

DIVERSE DISCIPLINES, ONE COMMUNITY

Biomedical Computation

Published by Simbios, an NIH National Center for Biomedical Computing

REVIEW

Computing THE History of Life

Fall 2013

PLUS:
THE BIOLOGY OF
INTERACTING THINGS:
The Intuitive Power of
Agent-Based Models

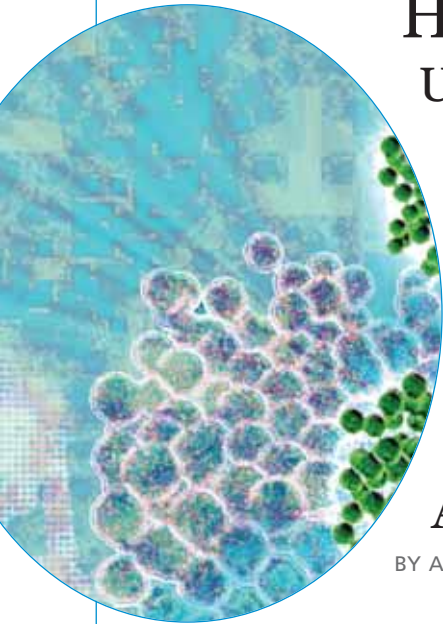
ALSO THIS ISSUE: SEE PAGES 1 AND 2 FOR PRIZE AND AWARD ANNOUNCEMENTS

10 Computing the History of Life: Using New Data and New Models to Tackle Old Puzzles

BY KRISTIN SAINANI

20 The Biology of Interacting Things: The Intuitive Power of Agent-Based Models

BY ALEXANDER GELFAND



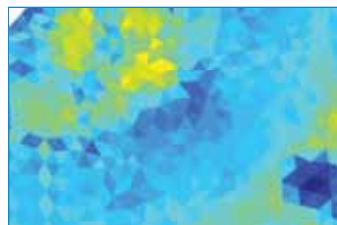
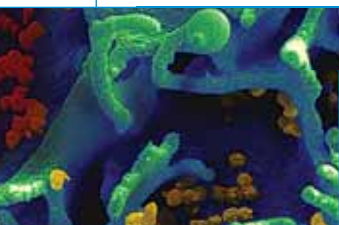
DEPARTMENTS

- 1 GUEST EDITORIAL | THE MISSING LINK: A SUSTAINABILITY PLAN**
BY GWEN JACOBS, PhD
- 2 SIMBIOS NEWS | A BALANCED APPROACH TO DESIGNING FORCE FIELDS**
BY KATHARINE MILLER
- 3 CURRICULUM AND BIG DATA: REVAMPING TO OPEN THE BOTTLENECK**
BY KATHARINE MILLER
- 5 VACCINES BY THE NUMBERS: COMPUTATIONAL APPROACHES TO DESIGN VACCINES FASTER** BY AMBER DANCE
- 7 CANCER'S CRYSTAL BALL: PERSONALIZED TUMOR MODELS TO GUIDE TREATMENT** BY SARAH C.P. WILLIAMS
- 28 UNDER THE HOOD | AGENT-BASED VIRTUAL-TISSUE SIMULATIONS**
BY MACIEJ SWAT AND JAMES A. GLAZIER
- 30 SEEING SCIENCE | DIGGING DEEP INTO THE TREE OF LIFE**
BY KATHARINE MILLER

Cover and Page 10 Art: Created by Rachel Jones of Wink Design Studio using: in-house images and phylogenetic tree images. See full credit in caption, page 10.

Page 1 Art: Created by Rachel Jones of Wink Design Studio using: sign image, © 72soul | Dreamstime.com.

Page 20 Art: Created by Rachel Jones of Wink Design Studio using modeling images. See article captions for full credits.



Fall 2013

Volume 9, Issue 3

ISSN 1557-3192

Executive Editor Russ Altman, MD, PhD

Advisory Editor David Paik, PhD

Associate Editor Joy Ku, PhD

Managing Editor Katharine Miller

Science Writers

Amber Dance, Alexander Gelfand, Katharine Miller, Kristin Sainani, Sarah C.P. Williams

Community Contributors

Gwen Jacobs, PhD, Maciej Swat, James A. Glazier

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 S221
Stanford, CA 94305-5444

Biomedical Computation Review is published 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)
Grace Peng, PhD (NIBIB)
Jim Gnadt, PhD (NINDS)
Peter Highnam, PhD (NCRR)
Jennie Larkin, PhD (NHLBI)
Jerry Li, MD, PhD (NIGMS)
Nancy Shinowara, PhD (NICHD)
David Thomassen, PhD (DOE)
Janna Wehrle, PhD (NIGMS)
Jane Ye, PhD (NLM)

BY GWEN JACOBS, PhD, DIRECTOR OF CYBERINFRASTRUCTURE FOR THE UNIVERSITY OF HAWAII SYSTEM AND PROFESSOR OF NEUROSCIENCE

The Missing Link: A Sustainability Plan

As the NIH National Centers for Biomedical Computing (NCBC) program enters its final years of support, there is an opportunity to reflect on how this program has made a lasting impact on the research community. The NCBCs were launched in response to the BISTI report¹ which called on NIH to make significant investments to train a new community of biomedical computational scientists; develop new methods in computational research; support efforts to make data available and useable; and foster a scalable national computer infrastructure to support biomedical research.

The NCBCs have succeeded brilliantly in their mission by focusing on important biomedical research questions, publishing volumes of high impact research articles, training scores of new computational scientists, and producing professional quality open-source software and resources that many research and clinical groups now depend on as an integral part of their research framework. However, the community-based research infrastructure developed by the NCBCs is now in jeopardy as the program winds down.

Now, with a sense of *déjà vu*, the cycle is starting anew, led by a report by the Data and Informatics Working Group² with a set of recommendations that parallel those of the BISTI report, but with a stronger focus on data management, integration and sharing designed to address the huge challenge of making better use of the deluge of data generated by biomedical researchers.

In response to the report, NIH has launched the Big Data to Knowledge (BD2K) Initiative and set a high bar for the long-term impact of this program: “A BD2K Center application is expected to propose the development of specific and substantive “products”—e.g., approaches, methods, software, tools, and other resources to analyze data—and then distribute the products to the user community to dramatically enhance the research community’s capabilities for using Big Data in biomedical research.”³

This bold vision is a tall order and lessons learned from the NCBC program indicate that new centers will face a significant challenge to ramp up and deliver, given the shortened time frame and reduced funding level of this new program.

The NCBCs have already provided innovative solutions to many Big Data challenges, as will the BD2K Centers in the future. What’s missing from both of these programs is a mechanism to address the major challenge of sustaining the

infrastructure, software products and services necessary to support biomedical research communities.

There are no easy answers to the question of who will pay for the support of public access to research data, software and the infrastructure to support it. Fran Berman and



Vint Cerf laid out several possibilities recently including public-private investments, government support for some community data collections and new economic models such as a small fee for downloads of data or software.⁴

The NIH has the opportunity to take on this challenge with the BD2K program by tasking the centers to develop sustainability plans for data collections and/or software in the first year of the project. Collectively the consortium could evaluate the feasibility of these plans with feedback from the community over time. If successful, the BD2K program could develop a realistic sustainability model for research resources that would be a huge benefit to the biomedical community. □

1. The BISTI report: http://www.bisti.nih.gov/library/june_1999_Rpt.asp

2. The Data and Informatics Working Group Report: <http://acd.od.nih.gov/Data%20and%20Informatics%20Working%20Group%20Report.pdf>

3. Centers of Excellence in Big Data Computing FOA: <http://grants.nih.gov/grants/guide/ifa-files/RFA-HG-13-009.html>

4. Berman, F and V Cerf (2013) Who will pay for public access to research data? *Science* vol. 341 pp 616 - 617

BY KATHARINE MILLER

A Nobel for One of Our Own

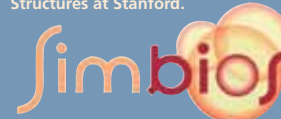
Michael Levitt, professor of structural biology at Stanford Univer-

sity, has received the 2013 Nobel Prize in Chemistry in recognition of his pioneering work in computational biology.

“This is wonderful for computational biology and a victory for physics-based simulation,” said **Scott Delp, PhD**, professor of bioengineering, mechanical engineering and orthopaedic surgery at

Stanford and co-PI for Simbios, the National Center for Physics Based Simulation of Biological Structures, of which Levitt is a part.

To read more about Levitt’s life and accomplishments, go to <http://news.stanford.edu/news/2013/october/levitt-nobel-chemistry-100913.html> □



BY KATHARINE MILLER

A Balanced Approach to Designing Force Fields

When simulating the movements of large molecules on a computer, researchers typically rely on an approximation of the force fields at play. That's because a truly correct simulation of those forces would require complex quantum mechanics calculations that would take years to simulate.

Many researchers use off-the-shelf force fields without knowing whether they are optimal for a specific situation. To address this issue, **Lee-Ping Wang, PhD**, a Simbios postdoc, created ForceBalance, a software program that makes it easier for researchers to efficiently develop and optimize their own force fields. "ForceBalance lets researchers pursue a scientific problem with greater confidence that their force fields are accurate," Wang says.

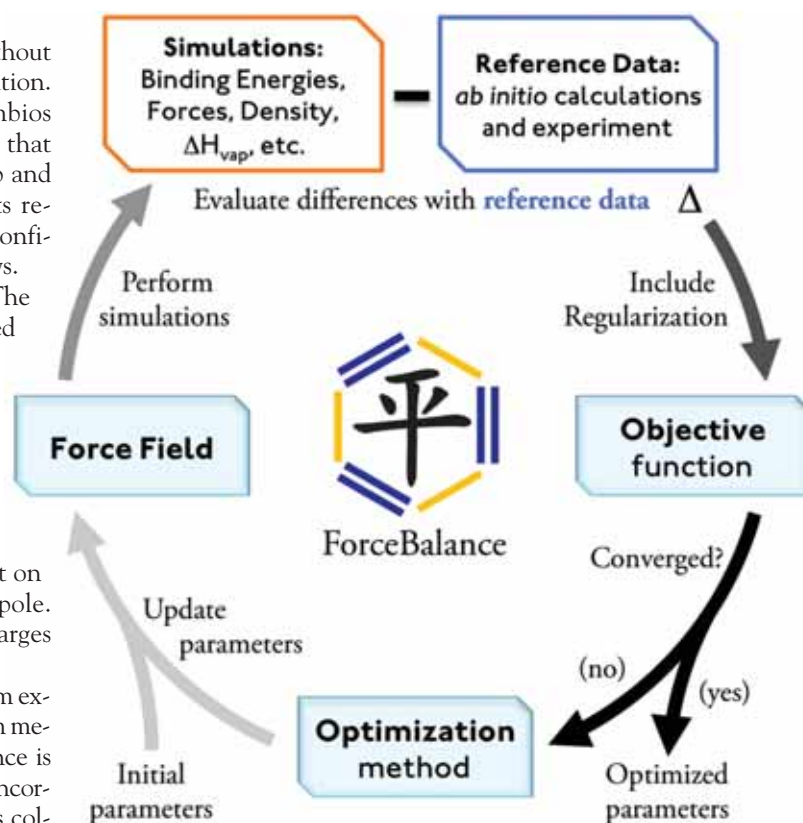
Force field design depends on three ingredients: The functional form of the forces—essentially a simplified description of the cloud of electrons; data representing the pieces of reality the force field should reproduce; and a method for optimizing parameters.

ForceBalance allows researchers a good deal of flexibility with respect to each of these three ingredients. For the functional form, for example, researchers have the freedom to represent the electrostatics as one positive and one negative point on each atom (a monopole), or as a dipole or multipole. Among other things, they can also model the charges moving around (induction or polarization).

In addition, the data for ForceBalance can come from experiments, theory (based on a small number of quantum mechanics calculations) or both combined. "ForceBalance is the only force-field software that can simultaneously incorporate multiple types of data," Wang says. He and his colleagues used this capability to generate an extremely accurate model of water molecules that was published in November 2012 in the *Journal of Chemical Theory and Computation*.

ForceBalance users can also choose among three differ-

ent optimization techniques: grid scan (which tests all the possible parameters); simulated annealing (which tries random jumps, honing in on the best parameters); and the Newton-Raphson method (a derivative-based approach).



ForceBalance fully automates the force field optimization process.

Despite offering three methods, Wang says, it's the Newton-Raphson approach that routinely finds the best solution in the least amount of time (taking only about 10 iterations to converge on the optimal parameters).

"With ForceBalance," says **Pengyu Ren, PhD**, associate professor of biomedical engineering at the University of Texas at Austin, who collaborates with Wang, "you can quickly evaluate or compare a number of different physical models to observables and determine what physics are more important to what." Researchers can in turn test more models for accuracy. "ForceBalance will accelerate improvements in models and make things more accurate," Ren says. □

DETAILS

ForceBalance is described in Lee-Ping Wang, Teresa Head-Gordon, Jay Ponder, Pengyu Ren, John Chodera, Peter Eastman, Todd J. Martinez, and Vijay Pande, et al., "Systematic Improvement of a Classical Molecular Model of Water", *J. Phys. Chem. B* 117:9956-9972 (2013), and can be freely downloaded from <http://simtk.org/home/forcebalance>.

In Other Simbios News...

The Duke's Choice Awards from Oracle and the Java community honor organizations and developers for their creative and innovative uses of Java technology. Ayman

Habib accepted the award on behalf of the OpenSim team at the recent JavaOne conference, attended by 60,000 people in San Francisco from September 22-26. □

CURRICULUM AND BIG DATA: Revamping to Open the Bottleneck

By Katharine Miller

Things change quickly in the fields of computational biology and bioinformatics. “New technology comes along and whoa! You need to design a new course!” says **Claudia Neuhauser, PhD**, director of graduate

studies, biomedical informatics and computational biology at the University of Minnesota, Rochester.

big data will be quite different from imaging or electronic medical records (EMR) research using big data, it’s hard to imagine a generic “Big Data for Biomedicine” curriculum.

The NIH workshop participants therefore

discussed many different types of training opportunities that could help open up the bottleneck, says **Karen Bandeen-Roche, PhD**, professor of biostatistics at Johns Hopkins’ Bloomberg School of Public Health, who led the workshop discussion together with **Zak Kohane, MD, PhD**, director of the informatics program at Children’s Hospital, Boston, and co-director of the Center for Biomedical Informatics at Harvard Medical School. All agree that there is no one-size-fits-all solution. So here’s a sampling of potential programs, some of which are already being piloted at various institutions around the country while others may require funding from BD2K or other sources.

ular courses in data analysis. And one of Kass’s colleagues, **William Cohen**, has developed a course in machine learning with large datasets. “It’s a class that specifically talks about how to scale things up,” he says. Courses of this type should be more widespread, Kass says.

“New technology comes along and whoa!
You need to design a new course!”
says Claudia Neuhauser.

Today, she says, biomedical “big data” are putting pressure on bioinformatics curriculum. Be it genomic or molecular data, imaging data, or electronic medical records, big data adds a new level of complexity that requires a shift in training.

“We need new research and we need revamped training programs,” says **Rob Kass, PhD**, professor of statistics and machine learning at Carnegie Mellon University.

In July 2013, the National Institutes of Health (NIH) hosted a workshop to discuss possible training initiatives to help people take full advantage of big data. The workshop, which was part of the Big Data to Knowledge (BD2K) initiative, generated plenty of ideas (some of which are described below) that may soon find their way into a grant program.

But even without new grants, the bottleneck in big data training requires academic institutions as well as society at large to ponder a difficult question: What kinds of training opportunities are needed to ensure that researchers can extract knowledge from biomedical big data?

As one might suspect, an individual’s background is a huge factor in determining the kinds of training needed as well as its duration (short- vs. long-term). The specific research question being addressed also affects curriculum. Because genomics research using

discussed many different types of training opportunities that could help open up the bottleneck, says **Karen Bandeen-Roche, PhD**, professor of biostatistics at Johns Hopkins’ Bloomberg School of Public Health, who led the workshop discussion together with **Zak Kohane, MD, PhD**, director of the informatics program at Children’s Hospital, Boston, and co-director of the Center for Biomedical Informatics at Harvard Medical School. All agree that there is no one-size-fits-all solution. So here’s a sampling of potential programs, some of which are already being piloted at various institutions around the country while others may require funding from BD2K or other sources.

Build on a Data Science Foundation

For researchers who already consider themselves data scientists in biomedicine, the leap to big data isn’t a huge stretch, says Kass. “People already good at data analysis have an easy transition to bigger datasets because the principles haven’t changed,” he notes. Still, the growing size of biomedical data sets means students need an appreciation for computer systems and software engineering as well as algorithms and statistical methods, not to mention all the issues associated with data warehousing, standardization, access, security, and confidentiality, Kass says.

Carnegie Mellon is already building more and more references to big data into its reg-

Use Case Histories at the Cutting Edge

Kohane would like to see training provided “right at the cutting edge of where the experts are.” He supports the use of case studies that involve problems created by the size of the data set and the limits on computational resources and bandwidth. As he sees it, learning happens best when there is a problem that biomedical domain experts believe is important, and they have the methodological people working on it with them. “Then it’s not make-work, and it’s not tangential,” he said during the workshop.

Train Team Members

People working at the interface of big data and biomedicine will inescapably work in teams of individuals with diverse sets of skills, Bandeen-Roche notes. The question then becomes, she says, “How do you create a community well-trained to be team members?”

“How do you
create a community
well trained to be
team members?”
Bandeen-Roche says.

In Bandeen-Roche’s own training program on the epidemiology and biostatistics of aging, people from different fields gain expertise in their own areas but are also

trained in shared activities—common curriculum and shared research projects—where they learn to work together and

graduate program in biomedical informatics and computational biology. “We’re preparing people for something that would have

“There is a need to train the vast numbers of people in the biomedical field who might need to talk to patients with a genome on a stick,” Neuhauser says.

communicate across disciplines. A similar approach might work for big data, she says.

Train for Real World Data

Colin Hill, PhD, chairman and CEO of GNS Healthcare, a big data analytics company, says there are really two buckets of healthcare big data: the bioinformatics/genomics side for drug discovery and development; and then the real-world data side dealing with mash-ups of EMRs, claims, and genomics data. “Our biggest growth (and the biggest growth in the field),” Hill says, “is in the real-world data side.” For jobs in this arena, he says, current training programs in computational biology or bioinformatics lack the necessary epidemiological training while current epidemiological training “typically doesn’t cover the new math of causal inference/Bayesian network inference and gives little exposure to claims data and EMR data.” Curriculum developers, he says, should take this job market into account.

