PICKING WINNERS
DATA MINING FOR DRUG DISCOVERY

In June of 2013, GlaxoSmithKline ($GSK) and software giant SAS announced a new project designed to make pharmaceutical research data more transparent—and shareable—in ways that would have been unheard of just a few years ago. SAS, which is based in Cary, NC, is working with GSK to build a secure website where pharma developers can go to obtain data from other companies' clinical trials and then use a host of tools to crunch that data in myriad ways. "We fundamentally believed the ability to interrogate vast stores of information would be useful and valuable for society," says Perry Nisen, senior vice president of science and innovation at London-based GSK. "This is an exercise we've been working on for quite some time, and we wanted to make it available for others to explore."

The GSK/SAS alliance is one of several new initiatives being undertaken in the biopharmaceutical industry that are aimed at making better use of the vast terabytes of information flowing out of research labs around the world—everything from whole-genome sequences to newly discovered disease biomarkers to data from clinical trials. The scientists and techies who are working together to ease access to this data have a shared goal of using informatics to make R&D more efficient. So they're improving not just the access but also the data-analysis technology that can help the industry assess new molecular pathways, predict side effects of experimental drugs, and even analyze potential future markets.

"Companies realize the value of this information is greatly enhanced if a single researcher can access multiple companies' information for a single purpose," says Matt Gross, director of the health and life sciences global practice at SAS. "The challenge is for them to agree on the processes and the standards by which they make information available, so they can aggregate the information across companies."

BY ARLENE WEINTRAUB

The platform that SAS and GSK are building is designed to house all the information that's generally not included in published clinical trials—the specific patient-by-patient numbers related to drug response, side effects, and so forth, anonymized to protect patient privacy. SAS-based analytics tools are built in, Gross says, as is a system for accessing alternative analysis tools and sharing documents from commonly used programs like Microsoft Word and Excel.

For GSK, the push to make data more useful and universal started many years ago as an internal project, Nisen says. The company worked with SAS to build a platform in which clinical trial data could be made available to researchers across the company. GSK has used the resulting platform for a range of projects, from scrutinizing placebo responses in multiple trials to exploring data from one study for signals that might show whether a drug could be useful in another indication, Nisen says. "We could potentially phenotype patients based on the constellation of measurements, both laboratory and clinical," which may help stratify patients more precisely for clinical trials, he says. In describing all the potential uses for the technology, Nisen adds, "I could go on for hours."

GSK made its data-sharing platform available to the public in May and is now working with SAS and others to devise a plan for incorporating data from other drug companies.

One of the open questions about the GSK/SAS platform is how requests for access to the data will be vetted. The companies have agreed that there should be some sort of independent review board to ensure that researchers requesting access are doing so for legitimate scientific purposes, but exactly what such an entity should look like is a matter of some debate. "What is the right process of looking at and approving a request in a way that lowers the barrier to getting information? And is there a way that a common review panel could analyze a request on behalf of multiple companies?" Gross asks. The answers have yet to be determined.

Nevertheless, virtually everyone in the life sciences industry agrees that big data—and new tools to analyze it—will enhance the entire R&D process, from discovery all the way through to marketing. "We believe high-throughput capabilities, coupled with analytical and computational methods, offer the guiding light for at least the next decade of new discoveries," says Alexander "Sasha" Kamb, senior vice president and chief of discovery research at Amgen ($AMGN) in Thousand Oaks, CA.

"We fundamentally believed the ability to interrogate vast stores of information would be useful and valuable for society."

