The information gold rush

Data mining for life sciences and drug discovery research

BY DAVID BRADLEY

The forty-niners had it easy, relatively speaking, panning for gold and blasting their way through the Rockies to search for fragments of a shining noble element. These days, gold comes in many shades and forms, not all of them as tangible as a bag of 24-carat dust. Information is the new vein to be tapped, and there is no richer mother lode than the information nuggets to be found in drug discovery.

Put simply, data mining is the process of selecting and exploring data to find new patterns. This is knowledge generation at its keenest, converting raw data into useful information. Keyword text searching is relatively straightforward and is at the heart of literature searching in every discipline—from financial to zoological text mining. Natural-language processing, on the other hand, is a very complicated problem in its own right. But for life sciences and drug discovery research, mining is complicated by the broad variety of topics and types of information to be mined, from small molecules to nucleic acid sequences and from cellular traffic to enzymatic active sites. Such disparate entities and their associated journal articles, annotations, and spectral and physical properties are found in diverse resources. Mining these resources requires prior knowledge of the different data formats, entity labels, and topic-specific vocabulary of the individual systems. Life would be simple if it were not so complex.

Indeed, biological information is, according to Lynette Hirschman, chief scientist in the Information Technology Center at not-for-profit Mitre Corp., "exploding," with an estimated 1 terabyte of new information being unearthed each week. New information hopefully brings new knowledge, and Hirschman has coined the term "biolinguistics" to refer to the exploitation of computational linguistics methods applied to problems in biology and, in turn, drug discovery. Where bioinformatics refers to the broad use of information science in the life sciences, biolinguistics provides a new framework within which researchers in academia and industry can use the automated and semiautomated tools of computing to help manage their information needs.

PHOTODISC

New modes of information management based on biolinguistics are now translating into new approaches to mining the data from countless microarray experiments, myriad model organisms, and the plethora of biological information pouring out of laboratories the world over. If gene databases like Genbank, FlyBase, Mouse, and Yeast, and protein databases such as SWISS-PROT and PIR, are the raw ore for the data miner, then biolinguistics is the tools and techniques for "refining" that ore. Biolinguistics provides the nomenclature and language that give researchers a set of standardized terms for a wide range of biological entities and topics, including gene names, protein nomenclature, disease symptoms, and biological functions.

Hirschman and her Mitre colleagues, Alexander Yeh and Alexander Morgan, are working on the development of interactive information extraction techniques for the online biomedical literature. Their work will help support the interactive, semiautomated selection of entries for inclusion in a biological database, or curation, using natural-language tools rather than arcane computer terminology and language. Hirschman explains that their focus is on maintaining the currency and consistency of existing databases.

The Mitre team implemented an open challenge program known as the Knowledge Discovery and Data Mining Challenge Cup, which took place in 2002 and provided a training and test ground for groups to display their data-mining prowess. Various teams showed how a body of 862 journal articles could be mined effectively and each paper associated with relevant data fields from FlyBase as part of the much larger task of curating such papers. The result of such developments at Mitre means that biologists are one step closer to being able to use a natural-language query to answer deep biological questions, such as "What diseases are caused by prions?" or more profoundly, "What peptide sequences are known to interact with prions?"

There are many other deep questions to be asked in biology, surrounding issues such as protein folding, structure, location, dynamics, sequence, interaction, and evolution. Answers to these questions may ultimately provide researchers with the means to simulate the living cell. Such a simulation is the long-term aim of the Blueprint Initiative, a public-good research program of the Samuel Lunenfeld Research Institute at

Mount Sinai Hospital affiliated with the University of Toronto.

Machine-learning methods and mining techniques could be the key. According to Blueprint's Ian Donaldson, writing in *BMC Bioinformatics* (**2003,** *4,* 11), machine-learning methods have proved themselves to be useful as tools in backfilling direct interaction and pathway databases, but it will be the coupling of this with human review and entry of data from journal archives, like the 14 million PubMed abstracts, into a factual database such as BIND (Biomolecular Interaction Network Database) that will convert simple data into information and usable knowledge. He and his colleagues have now developed the PreBIND and Textomy systems, which provide such coupling, allowing data mining of human, mouse, and yeast protein-interaction information with great accuracy. The curation and archiving of such data from the research literature can be done with a standard approach to data representation that will assist knowledge discovery as never before.