Take It to the Users

Bioinformatics and computational biology programs are a diverse lot, Bandeen-Roche says. “Some might approximate what is needed to handle big data while others don’t,” she says. And only a few are dedicated to training the users of big data rather than PhDs. “It’s important to think about training the huge number of people needed to do the standard day-to-day stuff—interpreting and explaining the data and translating it to clinical practice,” she says. “At the end of the day, it somehow has to help patients.”

Neuhauser’s program at the University of Minnesota has a huge focus on educating the local workforce of people at Mayo Clinic who are working in the labs where big data is already being used. Mayo employees have come to Neuhauser’s program seeking new skills. Because many have biology backgrounds, she established a sequence of three online quantitative courses (an introductory computer science course, as well as separate algorithms and programming courses) that enable them to enter the

been closed off to them a few years ago,” Neuhauser says.

Brush Up Professionals’ Skill-Sets

Several workshop participants noted their own need for big data training. “We all need to be retooled, PhD students as well as the rest of us,” said **Elaine Larson, PhD**, associate dean for research at Columbia University School of Nursing. Current professionals could benefit from short-term training provided online, at workshops,

“We all need to be retooled, PhD students as well as the rest of us,” says Elaine Larson.

bootcamps, or summer programs. Fellowships could help medical doctors learn big data informatics. And team challenges and competitions can provide training with the extra dose of reality that only a true problem-oriented experience can provide.

Get Creative—MOOCs and Modules

We live in a new era of education where MOOCs (massive online open courses) allow the possibility for scaffolding courses and making them broadly available. **Andrew Laine, PhD**, professor of biomedical engineering at Columbia University, noted during the workshop that the NIH could require grantees to contribute modules to a broadly shared resource. “If it can be done once and done really well, it can be a commodity to the community,” he said.

As Bandeen-Roche notes, “Extracting knowledge from big data is more difficult than one might have hoped.” Hopefully, efforts to design new training programs will allow researchers meet the challenge. □

BIG DATA PROGRAMS OUTSIDE BIOMEDICINE

If programs focused on biomedicine don’t offer enough big data training, researchers can look for opportunities elsewhere. For example, several organizations already offer novel training programs in big data.

Berkeley’s Simons Institute is currently running a four-month program called “Theoretical Foundations of Big Data Analysis.” Although not specifically designed for biomedical researchers, the program covers the same big data turf that a biologist would find useful, such as succinct data representations; parallel and distributed algorithms; and big data privacy. And several of the instructors have experience using big data in biomedicine.

Insight Data Science also offers a 6-week post-doc training fellowship to help “bridge the gap between academia and a career in data science.” **Jenelle Bray, PhD**, who recently completed her post-doc at Stanford using machine learning to study protein structures, entered the Insight program in September. She saw it as the best way to get the training she needed to lead a data science team—in biomedicine if possible. “Sometimes the newest technologies take a while to permeate academia,” she says. “You can learn them more quickly going into industry.”

VACCINES BY THE NUMBERS: Computational Approaches to Design Vaccines Faster

By Amber Dance

Like a “Wanted” poster distributed to a posse, peptide vaccines show the immune system a small sample (about eight amino acids) of a pathogen, training the body to seek and destroy viruses, cells or bacteria that tote identical peptides. But just as

rendering the vaccine impotent.

To design effective vaccines, Chakraborty says, scientists need a deeper understanding of how each amino acid in a key protein contributes to HIV fitness, and how multiple mutations work together. It’s no simple prob-

example of a fitness landscape: Suppose a virus contains just two amino acids. For each potential amino acid pair, the virus may be fitter or weaker. When mapped, there will be peaks for fit pairs and dips for poor ones. In designing an effective vac-

“One of my own goals...is to see how we can make vaccine design a systematic discipline,” says Arup Chakraborty.

a criminal might dye his hair and get a nose job to avoid recognition, some vaccine targets can frequently mutate those peptides, rendering them invisible to the immune system posse. Designing vaccines to handle these shape-shifters has proven challenging using the traditional trial-and-error approach. Today, computer and physical scientists are trying to change that by developing computer programs and simulations that identify the likeliest vaccine candidates and test out their capabilities *in silico*.

“One of my own goals...is to see how we can make vaccine design a systematic discipline,” says **Arup Chakraborty, PhD**, director of MIT’s Institute for Medical Engineering and Science in Cambridge. Two recent studies take steps in that direction—and offer hope of vaccines to treat HIV infection and cancer.

Hitting HIV at Its Most Vulnerable

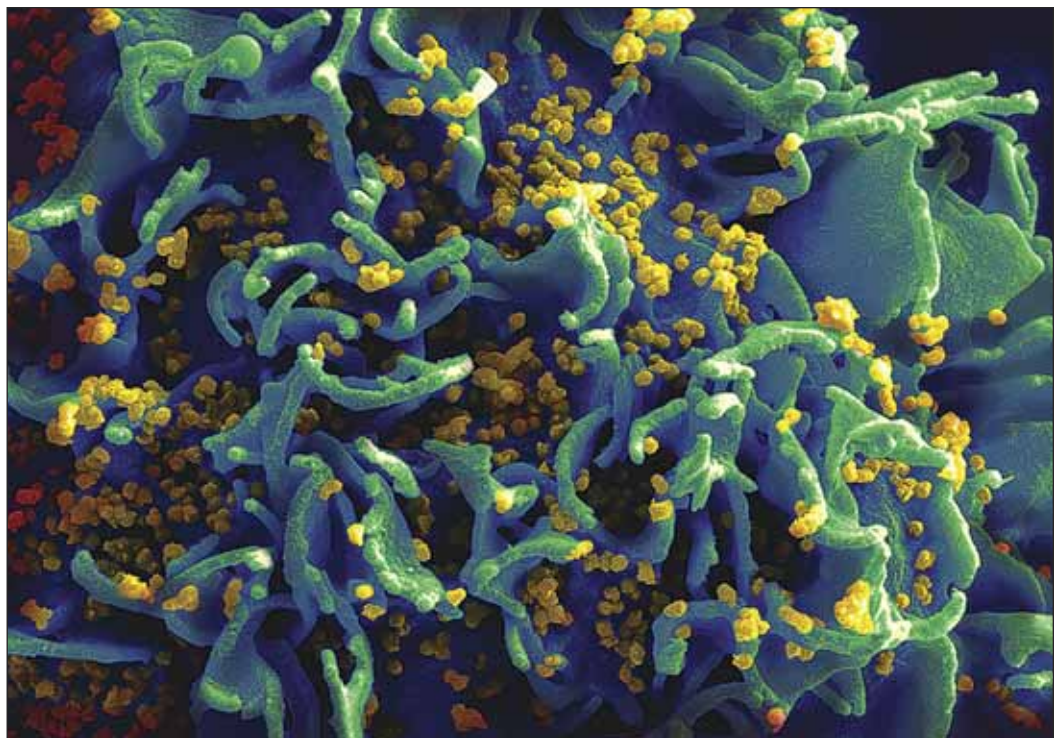
As soon as immune cells learn to recognize HIV, the rapidly mutating virus tweaks its peptides and becomes invisible once more. Vaccine-makers have tried to foil HIV by training the immune system to recognize the virus’s most crucial peptides—those that, if mutated, would weaken the virus. Thus far, this approach has failed because the virus can often make additional compensatory mutations, wiping out the disadvantage caused by the first mutation—

lem: Each amino acid can mutate to 19 different alternatives. Bioengineering each possible combination would be preposterous. That’s where computers can help.

In a paper published in *Immunity* in March

cine, Chakraborty says, “You want to push the virus off the hills and into the valleys.”

Of course, for the Gag proteins, which together encompass 500 amino acids, this computation is more complex. The first in-



HIV particles attack a human T cell in this scanning electron micrograph. Courtesy of National Institute of Allergy and Infectious Diseases (NIAID).

2013, Chakraborty and his colleagues—**Bruce Walker, MD**, **Andrew Ferguson, PhD**, and **Thumbi Ndung’u, PhD**—computed the fitness landscape for the HIV polyprotein Gag, which contributes the main structural elements of the virus. Here’s a simplified

carnation of the model (in the *Immunity* paper) calculates the fitness of Gag sequences made up of various combinations of wild-type amino acids and alternative (mutated) residues. It considers not only Gag with single mutations, but Gag with

every possible pair of mutations, or three, four or more mutations at once. The group is currently expanding the model, calculating fitness for not just wild-type or mutation, but for any of the 20 possible residues at each place in Gag. To determine fitness, the researchers measured the prevalence of different mutations in HIV DNA sequences collected from patients. They inferred the protein sequence from the DNA and assumed that more predominant strains were the fittest. The result is a multidimensional topographical map, with peaks where the virus does well and valleys where it's weak.

To test the model *in vitro*, the researchers engineered a handful of viruses with different sequences and infected human cells with the strains. Sure enough, the viruses that the computer predicted to be least fit replicated the slowest.

Using fitness landscapes, Chakraborty can identify places where mutations would cripple HIV, and the virus could not easily

Modeling Cancer Vaccines

Therapeutic vaccines can train the immune system to attack not only infections, but also cancer. “The tumor tries to hide from the immune response,” says **Robert Preissner, PhD**, of the Medical University of Berlin. The “Wanted” poster elements of cancer vaccines, designed to target individual cancers, are peptides that represent abnormal proteins on a tumor’s surface.

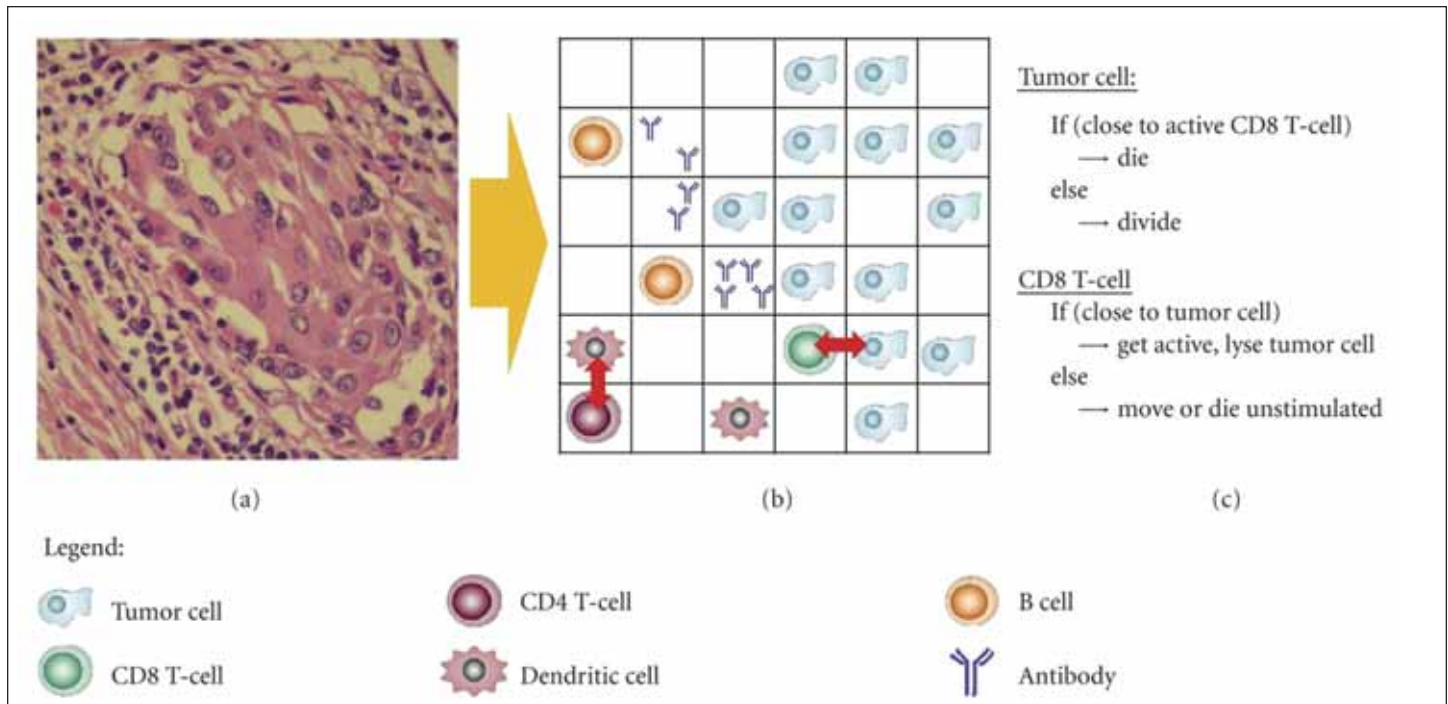
For a vaccine peptide to work, it must interact with two immune system proteins: the major histocompatibility complex (MHC) found in all cells and the T-cell receptors found on the surface of white blood cells. MHC molecules stick out from the surface of cells, displaying sample peptides from inside the cell—including vaccines—and signaling whether the cell contains foreign or otherwise undesirable material. A therapeutic cancer vaccine trains T cells to recognize the native but anomalous proteins expressed in cancer, so

whether a given peptide would help the immune system fight off cancer.

In the VaccImm model, cancer cells, immune cells, antigens and antibodies interact according to set rules. For example, if a T cell recognizes an antigen, it becomes activated and kills tumor cells. Users can input different peptide vaccines, and the program will calculate their effect on tumor growth. VaccImm is freely available online at <http://bioinformatics.charite.de/vaccimm/>.

In his simulations, Preissner has noticed that multiple vaccine peptides—four or more—work best. This matches the experience of immunologists using multiple peptides *in vivo*. However, only clinical trials will show if peptide cocktails that work in the simulation will work in people, Preissner noted.

VaccImm is still missing useful parameters, Preissner says. For one, there are many different types of MHC molecules. He would like to extend the list of MHC types



Two-dimensional representation of the players in a VaccImm simulation. Image credit: Robert Preissner.

compensate and evolve back to full fitness. “If you make those mutations, the virus is screwed,” he says. The group has designed therapeutic vaccines that should force HIV to make just those mutations, and is working toward a trial in monkeys.

The approach combines two technologies, DNA sequencing and computation, that are becoming ever cheaper, Chakraborty notes. “I think this will be useful for any mutating virus that we have today, or that will emerge in the future,” he says. Influenza, for example, is another rapidly mutating virus.

they will attack a tumor.

Researchers have made computer models of cancer vaccine action before, but only with simplified yes-or-no interactions between the vaccine, MHC proteins, and T cell receptors. In a study published by *BMC Bioinformatics* in April 2013, Preissner and his colleagues present an updated model, VaccImm, which calculates, in greater detail, the interactions between the specific amino acids in the vaccine peptides and those in the T-cell receptors and MHC molecules. The model should better predict

available to users. In addition, the current incarnation does not allow the cancer cells to mutate, but it should be possible to add this feature.

Francesco Pappalardo, PhD, of the University of Catania in Italy, co-developed the original framework on which VaccImm was based. He says computational vaccine development will save time, money and the lives of experimental animals. Moreover, he says, it will help immunologists understand the biological processes that underlie vaccine success. □

CANCER'S CRYSTAL BALL: Personalized Tumor Models to Guide Treatment

By Sarah C.P. Williams

When Kristin Swanson's father was being treated for lung cancer, doctors collected no shortage of data on his disease. They scanned his chest, regularly drew blood, and biopsied his tumor to study the cancerous cells. But each test told a different, sometimes contradictory, story about the cancer. And when Swanson asked the doctors about her dad's prognosis, their predictions often seemed rooted in averages for all lung cancer patients, rather than being informed by any of the test results.

"I realized," says Swanson, who started her career as an applied mathematician, "that there were all these different pieces of data, and no one was bringing them together."

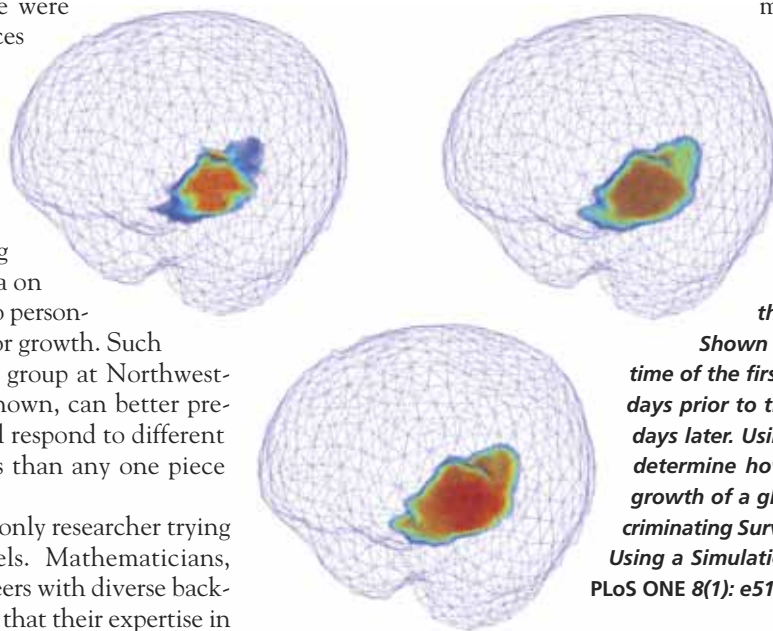
The experience motivated Swanson to focus her research on developing ways to integrate data on a patient's cancer into personalized models of tumor growth. Such models, her research group at Northwestern University has shown, can better predict how a tumor will respond to different treatments and drugs than any one piece of data alone.

Swanson isn't the only researcher trying to build such models. Mathematicians, physicists, and engineers with diverse backgrounds have realized that their expertise in studying complex systems can help them make sense of cancer. So they're creating models of the physical forces on tumors; developing equations to describe how cancers grow and spread; and using mathematical approaches to study how the molecular pathways in cancer cells interact.

"The reality is that no matter how complicated the molecular biology is, tumors are physical systems that obey the laws of physics," says Vittorio Cristini, PhD, pro-

fessor and director of computational biology in the pathology department at the University of New Mexico Cancer Center.

The power of the models and equations lies in the fact that data on any given patient's tumor can be plugged into the formulas and the resulting output—whether it's a prediction of a drug's benefit, a survival prognosis, or a description of the tumor's growth—will be personalized to that patient. The models haven't yet led to major changes in how doctors treat cancer outside of clinical trials, but they're poised to make a difference.



scans can reveal some aspects of the tumor size, but little else.

"Cancer is by definition a dynamic disease," says Swanson. "So it doesn't make sense to judge it with scans at single time points."

In 2010, Swanson reported in *Physics in Medicine and Biology* that by creating a growth model of a patient's glioma from a series of brain MRIs, she could predict whether the tumor would shrink in response to radiation therapy. She's now working with clinicians at Northwestern and other medical institutions to optimize how this model can guide the radiation therapy dose chosen for each patient and to create an iPad app that would put the models into the hands of doctors.

