Big Biological Impacts from Big Data

In the life sciences, data can come in many forms, including information about genomic sequences, molecular pathways, and different populations of people. Those data create a potential bonanza, if scientists can overcome one stumbling block: how to handle the complexity of information. Tools and techniques for analyzing big data promise to mold massive mounds of information into a better understanding of the basic biological mechanisms and how the results can be applied in, for example, health care. By Mike May

"Big data" is one of today's hottest concepts, but it can be misleading. The name itself suggests mountains of data, but that's just the start. Overall, big data consists of three *v*'s: volume of data, velocity of processing the data, and variability of data sources. These are the key features of information that require big-data tools.

Although biologists have spent decades struggling to collect enough data, says Keith Crandall, director of the Computational Biology Institute at **George Washington University** in Ashburn, Virginia, "the new bottlenecks in biology are big-data issues." As an example, he points out that the first human genome sequence, announced in April 2002, utilized the expertise, infrastructure, and people from 20 institutions and took 13 years of work and about $3 billion to determine the order of approximately three billion nucleotides. Now, says Crandall, "We can sequence a human genome for $1,000, and we can generate more than 320 genomes per week!"

As life scientists explore more ways to deal with big data's volume, velocity, and variability, they are starting to develop new approaches to analyzing information.

Ever-Expanding Volume

When it comes to collecting large volumes of information about human biology, the pharmaceutical industry started battling large data sets decades ago. As Jason Johnson, associate vice president for scientific informatics at **Merck Research Labs** in Boston, Massachusetts says, "Merck has for many years had clinical trials with thousands of patients, and the ability to query millions of de-identified patient records, and now we have next generation genomic sequencing that can create a terabyte of data per sample."

To deal with so much data, even large pharmaceutical companies need help. For example, Bryn Roberts, global head of R&D operations at **Roche** in Basel, Switzerland, says, "A century's worth of Roche R&D data were more than doubled in 2011–2012 in a single large-scale experiment to sequence hundreds of cancer cell lines." Roberts and his colleagues want to derive more value from these data sets and others collected years ago. So they are collaborating with PointCross in Foster City, California, to create a data platform that allows flexible searching of data from the past 25 years of Roche studies, including those outsourced to contract research organizations. Those data, along with information about thousands of compounds, will be mined to use the existing knowledge to develop new drugs.

To generate large volumes of data, though, a biologist does not need the infrastructure of a large pharmaceutical company. For example, consider the specifications of the Ion Personal Genome Machine (PGM) System from **Life Technologies** in Carlsbad, California (now a part of **Thermo Fisher Scientific**). This next generation device can sequence up to two gigabases in less than eight hours, and this is marketed as a "personal genome machine" that can go on a scientist's benchtop. Life Technologies' larger Ion Proton machine pumps out up to 10 gigabases in four hours or less.

In general, for academic and industrial life scientists, next generation sequencing supplies a bonanza and a bottleneck. As Crandall explains, "We cannot effectively study this volume of genomes until our computational software scales up to these big data needs." So his team is working with W. Evan Johnson, an assistant professor of medicine at **Boston University School of Medicine,** to develop software, PathoScope, that can handle the data from today's next generation sequencing (NGS) platforms, which turn information on gigabases of DNA into gigabytes of computer data—the exact ratio tends to be about linear, depending on the NGS platform being used. This software compares DNA samples to reference genomes in an effort to identify a pathogen. Crandall says, "Our data sets can

Upcoming Features

Digital Lab Management—July 25 ■ *Metabolomics—September 19* ■ *Genomics—October 3*

PHOTO: MAKSIM KABAKOU/SHUTTERSTOCK.COM

be 20 gigabytes of data per sample for hundreds of samples with downstream analyses generating 100's of gigabytes of data per sample."

Such large volumes of data can be especially useful in health care, where pharmaceutical scientists must take into account the variability among people when designing their experiments. "You can't draw any reasonable conclusion by studying only 10 people, but you can come up with important lessons by studying say 500,000," says Chas Bountra, professor of translational medicine at the **University of Oxford,** United Kingdom. Large-scale studies might reveal genetic contributions to disease, whether a drug could help a subset of patients, or which individuals are likely to develop a particular disorder.

