Creating an Integrated Portal for Biological and Chemical Information

Jo Whelan

PHOTOS FROM TRIPOS

Pharmaceutical companies are generating more biological and chemical data than ever before but have yet to capitalize on the data's full potential. The ability to use data to make better, faster decisions has become a key factor in a company's competitive position. Informatics solutions that can integrate data from many sources are essential if the data are to be transformed into useful knowledge.

Thanks to a raft of technological advances, the capacity of the pharmaceutical industry to generate data is growing at an unprecedented rate. Computer giant IBM reportedly estimates that the volume of life science data is doubling every month, driven by the explosion of information coming from genomics research, high-throughput screening, and combinatorial chemistry. Also adding to the avalanche of data are the emerging areas of proteomics and pharmacogenomics and the creation of virtual compound libraries, which can contain trillions of chemicals. In the late 1990s, the then Glaxo Wellcome anticipated that internally generated data alone would soon cause a 100-fold increase in network traffic on its computer system.

However, drug companies are painfully aware that the data-gathering revolution has not, as yet, transformed the productivity of the drug discovery process. Research by Andersen Consulting (now Accenture) in 1997 revealed high expectations among senior pharmaceutical and biotechnology executives (1). They believed that by 2000, drug discovery times would be

Jo Whelan is a freelance writer based in London, jo.whelan@virgin.net. This article was written in conjunction with Tripos, Inc. Please contact Heather Hunter at Tripos for more information, tel. 314.647.1099, heather@tripos.com.

halved, and there would be a three-fold increase in the number of new chemical entities (NCEs) being delivered to development. These improvements did not materialize; a follow-up report in 2000 found the situation barely changed (2).

It still can take 10–15 years and $500 million to bring a new drug to market. Each new molecule delivered to development costs about $70 million (2). The attrition rate between lead discovery and market launch still is unacceptably high, and too many compounds still fail relatively late in the development process at enormous cost. At the same time, sales from established blockbuster products are eroding steadily as more and more come to the end of their patent lives. To maintain the high returns its shareholders have become used to, the industry is under intense pressure to bring more NCEs to market while keeping a check on escalating research and development (R&D) costs. Furthermore, a high proportion of these drugs must have the sales potential to offset high development costs and replace the aging blockbusters.

In the current business environment, described by consultants PricewaterhouseCoopers as more hostile than anything the pharmaceutical industry previously has encountered (3), the ability to manage the data mountain and turn it into useful knowledge is crucial. Information technology (IT), in the shape of informatics systems, has moved from being a background operational tool to one on center stage. A company's informatics infrastructure is now a key issue when evaluating competitiveness, and IT development is an important element of strategy. Even the largest companies now are turning to specialist providers of informatics solutions to meet their data-handling needs.

Data sources

Data generated (and brought in) by pharmaceutical companies derive from a variety of activities within the drug discovery process, which can be broadly grouped under "biology" and "chemistry." *Bioinformatics* and *chemoinformatics* refer to the handling and manipulation of such data, respectively. In biology, genomics programs generate data about gene sequences and gene expression and now are complemented by proteomics. This work will generate unprecedented numbers of new biological targets. Another emerging discipline is pharmacogenomics, which heralds a move away from the one-drug-fits-all model toward a future in which various drugs are targeted at various subpopulations of patients, escalating the amount of data involved still further.

Chemical data — those centered around compounds rather than targets — derive from combinatorial chemistry and the creation of compound libraries; high-throughput screening; lead identification and optimization; toxicology testing; absorption, distribution, metabolism, and excretion (ADME) studies; and preclinical experiments. Data comprise both chemical structures and numerical and descriptive information about the properties of compounds. The amount of chemical data available has exploded with the advent of in silico research, where trillions of theoretical compounds are created by computer to form virtual compound libraries.

Changes in data-handling requirements also are being driven by business developments. The wave of mergers and acquisitions that has swept the pharmaceutical and life sciences sector has left many companies with a legacy of multiple, independently created, and incompatible data stores at locations around the world. The anticipated R&D synergies will not materialize unless data from the formerly independent partners can be shared effectively. Indeed, evidence exists that consolidation, in many cases, has been tremendously disruptive to R&D programs. In addition, the growing number of strategic alliances means that companies of all sizes must manage data generated by their partners as well. An increasing awareness also exists that market-related data must be brought into the equation at an earlier stage of the drug discovery process.

