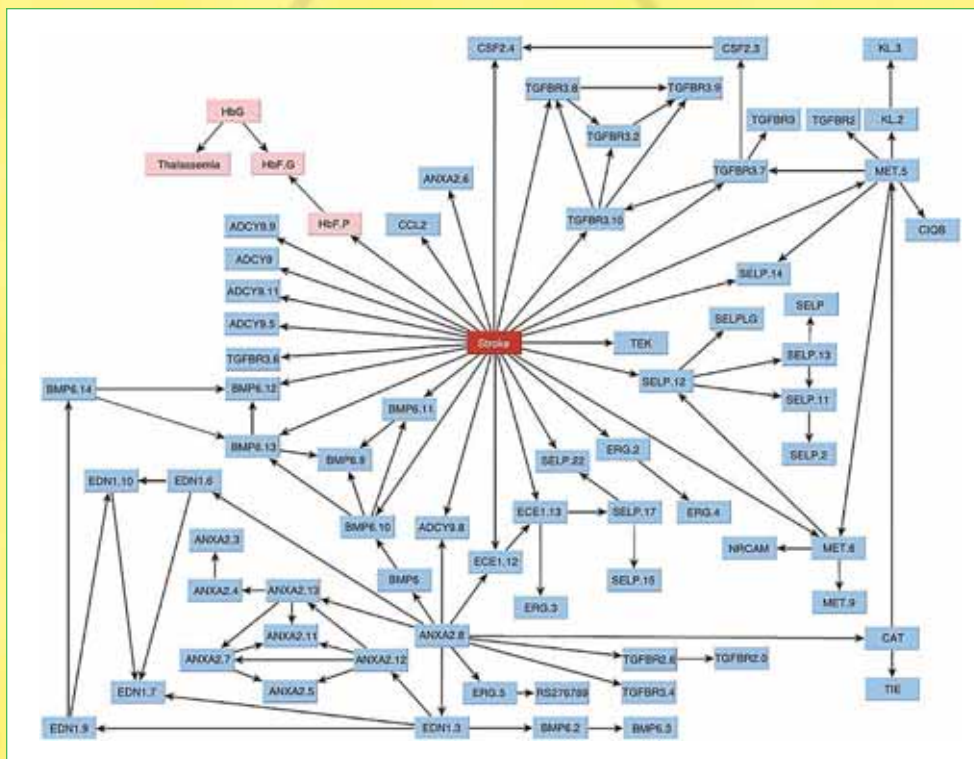
While serious drug reactions such as Stevens-Johnson syndrome may have a simple genetic cause, normal drug response may arise from a more complex interplay of genetic factors. Association studies that compare individual SNPs could miss these factors. “The problem is that most of these SNPs and genes by themselves explain only a small amount of the variation,” says **Scott Weiss, MD**, a Harvard Medical School researcher who studies the genetic basis of response to asthma drugs. “I am a firm believer that modeling epistasis—gene-gene interactions—is going to be necessary.” UCLA systems biologist **Steve Horvath, PhD**, agrees. “Often the different SNPs in a drug response pathway each have only a small effect, and would be terrible biomarkers,” he says. “It is only when they interact together that the effects add up to clinical relevance.”

However, a naive accounting for SNP interactions would be disastrous. If the SNP-by-SNP study is error prone, imagine what happens when you compare SNP pairs rather than single SNPs—the 100,000-SNP example would now entail nearly 5 billion such tests. The Bonferroni adjustment would run haywire—associations less than 99.9999999998% reliable would

be rejected. Compare triplets, quadruplets, or larger SNP cliques, and the combinatorics get hairier, and the reliability requirement even more unreasonable. Goodbye, subtle pharmacogenetic associations! If one does manage to find the causative SNPs or genes, the problems don't end—one now needs to fit them to a model to predict drug response. Standard methods for achieving this, such as logistic regression, wind up with a huge number of possible solutions. Overfit happens.

“Often the different SNPs in a drug response pathway each have only a small effect, and would be terrible biomarkers,” says Steve Horvath. “It is only when they interact together that the effects add up to clinical relevance.”

To tackle this challenge, some researchers are turning to a classic computer science formalism, the Bayes network. This approach provides an elegant way of exploring a space of gene interaction networks to find the one that best predicts disease or drug response based on the association data. “Bayes methods are almost a hundred years old, but they're designed to handle exactly this kind of problem,” says **Marco Ramoni, PhD**, who directs the Biomedical Cybernetics Laboratory at Harvard Medical School. In 2005, Ramoni and his team used this type of analysis to determine which sickle-cell anemia patients were at a high risk for stroke. About one patient in 10 has a



Stroke Risk in Sickle Cell Anemia. This Bayesian network shows how the risk of stroke among people with sickle cell anemia depends on 69 SNPs (blue nodes) in 20 genes and four clinical variables (pink nodes). Twenty-five SNPs on 11 different genes have a direct connection to the trait. By accurately capturing the interaction between SNPs, the network achieves high accuracy (98.2 percent) in predicting risk. Ramoni suggests a similar approach to predict gene-drug associations instead of the SNP-by-SNP comparison between cases and controls that many genome-wide studies employ. Reprinted by permission from MacMillan Publishers LTD, James F Meschia & V Shane Pankrat, *Defining stroke risks in sickle cell anemia*, *Nature Genetics*, 37: 435-400 (2005).

stroke before they reach 25 years old, but doctors don't know why, and typically medicate everyone. "So 90 percent of them unnecessarily get the therapy, and it's not a pleasant one." Using Bayesian analysis on genetic association data, Ramoni's group found a network of 25 SNPs and 4 clinical factors that could predict the risk with 98.2 percent accuracy. Recently, they used the same method to find a network of 37 SNPs in 20 genes that is 86 percent accurate in predicting the risk of a common type of stroke in the general population. While Bayesian analysis can easily incorporate previous biological knowledge, Ramoni for one doesn't use any. "We make the greatest effort to minimize the amount of prior information that we get," he says. "The process is entirely and happily data driven."

Some researchers are questioning the rationale for going genome-wide in the first place. They point out that such studies implicitly assume that a handful of common gene variants account for most of the differences in disease (or drug response) susceptibility. There is

increasing evidence that this "common disease/common variant" hypothesis is not valid, even for classic "single gene" disorders such as phenylketonuria (for which 531 genetic variants have been found so far). Drug response, being a complex trait, is likely to be even more diverse. "Most studies tend to ignore rare variants completely," says **Robert Elston, PhD**, a biostatistician at Case Western Reserve University. "I find that unrealistic."

Finding rare variants may need a different strategy, one that leverages prior biological information to look at specific areas of the genome at a resolution 10-fold or greater than current genome-wide studies—trading breadth for depth. Made possible by dramatic recent improvements in the speed and cost of DNA sequencing, this "deep resequencing" strategy is rapidly emerging as one of the most exciting tools for pharmacogenetics. In a sense it is a return to candidate gene approach, but with more powerful technology. "It is cheaper and more accurate than doing genome-wide studies," says Duke

University researcher **Allen Roses, MD**, who has used this method to find genetic variants linked to Alzheimer's disease. "We are getting spectacular results from it."

PUTTING IT TO WORK

Thanks to these rapid advances in pharmacogenetics, we could soon have the technology to predict drug response more accurately and for a wider range of medications. However making this technology available for routine clinical use could be a challenge.

Consider the poster child of drug response prediction, warfarin. The marriage between the gene pair (*VKORC1* and *CYP2C9*) and warfarin response is blessed with near-perfect statistical scores, clear-cut biological explanations, consensus among researchers, a simple dosing scheme, and the FDA's approval. Moreover, a 2008 report by the American Enterprise Institute-Brookings Joint Center estimates that routine genetic testing prior to warfarin therapy would prevent 85,000 serious bleeding events and 17,000 strokes, saving \$1.1 billion a year. It's a pharmacogenetic dream scenario.

Yet the concrete achievement—an increase in predictive value from 30 percent to 50 percent—doesn't impress everyone. Some feel that this is not worth the additional cost of genotyping each patient (about \$400 currently), disputing the bullish prediction in the Brookings report. For instance, a study in the *Annals of Internal Medicine* this January estimates that pharmacogenetics-based warfarin dosing would cost more than \$170,000 per quality-adjusted life year gained, or QALY, for patients with a certain type of heart condition. (Medical interventions that cost more than about \$50,000 per QALY are typically not considered cost-effective.) "Genetic Testing for Warfarin Dosing? Not Yet Ready for Prime Time" argues another paper published in *Pharmacotherapy* in May 2008. The authors point out that the theoretical benefits of genotype-based warfarin dosing have yet to be backed up by clinical data that demonstrate practical

benefit. They express concern that clinicians may place blind faith in the new dosing scheme and ignore the time-tested methods of monitoring the patients' response to the drug.

The latest blow is a rejection from Medicare. While acknowledging that “there is good evidence that persons who have these variant [CYP2C9 and VKORC1] alleles have heightened warfarin responsiveness,” the Centers for Medicare and Medicaid Services ruled on May 4, 2009, that “the evidence for improved health outcomes attributable to pharmacogenomic testing to determine warfarin responsiveness fails (as of this writing) to meet the standards of evidence to establish a basis for coverage.”

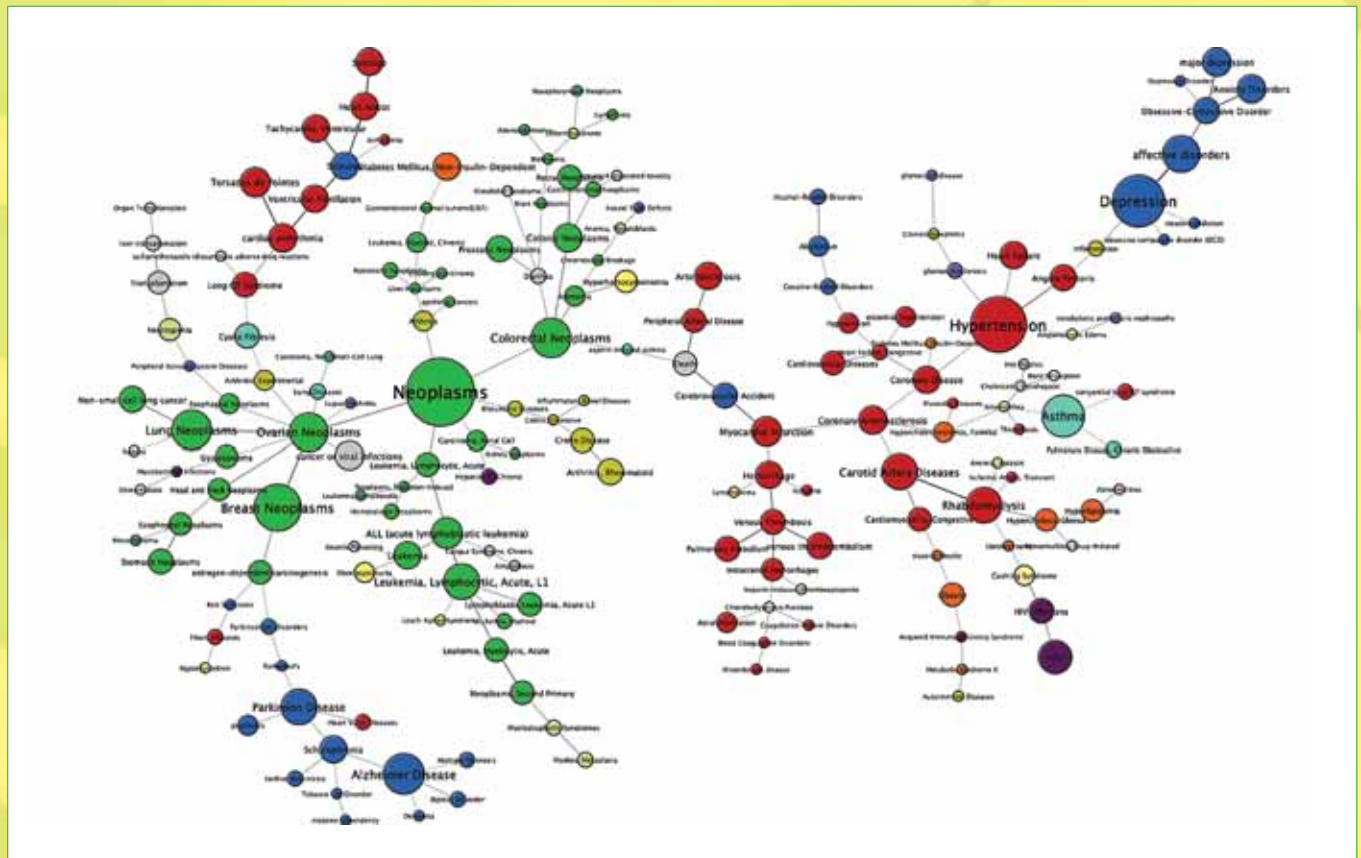
Altman finds this attitude “disappointing.” He feels that new medical advances with great potential to save lives should not be held back due to lack of clinical data. He points out that many medical practices are based on sense and evidence, not on randomized clinical

