Building a framework

High-throughput analytics tackle the bottleneck in data handling.

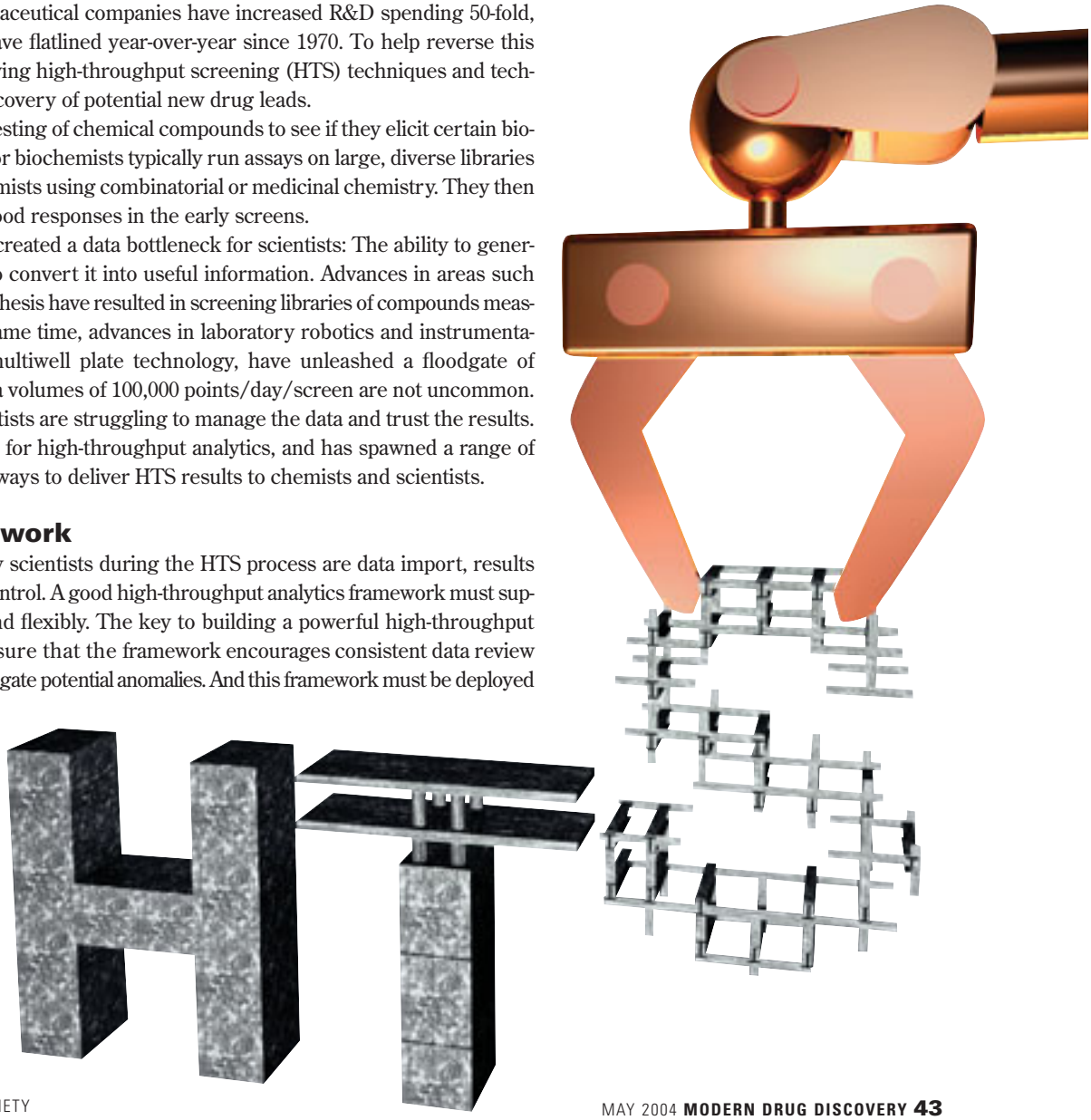
BY BILL LADD

In the past 30 years, pharmaceutical companies have increased R&D spending 50-fold, yet new drug approvals have flatlined year-over-year since 1970. To help reverse this trend, companies are applying high-throughput screening (HTS) techniques and technologies to accelerate the discovery of potential new drug leads.

