Bio·IT World
Technologies Driving Discovery, Development, and Clinical Trials

Welcome to your digital edition of *Bio-IT World* magazine!

We hope you are enjoying your subscription! As a member of the Bio-IT World community, you also have complete access to all online Archives, Buyer's Guides, Events, and Career Centers. Be sure to bookmark WWW.BIO-ITWORLD.COM for daily delivery of fresh insights and critical resources – or place Bio-IT World headlines on your desktop (or your web site) with our XML news feeds.

In addition, you can select from a wide variety of FREE enewsletters covering such targeted topics as Genomic Medicine, Drug Discovery, IT Trends, Regulatory Compliance, Technology for Healthcare and Clinical Trials. For a complete list of enewsletter titles, please go to http://www.bio-itworld.com/newsletters/sign-up.html

And as always, we welcome any and all comments or suggestions-just send us an email with DIGITAL SUBSCRIBER as the subject line.

>> CLICK HERE for technical assistance with your digital magazine
>> CLICK HERE for all other subscription customer service requests.

Bio-ITWorld reports the following digital circulation figures as verified by the Qmags Self-Audit procedures:

Dec/Jan 2008 Issue	
Total qualified digital subscribers:	6390
Biopharma/Chemicals	1908
Biotech/Bioinformatics	1790
Acad/Gov, Clinical Trial, CRO	1588
Hosp Lab/Healthcare, Other	1104
Nov 2007 Issue	
Total emails sent:	6358
Undeliverable:	88
Total delivered:	6262
Removed:	20

NAVIGATION KEY

- Previous/Next Page: mouse click (or keyboard) on left/right arrows to turn page
- Contents: takes you to the Table of Contents page
- Zoom In: top will zoom in to the top of page & bottom will zoom in to the bottom of page
- Zoom Out:takes you from single to double page view
- Front Cover: takes you to the Cover Page of magazine
- Search Issue: allows full search by keyword or phase (can also search archives!)

WE CAN. Yoh has been delivering high-impact talent to America's leading companies for the past seven decades – longer than any of our competitors. These customers count on our experience, time-honored processes, responsiveness and market intelligence. And our delivery of talent that is 24.7% better than the competition*.

Focus on your job. And your future.

YohClinical
Know greater talent.®
yohclinical.com

Cambridge Healthtech Institute

www.bio-itworld.com

Bio·IT World

Indispensable Technologies Driving Discovery, Development, and Clinical Trials

DEC. | JAN. 2008 • VOL. 6, NO. 10

The Route to Phase 0 and Microdosing 24

+

Visit the Thermo Scientific Exhibits at Pittcon 2008 at **Booths #1735 and 1741**

Count on Us.

Over 1500 customers worldwide already count on Thermo Scientific informatics solutions to manage one of their most important assets – laboratory data. These industry leaders have turned to Thermo Scientific LIMS and CDS software in their efforts to make the world healthier, cleaner and safer.

If you are looking for industry experience, domain expertise, innovative technology and world class after-sale support, you can count on us to meet your needs now, and well into the future.

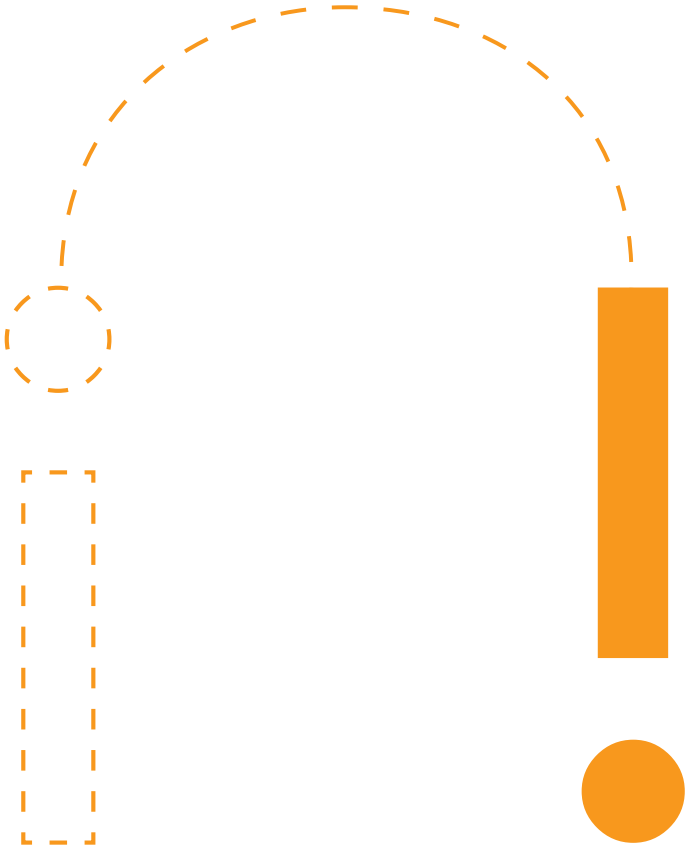
Find out which Thermo Scientific informatics solution is right for you. Visit **www.thermo.com/solutionfinder**.

+1 866-463-6522 • marketing.informatics@thermofisher.com

We have a solution for you, no matter what industry you are in:
pharmaceuticals, food and beverage, industrial chemicals, petrochemicals, metals and mining, water and wastewater and many many more.

Part of Thermo Fisher Scientific

Thermo SCIENTIFIC



Turn sequencing on its head.

It's here. Now you can perform a range of genomic experiments that you only dreamed of before. The Illumina Genome Analyzer changes the way you approach every kind of genomic experiment. The discoveries are happening, the wait is over, and our customers are publishing their results.

Change the way you think about sequencing.
Join the growing Illumina Community.

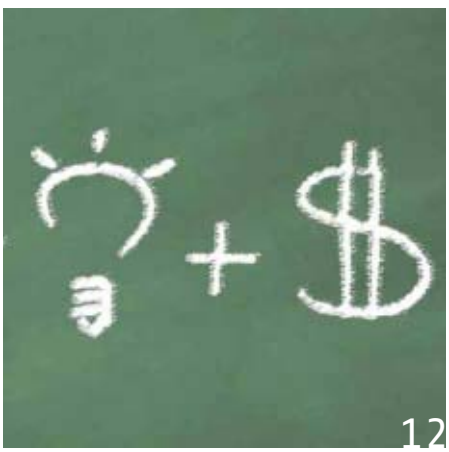
Do more with sequencing today:

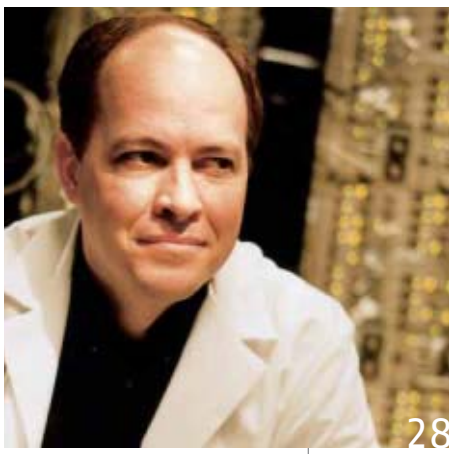
www.morethansequencing.com

illumina®

Contents [12·07 | 01·08]


6


12


28

Cover Story

Exploratory INDs: A Trial Before the Trials

Phase 0 trials and the use of tiny drug doses are gaining acceptance and winning converts. 24

Up Front

Cover illustration by Roger Harris

Computational Biology

Computational Development

In Every Issue

IT/Workflow

Clinical Research

Alternative Thinking About The Size of Blade Systems:

CAREFUL, DON'T TRIP OVER THE INFRASTRUCTURE.

Strength and power, now in a smaller, more convenient size. Introducing the HP BladeSystem c3000. All the technology of our larger BladeSystem in an efficient, compact, affordable package. Careful, watch your step.

Technology for better business outcomes.

Powered by the Quad-Core Intel® Xeon® Processor[1]

Get down to the smallest details. Visit hp.com/go/nocompromise69
1-866-625-1019

1. Intel, the Intel logo, Xeon, and Xeon Inside are trademarks or registered trademarks of Intel Corporation in the U.S. and other countries.

First Base

Inside Andy Grove's World of Bio-IT

KEVIN DAVIES

Bio•IT World frequently has cause to report the accomplishments of IT visionaries in life sciences. IDG founder Pat McGovern recognized the convergence of IT and life sciences earlier than most, bankrolling the launch of this magazine and, even more significantly, founding the McGovern Institute for Brain Research at MIT. Like him, other titans of IT are making profound contributions to medical science.

Which makes Andy Grove's recent rail against biopharma all the more surprising. In an interview on Sharon Begley's November 4 *Newsweek* blog coinciding with a speech at the Society for Neuroscience annual meeting, Grove, the former CEO and chairman of Intel, damned the drug industry for its lack of innovation, and the biomedical bureaucracy for not rewarding risky, translational research.

Since Grove joined Intel in 1968, the number of transistors on a chip has exploded from 1,000 to around 10 billion. By contrast, he says, "the same drug that was the mainstay of Parkinson's treatment in the 1960s, L-dopa, is still the mainstay today." The same is largely true for treating amyotrophic lateral sclerosis and Alzheimer's disease.

MICHAEL PRINCE

Grove is naming names.

Grove's grousing is personal. After surviving prostate cancer a decade ago, he is now battling Parkinson's disease. "In pharma," Grove says, "if a clinical trial doesn't work ... they just throw [the drug] away, when in fact the [trial data] averages may hide stuff that did work, and something that made patients different." A good drug "wrongfully convicted means the loss of benefits goes on forever."

By contrast, Grove states that everyone in the semiconductor industry shares "a deeply felt conviction that what matters is time to market, or time to money. But you never hear an executive from a pharmaceutical company say, 'Before the end of the year I'm going to have XYZ drug,' the way Steve Jobs said the iPhone would be out on schedule." That drive for understanding, he argues, "is just not present in pharma." Grove also lambastes the peer review process for "creating conformity of thoughts and values" — the modern equivalent of "a Middle Ages guild." The denial of funding and promotion to risk takers, he argues, leads to "more sameness and less innovation."

Life Story

Grove's opinions carry weight not merely because of Intel's success but his own remarkable life story. Andras Grof was born in Hungary in 1936, endured the Nazi and Communist occupations during and after World War II, before emigrating to the U.S. in 1957. He earned a Ph.D. in chemical engineering at Berkeley, and followed Robert Noyce and Gordon Moore from Fairchild Semiconductor to Intel in 1968.

Begley, the *Newsweek* interviewer, endorsed Grove's sentiments, which she said should be heeded by every congressman, big pharma CEO, and medical center dean who judges tenure and promotion decisions on a scientist's publication record "with no regard to whether the research is leading to something that can alleviate the suffering of humankind."

Grove is certainly right to endorse more risk-taking, translational, and clinical research. He prescribes "a cultural revolution in the research community" to give "wild ducks the opportunity to emerge and quack their way to success. But cultural change can be driven only by action at the top." The Howard Hughes Medical Institute is embracing that notion, but its annual budget is a fraction of that of the NIH.

But it would be daft to suggest that if biopharma simply followed the lead of the semiconductor industry, all would be well. Notwithstanding another mediocre tally of new drug approvals in 2007, big pharma has produced some superb drugs in the past decade. The semiconductor industry doesn't have the complex physiology of the human body — or the FDA for that matter — to contend with. And it would be ludicrous to reverse the improved funding of basic research over the past decade, which supplies the bedrock for all applied science, including drug research.

Many pharma insiders are understandably upset by Grove's dismissal of their dedication to the cause. Maybe the IT community should be doing more to assist pharma rather than hurling criticism. For example, in his excellent "In the Pipeline" blog, chemist Derek Lowe challenged Grove to put his money where his mouth is: "Start your own company," Lowe writes. "You've got the seed money; you can raise plenty more just by waving your hand. Start your own small pharma, your own biotech. Hire a bunch of bright no-nonsense researchers and show us all how it's done. Tell them that you're going to have a drug for Parkinson's by the end of the year, if that's what you think is lacking. Prove me and the rest of the industry wrong."

myALA conference planner

Join the Diversity of **LabAutomation2008**

LabAutomation2008 delivers the educational and informational solutions for the multi-disciplinary global community of scientists, engineers and professionals involved in research and discovery. Register today for the only world-wide conference and exhibition focused exclusively on the rapidly growing field of laboratory automation.

A Diversity of Industries Including:

- Drug Discovery & Development
- Clinical Molecular Diagnostics
- Agricultural & Food Sciences
- Forensics & Security
- Energy, Commodity Chemicals, and Polymers

LabAutomation2008 Key Education Tracks:

- Detection and Separation
- Micro- and Nanotechnologies
- High-Throughput Technologies
- Informatics
- New Frontiers *for* **LabAutomation**: Agricultural and Food Sciences

Registration Now Open!

Visit labautomation.org/LA08
or call +1.888.733.1ALA (1252)

Short Courses: January 26–27, 2008
Conference: January 27–30, 2008
Exhibition: January 27–29, 2008

Segway to Success campaign

Win a Segway PT

Submit your best work to ALA and you might be the lucky winner who takes home a Segway PT.

Visit **labautomation.org/journal/winasegway.cfm** for details.

LabAutomation 2008

Where Science, Technology and Industry Come Together

Premier Sponsor

VELOCITY11®
The Speed of Discovery™

ALA is a non-profit association committed to driving progress in laboratory technologies through high-quality education that benefits the global scientific community, including its membership of scientists, engineers, academicians, and thought leaders.

ALA Association for Laboratory Automation

Company Index

Advertiser Index

This index is provided as an additional service. The publisher does not assume any liability for errors or omissions.

VOLUME 6, NO. 10

Editorial, Advertising, and Business Offices: 250 First Avenue, Suite 300, Needham, MA 02494; (781) 972-5400

BioIT World (ISSN 1538-5728) is published monthly except July/August and December/January by Cambridge Bio Collaborative, 250 First Avenue, Suite 300, Needham, MA 02494. *Bio IT World* is free to qualified life science professionals. Periodicals postage paid at Boston, MA, and at additional post offices. The one-year subscription rate is $199 in the U.S., $240 in Canada, and $320 in all other countries (payable in U.S. funds on a U.S. bank only).

POSTMASTER: Send change of address to Bio-IT World, P.O. Box 3414, Northbrook, IL 60065. Canadian Publications Agreement Number 41318023. CANADIAN POSTMASTER: Please return undeliverables to Station A, P.O. Box 12, Windsor, ON N9A 6J5.

Subscriptions: Address inquires to Bio-IT World, P.O. Box 3414, Northbrook, IL 60065 (888) 835-7302 or e-mail biw@omeda.com.