trials (RCTs). Blood transfusion, for instance, is a standard medical procedure that has never been validated in a controlled trial. “I don’t think pharmacogenomics should be held to the RCT standard unless the rest of medicine is willing to be based on that,” Altman says. As for cost, he feels that with increasing use, genotyping will get cheaper and cheaper and eventually be almost free, “which means benefit/cost = infinity.” Many other experts agree. One study on HIV patients at a UK clinic, for instance, found that genetic testing for hypersensitivity to abacavir could save about 22,000 Euro (about \$30,000) per hypersensitivity reaction that was avoided. “Pre-prescription pharmacogenetic testing for this appears to be a cost-effective use of health care resources,” the authors conclude.

While the medical community continues to debate pharmacogenetic testing, regulatory agencies seem more positive. During the past few years, the FDA has approved genetic tests for

several medications including warfarin and the cancer drug irinotecan. “We’re seeing some development towards point-of-care testing, with results available in 45 minutes or so of the patient giving a sample,” says McLeod. However, he points out, clinical medicine doesn’t yet have the tools to put this type of lab result to good use. For instance, he asks, if a test shows that a patient has the CYP2C9*3 gene variant, is that good or bad? What does it mean for drug dosing? “Most practitioners don’t have a clue,” he says. “Just because we have a test doesn’t mean we’re smarter.”

For warfarin, researchers have created a web site, <http://warfarindosing.org>, that does answer some of these questions; McLeod believes it is high time biologists created more tools like this so that we can begin to reap the benefits of pharmacogenetics. “We have all the genetic information we need at birth,” he says. “In the ideal world we’ll carry that with us and use it when needed.” □



Pharmacogenetic Tree. Many diseases share pharmacogenetic interactions with the same drugs. Here, each node represents a disease and is shown proportional in size to the number of drugs available to treat it. Each disease is connected to another with which it shares the maximum number of drug-gene interactions. This information comes from the pharmacogenetic database PharmGKB housed at Stanford University. According to PharmGKB project director Teri Klein, PhD,

the database contains information on about 650 drugs with gene-dependent responses and 1,890 genes known to modulate drug response—of which 39 are rated as Very Important Pharmacogenes, or VIP genes, because of their broad impact. Reprinted by permission from MacMillan publishers: Clinical Pharmacology & Therapeutics, Hansen, NT, et al., *Generating Genome-Scale Candidate Gene Lists for Pharmacogenomics*, (2009).

FDA-Approved Drug Warnings with Pharmacogenomic Information

Abacavir (Treats HIV-1)

FDA Warning:
"WARNING: RISK OF HYPERSENSITIVITY REACTIONS . . . Patients who carry the HLA-B*5701 allele are at high risk for experiencing a hypersensitivity reaction to abacavir. Prior to initiating therapy with abacavir, screening for the HLA-B*5701 allele is recommended; . . ."

Variants listed in drug label: HLA-B*5701

FDA Requirements: Testing recommended, not required

Azathioprine (Immunosuppressant)

FDA Warning:
"It is recommended that consideration be given to either genotype or phenotype patients for TPMT."

Variants listed in drug label:
TPMT*2, TPMT*3A, TPMT*3C

FDA Requirements:
Testing recommended, not required

Carbamazepine (Treats epilepsy and neuralgia)

FDA Warning:
"Patients with ancestry in genetically at-risk populations should be screened for the presence of HLA-B*1502 prior to initiating treatment with Tegretol. Patients testing positive for the allele should not be treated with Tegretol unless the benefit clearly outweighs the risk . . ."

Variants listed in drug label:
HLA-B*1502

FDA Requirements:
Testing recommended, not required



Dasatinib (Treats leukemia)

FDA Warning:
The FDA requires testing for Philadelphia chromosome-positive status and resistance or intolerance to prior therapy prior to initiating treatment of acute lymphoblastic leukemia (ALL) with dasatinib.

Variants listed in drug label:
BCR-ABL

FDA Requirements:
Testing required

Irinotecan (Treats cancer)

FDA Warning:
"When administered in combination with other agents, or as a single-agent, a reduction in the starting dose by at least one level of CAMPOSTAR should be considered for patients known to be homozygous for the UGT1A1*28 allele..."

Variants listed in drug label:
UGT1A1*28

FDA Requirements:
Testing recommended, not required

Imatinib (Treats cancer)

FDA Warning:
The decision of whether to treat patients with imatinib is based on the presence of genetic biomarkers, including BCR-ABL (the Philadelphia chromosome), KIT, and PDGFR gene rearrangements.

Variants listed in drug label:
BCR-ABL, KIT:D816V

FDA Requirements:
Testing required

Warfarin (Treats blood pressure)

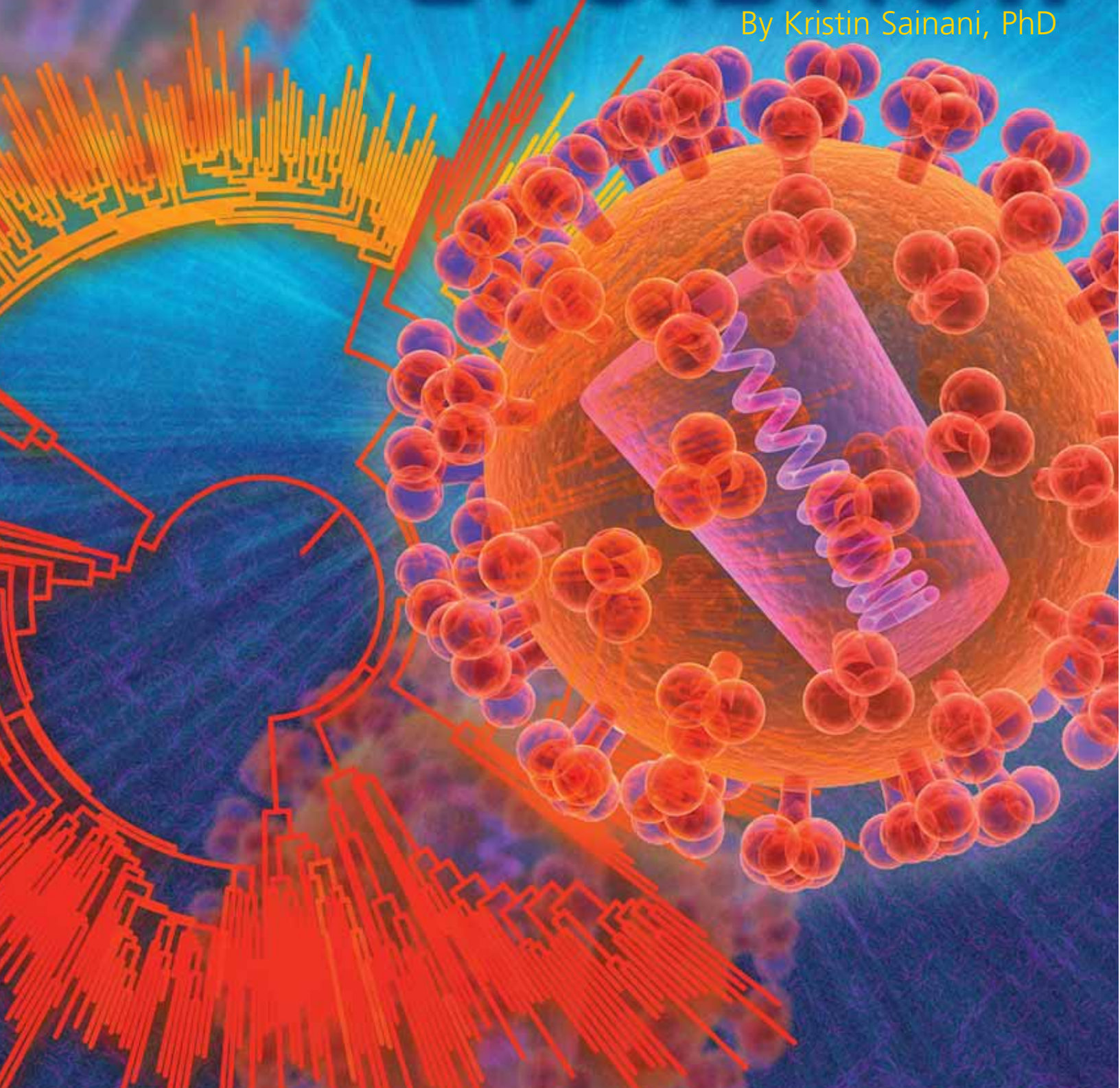
FDA Warning:
"The lower initiation doses should be considered for patients with certain genetic variations in CYP2C9 and VKORC1 enzymes . . ."

Variants listed in drug label:
VKORC1:G-1639A (rs9923231), CYP2C9*2 (rs1799853), CYP2C9*3 (rs1057910), CYP2C9*5 (rs28371686), CYP2C9*6 (rs9332131), CYP2C9*9 (rs2256871), CYP2C9*11 (rs28371685)

FDA Requirements:
Testing recommended, not required

Evolution

By Kristin Sainani, PhD



and HIV:

Using computational phylogenetics to close in on a killer

When Darwin published *On the Origin of Species* in 1859, it would be decades before HIV would jump from monkeys to humans and set off a devastating worldwide pandemic. But evolution is at the heart of HIV's biology, and Darwin would no doubt have marveled at the virus's evolutionary prowess. The virus evolves a million times faster than humans do. So fast, in fact, that a 2007 *Proceedings of the National Academy of Sciences* paper estimated that HIV and other retroviruses live just beneath the "evolutionary speed limit"—a notch faster and they would mutate themselves into oblivion.

Speedy evolution is HIV's secret weapon—allowing it to evade the immune system, resist drug treatment, and, thus far, remain impervious to vaccines. But it may also be the key to disarming the virus. The better scientists understand HIV evolution, the better they can contain the pandemic, improve treatments, design vaccines, and devise novel ways to fight the virus.

