target discovery and drug design: extracting the value from genomics

Pharmaceutical research and development is changing. The old model of drug discovery, based on a combination of imprecise candidate generation and broad physiological screens, has given way to more specific and intelligent approaches to target identification and drug design. Now, a vast influx of genomic information is set to revolutionise the range of targets available, at least to those able to navigate effectively through the bewildering immensity of the world's genomic data archives.

By Professor Ken Powell, Deputy Director, Wolfson Institute for Biomedical Research

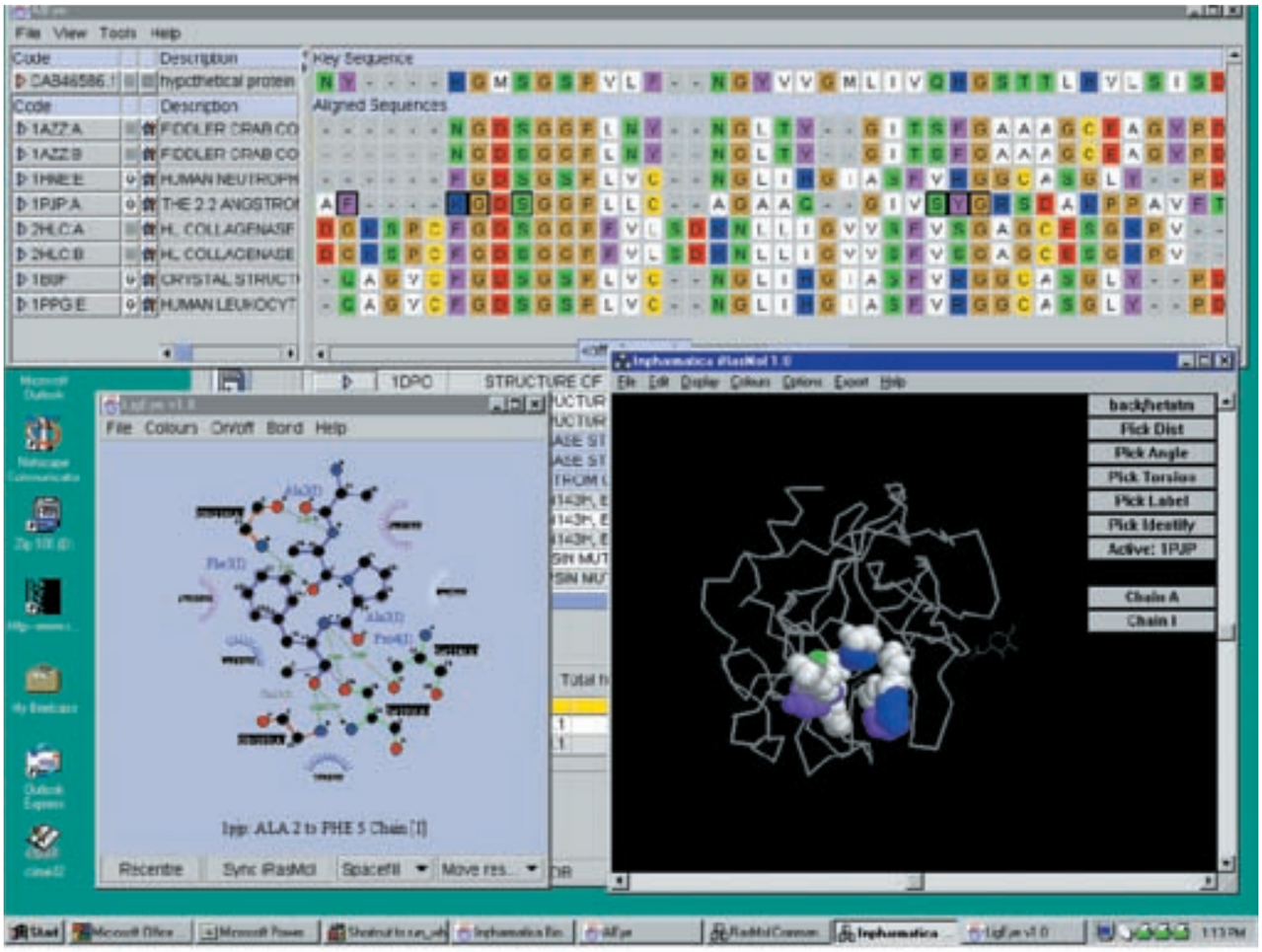
The excitement generated by genomics is in part a reflection of growing difficulties in the pharmaceutical sector. Historically the industry has grown at a rate of about 10% per annum but such figures have become difficult to sustain. In the United States, the patents for 11 of the top 15 best-selling drugs are due to expire by 2005[1]. Increasing cost containment is limiting pricing flexibility, while a more demanding regulatory environment has increased the clinical development costs of new products. It is now generally recognised that a step change in innovation will be necessary if the industry is to successfully introduce the raft of new products required to meet the expectations of financial markets during the next decade.