Daniel Weaver of Array Biopharma explained in the June 2004 issue of *Current Opinion in Chemical Biology* (**2004,** *8* (3), 264–270) how data-mining techniques can be exploited in library design, as well as in generating and optimizing new leads. Conceptually, he explains, drug discovery data are formatted for data-mining analysis into tables in which each row represents a compound and each column represents that compound's predicted or experimental properties. Data-mining techniques then yield a model of the chemical space that relates these various independent molecular descriptors with a single key attribute, such as efficacy, solubility, pK_a, and other properties.

This model can then be used to predict the same key property for new compounds (either real or hypothetical) and so show the data miner which of these might be useful new leads and which are likely dead ends. What it boils down to, Weaver says, is "given a set of compounds with measured activity or ADMET, find more compounds that should be synthesized and screened." Some models, however, can also provide a clearer understanding of quantitative structure–activity relationships (QSARs) and help in the design of new libraries.

Data-mining models, Weaver adds, can be simple, parametric equations derived from linear techniques or complex, nonlinear methods. Among the techniques currently finding favor are partial least squares, support vector machines, binary kernel discrimination, and recursive partitioning. Weaver says that most techniques can approach a classification accuracy of about 80%, and all reported techniques outperform a random baseline by a wide margin. Moreover, all the published techniques give fairly similar performance results within a few percentage points of each other, although early adopters of nonlinear techniques, such as support vector machines and binary kernel discrimination, are slightly ahead of the game, at least in the early stages of lead discovery.

"Nonlinear techniques may be appropriate for the early stages of lead generation and lead optimization, when overall model accuracy is paramount. Linear techniques may be more appropriate for later-stage lead optimization, when more easily interpreted models are desired," Weaver writes.

While drug discovery entered a phase of information overload long ago, the gold rush is happening now. "Most drug discovery is still being done with fairly old-school, informatics-uninformed techniques," Weaver tells *MDD*, "which is to say, most drug invention happens because you have a human looking at a large QSAR table, figuring it out." He suggests that truly informatics-driven drug discovery, with all the efficiencies that will bring, is still actively being sought. It is, he adds, "very slow to be adopted by the industry for lots of historical and cultural reasons." Regular use of truly predictive models of efficacy and ADMET are still to arrive. It is now that data- and text-mining techniques are maturing into systems that can be exploited to dig a far richer seam than any found by the forty-niners.

David Bradley is a freelance science writer at www.sciencebase.com. ■

Putting it into practice

British informatics company InforSense recently licensed its technology to consumer product company Unilever to develop a tailored text-mining service for business-trend analysis. Information scientists at Unilever R&D's Colworth Laboratory in England are set to pilot TextSense, InforSense's text-mining technology for free-text analysis. The system will help Unilever's R&D scientists and technologists build new knowledge from their research by analyzing patents, as well as published and internal documents.

According to Yike Guo, InforSense CEO and professor in computing science at Imperial College London, where he is director of the Imperial College Parallel Computing Centre, "Unilever's challenge to effectively use large-scale literature analysis is shared by organizations worldwide." He suggests that the integration of TextSense into the InforSense environment allows researchers to mine text alongside numeric data and so retrieve knowledge they can work on.

The InforSense system builds on research into Grid applications developed by Guo and his colleagues under the umbrella of DiscoveryNet (*www.discovery-on-the.net*), an e-Science project supported by the U.K. government through its Engineering & Physical Sciences Research Council (*www.EPSRC.ac.uk*).

Several users liken the Grid to a super-Internet, but it is much more than that, providing new ways for collaborators across the globe to share diverse information and instrumentation resources. DiscoveryNet allows data processing and knowledge discovery services to be published and applied as Grid services for real-time processing, interpretation, integration, visualization, and mining of massive amounts of time-critical data wrought from high-throughput devices. Such devices include biochips, high-throughput screening devices in biochemistry and combinatorial chemistry, environmental and energy sensors, and remote sensors.

InforSense points out that its system is uniquely horizontal in nature and so can help researchers exploit data in any format through the entire drug R&D chain, from target discovery to drug development.