CHI

Bio·IT World®

Indispensable Technologies Driving Discovery, Development, and Clinical Trials

EDITOR-IN-CHIEF
Kevin Davies (781) 972-1341
kevin_davies@bio-itworld.com

EXECUTIVE EDITOR
John Russell (781) 972-1342
john_russell@bio-itworld.com

ASSOCIATE MANAGING EDITOR
Allison Proffitt (781) 972-1345
aproffitt@healthtech.com

ART DIRECTOR
Mark Gabrenya (781) 972-1349
mark_gabrenya@bio-itworld.com

WEB EDITOR
Catherine Varmazis (781) 972-1344
catherine_varmazis@bio-itworld.com

DIRECTOR, BUSINESS DEVELOPMENT
Angela Parsons (781) 972-5467
aparsons@healthtech.com

VP SALES – CA, WESTERN US, MIDWEST, SOUTH EASTERN US, EUROPE, CANADA, PACIFIC RIM
Alan El Faye (213) 300-3886
alan_elfaye@bio-itworld.com

REGIONAL SALES MANAGER – NEW ENGLAND, NORTH EASTERN US
Kay O. Christopher (860) 693-2991
kchristopher@healthtech.com

SENIOR DIRECTOR OF MARKETING & OPERATIONS, PUBLICATIONS
Joan A. Chambers (781) 972-5446
jchambers@healthtech.com

CIRCULATION & MARKETING MANAGER
JoAnne Sperry (781) 972-1350
jsperry@healthtech.com

PROJECT/MARKETING MANAGER
Lynn Cloonan (781) 972-1352
lcloonan@healthtech.com

ADVERTISING OPERATIONS COORDINATOR
Kathrene B. Tiffany (781) 972-5440
ktiffany@healthtech.com

GRAPHICS AND PRODUCTION MANAGER
Ann-Marie Handy (781) 972-5493
ahandy@healthtech.com

Contributing Editors
Michael Goldman, Karen Hopkin, Deborah Janssen, Pauline Parry, Salvatore Salamone, Tracy Smith Schmidt, Mark D. Uehling

Advisory Board
Jeffrey Augen, Mark Boguski, Steve Dickman, Kenneth Getz, Jim Golden, Andrew Hopkins, Caroline Kovac, Mark Murcko, John Reynders, Bernard P. Wess Jr.

Cambridge Healthtech Institute
PRESIDENT
Phillips Kuhl

Contact Information
editor@bio-itworld.com

250 First Avenue, Suite 300
Needham, MA 02494

You Stand in Front of Our Instruments All Day... We Stand Behind Them...24x7

High Value, Substantially Expanded Services from Applied Biosystems

With over 25 years of experience in the development, manufacture, and service of innovative instruments, and with over 1000 highly trained, dedicated service professionals, AB Global Services is uniquely qualified to deliver a full suite of real-world lab services. From Remote Services to On-Site Application Consulting to Qualification and Professional Services and everything in between, AB Global Services is your value-added partner to help boost your productivity and maximize the return on your technology investment.

- Instrument Repair and Maintenance
- Smart Services
- Qualification Services
- Professional Services
- On-Site Application Consulting
- Training

To learn more go to http://info.appliedbiosystems.com/wwservice or contact your local Applied Biosystems sales representative.

Applied Biosystems

Sequencing Stories and Start-Ups

CHI's second next-generation sequencing conference showcases the first personal genomes and future technologies.

BY KEVIN DAVIES

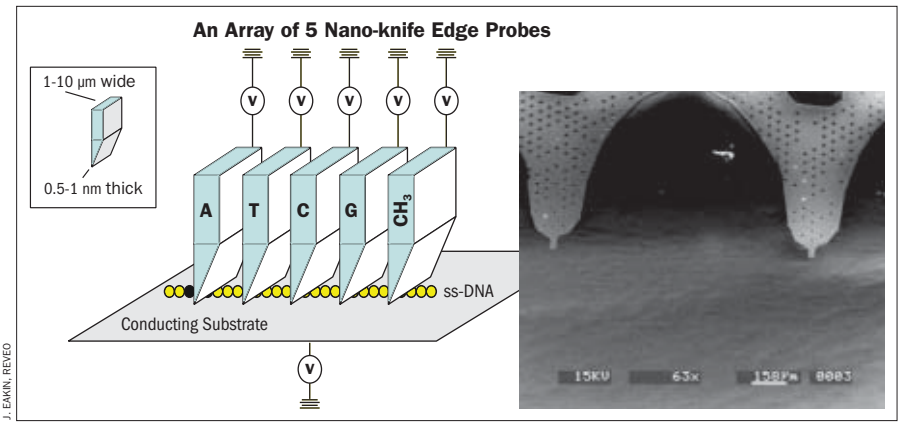
PROVIDENCE, RI — It's been 20 years since scientist-entrepreneur **Kevin Ulmer** launched the first single-molecule DNA sequencing company, SEQ. Since then, Ulmer has studied or consulted for just about every next-generation sequencing system. Now he has been born again. The title of his keynote address at CHI's Exploring Next Generation Sequencing conference*, said it all: "Sanger sequencing is dead: Long live Sanger sequencing!"

After two decades, Ulmer is returning to the personal genomics field with a new venture called Genome Corp. (See, "Genome Corp..."). Though the data in GenBank is doubling every 16 months, and the cost halves approximately every 25 months. Ulmer believes another dramatic shift in technology would be required to make the "$1000 genome" a reality before 2040.

Ulmer offered his the pros and cons of all four leading next-generation sequencing technologies: pyrophosphate diffusion and high consumable costs (454); read-length limits and image processing (Illumina, Applied Biosystems); and photobleaching, monochromatic signal detection, and cost (Helicos). (On that final point, Ulmer quoted a $2 million estimate published by the *Wall Street Journal*, whereas Helicos subsequently priced its HeliScope at $1.35 million.)

Ulmer has concluded that the proven track record, unparalleled read lengths, and well understood error properties of traditional Sanger sequencing serve as fundamental assets in building the equivalent of a "printing plant" for genome sequencing. Without divulging details, Ulmer said his variation on Sanger will involve single-molecule processing, "one pot" *in vitro* amplification, and a rethink on electrophoresis and imaging.

*Exploring Next Generation Sequencing: Applications and Case Studies; Providence, RI; October 17–18, 2007.



Reveo's nano-knife edge probe directly interrogates nucleotides to read DNA sequence.

Young Guns

The meeting was notable for presentations on the first two complete human genome sequences. **David Wheeler** (Baylor College of Medicine) discussed the first draft of James Watson's genome, based on more than 100 million fragment reads performed by 454. **Samuel Levy** (J. Craig Venter Institute) recapped the publication of the "first human diploid reference sequence," that of Venter himself, using Sanger methods. Both speakers emphasized the vast amount of genetic variation in these sequences, not merely SNPs but insertions/deletions and copy number variants.

Meanwhile, several start-up companies presented new platforms that hope to challenge 454, AB, Illumina, and Helicos in the coming years. One was Reveo, one of the first groups to enter the $10 million Archon X PRIZE for genomics. **James Eakin,** manager of business development, claimed that Reveo's nano-knife edge technology has the potential to sequence a human genome "error-free, in minutes, for pennies." He presented "a disruptive instrument concept" that uses physical, non-disruptive methods, rather than chemical, to directly interrogate chromosomal DNA sequences.

Reveo's solid-state instrument relies on concepts borrowed from the semiconductor, optics, and micro-fabrication industries. An array of electro-conductive, nano-knife edge probes, each about 10 nm long and less than 1 nm thick, interrogates the DNA, including epigenomic modifications. Eakin envisions a microwave-sized instrument costing about $5,000, with standard microprocessors reading sequence data at a speed of 1 base/nanosecond.

Steven Gordon, CEO of Intelligent Bio-Systems (IBS), is readying a platform based on sequencing-by- synthesis chemistry (SBS) developed by Jingyue Ju's lab at Columbia University. The target market is clinical "diagnostic sequencing," with a premium placed on "cost per test" rather than "cost per base." IBS' "pinpoint" system uses single-stranded DNA templates on an array and incorporates fluorescently labeled nucleotide analogs into complementary strands. The chemistry involves custom cleavable reversible terminators and dyes. After signal measurement, both terminator and dye are cleaved before the cycle repeats.

Gordon says his firm's bench-top instrument will produce about 5 GB sequence per day, and is looking for early access customers. The instrument will sell for about $275,000, enabling many

Genome Corp. Born Again

"There's nothing fundamentally wrong with Sanger [sequencing]," says Kevin Ulmer, founder of Genome Corp. "It needs some polishing and updating, but we believe it can outperform all of these next-gen methods by a substantial margin, without having to abandon what has generated 99.999% of everything that's in GenBank."

Ulmer's vision is to develop the most efficient production line for DNA sequencing. Without having to produce commercial instruments, Ulmer's focus is on developing an ultra high-throughput DNA sequencing factory.

"We will never sell hardware," he says. "We will be selling sequence information. It will be produced in a factory that will be designed and built with technical solutions and economies of scale that allow you to produce sequence far faster and cheaper than you could on a stand alone basis."

Ulmer says there are two main ingredients to reinventing Sanger to reach the necessary scale of affordability. First is "some form of *in vitro* amplification as the first step... so you have to move to essentially a one-pot amplification to give you tens of millions of reads."

Second is a major rethink of the downstream separation process. "Sequencing is an information services business, it's not an instrument reagent business. It will go the same way that oligonucleotide synthesis has gone," says Ulmer. To order oligos these days, "you just hop on the Internet, you tap away, and it comes in a tube... It's become a commodity... I can see sequencing going the same way."

The name Genome Corp. is borrowed from one of the first commercial DNA sequencing operations, launched by Harvard Nobel laureate Walter Gilbert 20 years ago. "My main concern was whether Wally would have problems with [the name], and he said, 'No, I can't object,'" says Ulmer. "It's somewhat gratifying to come full circle." — KD

labs beyond the genome centers to tackle whole genome analyses.

Another promising contender is NABsys (Nanopore array-based systems), a spin-out of Brown University. VP **John Oliver** explained how his start-up is developing hybridization assisted nanopore sequencing, combining the strengths of sequencing-by-hybridization (SBH), which "works but doesn't scale," with nanopore sequencing, which "scales but doesn't work." While others insist that nanopore sequencing could yield DNA sequences in real time, Oliver is content to use nanopores at a lower level of resolution — to detect the location of hybridization probes to the target sequence.

The key is to scale up the method by building nanopore arrays to permit rapid determination of positional hybridization information. The assembly of the target sequence relies on algorithms similar to those used in SBH.

Perhaps the most entertaining presentation came from **Leonard Bloksberg**, founder of New Zealand-based Cartesian Gridspeed (See, "The Quest to Make Sequence Sense," *Bio•IT World*, November 2006). "Biologists are feeling the pain of the massive data overload required to even participate in modern biology," said Bloksberg. His flagship program, dubbed SLIM Search, works up to four orders of magnitude faster than BLAST with high volumes of short sequences, and has affordable academic seat licenses. With BLAST laboring away in the background on his laptop, Bloksberg illustrated the acceleration of sequence analysis using his proprietary program. •

CHI's next Next-Generation Sequencing event switches to the West Coast and takes place April 23-24, 2008, in San Diego. www.healthtech.com/2008/seq

PUBLIC OFFERING

As part of their contributions to the NSF-funded **TeraGrid**, **Indiana University** is making public their Centralized Life Science Data (CLSD) service. CLSD is a platform that makes publically-available data sets, including PubMed, several NCBI BLAST databases, and dbSNP, accessible with a single SQL query, allowing researchers to merge search results from multiple sources. Researchers can also create their own queries. The University is currently offering CLSD in pilot mode, and is actively soliciting input and users to test and expand the service.

LARGEST SUPERCOMPUTER

SGI has built the world's largest Field Programmable Gate Array (FPGA) supercomputer configuration featuring 70 FPGAs. The supercomputer completed a BLAST-n query in 33 minutes, 900 times faster than the same application would run on a traditional cluster. The configuration is made up of off-the-shelf compentens including the SGI RASC appliance.

SOUTH AMERICAN CONTINGENT

Thermo Fisher Scientific has partnered with **Controval, C.A.**, a Venezuelan company, to open a sales center for Thermo Scientific informatics solutions in Venezuela and the Caribbean. Thermo hopes that the local office will better serve their existing South American customers.

COLLABORATION

NextGen Sciences and the **Harvard Institute of Proteomics** have announced a collaboration that marks the start of NextGen's Center of Excellence Program. The program will focus on integrating the company's information management software, orchestratorIMS, with Harvard's clone database to develop automated methods for multi construct design.

VCs Should Go Back to School

MICHAEL A. GREELEY

Venture capitalists need to go back to school. It is that simple. Some of our best, most innovative, most promising investment opportunities are coming from campuses of many of the universities in our back yards. In fact entrepreneurs would also be well-served going back to school.

In addition to generating wonderful ideas, faculty at leading universities are able to secure meaningful government and other non-dilutive grants to launch projects that have potentially compelling commercial applications. Many of these projects will become the great venture-backed companies of the future. The government has been aggressively funding many bio-IT projects; it is these grants which serve as an important source of validation and raises their visibility within the investment community.

An area of particular focus has been the molecular detection field. "With the sequencing of the human genome, the National Cancer Institute (NCI) has been very focused on being able to detect cancer better and sooner," observed Professor Bill Hancock at Northeastern University, who chairs the Chemistry Department. Hancock recently secured a meaningful grant from the NCI, part of the National Institutes of Health (NIH), which is funding a new $15.5 million, five-year initiative to discover, develop, and clinically validate cancer biomarkers by targeting the carbohydrate part of molecules.

The Northeastern team recognized that many of the genetic tests available today are inadequate and require greater insights from the proteomics field. Tissue analysis is acceptable but expensive and laborious. The challenge today is identifying proteins of interest which may be in low abundance. This analysis may be facilitated by glycosolation which researchers expect will lower the detection thresholds by 10-20 times. Ultimately this should meaningfully magnify the disease signal to make detection more readily available.

Particularly notable about this grant is that Northeastern worked across a number of labs and academic centers to secure this commitment. Thirty-seven labs competed for aspects of the overall grant. The Northeastern consortium developed a comprehensive proposal. This approach had important benefits: it allowed individual labs to focus on their core competencies and permitted the consortium to refine the overall strategy based on multiple inputs.

The complexity of the problems being studied demanded a multi-lab, multi-disciplinary approach. "Treatment will ultimately be a data-driven decision…it will be a combination of all the "*omics*" fields such as genomics, proteomics, glycomics, peptidomics which solve these issues around cancer, and bioinformatics will be in the center of it all," stressed Hancock. It is this lack of clarity around the complexities of the biology surrounding cancer which drives these types of grants.