HTS semiautomates the testing of chemical compounds to see if they elicit certain biological responses. Biologists or biochemists typically run assays on large, diverse libraries of compounds created by chemists using combinatorial or medicinal chemistry. They then test those compounds with good responses in the early screens.

But HTS automation has created a data bottleneck for scientists: The ability to generate data exceeds the ability to convert it into useful information. Advances in areas such as combinatorial chemical synthesis have resulted in screening libraries of compounds measured in the millions. At the same time, advances in laboratory robotics and instrumentation, such as high-density multiwell plate technology, have unleashed a floodgate of production data streams. Data volumes of 100,000 points/day/screen are not uncommon. As these volumes swell, scientists are struggling to manage the data and trust the results.

All of this drives the need for high-throughput analytics, and has spawned a range of new analytical tools and new ways to deliver HTS results to chemists and scientists.



An analytics framework

The basic tasks performed by scientists during the HTS process are data import, results calculation, and data quality control. A good high-throughput analytics framework must support these tasks efficiently and flexibly. The key to building a powerful high-throughput analytics framework is to ensure that the framework encourages consistent data review and the ability to quickly investigate potential anomalies. And this framework must be deployed across broad, dispersed teams that drive collaborative decision-making.

Crucial decision points, such as those that drive research projects, arise when scientists compare data across experiments, laboratories, and fields. But acquiring data from various sources is difficult. In the past, scientists had to master the intricacies of search

ILLUSTRATION: SCOTT NEITZKE

engines, databases, and IT infrastructure to query the data, which slowed or obstructed the scientific exploration necessary for good research. But new high-throughput analytical frameworks, many of them Web-services- and browser-based, now support truly integrated data access to and from any source, anywhere, at any time, and without end-user programming.

For example, researchers at one pharmaceutical company continually monitor assay screening results and update process databases with the previous day's results. They also regularly generate analyses of cumulative results for monthly project reviews. To make effective use of their database investments for HTS, the company needed a direct and secure conduit to these databases. The application would need to work directly with their existing Oracle database repository and be accessible using standard Web technology. Additionally, the researchers wanted to decrease the number of applications needed to support the daily workflow of the laboratory's HTS analysis process.

The company considered several existing applications to analyze and process its data and determined that the answer lay in building a custom application in-house, using proven resources already being applied at the company. In two weeks, the company's IT group built a Web application that allowed researchers to retrieve screening results without having to deal with the complex HTS database schema or the transformations required for effective analysis. With the system, researchers could quickly identify time-related trends and patterns or positional biases on their HTS screening plates. Through the application interface, they can immediately update problematic sample wells and mark them as "rejected" in the central database.

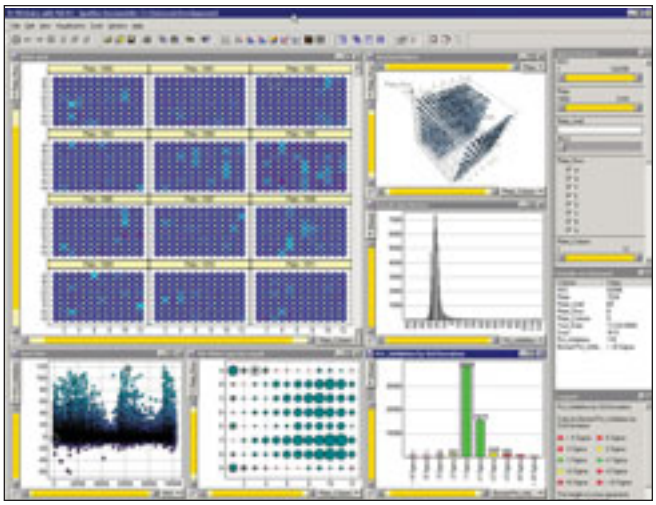


Figure 1. See here. By using data visualization software, scientists can identify patterns within complex data sets, greatly facilitating decision-making. (Image courtesy of Spotfire, Inc.)

