

DEALING WITH THE DATA DELUGE

Efficient use of all data is likely to be a discriminating factor between pharmaceutical companies that remain successful and those that don't.

Drug discovery is no longer the low-throughput, empirical science of 25 years ago. Molecular biology, high-throughput screening and computational chemistry have revolutionized the way in which drug discovery programmes are undertaken across the whole industry. Associated with this revolution has been a huge increase in the scale of data that need to be managed, processed and ultimately mined for value.

Against this backdrop, it would be surprising if informatics was not already heavily involved in the science of drug discovery. Whether the pharmaceutical industry in general has fully exploited the capabilities of informatics is perhaps more questionable. What is highly likely, however, is that over the next 5–10 years, efficient use of all data will be a discriminating factor between pharmaceutical companies that remain successful and those that don't. That isn't to say that informatics will replace or supersede existing drug discovery strategies; instead, there will probably be a growing reliance on informatics to improve efficiency, to allow more rapid decision making around existing projects and to more readily identify new ideas and targets. So, the pharmaceutical industry is faced with an increasing need for people with a knowledge of information technology (IT), and therein lies an opportunity for people with such skills.

Mirroring drug discovery in general, the allied informatics approaches are becoming increasingly multidisciplinary, not only in terms of scientific background, but also at a functional level in terms of the specific IT skills required (FIG. 1). Broadly speaking, there are four main areas of expertise in informatics: database administration; data processing and curation; data presentation; and data mining.

Database administration. Raw data — for example, from screening or genomics — are produced either through in-house research programmes or, increasingly, through access to public or proprietary databases. The scale, rate of data production and diversity of data types requires dedicated database designers and administrators to manage efficiently the flow and storage of information. Typically, people in this area will have very strong IT skills, including formal qualifications in database administration.

Data processing and curation. A consequence of the scale and diversity of the raw data is that a degree of processing and curation is increasingly becoming necessary, especially with genomics data. Although much of these data are well integrated, in order to link them efficiently across other disciplines and allow the data to be viewed easily, considerable effort is needed to reprocess them, curate them and to reduce complexity. Here, a knowledge of the relevant science becomes increasingly important, allied with IT skills, which typically include Perl and Unix, and perhaps some database design skills.

Data presentation. Development of mechanisms to view and present the useful data to scientists from a range of

disciplines is becoming increasingly important. Again, scientific knowledge allied with IT knowledge is essential; typically, web-authoring and strong skills with the programming language JAVA are necessary.

Data mining. The ultimate aim of informatics is to capture, curate and present the data in a format that will allow scientists to extract value from the underlying information. Not surprisingly, a scientific background is essential, and data mining is a process that can be undertaken not only by dedicated informaticians, but increasingly also by bench scientists. As the scale and breadth of data continue to grow, there will undoubtedly be an increasing reliance on skilled data miners to extract the maximum value from the underlying information.

Routes to a career in informatics

There are two themes running through all of the above. The first is the increasingly multidisciplinary nature of informatics and drug discovery, the consequence of which is a breadth of opportunities, both in terms of scientific background or IT skills set. The second is the need for some scientific understanding to support the application of the IT knowledge. Only database administrators work almost exclusively within the IT side of an organization; owing to the scientific knowledge involved, other informaticians tend to be integrated within discovery teams. For this reason, an increasingly common route for science graduates into informatics careers within the pharmaceutical industry is through an M.Sc. in either bioinformatics, cheminformatics or a more IT-focused conversion course after a scientific undergraduate degree. In response to the demand for training in informatics, many universities now offer such courses — good starting points for bioinformatics courses and cheminformatics courses are the websites <http://wbiomed.curtin.edu.au/teach/biochem/resources/Bioinformatics.html> and <http://www.indiana.edu/~cheminfo/informatics/cinformacad.html>. With an appropriate scientific background, there are several routes into informatics within the pharmaceutical industry and thereby on to rewarding and interesting careers.

Jerry Lanfear

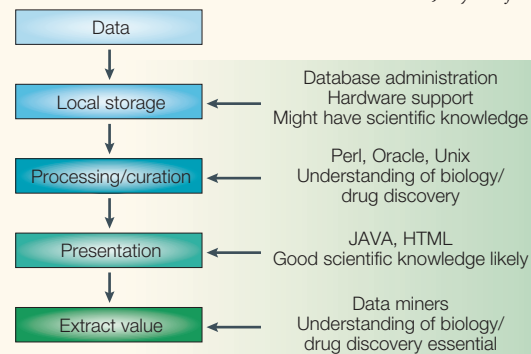


Figure 1 | Data management and associated skills.

Dr Jerry Lanfear, Manager and Head of Bioinformatics, Pfizer Global Research and Development, Ramsgate Road, Sandwich, Kent CT13 9NJ, UK. e-mail: Jerry_Lanfear@sandwich.pfizer.com