"Within a couple of years, informatics will generate "actionable" data and truly specific biological insights…my guess is that occurs over the next five years and then another five years from that point we may begin to see real clinical benefits. The black box may be a little more transparent with our work," said Hancock.

Ordinarily you don't expect a venture capitalist to be overly supportive of government's intervention in the funding of promising technologies but there may well be complex cross-discipline examples which underscore the value of such grants. The promise of bio-IT will only be realized by the interactions of various disciplines focused on solving a common set of problems. It is this broader ecosystem which the government and other academic/philanthropic entities can support.

Many VC's track grant recipients as well as encourage our portfolio companies to aggressively seek similar grants. Applications can be painfully long and decision insufferably slow to be reached but are often well worth it; many times winning is more important than the actual dollars received. The attention and notoriety with securing an award can serve to attract additional venture capital dollars as well as facilitate recruiting of other management team members and scientific advisors.

My advice to would-be entrepreneurs is to also pay close attention to these academic thought leaders. There is plenty of capital available to fund innovation, and there is a reasonably strong pipeline of innovative ideas coming out of academic environments; the most important missing ingredient in this equation is competent management talent to drive these projects forward. It would serve you well if you aspire to start a company to mine carefully the list of grant recipients at leading universities. You do not need to be the technical inventor to be a founder of a venture-backed company, but you do need to be there early.

Michael A. Greeley is managing general partner of IDG Ventures. E-mail: mgreeley@idgventures.com

OSINT and the Pharmaceutical Enterprise II

JAMES B. GOLDEN

Many OSINT (Open Source Intelligence) resources are available online, but the real art to OSINT analysis is in framing the right questions and creating the right sort of analysis for decision makers. While many free and commercial search engines and content aggregators exist in the public domain, a tailored retrieval, discovery, and refinement process must be created for answering specific intelligence questions around biopharma portfolios and practices from open sources.

Software for enabling this process includes: (1) OSINT search and content extraction tools (RSS aggregators, Web crawling and scraping); (2) Content classification, categorization, and clustering tools; (3) Entity and relationship extraction tools; (4) Taxonomy and ontology creation and management tools; (5) Presentation software including visualization; (6) Desktop search engines (personal knowledge refinement by individual analysts); and (7) Analytics tools, both quantitative (increased chatter, number of mentions, word co-occurrence, etc.) and qualitative (sentiment, opinion)

Many readers will have ample experience using (if not developing) such tools — whether building a corporate-wide search strategy, modeling data as taxonomies or ontologies, or building semantic-web applications. Many vendors supply these tools, often quite cutting-edge. However, a more complete approach to intelligence collection and analysis requires focusing on system integration and the final intelligence product, rather than the newest, coolest piece of technology. Big pharma research informatics groups probably license several interesting tools. Look at what you've got and think about linking them into an overarching process that addresses your specific questions.

The collection and analysis of OSINT requires a focused, synchronized set of processes. Open sources should be categorized by source type, content relevancy, source reputation and coverage quality. We routinely monitor over 150 reputable blogs that contain unique information to identify potential trends. I also keep a spreadsheet of online sources and rank them by their perceived accuracy and utility. When my content harvesting engines pull source data following a query, I check those sources against my list, update my sources, and iterate on my collection process. This also provides a head start in creating a taxonomy that can be passed to your search engine.

I use several entity and relationship extraction tools to aid in taxonomy creation, as well as manually collecting terms when web browsing. This has been especially fruitful in monitoring trends in biomarker discovery and validation. When traditional sources like Reuters or AP promote articles on personalized medicine, I can compare these articles against my biomarker taxonomy to make sure my automated systems haven't missed anything.

Monitoring and Analysis of OSINT Sources

Monitoring the open source ecosystem for real-time changes to Web content is critical. Relevant information should be received as soon as it's published and organized in a systematic way. Business analysts should be alerted to new blogs, articles, and featured sites with informative content and be able to add and organize blogs and news feeds without being technology experts. They should also be able to view content from one feed at a time, groups of feeds, or all related feeds at once.

Successful monitoring should include checks for:

- Automated web page surveillance and alerting relevant analysts when significant changes to relevant sources occur;
- Identification of new (or removed) content from open sources of interest;
- Sort and filter web page changes by date/time change detected, page content categories (blogs, chat, etc.) or watchlist groupings;
- The ability to rate and flag pages to suit analysis needs.

Analysis should include:

- Search by concept and example, as well as keywords;
- Discovery: the identification and extraction of items likely to contain relevant information for analysts, e.g. identifying trends, competitors, new entities of interest, etc.;
- Entity and relationship extraction of persons, groups, products, etc. and population of taxonomies and ontologies to understand relationships and associations to each other and to previously defined networks;
- Visualization of networks and non-intuitive relationships for improved understanding and trend-spotting;
- Calculation of quantitative metrics such as chatter volume, key word co-occurrence, new network connections, etc.;
- Calculation of qualitative metrics such as sentiment (person/organization's feelings or emotional response as manifested by descriptions in open sources) and changes in direction or opinion;
- Determination of cogent answers to questions that surfaced during structured discovery.

Jim Golden is a CTO at SAIC. He can be reached at goldenj@saic.com

Next Issue: Part III — Disseminating OSINT analysis to decision makers and some real world applications.

SNPedia: A Wiki for Personal Genomics

BY MICHAEL CARIASO

November was an historic month for personal genomics. Four companies announced details of new direct-to-consumer genotyping services. Google-backed 23andMe's kit sells for $999, while Iceland's deCODE Genetics launched its deCODEme service for $985. Knome began to seek clients for full genome sequencing, and Navigenics, announced it would launch in early 2008, offering a screening for 20 common diseases for $2500.

A few years ago, having gained familiarity with various microarray platforms, I figured out how to run my own DNA and extract the details. By cataloguing my single nucleotide polymorphisms (SNPs), I knew 500,000 facts about myself, but had no idea about their implications. As my resources were more technical than financial, starting a wiki made more sense than starting a genetic testing company. And so SNPedia (www.snpedia.com), a wikipedia for SNPs, was born. The site currently has information on nearly 2,000 medically relevant SNPs.

"23etAl" (23andMe, Navigenics, deCODEme, Knome, and other private companies) appear to be building high-quality curated walled gardens, whereas SNPedia is more of a public park. They may even use SNPedia, since they can continue to take customers' money to do the testing, but could use SNPedia to simplify some of the annotation and report generation.

Most consumers will be satisfied with results obtained from 23etAl, trusting everything is reliable, and will not have much use for SNPedia. Researchers may use SNPedia to increase the visibility of their work, but scientific journals will still be primary. It's the "recreational genomics" crowd that might be motivated to learn what an odds ratio or Bonferroni correction is. A wiki is a good format for that sort of information. They will (mostly) understand that SNPedia is a home of lower confidence interesting possibilities. And while there will be limitations to SNPedia's content — it is a wiki after all — pages will continually grow to add the missing information. As we like to say in the open-source software world, with enough eyeballs all bugs are shallow. The same holds true for the science.

The NCBI rs#s used to identify SNPs are the key to the whole system. The use of other nomenclatures is still wide-

A Stroll Through SNPedia.com

Use the search box to find the "Rs1799990" page. Clicking the history tab shows that this page was annotated entirely by the SNPediaBot (the wiki's meticulous and very capable librarian). The edit tab reveals:

```
{{ rsnum
| rsid = 1799990
| Gene = PRNP
| Chromosome = 20
| position = 4628251
| geno1 = (A;A)
| geno2 = (A;G)
| geno3 = (G;G)
}}
{{ omim
| id = 176640
| variant = 0005
| desc    = PRION DISEASE, SUSCEPTIBILITY TO
| rsnum   = 1799990
}}
{{ neighbor
| rsid = 16990018
| distance = 127
}}
{{on chip | Illumina Human 1}}
{{on chip | Illumina Human 1M}}
```

The SNPediaBot pulled down data from NCBI including the SNPs gene, chromosome, and position. It recognized the rs# in OMIM and recorded its existence and its link to OMIM. The bot identified that 127 nucleotides away is another SNP (for which additional information is provided), and that this SNP is found on two Illumina microarrays.

Technically SNPedia can be called a Semantic Web, which means authors can write programs that read, write, and understand the wiki. One of the goals of SNPedia is to create an ecosystem where people are encouraged to contribute. For example, if a researcher who has identified a SNP that varies across patient populations creates a page such as:

```
Title: rs12345
Body: The G allele is more common in prostate
cancer patients
```

the bot will reward his or her efforts by connecting this SNP to its neighbors and identifying its presence on any known microarrays. Perhaps a neighboring SNP is on a microarray and can be used as a surrogate for easier testing. This sort of information hasn't existed before in any accessible way.

The Categories page under Special pages (left hand toolbox) automatically reveals the latest statistics on the site, such

spread, but improving. I look forward to the day when copy-number-variations and mitochondrial SNPs have been similarly cataloged. Full genome sequencing remains the Holy Grail, but without using BLAST or other tools to reduce a full genome into discrete SNP-like categories, I doubt anyone will be able to make any actionable statements based on a full genome.

SNPedia also reveals which SNPs are present on the commercially available chips from Affymetrix and Illumina used by 23etAl. This provides an opportunity to compare what information is common to the respective platforms, and what SNP probes are unique. Because many of the current SNPs in SNPedia pertain to rare disorders characterized in OMIM (Online Mendelian Inheritance in Man), the wiki may also help suggest which SNPs should be included on the next generation of microarrays.

In a sense, SNPedia has been waiting for the day when enough people actually know their genotypes. 23etAl will bring that day much closer. Given the legal and ethical issues involved with sharing genetic information, I'm happy to let the 800-pound gorillas fight those battles. Few people currently know their genotypes, so our authorship is small. However, the author of a recent New York Times article on 23etAl said she got her rs#s as part of the 23andMe report, then found additional information via SNPedia. Hopefully more consumers will do just that.

Michael Cariaso is the senior scientific consultant for the BioTeam. He can be reached at cariaso@bioteam.net.

as the total number of SNPs and the number of SNPs located on various commercial microarrays.

In some cases, SNPs exist without an entry for the corresponding gene. For example, on the page for Rs28933101, notice that the gene MET is in red — the page about MET has not been created yet. Click on MET and you find a blank edit box. But even on a blank page, there is information. The What-Links-Here page (left hand toolbox) produces a list of six SNPs, and the entry for Autism. Even non-existent pages can be useful.

At the other end of the SNP spectrum is Rs1815739, a manually prepared entry that illustrates what most people hope to find at the site.

In addition to the wiki, there is also a chat room accessible from a tab on any SNP page. This allows people interested in a particular SNP or topic to talk in real time. For SNPs with a more academic interest, researchers across the globe may have a way to conduct a continuous virtual conference (akin to what some folks seem to be trying to do with Second Life). For SNPs of greater interest to the general public, the chat room may offer something between a genetic counselor and a peer support group.

Computational Biology

Swinging Through the Proteomic Data Jungle

HUPO stresses need for standardized analysis methods and a workflow focus.

BY KEVIN DAVIES

Data processing, validation, standardization, and protein quantifications were among the central themes of the 6th annual Human Proteome Organisation (HUPO) World Congress, held in Korea in October.

HUPO was founded in February 2001, the same week as the publication of the first draft of the human genome. The organization's council now has 48 members from 19 countries, with its headquarters located at McGill University and Genome Quebec Innovation Centre in Montreal. HUPO has 2000 founding members from 69 countries.

According to Bruker Daltonics' director of bioinformatics, Herbert Thiele, "We've all learned that in bioinformatics, we have to address different proteomics workflows. The extreme complexity of the proteome calls for different multistep approaches." These are usually combinations of electrophoresis and liquid chromatography (LC) techniques in combination with different MS and MS/MS methods.

"Any kind of software solution for data warehousing and analysis should address these different workflows in a flexible manner," says Thiele. Bruker's ProteinScape platform supports various discovery workflows through a flexible analyte hierarchy concept, as well as addressing scientists' needs in biomarker profiling and quantification. Thiele points out: "A database solution is the only way to compare experiments to one another and to extract knowledge based on past experiments."

"Quantitation is becoming more and more important," says Thiele. "All vendors are working hard on quantitation tools." ProteinScape fully supports all current label chemistries for protein quantification, and the software will handle future label technologies. Interactive validation of protein quantification based on raw LC/MS data is now simple and straightforward.

Recent improvements in MS instrumentation make a label-free MS-based quantification approach feasible. This technology has the potential to become a significant complement to current quantification methods, such as label based MS methods. The high throughput compatibility of a label-free approach allows large numbers of samples to be processed. Handling these workflows from data preprocessing to statistical validation of quantification results is a big challenge.

Thiele sees parallel processing as the next big challenge.

Brain Proteome Project

An important takeaway from HUPO 2007 was the need for standardized analysis methods and result validation techniques. One of nine official global HUPO consortia is the Human Brain Proteome Project, headed by Helmut E. Meyer (Medizinisches Proteom-Center, Bochum, Germany), aimed to map the "proteomic landscape of the brain" using mouse and human samples " to get deeper insights into neurodegenerative diseases, and produce an inventory of proteins in the human brain."

The brain consortium established guidelines for data processing for protein identification, which Thiele calls a "very important step forward." It will allow researchers to "compare results and statistical relevance for all generated data, within the huge jungle of proteomics data." A data warehousing system including a data processing pipeline is mandatory for data comparison and validation.

Thiele says: "The fundamental problem of protein identification is [that] you get a long list of potential protein identifications, but nobody tells you which proteins are actually correct. The decoy approach allows you to measure the rate of false positives by mixing artificial protein sequences into the database."

For protein identification and characterization, ProteinScape uses complementary tools. The use of different search engines provides automatic cross-validation of the identifications in parallel with improved sensitivity (resulting in more protein ID's). The resulting peptide identifications are analyzed by the ProteinExtractor tool. This can even merge data from different search engines as well as from different experiments (ESI and MALDI), producing an integrated result. The use of decoy strategies minimizes the need for manual validation.

"In the near future," Thiele continues, "it will be a must that all protein identifications will come with a statistical significance, so everyone can judge the validity of the information. We need reproducibility and standardized ways to create confidence in the generated results."

Much like the genomics arena, the variety of LC/MS mass spec techniques in proteomics is producing vast volumes of data, posing two major issues in bioinformatics. First, "Do we need all the raw data in the database?" To cope, the processing pipeline has to be able to condense the data, and dedicated software tools must validate the results. "These

ProteinScape uses duel search engines to cross-validate protein ID and characterization.

tools should be able to visualize selected raw data, and correlate the results," says Thiele. Especially for applying quantitation algorithms, access to MS raw data is mandatory to make sure the information contained in the raw data is not disturbed by processing."