PERRY NISEN, SENIOR VICE PRESIDENT OF SCIENCE AND INNOVATION, GSK

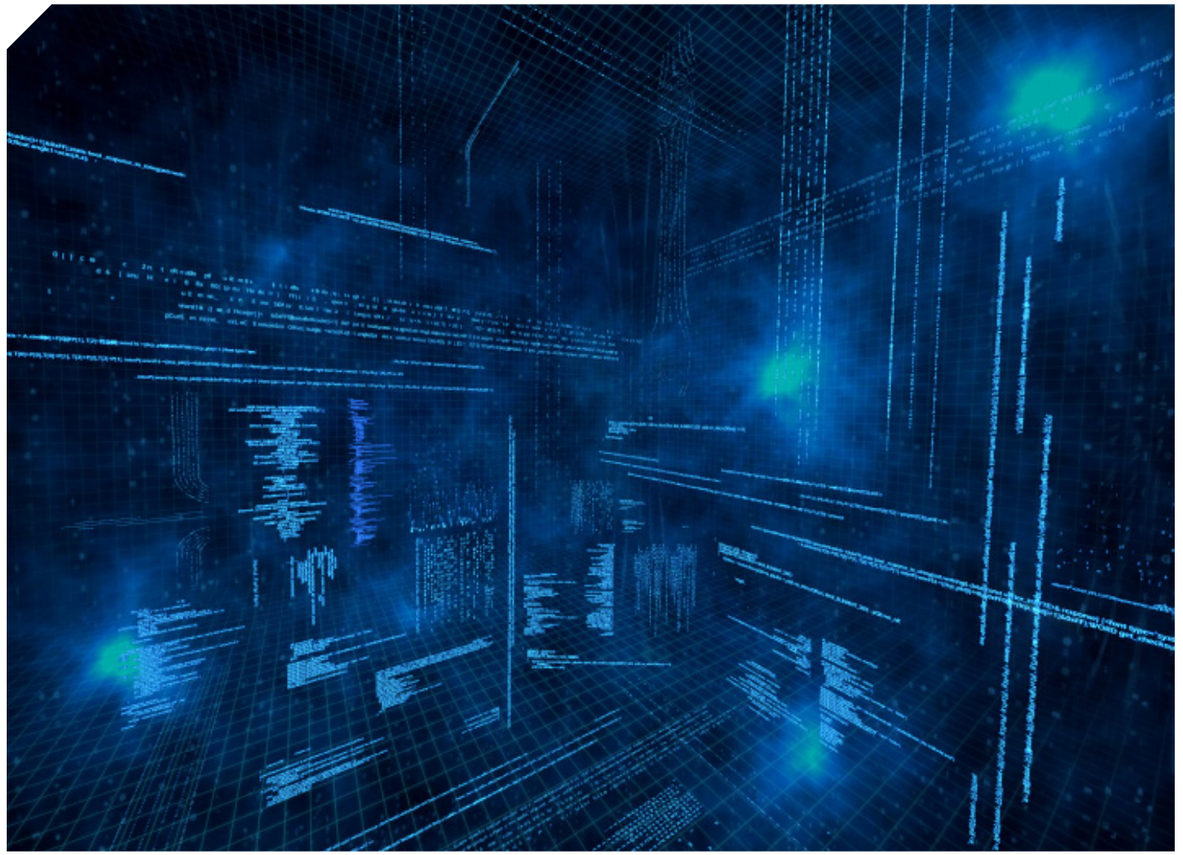
Amgen signaled its commitment to deploying big data toward drug discovery in December 2012, when it acquired Iceland's deCODE Genetics for $415 million. DeCODE had sequenced human genes from a half-million people and used technology it developed to read the 3 billion "letters" in each genome. That capability allowed deCODE to identify specific variants that played important roles in cancer and other diseases. Kamb says Amgen wanted to bring that technology into its discovery organization to improve the process of selecting the best targets for drug treatment.

"We reached a conclusion that what was holding the industry back was insufficient differentiation in [research] pipelines. People were swarming over the same set of targets that were usually published in the literature and often had been there for a while," Kamb says. "There was also a lack of predictability, so there were too many failures."

Amgen decided the only way out of that unproductive cycle would be to take advantage of the increasing understanding of human genetics, Kamb says. "There is sufficient, accessible human genetic variation in the population that does inform us about mechanisms that play in human disease," Kamb says.

"We believe high-throughput capabilities, coupled with analytical and computational methods, offer the guiding light for at least the next decade of new discoveries."