Her research team has also adapted

Using MRI scans of a patient's glioblastoma at multiple time points, Kristin Swanson created a personalized model of the tumor's growth and could determine the theoretical tumor size at any time point.

Shown (clockwise from top left) is the tumor at the time of the first pre-treatment MRI, the modeled tumor 19 days prior to the initial scan, and the modeled tumor 107 days later. Using this model as a baseline, Swanson could determine how much any given treatment affected the growth of a glioblastoma. From Neal ML, et al. (2013) Discriminating Survival Outcomes in Patients with Glioblastoma Using a Simulation-Based, Patient-Specific Response Metric. PLoS ONE 8(1): e51951. doi:10.1371/journal.pone.0051951

Modeling Tumor Growth

Swanson has focused her initial modeling efforts on gliomas, aggressive brain tumors with few treatment options. Gliomas rarely spread to other organs, making them an appealingly simple tumor type to model. But it's also notoriously hard to predict the prognosis for patients with glioblastoma multiforme, so there's lots of room for improvement in the clinical realm. Brain MRI

the model to be used in other situations. Based on two MRIs taken at least five days apart, they create a mathematical description of the kinetics and shape patterns of how a patient's tumor is growing. Then, they can use the model to project the size of the tumor at any later date.

Their most recent study, published in May 2013 in *Cancer Research*, used these modeled projections to study of the effectiveness of

"The reality is that no matter how complicated the molecular biology is, tumors are physical systems that obey the laws of physics," says Vittorio Cristini, a mathematician at the University of New Mexico.

chemotherapy and radiation combinations in 63 glioma patients. Standard response metrics comparing the size of a tumor before and after treatment are poor predictors of overall survival, Swanson says. But using models based on routine scans the patients already had, her group was instead able to compare the projected size of a tumor without treatment to the size of the tumor after treat-

ment. The resulting metric, dubbed "Days Gained," measured not just a net change in size of a tumor, but took its growth speed into account. Patients who had a "Days Gained" result of more than 117 days after their initial therapy were most likely to survive long-term.

"A dozen years ago, I gave presentations on modeling tumors and was routinely laughed at by oncologists," says Swanson. "Now that we're getting real clinical results and have cohorts of patients, we're being listened to."

Earlier this year, Swanson laid out the current status of the field that she calls mathematical neuro-oncology in a *Frontiers of Oncology* review. Models, she wrote, through metrics like "Days Gained," are helping identify patients who can better be treated with deviations from the standard of care. But it will take more doctors and institutions buying into the benefits of models before such model-based personalized care is routine. Already though, Swanson says she's seen more acceptance of modeling from clinicians.

"A dozen years ago, I gave presentations on modeling tumors and was routinely laughed at by oncologists," says Swanson. "Now that we're getting real clinical results and have cohorts of patients, we're being listened to."

While Swanson primarily models how tumors grow, Cristini is more concerned with mathematically describing how molecules from the outside can infiltrate a tumor. Whether or not drugs can reach the deepest, densest parts of a tumor, he thinks, is a driving fac-

tor in whether the drug can effectively fight the cancer.

"The physics of transport might be the single most important mechanism for drug resistance," he says.

By modeling the environment in and around a tumor, he's found, he can predict whether a treatment will be successful based on how drugs can perfuse into the tumor.

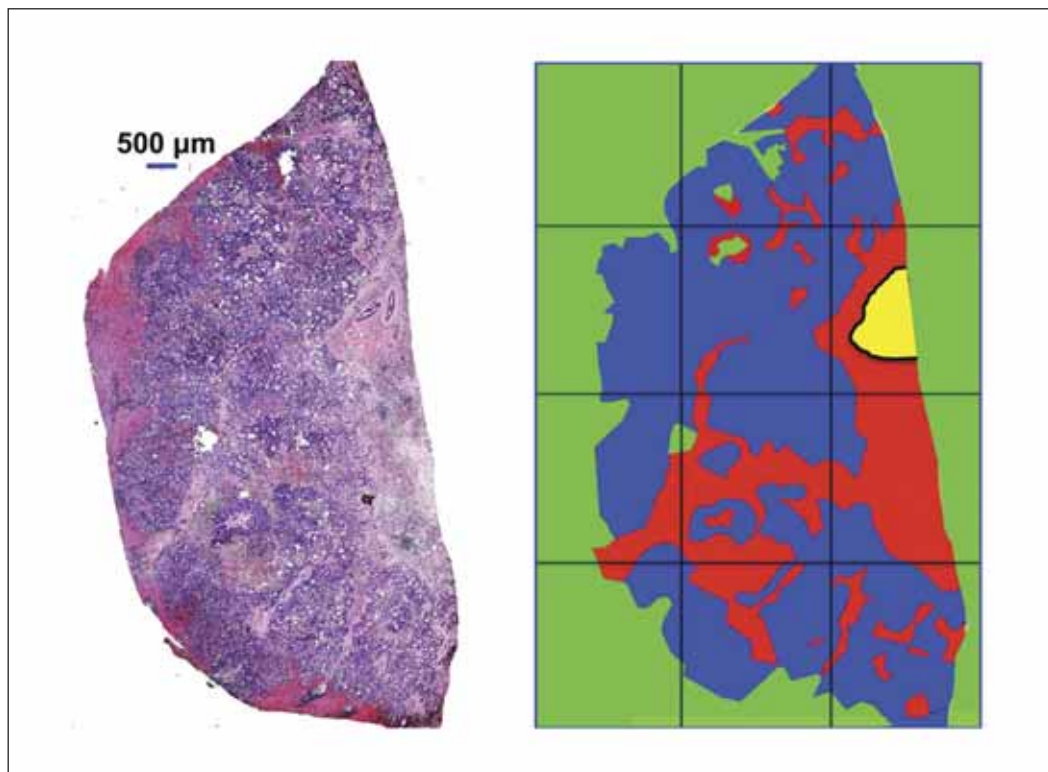
And such predictions, like those that Swanson has made based on her glioma models, can help guide clinician decisions between therapy options or dosages on a personalized level.

In August 2013 in *PNAS*, Cristini and his collaborators published the results of a study on colorectal cancers that had spread to the

livers of 10 patients. Using microscope slides containing samples of the liver tumors after chemotherapy, the scientists calculated the distribution and sizes of blood vessels that ran through each tumor. Then, they analyzed which tumor cells, and how many, fell into the so-called "kill radius," the zone where cells had been successfully killed by chemotherapy. Working backward, the team was able to generate a mathematical equation linking blood vessel characteristics to the resulting kill radius. The equation can now be used prospectively to calculate what dosage of chemotherapy is required to penetrate the entire tumor.

Cristini has applied his mathematical models of drug diffusion not only to the liver, but to tumors in the brain and breast as well. In *PLOS One* in April 2013, he showed that the inability of immunotherapy drugs to reach every part of a breast cancer explains why some tumors are unresponsive to the therapy. Most clinicians and biologists, Cristini says, had assumed that a molecular difference between how tumor cells respond to drugs—rather than a difference in the ability of drugs to reach tumor cells—was to blame for the differing outcomes.

Cristini's goal is to develop what he calls



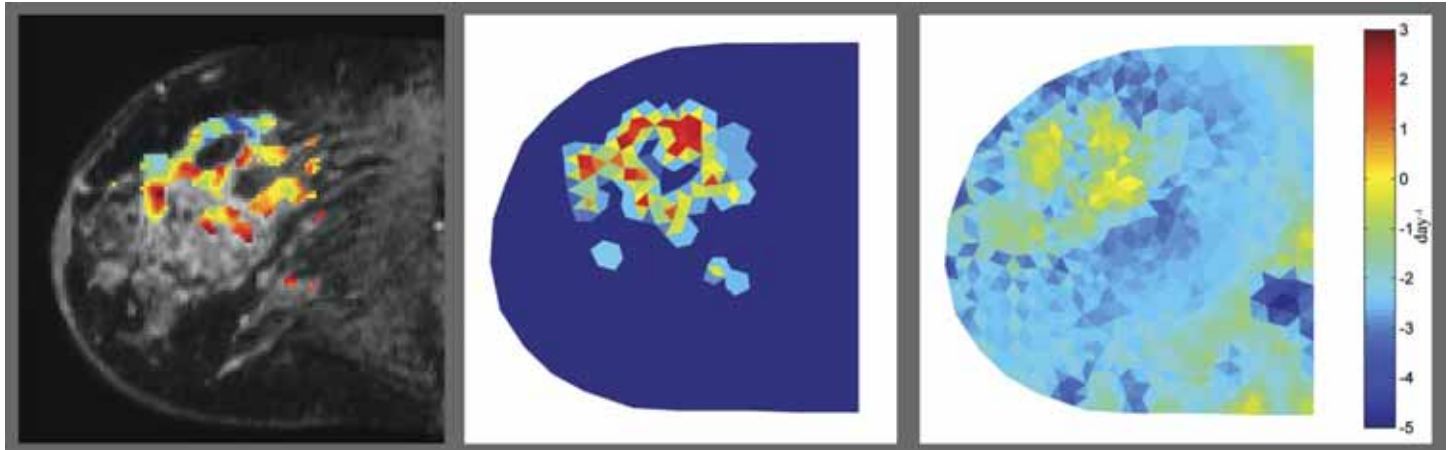
By taking histological samples from colorectal cancers that have metastasized to the liver, Vittorio Cristini's group determined which sections of the liver tumor had been killed by chemotherapy (left). Then, the scientists made a computer model (right) showing which areas of the tumor were alive (blue), dead (red) or part of the liver's portal triad and central vein (yellow). Using the model, they could determine how responsive a patient's tumor was to chemotherapy, and calculate how much drug would be necessary to eradicate the tumor. Reprinted from Pascal, J, et al., Mechanistic patient-specific predictive correlation of tumor drug response with microenvironment and perfusion measurements, *PNAS* 2013 110: 14266-14271.

“master equations of cancer.” Every physical parameter of a tumor, he says, can be described through physics and mathematics. And, as researchers like him are increasingly

the macroscopic and microscopic level.

The more data researchers integrate, though, the more data they have to store and process for modeling. And that presents a

challenge. “If a model only runs at Sandia or Los Alamos because it requires so much computing power,” he says, “then it’s not very practical for most clinicians to use.”



To gauge the response of a breast tumor to neoadjuvant chemotherapy, Thomas Yankeelov’s research group used scans of the tumors at various time points to determine tumor cell distribution. Together, these models from multiple time points could be used to reconstruct cell proliferation and diffusion model parameters. Following parameter optimization, the model is used to predict the tumor cell distribution at the final time point and compared to experimental observations. Courtesy of Thomas Yankeelov, Jared Weis and Mike Miga.

showing, many of these physical attributes are closely linked to differences between tumors and treatment success rates.

Modeling Molecular Pathways

But the complexity of cancers doesn’t just lie in physical properties that can be extracted from scans. Tumors are also diverse at the molecular and genetic levels. And modeling is ripe for understanding how the molecular attributes of a tumor influence

challenge. “If a model only runs at Sandia or Los Alamos because it requires so much computing power,” he says, “then it’s not very practical for most clinicians to use.”

Once multi-scale models are perfected, Deisboeck expects them to be used not only to guide decisions on individual patients, but to generate hypotheses on how novel cancer drugs will affect every aspect of a tumor.

“It’s all about target validation,” he says. “With a model, you can ask how targeting

of correlating available data to outcomes.

Yankeelov’s lab at Vanderbilt University is working with oncologists to model how breast tumors respond to neoadjuvant therapy—drugs given before surgery with the goal of eliminating the cancer. By scanning patients before and after a neoadjuvant drug is given, they’re developing equations that may be able to predict better than individual scans whether or not the neoadjuvant therapy will be effective at getting rid of cancer cells.

“Modeling can help us design better informed clinical trials and gauge better whether treatments are working,” Yankeelov says.

their physical properties.

With multi-scale modeling, **Thomas Deisboeck, MD**, associate professor of radiology at Massachusetts General Hospital and Harvard University says, researchers integrate data not only from scans, but also from isolated cells, tumor biopsies, and even blood samples. “But getting that type of data consistently even for one patient is spotty, let alone trying to get a big data set,” he says. And that’s what’s holding the field back.

To increase the power of existing data sets he’d like to see more collaboration within the field, and the establishment of standards and common markup languages that work for multi-scale models, he and his colleagues wrote in a 2013 commentary in *Cancer Informatics*. A new markup language called tumorML, he says, is poised to make a difference by working for models at both

a particular protein would change the behavior of the rest of a cellular system.”

If there are five drug options for a particular cancer, he explains, a multi-scale model could predict which drug or drug combinations—even what order and dosages to give the drugs in—would best help a particular patient.

Forecasting Outcomes

A hundred years ago, predicting a hurricane before it made landfall was nearly impossible. Today, sophisticated satellite measurements and a plethora of data are plugged into models that predict when and where and with how much force a hurricane will hit. That’s the metaphor that **Thomas Yankeelov, PhD**, associate professor of radiology and cancer biology at Vanderbilt-Ingram Cancer Center, uses to talk

Like the models others are developing, the ultimate goal of Yankeelov’s work is to find ways to better guide treatment decisions. And the challenge is getting physicians to buy into using the technology. Or at least, initially, integrating it into clinical trials.

“Modeling can help us design better informed clinical trials and gauge better whether treatments are working,” he says. His collaborations with physicians help bring the technology closer to those uses, transitioning from bench to bedside.

Today, a patient being treated for cancer will likely hear the same predictions that have been used in the past on the odds of treatment working for their tumor—based on averages. But as models make their way toward the clinic, these predictions will start to change. And for doctors and patients alike, that could prove a useful forecast. □

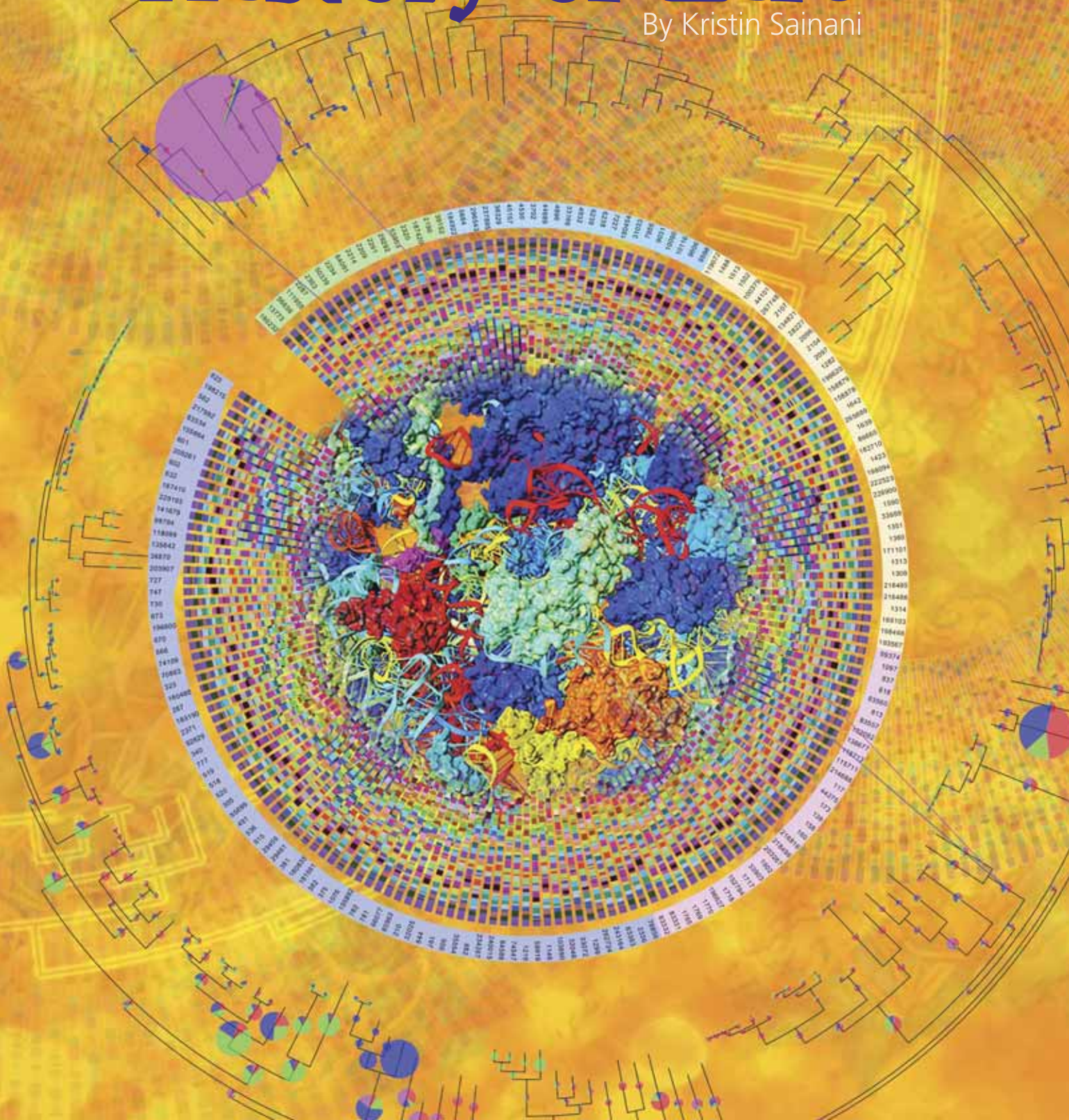
Computing

Using New Data
and New Models
to Tackle Old Puzzles

THE

History of Life

By Kristin Sainani



In 1977,

the late **Carl Woese, PhD**, shook up biology when he published the first tree of life based on genetic sequence data. His team showed that, contrary to popular belief, eukaryotes did not evolve from prokaryotes; instead, three distinct domains of life (bacteria, archaea, and eukaryotes) all arose from a common ancestor.

Woese's revelation is now considered one of the greatest biological discoveries of all time. But it was initially met with vehement skepticism from fellow scientists. He challenged a beloved paradigm in biology, and he initially paid the price. It took a decade for his findings to become widely accepted.

The question "where did we come from?" is one that philosophers, theologians, and scientists alike have been trying to answer for millennia. But reconstructing events that took place millions to billions of years ago is fraught with difficulties. The further back one goes, the less information there is; and the more people resort to filling in the gaps with ideas and stories. These ideas are often so neat and elegant and pleasing that it's hard to give them up, even when new data clearly contradict them.

Today we are in a data-rich era in evolutionary biology. For decades, computational biologists who work in phylogenetics have built evolutionary trees by inferring evolutionary distance from the similarities of DNA sequences for one gene. Now they can build trees using whole-genome sequences (currently available for numerous species). Armed with such data, as well as increasing computational power and sophisticated new computational models and tools, it finally might be possible to answer some of the toughest and oldest puzzles in evolution.