Genomics seems able to offer a solution – but can it be made to deliver?

The genomics gold-rush

The genomics revolution has been the outstanding development in the biological sciences during the past decade. Throughout the 1990s a rising torrent of genomic sequence data has flooded into the biopharmaceutical world on an almost unimaginable scale. More than half a million protein sequences (mostly predicted from the sequenced DNA) of numerous pharmaceutically important organisms, are already in the public domain. The 'first draft' of the human genome is expected within months. Within two years over 60 microbial genomes, several metazoan species and a robust version of the human genome will have been fully sequenced and their full repertoire of protein products predicted.

Hopes are high that the immense data repository now accumulating will solve the current problems of



the pharmaceutical industry, and consequently genomics has become the arena for major investment and intense competition between the major corporations, accounting for an estimated 10-20% of Big Pharma's R&D budget[2]. This confidence is impressive, but few even of the proponents of genomics have grasped its true meaning for the pharmaceutical world.

For the pharmaceutical industry, the implications of the genomics revolution are in fact difficult to overestimate. The core issue is this: once the full set of human sequence data is in the public domain, every pharmaceutical target will effectively be available to everyone. It has been estimated the human organism expresses some 5,000-10,000 potential drug targets[3], whereas the entirety of current drug therapy is based on fewer than 500 targets. During the next few years there will be a once-only opportunity to identify and patent each of these new targets as they emerge from the sequence of the human genome. Companies that fail to take their share of the spoils risk being left far behind in the drug development race.

Bioinformatics holds the key

The key to exploiting the riches of genomics is bioinformatics. Using appropriate techniques, it is now possible to identify candidate drug targets and even the starting points for potential drugs in silico, by searching through the multitude of gene and protein sequences now accumulating in public databases. It is also possible to study the detailed three-dimensional structure of candidate targets and identify the key structural elements that should be present in any drug raised against a given target. These two tasks – target discovery and drug design – form two consecutive,

Figure 1
From Inpharmatica's Biopendium the workbench – showing a unique alignment viewer, iRasmol structural viewer and Ligplot ligand viewing programmes. Using this facility the molecular biologist or chemist can gain an intimate understanding of the nature of a new target