ALEXANDER KAMB, SENIOR VICE PRESIDENT AND CHIEF OF DISCOVERY RESEARCH, AMGEN



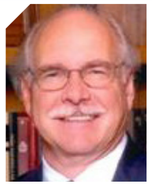
The challenge, Kamb admits, will be transforming that growing body of genomic information from a "parts list" to a concrete list of promising drug candidates—a process that's evolving at Amgen right now, he says. "We are aided by ultrahigh-throughput technologies and the application of human genome sequencing and genetics on a very wide scale. Now we have to take those targets, figure out which ones are feasible, and prosecute them with the tools of the trade that have gotten more and more sophisticated."

DATA ANALYSIS FOR THE MASSES

A growing selection of web-based tools for combing through big data is available to researchers who don't have the resources of big companies like Amgen and GSK. That selection is both a blessing and a curse. Michael Reich, director of cancer informatics at the Broad Institute in Cambridge, MA, estimates that at least 10,000 such programs are on the web, many of which cannot interface with each other. "When you have large genome-characterization studies that are creating many different types of data, you need many different tools to analyze it," Reich says. "You also need some kind of environment that will put the tools together and make that easily accessible to people who may not have the computational sophistication to

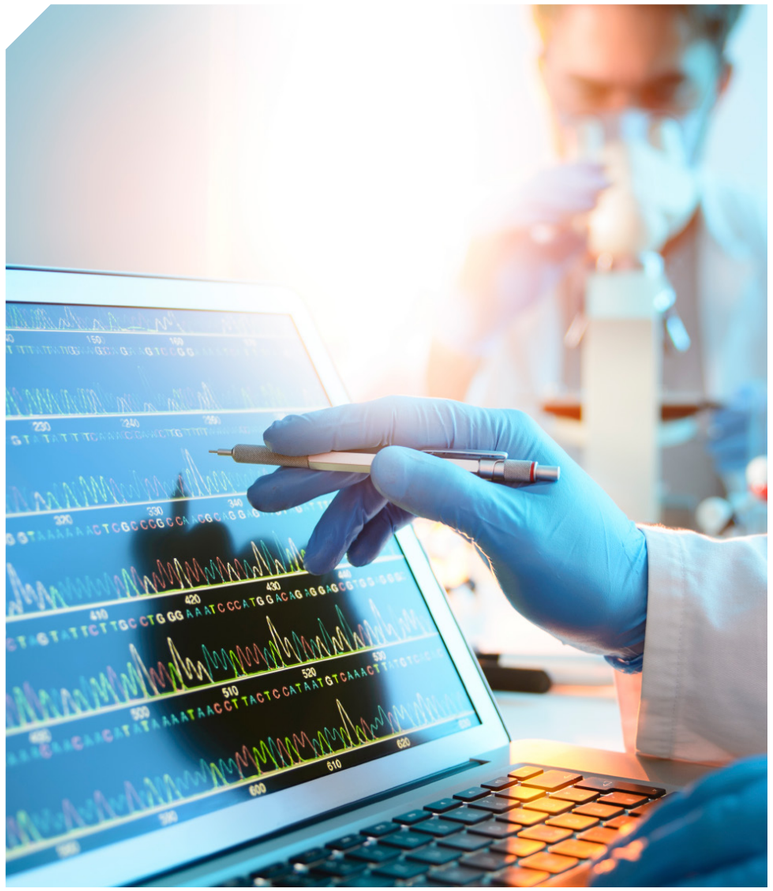
glue them together themselves."

The Broad recently updated a program it designed to do just that. GenePattern allows scientists to access hundreds of tools for gene-expression analysis, proteomics, RNA analysis and flow cytometry, all on one platform. Users can also capture and record all the steps they're taking in their analyses, so others can easily reproduce their experiments. The update allows the program to be run in the cloud, so scientists can store all the data they generate along with other material related to their research.



"We think that the focus in the future is to take this exponentially increasing volume of data, on the drug side and on the genomic side, and put it together."

