'. . .the drug industry is currently wasting as much as US$2.5 million in retrieving existing data'

editorial

Fiona Brown

Saving big pharma from drowning in the data pool

Spending on drug research and development keeps going up, but the number of new medicines reaching the market each year does not. There are various reasons for this imbalance, but it is accepted that one of the biggest bottlenecks in drug discovery is the escalating amount of information that confronts the researcher, not least the amount of data arising from the human genome project.

The data-pool dilemma is twofold: keeping a handle on the information already buried in in-house databases and making use of it; and absorbing the constant flow of new information. 'There is no point in collecting data and not developing drugs', says Lewis Jardine from Oxford-based Intellidos.

This article looks at some of the ideas emerging from a second generation of youthful informatics companies, which just might save big pharma from drowning in the data pool. This article is based upon discussions and presentations given at the Discovery Informatics Forum (London, UK) in November 2005 [1].

Crash and burn

In the past 15 years some big names in the scientific IT field, such as Lion Bioscience, Insight, GeneLogic and Affymetrix, abandoned the search for the intractable solution to data management. Some would say, and do say, that straddling different databases and IT platforms to pull all your data into one hopper just can't be done.

After big pharma companies sampled the expensive offerings of the first generation of bioinformatics companies, they largely turned inwards to find their own individual solutions. The new entrepreneurial bioinformatics boutiques hope that their new ideas might tempt pharma to look afresh at what the industry can offer. Many of the people behind these new incarnations have been through the 'crash and burn' of the old industry and have learned a thing or two.

Data mining

When mining for minerals a lot of low-grade material is sifted – to find the valuable content. This is also the case with data. The computer has allowed us to collect masses of information, but information has to be analyzed to be of any use.

From data we need to find relationships, connections and similarities. So a biotech company, for instance, might want to select appropriate molecules for screening. Using the SciTegic Pipeline Pilot they can analyze their corporate databases and eliminate undesirable compounds. What previously took three days can be reduced to a fully automated, error-free run taking just three hours.

Dr Rob Brown from SciTegic says, 'The overall data-mining problem is that different formats, different applications and different hardware are often tackled by adding yet more complications,' In other words he says, 'disparate data in different formats are stored on a variety of hardware; this in turn has to be processed by different applications with incompatible formats and platforms.'

The Pipeline Pilot can integrate data repositories and applications across platforms such as SGI, Linux and Windows, and can produce a single, seamless workflow process. Such an approach can lead to the generation of single intelligible reports or web interfaces.

Colourful data

Because experimental data can also exist in the form of digital images, mechanisms are needed to extract meaningful informa-

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2006.10.002

tion from them, which can then be fed into data analysis tools. Definiens, based in Munich and the UK, has developed Cognition Network Technology for this purpose.

In a recent project for Novartis, ~15,000 images of small-intestine crypts had to be studied as part of an extra toxicology study for the FDA. Using Definiens' preclinical Cellenger package, researchers at Novartis were able to reduce the time normally taken for such a task from 24 to six weeks, and get to market ten weeks earlier than predicted.

Text mining

Last year, Medline (an industry databank) had 14 million biographic units of information, which was increasing at the rate of 40,000 units every month. Professor Jun-ichi Tsujii comments, 'This amount of knowledge must be able to contribute to future drug discovery'. Prof Tsujii is the Director of the National Centre for Text Mining (NaCTEM), founded in 2004 in recognition of how important the whole process of efficient text mining, designed to cope with the relentless influx of information, is to the drug industry.

Phil Hastings of Linguamatics estimates that the drug industry is currently wasting as much as US$2.5 million in retrieving existing data. His company has produced a new text-search and mining system that is capable of using multiple word search questions to extract information that we are not aware we have; as Donald Rumsfeld put it, 'what we don't know, we don't know.'

Data integration

The pharmaceutical industry has invested massively in its databases in pursuit of tomorrow's drugs. A moderately sized company could well have as many as 1000 databases. This could represent more than a million gigabytes of data. Companies are now realizing the need to federate these databases and create a single access portal.

Through that portal, data have to be found and retrieved. Contained within the portal is a complex mix of text and image, buried away in spreadsheets, graphs, charts and scientific papers. Simple search engines can match the name of a molecule but fail to understand context or interrelationships.

'Also', says Intellidos' Jardine, 'people with different skills who need to access these data don't share the same dataspeak. Structured query language or SQL allows you to ask a question in a way that gets an answer, even when data are stored in different formats. Otherwise it's a bit like having the internet without the search engines.

'It's OK, for instance, to phrase a request such as ''Give me all the druglike compounds with this substructure'', but difficult to add in a supplementary ''only give me the ones already tested positive in assay at serotonin receptors.'' That is simply because you are already straddling several databases, jumping between chemistry and biology and looking for a mix of spreadsheets, graphs or digital images.'