The right tools

Despite the enormous volumes of data, the problem is not too much data but rather a lack of tools with which to turn these data into knowledge. Within companies, data remain localized in their departments of origin, often occupying diverse storage formats and operating systems. Failure to share data means effort is duplicated and valuable knowledge is not passed on. When shared, data often lack context, thereby making valid comparisons difficult. Weak trends that might provide vital signposts are obscured by "noise." In particular, the long-standing gulf between biology and chemistry in terms of data handling demonstrates the divide between them in the traditional, linear model of drug discovery. Both bioinformatics and chemoinformatics are rapidly growing disciplines, but

until recently little attempt had been made to integrate them. As a result, many companies are left with expensive systems that cannot communicate with each other.

There is an urgent need for IT solutions that can unlock the business value in the wealth of data held by pharmaceutical and biotechnology companies. One player in this arena is Tripos Inc. (St Louis, MO), which uses its platform of proprietary computational chemistry and data management technologies to provide both data management solutions and contract drug discovery services. For example, in 2000 Bristol-Myers Squibb (BMS) chose Tripos to design and implement a new integrated research informatics system in conjunction with the consulting group Accenture. BMS researchers and managers will use the desktop system as a primary drug discovery information portal designed to improve access to a wide range of scientific information and to provide access to calculation tools and the ability to analyze and compare calculated results with experiments.

Obtaining accurate answers to questions such as "What compound should I make next?" or "Do we continue with this series of leads?" can steer researchers away from directions that will ultimately prove fruitless, shaving vital months off drug development times. Maximizing the availability of data to support decision making is particularly crucial at the lead development and optimization stages, which are currently one of the main bottlenecks in the R&D pipeline. According to Dr. Paul Weber, senior vice-president of Tripos's Software Consulting Services, the company aims to enable pharmaceutical researchers to make decisions faster, more reliably, and from better information.

The aim of informatics solutions is to provide all personnel with seamless access to data from their desktops, regardless of the original storage format and location. Facilitating this goal are interfaces such as the Tripos MetaLayer Framework, a customizable integration tool based on a powerful combination of middleware and application development frameworks. The MetaLayer system provides a single desktop portal that allows any network client to access any distributed data or application. Biological and chemical information, including molecular structures, can be stored, retrieved, managed, displayed, and analyzed by means of one interface. More important, it enables legacy applications to be left in place. For example, ISIS, ActivityBase, and other Oracle-based systems can be integrated to enable seamless browsing and querying. Almost any data source can be added, including nondatabase sources. Data are read and reformatted and presented to the user to form a global decision-support system.

Viewing and analysis

Obtaining access to data is beneficial only if the data can be viewed and manipulated in a constructive and user-friendly manner. Using the MetaLayer system, the presentation format can be both customized for the client and personalized to reflect the needs and interests of various individuals within the company. The system can be used with a standard or customized viewing platform. Web-based viewing is available for all data, even those held in older systems that are not Web-enabled.

Graphical and statistical analysis capabilities allow users to search for and spot trends that may be weak and difficult to find. For example, a particular chemical subtype might have a tendency toward being active or being toxic; or a particular characteristic in a compound might be predictive of activity with a certain type of receptor. Discerning such a trend in an ocean of data can diminish time wasted pursuing unfruitful avenues or avert a costly late-stage failure by discovering problems early.

Data quality

Even the most sophisticated information management systems are only as good as the data they contain. "Without context, data can be useless," Weber warns. "It is vital to capture enough context to make the data meaningful. For example, one must know how an assay was run; an IC50 is not necessarily the same thing in every instance." To ensure that an organization is basing its decisions on meaningful data, scientists need systems that allow them to capture contextual information. This capacity is a feature of the ChemEnlighten comprehensive decision-support tool (Tripos) for the evaluation of chemical and biological data. The system is Intranet-based and specifically designed to handle the enormous volumes of data created by combinatorial chemistry and high-throughput screening. According to the company, it can assemble and store all the data for an entire research effort. The company's ChemCore data management technology system is designed to track commercially available reagents, reactants, products, libraries (both designed and synthesis), analytical data, and processes, thereby creating a knowledge database that can be used by researchers to support daily activities. Comments and analyses can be saved alongside struc-