DR. LLOYD EVERSON, CEO, MOLECULAR HEALTH



Now GenePattern users can also share their analysis methods, or "modules" as the Broad calls them, with other scientists. "It's an easy way to disseminate something you've done yourself to the worldwide genomics community," Reich says.

In April, the Broad released a new program called GenomeSpace, which eases the process of sharing data that's generated on GenePattern. For example, any user with a Dropbox cloud-storage account can send data from it to a GenePattern server. "We take advantage of a lot of the new cloud-based technologies to allow people to collaborate on big-data projects," Reich says.

Another company working to make genomic data more accessible via the cloud is NextBio, a Santa Clara, CA-based company founded in 2004. NextBio is essentially a curated collection of publicly available research data from clinical trials, molecular profiles of patients, reference genomes, and other sources. It's designed to make it easy for scientists to perform tasks such as validating new biomarkers or finding new uses for drugs that may have failed in past clinical trials.

Researchers use NextBio to interrogate publicly available data in ways that might not be possible on widely used sites like PubMed, says the company's co-founder and CEO, Saeid Akhtari. "Let's say a paper is published in PubMed that says 'We've compared breast cancer patients with healthy individuals and we identified these 10 genes that we believe are somehow

implicated in the development of breast cancer. To do that, they produce billions of data points," Akhtari says. "We don't just rely on the paper. We go back and we process the raw data and make that available. The publication may only include 10 genes, but they studied thousands of genes, and there can be orders of magnitude more information buried in that data."

Akhtari declines to elaborate on any specific projects that have employed NextBio's products, which are provided as software-as-service platforms, but the company has amassed an impressive list of more than 50 clients, ranging from academic institutions like Brown University to Big Pharma players such as Pfizer ($PFE) and Johnson & Johnson ($JNJ).

DEPLOYING DATA TOWARD DEVELOPMENT

The ability to analyze big data may be a boon not only to drug discovery but also to development. Several new analytical tools are designed to comb through events such as regulatory approvals or rejections, toxicity reports, and reimbursement decisions made by public and private insurers. The goal is to help drug developers predict—perhaps years in advance—the probability of a new product's success in the market, based on public information about similar products that have already completed the development cycle. Such insight may allow companies to make "go" or "no go" decisions much earlier in development, before they spend millions to develop a drug that may have limited potential to succeed.

One company developing such analysis technology is Molecular Health, a startup based in Basel, Switzerland, and backed by Dietmar Hopp, co-founder and former CEO of software giant SAP. In 2014, Molecular Health will launch a service for oncologists designed to assist them in choosing treatments by combining genomic information from individual patients with data released from genomics studies around the world. The company is also actively pursuing opportunities to deploy its big-data analysis know-how toward aiding biopharmaceutical development, says Dr. Lloyd Everson, CEO of the company's U.S. division, based in The Woodlands, TX.

The company is also working on a product called Molecular Analysis of Side Effects (MASE), which scans biomedical literature for data on the targets that particular drugs hit in the body, which genes are involved in metabolizing those drugs, and which adverse events such as dangerous drug interactions are occurring. The idea is to make predicting drug safety more precise. "We think that the focus in the future is to take this exponentially increasing volume of data, on the drug side and on the genomic side, and put it together," Everson says.

In 2012, Molecular Health embarked on a 5-year collaboration with the U.S. Food and Drug Administration, which has allowed the company to aggregate data from the agency's Adverse Event Reporting System and build it into MASE. By combining the

"If you're going to design a trial to meet FDA criteria, why not at the same time ensure those are the same criteria you would need to break into [worldwide] markets?"

S. YIN HO, CEO, CONTEXT MATTERS

toxicity reports with Molecular Health's genomic models of drug metabolism, the platform helps scientists spot patterns, which could in turn allow them to predict adverse events for drugs in development. Everson says the company should have a plan for introducing MASE to the biopharmaceutical market by the end of this year.