"It used to be that data were the limiting thing. But of course now, keeping up with the data is the problem. I've been around a long time and watched it all. It's been exciting," says **Russell Doolittle, PhD**, emeritus professor of molecular biology at the

"It used to be that data were the limiting thing. But of course now, keeping up with the data is the problem. I've been around a long time and watched it all. It's been exciting," says Russell Doolittle.

*Phylogenetic trees courtesy of Ivica Letunic and the Interactive Tree of Life, itol.embl.de; Letunic I, Bork P, Interactive Tree of Life (iTOL): an online tool for phylogenetic tree display and annotation, *Bioinformatics* (2007) 23(1):127-8. Ribosome foreground reprinted with permission from Harish A, Caetano-Anollés G, *Ribosomal History Reveals Origins of Modern Protein Synthesis*. *PLoS ONE* 2012; 7: e32776.*

University of California, San Diego. A pioneer like Woese, he reconstructed animal evolution using protein sequence data in the 1960s.

This article reviews seven history-of-life puzzles on which computational biologists and bioinformaticians are making headway: How did life begin? Which came first: RNA or proteins? Or did metabolism come first? Is there a fourth domain of life? How have proteins evolved since life began? Why did introns evolve? And what drives the evolution of form?

To answer these questions, many computational biologists are venturing beyond phylogenetics and simple Darwinian tenets by incorporating chemistry, physics, protein structure, epigenetics, morphology, ecology, and development into their algorithms.

The answers to these puzzles may surprise you, and some remain hotly contended. “People are still argu-

years ago, when the last universal common ancestor (the primitive cells that gave rise to bacteria, archaea, and eukaryotes) first appeared on Earth. “The earliest thing you’re ever going to see by direct sequence analysis is already an incredibly complicated organism. It had a lot more than DNA; it had RNA, proteins, RNA machinery, transport, homeostasis, and bioenergetics,” says **Eric Smith, PhD**, an external professor at the Santa Fe Institute. “So you have to dig back way further than that in time.”

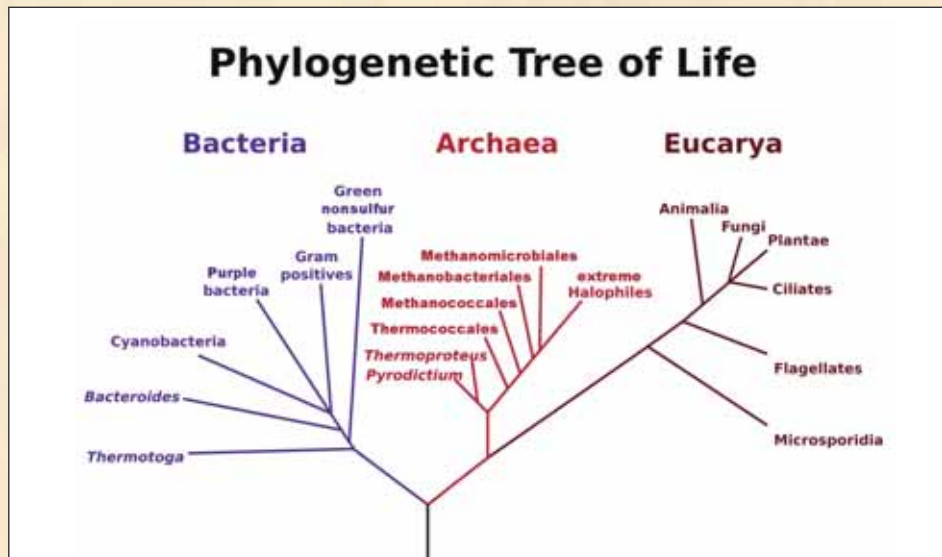
Experimental scientists have established several different scenarios for how organic molecules might have first appeared on Earth. For example, in 1953, **Stanley Miller** and **Harold Urey** famously created a “primordial soup” of amino acids by passing electricity (simulating lightning) through an airtight flask of water plus methane, ammonia, and hydrogen gases (which they believed, at the time, to be present on early Earth). After organic molecules first appeared, however, it is unclear how they joined together to build the basic machinery of life.

Some scientists have proposed that “autocatalytic sets”—groups of molecules capable of producing each other through mutual catalysis—were necessary to get things going. But others have argued that autocatalytic sets could not have arisen spontaneously. “Some say it’s equivalent to a tornado blowing through a junkyard and randomly assembling pieces of metal and plastic into a Boeing 747,” says **Wim Hordijk, PhD**, a computer scientist and owner of SmartAnalytix.com.

So, Hordijk and his collaborator **Mike Steel, PhD**, professor of mathematics and statistics at the University of Canterbury in New Zealand, decided to actually calculate the probability. “Nobody had ever looked at this in a concrete mathematical way,” Hordijk says.

“So that is what we’ve done. We proved mathematical theorems about it and ran computer simulations.” The mathematical framework integrates probability theory and graph theory—with molecules as nodes and interactions between them as edges in the graph.

In a 2012 paper in *Acta Biotheoretica*, they showed that autocatalytic sets do appear spontaneously with high probability. “In this simple model of a chemical reaction system where you have polymers floating around that could be glued together or broken apart and can do catalysis, it’s actually very likely that you will get these autocatalytic sets,” Hordijk says. Plus, smaller autocatalytic sets can team up together. “The smaller ones can grow into bigger ones. That’s necessary to get some sort of evolutionary process going,” Hordijk says.



Tripartite Life. In 1977, Carl Woese first proposed the radical idea that three domains of life arose from a common ancestor. He inferred evolutionary relationships by comparing sequence similarities in ribosomal RNAs across multiple organisms. His three-branch tree of life is now widely accepted. Created by Maulucioni from figure 1 in Woese CR, Kandler O, Wheelis M (1990). “Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya,” *Proc Natl Acad Sci USA* 87, and made available through the Wikipedia Commons at http://en.wikipedia.org/wiki/File:Phylogenetic_Tree,_Woese_1990.png.

ing many of the same arguments that they had before all of the data were there,” Doolittle says. But if there is one thing evolutionary biology needs, it’s a few renegades who aren’t afraid to challenge the status quo. Not all their revolutionary ideas will hold up to scrutiny, but those that do could forever change our understanding of life itself.

How did life begin?

Life on Earth began about 3.8 billion years ago. But exactly how this happened—how non-living chemicals transformed into organic building blocks and then living cells—remains a mystery.

Phylogenetics can only answer questions about what happened more recently than about 3.5 billion

It's difficult to create and study autocatalytic sets experimentally. But, in an October 2012 paper in *Nature*, experimentalists reported that small RNA molecules can spontaneously form a cooperative self-replicating network. Next, Hordijk and Steel simulated that system using their model, virtually replicating the experimental results, as published in a 2013 paper in the *Journal of Systems Chemistry*. They also made new predictions about the behavior of the system that the experimentalists are now testing. "The hope is that by doing these computer simulations, we can actually guide the experimentalists," Hordijk says. This particular experiment started with already assembled RNA, so it doesn't answer the question of how RNA formed in the first place, he notes.

Also, Hordijk and Steel's model makes no assumptions about the type of molecule involved; their mathematical framework could just as easily be applied to proteins or metal complexes. So, it doesn't answer the question of which type of molecule got life going.

Which came first: RNA or proteins?

Nucleic acids store the information that is needed to make proteins, but proteins are the workhorses that allow nucleic acids to replicate. So, scientists have long puzzled over which came first. In the 1980s, with the discovery that RNA can both store information and catalyze reactions, many scientists believed they had the answer: RNA came first (note that DNA is a more stable molecule believed to have evolved from RNA). The "RNA world" hypothesis—which purports that RNA got things going and was gradually replaced by proteinaceous enzymes and DNA—still prevails today. "I still accept the idea of an RNA world as real," Doolittle says. "There are RNA surrogates for many proteins. RNA could have easily been the intermediate that was gradually replaced by proteins."

But not everyone is convinced. For one thing, no one has ever synthesized ribose, the sugar backbone of RNA, in abiotic conditions, says **Jean-Michel Claverie, PhD**, professor of medical genomics and bioinformatics at the University of the Mediterranean in France. "I'm not in that field, but I had to review a book about the RNA world. And this is when I realized how weak the evidence is," he says. "The existence of an RNA world, although it would make a lot of sense and would elegantly explain the central role of the ribozyme in protein synthesis, is still not founded on anything solid."

In a 2012 paper in *PLoS ONE*, **Gustavo Caetano-Anollés, PhD**, professor of crop sciences at the University of Illinois at Urbana-Champaign (where Woese once worked), and his colleagues challenged the RNA world hypothesis. Caetano-Anollés builds evolutionary timelines by looking for similarities across organisms in 3-D structures—RNA secondary structures and protein folds—rather than in genetic sequences. "I have always been suspect of explo-

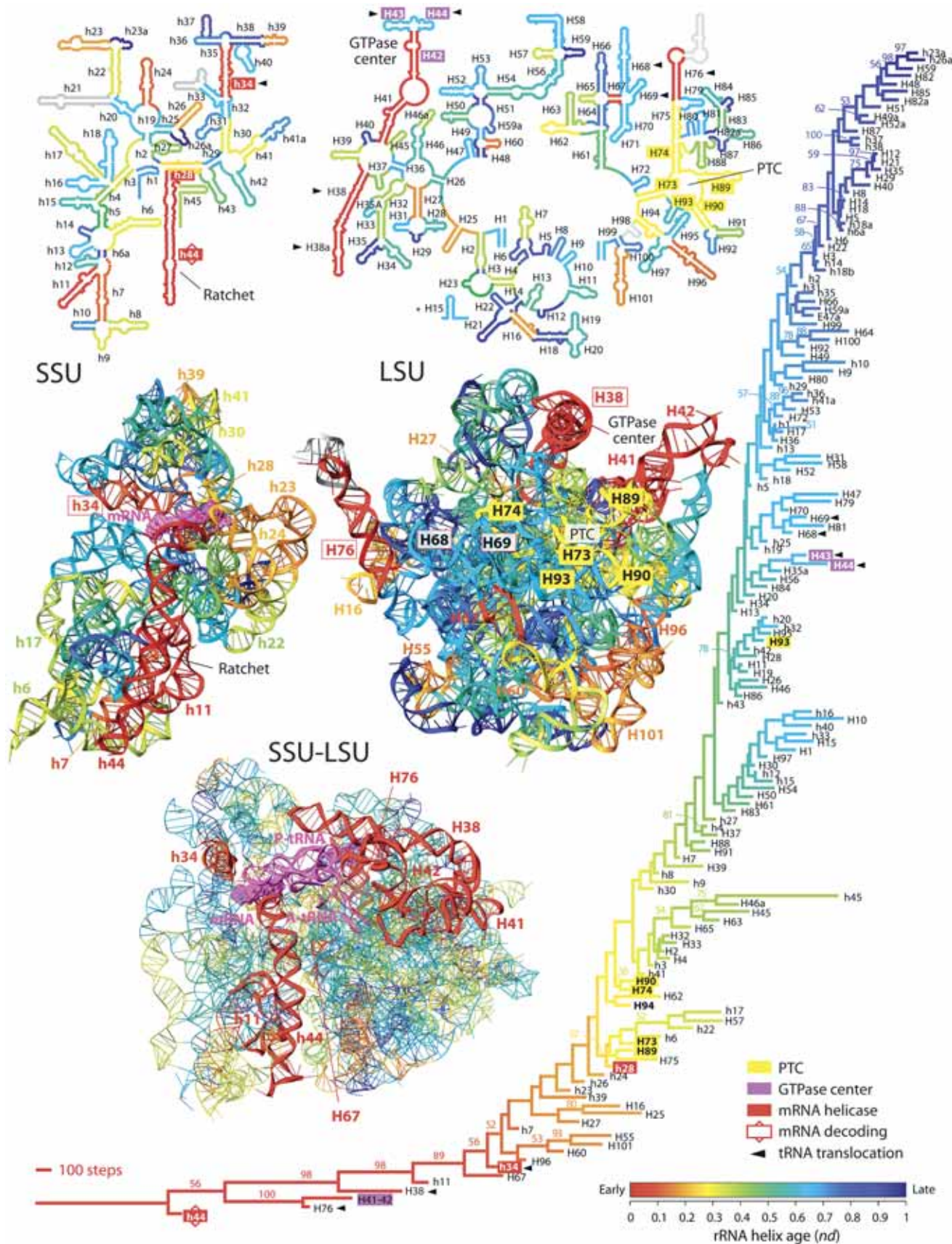
rations that come from sequence and target very deep evolutionary divergence," Caetano-Anollés says. "How can people make judgments about what happened so far back in time with something that is changing so incredibly fast?" Structures change at a much slower pace than sequence; structure comparisons are also much less sensitive to messy evolutionary

"I have always been suspect of explorations that come from sequence and target very deep evolutionary divergence," Caetano-Anollés says. "How can people make judgments about what happened so far back in time with something that is changing so incredibly fast?"

phenomena such as convergent evolution (independent evolution of similar features) and horizontal gene transfer (the exchange of genes between unrelated organisms), he says.

His team traced the evolutionary history of the ribosome using data from the SCOP (Structural Classification of Proteins) and CATH (Class, Architecture, Topology, Homology) protein structure classification databases (which group proteins into fold groups). They computationally compared ribosomal protein and ribosomal RNA structural domains across nearly 1000 organisms, including bacteria, archaea, and eukaryotes. The idea: structural domains that are the most universal are the oldest, whereas domains that appear in only a few organisms are the youngest. RNA-world proponents believe that the first ribosomes were composed solely of RNA. But Caetano-Anollés' team found that ribosomal proteins are just as old as ribosomal RNA; and that the two evolved together. Thus, early Earth was in fact a ribonucleoprotein world, he says. "My stance may not be popular with those who focus on sequence. However, structural genomic data have been analyzed and the interpretation is against the RNA world."

"Caetano-Anollés has certainly done some provocative stuff," Doolittle comments. "I think he's been mistaken about some of it, but his approach is so refreshing that I read all of his work, even though I'm skeptical of some of his conclusions." Doolittle wonders how the proteins could have been propagated without a memory system. "You can't have all the information for a particular kind of fold passed on from



A Structural History of Life. Gustavo Caetano-Anollés builds evolutionary trees based on the 3-D structures of RNAs and proteins. This tree reconstructs the evolutionary history of ribosomal RNA helices. The oldest structures are in red and the youngest structures are in blue. Similar analyses of ribosomal proteins (not pictured here) sug-

gest that ribosomal proteins and ribosomal RNAs coevolved, refuting the idea that RNA appeared on Earth before proteins (the so-called "RNA world hypothesis"). Reproduced from figure 2 of: Harish A, Caetano-Anollés G. Ribosomal History Reveals Origins of Modern Protein Synthesis. PLoS ONE 2012; 7: e32776.

one generation to another until you can explain how this information is stored. At the moment, that's still a fatal flaw," Doolittle says.

But, in a 2013 study in *PLoS ONE*, Caetano-Anollés' team provides evidence that protein synthesis occurred before there was a memory system (before there was a genetic code or ribosomes). "The ancestors of synthetases [the enzymes that load amino acids onto transfer RNA], are responsible for the specificity of the genetic code," Caetano-Anollés says. During transcription in the ribosome, tRNA molecules bring the correct amino acid into the growing protein sequence by matching their three-letter anticodons to codons in the messenger RNA. Synthetases contain two types of domains: those that perform the loading and those that read the anticodon to determine exactly which amino acid to load. Caetano-Anollés team found that the former are more ancient than the latter; this and other evidence suggest that these enzymes were originally involved in non-ribosomal protein synthesis. The genetic code only arose later, likely as a way to improve protein flexibility and function, Caetano-Anollés says.

Or did metabolism come first?

Smith also disputes what he terms the "radical RNA-first view." Though life may have gone through a stage in which RNA was the main molecule of both heredity and catalysis, he doesn't believe that RNA was the first mover. Rather, he says, metabolism began as a system of chemical reactions that did not involve RNA. Early metabolic networks could have arisen spontaneously and been catalyzed by minerals or perhaps simple small-molecule/metal complexes. "For early chemistry, we're not looking

In the metabolism-first view, the chemistry that eventually became life must have included methods for carbon fixation—converting inorganic carbon to organic carbon. Two carbon fixation pathways—the reductive citric acid cycle (also known as the reverse Krebs' cycle) and the Wood-Ljungdahl pathway—are believed to be the most ancient. But scientists have long debated which of these two evolved first.

Smith tackles these types of history-of-life questions by focusing on chemistry. "When you talk about the low-level chemistry, you don't need to refer to the genomic era of modern cells because whatever preceded them was also using low-level chemistry," Smith says. The metabolic networks that organisms use today are highly conserved. For example, modern autotrophs (organisms that can fix carbon and thus make their own food) use one of only six different pathways for carbon fixation. "This suggests that even the long-range evolution of complicated organisms has been strongly constrained by the principles of very low-level chemistry," Smith says.

To reconstruct the evolutionary history of biological carbon-fixation, Smith teamed up with **Rogier Braakman, PhD**, a fellow at the Santa Fe Institute. Braakman developed a novel computational technique called phylometabolic reconstruction, which integrates phylogenetics with flux-balance analysis, a type of metabolic analysis. In flux-balance analysis, researchers derive a series of equations to represent all the inputs and outputs in a metabolic network; then they simulate the flux of metabolites through this network, assuming constraints such as conservation of energy and mass. Braakman and Smith added the further constraint that early life must have been self-sufficient—able to make all its own building blocks. This limit confines the sequence-based phylogenetic reconstruction to a set of allowed configurations. "What we're doing here is saying: One thing

"For early chemistry, we're not looking for something that undergoes Darwinian adaptation, because the early chemistry is universal stuff that's never changed. We're just looking for stuff that will transduce energy, fix carbon, do the same things over and over again, and provide an ordered framework—out of which more molecular complexity comes later," Smith says.

for something that undergoes Darwinian adaptation, because the early chemistry is universal stuff that's never changed. We're just looking for stuff that will transduce energy, fix carbon, do the same things over and over again, and provide an ordered framework—out of which more molecular complexity comes later," Smith says.

that we know about autotrophs is that they made everything that they needed."

Their paper, published in *PLoS Computational Biology* in 2012, surprisingly concluded that neither the reductive citric acid cycle nor the Wood-Ljungdahl pathway evolved first; instead, primordial life contained both pathways. This redundancy may have

been an important failsafe since early life forms were probably fragile, Smith explains. Braakman and Smith also showed that further innovations in carbon-fixation were driven by the invasion of specific chemically novel environments (e.g., alkaline or oxidizing environments) more than by chance innovations in the genome.