Scientists study HIV evolution at multiple scales—globally, regionally, locally, and even within a single host. Using computational methods developed for the study of phylogenetics—the study of how organisms are genetically related to one another—they can date the age of ancestral viral sequences; unravel the virus’s travels across nations and among different populations; reconstruct networks showing which individuals transmitted HIV to one another; and identify genes under selective pressure. The methods are the same whether applied at the population or host level—for example, migration patterns between tissues in a single host are resolved in the same way

“Lots of tools have been developed to do evolutionary analysis of gene sequences. And a lot of these tools have cut their teeth on HIV,” says Eddie Holmes.

as migration patterns between countries. “This is a remarkable thing. We use the same underlying statistical and computational framework to tackle really quite different biological questions,” says **Oliver Pybus, PhD**, a research fellow in the department of zoology at Oxford University.

The study of HIV evolution is not only critical to fighting the virus; it has also driven advances in the computational tools used to study evolution in general. “Lots of tools have been developed to do evolutionary analysis of gene sequences. And a lot of these tools have cut their teeth on HIV,” says **Eddie Holmes, PhD**, professor of biology at Penn State University. “It’s like the space program; it’s this kind of glit-

tery prize of modern science. Some of the smartest people have worked on HIV to try and make these techniques.”

PHYLOGENETICS IN THE ERA OF HIV

At the heart of the study of evolution is the phylogenetic tree. Scientists align a group of sequences (either of the whole genome or of a particular gene) and compare the nucleotides at every position to establish how genetically distant the strains are. These genetic distances define the phylogenetic tree: the order of sequences in the tree as well as the branch lengths.

Computationally, it’s what’s known as an “NP-hard” problem. “Which means, as your dataset gets bigger, your solution space gets ridiculous,” says **Keith A. Crandall, PhD**, professor of biology at Brigham Young University. For example, the number of possible trees that you can build out of just 50 or 60 sequences exceeds the number of particles in the universe, says **Alexei Drummond, PhD**, associate professor of computer science at the University of Auckland.

Finding clever ways to search the tree space “is where a lot of the action is in phylogenetics,” Crandall says. There’s been a lot of advancement in this area in the past decade using Bayesian statistics, he says. For example, the Bayesian Markov chain Monte Carlo (MCMC) method is implemented in the popular program BEAST (Bayesian Evolutionary Analysis Sampling Trees, co-created by Drummond, <http://beast.bio.ed.ac.uk/>). “It doesn’t attempt to find a single best answer. It tries to give you a set of trees that are representative, that are plausible, given your data and the model,” Drummond says. Generating a set of trees has an added advantage—it contains inherent information about phylogenetic uncertainty. If 95 percent of the trees contain a particular feature, you can have 95 percent confidence in this feature.

Tree reconstruction assumes an underlying evolutionary model—which specifies, for example, whether A to G and C to T substitutions occur at the same or different rates. In the past, these models were over-simplified, says **Spencer Muse, PhD**, associate professor of statistics at North

Carolina State University. They assumed that changes at one nucleotide position were independent of changes in other positions, which is unlikely to be true within a codon; they also ignored evolutionary constraints imposed by quirks of viral biology such as overlapping reading frames (where multiple genes with different starting points overlap the same sequence). But “there’s a much richer class of models available now,” Muse says. He and **Sergei Kosakovsky Pond, PhD**, developed the popular program HyPhy (Hypothesis Testing Using Phylogenies, <http://www.hyphy.org/>), which among other features, allows users to flexibly specify evolutionary models. “If you can write the model down, you can put it in the package,” says Kosakovsky Pond, who is an assistant adjunct professor of medicine at the University of California, San Diego.

Many inferences can be gleaned from evolutionary trees once they’re built, as each unique pattern of evolution leaves a unique signature in the tree. For example, within a host, HIV evolution is primarily driven by natural selection (immune or drug pressures); one lineage survives at a time, and this gives rise to a tree with a single diverging branch. In contrast, at the population level, HIV primarily evolves by random mutations (genetic drift)—and this results in dense trees with lots of branches at each time point. In the past decade, important advances have been made in the computational and statistical techniques that are used to make inferences from evolutionary trees. These are highlighted in the examples that follow.

THE GLOBAL LEVEL: DATING HIV’S ORIGINS

Scientists have used phylogenetic analysis to detail the history of the HIV pandemic—including when, where, and how it got into humans, as well as when and how it spread throughout the world. “Computational analysis has been absolutely fundamental in understanding the origins of the virus. And it’s been a real success story,” Holmes says.

HIV can be divided into two types (HIV-1 and HIV-2) and three groups within HIV-1 (M, N, and O), but HIV-

1 M is the strain that predominates in the global pandemic. This strain descended from viruses found in chimpanzees in Eastern Cameroon, and appears to have first gained a foothold in what is now the city of Kinshasa in

date such events, scientists must convert from units of genetic distance on an evolutionary tree to units of time. Initially, they did this by assuming that all lineages of the tree evolved at the same rate, but this was biologically unrealistic.

diverged into ten unique subtypes (A, B, C, D, E, F, G, H, J, and K). A 2007 paper in *PNAS* showed that HIV travelled from Africa to Haiti in 1966 (likely in a single host); and then from Haiti to the U.S. in 1969 (also in a single host); these “founder events” gave rise to the B subtype which now predominates in North America and Europe.

THE REGIONAL LEVEL: MONITORING NATIONAL EPIDEMICS

At the population level, HIV evolution is driven primarily by random mutation (genetic drift), rather than particular selective pressures. So, how HIV evolves depends on how many people it infects and where it happens to spread—and the evolutionary trees reflect these factors. Thus, scientists can work backward from the trees to unravel the virus’s demographic history in a particular region (called “coalescent theory”), as well as its migration patterns (aptly named “phylogeography”). These computational techniques can complement or even stand in for traditional epidemiology, and can help guide intervention strategies.

Coalescent theory reveals how quickly a virus was sweeping through a particular region just from the shape of the evolutionary tree. “Different rates of transmission give rise to different shaped trees, and coalescent theory is an explicit mathematical formulation of that,” Pybus says. This mathematical framework is implemented in BEAST.

“You can find signatures in the tree

“Different rates of transmission give rise to different shaped trees, and coalescent theory is an explicit mathematical formulation of that,” Pybus says.

“Computational analysis has been absolutely fundamental in understanding the origins of the virus. And it’s been a real success story,” Holmes says.

the Democratic Republic of the Congo. The two oldest known HIV sequences were unearthed there—one from a stored 1959 blood sample and another from a 1960 tissue sample, which was just discovered last year.

It’s been hotly debated as to when HIV-1 M first crossed into humans. To

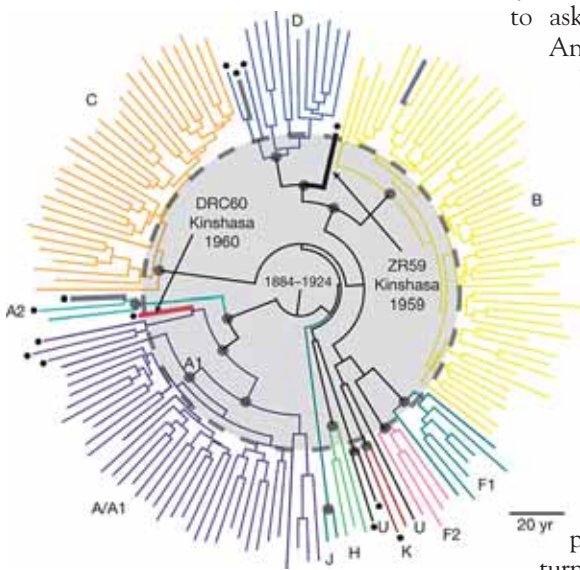
The development of “relaxed” molecular clock models—which relax that strict assumption—has improved the accuracy of dating. This model is implemented in programs such as BEAST. How accurate is it? To check this, you can pretend that you don’t know the ages of the oldest sequences and use the tree to date them, Pybus says. “We use the rest of the data to ask, exactly how old are they?

And the result comes out bang on, 1959 and 1960.”

In a 2008 *Nature* paper, scientists showed that HIV-1 M entered humans decades earlier than had previously been thought. Using BEAST software, they built an evolutionary tree from the 1959 and 1960 sequences and a sample of modern sequences, and then dated the root of the tree. Whereas earlier studies had pinpointed the date at around 1930, the new study put the estimate closer to the turn of the century (between 1884 to 1924, most likely 1908).

The *Nature* paper clearly refutes the contentious theory that HIV was introduced to humans during mass polio vaccination campaigns in Africa in the late 1950s. HIV was circulating in humans long before then. “So a lot of these evolutionary techniques pretty much put the squash on that hypothesis,” Crandall says.

From Kinshasa, HIV-1 M made its way out to the rest of the world, and



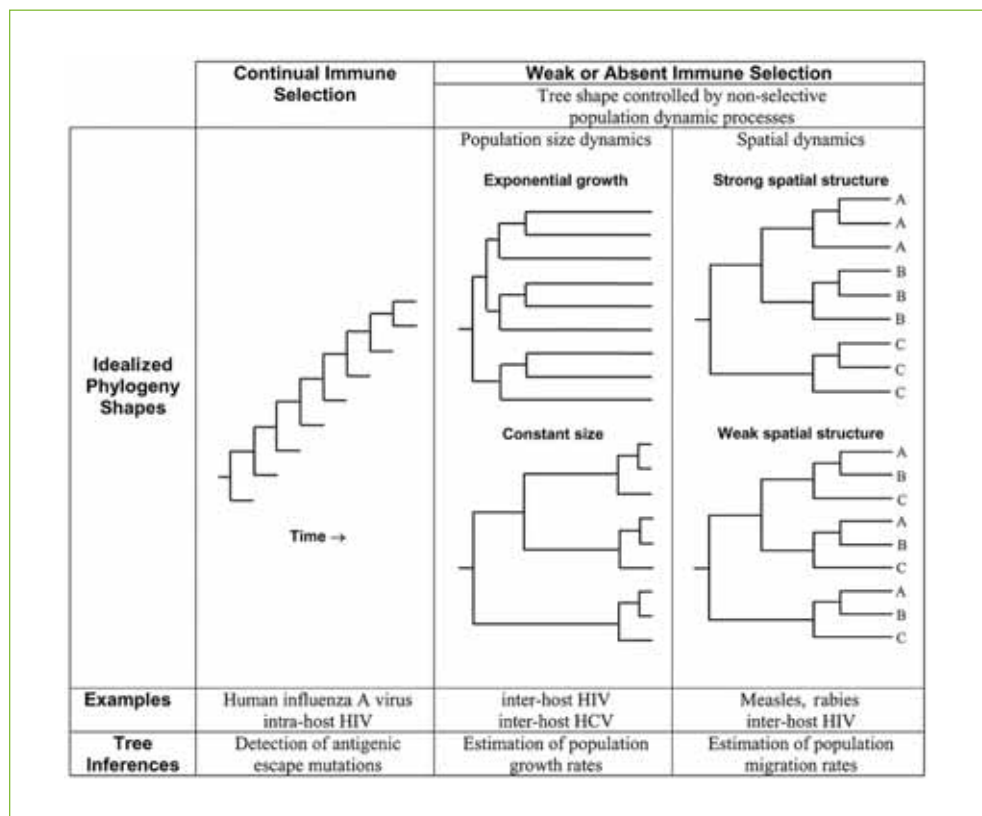
Dating the Birth of HIV. Scientists used a representative sample of modern sequences and the two oldest known sequences (from Kinshasa, 1959 and 1960) to date the origins of the virus to somewhere between 1884 and 1924, most likely 1908. Branch lengths are depicted in units of time in years. Different colors represent different subtypes of HIV. Reprinted by permission from MacMillan Publishers, Ltd., *Nature* 455: 661-664 (2 October 2008).

for not only what the size of the population was, but whether or not it was changing through time. So, exponentially growing populations give a different signature in the shape of the tree than populations with a constant size. That has been used in HIV to look at how rapidly the HIV pandemic has expanded throughout various parts of the world,” Drummond explains.