Other experts also expect to see a growing impact of genetic data on health care. "Genetics gives us a huge, powerful foothold into understanding how people get sick and what you can do about it," says Gil McVean, professor of statistical genetics at the **Wellcome Trust Centre for Human Genetics** in Oxford, United Kingdom. For example, genetic information might reveal biomarkers, or indicators of a specific kind of disease, like a molecule involved in a particular form of cancer. As McVean explains, "Genetics can tell you if a biomarker associated with a disease is worth going after as a [therapeutic] target." For instance, a molecule that drives a particular form of cancer could make a good target for treating the disease.

To apply this thinking, McVean and his colleagues are creating the Li Ka Shing Centre for Health Information and Discovery at Oxford University, which was launched on a huge contribution—about $33 million—from Chinese billionaire Li Ka-shing. This center will include a big data institute, which is currently under development. Overall, says McVean, the center "will bring together analytical data processing and genetics in one institute, so we can tackle some of the thorny but fascinating questions about collecting and analyzing big data sets."

Seeking High Velocity

The second *v*, velocity, depicts the speed of processing and analyzing the data. Scientists need high-speed processes to analyze the growing volumes of data.

In the past, analyzing gene-related data created a bottleneck. "Traditionally, these analysis platforms have posed productivity limitations on researchers," says Alan Taffel, president of **BioDatomics** in Bethesda, Maryland. "They've been difficult to use, often requiring bioinformatician support, and they've also been very slow in executing workflows." In fact, he says, it could take days or weeks to turn around a large DNA analysis. So BioDatomics developed its BioDT software, which provides more than 400 tools for analyzing genomic data. It integrates these tools into one software package to make them easier to use, and it can outrun any desktop computer.

BioDT runs on a computer cluster, which consists of machines, called nodes, inter-connected and working as one. "You need at least four nodes," says Maxim Mikheev, chief technology officer at BioDatomics. Yet BioDT can run on many nodes to process data even faster. "Scalability is theoretically unlimited," says Mikheev. "There are clusters that are using over 40,000 nodes." For users not inclined toward building a computer cluster, BioDT can also be accessed through the cloud.

Overall, says Taffel, BioDT "can execute workflows up to 100 times faster than traditional systems. Days or weeks become just minutes or hours."

"Our data sets can be 20 gigabytes of data per sample for hundreds of samples with downstream analyses generating 100's of gigabytes of data per sample."

Other experts also see the need for new tools driven by sequencing. According to Jaroslaw Zola, associate research professor in the Department of Electrical and Computer Engineering at **Rutgers University** in Piscataway, New Jersey, "The almost ubiquitous adoption of the next generation sequencing technologies requires new computational strategies to handle the data all the way from how it is stored, to how it is transferred, to how it is analyzed." This means that biologists must learn to use cutting-edge computational technologies, but, as Zola says, that "puts a pressure on information technology experts to deliver efficient solutions that are easy to adopt by the domain experts, yet hide the complexity of the underlying algorithm, software, and hardware architecture without sacrificing the efficiency." That, says Zola, requires novel algorithms, which he's working on now.

Versions of Variability

The third *v*, variability, creates a big challenge for biologists. As Bountra says, "We are now bringing together people from lots of different areas with lots of different data sets."

For one thing, a biology lab includes a variety of instruments, and they often collect data in specific file formats. So **ACD/Labs,** headquartered in Toronto, Canada, developed computing systems that integrate a wide range of data types when working with big data. As Ryan Sasaki, director of global strategy at ACD/Labs, explains, "We support more than 150 file formats from various instruments." He adds, "This lets us bring data into one environment, our Spectrus database, which can be made available through desktop-client software or accessed through the web as well as via other laboratory informatics systems."

Biology's big data also include new kinds of variability. Scientists at **Definiens** in Munich, Germany, for example, analyze what the company calls tissue phenomics, or information about a tissue sample's make-up, such as cell size, shape, the stains that they absorb, and which cells contact each other. This technique can be applied to a variety of studies, such as those designed to continued>

PHOTO: QUKA/SHUTTERSTOCK.COM

track the characteristic changes cells undergo during development, measure the impact that environmental factors have on an organism, or that quantify the cellular effects a medication may have on specific tissues.