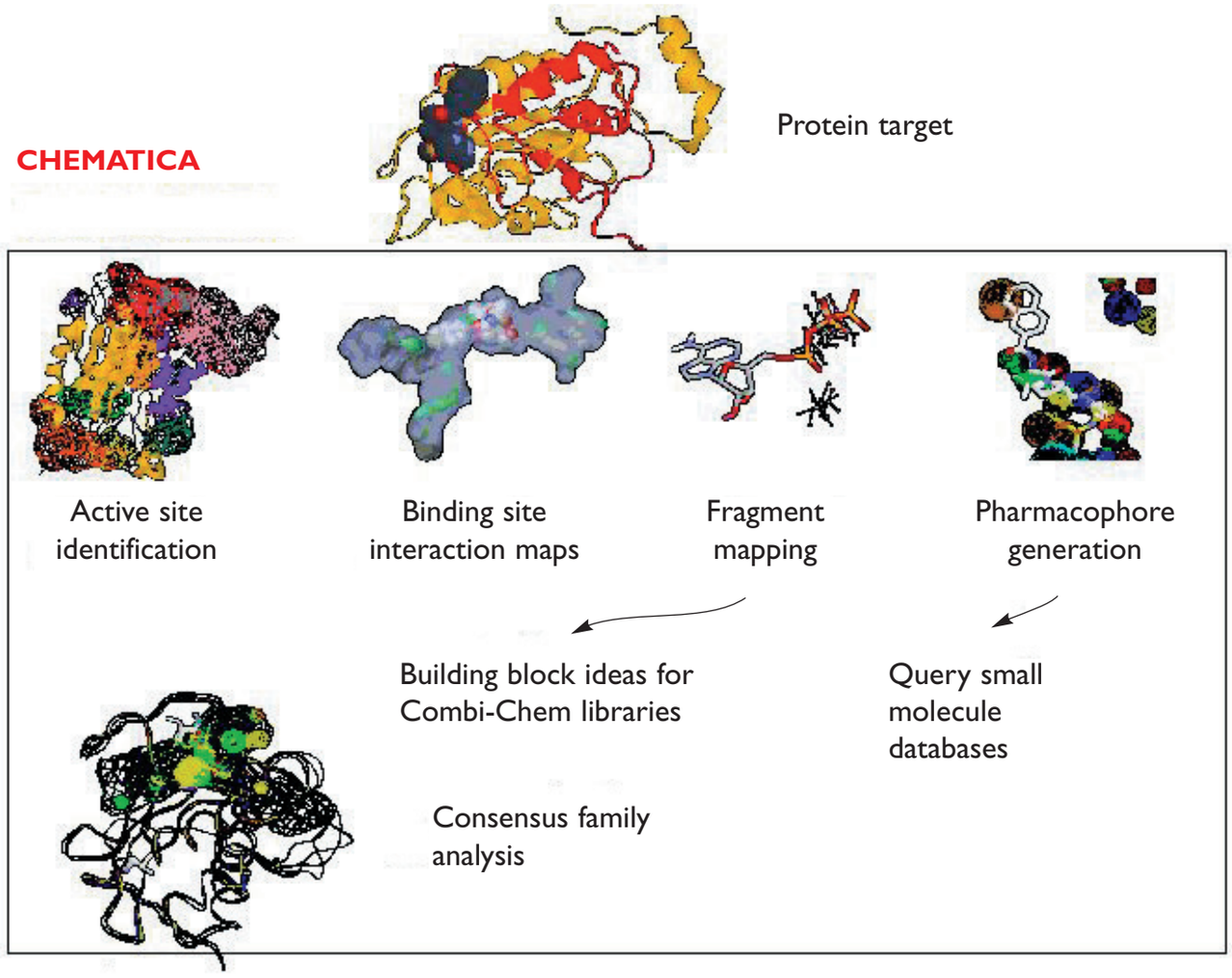


Figure 2
The Chematica suite supports drug design by using proprietary algorithms to map the active sites of drug targets, in addition a knowledge-based potential is applied to predict the precise chemical characteristics that drugs will need to bind effectively. The precalculated binding site information of a consensus family reveals the similarity and differences of the drug-binding requirement

critical stages in the pharmaceutical development process. By supporting the identification of valid targets and the development of drugs with appropriate binding characteristics, bioinformatics has the potential to dramatically improve not only the range of candidate drugs, but also the efficiency of the drug development process.

Finding the targets

High quality targets are the key to more specific and safer drugs, and thereby to improved success in clinical trials. Such targets must be physiologically relevant, representing critical points in a disease pathway. They must also be tractable to drug development, possessing binding sites suitable for drug-like molecules that will predictably modulate their function.