Using phylogenetics techniques including coalescent theory, Pybus and his colleagues showed that there were six independent introductions of HIV into gay men in the UK in the early 1980s; and each of those strains spread rapidly until the mid-1990s and then trailed off, corresponding to the introduction of effective combination therapy against HIV. Surveillance data from the UK showed similar overall

patterns, but missed the underlying genetic structure of the epidemic.

“Sometimes these methods are more accurate for tracking prevalence than surveillance can be—especially in places where surveillance is very limited or governments have reason to hide information,” says Sudeb Dalai,



Signatures of Evolution. When strong selective pressures (such as immune pressures) are driving evolution, less fit lineages die out and the most fit lineages survive, giving rise to a characteristic tree pattern (left panel). In the absence of strong selective pressures (such as at the population level for HIV), multiple lineages exist at once (center and right panels). Different growth rates for the virus also give rise to different tree patterns; the top tree in the center panel reflects exponential growth and the bottom tree in the center panel reflects constant growth. From Grenfell, et al., *Unifying the Epidemiological and Evolutionary Dynamics of Pathogens*. *Science* 303: 327-331 (16 January 2004). Reprinted with permission from AAAS.

“Sometimes these methods are more accurate for tracking prevalence than surveillance can be—especially in places where surveillance is very limited or governments have reason to hide information,” says Sudeb Dalai.

MS, an MD/PhD student at Stanford. “The genotypes are going to reflect the truth regardless of whether the surveillance actually does or not,” he says. Dalai’s team reconstructed the HIV epidemic in Zimbabwe—a place where surveillance data are shaky—and showed that the epidemic grew exponentially in the 1980s, correlating with political change and instability in

Zimbabwe, but reached a plateau by 1991, possibly reflecting effective intervention campaigns. When they back-calculated HIV incidence from mortality statistics, they got a similar trajectory for the epidemic.

With phylogeography, scientists blend phylogenetic information with geographical information to evaluate how the virus travels in space. One can

count migration events off an evolutionary tree as follows: if an ancestral sequence was sampled from region A and a direct descendent was sampled from region B, you can infer an A to B migration, explains Marco Salemi, PhD, assistant professor of pathology, immunology, and laboratory medicine at the University of Florida.

Using MacClade software, a popular

program for doing phylogeography, Salemi and his colleagues studied the HIV epidemic in Albania and Bulgaria, two countries where traditional epidemiologic data are lacking. Both countries—which were part of the Soviet Bloc during the Cold War—had explosive HIV epidemics during the early 1990s, likely related to the end of communism and the turmoil caused by nearby wars. The epidemics are dominated by subtype A, which is also prevalent in Russia and the Ukraine. But, surprisingly, Salemi and his team traced the source of the epidemics to Africa, not to Russia or the Ukraine. In Albania, HIV was introduced in the center of the country (in the capital) and then slowly spread to the periphery of the country, including its ports, which are just two or three hours to

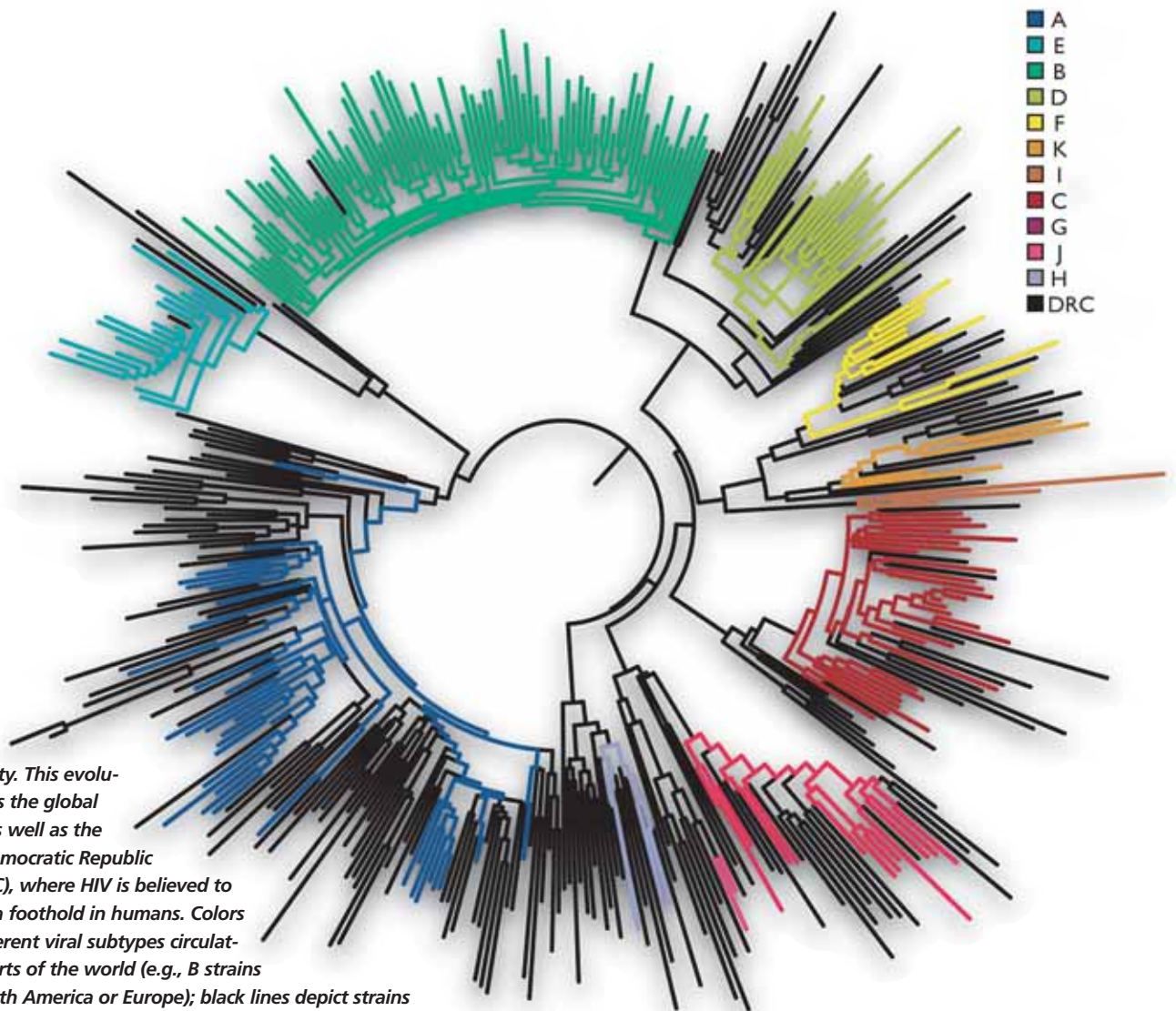
Italy by boat. “So it has the potential to modify the epidemic that is ongoing in Italy, in France, in Western Europe, and later on eventually in the U.S.,” Salemi says. In the last four years, the Italian authorities have found that 30 percent of new HIV infections are with non-B subtypes, up from only 5 percent before, he says. These findings may have implications for where to target interventions.

THE LOCAL LEVEL: LINKING INFECTED INDIVIDUALS

Phylogenetics can also identify transmission networks—people who likely infected one another—and thus may help direct local public health efforts as well as help prove guilt or innocence in criminal cases involving HIV transmission.

For example, a landmark 2009 paper in the journal *AIDS* describes an effort to use sequence data for local public health surveillance in San Diego. “We’re very interested in finding hotspots of HIV transmission—people who transmit to quite a few people, the nodes of the network so to speak,” says lead author **Davey Smith, MD**, assistant professor of medicine at the University of California, San Diego. “Then the next step would be to intervene on those individuals, or at least to figure out what characteristics are common to them.” This approach is routinely used to help control other reportable diseases such as syphilis and tuberculosis, but this is one of the first attempts to adapt it to HIV.

If you take HIV virus from two random HIV-positive people in San Diego,



Circulating Diversity. This evolutionary tree shows the global diversity of HIV, as well as the diversity in the Democratic Republic of the Congo (DRC), where HIV is believed to have first gained a foothold in humans. Colors represent the different viral subtypes circulating in different parts of the world (e.g., B strains sampled from North America or Europe); black lines depict strains from the DRC. Courtesy of: Andrew Rambaut, University of Edinburgh.

their HIV polymerase gene sequences will be about 5 percent different, says co-author Sergei Kosakovsky Pond (also of the University of California, San Diego). If they are less than 1 percent apart, then they are almost certainly linked—either through direct transmission within the pair or through transmission from a common partner. In their study of 637 individuals, they found that 25 percent were linked; the largest cluster comprised 12 individuals. Next, you can draw diagrams showing how all the sequences in the clusters connect, as well as incorporate information on “people factors”—for example, data on the patients’ sexual partners—to build a comprehensive computational model of the local transmission networks, Kosakovsky Pond says.

The UCSD researchers also track the transmission of drug resistant strains. About 20 percent of new HIV cases in San Diego are infected with a drug resistant strain, Smith says. This is a major problem because it takes three drugs to control the infection, and if a person is already resistant to just one of these drugs, they may quickly develop resistance to the other two. “So then we’ve just blown three drugs for this person,” Smith says. In a 2008 paper in the *Open AIDS* journal, Smith’s team showed that methamphetamine users

in San Diego have a high frequency of transmitted drug resistance.

Phylogenetic data have also been used as evidence in HIV criminal trials. For example, in a highly publicized case in Libya, six international medical workers were sentenced to death for allegedly infecting hundreds of children in a Libyan hospital with HIV. A month before the final appeal hearing in 2006, Pybus and his colleagues were asked to analyze viral sequences from the infected children. “That basically gave us a few weeks to actually do the analysis and write it up and get it published before this trial. We knuckled down and did the analysis in about seven days. We didn’t get much sleep,” Pybus says. Using BEAST software on a 60-processor computer cluster at Oxford University (which they tied up for the week), they built evolutionary trees and dated the most recent common ancestor of the outbreak. The paper was published in *Nature* with just days to spare before the hearing. Their findings unequivocally exonerated the medical workers: The outbreak arose from a single ancestor that predated the arrival of the medical staff to the hospital (March of 1998); and 40 percent of the diversity in the circulating strains was already present when the staff arrived. The epidemic was probably due

to a long-standing infection control problem at the hospital, Pybus says.