Is there a fourth domain of life?

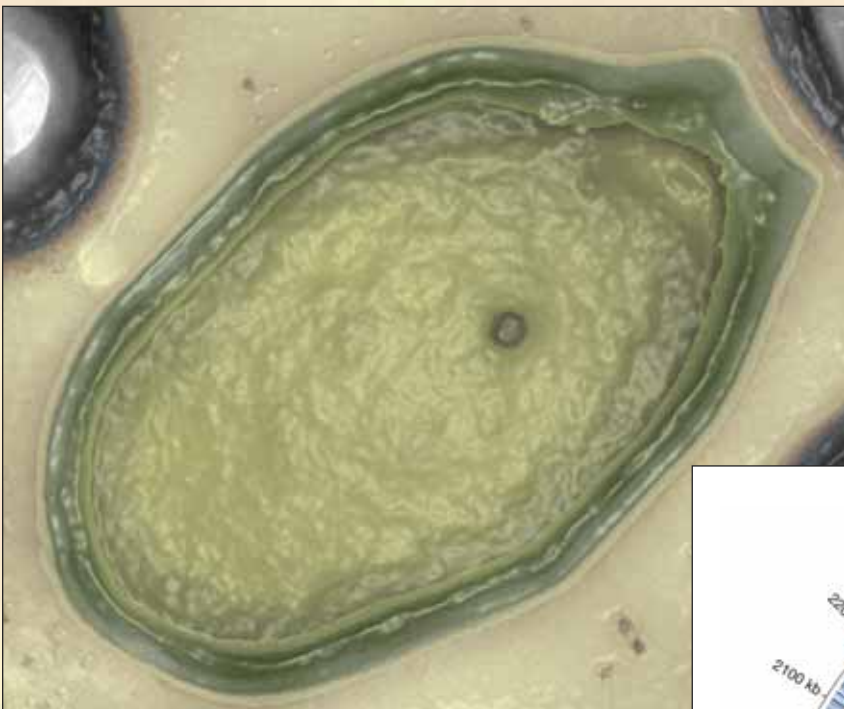
When it comes to reconstructing the history of life, viruses have traditionally been ignored. After all, it's not clear that viruses are even alive, given their lack of a cellular structure and dependence on cellular organisms. But with the recent discovery of giant viruses—which are as large and complex as some bacteria—viruses have suddenly taken center stage in evolutionary debates. Some researchers even argue that viruses comprise a fourth domain of life.

In 2003, French scientists identified the first giant

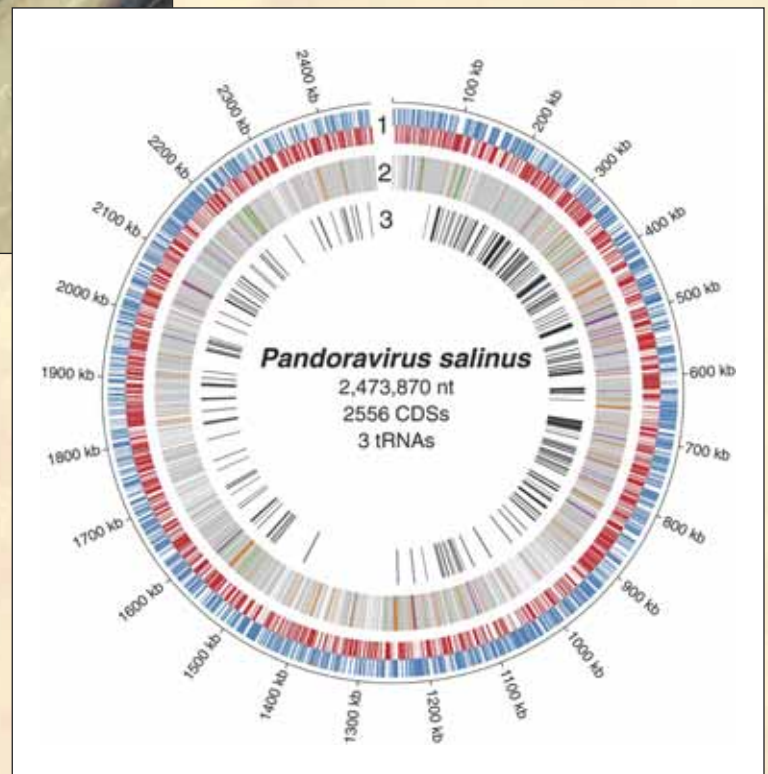
virus, which they named Mimivirus, short for “microbe mimicking virus.” The Mimivirus, which infects amoeba, can be seen under a light microscope and has more than 1000 genes, including some involved in protein translation and metabolism (hallmarks of cellular organisms). “This was a challenge for the classic paradigm of viruses,” says Claverie, who was involved in the discovery. Since then, Claverie’s team has uncovered several other giant viruses, including the Megavirus in 2011 and the most perplexing, the Pandoravirus, in 2013. The genome of Pandoravirus is twice as large as that of other giant viruses; and 93 percent of its genes resemble nothing ever sequenced before.

Mimivirus and Megavirus share certain protein translation genes, but are also highly genetically distinct. Claverie’s explanation: giant viruses descended from an ancient, cell-like common ancestor (one that has no modern cellular descendants). Over time, they lost genes and became parasitic. “We believe: the bigger the viral genome, the closer you are to the origin,” Claverie says. In phylogenetic reconstructions, Mimivirus and Megavirus wind up either at the base of the eukaryotic branch of life or on a completely new branch distinct from eukaryotes, archaea, and bacteria. Pandoravirus is so dissimilar to any known organism on Earth that its existence also challenges Woese’s tripartite tree of life. “It is an increasingly complicated story,” Claverie says.

Others strongly dispute this view, however. They believe that giant viruses are the ultimate gene robbers, and that their genomes are growing rather than shrinking. Giant viruses could have picked up their large and crazy genomes through horizontal gene transfer with their amoebal hosts (or other amoebal parasites). These looted genes may then

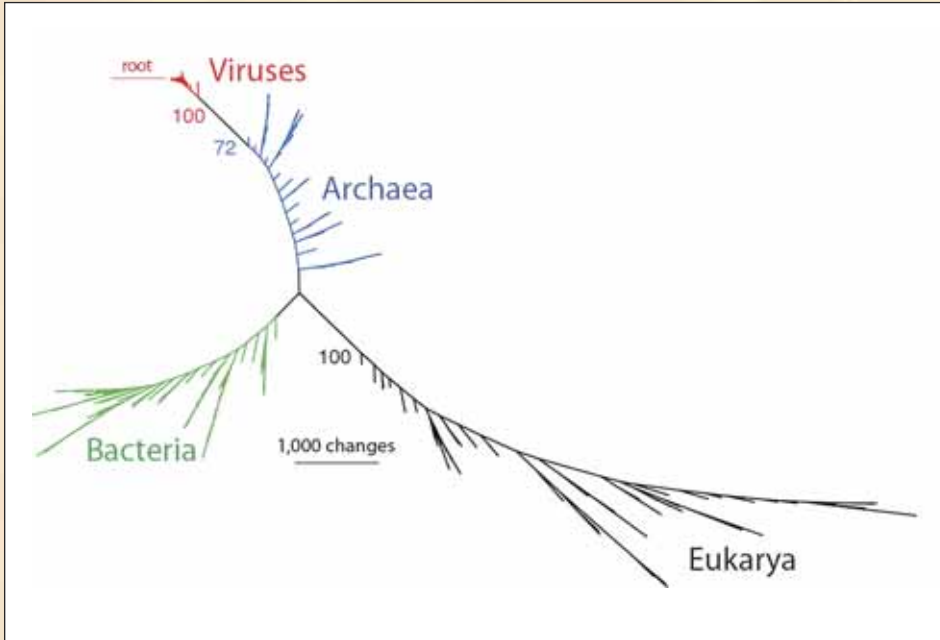


World’s Biggest Virus. Electron microscope image of Pandoravirus salinus (above) and a diagram of its genome (right). With nearly 2.5 million nucleotides (nt’s), the genome of Pandoravirus is as large as some eukaryotic cells and twice as large as any other known virus on Earth. But 93 percent of its genes resemble nothing ever sequenced before—opening up a Pandora’s box of questions about the history of life. In the genome picture, CDS=putative protein-coding sequences; CDSs on the direct (blue) and reverse (red) strands of DNA are indicated in the outermost circle. In circle 2, CDSs that match known genes or motifs are indicated in orange, green, purple, and white; CDSs with no match are shown in gray. Photo courtesy of: Chantal Abergel and Jean-Michel Claverie. Genome picture reproduced with permission from: Pandoraviruses: Amoeba Viruses with Genomes Up to 2.5 Mb Reaching That of Parasitic Eukaryotes. Science 19 July 2013; 341:281-286.



have evolved rapidly within the viruses, creating their puzzling genetic diversity. Using alternate models that account for such possibilities, other research

evolutionary history including giant viruses and other DNA viruses. And, like Claverie, he found that viruses clustered into a separate domain of life



A Fourth Domain? The discovery of giant viruses has raised the possibility that viruses comprise a fourth domain of life. Gustavo Caetano-Anollés' team built this evolutionary tree by comparing protein fold structures from the proteomes of archaea, eukarya, bacteria, and viruses/giant viruses (50 organisms each). They conclude that viruses are a distinct form of life that either predated or coexisted with the last universal common ancestor. Reproduced from: Nasir A, Kim KM, and Caetano-Anollés G. Giant viruses coexisted with the cellular ancestors and represent a distinct supergroup along with superkingdoms Archaea, Bacteria and Eukarya. *BMC Evolutionary Biology* 2012, 12:156.

groups have published phylogenetic reconstructions that place giant viruses squarely within the three domains of life, next to their amoebal hosts.

But Claverie isn't convinced by these arguments. "The thing is, if those viruses are picking up genes

that either predated or coexisted with the last universal common ancestor. "Until now, the universal tree is a tree of cellular lineages, not a tree of everything. From my point of view, that's an omission," Caetano-Anollés says.

"Until now, the universal tree is a tree of cellular lineages, not a tree of everything [including viruses]. From my point of view, that's an omission," Caetano-Anollés says.

from the environment, where are those cells? Because what has characterized those new viruses that we keep sequencing is that they don't look like anything else," he says. "They appear to steal genes from cells we haven't sequenced yet. And I don't think many people are prepared to believe that there is such a big loophole, such a big [set of] missing data."

Phylogenetic reconstructions are highly sensitive to models and assumptions, especially when dealing with viruses, as this debate reveals. But Caetano-Anollés also performed a structural reconstruction of

evolutionary perspective to this concept, she teamed up with Caetano-Anollés. "His way of mapping proteins structures on a timeline from four billion years ago to today was exactly what was needed to combine with our proteome-wide prediction of folding times," says Gräter, who is a group leader at the Heidelberg Institute for Theoretical Studies in Germany.

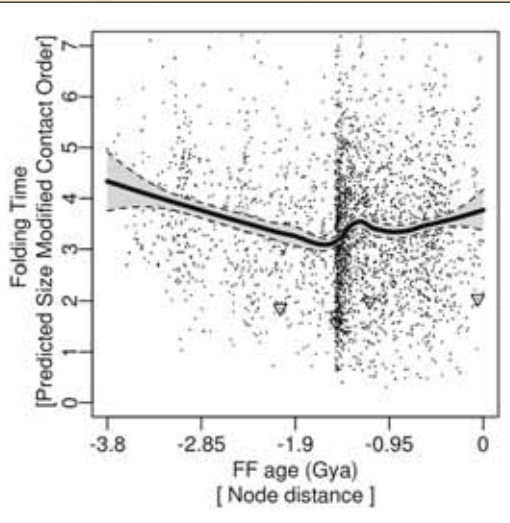
In a 2013 paper in *PLoS Computational Biology*, Gräter and Caetano-Anollés showed that protein folding became progressively faster from 3.8 billion to 1.5 billion years ago. (After this, alpha but not

How have proteins evolved since life began?

The earliest proteins to evolve were likely versatile but not optimized. Many researchers are trying to understand how proteins became optimized over the course of evolution. For example, what drove the evolution of different protein folds and of multi-domain complexes?

Frauke Gräter, PhD, an expert in protein folding, has long wondered about the evolution of folds. Her team made use of a model for predicting protein folding times for all proteins structurally known to date, based on the distance between contact points—amino acids that touch in the folded molecule—in the unfolded sequence. Contact points that start farther apart take longer to come together. To add an evolu-

beta folds continued to fold faster.) “Proteins were apparently folding faster and faster for most of the time during evolution. So there was pressure for efficient folding over time,” Gräter says. Faster protein folding likely prevents diseases that are caused by protein misfolding and aggregation, such as Alzheimer’s, she explains. “Once proteins are in their native fold, they are not prone to aggregation anymore.” Her team is now exploring evolutionary trends in other protein properties, such as floppiness and mechanical stability.



Folding Faster and Faster. By coupling a computational model that predicts protein folding times with a structural reconstruction of the history of different folds, Frauke Gräter’s team was able to trace how protein folding times have changed since the beginning of life. They found that protein folding became progressively faster from 3.8 billion to 1.5 billion years ago, at which time there was an explosion in protein fold diversity. After this, alpha folds continued to fold faster, but beta folds did not. Reproduced from: Debès C, Wang M, Caetano-Anollés G, Gräter F. Evolutionary Optimization of Protein Folding. *PLoS Comput Biol* 2013; 9(1): e1002861.

To achieve complex functions, proteins have evolved to work in multi-domain complexes that assemble after protein translation. Sarah Teichmann, PhD, program leader in genome evolution at the EMBL-European Bioinformatics Institute and Wellcome Trust Sanger Institute in the United Kingdom, wondered if the order of assembly is under selective pressure.

To test this theory, her team first developed a mathematical model that predicts the order in which protein complexes assemble based on 3-D structures and the surface area at the interfaces of different subunits. Then they looked for gene fusion events between genes encoding different subunits of the same protein complex. A gene fusion occurs when separate genes are shuffled into the same open reading frame, and thus become translated together in the order in which they appear in the genome. Teichmann reasoned that if the order of assembly of protein complexes is under selection pressure, then only certain gene fusions—those that preserve this order—would be favored in evolution.

“The neat computational trick here is that we are combining the structural bioinformatics with genomics. We go from the 3-D protein level to the 2-D genomic arrangement,” she says.

Indeed, she showed that fusion events that preserve the mathematically predicted order of assembly appeared statistically more frequently in the genome than those that did not. The results were published in *Cell* in 2013. “It’s intuitive in the sense that you want to have the subunits of a protein complex find each other quickly; you don’t want to have them floating around the cell in an unbound state for a long time,” she says. Unbound proteins could aggregate and cause disease.

Why did introns evolve?

One of evolution’s biggest puzzles is the intron. These extra pieces of DNA interrupt genes and have

to be spliced out before protein translation. When, why, and how did they evolve in the history of genes?

The question of “when?” has largely been solved, says Scott Roy, PhD, assistant professor of cell and molecular biology at the University of California, San Francisco. Though a few reputable naysayers argue that introns are as old as the genetic code itself (and helped make genes possible), “the consensus perspective is that a large number of introns arose for the first time in the last common ancestor of all eukaryotes,” Roy says. This would be about 1.5 billion years ago.

More perplexing is the why question. In higher eukaryotes such as humans, introns help create protein diversity through alternative splicing to produce more than one protein from a gene sequence. But until recently, scientists believed that alternative splicing was rare in lower eukaryotes and thus couldn’t be their *raison d’être*. “That turns out to be at best a gross simplification and in some cases just completely wrong,” Roy says. For example, recent microarray analyses showed that almost all of yeast’s 200 intron-containing genes are alternatively spliced, Roy says.

His team is hunting for examples of functional, evolutionarily conserved alternative splicing in fungi. Functionally important variants may represent only a fraction of transcripts, “so you have to sequence the heck out of the transcriptome,” Roy says. Analyzing the data is a major computational challenge because the transcripts have already had the introns removed, and the algorithm has to guess where these splicing events happened. “You get these short reads—about 100 nucleotides. And then you have this huge genome and you need to figure out where does this 100 nucleotide read come from in the genome,” Roy says. “There are a lot of programs out there that do it, but they’re not very consistent.” His team uses multiple programs as well as in-house software to arrive at a consensus.

They have found some alternative splicing events that appear to be conserved over long timescales and in different species; but “it remains to be seen whether it’s true conservation or just coincidence,” he says. “I don’t even know where my money is at this point. Which is exciting, actually,” he says.

The purpose of introns may also be related to the 3-D genomic architecture of eukaryotes, says Liya Wang, PhD, a research scientist at Cold Spring Harbor Laboratory. In eukaryotes, DNA is organized into nucleosomes: 140-base-pair stretches of DNA are coiled around proteins called histones. The DNA coiled around a histone is more likely to be an exon than an intron, suggesting that this 3-D structure helps to prevent introns from interrupting a functional stretch of DNA, Wang explains.

To study the mechanisms of intron gain and loss, he and Lincoln D. Stein, PhD, program director of informatics and bio-computing at the Ontario Institute for Cancer Research and a professor at Cold Spring Harbor, came up with a computational model that could recreate the distribution of exon sizes for the genomes of 14 different species. Surprisingly, their model predicted that the probability that an

exon will gain an intron is proportional to its size to the third power, suggesting a 3-D volumetric relationship rather than one based just on sequence. “One hypothesis is when the introns try to attack, they are attacking a ball that the exon occupies by its dynamic motion; the larger the ball, the higher the chance,” he says. The results were published in *BMC Evolutionary Biology* in 2013.

Wang and Stein are now modeling whether CG content (the frequency of cytosine/guanine nucleotide pairs, which is related to methylation), also affects intron insertion. Their work reflects a growing recognition of the importance of higher-order features, such as epigenetics and morphology, in shaping evolution.

What drives the evolution of form?

The first multicellular organisms appeared about 565 million years ago, followed by an abrupt explosion of body plans from about 550 to 530 million years ago (visible in the fossil record). Nearly all modern shapes appeared then; and there have been few innovations since. This observation has long puzzled scientists; how could gradual, Darwinian evolution result in such rapid changes in form?

Stuart Newman, PhD, professor of cell biology and anatomy at New York Medical College, believes that the answer lies in physics. In a 2012 paper in *Science*, Newman argues that genes that evolved for other purposes in unicellular organisms (such as those for adhesion), suddenly found new roles in the physical landscape of multicellular organisms. “You have a way through physics of generating radically new forms by very small genetic changes,” he says.