Discovering proteins that are candidate targets within sequence databases is not straightforward. This is because the three-dimensional structure and function of a protein cannot be predicted simply on the basis of its primary amino acid sequence. Unfortunately, primary sequence data is the only information that genomics provides, a limitation which thus far has held back full commercial exploitation of the data now becoming available.

The key breakthrough that has solved this problem stems from the realisation that all proteins are members of a limited number of families sharing essentially similar structures and functions. Thus, if the sequence for an unknown protein can be shown to be related to a known protein, the proteins are likely to belong to the same family and to have similar structural and functional properties. In this way, our knowledge of known proteins can be used to

predict the biological role, pharmaceutical potential and even the three-dimensional structure of novel proteins – provided a match can be found between their sequences.

This simple insight has already driven the development of new classes of drugs. In the development of anti-HIV drugs, for instance, the HIV proteinase was found to resemble half of an aspartic proteinase, a common enzyme class that includes molecules such as renin and pepsin. Since the 3D structures of these related molecules were already known and inhibitors were readily available, the development of new protease inhibitors, which are now integral to the treatment of HIV infection, was significantly accelerated[4].

Turning this simple insight into a method for efficiently searching genomic databases has proved a difficult challenge. Given the enormous size of the databases and the complexity of the sequences they contain, search procedures that are rapid and sensitive enough have proved hard to develop. As search tools have gradually come on line, pharma companies have not been slow to capitalise. For example, the journal *Science* recently reported that several new biologically active proteins found within databases are already entering clinical trials as biotherapeutics. These include a vascular growth factor that may improve circulation, a bone-strengthening factor for treating osteoporosis, a chemokine that may help protect bone marrow during chemotherapy and a stimulator of wound healing[5].

These early successes of data mining were derived from the comparatively small number of sequences that can be found without the need for great analytical power. For the majority of targets and drugs hidden deep within databases, much more sophisticated tools are needed. The greatest challenge in database searching occurs with the numerous proteins that are structurally related even when little or even no similarity remains between them at the level of the amino acid sequence. The three-dimensional structures of proteins are far more highly conserved than their amino acid sequences, consequently, conventional search techniques, which simply compare sequences, are unable to detect relationships and thus may fail to find important targets. Threading techniques which were introduced to database searching by Jones et al in 1992 which combine knowledge of sequences with knowledge of three-dimensional structures, enabling matches to be found between even the most distantly related proteins[6].

The introduction of threading represented a turning point for data mining. Truly powerful database search techniques, which incorporate high-speed threading, have become available just as the Human Genome Project nears completion. The next phase in the competition to discover novel targets will therefore be decisive.

From structure to drug design

Techniques such as threading have moved bioinformatics on to a new plane because they bring together the analysis of sequences and the study of three-dimensional structure. Only through a detailed knowledge of the protein's structure is it possible to ascertain its function, and hence its validity as a potential target. But equally, three dimensional structure is the key to the protein's chemistry and the possibility of designing drugs that will interact with it.

A new raft of computational techniques, falling under the general rubric of cheminformatics, are now emerging which support powerful and efficient structural analysis of targets and intelligent drug design. By analysing the structure of a target in detail, active sites (not necessarily with respect to natural ligands) can be identified and characterised, and small molecule databases queried to identify relevant molecular building blocks as the basis of rationally designed drugs. Using this knowledge, drugs that modulate the target can be engineered intelligently, significantly increasing the probability of a successful lead. Cheminformatics is a younger field even than bioinformatics, but robust, powerful and flexible tools are already being developed to support drug designers working at this level.

Beating the expertise crisis

How can pharma companies best exploit the opportunities now available? Most companies have accumulated proprietary sequence data in areas that interest them specifically, but to derive maximum value from such data it is necessary to probe against as large a database as possible to increase the chances of hitting a relevant homologue. In addition, databases can be mined for novel targets and drugs without the search being based on any existing company data. Most pharma companies will wish to adopt both these strategies, using database searches to interpret their own private sequence data and conduct *de novo* discovery and development programmes.