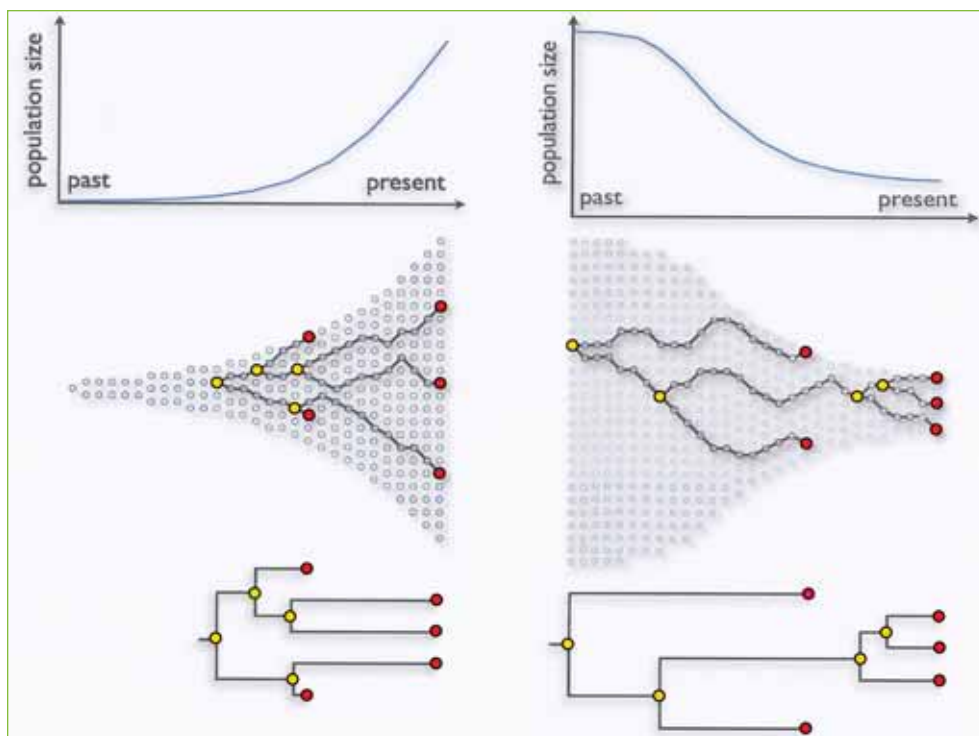
Though the paper was ignored at the 2006 trial and the medical workers were again sentenced to death, it apparently had an impact behind the scenes, Pybus says. “From what I’ve heard, our analysis did help in that it changed the tone of the diplomatic negotiations afterward.” Six months later, the medical workers were released. “That was enough politics for me for one lifetime,” Pybus says.

THE HOST LEVEL: GLIMPSING NATURAL SELECTION IN REAL-TIME

Scientists use the same techniques to study within-host evolution as they use to study population-level evolution. There’s one difference, however: natural selection plays a much bigger role in driving evolution within an individual, as the virus attempts to escape specific immune and drug pressures. “It’s beautiful natural selection, just like Darwin explained,” Crandall says. Identifying these escape mutations presents an additional challenge for modelers.

Evolutionary studies show that when HIV is transmitted to a new host, a single virus is often responsible for seeding the infection. A few weeks into the infection all the viruses have a single common ancestor that dates to the start of the infection. “HIV goes through a really severe evolutionary bottleneck when it gets transmitted from one person to another,” says **Bette Korber, PhD**, a laboratory fellow in the theoretical biology and biophysics group at the Los Alamos National Laboratory. The virus that is successfully transmitted may have unique characteristics, and could be specifically targeted by vaccines or early drug treatment.

Coalescent Theory Explained. *The shape of an evolutionary tree reflects the underlying population dynamics of the virus. The left panel illustrates the characteristic tree for exponential growth, whereas the right panel illustrates the characteristic tree for exponential decline. When the population size is small (e.g., early in the left example; later in the right example), branching events are more common. Courtesy of: Andrew Rambaut, University of Edinburgh.*

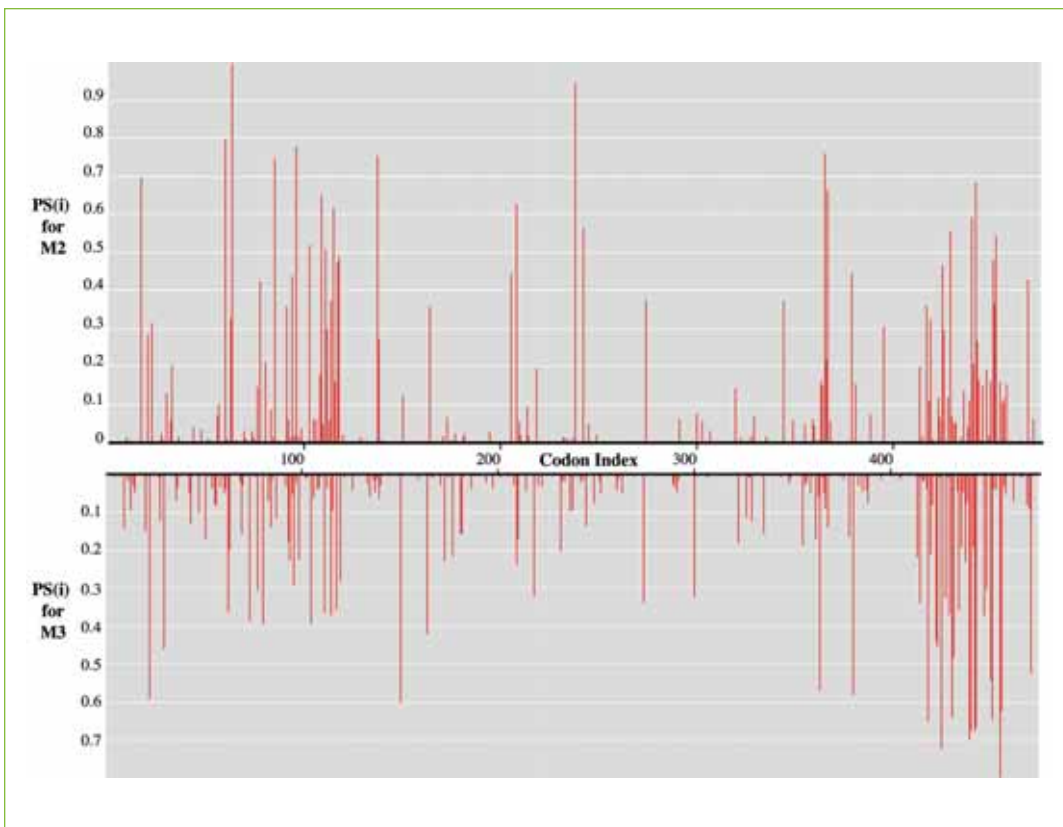


“You can actually see, if you look at the evolutionary trees, lineages keep dying out and only one survives; and that process occurs over and over, across five to ten years of infection,” Pybus says. “This arms race goes on and on; the only problem is, the immune system always seems to lose.”

After transmission, HIV undergoes weeks of rapid replication (acute infection) until the immune system finally wakes up and starts fighting back. In the absence of treatment, the virus and the immune system settle into “an evolutionary arms race,” Pybus says. “You can actually see, if you look at the evolutionary trees, lineages keep dying out and only one survives; and that process occurs over and over, across five to ten years of infection,” he says. “This arms race goes on and on; the only problem is, the immune system always seems to lose.”

The virus evolves to escape two different immune pressures—antibodies and killer T cells (cell-mediated immunity). To escape antibodies, HIV changes the shape of its envelope (surface) proteins. To escape cell-mediated immunity, HIV switches amino acids in epitopes, which are short snippets of viral protein that are displayed on the surface of HIV-infected host cells to alert killer T cells. Scientists can identify these escape mutations (using programs such as HyPhy) because genes undergoing positive selection leave a classic genetic signature—non-synonymous mutations (mutations that change the amino acid) occur more frequently than synonymous mutations (mutations that preserve the amino acid). Understanding these

Selection Detection. Scientists use computational methods to detect areas of the HIV genome that are evolving under positive selection (where nucleotide changes that alter the amino acid occur more frequently than changes that preserve the amino acid). Here, for a 500-codon stretch of the HIV genome, red bars indicate the probability that each codon is undergoing positive selection for each of two models of evolution—an over-simplified model (above) and a more biologically realistic model (below). The inferences from the two models differ considerably at several sites (for example, codons 65, 120, 230, and 470). Courtesy: Spencer Muse, North Carolina State University.



escape routes informs vaccine design (see sidebar on vaccine design).

HIV infects many different cell types in the body (not just immune system cells). As a result, the virus may become isolated in particular tissues and evolve independently from viruses in the rest of the body, giving rise to tissue-specific strains. For example, about 70 percent of patients exhibit near-complete phylogenetic segregation between sequences in the brain and in the blood, says **Satish K. Pillai, PhD**, assistant professor of medicine at the University of California, San Francisco.

In addition to the establishment of latent infection (non-productive infection of long-lived resting cells), this “compartmentalization” effect helps explain why we can control HIV infection with drugs but we can’t eradicate it, Pillai says. Drugs can push the virus to basically undetectable levels in the blood, but the virus may continue to thrive elsewhere. “The virus has a peaceful little sanctuary site that it can hide out in,” Pillai says. If scientists

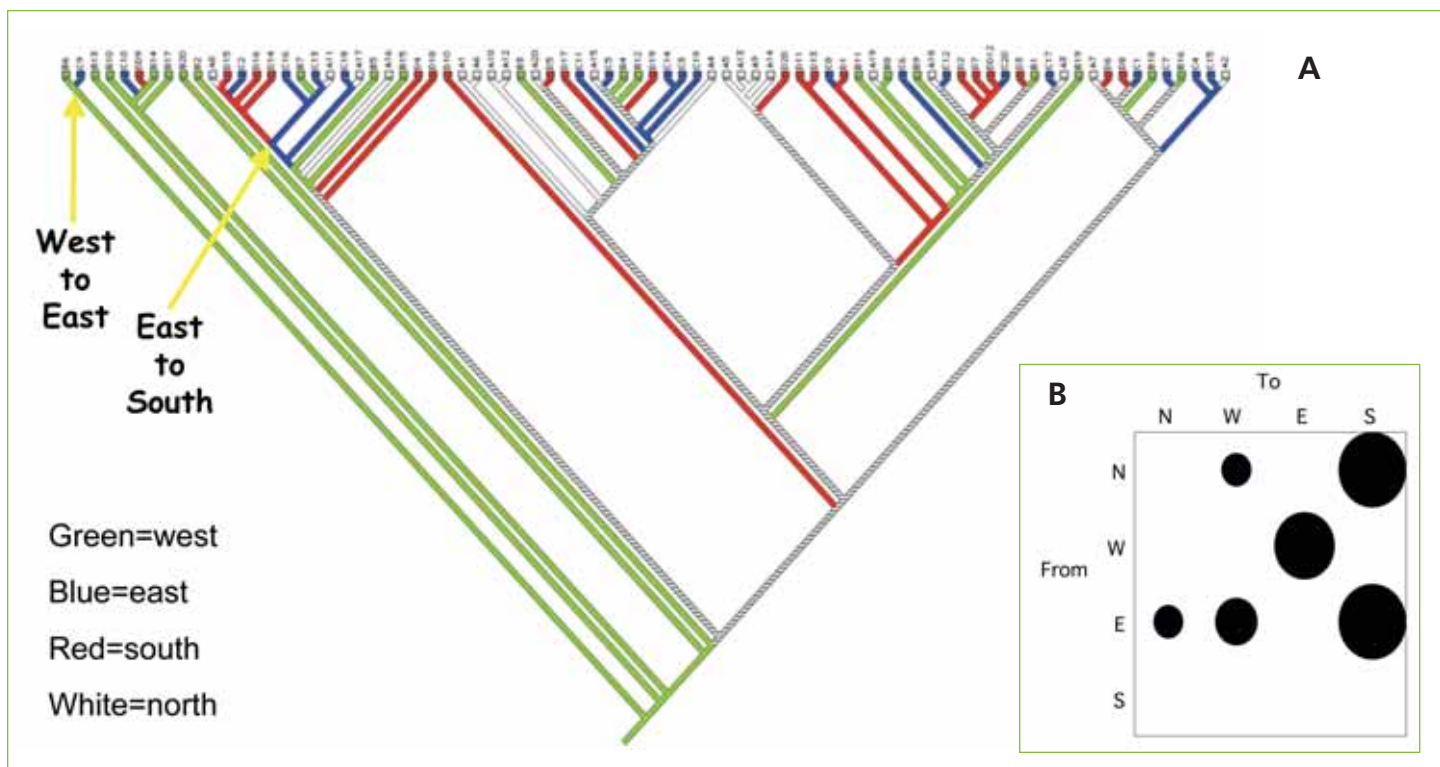
could figure out where the virus is hiding and continuing to replicate, they could design specific treatments to target those cells.