To address this issue, Intellidos has developed two interrelated products that allow users to ask complex, ad-hoc, cross-domain questions of their databases, and span many knowledge domains. Hyper Dossier mines the databases to find the search material and Query Constructor provides a question and answer route that avoids trawling through hundreds of pages of results.

Need for a generic system

Some in the informatics industry consider that there is need for a generic system capable of reading all the different presentations, algorithms and images that can be encountered within databases. To address this, InforSense has developed a workflow-based technology for integrating data in a way that is suited to the real-life practice of research, without demanding IT expertise. The Windber Research Institute found this ideal for merging patient data into one web-based resource for clinicians.

Amartus, by contrast, has adopted a different strategy, taking an 'open standards' approach to data search and capture. Employing their TargetWatch™ generic framework for single point of access, Amartus advises companies on 'best of breed' to use in their data integration, without producing a new interface.

GeneLogics' offering in this area is Proteus LIMS, a data-management platform that automates lab processes, and captures and integrates data from disparate bioinformatics tools to produce a 'data hub'. This particularly addresses the researcher's needs to warehouse easily retrievable experimental data.

Pharmacokinetics and pharmacodynamics

There are other new techniques capable of shortening the drug development process through computational approaches. In cancer, for instance, where new drugs are frequently used in combination with older, more well-established compounds, the variables that result can be more efficiently tested in the computer than with standard techniques.

Using published data from Vertex and AstraZeneca, Physiomics has developed a computer-generated model of cancer to predict the behaviour of diseased cells in response to two novel aurora kinase inhibitors. Such predictive tools can increase the efficiency of clinical trials and might even help to reduce the current 95% failure rate of novel anticancer drug candidates.

A database for pharmacogenomics

Finding out why drugs work in particular subsets of patients and not in others is the promise of pharmacogenomics. The Ingenuity Pathways Analysis (IPA) from Ingenuity Systems is a web-based software application designed to unravel disease pathways. After seven years in existence, the Ingenuity Knowledge Base houses 1.4 million expert extracted biological findings on human, mouse and rat genes – 275 journals are reviewed monthly and supplemented with relationships parsed from Medline abstracts. On average 50,000 findings on genes and proteins are extracted and stored every quarter.

The IPA can be used in several ways. For example, a microarray analysis can point to several hundred genes sufficiently perturbed to merit further investigation. The gene names can be entered into the IPA, producing a multilayered result. The top layer is a graphic of gene and protein interactions. By clicking points on this network a researcher can uncover information relating to functions, processes, diseases, canonical pathways and drugs.

At Inpharmatica, they concentrate on druggability. They use informatics to look at the types of targets against which the majority of compounds are successful, such as G-protein-coupled receptors (GPCRs), nuclear receptors or enzymes. Using a combination of technologies they can also pinpoint potential

adverse effects from drugs, something researchers don't usually go looking for.

Going forward

As an analyst who covered the first generation bioinformatics sector, Robin Campbell, from Jefferies, summarizes the sector, 'It was always fraught with difficulties: not least the problem of converting data into different formats to achieve homogeneity and ending up with a meaningless soup.

'The industry was always struggling with one major problem, namely that of getting big pharma to pay for whole new IT solutions when often they could stitch things together quite cheaply. This new generation of companies may be more targeted.'

'Pharma companies are at very different places in their current utilization of bioinformatics solutions for drug discovery,' suggests Dr Nigel Pitchford from 3i Investments. 'The customer base is overwhelmed with bioinformatics offerings that could provide incremental value-add, but still don't come close to a total solution.'

'At last November's Discovery Informatics Forum there was clearly an appetite for new solutions', says Jon Rees PhD, manager of the Forum. 'At these smaller focused meetings, the other advantage is that you get to meet the experts who have actually developed the systems.' The next mini expo on bioinformatics, BioSysBio [2], will be held in Manchester in January.

With so many valuable data already stored in the pharmaceutical industry's own databases, and more arriving by the minute, informatics is always going to be a vital tool in the process of finding successful drugs. Maybe this new revitalized industry is worth a closer look.

References

1 Discovery Informatics Forum, London, November 2005. Sponsors: UKTI and Simmons & Simmons (http://www.bioinformaticsforumuk.net/?nav=event.tem/HXELPMO0AK)

2 BioSysBio 2007. (Bioinformatics and Systems Biology) Manchester, Jan 2007 (http://www.biosysbio.com/)

Fiona Brown
Northbank Communications,
131–151 Gt Titchfield Street,
London W1W 5BB, UK
f.brown@northbankcommunications.com