Structured data, such as tables of numbers, do not reveal everything that is known about a medication or biological process. Much of what we know about living organisms exists in unstructured formats, like journal article text. As Johnson of Merck says, "There are thousands of ways to describe biological processes," and it is difficult to extract data from the literature.

At **IBM**'s Almaden Research Center in San Jose, California, analytics expert and research staff member Ying Chen and her colleagues have worked for years on creating technologies for mining text, which they now use for their "accelerated drug discovery solution." Their platform aggregates patents, scientific literature, basic chemistry and biology knowledge (such as how chemicals and molecules interact), more than 16 million unique chemical structures, and information about nearly 7,000 diseases. Using this system, researchers can search for compounds that might be useful for treating specific diseases.

Other companies also hope to mine existing resources to learn more about the biology of diseases and how to treat them. **NuMedii,** a big-data company in Silicon Valley, and Thomson Reuters, a provider of intelligent scientific information in New York, have teamed up to find new uses for existing drugs, known as drug repurposing. "Using genomic databases, integrated knowledge sources, and bioinformatic approaches, we can quickly discover novel uses for drugs," says Craig Webb, NuMedii's chief scientific officer. "We then leverage the safety data for the drugs in their original use to get to clinical trials faster and cheaper." NuMedii is contributing databases and analytics to the project, while Thomson Reuters is supplying in-depth knowledge on diseases and drugs.

One such project, Webb says, has researchers compiling gene expression data from more than 2,500 ovarian tumor samples and using several computer algorithms to predict whether any existing drugs could potentially treat ovarian cancer broadly or treat specific molecular subtypes. "Big data allows us to cast a wide net initially to identify leads, while 'big knowledge' allows us to quickly select viable compounds to test," Webb says.

Featured Participants

ACD/Labs
www.acdlabs.com

BioDatomics
www.biodatomics.com

Boston University School of Medicine
www.bumc.bu.edu/busm

Definiens
www.definiens.com

George Washington University
www.gwu.edu

GNS Healthcare
www.gnshealthcare.com

IBM
www.ibm.com

Life Technologies
www.lifetechnologies.com

Merck
www.merck.com

Novartis Institutes for BioMedical Research
www.nibr.com

NuMedii
www.numedii.com

Roche
www.roche.com/index.htm

Rutgers, The State University of New Jersey
www.rutgers.edu

Thermo Fisher Scientific
www.thermofisher.com

University of Oxford
www.ox.ac.uk

Wellcome Trust Centre for Human Genetics
www.well.ox.ac.uk

Complexity from Combinations

To the three *v*'s of big data, Stephen Cleaver, executive director of informatics systems at the **Novartis Institutes for BioMedical Research** (NIBR) in Cambridge, Massachusetts adds complexity. He says that scientists in the pharmaceutical industry analyze the data by "patients individually and then as a group, and then we integrate everything we have." That gets complex.

In health care, the complexity of big data analysis also arises from combining different types of information, such as data from genomics, proteomics, cellular signaling, clinical research, and even environmental studies. The results could reveal entirely new approaches to treating diseases. But Iya Khalil, cofounder of **GNS Healthcare** in Cambridge, Massachusetts, asks: "How do you make sense of those data and get insights from those data that will advance our understanding of the disease mechanism?" For Khalil and her teammates, the answer comes from machine learning, mathematics, computational algorithms, and supercomputers—all combined to explore the underlying pathways of disease and to follow a patient's likely response to a particular treatment.

At GNS Healthcare, such big-data analysis depends on a computational platform called REFS, which stands for reverse engineering and forward simulation. In short, the software analyzes data to construct possible molecular networks underlying a specific disease—that's the reverse part—and then it uses that information to simulate the impact a particular compound would have upon the pathway—the forward aspect of the process.

In addition to health care, REFS can be applied to basic biology. For example, Khalil and her colleagues have used this technology to make a molecular model of part of the cell replication cycle.

For Khalil and other scientists, the key is using big data in ways that move science forward. At NIBR, for instance, Cleaver and his colleagues want to make sure that the data is informative, first and foremost. "It's great to run advanced data-mining methods, but it must suggest the next scientific hypothesis," he says. That way, today's big data will change tomorrow's biology and medicine.

Mike May is a publishing consultant for science and technology.

DOI: 10.1126/science.opms.p1400086