Understanding how the virus evolves in different parts of the body has other clinical implications as well. “It’s not purely an academic pursuit,” Pillai says. Evolutionary pressures differ between tissues, which may cause HIV to evolve in specific, predictable ways. In a February 2005 *Journal of Virology* paper, Pillai and his colleagues documented unique genetic signatures associated with viruses in the male genital tract. This could open the door to vaccines or microbicides that specifically target genetic variants that reside here, Pillai says.

In the brain, viral evolution may be related to HIV-associated dementia—a debilitating condition that occurs even in those on effective drug treatment, Salemi says. In a September 2005 paper in the *Journal of Virology*, Salemi and his colleagues used phylogeography to track the migration of HIV through the brain of a patient who had severe HIV-associ-

ated dementia at death. They found that the virus entered the brain through the meninges and then spread to other brain areas, including the temporal lobe. Viruses in the temporal lobe were evolving at a faster rate than elsewhere in the brain, which could be a clue to the cause of dementia, Salemi says.

Pillai’s team also studies HIV evolution in the brain, using virus sampled from the cerebrospinal fluid of living patients. They are hunting for genetic signatures of brain-associated HIV and for mutations that correlate with cognitive impairment. For example, using machine learning techniques to correlate particular mutations with scores on a cognitive deficit test, they identified a serine residue in a particular loop of the envelope protein that is “very significantly correlated with severe cognitive impairment,” Pillai says. It may be possible to design a therapeutic vaccine to steer HIV evolution away from acquiring such harmful mutations during the course of infection, Pillai says. “That would be really cool. We don’t have that technology yet, but I



Phylogeography Explained. By constructing an evolutionary tree from HIV sequences collected from different geographical regions, scientists can determine the pattern of gene flow from one region to another. In panel A, colors represent strains from different geographical regions (green=west, blue=east, red=south, white=north);

migration events can be directly counted off the tree as indicated. This information is compiled computationally and translated into a bubble plot (panel B) which quantifies the gene flow between different regions. Pictures were generated using MacClade software. Courtesy of: Marco Salemi, University of Florida

Vaccines get an evolution lesson

The hunt for an HIV vaccine has been marked by highly publicized failures and enormous disappointments. HIV presents an immense challenge to vaccine designers because the organism is so diverse. As a benchmark, consider the flu vaccine—it must be reformulated annually because flu strains diverge by about 1 to 2 percent per year in the population. In comparison, HIV mutates by about 1 percent per year within just a single person; and, across the globe, different subtypes of HIV differ by up to 35 percent. Designing a vaccine to cover all (or even a useful fraction) of this diversity has turned out to be, thus far, insurmountable.

To meet this challenge, evolutionary scientists are designing viral proteins *de novo* in the computer that attempt to summarize HIV's variation. Thus far, they can make a "consensus" sequence by determining the most common amino acid at each position from a broad sample of sequences; reconstruct HIV's ancestral sequence (which is genetically between all modern strains); or build a "center of the tree" sequence, which has the lowest total genetic distance to all other sequences in an evolutionary tree. The resulting computer-generated proteins serve as immunogens in vaccines.

"It took a while to get everyone accustomed to the idea that you might want to artificially design a protein on the computer rather than use a natural protein. People didn't know if it would fold properly, if it would be antigenic, or if it would have the same sites that

were relevant for an immune response as a natural strain. As it turns out, it does," Bette Korber says. Korber runs the Los Alamos National Laboratory HIV Sequence Database (<http://www.hiv.lanl.gov>), which provides the datasets often used for these approaches. In a 2008 paper in *PNAS*, her group showed that a consensus envelope protein stimulated three-to-four fold higher cell-mediated immune responses in monkeys than a natural envelope protein.

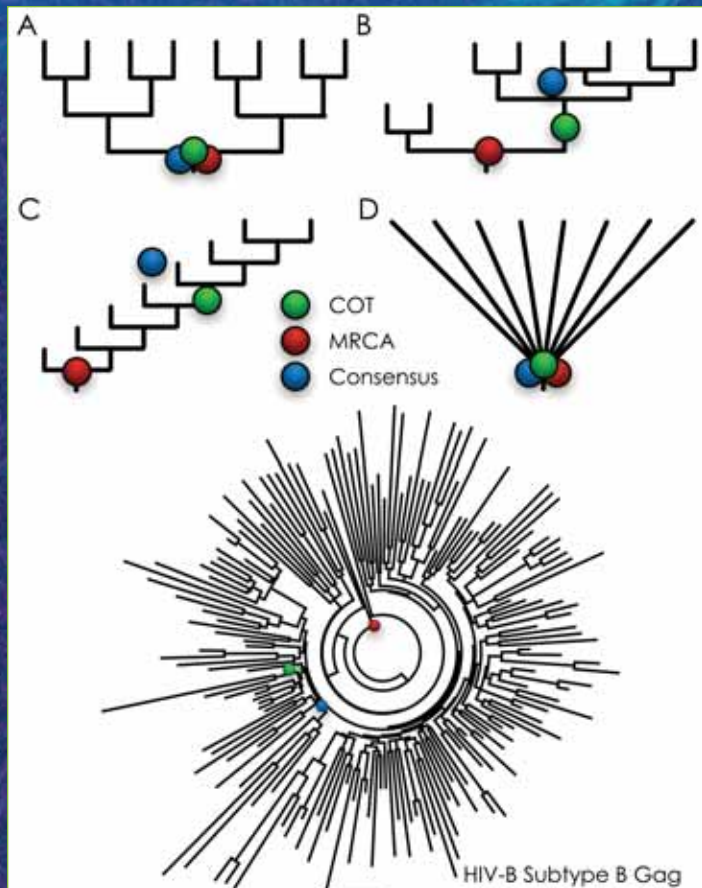
Another tactic is to design a vaccine from a diverse set of viral epitopes—the small fragments (typically nine amino acids) of virus recognized by killer T cells. Korber is piecing together the most common epitopes (cataloged in the Los Alamos database) into a small set of composite proteins. "I create sort of little Frankenstein proteins that look and feel like HIV proteins but they don't exist in nature," Korber says. So far, the proteins are showing good immunogenicity in animals, she says. "We're getting really good signals in the mice and the monkeys—which makes us delighted; they're doing a lot better than just natural proteins."

Despite Korber's success in animal models, other groups report poor results with similar strategies. Both the center of the tree and a "center of the tree plus" method (which added diverse epitopes) came up negative, says **David Nickle, PhD**, a senior research biologist at Rosetta Bioinformatics. "I've become convinced that none of those [approaches] are going to be adequate," agrees James Mullins, who collaborates with Nickle.

Mullins and Nickle believe that HIV tricks the immune system into mounting an initial response against "decoy elements"—pieces of the virus that are easily mutated. The immune system then becomes trapped by its initial response (a phenomenon called "original antigenic sin"), and ends up focusing only on these variable regions, to HIV's benefit. "My prediction is that all vaccine approaches will fail until we take into account and remove these decoy elements," Mullins says.

They are designing a "conserved elements" vaccine that contains only segments of HIV that are highly conserved—regions that don't evolve much and may not tolerate variation. "The more conserved an amino acid is in viral evolution, the more likely it is that it plays a critical role in the function of the virus," Mullins says.

Getting the immune system to attack these areas first may be the key to crippling the virus, they believe. "If you let the immune system choose what to mount an immune response to, maybe it chooses badly. So we want to redirect the immune system to mount a response to these conserved regions, even though it may be harder," Nickle says. The approach is still in the early stages of development.



Computational Vaccine Design. To tackle HIV's diversity, evolutionary scientists are designing artificial "summary" HIV proteins in the computer that may stimulate a broader immune response than natural HIV proteins. This picture compares three approaches—center of the tree (COT, green), ancestral (MRCA, red), and consensus (blue). The center of the tree approach constructs a sequence with the lowest total genetic distance to all variants in the tree; the ancestral approach reconstructs the sequence of the most recent common ancestor of the tree; and the consensus approach chooses the most common amino acid at each position from all variants in the tree. Depending on the evolutionary history, the three approaches may yield very similar or very different proteins (upper panels A-D). The bottom panel shows results from the three approaches for the HIV gag protein (lower panel). Courtesy of: David Nickle, Rosetta Bioinformatics.

think it's a possibility."

Unlike the immune system, which eventually loses out to HIV evolution, drug treatment can keep the virus in check indefinitely. But keeping ahead of drug resistance is a major undertaking. "The virus is fantastically plastic and adaptable. It can evolve resistance to all these drugs that we've developed against it. And it pretty much tends to evolve resistance to them in clinical trials—before they even get put on the market," Pybus says. HIV-1 protease (one of the nine viral proteins) has 99 amino acids, and in very heavily treated people as

ance, which researchers use to identify partial and full resistance mutations. Physicians can also enter sequence data and retrieve detailed information about their patients' mutations. "Helping clinicians interpret drug resistance tests is what's given the database the most recognition," Shafer says.

Computational scientists are working on providing new tools for physicians—for example, algorithms that predict the optimal drug regimen for a patient based on sequence data.

the sequence of mutations that may develop and the likely time frame. The resulting "mutagenic trees" are incorporated into their genotype-to-phenotype prediction algorithm

"That never ceases to amaze me: that one quarter of the amino acids can be mutated—and it still functions," Robert Shafer says.