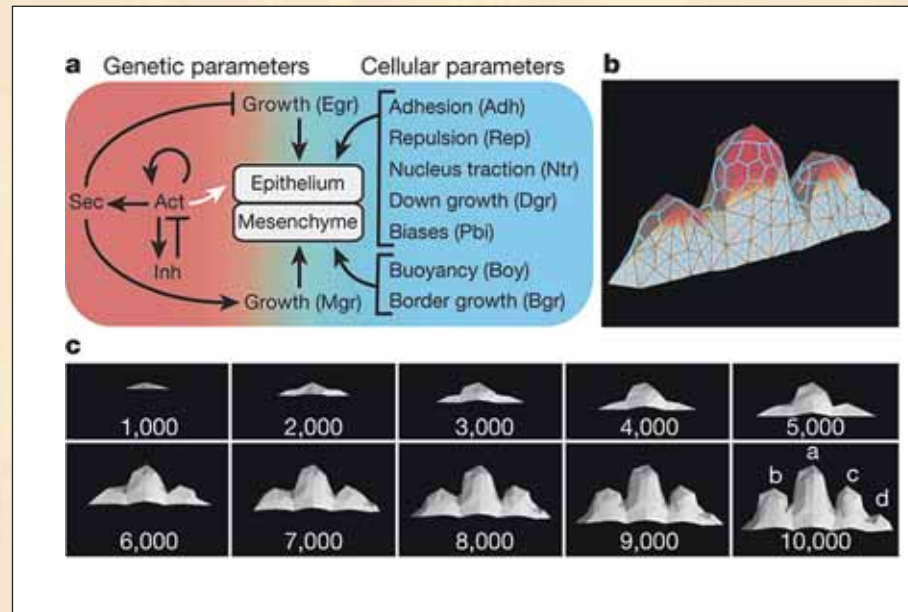
“If you look at the logic of the Darwinian perspective, it says you can’t have abrupt change. But this is a 19th century view. We now know with 20th century advances in the physics of materials that things like tissue masses can change abruptly and discontinuously,” he says. Physical laws also limit what morphological motifs are possible, which explains why there’s been little diversification in form in the past half billion years.

Newman’s team simulates limb development using a finite-element model. When they virtually evolve limbs, they end up with a variety of shapes that never existed in any animals, but that still resemble natural limbs. “So there’s both a great plasticity but then there’s also a constraint in that. With the Darwinian paradigm you can in principle get from anywhere to anywhere by adaptation, but this kind of mathematical modeling approach shows that there are really deep constraints in the kinds of forms you can come up with. You can’t get just anything.”

Isaac Salazar-Ciudad, PhD, a senior researcher at the University of Helsinki and the Autonomous University of Barcelona in Spain, also looks beyond Darwin to study the evolution of form. His team has developed a computational model of tooth development. “We have a set of cells and those cells have genes inside; those genes affect each other in gene

networks,” Salazar-Ciudad says. “Then at the same time, those cells are actually moving and interacting mechanically with each other.” This is one of the first models to combine these two components, he says.

In a 2013 paper in *Nature*, Salazar-Ciudad used his model to explore the relationship between genotype and phenotype in the evolution of morphology. He

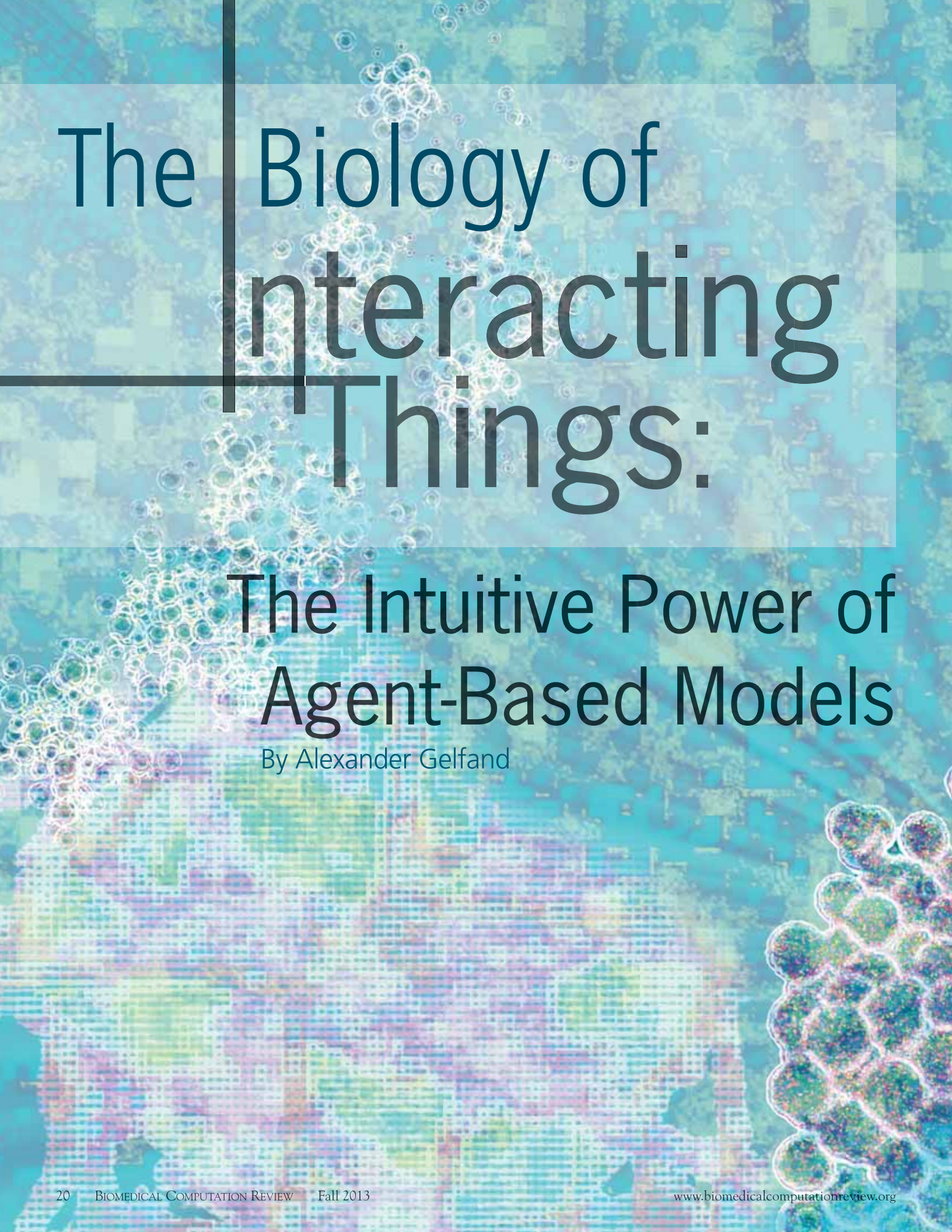


A Model with Teeth. Isaac Salazar-Ciudad’s team created a morphological model of seal tooth development and evolution. Panel (a) shows the cellular and genetic parameters included in the model. Panel (b) shows how tissue morphology is modeled in three dimensions; cells are allowed to move and interact with each other, creating shape. Panel (c) shows how the tooth shape evolves from the initial conditions until 10,000 time points. Salazar-Ciudad uses the model to study the evolution of teeth as well as their development. Reproduced with permission from: Salazar-Ciudad I, Jernvall J. A computational model of teeth and the developmental origins of morphological variation. *Nature* 2010; 464: 583-586.

virtually evolved teeth by gradually mutating them, and then explored the resulting 3-D phenotypes. “We found that the mapping between genotype and phenotype is so complex that natural selection cannot fine tune every aspect of morphology,” he says. “We say that natural selection is indeed acting all the time and it is very important, but there is a restriction on what kinds of things it can do.”

And more puzzles remain...


History-of-life puzzles spur passionate debate precisely because the scientific questions are so tied to existential ones—who we are, where we came from, why we’re here. But answering these questions isn’t just about satisfying deep-seated human curiosity; it’s also about practical ends. “Obviously there’s just a big curiosity behind it. People want to know where did we come from, where did it all start?” Hordijk says. “But, besides that, I think great medical things will come out of this. If we understand how life started, that automatically gives us a better understanding of how life works. That will certainly have a lot of important medical implications.” □



The Biology of Interacting Things:

The Intuitive Power of Agent-Based Models

By Alexander Gelfand



In the early 1990s, when **James A. Glazier, PhD**, first became interested in using agent-based modeling to simulate biological phenomena, the field was so new that he had to borrow ideas from the study of metal and soap.

Times have changed: Over the last 10 years, agent-based models (ABMs) have become an important component of the biomedical researcher's toolkit.

By their nature, ABMs would seem to be a perfect fit for biology. First developed in the 1940s, they simulate complex systems by having autonomous virtual agents (cells, anyone?) interact with each other and their environment according to preprogrammed rules, often with a degree of built-in randomness. Yet when agent-based programming languages and modeling software came along in the 1990s, they were slow to gain traction in biomedical circles. Representing the many cells in a biological system using ABMs can be expensive compared to running biological simulations based on differential equations, and the mathematical techniques used to analyze and optimize equation-based models do not necessarily work on ABMs.

Yet ABMs do carry advantages for biomedical research. Among other things, they are intuitive, work well in three dimensions, and can reproduce complex behaviors with just a few simple (even incomplete) rules. Moreover, progress toward hybridizing ABMs with other approaches, such as differential equations, is making them more powerful than ever. These pluses, along with increased computing power, are helping biomedical applications of ABMs take off, as scientists use them to investigate everything from tumor formation to bacterial growth.

Soap to Cells

As a doctoral candidate in physics in the late 1980s, Glazier, who now directs the Biocomplexity Institute at Indiana University Bloomington, studied the evolution of bubbles in soap froth. The surprisingly broad implications of that work led him to collaborate with a group of researchers at Exxon who were using computational models derived from statistical physics to investigate the related phenomenon of grain formation and growth in metals. A few years later, while working as a post-doctoral fellow in Sendai, Japan, in the laboratory of **Yasuji Sawada, PhD**, Glazier met **François Graner, PhD**, who was studying the microscopic fresh-water creatures known as hydra. Hydra are renowned for their regenerative capabilities—chop them into hamburger, and the cells will rearrange themselves to form a whole

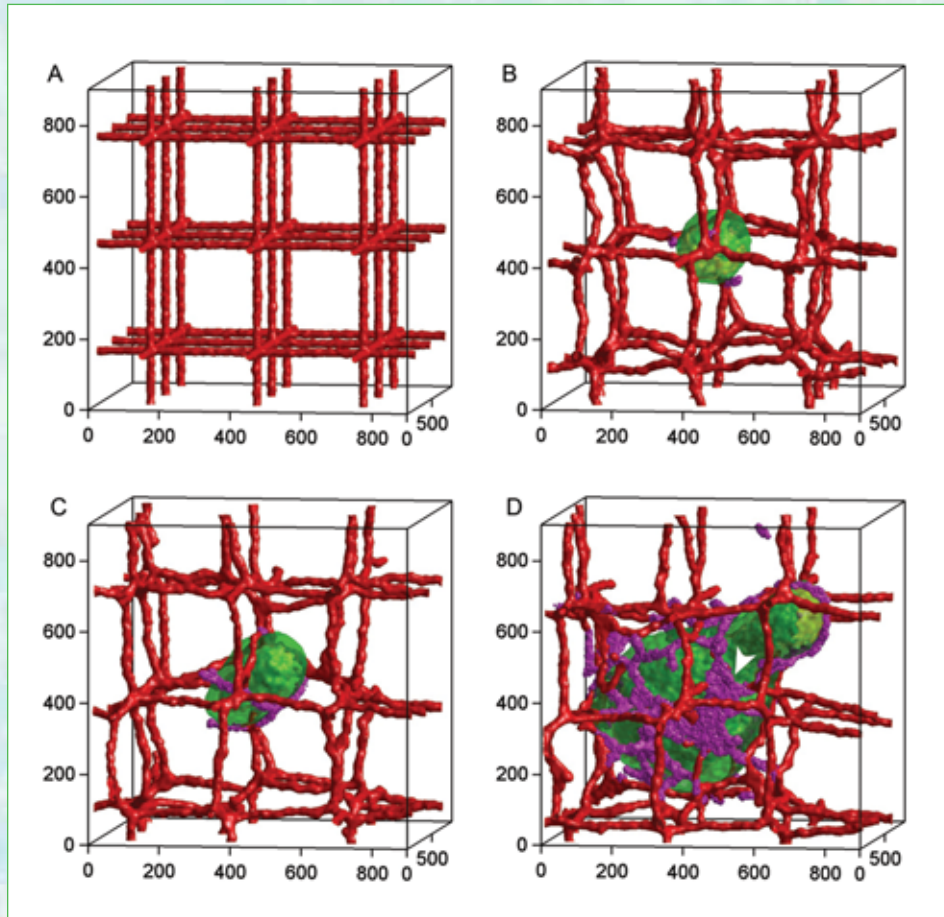
new organism—and Graner wanted to test the hypothesis that cell adhesion allowed a hydra’s two main cell types to sort themselves into larger structures during regeneration. Glazier realized that a modified form of the model that he and the Exxon researchers had been using could also simulate cell-sorting by treating each cell as an individual unit subject to basic physical

for biological purposes. Over the past two decades, scientists have used increasingly sophisticated forms of it to simulate multicell phenomena as diverse as wound healing, stem-cell differentiation, and skin pigmentation. Often, they have relied on CompuCell3D, an open-source modeling environment that Glazier and his collaborators **Mark Alber, PhD**, and **Jesus Iza-**

equations to describe the chemical fields that influence cell migration and differentiation, or ones that use ordinary differential equations to describe the dynamics of biochemical networks inside cells and the distribution of chemicals at the whole-body level. (Such hybrid models, which combine agent-based and equation-based methods for greater efficiency and multi-scale capability, are becoming increasingly popular.) Once users have made their selections, the software generates draft code that can be manually edited.

Despite its user-friendly interface, some very sophisticated computation is taking place under the hood. Most cell properties, behaviors, and interactions are bundled together in a single function, called the “effective energy” (the terminology harks back to the model’s roots in physics), which incorporates all of the forces acting on whatever agents are being simulated—cells, parts of cells, environmental features—and the rules that govern how they will respond. The cells live in regular 2-D or 3-D lattices, like pixels in a digital microscope image—a feature that links the GGH model to simpler cellular automata that represent cells as points on grids. Unlike automata, however, cells in the GGH model have volume, are deformable, and are affected not only by their immediate neighbors but by a host of other factors. They can also move about in three dimensions, providing a degree of spatial realism that is extremely valuable for tissue simulations.

The GGH model is also inherently stochastic: cells move about by randomly exploring their environment, responding to whatever forces have been programmed into the simulation, and moving on average towards a state of least energy. That randomness, says Glazier, is what gives cells the freedom to reorganize themselves. It also gives rise to very complex and even unexpected aggregate behaviors—behaviors that could not necessarily be predicted from the underlying rules. This quality, known as emergence, is both the hallmark of agent-based modeling and the secret to its success. “That complexity is the reason that this kind of modeling works,” says



Glazier and his colleagues have used ABMs to simulate sprouting angiogenesis as shown here, where an initial cluster of adhering endothelial cells forms a capillary-like network over the course of 18 hours ((A) 0 h; (B) ~2 h; (C) ~5 h; (D): ~18 h). Reprinted from Shirinifard A, et al., 3D multi-cell simulation of tumor growth and angiogenesis. PLoS ONE. 2009; 4:e7190.

forces and constrained by a few rules. **Paulien Hogeweg, PhD**, a Dutch theoretical biologist at the University of Utrecht who helped coin the term “bioinformatics” in the 1950s, later elaborated on Glazier and Graner’s initial modeling efforts, adding biological mechanisms like cell dif-

guirre, PhD, at the University of Notre Dame began developing in 2000 and whose development is currently led by **Maciej Swat, PhD**, at Indiana University.

CompuCell3D is meant to help researchers concentrate on the biology behind their simulations rather than on the

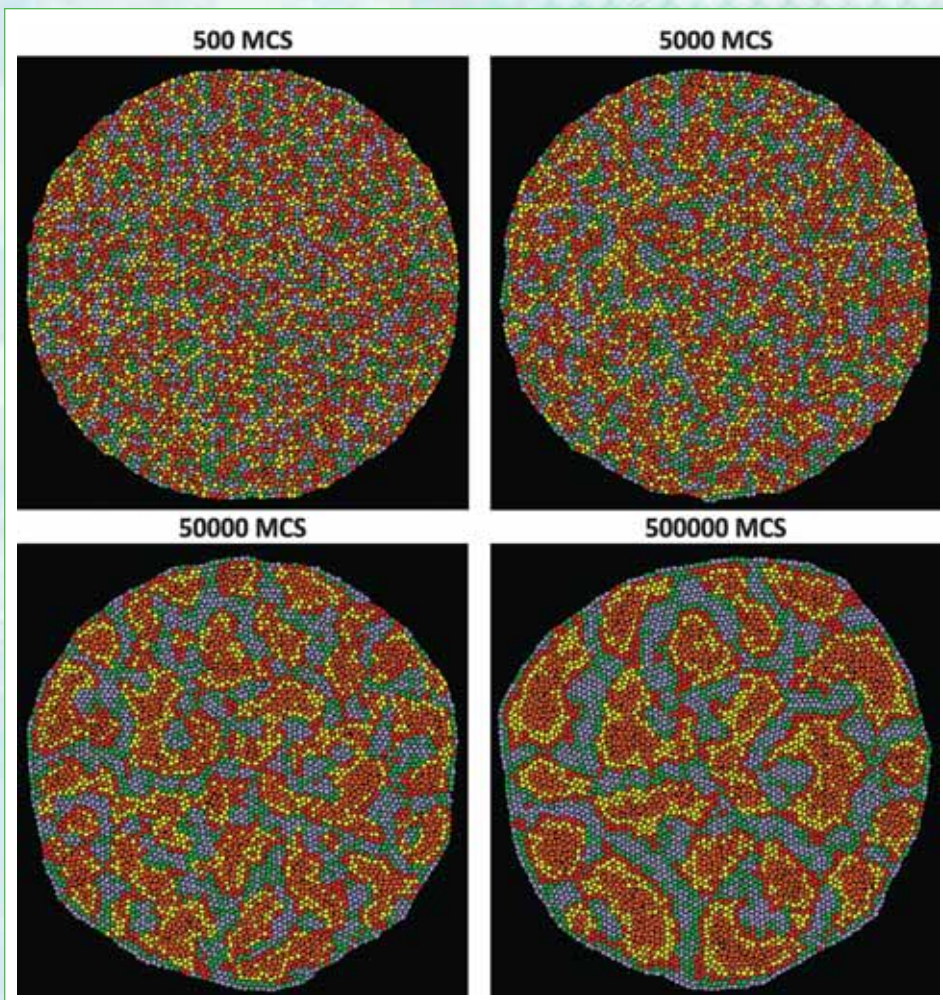
Emergence is both the hallmark of agent-based modeling and the secret to its success.

ferentiation and chemotaxis to create what became known as the Glazier-Graner-Hogeweg (GGH) model.