The other data handling issue concerns software for data visualization based on different workflows, for example, gel images and LC/MS data sets. "For the huge map of LC/MS data, a user-friendly navigation through large volumes of data is needed," says Thiele. "You need visualization tools for fast multi-resolution visualization of the data as an image ensuring seamless transition from a global overview of all spectra to selected isotopic peaks." Examples include MSight from Gene Bio, and SurveyViewer from Bruker.

Machine-Readable Experiments

Of course, producing large protein lists is not the end point in proteomics research. To enable result assessment and experiment comparison, the experimental conditions must be documented in a concise, reproducible, and also machine-readable way. This is done by PRIDE (PRoteomics IDEntifications database at the European Bioinformatics Institute, http://www.ebi.ac.uk/pride).

Thiele says: "The ideal would be to handle, distribute, and archive proteomics data in a data repository and incorporate the publishers of science journals to set up specific guidelines. In the past, all manufacturers had their own file formats, with software running just on the vendor's machine. Nowadays, the vendors are participating with consortia to support initiatives in data standardization." That helps researchers generate data on one instrument and use dedicated software tools to turn data into knowledge.

The European Commission-funded ProDaC consortium (Proteomics Data Collection) will finalize data storage and standards, implement conversion tools, and establish standardized submission pipelines into the central data repository. For example, the Brain Proteome Project has already uploaded the ProteinScape data reservoir into PRIDE.

In Thiele's opinion, IT has an important role both in "computer clustering and computer grid technology. Automatic parallel processing of large MS data sets in a distributed computing environment, combining compute resources at different locations to do specific tasks, is the great challenge in near future." •

Gold Standard

Another HUPO-related data validation initiative involves Invitrogen, which is launching the HUPO Gold Mass Spectrometry protein standard sampling program.

Designed to serve as the first commercial, all-recombinant human protein standard for mass spec, the HUPO Gold MS standard is a defined mixture of known human proteins that can serve as a benchmark to judge data quality and allow researchers to cross-reference their results. The standard will work regardless of the type of mass spectrometer used.

"With a variety of published mass spectrometry workflows, as well as the large number of instruments and data-analysis software packages available for use, researchers today face major challenges validating and comparing their published data," said John Bergeron, chair of HUPO scientific initiatives. The new standard, he says, together with HUPO training, "should lead to field-generated data of greater run-to-run accuracy and reproducibility."

Paul Predki, Invitrogen's VP of R&D, adds that current mass spec standards could contain contaminants or vary slightly in mass based on natural genetic variations. "We have designed a valuable resource that will aid scientists in making their substrate identification more definitive and will allow them to reference their efforts on a global research scale," says Predki.

The HUPO Gold MS Protein Standard samples are available to HUPO members, with the full release set for early 2008.

Computational Development

Extracting Information in Life Sciences

Through the BioCreative competition, MITRE's Lynette Hirschman helps raise the performance and relevance of academic text mining solutions.

BY KEVIN DAVIES

The MITRE Corporation is a nonprofit, public-interest organization that manages three federally funded research and development centers (FFRDCs), and as part of its mission, conducts its own R&D programs. One FFRDC supports the Federal Aviation Administration, another the Internal Revenue Service. The third, and oldest, supports the Department of Defense (known as the DOD Command, Control, Communications and Intelligence FFRDC). MITRE's main sites are in Virginia and Massachusetts, with smaller sites colocated with key sponsors.

"We do not make products, we cannot compete, because that would interfere with our primary mission which is to provide unbiased advice to the government. We are a very different kind of entity," explains MITRE's Lynette Hirschman.

Hirschman joined MITRE in 1993, and works at the Center for Intelligent Information Systems (CIIS), part of the DOD FFRDC. She admits to having a "strange background:" an undergraduate degree in chemistry, a Master's in German literature, and a Ph.D. in computational linguistics. She joined MITRE after stints at Unisys, NYU, and MIT, working primarily in language understanding.

Since 2000, Hirschman's interests have hovered "increasingly on the intersection of natural language and bioinformatics."

"Biology is a great place to apply information extraction — the ability to pull out facts from free text material — because unlike computer scientists, biologists actually read the literature!" says Hirschman. "They're very dependent on knowing who's doing what and what the state of the art is. If you can give them tools to make that easier, then you don't have to have post docs spend their first six months extracting information from the literature on their projects."

The other advantage, she says, is the rise of genomics: "Biology is a language; it's the language of DNA and protein. These are linear strings of molecules, they encode information. That is the principle you can apply to determine where to find genes and find repeating patterns — extremely similar to the techniques used in modern natural language processing."

Lynette Hirschman fosters creativity.

Creative Vision

Today, Hirschman leads MITRE's efforts and know-how in information management and data mining to the life sciences community, developing tools to help

BioCreative II attracted groups from multiple countries representing universities, research institutes, and private companies.

The Language of Text Mining

The fifth Fraunhofer Symposium on text mining in the life sciences last September was supported by TEMIS, a French text mining company, with offices in Paris, Grenoble, Germany, and Philadelphia.

Last year, TEMIS released a new text mining platform called Luxid, replacing Insight Discoverer Suite. TEMIS subsequently released Luxid for Life Sciences, which is used in three major areas: accelerating drug discovery; analyzing and monitoring company IP assets (especially patents); and reducing adverse event-related risks.

"Luxid for Life Sciences integrates all our knowledge and experience in the field of life sciences. It integrates most of our experience through cooperation with larger pharma organizations and institutes. The key feature is basically the combination of three domains of knowledge — chemistry, biology, and medicine," says Charles Huot, TEMIS' cofounder and COO.

The list of TEMIS' pharma customers has grown steadily over the years, including Sanofi-Aventis, Roche, Novartis, and Pfizer. The company recently announced a large deal with Bayer HealthCare to support drug discovery efforts. "The objective for all these projects is really to help scientists read the scientific literature faster, whether it is Medline abstracts or full-text. The goal is always the same — the need to accelerate the discovery process," says Huot.

A major application of text mining is to understand the druggability of a particular target. "The idea is to look at combination between the chemistry model and the biological model," says Huot. "To do that, you need to discover and understand both the chemical compounds within documents, whether text or chemical formulae, and identify the associated gene or protein." Extracting information on both chemicals and biological products "needs two features that are hard to obtain together," says Huot.

Another key feature is normalization, says Huot, so everyone knows the drug names. But Luxid also allows investigators to "draw the chemical compound and search with the drawing." Huot claims that this dual search system — text and chemical structure — is unique.

One area that TEMIS has exploited successfully is adverse events. Huot points to a major U.S. pharma company that was required by the FDA to comply with Sarbanes-Oxley by managing information received in unsolicited emails. "If you have a patient who sent an email to the website of a pharma company, they have to read this email — they can't just throw it away. They have to see whether you might find a potential adverse event," Huot explains.

Such unsolicited emails might say: 'I'm taking this particular drug,' 'I feel sick, it seems to be worse than before,' or 'I get vertigo when I stand up.' "This email must be identified as a potential adverse event. It should be posted into a special FDA form and reported to the FDA in a very short time," says Huot.

Charles Huot is using text mining to record and track adverse events.

Many pharma companies thus employ "an army of people" to read those emails. If anything strange is detected, "you have to report this information to another level," says Huot. "We are working on a system that replaces the human being to detect the potential association by reading the email. This is why we are growing our unit in Philadelphia," says Huot. "We put the system in production May 2007. It's a savings of hundreds of thousands of dollars per year" for the pharma client.

Although Luxid for Life Sciences encapsulates knowledge from chemists, biologists, and physicians, Huot acknowledges the importance of expert domain knowledge. The new Luxid release in early 2008 will feature social tagging — "the ability for Luxid to learn from its users," says Huot. "All automatic systems make errors — it's not a problem so long as you can learn from those errors and modify dynamically, and provide to your end user the correction very quickly. Luxid will integrate this ability to quickly integrate the knowledge from hundreds of scientists directly into the system."

— KD

researchers manage databases through text mining of the scientific literature. "MITRE tends to work at the infrastructure levels, we are typically system engineers and chief architects on large projects. On the research side, we tend to provide an architecture or evaluation," she says. "You get to leverage everybody else's progress, you get to watch and see what works... putting out the milestones for the researchers and reaping the results for the end users."

To that end, in 2003 Hirschman launched a community challenge cup for the systematic evaluation of biological text mining solutions. Her inspiration was hearing a report on CASP (the structural biology competition to evaluate methods of protein structure prediction) at a conference in 2001. Text mining systems were not directly comparable at that point, because they had not been trained on the same data sets or applied to real biological applications.

In collaboration with Spain's Alfonso Valencia and Christian Blaschke (now at bioalma), Hirschman created BioCreative: Critical Assessment of Information Extraction for Biology (with funding from the National Science Foundation).

Commercial/Open Source Pros and Cons

"Commercial [text mining] software is usually easier to use, often easier to integrate with other tools, and generally has someone to call for support issues," says Biogen Idec's William Hayes. "Academic software, on the other hand, is often leading edge but very difficult to use in a lot of cases, with limited recourse regarding bug fixes other than doing it yourself."

Academic software can be targeted to needs that are not commercially viable, and is "not often designed to interoperate with other tools," says Hayes. Of course, academic software is usually open-source, which allows for customized fixes and changes. This leads to the standard bio-IT question — pay upfront in commercial licensing fees or pay for more staff to do custom development with academic tools.

FILE PHOTO/FURNALD & GRAY

Hayes finds strengths in both commercial and open source.

Another issue, says Hayes, is scalability. "It can be very difficult to get the commercial solutions to scale, much less academic solutions. Many of the academic solutions I've seen are not designed for scalability but only to run against a research data set."

For competitions such as BioCreative, performance requires significant tuning to a specific use case to achieve results. "A commercial entity could get bragging rights if they came in first," says Hayes, "but it doesn't result in a strong competitive edge."

Most commercial products stress "features, functionality, and subjective performance" over accuracy, says Hayes. "I'm okay with that, because accuracy is incredibly dependent on the corpus (document collection) and what one is trying to extract from the literature. It's often more the accessibility of the information we are after — interface design coupled with reasonable accuracy without requiring a degree in computational linguistics to use it — and whether, with limited resources, we can get and keep an application up and running."

Nevertheless, Hayes says there are several academic/open-source technologies integrated with commercial applications for text analytics at Biogen Idec, or under consideration. These include:

- Cytoscape for network/graph visualization of extracted information;
- Exhibit (Semantic Web Simile project) for visualizing integrated competitive intelligence data mined from the web and databases;
- iHOP for protein co-occurrence analyses (Hayes says he's also interested in a commercial application, BioVista, which compares billions of co-occurrence relations between proteins, drugs, methods, cell lines, companies, adverse events, etc. from Medline, full-text journals, and patents);
- Protege for collaborative thesauri management;
- Abner (Ab initio named entity recognition).

"We'd like to use quite a few academic tools we read about, we just don't have enough bandwidth to integrate them," says Hayes. "The biomedical text mining community needs to encourage more interoperability of tools and applications." He points to the excellent example of the Jena University Language and Information Engineering Lab (JULIE Lab; http://www.julielab.de), which wraps its tools in IBM's open-sourced UIMA text analytics framework for interoperability.

Hayes appreciates the trend of leading academic work that migrates into the commercial space, "giving us the best of both worlds."

— KD

The first BioCreative workshop, in 2004 in Madrid, attracted 27 groups from 10 countries, including a few private companies, applying text mining tools to real biological challenges. "We tried to frame this as a collective endeavor to do well, as opposed to [focus on] winning," says Hirschman.

The groups could chose any or all of three tasks:

1. Extracting gene/protein names from text;
2. Translating those names into standardized gene identifiers for inclusion in three model organism databases (the most successful systems extracted gene names from Medline abstracts with about 80-90 percent accuracy);
3. Finding evidence for Gene Ontology (GO) annotations for protein function, biological process, and localizations (this task was the hardest, at 20-30 percent accuracy, although follow-up results have been encouraging).

In September 2007, Hirschman presented the results of BioCreative II (See, "Top Systems...") at the Fraunhofer Text Mining Symposium in Germany (See, "The Language of Text Mining"). The first two tasks in the second workshop were the same; the third was reproducing the steps of a biological database curation pipeline for protein-protein interactions, working with two databases, MINT (University of Rome) and IntAct (European Bioinformatics Institute).

The benefits of BioCreative include fostering collaborations and a detailed comparison of approaches. Says Hirschman: "One of the very interesting results in the gene mention and gene normalization tasks was if you pool the results from multiple systems, you got better results than from any one system." Even low-performing systems occasionally

found gene mentions that higher-ranking systems didn't find, boosting the tally to approximately 90 percent.

In another lesson from CASP, Valencia and Hirschman are trying to make text-mining tools more widely available to biologists, and easier to insert into data pipelines. That's not a trivial undertaking, particularly in biopharma organizations (See, "Commercial/Open Source Pros and Cons").

Hirschman chuckles when asked about the poor participation of commercial text mining providers in BioCreative. "They have not expressed interest at this point," she says. "There is limited financial incentive unless they're sure they will do well." But a few companies including bioalma did participate in the most recent contest.

Poised for Success

Hirschman believes that "text mining is *poised* to become an important tool," although she acknowledges that, "The uses of text mining on the pharma side are much bigger than the research side." She was particularly impressed by a Fraunhofer symposium presentation from Novartis' Thérèse Vachon, deploying a powerful system able to index and text mine various data (See, "Novartis' Answer to Harry Potter," *Bio•IT World*, Dec 2006).

"On the research side, especially if you're dealing with database curation, the information has to be really correct if it's going to be used and reused," says Hirschman. "On the pharma side, getting partial information is OK. If you can get it cheaply, that's good. So text mining tools in some senses may be a better match for certain kinds of pharma applications than they are for database curation applications."

Hirschman continues: "Pharma is ahead of the research community, and the reason why academic groups don't use commercial tools is they're very, very expensive. They've got a good market in pharma. We need to figure out how to do a better job of leveraging the progress on the commercial side (and the commercial tools that are available), to help the research community and the publicly funded databases — which are heavily used by pharma."

"The NIH needs to be worry about this [disparity]," says Hirschman. "Expert curated biological databases provide a critical resource — NIH needs to invest in affordable tools to cut the cost of curation." •

Top Systems at BioCreative II

According to MITRE's Lynette Hirschman, the top performing entries for the gene mention task at BioCreative II included an entry from IBM, led by Rie Ando Johnson, and two systems from Taipei: National Yang-Ming University and Academia Sinica. All three systems used advanced machine learning techniques.

The best performances in the gene normalization task all came from Germany: Jörg Häkenberg (Technical University, Dresden), Katrin Fundel (Ludwig-Maximilian University, Munich), and Juliane Fluck (Fraunhofer Institute). Häkenberg's system used synonyms and contextual information to associate gene identifiers with mentions in text. The other groups used large-scale synonym lexicons coupled with string matching techniques.