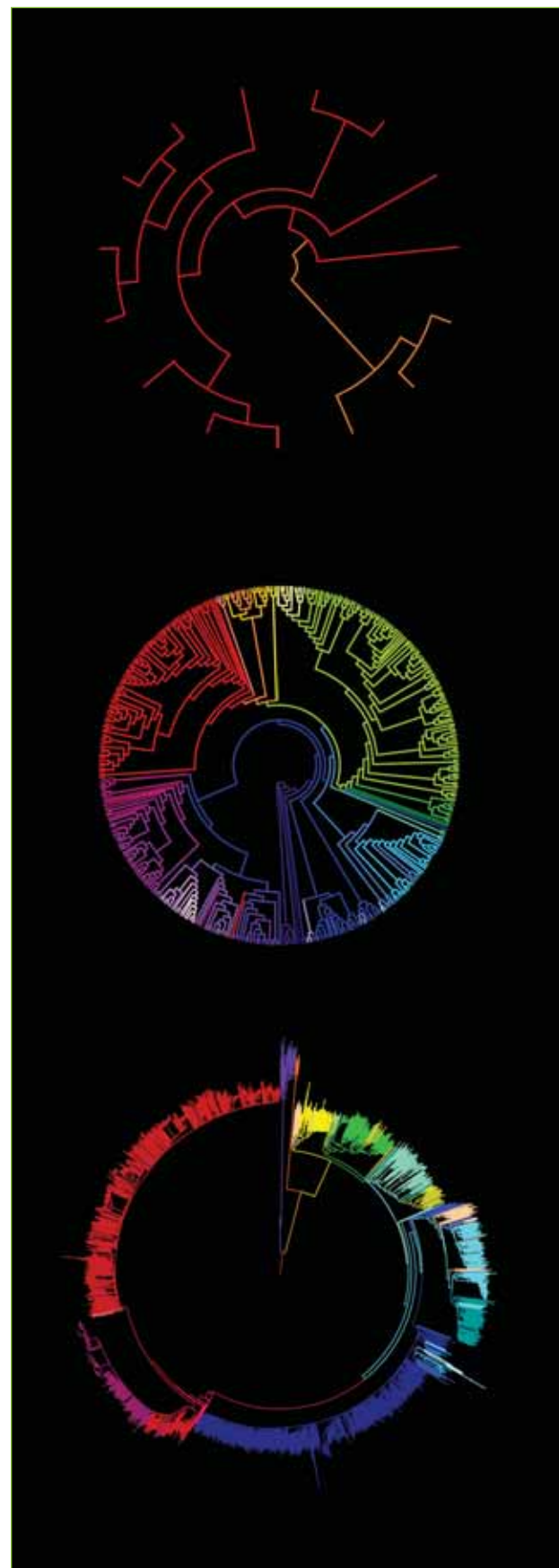
many as 25 of the positions can be mutated, says **Robert Shafer, MD**, associate professor of medicine and pathology in the division of infectious diseases at Stanford University; Shafer runs the Stanford University HIV drug resistance database (<http://hivdb.stanford.edu/>). "That never ceases to amaze me: that one quarter of the amino acids can be mutated—and it still functions," he says.

The virus mutates in predictable ways in response to particular drugs, so the challenge is to document and keep track of these mutations to help physicians, epidemiologists, and drug designers. The Stanford database contains about 100,000 viral sequences, linked to data on *in vitro* and *in vivo* drug resist-

"We infer rules for selecting optimal treatments from clinical databases; it's a big machine learning problem," says **Niko Beerenwinkel, PhD**, assistant professor of computational biology at the Swiss Federal Institute of Technology Zürich.

Beerenwinkel and his colleagues try to predict not only current drug resistance (both to single drugs and drug combinations), but also the potential for the virus to develop resistance in the future. To do this, they reconstruct the typical evolutionary paths of HIV under certain drug pressures—

Snapshots of a Pandemic. Evolutionary trees were created from all the HIV whole genome sequences available in 1993 (n=15, at top), 2003 (n=397, at center), and 2009 (n=1885, at bottom) from the Los Alamos HIV database (<http://www.hiv.lanl.gov>) and GenBank. Different colors depict different subtypes and recombinants of HIV—which is the most sequenced organism ever. This picture shows the increasing availability of whole genome sequences as well as HIV's increasing diversity. Courtesy of: Keith Crandall and Matthew Bendall, Brigham Young University



(<http://www.geno2pheno.org>). If the virus only needs one mutation to escape drug regimen A and five to escape B, we can predict that B will suppress the virus for a longer period, Beerenwinkel explains.

EVOLVING PHYLOGENETICS

Despite the progress made in HIV evolution and phylogenetics, some challenges remain. One issue is how to deal with recombination—where two different viral strains infect the same cell and exchange genetic material, so-called “viral sex.” HIV actually evolves more rapidly by recombination than by point mutation, says **James Mullins, PhD**, professor of microbiology and of medicine at the University of Washington. But most tree-building programs don’t account for recombination—which can lead to mistakes (especially when dealing with whole genome sequences) since a recombinant sequence actually has two separate lineages. “No one’s attacked that problem really effectively in phylogenetics; I would say that’s an understatement,” Mullins says.

Several programs identify recombinant sequences and remove them prior to tree building. But what’s really needed is a program that can detect recombination, figure out the breakpoints, and incorporate that history into the tree. This is a difficult task, because it greatly increases the number of possible trees. “If the phylogeny problem is NP-hard, one could say the recombination problem is NP-harder,” Pybus quips.

Predicting Evolution. HIV develops resistance mutations to particular drugs in predictable ways. By linking genotypic and phenotypic data from large HIV databases, researchers can tease out these mutation pathways (called “mutagenic trees”). This picture illustrates the typical amino acid changes that HIV may undergo to develop resistance to the drug AZT. In these panels, the numbers shown along the arrows indicate (a) the probability of a particular mutation and (b) the average number of days it takes for each such mutation to occur. These mutagenic trees are incorporated into algorithms that predict optimal drug combinations for patients based on their viral genotypes (www.geno2pheno.org). Courtesy of: Niko Beerenwinkel, Swiss Federal Institute of Technology Zürich.

Nobody has solved the problem adequately yet, but BEAST developers are working on it, Drummond says.

Another challenge is the rise of next generation sequencing platforms, such as 454 pyrosequencing, which increase the speed of sequencing by orders of magnitude. Besides providing a wealth of data for building evolutionary trees, the technology allows “deep sequencing,” the ability to detect viral variants within a single patient that are present at very low levels—including low-level drug resistant variants—rather than just the dominant clones. This information may improve our ability to predict drug failure.

“But there’s a lag between the sequencing technology and our methodology that processes these sequences,” Kosakovsky Pond says. Current phylogenetics programs can handle hundreds of sequences, but next generation sequencing may provide thousands or tens of thousands of complete HIV genomes at once. “It’s going to be a bit of a tidal wave for those of us who do the analysis,” Pybus says.

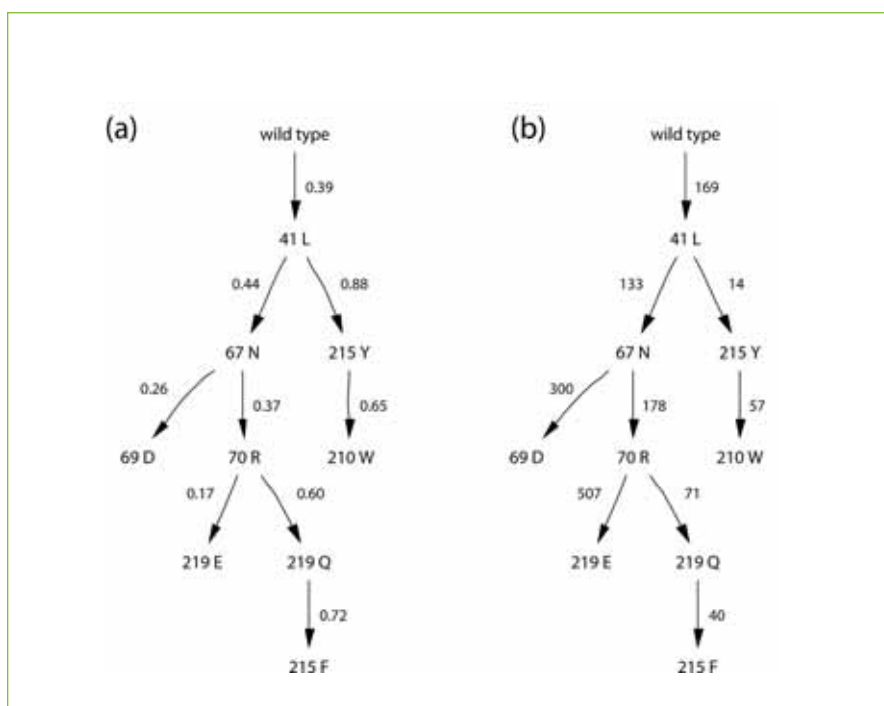
Besides the sheer volume of data, the technologies present new bioinformatics problems, says **Allen Rodrigo, PhD**, professor of computational biology

and bioinformatics and director of the Bioinformatics Institute at the University of Auckland in New Zealand. They yield short reading lengths, which have to be assembled; and they also have high error rates—which means it can be difficult to differentiate technical errors from real mutations in HIV. “This is going to

“If the phylogeny problem is NP-hard, one could say the recombination problem is NP-harder,” Pybus says.

open up a whole new set of computational challenges that we’re just starting to look at,” Rodrigo says.

Researchers in HIV are once again at the forefront, driving forward these advancements in the study of evolution. Hopefully, these tools will, in turn, help drive HIV into extinction. □



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

Simplifying the Science and Art of Molecular Dynamics

Using molecular dynamics (MD) software, scientists can simulate molecular movement to study biological phenomena that currently cannot be observed experimentally.

But the value of MD software can be outweighed by its steep learning curve. For example, to run GROMACS, a popular MD software package with a 300-page user manual, users typically deal with four different programs—each with at least 15 different options that are run via a com-

putational biologist and programmer for Simbios.

OpenMM Zephyr is built on a version of GROMACS that incorporates the OpenMM libraries (see Summer 2008 issue), enabling it to run on graphical processing units and making it possible to run larger and/or longer simulations than can be run on CPUs. But Zephyr only exposes a tiny fraction of the numerous GROMACS parameters—often via pull-down menus—narrowing the choices so that a user can quickly get a simulation running. Zephyr is also integrated with the widely used molecular dynamics viewer VMD; simply selecting the checkbox for VMD launches it and displays the simulation as it is running.

Bruns points out that although OpenMM Zephyr is easy to use, it is not a black box. “Complicated things are done by Zephyr and we don’t want the user to just be a monkey turning the crank,” says Bruns. OpenMM Zephyr is therefore designed around three guiding principles: discoverability—exposing information so that the user can drill deeper to understand what is going on; feedback—communicating to the user when things go wrong, as well as when things go right; and expert convention—building in the parameter choices, workflow, and other best practices of expert MD users to guide an individual.

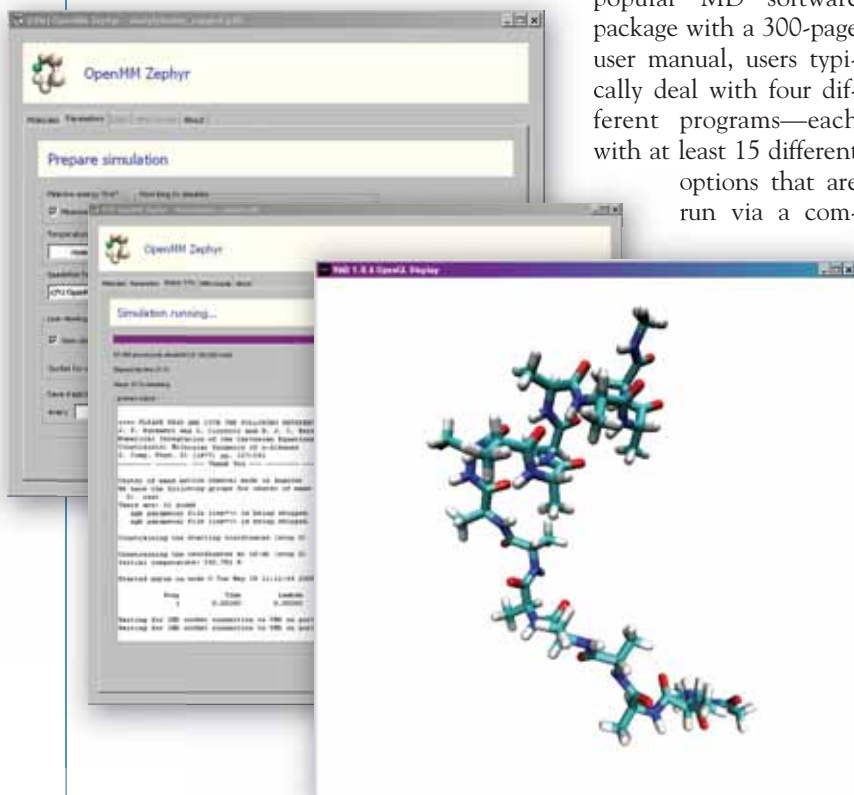
The approach seems to be working. “From a cost-benefit analysis, Zephyr is very attractive,” says **KC Huang, PhD**, an assistant professor in bioengineering at Stanford University. For people like him who aren’t well-versed in using MD on biological molecules, he says, OpenMM Zephyr is a great way to explore whether it will be helpful or not, without making a huge initial time investment in learning.