The GGH model was one of the first agent-based models designed specifically

nuts and bolts of model building. To that end, the software allows users to select the cell types and behaviors they want from a series of drop-down menus. It also lets them add modules that use partial differential

Glazier, who adds that modelers can simplify many of the rules governing the individual agents in an ABM and still generate realistic global behaviors, so long as they include the key biological mechanisms—



CompuCell3D can be used to simulate cell sorting using various rules for cell adhesion. These snapshots show the dynamics of cluster formation during a 5000-cell aggregate simulation with five different levels of cadherins (represented by the five different colored cells). Reprinted from Zhang Y, et al., Computer Simulations of Cell Sorting Due to Differential Adhesion. PLoS One 2011; 6(10):e24999. doi: 10.1371/journal.pone.0024999.

an especially handy trick in cases where quantitative data (e.g., rate constants, physical forces) remain spotty, if only because “no one thought to try to measure it.”

Emergent Rules

According to **Gary An, MD**, the ability to generate complex emergent behaviors in the absence of comprehensively detailed knowledge makes agent-based modeling ideally suited to testing hypotheses and conducting *in silico* trials.

An, associate professor of surgery at the University of Chicago, first came to agent-based modeling while working as a trauma surgeon at Cook County Hospital in the 1990s. Frustrated by the lack of medications he and his colleagues had for treating sepsis, a potentially fatal condition that occurs when the immune system’s own response to injury or infection triggers inflammation throughout the body, An began building agent-based models of sep-

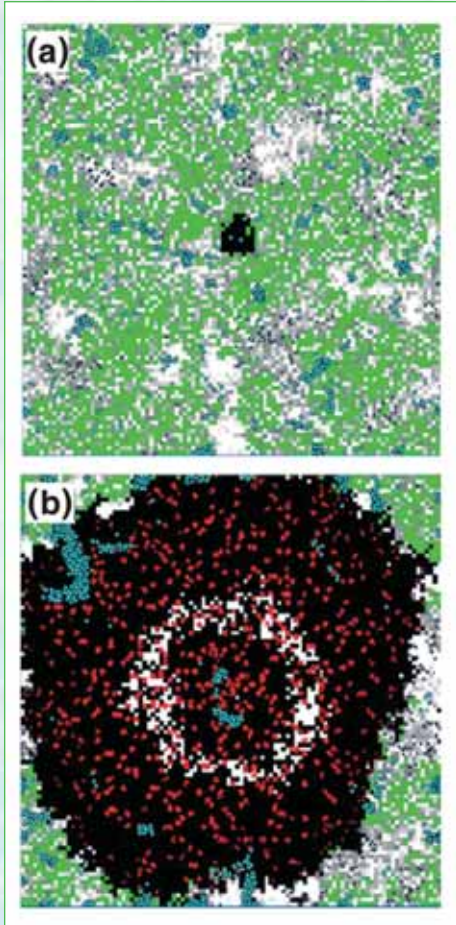
sis using SWARM, a software platform for multi-agent simulations of complex systems developed by the Santa Fe Institute. Since then, he has continued to use ABMs to investigate acute inflammation, often in collaboration with his friend and colleague **Yoram Vodovotz, PhD**, an immunologist and professor of surgery at the University of Pittsburgh. He has also helped others apply similar models in their own research.

An was first attracted to ABMs because he found them to be more intuitive than equation-based models. “I wasn’t a math guy,” he says. “I didn’t think in terms of differential equations and calculus. I thought in terms of things doing things”—i.e., cells interacting with other cells—“and things doing things is agent-based modeling.” But he has come to appreciate ABMs as tools for dynamically embodying what we know (or think we know) about biological systems and processes, and as platforms for testing hypotheses that can yield unexpected insights

“I didn’t think in terms of differential equations and calculus. I thought in terms of things doing things,” An says. “And things doing things is agent-based modeling.”

into biomedical problems.

Recently, An helped a group of researchers at the University of Chicago build an agent-based model—the Ductal Epithelium Agent-Based Model (DEABM)—to simulate how cancerous tumors form in



*This ABM, created using SPARK, simulates how liver inflammation caused by a hepatitis C infection accelerates the progress of tumor formation from a patch of a few hypothetical cancer stem cells in black (a) to the formation of a hypoxic core in the center of a growing tumor (b), as well as tumor angiogenesis (red dots). Reprinted with permission from An, G, et al., *Agent-based models in translational systems biology*, Wiley Interdiscip Rev Syst Biol Med. 2009; 1(2): 159–171. doi:10.1002/wsbm.45.*

breast tissue. The agents in the model consisted of the various cell types found in the mammary duct epithelium (luminal and myoepithelial cells, fibroblasts, stem and progenitor cells), all of them programmed with rules defining how they grow and differentiate, mutate and die. Drawing on data from textbooks and review articles, the Chicago group also equipped their agent-cells with variables representing internal, molecular-level components, including seven genes known to play a role in both cell function and tumor formation. The simulations used

three virtual populations of 500 individuals and ran for 15,000 time steps, corresponding to approximately 40 years. Genetic mutations were allowed to accumulate over time, ultimately impairing cell function and resulting in cancer.

The first version of the model did a good job of accurately simulating normal cell population dynamics and breast physiology. But it could not generate estrogen receptor-positive (ER+) tumors, which are in fact the most common kind. This “huge fail,” An says, not only indicated a serious flaw in the model (because the rules governing the agents were based on the best available knowledge concerning breast cancer), but also pointed to a serious gap in researchers’ understanding of the pathogenesis of ER+ tumors.

The clue to solving this mystery lay in the model itself. Since ER+ cells are normally prevented from proliferating by the suppression of the receptor c-Met, the agent rules specified that ER+ cells were not allowed to divide. That, in turn, meant that mutations to the cells couldn’t accumulate to be passed on to future generations and lead to cancer. So An and his colleagues began looking for something that would allow ER+ cells to proliferate—something that would ordinarily be responsible for suppressing c-Met, but that could be impaired. A literature search identified the gene RUNX3 as a possible candidate; and once it was incorporated into the model and permitted to mutate, ER+ cells acquired the capacity to replicate and accumulate damage, resulting in the appearance of ER+ tumors.

The discovery that RUNX3 might play a role in breast cancer by regulating ER+ cell proliferation could be clinically useful. For example, An raises the prospect of one day screening for decreased expression of RUNX3 as a warning sign of increased risk for ER+ tumors. But the discovery process also highlights one of the advantages of agent-based modeling. An equation-based model, An says, might simply have been designed to reproduce the rates of ER+ tumor occurrence seen in the real world, and would therefore have masked the underlying mechanism. The agents in the DEABM, however, could not reproduce those rates without having the proper mechanism written into their rules in the first place—making the absence of that mechanism painfully clear. As Glazier says, “An agent-based model is constructive—it

includes only what you put in. If you leave out a key mechanism, you will never replicate the biology.”

“An agent-based model is constructive—it includes only what you put in. If you leave out a key mechanism, you will never replicate the biology,” Glazier says.

If It Grows Like Skin, and It Looks Like Skin...

Even when the underlying rules for a model are incomplete, researchers can use ABMs to test hypotheses “before killing rats or growing cells,” An says. Robert Isfort, PhD, and his colleagues at Procter & Gamble have made the most of this capability. Working together with researchers at the University of Sheffield in England, the Procter & Gamble group employed agent-based modeling to test no less than three competing theories of how epidermal tissue maintains and renews itself over time. In the process, they resolved a central question in skin biology and helped advance the field of stem cell research.

According to the oldest hypothesis, known as asymmetric division, stem epithelial cells drive epidermal regeneration by dividing either to form new stem cells, or to form progenitor cells that go on to produce the differentiated progeny that make up the outer layers of the skin. Another, more recent hypothesis, known as population asymmetry, holds that progenitor cells are primarily responsible for skin renewal through stochastic differentiation, with stem cells playing only a secondary role. The third and latest hypothesis, population asymmetry with stem cells (PAS), contends that stochastic differentiation of both stem cells and progenitor cells is required to maintain and regenerate skin tissue. With experimental data to support all three, the question remained: which hypothesis was correct?

Using a modified form of a human skin model that the Sheffield group had developed with an agent-based modeling platform called FLAME, the international team

of researchers translated all three hypotheses into separate, stochastic ABMs, each with slightly different rules for cell division and differentiation; the probability of stem cell division, for instance, changed from model to model. They then ran each simulation for the equivalent of three years. In addition, the virtual skin in the PAS model was wounded at the three-year mark, and the simulation was run for the equivalent of an additional year to see how it would respond.

Although the physical forces acting on

differentiation produced strikingly different outcomes. Most surprisingly, says Isfort, the models derived from the first two hypotheses were unable to produce colonies of mother and daughter cells that behaved realistically over the long term. Consequently, while all three models yielded mature epidermal layers with similar cellular organization after three years, only the model instantiating the PAS hypothesis, according to which both stem and progenitor cells divide and differentiate stochastically, was able to generate tissue

been wounded, or suffered damage through the normal aging process.

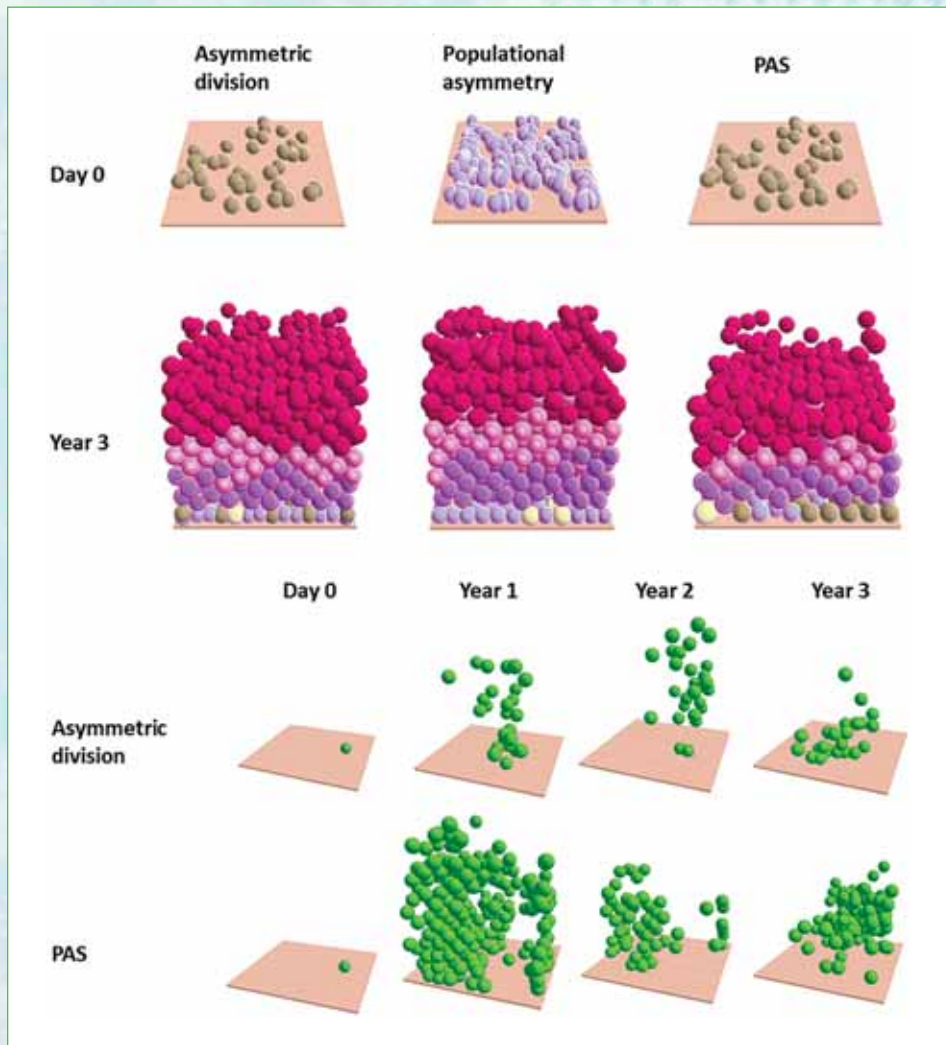
Hybrid Vigor

Despite its strengths, agent-based modeling can, at times, be slower and less efficient than equation-based models. That is why some researchers are creating hybrid approaches that combine the two methods.

Yoram Vodovotz, who has co-authored a number of papers on agent-based modeling with Gary An, says that while ABMs can often be assembled more quickly than equation-based models, their stochastic and emergent properties sometimes make it difficult to relate outcomes to specific causal factors. In 2004, for example, Vodovotz and An both published papers in *Critical Care Medicine* on sepsis, with Vodovotz simulating a population of patients using a deterministic model based on ordinary differential equations, and An using an ABM. In that case, says Vodovotz, the mathematical model allowed him to trace individual patient outcomes to particular configurations—to say that patient X, for example, died because of a specific pathogen load, or a genetic predisposition to acute inflammatory response—whereas the ABM could only indicate that a certain percentage of virtual patients hadn't responded to treatment, without revealing precisely why.

Emergence also makes it harder to set and optimize parameters in ABMs than in mathematical models. In an equation-based model, for example, the modeler can simply program parameters like rate constants, which characterize the rates of biochemical reactions in a system. In an ABM, however, rate constants must emerge from the individual interactions of the agents; one must run the simulation first, then measure the rate constants and tweak the model if the numbers don't match experimental data. Vodovotz says that this makes parameter optimization in ABMs “very nontrivial,” adding: “It's one of those grand challenge-ish types of problems.”

Moreover, while equation-based models can be analyzed using well-established mathematical techniques, the complex patterns that emerge from ABMs can be difficult to quantify and analyze with the same degree of rigor. And while agent-based methods are very good at simulating local interactions between heterogeneous populations of cells at multiple scales and with a high degree of spatial realism—e.g., simulating how different kinds of cells migrate from place to place, adhere to one another, and arrange themselves in the macroscopic patterns found in real-life tissues—differential equations provide a



*Isfort's team used an ABM platform called FLAME to simulate the growing epidermis (skin) from an initial seeding with just a few cells to maturity (at year three) under three hypothetical cell division scenarios—*asymmetric division*, populational asymmetry, and a combination called PAS. The mature epidermis looked approximately the same (top), but that similarity belied some significant underlying differences. For example, when the ABM followed the division and movement of a single stem cell in the asymmetric division case, the offspring formed a column around the stem cell with several progeny lingering in the basal compartment. In contrast, individual colony shape changed dramatically in the PAS hypothesis, with substantial lateral movements. Reprinted with permission from Li X, et al.,* Skin stem cell hypotheses and long term clone survival—explored using agent-based modeling*, Sci Rep. 3:1904 (2013).*

the cells as they adhered to the lower level of the epidermis or migrated to the upper surface of the skin remained the same in all three models, variations in cell division and

that acted “like the real stuff.” In addition to fueling future experimental research on stem and progenitor cells, this work could lead to new therapies for repairing skin that has

more efficient and less expensive way of modeling well-mixed systems and other phenomena that can be adequately represented at the continuum level, such as blood flow.

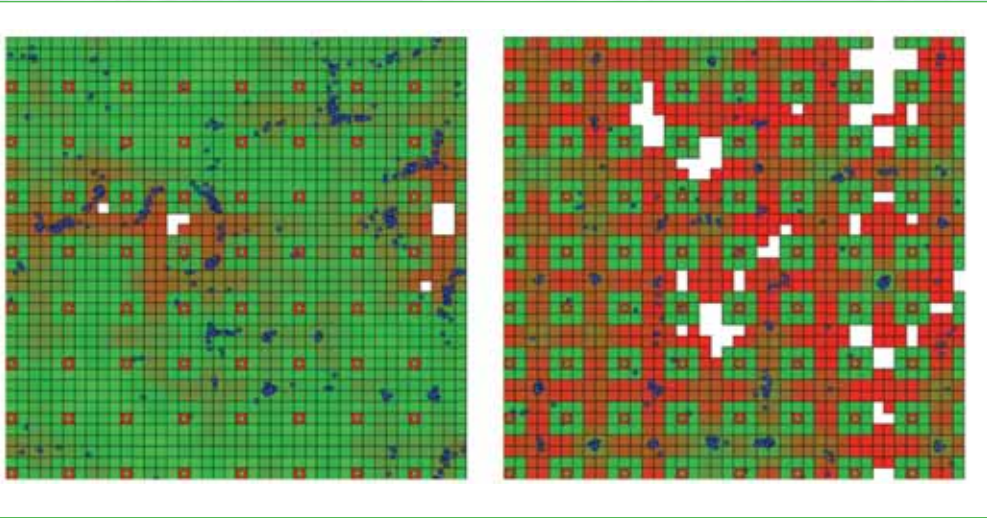
Just as Glazier and his team at Indiana University have gradually expanded Com-

techniques, and have used it to demonstrate how basic inflammatory mechanisms can lead to both positive and negative outcomes in various kinds of tissue. In a paper published in May 2013 in *PLoS Computational Biology*, for example, Vodovotz employed a

pro- and anti-inflammatory cytokines) that are involved in the formation of pressure ulcers. According to the rules governing the agents in the ABM, damaged epithelial cells released inflammatory cytokines that caused further damage; but they could also be healed by anti-inflammatory cytokines at a rate that depended on the amount of oxygen delivered by the blood. The simulation produced realistic-looking pressure ulcers at rates suggesting that people with spinal cord injuries are more likely to form them than people without such injuries, perhaps due to changes in vascularity—a finding that could lead to tools for predicting the risk of ulcer formation based on non-invasive measurements of blood flow.

Christian Jacob, PhD, a computer scientist at the University of Calgary who has used ABMs to simulate everything from ant colonies to traffic congestion, has also built a platform for constructing hybrid models, albeit one that takes the whole body as its canvas. Developed by PhD student **Tim Davison** and other graduate students in Jacob's Evolutionary & Swarm Design Lab, the software suite, which goes by the name **LINDSAY Composer**, can be used to create interactive 3-D simulations and visualizations of human physiological processes across multiple scales, from systems and organs to cells and sub-cellular structures. Users can drag and drop objects into their simulations from a component library that contains templates for various agents (e.g., cells, pathogens), all of which come with their own customizable sets of properties and interaction rules. Like SPARK and CompuCell3D, LINDSAY Composer can also combine mathematical and agent-based models—feeding data, for instance, from a mathematical model of molecular concentration gradients to an agent-based model of cell development. Jacob's ultimate goal is to create a comprehensive 3-D interactive model of human anatomy and physiology, called **LINDSAY Virtual Human**, which will enable users to zoom seamlessly from the whole-body scale right down to the molecular level for both medical education and research purposes.