Among the top groups in the protein-protein interaction annotation task were Larry Hunter's lab (Univ. Colorado School of Medicine) and Claire Grover's group (University of Edinburgh). Both combined machine learning with manually derived rules to create pipelines to flag articles for curation, find relevant genes and proteins, and identify specific interacting proteins.

BioCreative
Critical Assessment for Information Extraction in Biology

Register by January 25 and Save up to $500!

Cambridge Healthtech Institute's Sixth Annual

Bio·IT World Conference & Expo

April 28 – 30, 2008 • World Trade Center • Boston, MA

Keynote Presentations by:

Linda Avey
co-Founder
23andMe, Inc.

Joshua Boger, Ph.D.
President & Chief Executive Officer
Vertex Pharmaceuticals, Inc.

John Reynders, Ph.D.
Vice President & Chief Information Officer
Johnson & Johnson, Pharma R&D

2008 Benjamin Franklin Award presented by Bioinformatics.org

Keynote Panel:
The Future of Personal Genomics: A special plenary panel discussion featuring
George Church, Ph.D., Harvard Medical School, Dietrich Stephan, Ph.D., Navigenics, and more

Concurrent Tracks:

1. Informatics and IT Infrastructure & Operations
2. Informatics for Genome Analysis, Biomarkers, and Target Discovery
3. Predictive and *in silico* Science
4. eChemistry Solutions
5. Clinical and Medical Informatics

A Must-Attend Event for the Life Science IT & Informatics Community!

Pre-Conference Workshop Update:

Monday, April 28, 1:30-4:00pm

Next-Generation Sequencing
Data Management Workshop
organized by the BioTeam

Featuring George Church (Harvard Med/Knome), Matt Trumbull (Broad Institute) and leading vendors.

Corporate Sponsors: BIOTEAM Enabling Science, SCHRÖDINGER, TATA CONSULTANCY SERVICES

Official Publication: Bio·IT World

Lead Sponsoring Publications: GEN Genetic Engineering & biotechnology News, Digital HealthCare & Productivity, Bio·IT World, eCliniqua, Science AAAS, TheScientist

Organized & Managed by:

Cambridge Healthtech Institute
250 First Avenue, Suite 300, Needham, MA 02494
Phone: 781-972-5400 • Fax: 781-972-5425
Toll-free in the U.S. 888-999-6288

www.Bio-ITWorldExpo.com

Event Features:

- **Access** All Five Tracks for One Price
- **Network** with 1,500+ Delegates
- **Hear** 85+ Technology and Scientific Presentations
- **Choose** from Four Pre-conference Workshops, the Oracle Life Sciences and Healthcare User Group Meeting, and Symyx Software Symposium
- **Attend** Bio-IT World's Best Practices Awards **NEW**
- **Connect** with Attendees Using CHI's Intro-Net
- **Participate** in the Poster Competition
- **See** the Winners of the following 2008 Awards:
  Benjamin Franklin
  Best of Show
  Best Practices
- **View** Who's Who & What's New in the Exhibit Hall
- And Much More!

Presentations from:

Aditya Phatak, Persistent Systems Inc.
Alan Higgins, Ph.D., Viamet Pharmaceuticals
Alan Ruttenberg, Principal Scientist, Science Commons
Anton Petrov, Ph.D., Cytogenetics, infoQuant, Ltd.
B. Woody Sherman, Ph.D., Schrödinger, Inc.
Barry Bunin, Ph.D., Collaborative Drug Discovery, Inc.
Belinda Chen, Ph.D., Institute for Information Industry
Bernard Wess, ProtonCare and PERSEID Software
Chris Dagdigian, BioTeam, Inc.
Chris Dwan, BioTeam, Inc.
Christian Reich, Genstruct, Inc.
Craig Lipset, Pfizer
David Chiang, Sage-N Research, Inc.
David de Graaf, Ph.D., Pfizer
David DeCaprio, Broad Institute/MIT
David Jordan, Lineberger Comprehensive Cancer Center
Dietrich Stephan, Ph.D., Navigenics
Dinesh DeAlwis, Ph.D., Lilly & Co.
George Church, Ph.D., Harvard Medical School
George Redmond, National Cancer Institute
Georges Grinstein, Ph.D., UMass Lowell
Peter V. Henstock, Ph.D., Pfizer Research Technology Center
James Reaney, BlueArc Corporation
Jason Johnson, Ph.D., Inpharmatics/Merck & Co., Inc.
Jean Peccoud, Ph.D., Virginia Tech Bioinformatics Institute
John Quackenbush, Ph.D., Dana-Farber Cancer Institute
John Reynders, Ph.D., Johnson & Johnson, Pharma R&D
Jomol Mathew, Ph.D., Dana-Farber Cancer Institute
José S Duca, Ph.D., Schering-Plough Research Institute
Josef Scheiber, Ph.D., Novartis Institutes for Biomedical Research
Joshua Boger, Ph.D., Vertex Pharmaceuticals, Inc.
Keith Boyce, Immune Tolerance Network / UCSF
Kenneth Buetow, Ph.D., National Cancer Institute
Kevin Davies, Ph.D., Bio-IT World
Khaled El Emam, Ph.D., TrialStat Corporation & University of Ottawa
Knut Wittkowski, Ph.D., D.Sc., The Rockefeller University
Les Jordan, Microsoft
Linda Avey, 23andMe, Inc.
Linda King, Eli Lilly and Company
M. Vidyasagar, Ph.D., Tata Consultant Services
Marco Ramoni, Ph.D., Harvard Medical School
Marti Jett, Ph.D., Walter Reed Army Institute of Research
Michael N. Liebman, Ph.D., Windber Research Institute
Olive Yuan, Ph.D., Regeneron Pharmaceuticals. Inc.
Patrick Aloy, Ph.D., Anaxomics Biotech
Philippe Stas, AlgoNomics
Phillips Kuhl, Cambridge Healthtech Institute
Prof. Zvia Agur, Optimata
Rakesh Nagarajan, Ph.D., Washington University School of Medicine
Richard Cramer, Ph.D., Tripos
Richard Gill, Ph.D., ProNAi Therapeutics, Inc.
Robert Deans, Athersys, Inc.
Robert Feinstein, Ph.D., Kelaroo, Inc.
Ronald Waife, Waife & Associates, Inc.
Samuel (Sandy) Aronson, Harvard Partners Center for Genetics and Genomics
Saul Kravitz, Ph.D., J. Craig Venter Institute
Suresh Singh, Vitae Pharmaceuticals
Tatyana Zhukov, Ph.D., H. Lee Moffitt Cancer Center & Research Institute
Towia Libermann, Ph.D., Beth Israel Deaconess Medical Center
Ulrik Nielsen, Ph.D., Merrimack Pharmaceuticals
Vadim Sapiro, J. Craig Venter Institute
Vipul Kashyap, Ph.D., Partners Healthcare System
Yingyao Zhou, Ph.D., Genomics Institute of the Novartis Research Foundation
and more!!!!!

Make sure to mention keycode U02 when registering!

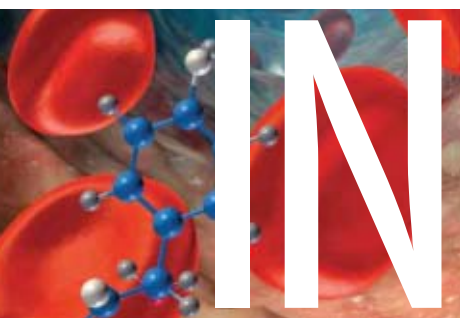
Bio-IT World will present the Awards for its 2008 Best Practices competition at a special evening ceremony during the Bio-IT World Conference & Expo. Entries are currently being sought for outstanding examples of novel technologies and strategies enhancing the success of operations in any part of the research/drug discovery pipeline. Winners will be selected by a peer-review expert panel in early 2008. Complete information, including online entry forms, is available at **www.bio-itworld.com/bestpractices**

www.Bio-ITWorldExpo.com

Exploratory INDs: A Trial Before the Trials

Phase 0 trials and the use of **tiny drug doses** are gaining acceptance and winning converts.

By Mike May



IN 2006, Merck created a new department dedicated to experimental medicine. This is not a new idea — not even at Merck, where this approach has been used for several years — but the priority is. Within one year, Merck went from zero to more than 50 people dedicated to experimental medicine, and the department is widely credited with accelerating the approval in 2006 of Januvia, a promising anti-diabetes drug. "We want the best scientists living and breathing this 24 hours a day," says Gary Herman, head of Merck's department of experimental medicine.

Herman defines experimental medicine as "small clinical trials in a limited number of subjects or patients — who are highly controlled — to get an assessment of the pharmacological activity for compounds. The group is committed to creating innovative ways to establish proof-of-concept for new mechanisms by exploring clinical models that could yield better, faster, and more cost-effective ways to prioritize drug development."

Biotechnology and pharmaceutical experts have a number of pet names for what to call a small-sample size study on humans before traditional clinical trials: exploratory investigational new drug (IND), Phase 0 (See, "When, Why and How"), pre-Phase I. But they all mean pretty much the same thing. "It's probably best to use the term exploratory IND," says Chris Elicone, senior marketing manager at Applied Biosystems. "At least the FDA has provided some guidance there."

Elicone refers to the FDA's "Guidance for Industry, Investigators, and Reviewers: Exploratory IND Studies," which underscores that these studies have "no therapeutic or diagnostic intent." Instead, exploratory INDs can be used, for example, to compare similar compounds to select the preferred one to move ahead. To gather pharmacokinetic data in an exploratory IND, scientists can use microdosing, which involves one compound tested through a single dose — which must be the lower of two options: 100 micrograms or 1/100th of the pharmacological dose based on animal studies.

By any name, though, exploratory INDs aim to push the more promising drugs into clinical trials and increase the odds of success. Negotiating that path, though, depends on solving many informatics and IT challenges.

Getting to Phase I

From an informatics perspective, the crucial question is: How do companies use data from an exploratory IND to decide if a drug will move ahead?

"Multiple factors go into that decision," says Mark Schmidt, director and clinical expert for CNS at Johnson & Johnson Pharmaceutical Research and Development (J&JPRD). "The largest factor is the risk-benefit ratio." A little more risk can be accepted if a compound potentially fills an unmet medical need. Likewise, the strength of the complete preclinical package comes into the decision-making process regarding moving ahead a compound.

Anastasia Christianson, senior director of discovery medicine informatics at AstraZeneca Pharmaceuticals, agrees that results from a short, small study seldom determine whether a compound moves into clinical trials. "We'll ask: How does this study relate to the studies done before and the biology of the disease and the unmet medical need?" Christianson explains. "The decision is based on all of the information that we've collected." She adds, "It's a big information-management and -utilization problem, because we collect more and more information along the way."

When Merck applied an experimental-medicine step to Januvia (sitagliptin), its recently approved DDP-4 inhibitor for Type II diabetes, the researchers developed a crucial single-dose study in patients. "Then, we modeled various concentrations of this diabetes drug based on the response of biomarkers, and this let us narrow the dose range to study in later stages," says Herman. "It cut at least a year off our timeline." In fact, Herman says the predicted dose from the early Phase I study ended up being the clinically approved dose. In all, the drug spent less than four years in clinical trials.

Dealing with the Data

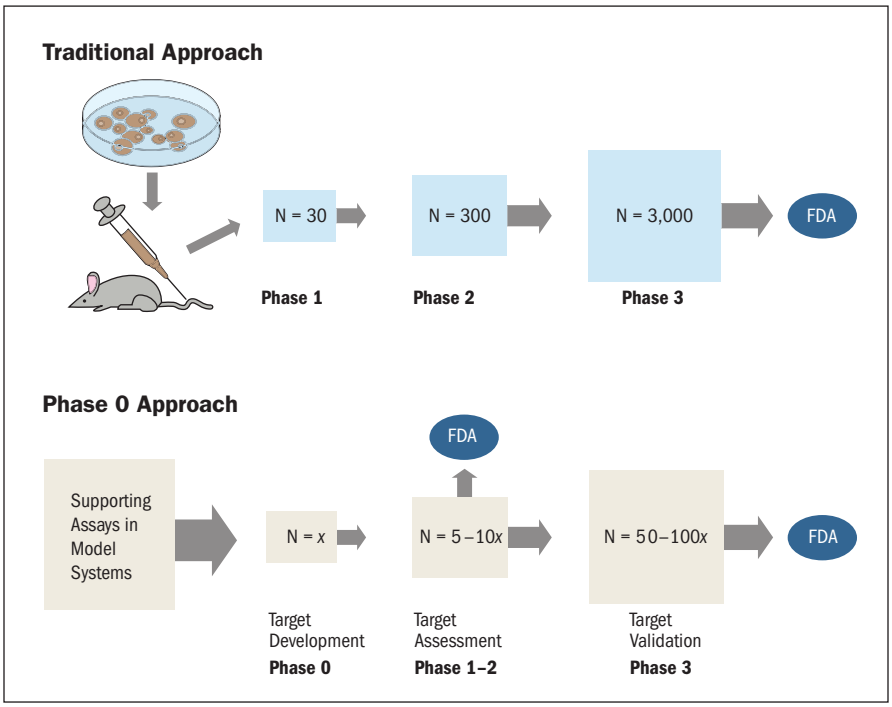
"The data analysis tools for these studies are similar to the ones that we apply in clinical trials," says James Bolognese, senior director of clinical biostatistics at Merck. Common approaches used at Merck rely on SAS software and some tools developed with S+. Moreover, Merck's genetic-profiling researchers use network-analysis software. Bolognese adds, "People in our Rosetta [Inpharmatics] group are experts in the systems biology approach to analyzing problems."

Merck scientists are also working on adding some Bayesian techniques. "In conventional analysis," explains Bolognese, "you compute the probability of false positives and negatives based on the data in a trial. With Bayesian techniques, we incorporate prior information — information that existed before the trial." That previous knowledge might come from animal studies or published data in the literature. "This can tighten the predictive ability of data gathered in a trial," says Bolognese.

At AstraZeneca, Christianson notes, "One of the biggest challenges is making sure that you have the information integrated in the right way to use it. There is a technical component, a policy component, and a human component." On the technical end, Christianson points out that data must be stored where users have easy access and can integrate data from various experiments. "The policy component is around what you can and cannot do ethically, such as what kind of analysis can be done," she says. "On the human side, the analysis needs to be interactive, because you have people with different areas of expertise doing the analysis together."

Christianson and her colleagues take on these challenges with some in-house products and other off-the-shelf ones, such as SAS and Spotfire. "We also might take something off-the-shelf and modify it," she says.