Given these objectives, what databases should pharma companies use, and how should they mine them? Currently biotech companies such as Incyte (Palo Alto, CA, USA) offer access to subscription databases that contain significant numbers of sequences not yet in the public domain. With each passing month, however, as more sequences become available in public databases such as GenBank and SWISS-PROT, it is the

References

1 Derwent Information
2 PriceWaterhouseCoopers: Pharma 2005, an industrial revolution in R&D.
3 Drews, J. Human disease – from genetic causes to biochemical effects. Blackwell, Berlin, 1997.
4 Toh, H, Miyata, T. Is the AIDS virus recombinant? Nature 1985; 316: 21-22.
5 Wickelgren, I. Mining the genome for drugs. Science 1999; 285: 998-1001.
6 Jones, DT, Taylor, WR, Thornton, JM. A new approach to fold recognition. Nature 1992; 358: 86-89.

ability to search data, rather than access to data per se, that is becoming the decisive determinant of which companies will succeed.

The Achilles heel of pharmaceutical bioinformatics is an acute, world-wide shortage of true expertise. Sophisticated knowledge is required to select databases appropriately, to sift and clean what is in some cases unreliable data, to collate data for analysis from different sources and to use data mining tools effectively. Equally, cheminformatics procedures require not only a detailed knowledge of protein structures and protein-protein interactions, but highly specialised skills in the development and use of structural analysis algorithms.

Consequently, whereas most pharmaceutical companies are rightly attempting to develop in-house expertise, there has also been growing demand for specialist products and resources, leading to the emergence of a thriving informatics sector within the wider biotech industry. Off-the-shelf sequence and structural analysis tools are provided by several companies such as Oxford Molecular Group (Oxford, UK) and InforMax (Maryland, USA). Tripos (St Louis, USA) and Molecular Simulations (San Diego, USA) provide extensive small molecule design software. A number of companies are now offering both tools and processed data to meet specific needs. For example, Lion Bioscience (Heidelberg, Germany) concentrates on providing companies with data on the analysis of DNA and protein sequences at the primary sequence level, while Structural Bioinformatics (Palo Alto, CA, USA) concentrates its expertise on structures that are amenable to 3D modelling.

Inpharmatica (London, UK) has taken the challenge of the expertise gap to its logical conclusion by developing resources that require virtually no specialist bioinformatics expertise, freeing users to concentrate on the biology and chemistry of the problems they are studying. In the Biopendium (**Figure 1**), the world's entire complement of public sequence and structural data have been pooled, integrated and analysed in advance to support target identification, prioritisation and selection. The sequence similarity calculations have been performed using extremely powerful proprietary threading techniques; it is the only database of its kind currently available. Downstream of the Biopendium, a second resource, Chematica, bridges the interface between biology and chemistry in the support of rational drug design (**Figure 2**). The Chematica knowledge base enables Inpharmatica, its partners and subscribers to discover potential lead molecules based on the protein data available in the Biopendium.

Conclusions

Genomics is central to the future of pharmaceutical R&D, and has become a key battleground in the struggle for sustained growth and market leadership. Success will go not simply to those companies with the greatest repertoire of privately held sequences, but, since all human sequences will shortly be in the public domain, to those companies with the greatest ability to mine the value locked within these data archives. Target discovery, structural analysis and intelligent drug design all now depend critically on powerful bioinformatics and cheminformatics. Those companies which recognise this and make best use of the new information-driven paradigm will be the winners in the race to dominate the post-genomic world. DDW

Following a successful academic career in virology including a Fulbright scholarship, ***Professor Powell*** *joined the Wellcome Foundation Ltd as a senior virologist and was quickly promoted to the Head of Antiviral Research and then to Head of (all) Biological Research for the Company. He was a key player in the licensing of the HIV drug Aggrenase, recently launched by GlaxoWellcome. Following the merger of Wellcome with Glaxo he preferred to join an entrepreneurial group setting up a new research institute at University College London with the specific aim of incubating young start-up enterprises. He has been responsible for the nurturing of four start-up companies and has steered them through their initial creation.*