For **Victor Pinks, PhD**, the science chairman and technology director at Marmion Academy, a high school in Aurora, Illinois, Zephyr’s accessibility and ease-of-use were compelling reasons to switch to it. He and his student **Timothy Hagerty** had been designing laboratory exercises using another MD program. “When we saw OpenMM Zephyr, we thought, ‘That’s exactly what we want.’ Now we can go right into the science,” says Pinks.

Klaus Schulten, PhD, professor of physics at the University of Illinois at Urbana-Champaign, says, “Molecular dynamics is becoming more and more of a tool for use by biomedical scientists, including both clinical and experimental investigators. So it is really wonderful that Simbios is simplifying the science and art of it.” □

mand line interface. The first hurdle though is getting through the multi-step installation process. And other MD packages are just as complicated. Installing and learning to use any of them can be daunting. To address that problem, Simbios has just released a new version of OpenMM Zephyr, an application to simplify the MD process. It has a one-click installation process and provides a graphical user interface that guides an individual through the workflow for setting up, running, and viewing an MD simulation.

“With OpenMM Zephyr, we’ve created a software that is a good educational tool for learning how to use molecular dynamics. It also makes it more comfortable for even an expert user to get things done quickly,” says **Christopher Bruns, PhD**,



DETAILS: OpenMM Zephyr can be freely downloaded from <http://simtk.org/home/zephyr>. It currently runs on Windows and Macintosh platforms. The Linux version is expected to be released later this year.



BY ALBERT GOLDFAIN, PhD



Canonicity and Disease Ontologies

Ontologies provide biomedical researchers with an inventory of the universal features of reality across organisms, biomedical disciplines, and levels of granularity. In capturing what is universal, there is often a need to refer to what is prototypical, representative, true-by-default, and statistically expected. In other words, we often need a reference for *canonical* entities and relationships. Clinically speaking, canonical facts for human beings

and other forms of departure from the norm. Treatment decisions are made on the basis of how (and to what extent) a patient deviates from a canonical life plan. Thus, a reasoner needs a coherent representation for the canonical that is compatible with biomedical ontologies.

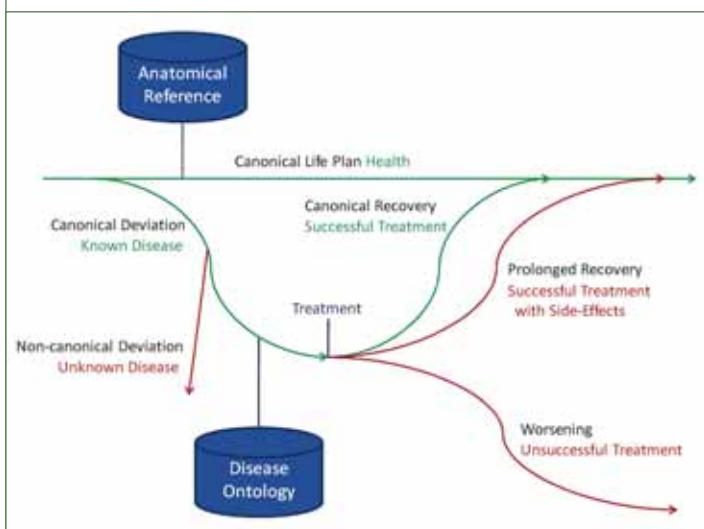
If ontologies are to be used in service of *computational* reasoning (e.g., in clinical decision support), the notion of canonicity must be formalized and given a precise semantics. Some work has been

done along these lines for anatomical ontologies [1], in which the range of quantified variables in logical formulae is restricted to anatomical entities. Such ontologies capture information about properly ordered parts undergoing proper functions, manifesting healthy dispositions, and playing proper roles. This approach can also be extended to disease ontologies, which capture disordered parts along with

references computationally amounts to using a similarity metric in a space of clinical measurements. Ontologies must include definitions for what these measurements are measurements of.

If I break my arm in a way that has never been clinically observed; or after treatment, my arm fracture worsens, does not heal in the expected amount of time, or does not heal at all, these would be *non-canonical* deviations from health, treatment, and recovery. Such instances can then be compared with a canonical reference regarding the clinical expectation. Non-canonical deviations, then, can serve as important signposts of unknown diseases or disorders, unsuccessful treatments, or erroneous outlier data.

Without a formal description of the canonical, a computational reasoner can only compare instance data to more instance data. This may be fine if the instances are sufficiently large in number and are drawn from a sufficiently representative population, but this is often an idealized situation. A reasoner who is given a dataset of human arms fractured playing football will not compute a sufficiently general prototype of a fractured arm because the instances are more likely to be the same sorts of high impact fractures. Such a reasoner would ignore the existing general knowledge about different types of arm fractures and the relations between them. Canonical entities serve as a compact summarization of general knowledge. They enhance ontologies by providing a baseline from which a deviation can be logically described, quantified, and measured.



include: body temperature is 98.6 degrees Fahrenheit, pregnancy lasts nine months, and adults have 32 teeth. These are context-independent clinical expectations that are the result of broad scientific consensus. Ontologically speaking, there are no instances of canonical humans with canonical parts functioning in canonical ways. Nevertheless, the canonical representation serves as a useful and economical starting point for describing and understanding disease

their associated malfunctioning, and the dispositions for disease that they give rise to. Clinical treatment is only possible because there are known patterns in how disorders become the physical basis for disease—essentially, *canonical* deviations from health. It is this sort of information that is covered by ontologies of disease and disorder. So, for example, my right arm currently resembles the human right arm as described in an anatomical reference. If I break my arm tomorrow, it may resemble a canonically fractured arm (if the fracture is of a known type). We cannot say that my arm is an instance of either the canonical human arm today or the canonical fractured arm tomorrow, but these reference points are essential for reasoning about my arm (as an instance of the universal arm) and its change of state. Comparing my arm to either of these

DETAILS

Albert Goldfain, PhD, is a researcher for Blue Highway, LLC. [www.blue-highway.com]. He is currently working as a postdoctoral associate on the Infectious Disease Ontology [www.infectiousdiseaseontology.org].

REFERENCES

- [1] Neuhaus, Fabian and Barry Smith (2007), “Modeling Principles and Methodologies—Relations in Anatomical Ontologies”, in Albert Burger, Duncan Davidson, and Richard Baldock (Eds) *Anatomy Ontologies for Bioinformatics: Principles and Practice* (Springer: New York), p. 289-306. □

Biomedical Computation Review

SIBIOS AN NIH NATIONAL CENTER FOR BIOMEDICAL COMPUTING

Stanford University

318 Campus Drive

Clark Center Room S231

Stanford, CA 94305-5444

seeing science

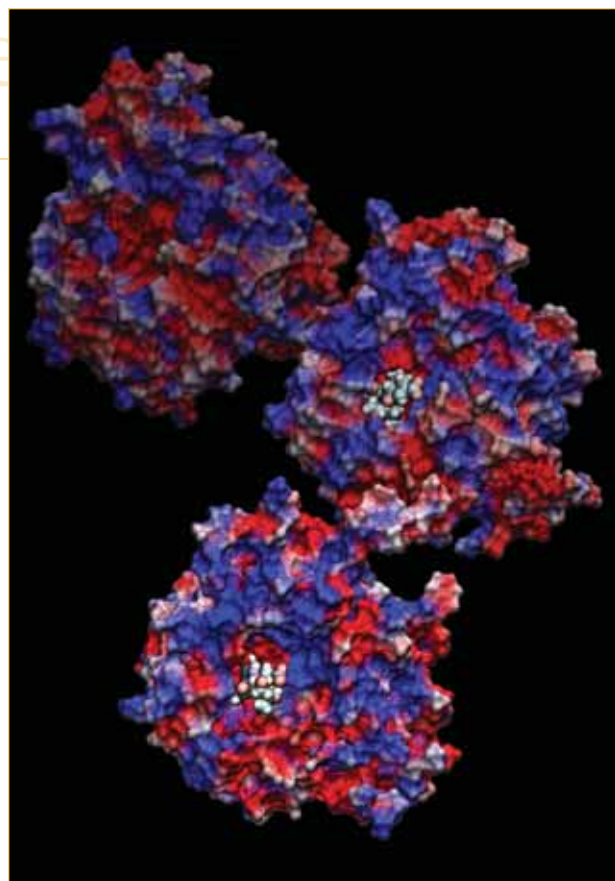
SeeingScience

BY KATHARINE MILLER

Swine Dynamics

The antiviral drugs Tamiflu and Relenza target a key flu protein—neuraminidase—preventing it from doing its job of releasing virus particles from infected cells into the body. The type of neuraminidase protein (N1) in the 1918 Spanish flu (H1N1), the 2003 avian flu (H5N1) and the 2009 swine flu (A/H1N1) is responsive to these drugs. But a few mutations in a key part of the N1 protein can render the drugs useless. To better understand why, a team of researchers at the University of Utah and the University of Illinois, Urbana-Champaign used molecular dynamics simulations to observe how Tamiflu and Relenza bind to the N1 protein of each of these three viruses. The work has not yet been published.

Because the swine flu virus is so new, the researchers had to first create a model of that flu's N1 protein. Next, they simulated the antiviral drugs binding to the neuraminidases from all three viruses. The stable binding observed (in the simulations) reflected the fact that these drugs are effective for the wild type strains. The researchers were also able to observe which specific atomic interactions within and near the binding pocket were most important—and hypothesized which mutations in some of these areas might cause drug resistance. Such hypotheses will form the basis for further simulations as well as for experimental work on antiviral drug resistance in flu. "Our observations help to establish a baseline set of drug-protein interactions that one can compare to the case of drug resistant mutants," says **Eric H. Lee, PhD**, a postdoctoral scientist at the University of Illinois, Urbana-Champaign. The group is already running new simulations involving such mutations. □



*These three neuraminidase protein structures show (from bottom to top): H1N1 swine flu with Relenza bound, H5N1 avian flu with Tamiflu bound, and H1N1 of the Spanish flu without a drug bound. The simulations compare the binding of FDA-approved drugs for the different flu strains and also characterize neuraminidase mutants of flu strains that developed Tamiflu resistance. Image courtesy of **Thanh Truong, PhD**, professor in the department of chemistry at the University of Utah, **Ly Le**, a graduate student in his lab, **Klaus Schulten, PhD**, professor of physics at the University of Illinois, Urbana-Champaign, and **Eric H. Lee, PhD**, a postdoctoral scientist in Schulten's lab.*