Schmidt at J&JPRD points out the need also to do pharmacokinetic and pharmacodynamic modeling. He discusses the need to compare results from exploratory IND with data from preclinical animal studies, such as ADME



Comparing the traditional approach to developing an oncology drug (top) versus adding Phase 0 (bottom), James Doroshow and his colleagues at the National Cancer Institute illustrate that significantly fewer patients can be used in the latter approach. In the lower development plan, x represents the number of patients needed for the Phase 0 trials, which the authors set at 10–15. Adapted from Kummar, S. *et al.* 2007. *Nature Reviews: Cancer* 7:131–139.

When, Why, and How

Phase 0 is "not just another clinical trial," says James Doroshow, director of cancer treatment and diagnosis, National Cancer Institute. "It's really the end of pre-clinical" research where the "intent is biomarker development."

Speaking at a recent American Association of Cancer Research conference in Singapore, Doroshow discussed the challenges and promise of Phase 0 trials culled from the "first Phase 0 oncology trial." His goals were to determine a non-toxic dose range at which a candidate inhibited PARP (poly (ADP-ribose) polymerase) in tumors and in peripheral blood cells.

"We knew in three to four weeks that we had results," he said. The Phase 0 study established that the drug inhibited the target at clinically achievable concentrations. Researchers were able to perform real-time pharmacokinetic (PK) and dynamic (PD) analyses, netting results within 72 hours, and develop standard procedures for processing tissue samples and gathering PK/PD data.

"Why would you not do it?" Doroshow asked rhetorically. He concedes that Phase 0 studies require "significant investments of resources... a multidisciplinary team is crucial." But Doroshow considers the expenditure worthwhile. Phase 0 "helps point the way to data needed for phase 1." It saves time and offers toxicity data early in the process. As more money flows into biomarker assays, researchers will be designing "smaller but smarter trials," he said.

There are challenges, both statistical and ethical. Barriers to enrollment include studies with no therapeutic intent, and sometimes the need for pre- and post- study tissue biopsies. But these drawbacks can be overcome with a well-run informed consent process and a clearly expressed rationale. Statistically, limited sample sizes can cause problems, but Doroshow believes those can be managed by obtaining a measure of intra-patient variability for the pre-treatment endpoint value.

Phase 0 trials are not always appropriate, but Doroshow believes they should be considered for a targeted drug with a wide therapeutic index under development for chronic or multi-dose oral administration. He cautions that Phase 0 will not generally be appropriate for cytotoxic agents with very narrow therapeutic indices delivered intravenously. — Allison Proffitt

(absorption, distribution, metabolism, and excretion). But then he says, "The technical challenge is minor; the greater challenge is institutional."

"We need to reshape how people report data," says Schmidt, "and how that can be harmonized between different functions to make them compatible." For example, if studies on a new compound lead to 30,000 data points from proteomics and even more voxels of imaging data, "you can't do a correlation analysis on that," says Schmidt. "Comparing those data requires some judgment, and that requires IT bringing different experts together."

To help IT solve such problems requires strong communication. "When working with IT internally or externally," says Schmidt, "the main thing that I see is them wanting to know more about the nature of the clinical phenomenon. For me, I need to know how they will use the various databases, how defining a field parameter matters — things like that."

Downsizing Doses

In microdosing, scientists use such small amounts of the drug candidate that it can be difficult to measure levels in patient samples. With accelerator mass spectrometry, for example, Colin Garner, CEO of Xceleron, says that compounds can be detected at levels as low as 1 attogram, or $1x10^{-18}$ grams. This technique arose from carbon dating, so some computation takes place to turn data into pharmaceutical terms. For pharmacokinetic analysis, Garner says, "We typically use the Pharsight product, WinNonlin."

To detect very low levels of a compound in a sample, scientists can also turn to liquid chromatography (LC), followed by two sequential stages of mass spectrometry (MS). "From the LC/MS/MS perspective," says AB's Elicone, "several pharmaceutical companies have reported success in proof of concept using the API 4000 system in exploratory INDs." Another company, he says, purchased an API 5000 system (also an LC/MS/MS instrument) just for microdosing studies. "The LC/MS/MS approach benefits pharma," says Elicone, "because they already have the technology in-house, and they are very familiar with the operational workflow." Such techniques can detect compounds at levels as low as a few picograms/milliliter, which is sufficient sensitivity for microdose studies.

Firms such as Xceleron work with biopharma customers on a consulting basis. "We assist companies in the candidate-selection process," says Garner. "Microdosing can give them a very early read on the human metabolism of a compound, which is a crucial parameter for a new drug."

In some cases, a microdosing approach can also reduce the time and cost of getting a compound to first-in-human studies. According to "Microdosing in Translational Medicine: Pros and Cons," a 2006 Insight Pharma Report authored by Hermann Mucke, conventional approaches require 12–18 months and $1.5–3 million to get a compound into humans; microdosing takes just 5–8 months and $300–500 thousand.

Expansion of Exploratory INDs

No one knows how many companies have used exploratory INDs or on how many drugs. Mucke says he was unable to find any hard figures for the Insight Pharma report. He estimates that several dozen companies have used the approach, although "most companies who do microdosing decide not to announce it for some reason."

One case Mucke did cite in the Insight Pharma report was at Millennium Pharmaceuticals, where scientists used LC/MS/MS "to evaluate dose proportionality under microdosing conditions for two established drugs (fluconazole and tolbu-

Upsides and Downsides of Microdosing vs. Conventional Pathway to First-time-in-Human Milestone

	Conventional Approach	Microdosing Approach
Time and cost from selection of preclinical candidate to finalized first-in-human study	12–18 months $1.5–$3.0 million	5–8 months $0.3–$0.75 million
Approximate minimum amount of compound required for first-in-human-study and pre-qualification (depends on potency and bioavailability)	100 grams	0.1–1 grams
Predictive power for pharmacokinetic parameters at pharmacologically effective doses	Definite	Generally good if mass effects and/or protein binding make no significant contributions
Need for ^{14}C-labeled compound for first-in-human study	No	Yes (if AMS is used) No (if LC/MS/MS is used)
Available options for outsourcing	Huge number of certified preclinical and clinical-stage CROs and analytical laboratories in all major pharmaceutical markets	Use of AMS requires certification of clinical CRO for ^{14}C work; analytics restricted to a handful of highly specialized providers
Standardization and degree of establishment of regulatory path	Firmly established and internationally harmonized through ICH guidelines; few if any variations possible	Very new — authorities and developers are on a learning curve; U.S. and European regulations not identical in some points

Source: Cambridge Healthtech Associates

tamide, chosen because of their similar pharmacokinetics characteristics in rats and humans) and an investigational Millennium compound identified as MLNX in rats." Another example is CRO Radiant Research, which used microdosing to develop a new pharmacokinetic profile for AZT (azidothymidine), the anti-HIV drug. Mucke also reported that GlaxoSmithKline "has a traditional involvement in microdosing that stems from its role as one of the cofounders of CBAMS, the company that later became Xceleron."

Perhaps most important, companies are learning when to turn to exploratory INDs. "It's not for all situations, and not for everybody," says Mucke. "It doesn't really save time overall, because you must still do Phase I."

AB's Elicone sees companies becoming more aware of what exploratory INDs can really do. "A first, it seemed like everyone was talking about doing exploratory INDs with all compounds that met preclinical criteria," he says. Now, he says the trend is that "pharmaceutical companies are becoming much smarter about when to apply it." (See, "Not for Everyone...")

Some companies, though, are forging ahead in a broad way. Although Merck put much of it's first concentrated experimental-medicine effort into diabetes, that's just a start. "We want to use experiment medicine wherever possible," says Herman. He points out that Merck is working on platforms for oncology and neurological areas, putting a premium on imaging technologies (and the IT to handle them).

Even if companies can't agree on everything about this approach, most see the value of exploratory INDs. It could cut costs, get promising drugs into clinical trials faster, reject doomed drugs sooner, and invigorate the pharmaceutical pipeline once more. •

Not for Everyone ... Yet

"We haven't done Phase 0 or microdosing studies," says Dave Bearss, chief scientist at SuperGen, a Dublin, Calif.-based cancer biopharma. "But we've spent a lot of time trying to figure out how they might help us." Says Bearss: "By the time we did a toxicology package needed for a Phase 0 study, we'd almost have what we needed to do a Phase I study. What's the cost-benefit of the Phase 0?"

For SuperGen, the key to microdosing studies would be getting into humans faster. "I mistrust animal models," says Bearss, although he's spent much of his career trying to make better ones. "Mice and rats don't make reliable human models, and even a monkey doesn't make a good human model."

Still, animal models do lots of work. "Exploratory INDs don't change how we would view using animal studies," says Brian Zambrowicz, executive VP and chief scientific officer for Lexicon Pharmaceuticals, "but it could help with ADME or pharmacokinetics to make a potentially difficult choice between promising drug candidates." He adds, "You need animal pharmacology to see if it is useful to take a compound into humans."

SuperGen might turn to Phase 0 studies if it needs to choose priorities from a group of similar compounds. In fact, SuperGen might face that situation soon. "We have three leads — two of them from the same chemical scaffold, and they are all pretty equivalent in the lab," says Bearss.

Zambrowicz and his colleagues see that benefit, too. "We considered using an exploratory IND in a case where we had two compounds, and all of our toxicology and animal pharmacology studies made these compounds look equivalent." But Lexicon hasn't turned to exploratory INDs yet. — MM

2002 FILE PHOTO/CHRIS HARTLOVE

Reynders Takes the CIO Reins at J&J R&D

He lists translational science and integration among the key challenges in his new position.

Last summer, John Reynders became CIO for Johnson & Johnson's Life Sciences Division (succeeding John Barbano, who is now managing J&J's medical devices and diagnostics group). Reynders' rapid ascent in life sciences started in 2001 when he joined Celera as VP of Informatics. Celera was sitting on roughly 100 terabytes of genome data, and Reynders was suddenly responsible for all supercomputing capabilities, discovery software engineering, and enterprise system infrastructure. (See "Scaling Celera's Mountain of Data," Bio•IT World, April 2002). Not that he was likely to be intimidated. Reynders joined Celera from Los Alamos National Laboratory where he'd been director of the Advanced Computing Laboratory. He earned his Ph.D. in applied and computational mathematics from Princeton University.

Celera's fortunes, of course, took a downward turn but Reynders' trajectory did not. In 2004, he moved to Lilly Research Laboratories as information officer, discovery and development informatics. There, he encountered a vastly greater diversity of informatics tools, and he now says one of his most satisfying achievements was leading the effort to build a much more integrated informatics environment at Lilly. Reynders, who will deliver the opening keynote at the 2008 Bio-IT World Conference & Expo in Boston, recently spoke with Bio•IT World editor-in-chief Kevin Davies and executive editor John Russell about his new role and plans at Johnson & Johnson.

BITW: Tell us how the J&J opportunity came about?

Reynders: As with most things, it started with a phone call. I knew of J&J from afar, and was quite happy at Lilly and plugging away on the programs there. But when the opportunity presented itself, being able to look at not just the pharmaceutical perspective but at pharmaceuticals in the context of devices, in the context of diagnostics, and the role of technology at that intersection, that was quite exciting and intriguing.

You spent three years at Lilly — what do you consider your most gratifying accomplishments there?

There were many things such as the people and the company. If I were to point to a programmatic activity, I'd say it was the deployment of an integrated informatics program. This really involved how you think about [the role of] informatics across the pipeline and across multiple disciplines... Informatics as a discipline really shouldn't be bioinformatics by itself, cheminformatics by itself, or medical informatics by itself, but really our challenge is more the connection between these domains.

Biomarkers, for example, are getting to be very challenging. Gone are the days of, here's an SNP and here's a molecular diagnostic and off we go. It's really now a matter of finding signals in a combination of imaging data, expression data, and SNP data. So deploying that program and getting it underway at Lilly and seeing it actually come to life and impacting the pipeline is probably the thing of which I was most proud.

What is the scope of the IT group you've inherited at J&J and some of the challenges you'll face?

It's quite a global organization. J&J Pharmaceuticals is structured into three franchises. One focused on biotechnology, which is large molecules and biologics with a strong oncology focus, as well as inflammation. Then there's the CNS/IM focus, which is central nervous system and internal medicine (cardiovascular, arthrosclerosis, and endocrine) and a strong virology franchise — infectious diseases, HIV, hepatitis C. My responsibilities are

the IT and informatics across these three organizations, not the central IT. We have a central organization that does the data center, the networks, the PDAs, security, things like that.

My challenge — and the opportunity — is how you make sure you're bridging between the science and technology. I remember giving a keynote at your first Bridging IT and Discovery conference (See, "Bridges and Boundaries," *Bio•IT World*, November 2005). I could whip out that deck and go through the same list: how you blend the community, how you have common goals, how do you form teams that can do everything from the science to the technology and are agile in terms of moving across these domains, how you move decision making close to the projects…

I was delighted to see there were a lot of pieces in place when I arrived here, so I can focus on the value creation part sooner. I think the extra challenges are there's a broad set of therapeutic areas, each of which has nuances. You'll have a lot of imaging, for example, in some therapeutic areas, you'll have more gene expression and proteomics in others. How do you find connections between these therapeutic areas?

Finally, J&J is a very distributed organization, so that's exciting. It's almost like going from polar opposites, where Lilly is probably one of the more centralized pharmaceutical companies you could find, and J&J would certainly be one of the more distributed pharmaceutical companies. So, it's navigating in this culture of innovation spread into many centers. There's also being part of a larger company where we can reach out from pharma into diagnostics, devices, and consumer.

How is this position different from your role at Lilly?

My role at Lilly was focused more in discovery and CMC (chemistry, manufacturing, and controls), and building a biomedical informatics capability to support translational science in partnership with the clinical information officer at Lilly, whereas here, my accountability spans the whole pipeline. I find that exciting because a lot of the future in terms of informatics is going to be in translational science, going from bench to bedside and back to bench, but also health outcomes. How do you go from the clinic out into the real world and from the real world back into the clinic, and frankly, back into discovery? So, that entire closed loop between health surveillance and health outcome into the clinic and into discovery is what's different.

In terms of roles here, there are a few different CIO titles. My scope of responsibilities is all of IT for pharma R&D.

My challenge — and the opportunity — is how you make sure you're bridging between the science and technology.

I have, for example, a colleague who is responsible for all of operations at the supply chain or manufacturing. I have another colleague who's responsible for all of commercial operations and the three of us report up to a single CIO for all of pharmaceuticals.

What will you do differently at J&J?

I'll be looking more opportunistically. The value creation opportunity at J&J is finding connections between entities, so for example, reaching out into Veridex or Virco [Labs], which are [J&J-owned] diagnostic companies and seeing where we might find new products. In other [companies], I think it's more, here's the portfolio, execute the portfolio, reprioritize, put in new projects, and off we go. That's a critical part of how the work is going to be done hertze. But it's certainly the case that J&J being very distributed, there's more energy that I think needs to be spent in finding the value creating opportunities between the operating businesses and between the different sectors here: pharmaceuticals, medical devices and diagnostics, [and] consumer.