New York-based Context Matters is focusing on how data can help companies gauge a product's chances of succeeding in the market. The company was founded in 2010 with the mission of helping drug development executives predict whether insurers would pay for their drugs. The company's Web-based platform, called Reimbursement Risk Tracker, collects data from payers around the world, curates it, and then presents it in a way that makes it easy for development professionals to look for reimbursement patterns.

Context Matters' CEO, S. Yin Ho—a former physician who once worked in Pfizer's e-health division—believes biotech and pharma companies are starting to catch on to the importance of understanding reimbursement patterns early in the development process. "They're utilizing the data to reduce reimbursement uncertainty," Ho says. "But some companies have expressed interest in bringing this into clinical-trial design, which conceptually makes sense. If you're going to design a trial to meet FDA criteria, why not at the same time ensure those are the same criteria you would need to break into [worldwide] markets?"

For example, one of Context Matters' clients is aggregating reimbursement data and then analyzing it in multiple ways to see if patterns are starting to appear in various countries, Ho says. "It's the elements to create a predictive model," which could help the company decide where to apply for approval first. "They're playing out multiple scenarios at one time and trying to understand what the result would be."

WORKING OUT THE KINKS

The move toward making data more transparent and analyzable is laudable, Ho believes, but only if the life sciences industry has a clear endgame—meaning companies define the problems they're trying to solve before jumping onto the big-data bandwagon. Many researchers say analyzing patterns in clinical trial data will help them avoid wasting resources on molecules that are likely to fail as drugs, but Ho isn't convinced. "Research involves going down multiple blind alleys," she says. "There is a hope that if you can document [failures], you have the ability to prevent someone else from going down a rabbit hole. But I don't think you'll necessarily eliminate that many rabbit holes."

Others worry that without widespread sharing among many companies in the industry, the utility of the data might be limited. "Not everybody wants a shared environment," says SAS's Gross. "But imagine a single researcher working on a diabetes drug [who] wants to aggregate information from two or three companies working on a similar compound, to see if there are certain types of people who are responding well. Having only one company's view isn't going to give you that rich picture."

And then there's the challenge of protecting patient privacy—a major administrative hurdle, says Eric Perakslis, the FDA's former chief information officer, who is now executive director of the Center for Biomedical Informatics at Harvard University. "Who is ensuring identifiable information [about patients] isn't getting into the wrong places? Who's keeping track of that?" Perakslis says. "There's a governance and compliance function that these folks have to think about. The last thing I want to see is somebody getting a big fine from the government."

Once all the kinks are worked out, though, Perakslis predicts life sciences companies will find multiple uses for big data that will ultimately improve R&D. For example, he says, the industry has started to look at sharing placebo data from clinical trials, so companies will no longer have to duplicate control arms that other companies have already done. "In a consortia style they're sharing [control data]," he says. "If you've got a bunch of companies working in XYZ disease and none of them are making headway, you're seeing them come together" so they can move the most promising therapeutic candidates ahead faster, he says.

"There's a governance and compliance function that these folks have to think about. The last thing I want to see is somebody getting a big fine from the government."

ERIC PERAKSLIS, EXECUTIVE DIRECTOR, CENTER FOR BIOMEDICAL INFORMATICS, HARVARD UNIVERSITY

One example is the CEO Roundtable on Cancer, a group of more than 30 life sciences executives who are working together to accelerate oncology drug development. In late 2012, the group launched Project Data Sphere, an effort to create a technology platform for sharing oncology clinical trials data.

As for Glaxo's efforts to make its own data more transparent, Nisen says the company's new website has seen a "reasonable uptick" in hits since it was launched in the spring. The site currently lists more than 200 clinical studies that Glaxo has started since January 2007, but the company intends to expand the site by adding all the studies it has conducted since 2000, including those that were never published. "What we put in place was the first step," Nisen says. "We wanted to make it available and make it as easy as possible to assemble the data and analyze it. It's an evolution that ought to be based on experience. There may be pieces we need to make more amenable for users. We'll learn as we go." ●