tures and experimental results as they are generated in the lab. Both corporate and commercial databases can be searched from the ChemEnlighten interface, and data can be imported and exported in industry-standard formats, including the company's SYBYL molecular structure analysis platform and UNITY software environment. Data are viewed in spreadsheet format, allowing properties to be calculated and subsets selected and analyzed by means of graphical and statistical tools. By simplifying the visualization and analysis of hundreds of thousands of compounds, tools such as ChemEnlighten and ChemCore may help chemists with decisions such as what compounds to purchase, synthesize or screen, or carry forward to the next stage of research.

Enhancing standard systems

Pharmaceutical companies have made extensive use of standard data-handling systems such as Oracle relational databases. These databases often act as corporate repositories, which contain records of all the compounds an organization has made or purchased together with their analytical and bioassay results. Scalability allows companies to cope with the million or more new compounds that they make each year. The nature of relational databases also makes them a potential source of guidance on questions that require insight into the link between the properties of molecules and their structure. Researchers can be interrogated with queries such as "What is known about the toxicity of compounds containing a particular chemical group?" or "Do we have molecules in our database whose shape would allow them to bind to a target enzyme?" However, because relational databases cannot store chemical structures, structural data have been stored separately from data about properties in databases (e.g., the Unity software package [Tripos] and the ISIS/Host system [MDL, San Leandro, CA]). However, this requires one to run separate structure and property searches, with the results of one search being used to filter the other.

Oracle's recently introduced data cartridge technology has created the potential to make relational databases even more useful in specialist applications. Data cartridges are modular extensions that enable Oracle to handle new data types. On the basis of this technology, Tripos has made available its AUSPYX system, a chemistry data cartridge that is designed to make Oracle chemically literate. The AUSPYX system extends Oracle's syntax to provide access to the two-dimensional, three-dimensional, and similarity-searching capabilities of the Unity software package. Oracle 8i databases thus can be mined for both structural and property data in one query, enabling researchers to unlock relationships hidden in their data. The AUSPYX system is designed for companies that need a solution for storing and searching chemical and relational data or that need to convert a legacy system not equipped to handle both types of data, according to Weber. ISIS/Direct Molecules (MDL) is another such cartridge.

The information needs of pharmaceutical and biotechnology companies are diverse and constantly changing, making it impossible for off-the-shelf software solutions to fit every organization's workflow. AUSPYX can be adapted by the purchaser using standard development tools such as Java, Visual Basic, or C++. Alternatively, customized applications can be developed.

Changing the information culture

To meet the demand for faster and cheaper development of new compounds, the drug discovery process must move away from the traditional linear model in which compounds are passed in isolation down a conveyer belt of independent departments. It must be replaced by a web-like structure in which various disciplines interact throughout the process, with data from each stage available and understandable to all.

In the traditional model, biologists rarely talked to chemists, and lessons learned at the ADME stage never made it back to the people churning out the next batch of compounds. In the new climate, such waste and fragmentation of knowledge is not an option if companies want to remain competitive. Integrating data from the disparate sources found in today's globalized, consolidated Big Pharma companies, or from the increasingly dense network of alliances involving companies of all sizes, is a major challenge for IT and software providers. IT solutions also must allow context to be captured with raw results and must add value to data through powerful yet user-friendly viewing and manipulation tools. In 1998, PricewaterhouseCoopers predicted that by 2005, all personnel would have plug-and-play access to Web-based, global information portals containing all available data in indexed, integrated form (3). Increasingly, technology is available to harness the power of the data explosion. Companies and individuals must now embrace the cultural change required to take advantage of this potential and translate data into knowledge at every stage of the drug discovery and development process.

References

1. Andersen Consulting, *Reinventing Drug Discovery: the Quest for Innovation and Productivity,* company publication, 1997.
2. Accenture, *High Performance Drug Discovery: An Operating Model for a New Era,* company publication, 2001.
3. PricewaterhouseCoopers, *Pharma 2005: An Industrial Revolution in R&D,* company publication, 1998. **PT**