Data visualization

Visualization—that is, a way to display the results of a computer function—has often been an afterthought to analysis. Yet modern computing tools can do more than simply display results; they can provide interactive, fast, and flexible data visualizations that help and even enhance human thought processes (Figure 1). The human brain, for instance, is adept at perceiving patterns, shapes, and colors. Modern data visualization tools can present complex data sets in unique ways that engage these innate perceptive abilities. In addition, interactive visualizations can be made, encouraging researchers to note, investigate, and explore unexpected behavior in their data. Interactive visualizations let machines focus on what they do best—data processing—while humans focus on their talent—making decisions.

Researchers at one prominent U.S. medical research university, for instance, have been able to interactively explore data through a visual interface that helps them recognize intricate patterns they otherwise might not have discovered. Scientists and geneticists use visualization software to display clusters of color-coded data that are cross-referenced with drug patent data and nearly every known toxin and therapeutic compound.

The software integrates these various data sources and turns the information into interactive, multidimensional images that immediately reveal patterns and trends that numerical readouts obscure. Researchers also use visualization to analyze gene expression data to determine how cellular structures, often in a disease state, carry out specific genetic instructions. As a result, the researchers can visually probe the data with hypothetical "what if" questions as never before.

Usability

A high-throughput analytics framework may connect to all the right data sources and incorporate impressive visualization and analysis options, but if it is not easy to use, it will not be used. IT specialists may enjoy the process of matching the right data to the right algorithm and the right visualization, but laboratory scientists would rather be at the bench than at a computer. The HTS framework should cater first to the needs of the average scientist, offering more sophisticated tool sets as an option for those scientists who want greater control.

Guides, for instance, have emerged as a popular way of simplifying complex tasks. In common desktop tools, such as Excel, Adobe Photoshop, and Web applications, guides step users through common tasks. Rather than searching through menus to find the option needed to complete a task (or navigating through a series of options spread out over many areas of an application), users can consult a guide, which combines the tools needed in one interface. Guides are useful because they can serve both novices and advanced users and, when applied to discovery research, can also greatly accelerate many aspects of data analysis, from transacting with or entering experimental information to configuring visualizations to setting the parameters of a complex clustering method or data reduction algorithm.

Extensibility

Because the discovery environment changes rapidly, software must adapt just as quickly. Implemented software solutions supporting analytics must be readily configurable to support access to new, emerging data sources, new analytical tools, and new analysis processes. And because one never can predict what form a new analytical approach will take, the platform architecture

should be able to incorporate methods developed in many different architectures and run-time environments.

Many analytical applications claim extensibility but place such strong restrictions on the approaches used to add functionality that, in the end, they are quite limited in scope. When evaluating a solution's ability to be integrated with other analytical tools, it is important to note whether extensibility is a core development principle or a marketing-induced afterthought. When a high-throughput analytic framework is well planned, the application infrastructure itself becomes an important part of the application's functionality.

For example, an HTS center at a leading European pharmaceutical company processes requests from the company's research centers throughout its home country and beyond to test compounds for potential applicability as drug candidates. As part of the screening process, the center's technicians perform assay development and screen optimization procedures and then administer the tests. The results are analyzed by a team of about 20 quality control (QC) researchers who weed out "bad data" before reports are sent on to the company's global data warehouse.

To drive the QC process, the company designed its HTS framework around a visual analytical application with open application programming interfaces. Thus, from within the single application interface, users can access a variety of in-house statistical calculations and algorithms, as well as analyses from other external software programs.