One thing I find that's very intriguing is J&J is more of a virtual company. I can do quite a bit virtually and that makes it easier to lead and manage… The meetings are more done by communication, by email or by teleconference. For me personally, I am able to move more quickly with coordination so distributed and asynchronous.

Have you thought about a near term strategy and specific goals?

The main thing I'm focused on now is getting my hands around the various biomarker programs we have. Again, if we're going to talk translational science and health outcomes, it really boils down to what are the biomarkers — be it for pharmacodynamics, be it for translational science, be it for patient stratification, be it identifying signals out of the population in terms of health outcome. All of that is materialized in a biomarker. This in turn drives an informatics strategy; what do we need to do in terms of imaging, pharmacogenomics, metabonomics, all the way from discovery to clinic. A lot of my early focus is understanding the biomarker platforms, kind of doing an inventory, understanding the scientific capabilities, and that'll in turn drive my informatics strategy.

Are there major projects either you're inheriting or considering starting?

It's a little early for that. I'm 60 days in, so I'm still in data gathering mode, but if I were to give you a general view, it's going to be something that looks like how do we integrate? Integration is going to be a theme I'm going to be saying again and again. How do we integrate across the pipeline and across platforms?

What's your view on the right mix of internally developed versus commercial tools?

I'm a big fan of frameworks because of what they allow you to do. If you have an SOA and frameworks, it allows you to be targeted in what your components are and to decide which components are commodities that you just want to buy versus which are strategic that you want to build. So, my approach with third party vendors and partners is [to ask] how can we work together to build systems? How

can we take pieces and parts of what we do and combine it?

I'm not a fan of the large monolithic vendor solution. I think [that's] the same thing a lot of my CIO colleagues in other places are saying. I remember actually sitting in a Bio-IT World Conference panel with colleagues from Merck and Pfizer saying the same thing — gone are the days of companies buying large standalone third-party platforms. So, if I need to be integrating my data, integrating my information, integrating my algorithms, the partners I'm going to be working with are those that can work in an integrated environment.

Are there cultural areas where you think J&J can improve?

What's helped make this transition smooth is that I find a lot of cultural similarities between Lilly and J&J. They're both companies that focus upon integrity, excellence, and the people. If I look at J&J, its credo is something that I found very compelling in terms of the kind of environment I want to work in and build high-performance teams that deliver excellence. The culture is collaborative, so it is possible to bridge across the science and technology. I see many examples of that working. There are areas where I'll need to build those bridges that haven't been built yet, but the culture of J&J was something that was a big attractor to me in terms of coming here.

Is J&J more focused on developing successful drugs through the pipeline, or vetting the candidates you have so they fail sooner to save money?

As with most any other pharma, no one's going to say no to a blockbuster, but we are realizing that patient populations are different. Again, getting back to biomarkers, it's about how do you find the right patient for the right drug, the right dose, the right time; all that targeting. At the end of the day, the bars are going up across the industry in terms of what the patients expect in terms of the investment they make in pharmaceuticals. So we want to have a lot of clarity around the health outcomes that are being delivered, not just in a general sense, but the heath outcomes to specific patient population.

Is the concept of translational medicine working, or are there still challenges in getting clinical researchers to communicate with discovery teams?

There are a couple of challenges with translation. First, translation is two-way. Sometimes the mistake is made that it is one-way; that you find a way to have an improved animal model or an improved *in vivo/in vitro* set of models that can help you with a pharmacodynamic endpoint that brings things into the clinic. But what tends to be missing in translation is how you go back from the clinical result to help improve the pharmacodynamic endpoint or the biomarker in the first place. So, I think the general challenge in translational sciences is how do we have it not be a diode on the pipeline where it only goes one-way.

You mentioned that J&J has a lot of subsidiaries in the medical devices and diagnostics field. What about that appeals to you?

I really think the future is going to be how you target drugs, how you get the right drug to the right patient. What is exciting about J&J is the ability to converge all these technologies. It's not like a pharmaceutical company that would then have to go partner with a diagnostic company and ask the diagnostic company why [something works].

Information is the back plane that makes that happen. How do you understand the patient population, how do you understand the nature of the composite biomarker, how do you instantiate that into a diagnostic? Working with different parts of J&J to bring targeted therapy, personalized medicine to life — that's probably what I found most exciting in this opportunity. •

Emerging Trends in Top 500 Supercomputer List

BlueGene once again finds itself on top.

BY SALVATORE SALAMONE

The 30th edition of the **Top500** list of the world's most powerful computers, released in November at the SC07 International Conference for High Performance Computing, Networking, Storage, and Analysis, further validated IBM's much vaunted Blue Gene system.

Topping the list again was the BlueGene/L System at the Lawrence Livermore National Laboratory. This system has held the top spot since November 2004. Its current Linpack benchmark performance is 478 teraflops, up from 280 teraflops just six months ago. In all, IBM placed four Blue Gene systems in the top 10. They are joined by **Cray**, with three entries in the top 10, **HP** with two, and **SGI**.

Cluster systems dominate the Top500 list, with 406 systems in the new ranking. Most life sciences organizations have shifted much or all of their research off monolithic mini-computers and onto clusters. Similarly, the use of standards-based, high-performance interconnection technology continues to grow. Replacing proprietary solutions, Gigabit Ethernet is now the most common internal connection technology, in use by 270 of the systems. Multi-core processor-based systems are now the dominant chip architecture in the Top500 list. According to the list's organizers, "The most impressive growth showed the number of systems using the **Intel** Clovertown quad-core chips which grew in six months from 19 to 102 systems." Most of the remaining systems use dual-core processors. •

BlueArc's Titan 1100 Targets Cost Conscious

Newest model cuts cost for smaller customers.

BY KEVIN DAVIES

Although BlueArc continues to "double down" and push the performance of its data storage devices, the company hopes its latest model, the Titan 1100, will appeal to an entirely new customer base.

"For those customers who are absolutely price sensitive, and need to get in around say $75K, instead of $100K, we have an answer," says Louis Gray, BlueArc's director of corporate marketing.

San Jose-based BlueArc, which recently filed for an initial public offering, currently has a customer base of more than 225 clients, of which about 30-50 are in the life science/higher education field. Among its notable life science clients are the Washington University of St. Louis genome sequencing center, and the Wellcome Trust Sanger Institute in the U.K. Cray is a supercomputing partner with BlueArc.

Introduced in 2006, the Titan series exists in three formats — 2100, 2200 and 2500. The architecture of the new Titan 1100 model is "fundamentally unique, it's patented, [and] designed for very intense high-end computing," says Jon Affeld, senior director of product marketing and business development.

"To bring that down to the midrange is actually very straightforward, it's a matter of cost reduction," Affeld says. Those cost reductions come chiefly from the removal of two Ethernet ports (four instead of six) and a reduction in processing power and memory. What the 1100 does offer is support for two-node clustering and capacity of up to 128 terabytes.

Affeld explains that BlueArc is taking the same software, data management, and virtualization features from its high-range systems and deploying them on the midrange 1100 product. The goal is to open up new markets, particularly smaller biotech companies, university departments, and research centers.

"I think the 1100 will be very popular in the universities that are doing research," says Affeld. "You can start with the 1100, it's a very powerful, capable system. But what if you were wrong in terms of sizing it? It's very elegant to upgrade." Indeed, upgrading from an 1100 to a 2500 is made possible by the common software feature set and modular design. Physically substituting new blades make for a relatively simple upgrade process. •

Information Capture Across Organizations

Org changes are crucial to connect discovery teams.

DANIEL WEAVER & MICHAEL CAPSAMBELIS

Drug discovery doesn't necessarily accelerate when companies integrate formerly stovepiped databases. While overcoming technical barriers that segregate data streams is an important prerequisite, accelerating drug discovery ultimately hinges on bridging the organizational divides that isolate people, projects, and knowledge.

For example, executives, project managers, and scientists must routinely interrupt discovery progress to track down data and prepare progress reports. Yet, often, they must still take action without all necessary information, even when it exists within their own organization.

Often, observations and intuitions arising from one person's research are pertinent to work happening elsewhere in the company, but most team members are unlikely to know when their own ideas might augment others' hypotheses. Currently, the understanding one team has of another's thinking is confined to periodic, "batch" updates that cover only the highlights. This limits prospects for impromptu collaboration that hasten drug discovery.

Life sciences informatics applications are evolving to address organizational integration. People use these applications to explore data and to get an up-to-date, dynamic picture of what others involved with the discovery pipeline are doing. The goal is deep, ongoing situational awareness: a mutual understanding among executives, managers, and scientists of the many issues that affect discovery. This shared awareness relies on both traditional, structured data (scientific and financial) and human knowledge and interpretation, such as team members' ideas, hypotheses, and plans.

Applications that integrate primarily traditional data streams are effective for such tasks as seeing and exploring an outline of the portfolio's recent performance. However, these structured drug discovery data represent the endpoints of past action: last week's assay results, the past quarter's earnings, etc. In contrast, applications that combine traditional data streams with researchers' questions, observations, and goals are innately future-oriented. They furnish fluid, forward-looking pictures of the discovery pipeline and its component projects that experienced people use to coordinate more effectively with others throughout the organization.

Knowledge and Composability

Capturing a continuous, up-to-date flow of human knowledge in a useful form is challenging. Applications that accomplish this generally have two important characteristics.

First, they collect diverse human insight and expertise in ways that can be pooled with structured, visual data, and shared and explored in the same context. The knowledge-capture mechanism must be flexible. Because it's impossible to define in advance what information will lead to spontaneous insights, conventional tools such as templates are too limiting.

Second, they capture information automatically, as a natural outgrowth of people's daily activities. Processes that require people to stop work and explicitly summarize their thoughts for inclusion in a database are inefficient and ineffective, interrupting discovery progress and capturing only a superficial overview of researchers' ideas. Some informatics applications now feature tools such as annotations — ink, text, and sticky notes that team members use to attach questions and ideas to other data as they work. This provides a current, searchable view of team members' collective analyses and thoughts.

> Accelerating drug discovery hinges on bridging the divides that isolate people and knowledge.

Composability is another emerging application feature designed to illuminate people's thinking and knowledge. In drug discovery informatics, individuals use composable workspaces to flexibly combine and analyze different pieces of visual and/or textual information. For example, a scientist may place a subset of assay results and a few promising compound structures that (s)he wants to synthesize (with accompanying notes) into a single workspace. This composition helps clarify thoughts, and also offers a meaningful view of the scientist's progress, so other team members can see what she is thinking and where she is headed.

The result is a more tightly coordinated organization in which those involved with the pipeline maintain insight into what others are doing and thinking. Shared situational awareness makes it possible to reallocate time to those that more directly push discovery forward, and to take action based on better information. Discussions shift from factual "who," "where," and "what" details to more discovery-oriented "how's" and "why's." And, such collaborative analyses are captured for future reference.

Ad hoc hallway conversations will continue to be an important source for drug discovery collaboration. However, informatics applications people use to pool disparate knowledge encourage similar collaboration to happen much more frequently throughout the organization. Ultimately, this opens up many new opportunities to accelerate drug discovery. •

Daniel Weaver is associate director of scientific computing at Array BioPharma; daniel.weaver@arraybiopharma.com. Michael Capsambelis is director of product management at Viz, a business of General Dynamics: michael.capsambelis@gdviz.com.

[GUEST COMMENTARY]

Harnessing Multi-Core Personal Computers

Scientists and IT groups create return on investments.

BY JIM TUNG

Several trends over the past 25 years have revolutionized how scientists and engineers work: PCs replacing minicomputers for data acquisition and analysis, high-level environments replacing Fortran for technical computing, and modeling and simulation playing a key role in embedded system development. Those trends reflected scientists' thirst for more speed and the need to handle ever-larger data sets. It was never simply a question of CPU quantity or clock speed or addressable memory, but rather the user's ability to harness that compute power to do useful work. Nowadays, as I talk with biomedical researchers about their challenges and computing needs, I have a sense of déjà vu, as I see some familiar patterns mixed with some new twists.

Back in the centralized-computing days of VAX, Convex, Alliant, and Cray, the computer's OS and system administrators bore the administrative burdens while end-users struggled at their terminals to set up their work. As standalone personal computers (PCs) took over, the end users flourished: they controlled what their computer did, high-level environments enabled them to develop algorithms and analyze data without the need for low-level programming, and ever-faster processors gave them power boosts without additional effort. However, system administrators struggled to manage and support those dispersed computing activities.

At that time, it was groundbreaking to provide a technical computing environment that let the end user work interactively while insulating them from computer variations, running on a PC, workstation, or supercomputer without change. There was no need to recompile or figure out the byte order of the CPU with which you were working.

Today's landscape has some familiar aspects and other attributes that are new. End users still use their PCs, except they now have dual- and quad-core processors. Today's multiprocessor clusters feel a lot like the old mainframe/supercomputer paradigm, except cheaper and based on "standard" processors and operating systems. And end users still desire more speed, better data handling, and improved productivity.

> Somewhere along the way, scientists and IT groups seem to have lost track of each other.

One twist is the mixed composition of users in biopharma: biologists and chemists who rely on data and want turnkey, easy-to-use software; statisticians and mathematicians who need to create, refine, and deploy new algorithmic approaches; and computer scientists and programmers who want tools for rapidly generating production-computing applications that work on huge volumes of data.

Lost Track

Somewhere along the way from mainframes and minis to PCs to today, scientists and IT groups seem to have lost track of each other. End users crave speed and power, but don't talk to their IT groups (except to badger for a faster PC with more memory). Meanwhile, IT groups are buying more multi-core PCs (there really isn't any other option nowadays) that users can't fully take advantage of. And they set up server farms and have to search in-house for projects and end users interested in using them. It is ironic!

Hidden in this bizarre situation is a very interesting opportunity. The multi-core PC enables end-users to do parallel and distributed computing without impacting IT groups. The improved OS, scheduling, and administrative tools of compute servers (along with the fact they're based on the same processors and operating systems as the PCs) mean that more compute power is available — and more affordable — than ever before.

Recent enhancements in technical computing environments provide greater consistency in how users can capitalize on today's variety of computing systems. High-level technical computing tools that can distribute work on a server farm, without the need for low-level message passing programming, allow users to harness multiprocessor clusters for applications such as Monte Carlo simulations and sequence analysis. Scientists can also work with larger data sets, as the number of processors and memory space can scale up. And these same environments can also take advantage of multi-core PCs, enabling end-users to make full use of the computing systems at their disposal, with minimal changes to what they do on their own PC.