Jacob's introduction to agent-based modeling came during the 1980s when he first encountered computer simulations of flocking birds. He was immediately impressed with the method's capacity to handle mixed populations of agents in three-dimensional space—a capacity that proved crucial to a project that Jacob and



*In a hybrid model incorporating both differential equations of tissue ischemia and an ABM of stochastic pressure ulcer formation in healthy controls (left) and people with spinal cord injuries (right), Isfort and his colleagues predicted that, as expected, the latter population is more prone to pressure ulcers. In these simulation snapshots after 2000 steps, green squares represent healthy epithelial cells, red squares represent damaged epithelial cells, red circles represent blood vessels, blue circles represent macrophages, and white squares represent dead cells. Reprinted from Solovyev A, Mi Q, Tzen Y-T, Brienza D, Vodovotz Y (2013) Hybrid Equation/Agent-Based Model of Ischemia-Induced Hyperemia and Pressure Ulcer Formation Predicts Greater Propensity to Ulcerate in Subjects with Spinal Cord Injury. *PLoS Comput Biol* 9(5): e1003070. doi:10.1371/journal.pcbi.1003070.5*

puCell3D to incorporate equation-based models of the biochemical pathways inside cells and the chemical flows between organs and tissues, Vodovotz has also been building hybrid models that offer the best of both worlds. He and his colleagues at the University of Pittsburgh have developed an open-source software package called SPARK (Simple Platform for Agent-based Representation of Knowledge) that can integrate mathematical and agent-based modeling

hybrid model to simulate the formation of pressure ulcers, or bedsores, on the skin of patients with spinal cord injuries, a common and potentially life-threatening occurrence.

The model used ordinary differential equations to simulate blood flow in the skin based on non-invasive measurements taken from injured individuals and an uninjured control group; and a stochastic ABM to simulate the blood vessels, cells, and signaling molecules (epithelial cells and macrophages;

ABMs for Biomedicine

Many possible ABM software programs exist (including NetLogo, which might be best for the ABM novice; <http://ccl.northwestern.edu/netlogo/>), but the five listed below are featured in this story. Although they all accomplish somewhat the same thing, they were developed using different programming languages, possess varying levels of support and documentation, and have been used to build different models.

Curious investigators are encouraged to visit their respective websites to learn more about them.

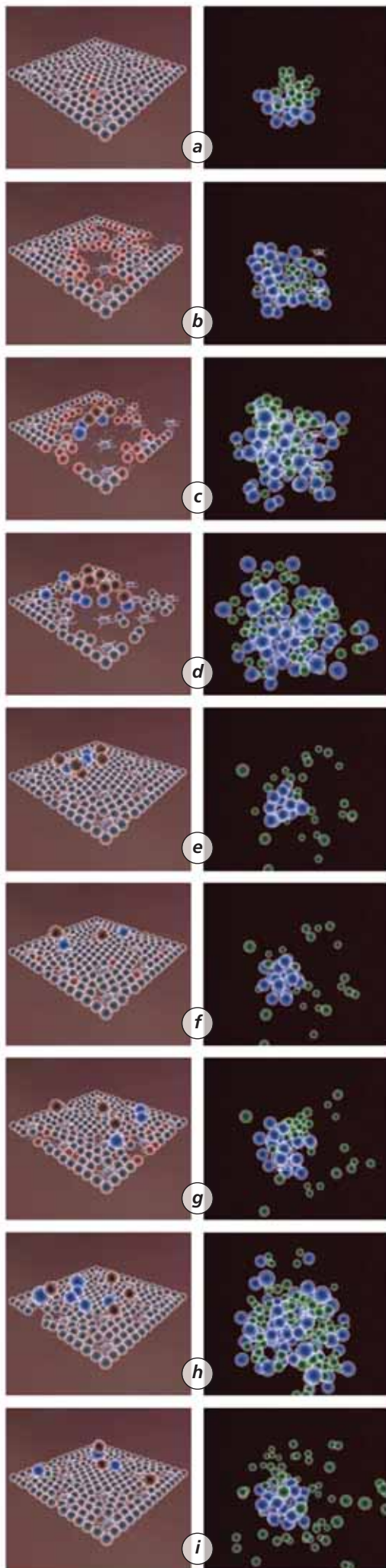
CompuCell3D: [www.CompuCell3D.org](http://www.compuCell3d.org)

SWARM: www.swarm.org

FLAME: <http://www.flame.ac.uk>

SPARK: <http://www.pitt.edu/~cirm/spark/>

LINDSAY Composer:
<http://lindsayvirtualhuman.org/?q=node/59>



a former student, **Vladimir Sarpe**, MSc, recently undertook using LINDSAY Composer.

In a paper published in *BMC Bioinformatics*, Jacob and Sarpe describe how they used a three-dimensional ABM of the human immune system to simulate and visualize the body's response to influenza A virus, from the initial infection of epithelial cells in the lungs to the destruction of the virus by lymphocytes. The model included such agents as T cells, B cells, viruses, and antibodies that were programmed to interact according to various rules in two distinct 3-D environments: within the lung tissue, and inside a lymph node. From a computational perspective, each environment was treated separately—the lymph node and lung tissue simulations were in fact executed on different computing nodes—but they communicated with one another via “controllers” that shared information as necessary. A dendritic cell in the lung that encountered a virus, for example, would engulf the pathogen and transport it to the lymph node to activate the T and B cells. They in turn would produce killer T cells and antibodies that would travel back to the lung tissue in order to neutralize

*In Sarpe and Jacobs' simulation of an immune response to influenza A infection over time, agent interactions occur in both lung tissue (left column) and lymph node (right column). Initially, when the virus infects the lung cells (a) (with red cells representing infected cells) there is not yet any immune activity in the lymph node (b). As the infection progresses, the immune response can be observed both in the tissue (cell-mediated) and in the lymph node (humoral) (c). Eventually, the initial infection is eliminated (d,e) and the simulation reaches a steady state (f). Upon reintroduction of the virus (g), the immune reaction is bigger and faster (h,i) because the immune system remembers the virus. Reprinted from Vladimir Sarpe and Christian Jacob, *Simulating the decentralized processes of the human immune system in a virtual anatomy model*, *BMC Bioinformatics* 14(Suppl 6):S2 (2013).*

the virus and destroy the infected epithelial cells. The simulation even generated “memory” T and B cells that stuck around after the initial infection to enable a faster response upon subsequent exposure to the virus.

Getting ABMs Into More Hands

The high computational overhead incurred by ABMs remains a challenge. In the case of their immune-system simulation, Jacob and Sarpe sidestepped the issue by relying on a relatively small number of

The high computational overhead incurred by ABMs remains a challenge.

agents—a few thousand, far less than the actual number of cells and viruses that would really be involved—and used probabilities (of becoming infected, of releasing antibodies, of reproducing) to generate the kinds of emergent behaviors that would arise with more realistic numbers of moving parts. As a result, the model was able to produce outcomes that accorded both with clinical data, and with the results of a robust equation-based model.

That approach might not always be ideal, however; so for Jacob, driving down the computational expense of agent-based modeling has become an area of research unto itself. In a paper published this year in the journal *Simulation*, he and his colleagues reported that they were able to reduce the number of agents in a simulation by creating so-called “observers” that recognized patterns in the behaviors of groups of agents, and replaced those groups with single meta-agents that subsumed their behaviors. When applied to a blood-clotting simulation in which 12 different blood factors were represented as agents, average run-time was cut almost in half.

By making agent-based modeling more affordable, such advances could also help put ABMs into the hands of more scientists. And that would be good news for biomedical researchers who do not necessarily know much about machine-learning algorithms or parameter optimization, but who do find it easy to grasp a modeling technique that so faithfully reproduces the kinds of objects, interactions, and behaviors that they observe in nature.

“They actually think of these agents,” Jacob says, “without knowing it.” □

BY MACIEJ SWAT AND JAMES A. GLAZIER

Agent-Based Virtual-Tissue Simulations

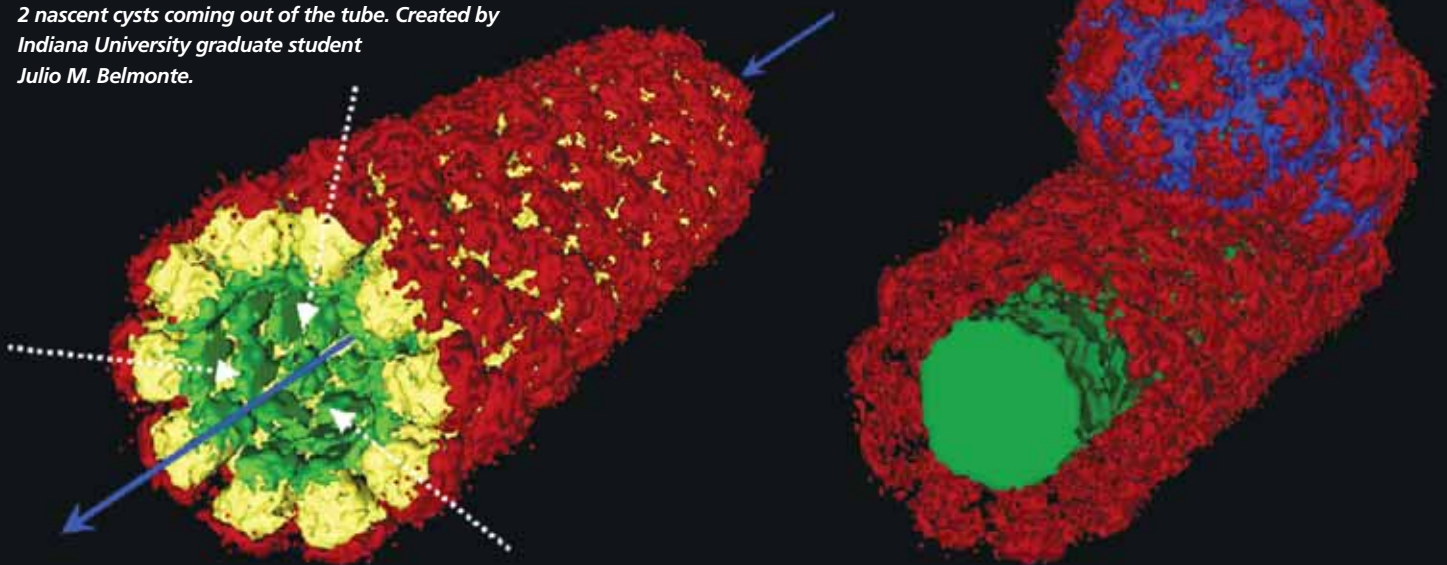


Cells have a limited repertoire of behaviors and interactions. They grow, divide, die, stick to each other, send and receive signals, change shape, polarize, differentiate (change behaviors), form sheets, secrete, absorb, pull on and remodel extracellular material, and migrate in response to signals in their environment. Despite their limits, cells nevertheless give rise to a wide range of tissue-level processes including embryonic development; wound healing; regeneration of a severed salamander limb; degenera-

These observations—that the interactions among simple cellular behaviors drive the emergent behaviors of tissues, organs and organisms—lie at the core of agent-based virtual tissues. Their empirical validity is the reason we can build predictive models using a limited number of relatively simple, universal biological mechanisms.

Agent-based models abstract key behaviors and interactions from the complexity of real biological components,

This 3-D agent-based model of the onset of polycystic kidney disease was created in CompuCell3D. It shows (left) the initial condition of the nephron and (right) a snapshot of a simulation with 2 nascent cysts coming out of the tube. Created by Indiana University graduate student Julio M. Belmonte.



tion of bone in osteoporosis; cancer metastasis; and lethal over-growth of the kidney in polycystic kidney disease.

Even the most detailed introspective examination of the properties of a single cell cannot reliably predict this variety of behaviors at the tissue- or organ-level. Moreover, cells themselves do not usually behave idiosyncratically, as the common biological definition of cell types indicates. And while the biochemical networks inside cells are capable of highly varied behavior *in principal*, the regulatory mechanisms active during particular developmental stages or diseases are often quite simple.

It turns out that emergent behaviors at the tissue level result from feedback—the way an agent (or cell) acts in response to its environment that in turn changes that environment. Indeed, it is the emergent interactions among classes of behaviors, rather than details of their control, that often leads to complexity of pattern formation.

embody them as computational agents and then run simulations to observe the emergent phenomena. They are especially useful in answering questions about the dependence of emergent properties on specific agent behaviors or environmental perturbations. For example, if we want to understand the factors determining the trajectories of birds in a flock, we can abstract the birds to motile *boids*, which attempt to maintain a fixed distance and angle with respect to their neighbors. To understand why antiangiogenic chemotherapies can lead benign tumors to metastasize, we can model tumor-cell agents that use nutrients to grow, mutate when they reproduce, and die when they starve. These agents also consume diffusible nutrients from the environment and, in the absence of sufficient oxygen supply, secrete diffusible signaling molecules to promote the proliferation of vascular endothelial cell agents, which in turn supply diffusible nutrients and oxy-



gen to their environment.

We might then compare how the velocities of flocks in a flock correlate, or how the distribution of cell motilities in the tumor changes over time for an unperturbed tumor versus one in which we temporarily kill off the nutrient-supplying vasculature. Both

simulations make useful, experimentally verifiable predictions: Increasing the inertia of the flocks causes a transition from gnat-like swarming to goose-like smooth flight, while loss of vasculature causes a pattern of nutrient deprivation which favors motile, potentially metastatic tumor-cell phenotypes at the expense of the non-motile benign phenotypes favored by a steady nutrient supply.

It is important to keep in mind, however, that we may easily overlook important mechanisms—a successful model shows sufficiency of mechanism, not necessity. As a consequence, we are most likely to identify new mechanisms when simulation results differ from experiment.

Agent-based virtual tissues come in two main types—multi-cell and continuum—that serve different purposes. Multi-cell virtual tissues are useful for examining emergent behaviors resulting from the movement and reorganization of hundreds of thousands of individual cells over volumes of cubic millimeters, as in the organization of organs in embryos. Continuum virtual tissues, in which the agents are tissue volumes aggregating the behaviors of tens of thousands to millions of cells, are useful for treating larger volumes, such as an adult heart or a multi-centimeter brain tumor. Jump-up/jump-down (or hybrid) virtual tissues combine continuum models with periodic multi-cell simulations of representative tissue-volume agents to update continuum model parameters.

The various multi-cell simulation methodologies (and there are many) trade off the level of detail per cell against the number of cells per simulation. **Cellular automata**, for example, represent cells as single, fixed lattice points, allowing the largest simulations but limiting the possible cell movements and interactions. **Center models** represent cells as point particles in 3-D space interacting via potential-energy fields, much like molecular dynamics simulations, allowing cell movement, but neglecting cell shapes. **Sub-element models** build individual cell agents out of collections of tens or hundreds of center-model subcomponents at proportionally greater computational cost. **Cellular Potts Model** (or **Glazier-Graner-Hogeweg**) stochastic models approximate complex cell shapes as collections of pixels on a regular lattice and define their behaviors and interactions through the local minimization of effective energies depending on cell and pixel configurations. And **finite element** and **immersed boundary** models allow detailed geometrical representation of the shapes and surface properties and forces of cells, at much greater computational load per cell. Ultimately, each simulation method should give the same results for the same biologically determined classes of objects, behaviors and interactions.

Until recently, coding complex virtual-tissue simulations required the creation of custom low-level computer code for each model. Now, virtual-tissue simulation environments simplify the construction, execution and analysis of agent-based models by providing libraries of cells, sub-cellular components, extra-cellular materials, intracellular biochemical networks, and fluid and diffusing chemical agents. Just as Matlab made sophisticated mathematical modeling accessible to non-specialists, domain-specific multi-cell simulation environments such as CompuCell3D, Morpheus, Simmune and CellSys democratize virtual-tissue simulations. By reducing the model-specification code from tens of thousands to hundreds of lines, these environments allow researchers to concentrate on the difficult

Just as Matlab made sophisticated mathematical modeling accessible to non-specialists, domain-specific multi-cell simulation environments such as CompuCell3D, Morpheus, Simmune and CellSys democratize virtual-tissue simulations.

problem of understanding the biology rather than on computational details. In these environments, the modeler only needs to specify high-level parameters, such as the agents and their properties and how these properties change over time; the modeling software then iteratively evaluates all of the interactions present in the current model configuration and updates the parameters of each agent.

Such agent-based simulations, like modern vital 3-D microscopy, produce cell-resolution 3-D time series results, which can then be compared against experimental results through the identification of characteristic metrics. While we are still learning how to extract biological meaning optimally from these simulations, they remain rich sources of information. □

DETAILS

Maciej Swat is an associate scientist and lead developer of CompuCell3D. James A. Glazier is professor of physics and director of the Biocomplexity Institute at Indiana University Bloomington. CompuCell3D (CC3D, www.compuCell3d.org) is an open-source, cross-platform, multi-cell simulation environment that provides a platform for compact, high-level specification of simulation agents and behaviors using predefined Python templates in a language-aware template-supporting editor (Twedit++), as well as simulation execution, visualization, post-processing and results tracking.

Stanford University
318 Campus Drive
Clark Center Room S221
Stanford, CA 94305-5444

seeing science

SeeingScience

BY KATHARINE MILLER

Digging Deep Into the Tree of Life

The awe-inspiring journey from the first cell some 3.5 billion years ago to the remarkable diversity of species we see today is now available in a tabletop display called DeepTree. “For the first time, people can explore the entire tree of life in one interactive visualiza-

tion,” says Chia Shen, PhD, senior research fellow in computer science at Harvard University’s School of Engineering and Applied Sciences.

DeepTree is part of a larger museum exhibit called *Life on Earth* that was put together by Shen’s team and is currently

stationed in four museums including the California Academy of Science in San Francisco and Chicago’s Field Museum.

To create DeepTree, Shen’s team merged vast public datasets of phylogenetic trees, common names and species images, as well as estimates for the times of divergence; selected a tree shape that would accurately reflect the way species diverge gradually over time; and studied how multiple museum visitors interact

with the displays simultaneously to enable cooperative learning.

“Our project is very carefully constructed so people can learn,” says Shen. Indeed, a research study carried out in two museum settings showed that by using DeepTree, young people have an increased understanding of common ancestry and the relatedness of diverse species.

Shen and her colleagues are also experimenting with rendering a large tree in the cloud. “Secondary school teachers are interested,” Shen says, “And we think we can do it.” □



Using the touchscreen DeepTree display, museum visitors (inset, left) can zoom through evolutionary history from its roots to fungi, plants, birds, fish (pictured), and mammals. Images courtesy of Life on Earth. For more information, visit <https://lifeonearth.seas.harvard.edu/>.