Deployability

The competitive pressures of pharmaceutical research have resulted in a widely distributed user base. Global project teams encompassing researchers from a variety of disciplines must be able to share data and reach consensus on research directions. Web infrastructures are ideal for these purposes and are the logical choice to underpin an analytics framework. Users gain access to the most up-to-date methods that apply to their research. And organizations can manage their analytics tools centrally. With one change to one part of the system, IT staff can deliver new features and functionality to researchers anywhere in the world.

Developing new drugs to fight infectious disease requires researchers at one company to sift through mountains of data, which are stored in an Oracle database. To run experimental queries and analyze the results, scientists were forced to rely on the company's informatics IT team and an arsenal of proprietary data tools or else manually query the databases themselves using structured query language. In either case, they would have to display and share the results in numerical form in Microsoft Excel spreadsheets, poring over the data for clues, trends, patterns, and insight. Although this is common practice in pharmaceutical companies, this company wanted to foster a more immediate, collaborative analysis environment that would speed up discovery and development.

To address this problem, the company brought in an analytical application that gave its scientists an interactive visual environment

for analysis in which they query databases with "what if" questions on-the-fly through a graphical user interface. They can then publish and communicate the analysis and decision data directly through the company's e-mail system using a "poster" system, thereby lowering the learning curve and ensuring rapid communication of analysis decisions. Company scientists no longer have to manually query a database every time they change a parameter or even one small aspect of a hypothesis, nor do they have to go through the painstaking steps of tracking down colleagues or building PowerPoint presentations to share their experimental results.

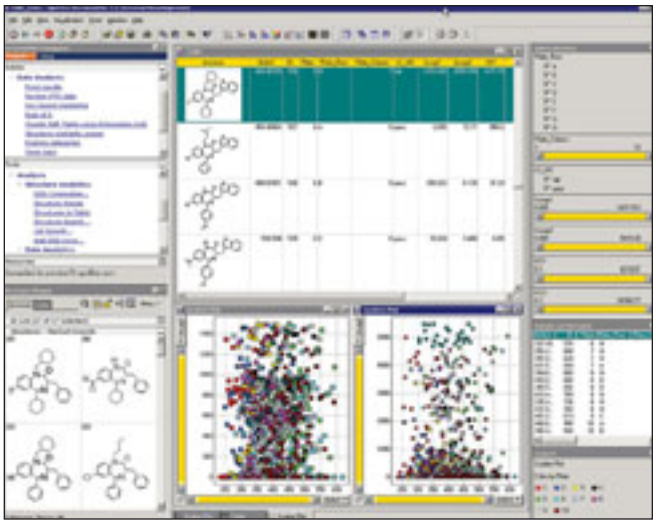


Figure 2. I came, I SAR, I conquered. An informatics framework allows researchers to pull data from various sources. (Image courtesy of Spotfire, Inc.)

Consistency

Well-planned high-throughput analytics frameworks not only meet the access, analysis, and usability demands of the individual researcher, they also provide consistency across the broader organization. It does not make sense for screening researchers to invent their own ways of reviewing and validating their data. On the other hand, investigators need to be free to follow up on the particular anomalies they see in their own data. An analytic framework needs to support both of these needs, providing a consistent best-practice-based analysis approach while also permitting additional analytic capabilities for the unanticipated findings. As different user communities configure the environment for access to data relevant to their specific work, new classes of scientific investigation will bring these previously disparate classes of data together (Figure 2). This is seen today in high-throughput ADME, where traditional development work is facing the new challenges of using high-throughput techniques.

Research is not slowing down, and the rate at which experimental data are being generated stands only to increase. In evolutionary terms, the fittest research organizations in the future are likely to be those that can efficiently analyze data to come to decisions about what to do next. Analytics, therefore, must itself evolve. Industrial-quality research requires industrial-quality software that lets analytical researchers focus on their specialty rather than on data management, reporting, or storage.

Bill Ladd is senior director of analytical applications at Spotfire, Inc. ■