While speeding up existing applications is unquestionably good news, the most exciting opportunity is still to come: providing tools so that algorithm and application developers can more easily create techniques that make explicit and optimal use of parallel-computing systems, regardless of whether it's a dual-core PC or a server farm with hundreds of processors. The programming and system administration tools are reaching the point where end users can really focus on the problems and applications, taking advantage of available hardware without the need to deal with it explicitly.

But so what? Are you doing better drug discovery or advancing your understanding of systems biology better than the next guy? Now you and your users have the opportunity to create that return on your investments in your computing resources. •

Jim Tung is a MathWorks Fellow. He can be reached at jim.tung@mathworks.com.

Clinical Research

Standards Questions

CDISC's Becky Kush looks forward to new FDA rules.

BY ANN NEUER

Celebrating its 10th anniversary this year, CDISC — the Clinical Data Interchange Standards Consortium — has become a major voice in the pharmaceutical industry for its work in developing platform-independent data standards for the electronic acquisition and exchange of clinical trial data and metadata. The group was recently voted into the International Organization for Standardization (ISO) with Liaison A status — a level enjoyed by only a few organizations, and maintains a charter relationship with HL7 (Health Level Seven), the health care standards body.

Spearheading CDISC is president and CEO Rebecca (Becky) Kush who, with her staff of ten, provides vision to dozens of dedicated volunteers from its roughly 200 member organizations. Those organizations have implemented many clinical research standards and are working on additional initiatives with partners such as FDA and the National Cancer Institute (NCI).

"We are at an exciting juncture," Kush told *Bio•IT World*. "FDA is writing a proposed rule to require electronic submissions of clinical data in New Drug Applications (NDAs), Biological License Applications (BLAs), and Abbreviated New Drug Applications (ANDAs) in a format that FDA can process, review, and archive." The FDA has issued a guidance on eSubmissions using the electronic common technical document (eCTD), which specifies use of the Study Data Tabulation Model (SDTM). Kush says, "There is every reason to believe that the rule would require its use."

> "We are at an exciting juncture. FDA is writing a proposed rule to require electronic submissions of clinical data ... in a format that FDA can process, review, and archive."
>
> **Rebecca Kush**, CDISC

Training the Corps

FDA has asked CDISC to train its reviewers on SDTM. Kush says her team has trained almost 200 people so far. There is also growing demand for education courses, both in public forums and privately for constituents. CDISC is recruiting additional instructors to keep up with demand. Most of the courses have been face-to-face training, but at the CDISC International Interchange in October, CDISC launched a new virtual learning tool on SDTM.

An exciting initiative is CDASH (Clinical Data Acquisition Standards Harmonization). Says Kush: "This opportunity grew out of the FDA Critical Path initiative to develop standards to support data collection in the case report form (CRF) at the level of the investigative site." In 2006, FDA and the Association of Clinical Research Organizations (ACRO) requested CDISC take the lead. "CDASH focuses on defining standardized data collection fields in the CRF to collect safety data and medication information across applications and sponsors," Kush says. "The data fields are mapped to CDISC's SDTM, and there will be an XML version based on the Operational Data Model (ODM) transport standard used to collect data electronically."

Another initiative called BRIDG — the Biomedical Research Integrated Domain Group model — is an effort to link CDISC clinical standards to healthcare standards developed by HL7. "This effort is intended to reach consensus around the meanings of the elements used in clinical research," says Kush. "For example, in terms of semantics, how is 'enrollment' defined? How is 'adverse event' defined?" Last June, CDISC and partners released BRIDG model version 1.1 — the fruits of a three-year collaboration with FDA, NCI, and HL7. The National Cancer Institute's caBIG (cancer Biomedical Informatics Grid) initiative is also part of the BRIDG effort, seeking to achieve interoperability among clinical trials research in cancer.

Catching On

There is growing acceptance of CDISC standards internationally. "We have incredibly active groups in Japan and Europe, where we hold interchanges every year," Kush says. These European and Asian coordinating committees are "a very integral part" of CDISC, she says. Several user networks have sprung up in Europe in various languages, "because people wanted to get together and discuss implementation issues."

Ten years post launch, Kush says her main focus is to "hold everything together, ride the political waves. Seriously, my role is to continue to paint the CDISC vision to make sure we see that shining star and see where we are headed. The technical people need to look at the details but unless they can also see the grand picture, they may get lost as to why they are doing what they are doing." •

New Mind-Set About eClinical Technologies

EXCLUSIVE: Tufts CSDD survey shows positive attitude shift toward technologies.

BY DEBORAH BORFITZ

Over the past few years, there has been a seismic shift in the way biopharmaceutical companies and clinical research organizations think about and use electronic clinical research technology. The changes have been almost entirely positive, according to Ken Getz, senior research fellow at the Tufts Center for the Study of Drug Development (Tufts CSDD) in Boston.

That's the picture painted by a just-released study conducted by Tufts CSDD assessing adoption and attitudes about eClinical technologies and standards. The results are being presented for the first time by *Bio•IT World* and *eCliniqua*.

Adoption concerns and barriers today spring more from an "implementation mindset" than worries about management buy-in and lack of a market leader characteristic of a nascent, emerging market, says Getz. "Management cares most about return on investments already made, integration, and internal coordination."

Fears about choosing the wrong vendor have also receded, owing largely to a maturing market with fewer, stronger players accepted as drivers of future innovation. Growing acceptance of data interchange standards promulgated by the Clinical Data Interchange Standards Consortium (CDISC) has raised the "comfort level" in selecting a particular solution. "There's less of a sense of risk now in terms of compatibility," says Getz.

Increased Adoption

For both solutions and standards, familiarity breeds adoption, says Getz. Electronic case report forms, among the earliest clinical research technologies launched, are also the most utilized (by 89% of respondents), closely followed by statistical analysis (82%), solutions for site management (56%), document exchange portals (54%), and electronic patient reported outcomes (47%).

Adoption of CDISC standards likewise mirrors the sequence with which they were introduced. The data submission model, with the earliest produced standards, is used by 69% of surveyed companies, 90% including pilot programs, Getz adds. Standards for the exchange of non-clinical data are all still in the pilot stage. Standards for operational data and lab data are currently being piloted by companies using them, 32% and 26% respectively.

Support for the CDISC mission, overall, is high. More than 70% of respondents were members of the organization, giving input to collaborative efforts to build industry-wide data interchange models.

> "There's less of a sense of risk now in terms of compatibility."
>
> **Ken Getz**, Tufts CSDD

Companies now recognize many perceived benefits from adopting CDISC standards, due in part to the growing diversity of data interchange models, Getz says. The benefit of improving partner data exchange, named by just one-third of companies three years ago, was cited by 59% in the new survey. Improving data quality early was also cited much more frequently (36% from 11%).

The survey indicates users' growing enthusiasm; experience with the oldest solutions for clinical data informatics and statistical analysis are "very positive," says Getz. Experience with newer solutions for site oriented project planning was rated as positive by 30% of respondents and negative by 20%. "This is also the spottiest area in terms of overall usage. It's done more project to project, so it makes sense that the experience level is more variable as well," says Getz.

Four out of ten companies report having achieved a return on their investment in eClinical technologies, says Getz. The "realized impact" is being felt in more than a dozen different areas, most often in terms of quicker database lock, real-time access to data, and in monitoring program status. Again, the biggest punch is coming from the more "established solutions."

The difficulty in sharing data between various technology solutions remains the chief frustration of companies today, affecting more than two-thirds of respondents. Yet even that is a "positive signal," Getz says, because it implies that fewer companies are still piloting solutions. "They're using them throughout their organization with higher levels of satisfaction and demonstrated benefit." •

By the Numbers

Clinical Research Technologies Used by Survey Respondents:

- 89% use electronic case report forms
- 82% use statistical analysis
- 56% use eClinical solutions for site management
- 54% use document exchange portals
- 47% use electronic patient reported outcomes

CDISC Membership

- 70% were members of CDISC

Perceived Benefits of Adopting CDISC

- 59% cited improving partner data exchange
- 35% listed improving data quality

New Products

Excel Add-in

Gentelligent has introduced BasePro, an Excel add-in that adds over 40 custom functions to Excel. The functions address common informatics needs when working with DNA, RNA, and protein data. Functional coverage includes complements, melting temperature, molecular weight, sequence repeats, coding sequence, and general sequence cleaning and formatting. BasePro includes complete online help and free email support, and works with Excel 2002 (XP), 2003, and 2007.

Product: BasePro
Company: Gentelligent
Available: Now
For More Information: www.gentelligent.com

MicroRNA Assays

Luminex has launched FlexmiR Select, a new microRNA (miRNA) assay designed to allow researchers to further advance understanding and enhance the analysis of miRNAs. The latest addition to Luminex's FlexmiR line of products, FlexmiR Select allows researchers to create customized miRNA panels for more efficient and focused miRNA analysis, especially of small subsets of microRNAs, which is essential to understanding more about the role microRNAs play in disease.

Product: FlexmiR Select
Company: Luminex
Available: Now
For More Information: www.luminexcorp.com/microrna

Chemistry Software

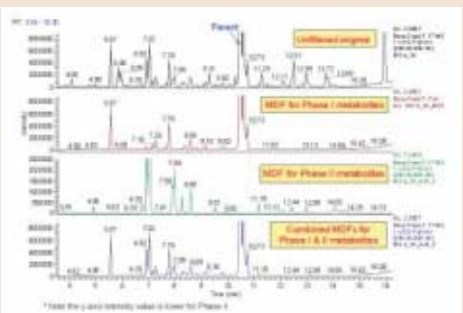
Advanced Chemistry Development, (ACD/Labs) has announced the release of its latest version of ACD/Labs PhysChem software. Version 11 provides updated molecular property predictions for a broader variety of chemical classes through enhanced models, and offers a new customizable interface for instant interactive review of results. Updates to version 11 include new logP prediction model based on experimental data from more than 25,000 compounds, substantial increases of the internal database, expanding and diversifying chemical space coverage, particularly for compounds of pharmaceutical interest, and the addition of over 2,500 new compounds and experimental data into the training database, resulting in increased chemical space coverage and greater prediction accuracy.

Product: PhysChem version 11
Company: ACD/Labs
Available: Now
For More Information: www.acdlabs.com/physchemlab

Screening Services

Caliper Life Sciences has launched a new screening service to help researchers determine the efficacy of combination drug therapies and identify potentially harmful combinations before therapies enter clinical trials through their Caliper Discovery Alliances & Services (CDAS). Cellular assays are used to screen for synergistic compound combinations and flag potentially harmful toxicity levels. Assays include Caliper's anti-cancer cell proliferation panel, cellular cytokine/chemokine release assays, or using *in vivo* efficacy models.

Product: Drug Combination Services through Caliper Discovery Alliances and Services
Company: Caliper Life Sciences
Available: Now
For More Information: www.caliperls.com



Metabolite Upgrade

Thermo Fisher Scientific has released MetWorks 1.1.0, an updated version of their metabolite identification software. The software facilitates automated acquisition, processing, and reporting of liquid chromatography/mass spectrometry data in support of biotransformation studies. Key improvements in v. 1.1.0 include flexibility to apply up to six multiple mass defect filters, component filtering to exclude duplicates, and automatic generation of Data-Dependent Parent Mass Tables.

Product: MetWorks 1.1.0
Company: Thermo Fisher Scientific
Available: Now
For More Information: www.themo.com/metabolism

Bio•IT World Best Practices 2008

JOHN RUSSELL

Defining and sharing best practices in drug discovery is challenging. The mix of technologies being put to work is ever-changing and the speed with which those technologies advance is dizzying. Stir in competitive zeal which inclines companies to conceal, not share, their winning ways, and you get a sense of the challenge. This picture is rather different from pure academic research, which is every bit as competitive, but whose laurels are earned through the peer-review publishing cycle.

Yet difficult doesn't mean impossible and even if today's best practices are unlikely to remain so for long, they nevertheless represent excellence to be celebrated and the kind of vanguard thinking from which the next generation of best practices will spring.

In 2008, *Bio•IT World* is again sponsoring a Best Practices Awards program to recognize innovation and excellence in the use of technology to advance drug discovery and development and biomedical research. This is our fourth Best Practices Awards effort, following a two-year hiatus during which, among other things, *Bio•IT World* was acquired by Cambridge Healthtech Institute, and became the core of the Cambridge Healthtech Media Group.

I urge you to review *Bio•IT World* Best Practices program (www.bio-itworld.com/bestpractices), consider whether your organization has an appropriate example of technology well-used, and submit an entry to Best Practices 2008. It's good for you, good for your organization, and good for the industry.

Moreover, *Bio•IT World*'s special report to be published next summer and its awards ceremony dinner at the Bio-IT World Conference and Expo (April 29 in Boston) will shine a spotlight on innovative activities at a time when the drug industry receives so much criticism.

This year's entrants will be in good company. Past Best Practices winners include: Baylor College of Medicine; Children's Memorial Hospital, Chicago; Endo Pharmaceuticals; GlaxoSmithKline; Harvard Partners Center for Genetics & Genomics; Iconix Pharmaceuticals; Locus Pharmaceuticals; Millennium Pharmaceuticals; National Cancer Institute; Perlegen Sciences; Pfizer PGRD; Roche Diagnostics; St. Jude Children's Research Hospital; Surromed; TGen; Vertex Pharmaceuticals; and Wyeth Research.

As a general rule, we're look for examples of work that was done in the last two years and is substantially complete (though there is some wiggle room.) Nominations are open to R&D and scientific facilities and labs in pharmaceutical companies, biotech companies, academia, government, medical or related institutions and organizations, as well as public and private research organizations. The deadline for entry submission is **January 31, 2008.**

There are seven categories: basic research and R&D; drug discovery and development; clinical research and trials; translational and personalized medicine; IT infrastructure and knowledge management; business and economic development; chemistry and compounds.

For examples of the kinds of projects we're looking for, as well as downloadable entry forms and guidelines, please visit www.bio-itworld.com/bestpractices.

The judging process is rigorous. We assemble a panel of experts drawn from industry and academia, across a number of disciplines, and convene them to evaluate and rate all entries and pick winners. The winners will then be announced on April 29 at a gala dinner held in conjunction with Bio-IT World & Expo at the World Trade Center, Boston.

Take advantage of this opportunity to showcase your organization's efforts and to contribute to the industry's broader knowledge base. Feel free to contact me with questions at john_russell@bio-itworld.com.

WE CAN. Yoh has been delivering high-impact talent to America's leading companies for the past seven decades – longer than any of our competitors. These customers count on our experience, time-honored processes, responsiveness and market intelligence. And our delivery of talent that is 24.7% better than the competition*.

Focus on your job. And your future.

YohClinical
Know greater talent.®
yohclinical.com